# A SURVEY OF MULTIVARIATE METHODS FOR SYSTEMATICS

Nancy A. Neff

American Museum of Natural History, and
City College of the City University of New York

Leslie F. Marcus

American Museum of Natural History, and
Queens College of the City University of New York

printed at the American Museum of Natural History, New York

supported by National Science Foundation Research Grant DEB 79-17382

copyright © 1980 by Nancy A. Neff and Leslie F. Marcus

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the copyright owners.

Nancy A. Neff
Dept. Vertebrate Paleontology

Leslie F. Marcus
Dept. Invertebrates

American Museum of Natural History
Central Park West at 79th Street
New York, New York 10024 USA

This publication may be cited:

Neff, N. A. and L. F. Marcus. 1980. A Survey of Multivariate Methods for Systematics. New York: privately published

This manual was prepared for distribution at a workshop on Numerical Methods in Systematic Mammalogy, held on 9 June 1980 during the 1980 Annual Meeting of the American Society of Mammalogists in Kingston, Rhode Island.

If additional copies are still available, they may be obtained by sending:

1. a self-addressed mailing label, and
2. a check or money order for $1.00 made out to the American Museum of Natural History, for postage and handling,

to one of the authors at the above address.

## TABLE OF CONTENTS

# P R E F A C E

That the biological world is multivariate has been widely recognized. The importance of quantifying one's observations for analysis and for communication has also been recognized. These facts, plus widespread computer availability has led to the near exponential growth in recent years in the applications of multivariate analyses in systematics and other areas of biology. Multivariate analyses are numerical techniques used to study and describe the covariation between variables, individuals, or both together. In systematics, the variables are usually characters measured on a set of individuals. The goals are varied: some possible purposes include data or dimension reduction, searching for structure, testing the fit of one's data to a model, or discriminating between populations. The strength of multivariate analysis lies in the ability it confers on the user to examine many variates simultaneously and quantitatively.

Multivariate methods have not yet fulfilled all expectations, and will not without further development of the statistical methodology. Demands by the practitioners for more relevant methodology will encourage the statisticians. Also there is much for most of us to learn about the potential of the methods already developed. All of the questions have not been asked and all informative structures and models have not yet been explored.

Our manual, and the American Society of Mammalogists workshop which it accompanied, was conceived in response to the growth and interest in this methodology. We trust that this initial effort will shortly be made obsolete by more complete discussions reflecting the results of continued growth and exploration.

Although the increase in use of numerical methods is sometimes cited as evidence of increasing objectivity in systematics, the use of measurements for data, or mathematical models for hypotheses, does not obviate a continuing need for close observation of the specimens themselves and clear thinking about the biological hypotheses being investigated. Numerical methods are tools only--powerful tools, yes, but only means to an end. We have written this manual in the hope that concentration on how numerical methods can be used in systematics will help the systematists be demanding of the methodology and search for the numerical methods to fit the biological questions, and never vice versa.

We believe that numerical methods, even in their present state of development, have a great potential for usefulness in systematics. The rapid increase in the employment of multivariate studies indicate that others are of this opinion also. As interaction between systematists and statisticians increases, we expect even more useful design of methods, permitting closer tailoring to the biological structures and relationships we study, and fitting more exactly the models and hypotheses which systematists wish to test.

Just as drawing a specimen usually increases the amount one sees on it, so defining and taking measurements can often be instructive because of the time spent thinking hard about and closely looking at the specimens themselves. Unexpected relationships among values or individuals should lead one back to the specimens for a fresh look. In sum, numerical methods are certainly not an alternative to looking at the specimens themselves!

This manual is not about numerical taxonomy. We have discussed methods of multivariate analyses with respect to a diversity of goals in systematics. We have written from a fairly strongly held philosophical position about how scientific studies are most effectively done, but not with a narrow definition of appropriate goals or topics of study. Numerical taxonomy, in our use of the term, includes the philosophical positions well discussed in Sneath and Sokal [1973]; in contrast, we wished to survey methodology in a neutral fashion, divorced from a specific choice of goals. Also, because of our interest in morphology, we have emphasized ordination and descriptive techniques, rather than clustering techniques which are well covered by several books to date.

One final point about our approach may be made explicit: what was included in this survey, and the nature of the discussion, was largely motivated by initially strong feelings on the part of at least one of us about how numerical methods are most wisely used in systematics. The research and writing have not substantially changed the major tenets of our philosophy, although some specific details of belief have certainly changed. However, we have come to realize that this methodology is in fact complex, and difficult to present and discuss from the orientation of biologists or systematists rather than statisticians. We are <u>not</u> attempting to set standards for numerical studies in systematics. We do urge anyone, inexperienced or experienced, to practice a healthy scepticism towards the literature and the recommendations therein, including this text.

# A C K N O W L E D G M E N T S

We have been aided in the preparation of this manual by numerous and diverse people. Dr. F. James Rohlf, Department of Ecology and Evolution, SUNY at Stony Brook, gave generously of his time and expertise. He critically read early drafts of most sections and greatly improved the product. We are grateful for both his criticism and his encouragement.

The National Science Foundation supported the production of this manual, and the American Society of Mammalogists workshop at which it was distributed, through a research grant DEB 79-17382 to the authors.

We thank the American Museum of Natural History, most especially Ms. Diane Menditto, for help with and review of the grant application, and subsequent administration of the grant. The staff of the museum library was, as ever, cheerfully helpful. The Department of Invertebrates, in which one of us (LFM) is a Research Associate, was extremely helpful in supplying workspace and use of a CRT and Diablo terminals; we especially thank Ms. Julia Golden of that department for permitting us incredible license in the use of her office. The Department of Vertebrate Paleontology also supplied space, supplies and support. The Print Shop of the American Museum, in the person of Mr. Vincent Tumillo, performed heroically in printing this manual under too short a time estimate from two inexperienced publishers!

This manual will accompany a workshop on 9 June 1980 at the Annual Meeting of the American Society of Mammalogists. We thank Drs. F. J. Rohlf, M. A. Bogan, D. J. Schmidly, D. A. Schlitter, and G.D. Schnell for agreeing to serve as discussants on a panel at this workshop. We appreciate ASM's interest, and especially the support and help of Dr. J. Mary Taylor. Drs. Duane A. Schlitter and Sydney Anderson provided early comments and encouragement.

We thank the City University of New York, University Computer Center for making available to us text-editing (WYLBUR) and word-processing (Waterloo SCRIPT) capabilities.

# ABBREVIATIONS USED IN THIS TEXT

| | |
|---|---|
| ANOVA | analysis of variance |
| BMDP | Biomedical Computer Programs, P-series |
| CCA | canonical correlation analysis |
| CVA | canonical variates analysis |
| MANOVA | multivariate analysis of variance |
| ML | maximum likelihood (factor analysis) |
| MST | minimum spanning tree |
| NMDS | nonmetric multidimensional scaling |
| NT | numerical taxonomy |
| NTSYS | Numerical Taxonomic System of Multivariate Statistical Programs |
| OTU | operational taxonomic unit |
| PCA | principal components analysis |
| PCORD | principal coordinates analysis |
| PFA | principal factor analysis; PFA(iter) refers to the iterative solution to PFA |
| SAS | Statistical Analysis System |
| SPSS | Statistical Package for the Social Sciences |
| VIF | variance inflation factor |

# SYMBOLS USED IN THIS TEXT

Capital letters usually refer to matrices, and the corresponding letter with subscripts usually refer to elements of that matrix. We have attempted to be consistent in our use of symbols within this text, have not used the same symbol for different entities within the same section of the text, and have tried to continue the same symbol for the same or parallel meanings throughout the text. However, the symbols used for the various matrices and coefficients vary widely among texts and papers; there is no standard usage, with the exception of a very few widely used symbols adopted as extensions from the univariate case.

A original data matrix, usually of order n x m (e.g. m variables measured for n OTUs)

B mean-centered data matrix; $b_{ij} = a_{ij} - a_{\circ j}$ (see page 10 in INTRODUCTION for a further explanation of this notation). Also, the estimates of $\beta$'s in multiple regression or Fourier analysis, or one of the two matrices of coefficients in canonical correlation analysis.

$\beta$ vector of coefficients which are parameters in the regression model or the Fourier model

C standardized data matrix; each element of B (mean-centered data) is divided by the standard deviation for the corresponding variable. Also, one of the two matrices coefficients in canonical correlation analysis.

D n x n matrix of distances among OTUs

$D^2$ Mahalanobis distance squared

E m x r matrix in which the columns correspond to r ($r \leq m$) eigenvectors or sets of loadings; $E_r$ is an m x r matrix of r ($r<m$) retained eigenvectors in a canonical analysis

F n x r matrix of factor scores, $r \leq m$

G association matrix, comprising values of a similarity or distance measure among OTUs

H association matrix which has been simultaneously mean-centered by variables and OTUs

$$h_{ij} = g_{ij} - g_{i\circ} - g_{\circ j} + g_{\circ\circ}$$

L square matrix with eigenvalues on the diagonals and zeros elsewhere; the variance-covariance matrix from the corresponding eigenvectors

Q matrix of correlations among factors

R matrix of correlations among variables

$R^2$ multiple correlation coefficient squared, or coefficient of determination; canonical correlation coefficient squared

S variance-covariance matrix among variables

$S_A$ variance-covariance matrix among groups

$S_W$ pooled variance-covariance matrix within groups

T specially row- and column-scaled matrix in correspondence analysis

U diagonal matrix of uniquenesses in factor analysis

V matrix of residuals or remaining variability, or "error"

W matrix of scores on one set of the canonical variate axes

X one of two groups of variables in partitioned data sets

Y one of two groups of variables in partitioned data sets

Z matrix of scores on the principal component axes, or on one set of the canonical variate axes in canonical correlation analyses

$b_j$ estimated coefficients of the X variables in canonical correlation analysis or multiple regression

$c_j$ estimated coefficients of the Y variables in CCA

$d_{ij}$ Euclidean or Pygathorean distance between individual or OTU i and individual or OTU j

i subscript indicating the row in which a subscripted element occurs

j subscript indicating the column in which the subscripted element may be found

k number of groups into which specimens or OTUs are partitioned

$l_i$ eigenvalue associated with the ith eigenvector

m number of variables in the analysis; restricted to the number of X variables in an analysis partitioned by variables

n number of specimens or OTUs in the analysis; subscripted by group number when restricted to the number of specimens in a group in an analysis partitioned by variables

p number of variables in the Y set in an analysis partitioned by variables; also the probability associated with some event

r number of components retained in a canonical analysis

# I N T R O D U C T I O N

This manual is a survey of multivariate methods of current or potential use in morphometric and systematic mammalogy. It is perhaps more easily characterized by stating what it is not: this manual is not a "cookbook", it is not a "how-to" manual. This review is intended to be neither a comprehensive literature review, nor a detailed mathematical explanation of methods. Rather, for each method, we aimed to provide reference to published discussions which we have found useful and comprehensible, and to briefly summarize pertinent information which appeared to us scattered or not readily accessible. We hope that, as use of multivariate analyses continues to increase, others will be encouraged to expand on this first attempt. We also hope that systematists will be continually more demanding of the methodology, its purveyors, and themselves as its practitioners.

Methods have deliberately received unequal treatment; the amount of detail and discussion is proportional to the complexity of the method and to our judgment of the degree of present or potential usefulness to systematists. Since usefulness depends on both the appropriateness of the method and its availability, some newer or less extensively developed, or otherwise less available methods are not discussed. In each section, the discussion generally follows this outline:

1. In general terms, what does the method do?
2. The model: the structure of the method and how it treats the data.

3. Variations in the model.
4. Terms and nomenclature (in summary form).
5. Comparison with closely related methods (including some discussion of problems and limitations of the method).
6. Applications--(with further discussion of problems and limitations of the method).
7. Computational requirements--what a priori structure is imposed? (assumptions and structure necessary even if not hypothesis-testing).
8. Statistical assumptions--what assumptions become necessary if statistical inferences are made; what statistical tests are available for this method?
9. Biological assumptions--the biological implications of the assumptions made for computation or statistical inference; examples of the assumptions made by typical applications.
10. Statistical packages and computer programs--notes on the method's availability and important options, in the major packages (some discussion of problems and limitations here as well).

The section describing the methods is organized by the structure imposed on the data by the methods--partitioning of variables and number of groups. The following section contains a discussion of the various purposes for which one may employ multivariate analyses. This provides a more comprehensive survey complementing the section on applications within the discussion of each method.

The annotated bibliography contains a list of important journals containing articles on biometrics, major books, and the literature cited in the text. The first appendix contains some suggestions for publication of the results of multivariate studies.

The second appendix surveys some computer packages, discussing briefly the major ones in use in systematics in the United States; this is a more general discussion complementing the specific notes under each method.

Even with the manual in hand and its organization described, the question of how best to use it probably still remains. (We expect most readers to find the wide right margins useful for annotations--comments, further references, questions, argument, and so forth. The purpose of the unbound, 3-hole format is to allow further additions by interleaving sheets in a ring binder.) If readers are experienced in some or all of the methods mentioned here, they may find the discussions of biological assumptions most useful, the discussion of problems or limitations of each method a stimulus and perhaps a source of ideas or argument. Certainly there is no single "right" way to do many of the analyses described here; most methods have many variants. And some potentially useful approaches to examining one's data have probably been neglected. Our goal was to survey the most important or widely used methods, and to discuss some topics which seem to us very important and relevant to many uses of multivariate methods.

The reader with little previous knowledge may have one or more of several possible objectives: one may be questioning whether to learn about these usually difficult methods at all. If so, the discussion later in this introduction should prove relevant, as also should study of some of the fruitful applications cited under each method or listed in the bibliography. One may be trying to match a method to one's problem: in that case, the discussions of assumptions may warrant closest study.

Finally, the reader may already have decided to learn about one or many multivariate methods, and may find this manual useful as an introductory discussion and source of references. BEWARE: This manual will not read like a novel. Most readers will probably not feel that they understand a method even if they read each of our sentences about that method most carefully. As with any other new skill, most people gain an understanding only through extensive practice with the concepts and terminology, both from the literature and from personal experience with applications to more or less realistic problems. From our personal experience in learning about these methods, we feel that the following specific recommendations will be generally useful. Although they may sound obvious or perhaps discouraging, we feel that the overall effect will be a more realistic approach to multivariate analysis, with ultimately less discouragement.

1) Read several discussions, in different references, of the method you are studying, and reread useful ones subsequently. Read for increasing detail and comprehension--don't expect to find everything or even very much clear the first time through.

2) Use a large stack of scratch paper and play with the matrix algebra or with the geometric representation of the algebra or both. Write out term-by-term representations of matrix equations, or work through by hand a very small example, if you would like that much facility with a method.

3) Try a medium-sized data set in a couple of packaged programs. (At this stage, don't expect publishable results! It is perhaps more useful for learning purposes to (re)analyze an already published data set.) Talk with someone--a fellow 'student' or a biometrician--about your results.

4) Decide for yourself how much you need to know about a method in order to use it in your research. The risk of mistakes or misdirected effort clearly decreases with increasing understanding (but neither, of course, can the risk ever be zero). Unfortunately, the literature does not provide an obvious "easy introductory package": in reading about a method, one can easily become bogged down in more statistical and mathematical detail, and more extensions and variations of the method, than is efficient to learn about for a specific biological problem.

Diverse opinions have been voiced about the level of expertise necessary to use any method responsibly. The problem appears to be that adequate review by colleagues of papers with complicated methodology is difficult and time-consuming, with the result that authors cannot depend on critical review prior to publication to catch areas of major confusion in terminology or methodology. The wide availability of "canned" computer programs, especially in the relatively easy to use statistical packages, is widely considered a danger to the quality of numerical research. Corruccini [1975b:14] quotes a series of warnings from biometricians about the hazards of the naive use of multivariate procedures, and goes so far as to suggest that eventually authors may have to be responsible for their own computation and programming. This seems to us an extreme position which would be wasteful of the time and effort required to duplicate (often inefficiently) previous programs. A sense of some of the potential hazards of the methodology, and the various points in the debate about a responsible use of these methods, can be obtained from Kowalski [1972], Corruccini [1975b], and references cited therein.

5) Make use of consultants who complement your

areas of expertise. You may find it useful to work closely with a biometrician or statistician. Make the nature and details of your study clear to the person so that you can learn from them about the behavior of the methods you are using under conditions found or expected in biological data. For example, how powerful are your tests for your sample sizes? What are the assumptions being invoked by the application of the method you have chosen? Are one or more of the assumptions violated, and if so, can you "correct" the situation by transforming your data or choosing another type of test or procedure? Check with someone familiar with the statistical package or program you are using about the accuracy of the operations you are performing. The aim of much of the following discussion is to alert the systematist to hazards and potential problems encountered in using these methods, and to indicate some potential solutions to diagnosed problems.

6) Users at all levels of expertise will find it extremely useful to have a couple of medium-sized data sets with which they are very familiar, for use as benchmarks or standards. When you use a statistical package new to you or a new program (even one you wrote yourself) testing it on some very familiar data will enable you to assess the behavior of the package or program.

A final, editorial comment is required, about whether or not we consider multivariate methods to be (currently or potentially) widely useful in systematic mammalogy. This is mentioned because, in our discussions with fellow systematists, we have thought at times that we could detect a note of defensiveness in the voices of those who, while declining to use

numerical methods themselves, characterized those who did as either fanatics who wish everyone to use such methods, or mediocre systematists hiding behind a front of numbers. On the other hand, among the practitioners of numerical methods, the less experienced seem frequently to expect many more definite answers than the methods can provide (and may sound evangelist while voicing these expectations). And the very experienced often seem to become entranced with the details and mechanics of the methods, and to relegate the biological sources of data, the very questions themselves, to second place. We hope we have avoided most extremes. Although this manual is certainly focussing on numerical methodology, we have tried to discuss applications for each method, and some biological representations of various assumptions and interpretations. Clearly no such discussion can be definitive; rather our goal was to provide examples of such an analysis of assumptions in the expectation that it will be rapidly expanded and improved upon.

There is still the larger question whether numerical methods are useful or not. "Should I use them?" Dare we answer--if it feels good, do it? Outdated, perhaps, by nearly two decades, it still may not be a bad idea in some cases. The primary value of numerical methods is exploratory, heuristic, or communicatory. That is, these analyses are a possible way of learning about your data, seaching for relationships, and summarizing or describing the information in your data to others. The role of hypothesis-testing is actually smaller, and will probably remain so as long as the assumptions required for statistical inference are so often so unrealistic for biological data. An approach intermediate between the usual statistical inference and the purely exploratory use may prove increasingly useful. This is confirmatory analysis, in which one tests the fit of the data to a model structured

specifically for the particular biological problem being studied. Since the exploratory or heuristic uses are rather personal, currently the answer to the question "Should I use numerical methods?" is the answer to "Do you find the method(s) useful for you?" There is a certain amount of positive feedback in this, however, since multivariate methods usually become more useful as the user's skill, sophistication and interest increases.

We certainly are not recommending universal or even necessarily widespread application of multivariate analyses in systematic studies [see Kowalski, 1972, for one discussion of the general usefulness of multivariate analysis]. While the frequency of use is certainly increasing at present, the results may still not justify the effort for some to learn enough such that they feel comfortable using the methods. One accomodation may be to work closely with someone who does enjoy the numerical part.

# BACKGROUND FOR USE OF THIS MANUAL

A working knowledge of only a restricted part of matrix algebra is necessary to follow the vast majority of discussions of multivariate methods. Fortunately, most matrix algebra is relatively simple to understand; the necessary matrix algebra is little more than ordinary algebra done repeatedly in shorthand (but capable of representing some very useful geometric manipulations). The only major difficulty is that it requires some practice and use before one is really at home with it. A reasonable level of understanding can be gained by working through the introductory chapters of some multivariate analysis texts; the first seven chapters of Van de Geer [1971] are especially clear. A more thorough introduction (and one useful for gaining a feel for the geometric representation of various operations) is provided by Green and Carroll [1976]. Other introductions may be found in Jöreskog, Klovan, and Reyment [1976: Chapter 2] and Davis [1973: Chapter 4]. there are also a large number of introductory linear algebra texts (check used college texts) which are not particularly orientated to multivariate analysis, but which go through the fundamentals of matrix algebra with the speed and thoroughness of an undergraduate math course.

For the discussions in this manual, some conventions and a few basic concepts are needed to understand even the interpretive discussions:

The original data matrix, the array of measurements for several variables taken from a sample of individuals or specimens, is always oriented so that the rows correspond to individuals, specimens, or taxa, i.e. Operational Taxonomic Units (OTUs) [Sneath and Sokal, 1973:69], and the columns to variables or characters. The number of rows will be the sample

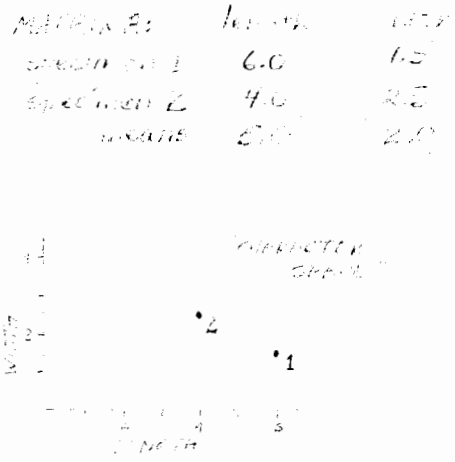
size, n, and the number of columns will be the number of variables, m. Matrices of eigenvectors and matrices of factor loadings will be arranged such that each column is an eigenvector, or a vector of factor loadings. The m rows correspond to the m original variables, and the r ($r \leq m$) columns correspond to the r eigenvectors or sets of coefficients. A capital letter alone refers to the entire matrix: e.g. A for the original data, E for the eigenvectors. Lower case letters refer to an element or subset of elements from a matrix, referenced by subscripts:

A = the n x m original data matrix
$a_{ij}$ = element of A in the ith row and jth column; the measurement of the jth character on the ith specimen
$a_{i\cdot}$ = mean of the ith row; mean of all measurements for an individual
$a_{\cdot j}$ = mean of the jth column; mean of all values for a character.
$a_{\cdot\cdot}$ = grand mean; mean of all $a_{ij}$'s
A' = transpose of A; the rows and columns are exchanged, producing an m x n matrix; necessary for some matrix operations.
E = m x r matrix of eigenvectors, $r \leq m$
$e_{ij}$ = loading of the ith character on the jth component or factor
$e_j$ = the jth column of E; a column vector which is the jth eigenvector
$e'_i$ = the ith row of E; a row vector comprising all the loadings for variable i.

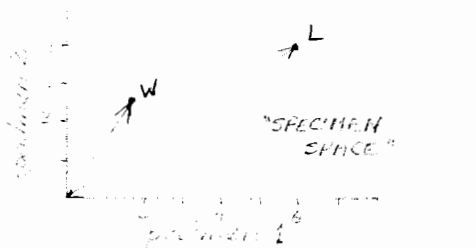
(A table of symbols is given following the Table of Contents and a table of abbreviations.)

Multivariate methods are frequently discussed in terms of geometric transformations. The original data can be considered a cloud of n points in an

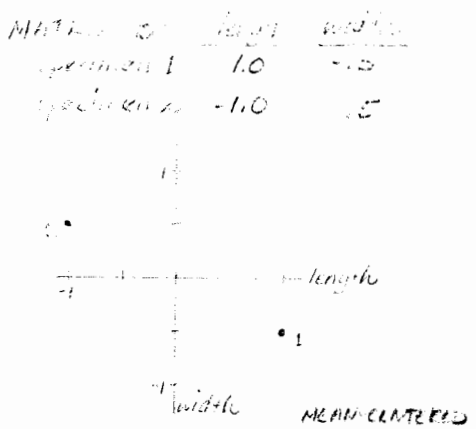
m-dimensional character space, in which each axis represents a character. The coordinates of a point's (individual's) position in this space are its values for each character. Clusters of points would then represent groups of similar specimens.



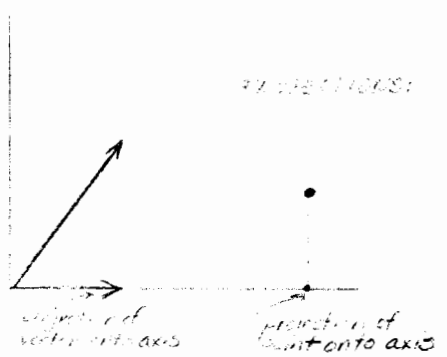
The same data matrix may also be thought of as a cloud of m points, now representing characters, in an n-dimensional specimen space. In this case each axis is representing an individual, and the coordinates of each point (each character) are that character's values for the specimens. (Many people find it more difficult to conceive of this space; it is generally used less in morphometrics than in numerical taxonomy.) One useful feature of this space is that clusters of character vectors (where a vector is a line segment connecting each point with the origin) represent highly correlated characters, or character complexes.



The initial data matrix is frequently <u>mean-centered</u> by variables. Algebraically this means that each value in the data matrix is replaced by its deviation from the mean over all specimens or observations for that variable: $b_{ij}=a_{ij}-a_{\cdot j}$ Geometrically, this is equivalent to moving the cloud of points in character-space so that the mean for each variable is at the origin.



A few other terms, which are used repeatedly, are most easily thought of geometrically. The <u>projection</u> of a point onto an axis (or other line) is the intersection of the axis and a line from the point perpendicular to the axis. The projection of a vector onto a line is another vector ending in the the projection of the point representing the tip of the original vector. The projection of a point onto a plane is, again, the intersection of the line, which both contains the point and is perpendicular to the



plane, with the plane. This concept is generalizable to any dimension. Two vectors, lines, or axes are said to be orthogonal when the projection of one on the other has zero length, i.e. they are mutually perpendicular. Two variables are orthogonal if they are uncorrelated.

The centroid of a cloud of points is that point whose coordinates are the means of the variables for that cloud. In 2- or 3-dimensional space, the centroid is the center of gravity of the cloud of points.

The order of a matrix refers to the size of the matrix in terms of the number of rows and columns. However, some of the information represented by the elements of the matrix may be redundant. There is a somewhat technical shorthand in matrix algebra for dealing with the questions "how many variables do we really have?" or "what is the dimensionality of my data?". The rank of a data matrix (or in fact any matrix) is the number of linearly independent axes required to define completely the coordinates of the cloud of points representing the OTUs. Rank is simple to determine by a number of matrix operations (available in most statistical packages). For example, the rank of a data matrix is reflected in the number of eigenvectors and eigenvalues which can be extracted extracted (see "Principal Components Analysis" in METHODS): the number of eigenvalues greater than zero (when there are no negative eigenvalues) gives the rank of the variance-covariance matrix and hence of the data matrix. The rank of a data matrix is always less than or equal to the number of rows or columns, whichever is smaller. Usually one measures more OTUs than variables; the rank in such cases would be less than or equal to the number of variables. If the rank is less than the total number of variables, then the determinant (a measure of "volume") of the

variance-covariance or correlation matrix (among variables) will be zero, since the group of vectors (from the origin to each data point) will fit in a space of lower dimensionality than that defined by all the variables. Negative determinants do not arise for most multivariate computations as long as there are no missing data.

Mean-centering may have an effect on the rank of a matrix. If there are as many as or fewer specimens than variables ($n \leq m$), and thus the rank is equal to n, mean-centering over variables will reduce the rank by one. (A simple case would be three points in 3-space which lie on a plane not passing through the origin. The rank is three since three dimensions are needed to contain the points and the origin. After mean-centering, however, the origin will lie on the plane containing the three points, and the points plus the origin can be contained in a 2-dimensional space--that plane.)

Matrices of less than full rank result if variables which are invariant (i.e. all OTUs have the same value for the "variable") are inadvertantly left in the data matrix, or if some of the variables are redundant. Redundant variables are those which can be expressed as linear combinations of other variables: for example, if variables are expressed as percentages, such as allele frequencies in genetic polymorphisms, and the percentages of the alleles recorded for each OTU add up to 100%, then there will be redundancy in this data matrix. Other redundancies are less obvious--i.e. characters which are linear functions of other characters, but not obviously so. For example, a complex of characters may be simply linear combinations of a smaller set of characters. If the redundant variables can be identified, then their removal would result in a matrix of full rank, i.e. of rank equal to

the total number of remaining variables.

There are several practical aspects of rank. Even though redundant variables may be measured, they may be measured with error. Additional variability due to error, no matter how small, can give rise to additional small eigenvalues greater than zero, which give the impression of greater rank. One should keep this in mind when trying to determine the "rank" of one's data (see section on "Dimension Reduction" in PURPOSES). Outliers will also tend to inflate rank if they are in dimensions not represented by the mass of the data. Methods for outlier detection (see PURPOSES) will help to find these observations. Van Valen [1974; 1978] has discussed "dimensionality of variation" in a biological context. He has attempted to define dimensionality as a non-integer measure based on the multiple correlation of each variable with all of the others. His measure reduces to the average communality in some forms of factor analysis. It seems to us that a factor analytic approach (not mentioned by Van Valen [1978]) might be more useful.

DESIGN OF STUDY

Assumptions:

One of our primary concerns in this text was to discuss the assumptions associated with some of the numerical methods used in systematic problems. Assumptions are conditions which are taken as axiomatic within the narrowly defined bounds of one's analysis; they are not themselves tested by the analysis. If one were dubious about the truth of a condition assumed in an analysis, one could generally construct another analysis designed to test that condition, which would then be treated as a hypothesis in that new analysis. One does not necessarily believe in the assumptions made in an analysis, since one can do analyses with a more exploratory intent: what results emerge given certain assumptions? However, acceptance of specific results as valid (e.g. that the probability found of achieving that value of a test statistic by chance alone is the true probability) will necessarily imply certain beliefs about the conditions taken as assumptions.

Several assumptions are frequently made in multivariate statistics, especially in hypothesis-testing: multivariate normality, equal variance-covariance matrices of groups, and independence of error terms or residuals, are three common ones. Other assumptions are more specific to certain types of analyses--for example, the assumption in most discriminant analyses that the initial observations have been correctly classified into known a priori groups. Finally, the application of any method to a biological problem requires certain assumptions about the structure of the problem and the nature of the evidence. The interpretation of the results will similarly invoke assumptions; all of these latter assumptions are much more specific to the nature

of the biological questions being asked.

One way to think of the structure of a study, and the division between assumptions and hypotheses, is to consider the number of "equations" and the number of "unknowns". Frequently studies or steps in studies seem to be basically one "equation," which restricts the number of "variables" or "unknowns" which can be solved for to one. Let us take a very simplified example of hybrid identification: in Case 1, one has some mice which have some of the characters of two distinct species of mice, and are clearly not referrable to any single, known species. Perhaps they are hybrids between the two species they resemble. To check whether these putative hybrid mice are really hybrids or not, then one must decide a priori criteria by which such hybrids can be identified as such and apply those criteria to the mice in question. If they fit the criteria, they are hybrids, and if they don't, they aren't; our unknown is the answer to the question "are those mice hybrids?" and the criteria by which hybrid mice are recognized as such are not tested, but rather are known or assumed in this study.

Now, Case 2 would be if one wished to question the accuracy or usefulness of the criteria for identifying hybrid mice; one could not do so with the putative hybrids analyzed in the first study. To try would introduce two "unknowns" into our one equation: if a mouse were identified as not a hybrid then one would not know if this were true or false, since one would not know if the criteria, now also being evaluated, were holding or not. To evaluate the critera, one would have to have mice which one knew, or were willing to assume for this study, were hybrids. Then if, say, 20 out of 100 hybrid mice were identified as not hybrids, one could make some statement about the reliability of those criteria for identifying hybrid

mice.

Another way to state this important point is that, since one cannot evaluate more "unknowns" than the number of "equations" permit, one cannot use the outcome of an analysis to evaluate the validity of an assumption. The rejection of 20 mice of hybrids in Case 2 is not evidence that our assumption that all 100 mice were hybrids was wrong; to reject that assumption we would have to know (i.e. assume) that our criteria were right without fail, and that would put us back into the framework of Case 1. Similarly, if all 100 of the mice we assumed were hybrids in Case 2 were indeed identified as hybrids by some set of criteria we were evaluating, we would not be justified in asserting that such identification supported our assumption that the mice were hybrids. It might be that we were wrong in our assumption, and none of them were in fact hybrids, but that our criteria were so broad that they were all identified as hybrids anyway; we could not tell that this was so from our single analysis. This possibility illustrates the close dependence of the accuracy of the results (here the efficacy of our criteria) and the truth of the assumptions (that those 100 mice were really hybrids).

This discussion may seem too basic to need stating. It is certainly of wider relevance than merely to numerical studies. And biological studies are, of course, usually more complicated than this simple example. One generally asks a series of questions, including questions based on the outcome of initial steps in the study, and sometimes questions about the validity of the conditions taken as assumptions in earlier steps. The structure of each step may be more complicated, in that several unknowns may be estimated if one has a sufficient number of known (or assumed) parameters and relationships. The structure of most

studies will correspondingly be much more difficult to determine.

We are of the opinion that such analysis of "unknowns" and assumptions is very necessary, for several reasons. Most importantly, the thought required to recognize and state assumptions is an excellent way to sort out a problem and evaluate one's evidence. This heuristic value is a major reason to discuss assumptions and other parts of the structure of one's study. Also, no matter how obvious some assumptions may seem, it is useful to state assumptions explicitly. This benefits the reader who may not have considered the problem as thoroughly as the author. Such discussion also will dispel the connotation that assumptions are to be avoided. They are useful and indeed necessary in any study, merely because we cannot study everything at once.

Finally, consideration of assumptions will make clear the dependency of the results on the assumptions. Some methods may be very robust to violations of some assumptions, which means that the results will frequently be right even if the assumptions do not hold for the data set being analyzed. Other assumptions may be quite critical, so that one would distrust all results obtained if the assumptions were shown to be invalid for the data. Practical decisions will thus be affected by an evaluation of the assumptions.

If the assumptions and structure of a study can be stated very explicitly, then they can often be used to formulate a model for a specific analysis. The fit of one's data to the stated model can be used to test how well one's model describes the phenomenon being studied. Such an approach is called confirmatory analysis, and is in varying stages of development for different methods and different problems. It has not

been widely applied in systematics, but has been useful in other fields [for example, in the social sciences: see Timm, 1975, and Mulaik, 1972].

Because of this importance which we attach to the structure of a study, we have included, in the discussion of each method, separate sections on assumptions. "Computational requirements" are those features required in a data set for the method to be applicable--e.g. two or more a priori groups in discriminant analysis. These are conditions which are necessary in order for the method to be used on a data set at all. "Statistical assumptions" are those required for statistical inference; this section includes discussions of the sort of statistical tests associated with each multivariate method. These assumptions are only made when statistical inferences are made, and will not be required in initial stages of exploratory studies. Generalizing from one's sample values (e.g. the values of the principal component coefficients for one's sample) to the population parameters (e.g. the principal component coefficients for the entire population from which one's sample was taken) will generally require at least some of the assumptions discussed under "statistical assumptions".

The third section, "Biological assumptions", is the most difficult, most amorphous, and most important. Here we attempt to cite examples of some biological implications of the use of that method, some biological translations of the computational requirements and the statistical assumptions, and perhaps some common assumptions invoked during interpretation of the results from a typical application. This section will necessarily be much less complete than the others, since biological assumptions are largely specific to individual studies. What we discuss will perhaps serve as an example and stimulus; our own experience suggests

that consideration and presentation of the biological assumptions is interesting and rewarding even though difficult.

Here in the Introduction, we have included a brief discussion of several widely employed statistical assumptions and related topics. The assumption that sampling is random is universally made (section 1). Two of the most commonly invoked statistical assumptions in multivariate analyses are the assumption of a multivariate normal distribution of the data being analyzed and, if more than one group are analyzed, equal covariance structures among the populations sampled. These are discussed in more detail below (sections 2 and 5). If these assumptions are violated, then the estimates of parameters obtained in the analyses may be invalid. (For example, the true distance between two population centroids may be very different from the estimate based on our sample in a discriminant analysis if there is heterogeneity among variance-covariance matrices.) The other possible effect of a violation of the assumptions is that the computed probability of achieving the observed value of a test statistic may be incorrect.

These two potential problems and some possible solutions are discussed in more detail in the following sections, in the following steps: The Central Limit Theorem is frequently invoked to argue that linear combinations of even non-normal variables are more normally distributed (section 3). Raw data which are clearly non-normal can sometimes be transformed to make its distribution more nearly normal (section 4). If the assumption of normality is still violated, or if homogeneity of variances clearly is invalid, then one considers the robustness of the estimators or tests to violations of the assumptions (section 6). Estimators which are not robust may be replaced by estimators

which do give more valid results; these may be computed in various ways, including so-called jackknife techniques (section 7). Tests which are not robust to violations of the assumptions may be replaced by randomization tests (section 8).

1) Independence and random sampling:

All statistical tests depend on the fundamental assumption of random sampling from the populations of interest in a study. This means, for a sample of individuals, that every individual in the population has the same chance of being in the sample as any other, and that the individuals are sampled independently [Snedecor and Cochran, 1967:10]. Independence means that the probability of any individual being sampled does not depend on the probability of any other individual being sampled. An important feature of random sampling is that the researcher has no control over which units appear in the sample. Inferential procedures based on probability can not be used for samples comprising individuals specifically selected as "typical" members of the population. In finite, well defined populations, it is possible to achieve random samples by enumeration of all of the members, and then using a table of random numbers or other randomization procedure to draw the sample. In studies of biological populations, in which the population of interest is hard to define, let alone enumerate, and the sampling procedure (e.g. trapping) may affect the probability of an individual being included in the sample, random sampling is seldom if ever possible.

In practice, exceptions to random sampling will not invalidate the results if the sampling behaves more or less randomly with respect to the test which will be

employed. Also, the generally large amount of variation in nature helps the biologist achieve apparently random sampling more often. However, the validity of this point for any specific example can only be checked by a special investigation [see discussion in Snedecor and Cochran, 1967:12]. As a general practice, the investigator should specifically guard against introducing systematic bias into the sample. (A simple example would be using one kind of trap in one locality and another kind at a second locality in a study of interlocality differences in some sort of mammal.)

An important point is that the inference can only be made about the population sampled (e.g. trappable individuals, if the sampling were by trapping); the extension of estimates or inferences to the members of the population not available must be on the basis of information external to the statistics used. Replication of studies or additional sampling allows us to extend our inferences.

2) Multivariate normality:

Although many measurements do seem to have a normal distribution, there are clearly exceptions to normality in biological data. In addition to binary and other discrete data, there are also situations involving polymorphisms, in which the values for one or more characters will be multimodal or otherwise non-normal. In practise, discrete characters with a substantial number of classes, will frequently approximate a normal distribution well enough to avoid invalid results in spite of the violation of the assumption. The assumption of multivariate normality is more than merely an assumption that each variable is itself normally distributed; it also must be true that the

projections of the data points on any line in the space are normally distributed [Morrison, 1976:90].

Almost all of the tests of hypotheses used in multivariate data analysis are generated from assumptions of multivariate normality; there is a paucity of alternative models for data which are not multivariate normal. Non-parametric multivariate analytic methods have been proposed but they have not yet found practical use for summarizing bodies of data [Gnanadesikan, 1977:137; Puri and Sen, 1971 is a book on non-parametric multivariate statistics for example].

The origin of tests in a multivariate normal framework would seem to imply that if one's data do not follow a multivariate normal distribution, then the results of tests or analyses based on these techniques are not valid. However, some of the tests are robust to normality assumptions (see discussion of robustness below and under specific methods); in other cases the data may be transformed to follow, at least more nearly, a normal distribution. We therefore need tests for multivariate normality. Gnanadesikan [1977:162-195] has an excellent discussion of this topic together with examples and an extensive bibliography. We will only attempt to mention a few of the relevant points.

Since a multivariate normal distribution has the property that all of the individual variables must be normally distributed, as well as all linear combinations of the variables, then univariate tests can be applied to the variables one at a time or to any linear combination. These are discussed in Gnanadesikan [1977:162-168] and include the ordinary goooodness-of-fit chi-square, the Kolmogorov-Smirnov test [Sokal and Rohlf, 1969:571-575], measures of skew and kurtosis, and gap tests. A general approach to

finding transformations which give closer approximations to normality also leads to tests for normality. Various plotting techniques for exposing non-normality are also discussed in Gnanadesikan. Another consequence of multivariate normality is that the regression of any variable on any subset of the others is linear, so that looking for linearity graphically, or testing for linearity also serves as a test for multivariate normality [Cox and Small, 1978]. Also note that transformations which yield symmetrical distributions will satisfy at least one aspect of normality.

All of the univariate tests, or the tests on regressions or on linear combinations of the data, will not guarantee multivariate normality; all these tests are on directions sampled from the multivariate space and the data may still be non-normal in a direction not sampled. However, if any one of these tests indicates non-normality the distribution is non-normal. Many of the multivariate procedures in Gnanadesikan [1977:168-177] are extensions of the univariate procedures. Some of the multivariate tests are large sample tests, time consuming to compute, and approximate at best [Reyment, 1971]. In many situations tests of these types will serve to expose the nature of the underlying distribution of the data; transformations to normality of the variables separately or transformation of various projections may be one way to achieve multivariate normality (section 4) [op. cit.; Dunn, in press]. Multivariate tests for normality are also derived from procedures for finding such transformations.

3) Central Limit Theorem:

As sample size increases, a linear combination of

measurements drawn from a statistical population of any distribution with finite variance will approach a normal distribution; this statement is a paraphrase of the Central Limit Theorem [Morrison, 1976:8-9]. For example, means of measurements from repeated random samples of 30 or more individuals will be close to normally distributed, even though the distribution of the measurements themselves is not normal; the normal distribution of the sample statistic will then permit reasonably correct probabilities to be obtained. In statistical analysis of data which are univariate, the Central Limit Theorem allows us to use the tabled normal distribution to find cutoffs which correspond to chosen significance levels for tests of hypotheses and confidence intervals, even for data which are decidedly not itself normally distributed. Sample sizes of 30 or more are usually sufficient for sample statistics like means to give us reasonably accurate probabilities. (One of us has found through extensive simulation that the distributions are very close to normal for sample sizes on the order of 30 or more.) The actual accuracy depends on the parameters of the true distribution of the data [see Pearson and Please, 1975, for an extensive and useful guide]. For example, tests based on the binomial distribution for frequencies near 0.5 give accurate results for sample sizes less than 30, although if the true proportion is near 0 or 1 larger samples are required. Thus, for univariate statistics the central limit theorem renders procedures based on the normal distribution widely applicable for statistical inference based on moderate to large sample sizes.

There are also multivariate versions of the central limit theorem which give generalizations for the univariate result [Rao, 1965:108]. These are sometimes invoked as rationalizations for use of normal theory in multivariate statistics. Morrison [1976:85] states

"that with the exception of rather pathological cases, the multivariate central-limit theorem would guarantee that the large-sample distributions of test statistics would lead us to similar conclusions about the state of nature" as would be obtained from the development of different sampling distributions for each different probability model for the multivariate data. However we are left with the question of how large is "large". Thus, the central limit theorem indicates that the means of principal component scores, taken from repeated random samplings number of some large number of OTUs from a statistical population, will be normally distributed, even though the measurements for the OTUs are not themselves multivariate normal. The univariate central limit theorem does allow us to test hypotheses on means (estimated from large sample of OTUs) of such multivariately derived scores.

More difficult to deal with is the invocation of the central limit theorem to suggest that principal component scores themselves, or discriminant function scores are normally distributed when the original variables are correlated and decidedly non-normal (e.g. discrete data). This appears to be part of the "lore" of multivariate statistics and we have found it difficult to track down such claims, or to find the theoretical underpinning for such statements. We also have observed that principal component scores or discriminant scores do appear more "normal" than some of the parent variables (obviously true when one or more of the variables is a 0/1 variable).

Consideration of the implications of correlated variables seems to be a useful approach to understanding the problem. Attempting to infer the expected distribution of principal component scores, for example, from the central limit theorem means that the set of measurements for one individual now

corresponds to the "sample" (analogous to the set of measurements from the 30 individuals in the univariate example above), and the score for that individual is the linear combination, analogous to the mean in the univariate case. If the variables are correlated, the measurements on an individual will not constitute a random sample of independent observations. Thus the central limit theorem as usually formulated appears to be inapplicable. Debate about the distribution of scores computed from increasing numbers of variables continues, however; the behavior of such scores has not been demonstrated, and thus a problem remains.

4) Transformations:

One way of dealing with the requirements of multivariate normality assumptions when the data are known or suspected not to be normally distributed is to transform the data to make them more nearly normal. There is a natural resistance to such transformations as their use seems to make the results of an analysis less interpretable to the research worker. This can be a difficulty. What do statistics and tests based on the logarithms of skull length mean? What does the arcsine transformation of allele proportions mean in a geographic study? The statistician has got to be kidding! On the other hand, research workers don't bat an eye at pH (which is the negative logarithm of the hydrogen ion concentration). Logarithms of linear measurements frequently do cause the standard deviations of such measurements to be similar across very differently sized forms, and are also useful for expressing the allometric relation between characters for studies of shape-related size changes within populations. Maybe a logarithmic scale is a "natural" scale after all?

Many of the transformations available for univariate distributions have the property that they make the data more normal (at least more symmetrical) and simultaneously, as in the log transformation of linear measurements, make the variance more homogeneous over samples. Univariate transformations may be applied to each variable with the result that each of their distributions becomes more normal, but they are not sufficient to guarantee multivariate normality. This procedure "seems a harmless pastime at worst" [Dunn, in press]. [See Sokal and Rohlf, 1969:380-387, for examples of univariate transformations for data known not to be normal; Dunn, in press, for an up-to-date review of the topic; Gnanadesikan, 1977:137-150 also gives a summary with examples.] Dunn [ibid.] and Gnanadesikan [ibid.] give general transformations for multivariate data to achieve multivariate normality. These methods are more complicated and are not available in the statistical packages, but must be programmed separately or within the packages. The usual univariate transformations can easily be applied in these packages.

5) Homogeneity of variability:

The statistical tests in discriminant analysis require the assumption of homogeneity of variance-covariance matrices among the groups, as well as the assumption of multivariate normality within each group. The tests are in fact more sensitive to heterogeneity of variability than to violation of the normality assumption. Van Valen [1978] has recently summarized tests for homogeneity of variance for univariate and multivariate populations. The latter involve generalizations of univariate tests. He also discusses multivariate generalizations of the coefficient of variation and comparisons of this

measure among populations. Van Valen says never to use the F-test to test equality of variances. While this admonition is a bit strong, it is true that the test is valid and optimal only if the populations are normally distributed. The 2-sample F-test and its k-sample extension for homogeneity of variance (and its multivariate analogue for homogeneity of variance-covariance matrices), depend on the normality assumptions, and are not robust to violation of this assumption [Pearson and Please, 1975 for a univariate study]. Thus the hypothesis of homogeneity of variance may be incorrectly rejected because of non-normality. Van Valen's warning is partly motivated from observation that samples are frequently too small to test adequately for normality, and therefore not to test for equality of variance is better than to employ a test which gives an unreliable answer. He does offer several alternative, more robust tests, including a jackknife procedure. (See the discussions under "Discriminant Analysis" for the effects of this assumption on tests.) Although one may wish to transform one's data to correct for heterogeneity of variance-covariance matrices in order to satisfy assumptions for certain tests, the heterogeneity itself is of biological interest. What does the difference in covariance structure or character correlations tell us about the several biological populations under study? It is possible that difference in covariance structure may be a better discriminator between populations than a difference in means [Corruccini, 1975b:4-6].

6) Robustness of estimators and tests:

The idea of robustness has been used in two difference senses: 1) for robust tests, and 2) for robust estimates. Robust tests are those for which the probability obtained is approximately correct even

though assumptions used to derive the test do not hold for the particular data set being studied. For example, both univariate t-tests and the analysis of variance are robust to violations of the assumption of normality. On the other hand, the analysis of variance is sensitive to (not robust to) violations of the assumption of equality of variances, especially for unequal sample sizes among groups [Scheffé, 1959:351-358, see especially Table 10.4.1; see Ito, 1969, for multivariate extensions].

The robustness of statistical tests may be investigated theoretically, or by Monte Carlo (sometimes called simulation) studies. In the latter, non-normal or heterogeneous populations are defined and then random samples are drawn. The random sampling is repeated usually a large number of times, and the robustness is measured in terms of the difference between the chosen significance level and the actual probability level obtained. Since there are infinite ways to violate assumptions, it takes cleverly designed Monte Carlo studies to explore violations of interest for potential users. It is also sometimes difficult to make a general statement as the robustness may depend on the nature of the violation as pointed out above for the analysis of variance. A useful example of such a study is that of Pearson and Please [1975] for univariate t-tests for one and two samples and tests for variance and the F-test for the comparison of two variances. The robustness of some multivariate procedures has been investigated and is discussed under the appropriate methods section. It is more difficult to draw general conclusions for multivariate robustness than for univariate robustness, as there are more ways the assumptions can be violated. Monte Carlo studies are scattered in the literature of many disciplines so that it is difficult to pull together all of the latest conclusions and lore. Kim and Mueller [1978b:78] cite

factor analysis simulation studies. Everitt [1979] reports results from Monte Carlo studies of the robustness of Hotelling's $T^2$ test.

The second form of robustness refers to estimates of parameters. Robust estimators are insensitive to the effect of extreme observations or outliers. For example, statistics such as the sample median and interquartile range are robust relative to the mean and standard deviation as estimates for location and variability. In general, robust estimation procedures give smaller weights (or in the examples given above no weight) to observations which are extreme in some sense. Estimators which leave out some of the largest and smallest values are called "trimmed" estimators [Dixon-BMDP:150-151; Dixon and Massey, 1969:331]. For example a trimmed mean might be computed from the data leaving out the 15% of the smallest values and 15% of the largest values. Other methods for robust estimation are discussed in Dixon and Massey [1969:330-332], the BMDP manual [Dixon, 1977:150-151], and Mosteller and Tukey [1977:Chapter 10]. Jackknife estimators (section 7) are said to be robust when compared to standard estimators. Robustness for estimators is more tractable to theoretical investigation than robustness for tests, and is a currently active area of statistical research. However, the results of this research are only just recently becoming available to the users in terms of easily implemented procedures. A use of robust estimates in discriminant analysis is given by Harner [in press].

7) The validity of estimates and jackknife estimators:

The majority of estimates and their associated confidence intervals available in multivariate analyses

are dependent on normality in single group studies, and additionally in multigroup studies, homogeneity of variance-covariance matrices among the groups. If the assumptions are violated, or if the sample size is small, the important question arises: how valid are my results? That is, am I close to the true values for the parameters that I am estimating? Confidence intervals and standard errors are what we usually use to answer this question for univariate statistics. However, they are not available for all of the multivariate statistics covered in this manual. There are, however some other approaches which allow us to obtain an estimation of the validity of the estimates.

Ideally we could apply the estimates from an analysis to a new set of data. An example would be to find the discriminant functions for assigning individuals to groups based on the specimens or OTUs already measured, and then compare the proportion of errors for those original data with the proportion of errors made when using the functions to assign newly collected data whose assignment to groups is also known a priori. The same procedure could be used with principal components: scores computed from coefficients determined from one set of data could be plotted and compared with those scores obtained by using the same coefficients with new data. Frequently, however, all of the specimens that the researcher can accumulate have already been studied and measured. In that case, the same results can be obtained by splitting the sample randomly into two groups. In the pattern recognition literature, these two samples are called the training data set and the test data set, but they can of course be used reciprocally: each set can be use to compute scores and statistics for the other set. This procedure is sometimes called cross-validation [Mosteller and Tukey, 1977:36-40].

The split need not be an even split, especially if a large sample data is available. The logical extreme of assymetrical splitting is a technique, developed by John W. Tukey, called the "jackknife" [Van Valen, 1978; Bissel and Ferguson, 1975]. The sample is split into two groups, one with one OTU and the other with all of the remaining OTUs. To use the jackknife in PCA, for example, to estimate the true coefficients and the true variance for each principal component, a PCA is done for n-1 of the OTUs. Then a different OTU is left out and the PCA is repeated. This is done a total of n times, once for each OTU left out. The jackknife estimates are then weighted averages of the coefficients and variances of the n analyses. Jackknife estimates have been shown to be quite free of distribution assumptions like multivariate normality. More importantly the n sets of statistics generated can be used to compute the standard errors of all of the statistics of interest in the analysis. Jackknifing can be applied to factor analysis, PCA, canonical correlation analysis, and estimating Mahalanobis distance among groups, to name a few applications [see Bissel and Ferguson, 1975, for some limitations]. A jackknife type procedure is available in BMDP for estimating the error probabilities in discriminant analysis.

The main drawback to the use of the jackknife is that computation time is increased by a factor of n. This can be prohibitively expensive for large data sets. However, all possible partitions of one and n-1 need not be run; the theory and estimates are also valid for a random sample of such partitions so that a practically sized subset of partitions may be used. An example of the use of the jackknife in discriminant analysis is given in Mosteller and Tukey [1977:148-162], and simpler univariate applications are discussed in Van Valen [1978] and Bissel and Ferguson

[1975].

8) Randomization tests:

The majority of tests available in multivariate analyses are dependent on normality in single group studies, and additionally in multigroup studies, homogeneity of variance-covariance matrices among groups. Also, some of the tests give correct results only for very large samples. (Such tests are asymptotic, which means that the test statistic approaches the value associated with the correct probability level only as the sample sizes become infinite.) When only small samples are available, or the assumptions are violated by the data in hand, one needs non-parametric tests, i.e. tests for which probabilities can be obtained without the problematic a priori assumptions about distributions. One form of non-parametric test of wide application in univariate and multivariate statistics is Fisher's randomization test [Sokal and Rohlf, 1969:629-637 for discussion of univariate applications; Snedecor and Cochran, 1967:132-134]. The only assumption required is that sampling was random. It is useful to define a randomization test for any test statistic whose distribution is unknown or known only for large samples, or whose distribution depends on assumptions which are not valid or about which the researcher is not sure.

For example if one is testing the null hypothesis that two multivariate data matrices are samples from one statistical population, versus the alternative hypothesis that they represent two different statistical populations for which the mean of one or more of the variables is different, then the following procedure would produce a randomization test. For the $n_1$ and $n_2$ individuals from population one and two

respectively, one defines a combined data array of the $n_1+n_2$ individuals. Random samples of $n_1$ and $n_2$ individuals are drawn, and a value for the chosen test statistic, (for example Mahalanobis $D^2$), is computed for the random partition. This is repeated for all possible samples sized $n_1$ and $n_2$, i.e. all permutations of the $n_1+n_2$ individuals into partitions of $n_1$ and $n_2$ individuals. Sometimes the randomization test is called a "permutation test" for this reason. If the total number of possible permutations is too large, then a reasonably large number of random partitions can be made (usually 1000 or more--the actual number depends on the desired accuracy of the probability determined for the test statistic). Then the proportion of $D^2$ values greater than or equal to the $D^2$ observed for the data partitioned into the original samples will give an estimate of the probability of exceeding the observed $D^2$ under the null hypothesis. If this probability is less than the chosen significance level then one would reject the null hypothesis in favor of the alternative that the populations sampled were in fact different.

There are two difficulties with randomization tests: 1) they take a large amount of computer time in comparison with the more usual tests (based on multivariate normality), although as computers get faster or cheaper, this may not be a limitation. 2) For any given testing situation, the test statistic used is ad hoc, or derived as an analogy to a test statistic (like Mahalanobis $D^2$ suggested above) based on multivariate normality assumptions. This does not affect the correctness of the test, but the statistic chosen may not be the most powerful one. Since the power of a test depends on the distribution of the data in the underlying statistical population, this is a kind of paradox. One can invent "useful statistics" for randomization tests, only limited by one's

imagination and the nature of the alternative hypothesis that one may wish to accept if the test result were to lead to a rejection of the null hypothesis. Randomization tests are nearly as powerful as the optimal normal-based test (i.e. they are efficient) in many circumstances for univariate statistical tests of hypotheses; therefore the main drawback to their use would be the amount of computer time required and the lack of procedures in the popular statistical packages.

There is an increased interest in randomization tests at this time; a new journal devoted to their use has recently been started [Good, 1979]. Sokal and Rohlf [1969:704-706] provide a FORTRAN computer program for two-sample univariate randomization tests. They are not difficult to program in FORTRAN or within statistical packages such as SAS.

Sample sizes:

One of the most common questions asked by a research worker is how large a sample do I need for the study I am about to undertake? This is one of the hardest questions to answer because it depends on so many things, including some only known when the study is complete. There are, however, guidelines for sample size adequacy in terms of some needs, such as how precisely one needs to estimate a parameter, or what size difference one is looking for when tests of hypotheses are proposed for comparing two or more populations sampled. In general, rules and guidelines are better known for univariate statistics than for multivariate statistics, and they do again depend on the probability distribution assumed for the data. For most statistics the precision, usually measured by the

standard error of the statistic (the standard error of the mean is perhaps the best known), is inversely proportional to the square root of the sample size. In other words as the sample size goes up the standard error gets smaller, or the precision increases--but it is a square root rule. In order to make the standard error one half of the value obtained for a sample of n observations, 4n observations (or an additional 3n) are required. For a specific example, the confidence interval length for a principal component's variance (i.e. the corresponding eigenvalue) may be deduced to be a function of the reciprocal of the square root of the sample size from its formula [Morrison, 1976:294]. This is however an asymptotic confidence interval, which means that the assumed distribution and associated probabilities only become correct with fairly large sample sizes (really only in the limit as the sample size becomes infinite).

In tests of hypotheses for several groups of observations, the sample size required depends on the magnitude of differences one wants to detect using significance tests. In this case the sample size required (in order to detect that difference one is interested in) is a function of the significance level and the power of the test for finding the specified difference [see Dixon and Massey, 1969:264-279, for univariate procedures for determining power for commonly used tests]. The power of a statistical test, for a specified alternative hypothesis, is the probability of rejecting the null hypothesis when the specified alternative is true; i.e. how often, when the alternative is really true, does one say it is true? For example, the null hypothesis that the percentage of males in a population is 50% against the alternative hypothesis that it is not 50% may be tested using the binomial distribution for the number of males in random samples from the population. The power, for a specific

sample size and significance level, can be computed for a specific alternative, for example that the proportion is 40%. If the difference between the value under the null hypothesis and the true value is actually larger, then the power will be larger, and if it is smaller then the power will be smaller. The power will be 1 or near 1 for very large differences, and equal to the significance level chosen when the null hypothesis of no difference is true (for most two-tailed tests). And for any given difference, the power will be greater for increasing sample sizes. Very small differences (even biologically insignificant differences) can be found to be significant for very large sample sizes.

In multivariate statistics the situation is more complicated as sample size determination depends on the number of variables, the sample size in each group, and the number of groups under study. There are a number of "rules of thumb" and much lore on the subject of sample size, but unfortunately little documentation for most such rules. We have usually chosen not to give the rules of thumb unless they are documented by studies or statistical theory. We will cite what we can under each method where we know of a source for the study or rule.

Many of the tests used in multivariate statistics are asymptotic while a few are exact. An exact test is one in which the reported significance level is obtained from some known probability distribution of the test statistic based on the assumptions from which the test was derived, regardless of the sample size. For example, a test for no difference in joint means of two populations for any number of variables is exact: the F table may be used to determine rejection values for observed F (assuming multivariate normality and equality of variance-covariance matrices). Asymptotic tests are test statistics whose distributions are known

only as the sample size becomes infinite; however, they may be good approximations for even relatively small sample sizes. Chi-square tests on proportions, based on the normal approximation to the binomial, are asymptotic tests, even though the approximation may give two-place accuracy to significance levels for samples as small as n=10. Adjustments frequently improve the approximation (Yates correction in this example), and some of the funny looking functions of m, n, and k in the formulae for asymptotic test statistics (such as some of Bartlett's tests) are such corrections.

Missing data:

Almost all of the multivariate techniques have been developed in terms of the simplifying assumption of a complete data matrix, i.e. a value of each variable is available for each individual OTU measured. The usual derivations and computations of multivariate techniques employ matrix algebra, in which operations are not defined for incomplete matrices. Defining new operations which reasonably accomodate missing values is difficult conceptually, since each specimen can be considered a vector in a hyperspace defined by axes representing each character, but can not be placed relative to characters (axes) for which a value is missing. Estimations of missing data estimate the specimen's position on an axis for which one does not have a value, so that the specimen's position could be defined in the total space containing the data. Missing data are also a problem in univariate statistics (for example in the analysis of variance) where partitions of variance are no longer additive and independent, and therefore tests are harder to interpret.

In biological studies it is common for data or measurements to be missing as a result of loss, breakage or, in the case of fossils, lack of preservation of parts of the specimens available. Data may also be missing as a result of comparing characters which are not present in some populations or OTUs (e.g. a comparison of antlers between sexes for some species of cervids). These "no comparisons" cannot be logically estimated in the same fashion used for missing values from incomplete specimens. No comparisons are not a problem for statistical tests or confidence intervals: we know the populations are distinct for those characters. Tests are only relevant for characters which are measurable or have some variability within the various populations. No comparisons are important in numerical taxonomy and cluster analysis [see Sneath and Sokal, 1973:178-182, for a full discussion].

The effect of missing data on results of analyses and tests of hypotheses has been an active reseach area for statisticians [Afifi and Elsahoff, 1969; Beale and Little, 1975; Dempster, Laird and Rubin, 1976]. Dempster [1969:260-264] and Morrison [1976:120-124] both treat simpler examples of missing or incomplete data. They both stress the importance of determining whether the fact that the data are missing is related to some property of the OTU. For example, if total length of a limb bone could not be measured because juvenile epiphyses were not fused, the very fact the individual is a juvenile might invalidate some of the missing data procedures; preferably a procedure developed from complete juveniles should be used. Pertinent studies on missing data are discussed under each method where appropriate, however some general approaches are available especially in BMDP(77 and later).

The BMDP manual [Dixon, 1977:333-336; also see Frane, 1976] has a clear discussion of many of the simpler options available for detecting patterns of and correcting for missing data. These include: 1) deletion of cases with missing data, 2) using the data available for the variables and pairs of variables for computing the variance-covariance matrix or other association measures, and 3) substitutions of various estimates for missing data in the original data array prior to the main multivariate analysis. A variety of methods for displaying the pattern of missing data and relating the pattern to age or sex, or other variables are available in BMDP procedures. They can also be used for replacing the missing values by one of a number of estimation procedures. Procedures 1) and 2) above are available in all of the commonly used statistical packages.

IMPLEMENTATION

Computers have played a major role in the recent increase in use of multivariate analysis. An understanding of the nature of that role is important in the ability to critically evaluate applications of multivariate techniques. The methods we will discuss can all be done "by hand" with pencil and paper, given enough time and enough paper, or a small enough data set. In other words, humans can and some do know exactly what the computer is doing to the data, since the computer is acting in response to programming by humans. (For example, when it prints out error messages in response to part of your program--a response which you certainly didn't tell it to make--the computer is acting under the control of a series of programs which tell it how to deal with the program you submitted.) The purpose of emphasizing the mechanical nature of computers is to emphasize their irrelevance to the structure, intent, and underlying theory of the study. For example, the use of a computer to calculate the principal components of a data matrix is as irrelevant to the results of the study as whether one calculated the mean of three numbers using a pocket calculator, or pencil and paper, or did it in one's head. If the correct formulae are used, if the data are entered and operated on correctly by each device, then the results will be the same except for possible rounding errors.

Errors and accuracy:

The complexity of multivariate methods, however, means that there are many opportunities for error in the use of a statistical program or package by a systematist, and also in the programs which constitute the statistical package. It is far easier using a

computer than using pen and paper to get results from a multivariate analysis without knowing what the analysis has actually done to the data, which means that the errors are less easily detected by the systematist. Computers can make mistakes, but they are very rare compared to operator or programmer (i.e. human) error.

Even the best and most widely used packages and programs have been known to produce incorrect or inaccurate results. The programmers and creators of the packages do their best to remove all "bugs" and use up-to-date computational procedures (algorithms) but mistakes do crop up from time to time. Users can protect themselves against such problems by having some "benchmark" data sets against which to test the programs. Small textbook examples will do for a start, but medium-sized data sets with which one is thoroughly familiar are better. Run the data set using the intended program, and better yet compare the results from more than one program for a given analysis. Discuss any possible errors or disagreements with your local computer center consultant or statistician. The creators of the packages want to know about errors in order to correct them and provide a better product.

Accuracy of computed results is another consideration in using computer packages and programs. Contrary to popular belief, computers are not necessarily accurate. The IBM 360-370 series of computers for example, only give about 7 significant digits when used in single precision mode (the default for FORTRAN programs). Other computers may give the same or more accuracy; some of the microcomputers or minicomputers give less. The accuracy of computation is a function of both the architecture of the computer and program used (part of the software). Matrix computations, used a great deal in multivariate analysis, are complex and repetitive, and rounding

error, when insufficient digits are used, may give answers that are very inaccurate. Some of the packages use double precision which gives about 14 digits of accuracy, and therefore less round-off error. Check the accuracy of the package or program you are using. In FORTRAN for example DOUBLE PRECISION is a program option.

Iteration:

Many of the computational procedures used to find solutions in multivariate statistics are iterative; that is the statistics are computed by successive approximations rather than by a direct algebraic solution. This method is used for one of two possible reasons: 1) for some problems, an iterative solution is the only one known, or 2) an iterative solution may be preferred in a computer program, even though a direct algebraic solution may be known. Those methods which include the computation of eigenvectors and eigenvalues are all iterative: principal components, canonical variates analysis and canonical correlation depend on this type of iteration. Most forms of factor analysis might be thought of as doubly iterative, since communalities must be repeatedly estimated through eigenvalue and eigenvector determinations which are themselves iterative. Some of the most time-consuming iterative techniques are those involved in certain cluster analyses in numerical taxonomy and tree or network analyses in phylogenetic systematics. In the latter case a large number of possible networks are possible even for relatively few OTUs, although there may be a single or only a few optimal solutions.

Considerable energy has been expended in formulating efficient algorithms for finding optimal solutions in iterative methods which do not automatically converge

to the correct solution. Most such algorithms are so-called "hill climbing" algorithms. They are trying to maximize some optimality criterion for which there are one or more optimal solutions or "summits" and a number of suboptimal "hilltops" or local optima. It is sometimes difficult to know whether the algorithm has reached a local optimum or one of the true "summits". New starts (corresponding to randomizing the order of presentation of the data to the computation routine) will protect against this difficulty, but there is usually no absolute guarantee that the optimal solution has in fact been obtained. Nonmetric multidimensional scaling is an example in which this type of iteration is used.

The user is sometimes free to choose the maximum number of iterations in statistical routines. Defaults are usually available and one might start with these and experimentally change the values to see the effect on the solution. Often a criterion used to judge the number of iterations required is the difference between values from successive iterations: if the absolute value of the difference is very small, increasing the number of iterations may use much computing time and will not change the solution very much. Some statistical routines allow the choice of such a difference interval.

Interpretation and publication:

The methodology discussed here has as its ultimate purpose the making of some aspect or structure of the data intelligible, whether by exploring for previously unseen structure, or by testing the fit of a model, or by finding the probability for the truth of a statistical hypothesis. Its ultimate value, then, lies in the extent to which interpretable results are

produced.

We heartily agree with Kruskal and Wish [1978:12]: "The process of interpreting the configuration is the central step in many applications, and is best learned by active participation." Simple generalizations about how to interpret one's data are not available. There is clearly a personal component here, in that one benefits from a flexibility of mind to be able to see the unexpected, plus a willingness to interact with the methodology. The degree of sophistication and experience with the methods will also play a role; an understanding of how the numbers and plots are obtained from the initial data matrix is clearly central to an ability to interpret the results.

The final step, publication, is still not far removed from the topic of this manual. We discuss the diversity within general methods (such as the varieties of factor analysis or discriminant analysis) with the partial intent of pointing out details which should be specified in a description of any application. Not only will several different methods be referred to in different references or programs by one name (or a confusing variety of names), but various details of the specific model may differ among programs (or different releases of the same program), or be options within a single program. The difficult task of specifying precisely which method was employed, and in the model those details which are subject to variability, is crucial for the useful publication of the results of one's analyses. A careful reader will need such details in order to digest one's study. An appendix to this manual contains a list of suggestions about publication of the results of multivariate analysis.

# M E T H O D S

We have organized the discussions of methods in this section according to the structure of the data matrix. The two dimensions of the structure considered here are the partitioning of variables into one or more sets, and the partitioning of items or specimens into one or more groups. Our approach is actually an oversimplification of some of the considerations necessary in the overall design of the study. The structure of the data should correspond to the structure of the problem being studied, as should the structure of the method one will use to analyze that problem and those data. Putting the question bluntly, is the way you are tackling the problem going to produce an answer relevant to your questions? This match between problem and method is in practice one of the most difficult aspects of multivariate studies.

Problem structure was discussed at some length in the INTRODUCTION (pages 15-19), in terms of the role of assumptions in a study. Other considerations are discussed by Dempster [1971:322-324] in his treatment of "epistemic structure". The logical structure of the data set itself, although an extremely important subject, is not discussed in our survey; it is assumed that the reader is familiar with some aspects of this topic, such as the distinction between the different kinds of response variables possible (e.g. dichotomous, meristic, continuous, etc.). We refer the reader to discussions such as Dempster [1971:318-322] or Gnent [1979] for important further considerations of data structure and acquisition.

The structure of the model in the method employed

should match the problem structure. The data analyst, to be able to make this match, must understand the structure of the available methods. Therefore, in our review of each method, we have included more description of the specific mechanics of the method than might at first seem warranted for a non-mathematical audience. The information we give, however, is that which seems to us necessary (although perhaps not yet sufficient!) for an understanding of the methods, and which is frequently not readily available to the biologist. The information is largely contained in the first part of each discussion. (See pages 1-2 of the INTRODUCTION for the general outline followed under each method). The shorter treatments follow this outline less explicitly; the longer discussions use some of these topics as subheadings.)

The sequence in which methods are considered in this section follows this order:

1) description of within-sample variation, with no partitioning of the variables.
2) description of within-sample variation, with the variables partitioned into two or more sets.
3) description of among-sample variation.

The second and third categories will be shown to be related, in that if one uses one set of variables as indicator variable(s) in a type (2) analysis to designate group membership, the result is equivalent to an analysis of among-sample variation. To some extent, an understanding of methods later in the sequence will be facilitated by a reading of the methods preceding them.

The distinction between categories (1) and (3), within-group analysis and among-group analysis, lies in whether or not the method takes into account within-group variation when applied to a multigroup

problem. Thus, principal components analysis or nonmetric multidimensional scaling may be applied to sample centroids (or a collection of vectors of character states for OTUs), but the method treats the data as a homogeneous set of observations and does not incorporate any analysis of within-group variation. Discriminant analyses and MANOVA, in contrast, weight the between-group descriptors by functions of within-group variation. But even this distinction is not entirely hard and fast, in that linear discriminant analysis (often called canonical variates analysis) is equivalent to performing a principal components analysis (a type (1) analysis) on group centroids after the data (or character space) have been transformed as a function of the pooled variance-covariance matrix, which transforms the common 95% ellipses within groups to circles [Rempe and Weber, 1972]. However, it is clear that in this case the within-group variation is indeed part of the analysis since it is that variation which determines the transformation of the space prior to the analysis on the group centroids, which puts the entire analysis in category (3).

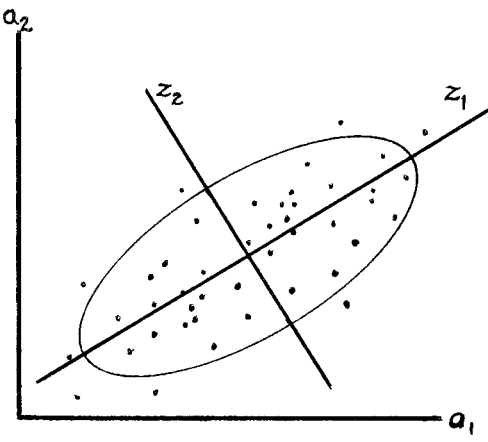
Up to this point, we have not discussed any distinction between univariate and multivariate analyses, or whether multi-variable methods such as contingency table or multiple regression analyses are considered part of multivariate analysis proper or not. Our diffidence reflects the lack of unanimity in the field. We are using "multivariate" in this text to refer to any method or problem which involves more than one random variable. Our philosophical orientation is that multivariate analyses are those used in the study of covariation and interrelationships among variables, and associations among individuals. Even when the design of a study includes more than one fixed or controlled variable (i.e. design variables), such a study is to us still univariate in its focus on one

variable of interest, one substantive variable in Dempster's terminology. Although we have not used the term so broadly (for practical reasons, in view of what is usually meant by 'multivariate'), we appreciate Dempster's point [1971:317]: "A preferable viewpoint would be to start with ordinary 'univariate' data as the simplest case of multivariate data--relating one substantive variable (like weight) to an indexing variable (labelling the animals weighed)--and to place no rigid limits on the varieties of data types to be called multivariate." [Dempster, 1971:317].

The most important point about 'multivariate' vs. 'univariate' is that there is no single right answer, only preferences and noticeable variation in textbooks and references. ANOVA is generally considered a univariate technique but its model is often written in its most general form in matrix notation and is a special case of multiple regression, which is frequently viewed as a multivariate technique. (We discuss two models for multiple regression, one of which is univariate by our definition, and the other multivariate. Both are considered here since they constitute a more complicated topic less frequently considered in introductory texts.) While more typical univariate procedures could be fitted into the categories given for the following discussion, it did not seem efficient considering the large number of good introductory statistics texts available. Therefore, the following discussion assumes some familiarity with univariate statistics, including ANOVA, as well as some general background in statistical inference.

# PRINCIPAL COMPONENTS ANALYSIS

If a data set is visualized as a cloud(s) or scatter of points (each point representing an individual) in an m-dimensional character space (each axis representing a variable), then principal components analysis (PCA) is a method that finds a new set of orthogonal (mutually perpendicular) axes in the directions of greatest variance among individuals. The first axis is a line in the direction through the cloud such that the projections of the individuals onto the line have maximum variance (or equivalently, such that the sum of squares of the distances between each point and the line is minimized). This is equivalent to finding the major axis of the best fitting ellipsoid determined from the cloud of points (e.g. a multidimensional football shape). The second axis is in the direction of greatest variance perpendicular to the first axis. Successive axes are found until all of the variance is accounted for. These axes are the principal axes, or the principal component axes. The bivariate case is pictured to the right: $z_1$ and $z_2$ are the new axes.



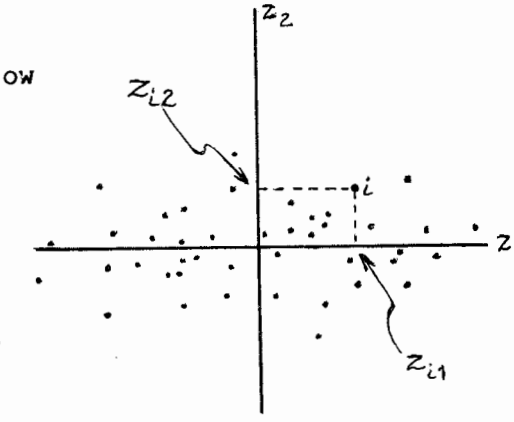

The first step in PCA is usually to mean-center the raw data set, A; this produces a matrix B. From B is determined the variance-covariance matrix, S. The eigenvectors of S are vectors in the directions of the principal axes (orthogonal axes in the directions of maximum variance among the individuals in B), and when appropriately scaled, the eigenvectors are also the sets of coefficients that are used in the equations for the principal components. (Descriptions of how to compute eigenvectors are found in most texts; particularly good descriptions are in Van de Geer [1971:62-74] and Jöreskog, Klovan, and Reyment [1976:63-78].) The first principal component is a linear compound of the original observations ($b_{ij}$'s) weighted by the elements of the first eigenvector

($e_{i1}$'s):

$$Z_{i1} = e_{11}b_{i1} + e_{21}b_{i2} + \ldots + e_{m1}b_{im} \quad \{1\}$$

Thus the scalar product of an eigenvector times the row of B (with elements $b_{i1}$ through $b_{im}$ for m variables) corresponding to a given individual i produces the score $z_{ij}$ of the ith individual on the jth principal component axis in the new coordinate system. ($z_{i1}$ on the first principal component is the example in equation {1} above.) The collection of eigenvectors for a data set may be thought of as a matrix E with as many columns or vectors as the rank of the matrices B and S, usually m vectors (if n>m) for real data. The matrix E as a whole is also sometimes referred to as the weighting matrix or coefficient matrix, because it contains all of the weights or coefficients for each original or mean-centered variable used to determine the position of each observation on each of the new axes. Each score results from multiplying each eigenvector times each observation vector, as in equation {1}. The entire set of scores ("Z" below) is then the entire set of coordinates of the original individuals in the new coordinate system defined by the principal component axes:

$$(Z = B * E) \quad \{2\}$$



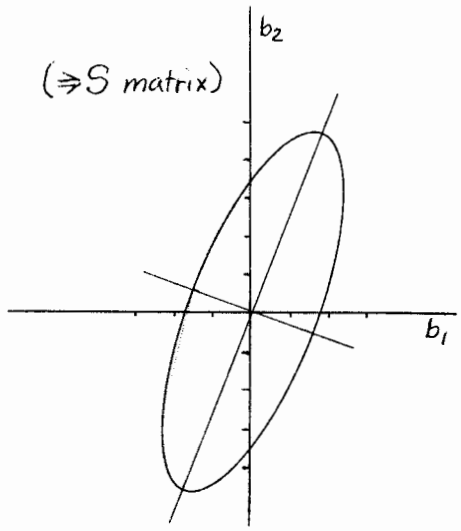

*Principal component scores of point i.*

The amount of the total variance among the observations attributable to or explained by each component is given by the eigenvalue associated with that eigenvector corresponding to the component. The sum of all the eigenvalues is equal to the sum of all the variances of the original variables. The eigenvalues are ordered such that the biggest eigenvalue is associated with the first eigenvector, since the first component axis is in the direction of maximum variance among the individuals. (The

eigenvalues are sometimes arranged as the diagonal elements in an m x m matrix, with off-diagonal elements equal to zero. This is the form of a variance-covariance matrix, with the eigenvalues as the variances of each new variable or component, and the covariances equal to zero since the principal components are uncorrelated.)

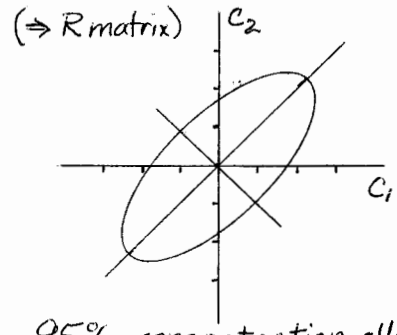
Alternative pictures or results arise from possible variations in the procedure just outlined. Let's consider first that one can also calculate the principal components of a data matrix (matrix A) that has not only been mean-centered (matrix B), but also standardized (producing matrix C) so that each variable has standard deviation 1.0 . In this case, the eigenvectors that are found are those of the correlation matrix, R. The resulting principal components will not be the same as the principal components resulting from the eigenvectors of a variance-covariance matrix, since the division by the variables' standard deviations (the standard deviation of the columns of B) in the standardization of the data is, in geometric terms, a shearing transformation rather than a simple rotation or translation. Thus, the positions of the points relative to each other are changed. (The transformation from B to the correlation matrix is changing the scale of each axis so that the transformed variables (before PCA) all have standard deviation or variance 1.0 .) The important conclusion is that PCA on C, using the correlation matrix, vs. PCA on B, using the variance-covariance matrix, results in solutions which are different, frequently very different, and which cannot easily be transformed from one to the other by any simple scaling; thus PCA is not invariant under such rescaling of the variables.

There is also variation in the scaling of the eigenvectors, and in the preferences held for different



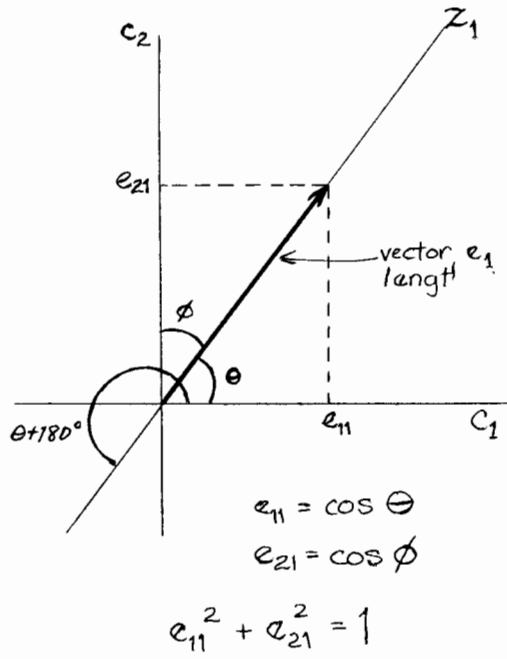

95% concentration ellipse of mean-centered data

correlation between $b_1$ and $b_2$ = .7



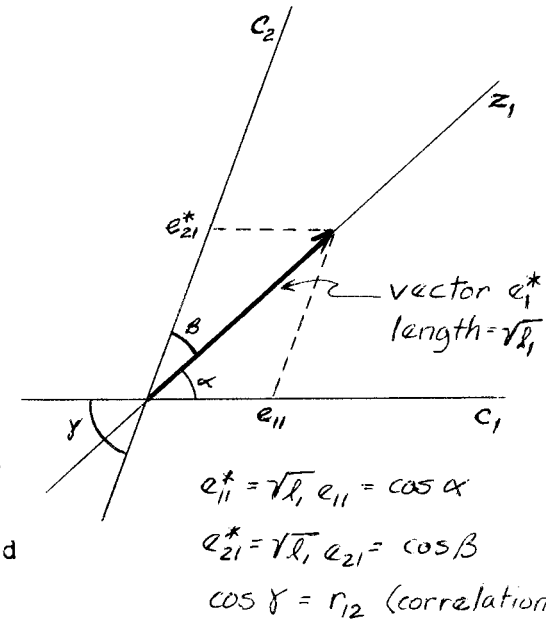

95% concentration ellipse of the same data after standardization; correlation between $c_1$ and $c_2$ still .7

methods of scaling. The coefficient matrix E, which is usually used to calculate the scores on components, is scaled such that the sums of squares of its rows (corresponding to the original variables) are 1.0 and similarly the sums of squares of its columns (corresponding to the components) are 1.0 . (This means that each eigenvector so scaled or normalized has length 1.0 .) When the coefficients of the components are expressed in this fashion, each element squared represents the proportion of the component's variance that is accounted for by a particular one of the original variables. The elements of the eigenvectors scaled in this fashion are frequently called weights; they are the coefficients $e_{i1}$ in the equation for calculating the score of each observation (represented by the vector of measurements $b_i$) on the first component, shown above. The elements of an eigenvector, scaled to have length 1.0, are also the cosines of the angles between the principal component axis corresponding to that eigenvector and the axes representing the variables. And they are the regression coefficients of the component on the original variables. The eigenvector is a unit vector on the corresponding principal component axis when plotted in the original variable space. Whether the elements of an eigenvector are positive or negative is in one sense arbitrary: the entire eigenvector can be multiplied by a -1.0, and will still be a unit vector along the same axis, but in the opposite direction. (This follows from the relationship: $\cos(\theta + 180^\circ) = -\cos\theta$.)




The elements of the eigenvectors are sometimes scaled in a fashion like that more frequently employed in factor analysis: the elements of an eigenvector scaled to have length 1.0 are multiplied by the square root of the corresponding eigenvalue, and each element is divided by the standard deviation of the

corresponding original variable. (Tne step of dividing by the standard deviation of the variables is unnecessary when C has been analyzed, since in that matrix the standardized variables have standard deviation 1.0 .) The result is that each vector is so scaled to have a length equal to the square root of the corresponding eigenvalue. The elements of these vectors are sometimes called coefficients also, but are more often termed loadings, or component correlations, since each loading is the correlation between an original variable and a component. They are again cosines of angles, in this case between the component axes and the original variable axes in a coordinate system constructed so that the angles between the variable axes equal the correlations between those variables. And each element squared is equal to the proportion or percent of the variance of each original variable that is accounted for by a particular component. This view can be very useful for interpretation, since the situation where variation in a particular variable is largely accounted for by one of the later components (with a small eigenvalue) would be more apparent; these later components are frequently examined in less detail or entirely ignored since they explain a small amount of the <u>total</u> variance, but may still contain information of importance with respect to a few variables or observations. (Morrison [1976:273-274] briefly discusses how to deal with this situation.)




If only some subset of the principal components with associated eigenvalues greater than zero is retained, one may then ask: what is left of the variance-covariance matrix S, the correlation matrix R, or even the mean-centered data matrix B, after the portion that is explained by the retained components is subtracted? Our earlier statement of the model for principal components analysis, that the scores are the

product of the original observations times the eigenvectors (Z=B*E), can be equally well written by expressing the original data in terms of the principal component scores and coefficients: B=Z*E'. If only the r largest components are retained (and the original matrix was of some rank greater than r), then the scores on those r components and the r sets of coefficients will not completely predict the original data. The differences between the original data and the amounts explained by the r components are called residuals (the matrix labeled "V" in equation {3}):

$$B = Z_r E_r' + V \qquad \{3\}$$

Similarly, the variance-covariance matrix of the original data equals the part of the variance-covariance matrix explained by the components plus the residual variances and covariances.

$$S = E_r L E_r' + U \qquad \{4\}$$

The matrix L is the r x r diagonal matrix of r eigenvalues retained. Both sets of residuals, that set from the variance-covariance matrix and that set from the original data matrix, may be examined for large values or structure among the values. Both U and V will have rank less than or equal to m-r, since r dimensions have been explained by the r components.

In summary at this point, it may be useful to consider the terms that have been used in this discussion of principal components analysis. The literature discussing PCA (including texts in multivariate statistics) shows great variation and considerable inconsistency in the use of some terms. First, let us reiterate that a principal component of a data matrix is a new variable produced from the linear combination of the original variables using the elements of the corresponding eigenvector (usually

scaled to have length 1.0) as the coefficients. The eigenvalue is a single value for each eigenvector, equal to the variance of the observations in the direction of the corresponding principal component axis. (The elements of an eigenvector are not eigenvalues.) Eigenvectors and eigenvalues are also called latent vectors and latent roots, or characteristic vectors and characteristic roots, respectively.

When the equation for a principal component is solved using the measurements for an individual (usually the mean-centered measurement from B, but not always), one obtains the score of that individual on that component. The axes on which the scores of the individuals are plotted are not the principal components themselves, but rather they are each an axis (or basis vector) providing a scale against which the scores for a component are plotted. (This is a fine but important distinction, and perhaps more obvious when expressed in terms of two of the original variables: each axis in a bivariate plot, of "length (L) by width (W)", for example, is not the variable length or width itself, but is the axis or scale against which the values of the variable in the data set are plotted.) The equation for a principal component (as {1} above) is not the equation of the principal component axis, but is instead the equation of an m-dimensional hyperspace since there are m+1 variables in the equation (the m original variables and the score).

There are two ways to find the equation (in terms of the original variables) of a principal component axis. One is to use the fact that, in any system of perpendicular axes, any axis is defined as that line for which the values of all the remaining variables are zero. (For example, in a bivariate case with axes for

X and Y, the Y-axis is the line X=0.) In component terms, each axis is a line of points with score of zero for all other components. So one solves a set of m-1 simultaneous equations: in this case, a solution of any m-1 principal components set equal to zero is the equation of the axis corresponding to the remaining component. This is more easily said than done, however, so we are fortunate that there is a computationally simpler method. Since the cosine of an angle between a vector and an axis is equal to the projection of the vector, scaled to have length one, on that axis, the elements of each eigenvector, being such cosines, give us the coordinates of a point on the corresponding principal component axis in the original variable space. Since each component also goes through the origin of mean-centered data, we have the coordinates of two points, which determine a line--the principal component axis. This is illustrated for a bivariate case in the figure to the right.

The elements of the eigenvectors suffer the most variation in terminology. One can find examples of almost any term--weight, coefficient, loading, and others--being used for the elements of eigenvectors scaled in a variety of ways. We would like to make two points:

(1) The term coefficient or regression coefficient seems most logically applied to an element of an eigenvector scaled to have length one, since such values are usually the coefficients in the equation for a principal component. The scores thus produced have a variance for each component equal to the corresponding eigenvalue. If one is performing principal components analysis within a factor analytic framework, then a set of coefficients, often called factor-score coefficients, may also be produced, which are instead the elements of the eigenvector scaled to have a length

equal to the inverse of the square root of the corresponding eigenvalue. The resulting scores will have a variance equal to 1.0 for each component. The term correlation coefficient refers to an element that has been multiplied by the ratio of the square root of the variance of the component (i.e. the square root of the eigenvalue) to the standard deviation of the original variable. Loading is frequently used rather ambiguously to refer to an element of an eigenvector regardless of scaling. In practice, "loading" most often refers to a correlation coefficient. Correlation coefficient is the term for (and the scaling of) eigenvector elements most often used in factor analysis; correlation coefficients (and therefore usually loadings) are not used as coefficients in the principal components equations.

(2) No matter what term or other scaling is used, it helps to state explicitly in publications the scaling employed if values of "coefficients" or "loadings" are discussed.

Much of the variation in the terms for the elements of the eigenvectors, and for the entire matrix E scaled in different ways, arise from the disagreements in the literature over whether or not PCA is a type of factor analysis. We think that it is most useful to distinguish between PCA and factor analysis because the latter has rather different goals and diverse underlying models, and is therefore much more complicated. PCA is unfortunately treated as part of factor analysis in many of the major statistical packages, with the result that most of the terminology is factor analytic, and more complicated and confusing than necessary. In spite of this orientation in the packages, we feel that it is heuristically useful to distinguish clearly between principal components analysis and factor analysis, and to avoid further

ambiguous use of terms. PCA and factor analysis are contrasted in some detail in the section on "Factor Analysis".

Applications:

Principal components analysis is widely used as a dimension-reducing technique, to summarize as much of the information (variation) in the data as possible in a few dimensions [Gnanadesikan, 1977:7-15; Rao, 1964]. A large series of measurements, for example, might be suspected of being highly redundant, but of containing information about a few important "factors" (size perhaps, and some proportional/shape features, for example). In this case, the first two or three principal components might explain a large proportion of the variation in the data set, and the remaining components might be considered measurement error or 'uninteresting' individual variation or both. If the data points were entirely contained within a subspace (of dimensionality, r, lower than the original number of variables, m), then the last m-r eigenvalues would be zero since the first r eigenvectors would explain all the variation in the data. (A 3-dimensional example would be if all the data points, each individual given by three coordinates, lay on a plane somewhere in 3-space, then PCA would find two axes on that plane, and the eigenvalue for a third axis would be zero since there is no variation in that direction, perpendicular to the plane.) As an aspect of the use of PCA to summarize data, the scores of the individuals on the first two or three components can be plotted, which can then be visually examined more easily than can all the numerous possible plots of the original data [Pizzimenti, 1975; Hoffman, Koeppl and Nadler, 1979; Best, 1978 with MST; Kennedy and Schnell, 1978 for a three-dimensional perspective drawing]. The

relationship between the original variables and the principal components can be visualized by projecting a vector (usually scaled to have length equal to the standard deviation of the variable) for each variable onto the plot of the scores [for example, see Jolicoeur and Mosimann, 1960: Fig. 5]. PCA can also be used to reduce the number of variables prior to further analysis; the scores on the chosen components would be used as the new pieces of data.

PCA is also a useful exploration technique: It provides convenient axes on which to plot the data for examination of the interrelationships among individuals. This may enable the detection of outliers among the data, either errors or truly aberrant individuals [Gnanadesikan and Kettenring, 1972]. (Such plots might not warrant publication in themselves, but their production is frequently a profitable step in the analysis.)

However, here we repeat the caveat we made above in a different fashion: the investigator should consider carefully what kinds of structure she/he is looking for or is interested in, since there are possible situations in which some interesting or significant structure will be expressed only in later components. One possibility is one or two variables not correlated with any of the others, and therefore expressed only in a later component; information in those variables would not be included in subsequent analyses based only on the first few components. Another possibility is one or a few individuals being distinctly deviant from the others in a direction that does not account for much of the total variation; such individuals would not appear "out of place", or separated from the others, in plots of only the first few components. One way to check for interesting structure or information in later components is to calculate residuals; these

are discussed briefly in Cooley and Lohnes [1971:104] and more extensively in Gnanadesikan [1977:260-263].

The scores of individuals on the principal component axes can be examined in several ways in addition to the usual scatter plots. The scores are sometimes used in t-tests, for example, or may be plotted on maps [e.g. Menozzi, Piazza and Cavalli-Sforza, 1978; Kennedy and Schnell, 1978]. See Thorpe [1976] for references to analysis of geographic variation using PCA. Carleton and Eshelman [1979: Fig. 17] plot scores for two components against time for fossil rodents.

The eigenvectors provide information about the dependencies among the measured variables, which may lead to the recognition of functional complexes. The most common such interpretation is the identification of a component with eigenvector coefficients all of the same sign and very nearly the same size as a "size component" [e.g. Bryant and Turner, 1978:764; see Blackith and Reyment, 1971:147-153 for discussion and further examples]. When several coefficients are much larger than the others (and usually of opposite sign) the component is said to be summarizing shape as expressed in the contrasts between those highly weighted characters. [Jolicoeur and Mosimann, 1960; Morrison, 1976:286-289.]

One aspect of interpretation in PCA may be worth mentioning, if only to stimulate a little argument: after the first component, there is no reason to expect the principal components to be a particularly meaningful combination of characters, and even the significance of the first could be questioned. Biologically interpretable combinations would not necessarily be expected to be orthogonal; this is the rationale for oblique rotations in factor analysis (see that section). The value of PCA is that it provides

the two- or three-dimensional space that contains as much as possible of the information (variation) in the total data. The 'interesting' directions within this space would not necessarily be expected to be parallel to one of the components, but should instead be visually assessed. For example, the size directions (not parallel between groups) and the directions of discrimination among groups, were not parallel to any principal component axis in a fairly typical data set for two parental species and their hybrids [Neff and Smith, 1979].

Computational requirements:

None. [Gower, 1966a:326.]

Statistical assumptions:

PCA is not primarily a hypothesis-testing procedure, but instead a data analysis method for displaying interrelationships among individuals or variables. Because of this intention, there have not been a large number of tests devised to use with the results of PCA. Even if the method is used descriptively, however, it is important to remember that the eigenvectors and axes that are found are those for that set of specimens measured. If the same variables were measured on another sample of individuals from the same population, the eigenvectors would be slightly different. In other words, there is sampling error, which limits the extent to which one can make precise statements about the eigenvectors of a population R or S matrix for a given set of variables based on the eigenvectors of a sample R or S matrix.

Hypotheses can be formulated, however, about the

components found, and there have been some tests devised. If the eigenvalues for two components are exactly the same, then the distribution of the data in the plane defined by those two components forms a circle. Intuitively, one can immediately see that for a circle, the major vs. minor axis is indeterminant: any two axes equally well describe the variation present. Therefore, the particular orientation that resulted would be a random event, and the associated eigenvectors without particular significance. It is extremely unlikely that any two or more eigenvalues for real data will be numerically precisely the same, but they may be quite similar. A test for sphericity can be used to test the null hypothesis that any two or more eigenvalues are the same, or even that the correlation matrix is equal to the identity matrix--i.e. all the correlations between pairs of variables are zero. If the latter is found to be true, then PCA is not applicable, since the original variables are already uncorrelated [See Cooley and Lohnes, 1971:102-103]. The test for the equality of any two or more eigenvalues permits one to determine the meaningful principal axes [See Morrison, 1976:294-295]. Cooley and Lohnes [1971:105] describe a test for the null hypothesis that the determinant of the population residual matrix (after a certain number of components have been extracted) is zero, which would indicate that subsequent components should not be used. Other possible inferential procedures include confidence intervals for the eigenvalues, and a test of the hypothesis that a particular eigenvector is equal to some specified vector (which can be specified to allow a test of a hypothesis of an allometric relationship among others). [See Morrison, 1976:292-299.]

Valid use of these tests requires the assumption that the data are multivariately normally distributed

(which implies that the data are homogeneous, i.e., not a mixture of sources of variation, such as, in biological terms, both sexes, more than one species, etc.). Many of the tests also assume full rank for the data, since in those tests the inverse of the coefficient matrix E is used, but will be undefined if E is less than rank m. Some of the tests (including all those discussed in Morrison [1976]) require large sample sizes, since the probabilities for the test statistics are true as n approaches infinity. Harris [1975:175] recommends that the sample size be such that the difference between n and the number of variables is greater than 30: n-m>30. (We do not know the basis for this recommendation however.) Fewer tests are available for PCA results calculated from the correlation matrix than from the variance-covariance matrix.

Gnanadesikan [1977:203-207] discusses various graphical techniques for examining the entire set of eigenvalues from an analysis, but notes the dearth of formal inferential procedures in this area.

Biological assumptions:

The use of PCA, as described above, to look for relationships among variables is limited to looking for linear relationships since the principal components are themselves linear combinations of the original variables. Linear relationships are not always to be expected, however (especially if variables are measured on different scales--e.g. angles, linear measurements, weights, percentages, or ratios). If one has some reason to expect a specific sort of nonlinear relationship, then an appropriate transformation of the original data might be useful (such as logarithms of lengths and weights). Another, little explored

possibility is non-linear or generalized PCA for finding the non-linear coordinate system most closely in agreement with the data. [Gnanadesikan 1977:48-62.]

An implicit assumption frequently made in the use of PCA is that the important biological phenomena will be represented most clearly by the components in the directions of greatest variance, i.e. the first few components. Inferences about the entire data set are frequently based on the first few components, especially when they explain a very large proportion of the total variance. This assumption can be tested by examination of the loadings of the variables and the scores of the observations on all the principal components.

The assumption is frequently made when PCA is used descriptively that the results of a PCA on a sample can be taken as representative of the directions of variation and interrelationships among variables found in the population. The extent to which this is valid depends, of course, on the size of the sample and the degree to which it is representative of the population.

When the data are known to represent more than one group, and are therefore heterogeneous, the additional step is sometimes taken of comparing patterns of variation among the groups. The implicit assumption is sometimes made during such interpretation that the directions of greatest variance are the same for all groups, or the same for each group as for the data taken as a whole (see also Applications under "Discriminant Analysis"). The validity of this assumption can be examined by PCA of each group separately, or some subset of the groups, and a comparison of the size of the eigenvalues and the loadings on each eigenvector.

The choice made between the correlation and the variance-covariance matrix implies an assumption about the variance in the original data. When PCA is done on the variance-covariance matrix, those variables with high variance in the data are given greater weight in the analysis. The correlation matrix results from standardizing the deviations from the mean, i.e. dividing each deviation from the mean by the standard deviation of the variable, so the resulting values have a mean of zero and a standard deviation of one for each variable. To use the variance-covariance matrix is then to attribute importance or meaning to the different amounts of variation in the variables. One should remember that the relative amounts of variation are at least partly a function of scaling; if a variety of scales were employed (millimeters, grams, and meristic counts, for example), the correlation matrix is generally preferable. If all measurements were made in one unit, in millimeters for example, and the variation in shape were being examined, then the variance-covariance matrix may be more suitable. Performing a log-transformation of the data prior to PCA on the variance-covariance matrix would "correct" for difference in variance related to large differences in the mean value of variables measured on the same scale.

Interpretation of the results of PCA frequently involves an assumption that variables that are highly correlated are in some fashion functionally linked.

If PCA is used to evaluate interrelationships among individuals or groups of individuals, the assumption is often made that the relevant groups are indeed present. For example, in analysis of putative hybrids, it is frequently assumed that the possible parental species have all been identified and included in the analysis. (See Neff and Smith [1979] for a discussion of the

assumptions required by PCA in hybrid analysis.)

As mentioned previously, the assumption of multivariate normality required for appropriate statistical inference translates into a variety of biological assumptions, such as lack of sexual dimorphism, presence in the sample of only one species or morph, and so forth. Such homogeneity is not usually assumed in a descriptive or exploratory use of PCA in which statistical tests are not employed. Even in a descriptive use of PCA, however, the method is often sensitive to the inclusion of one or two characters with very few states (0/1 coded, for example) in a data set comprising largely continuously distributed characters. The result may be a component almost entirely devoted to the two- or three-state character; this may produce two or three clusters in the plots of the scores [Neff and Smith, 1979:186,192].

Statistical packages and computer programs:

MIDAS has a convenient PCA routine, under that designation.

SAS allows a PCA using the correlation matrix in PROC FACTOR. Scores can be produced using PROC SCORE, and plotted using PROC PRINT. PROC MATRIX of SAS permits the most flexibility, since the user can program PCA in a few lines, and exercise much greater control over variations in the method and subsequent manipulations of various results.

SPSS will allow PCA, also within the procedure FACTOR. Analysis is limited to correlation matrices. Scores can be computed for observations with missing data; various options are available.

BMDP offers perhaps the most versatile of the generally available "canned" PCA procedures, as part of the factor analysis procedure. Correlation or variance-covariance matrices may be used, as well as other matrices. The data may be mean-centered or not, and cases may be differentially weighted. Distance statistics useful for detection of outliers are available. (This is one form in which residuals can be reported.) One can also use coefficients from one set of data to compute scores using another set of data. Various plots may be obtained using options within the procedure.

NTSYS has PCA as an option in its procedure FACTOR for both correlation and variance-covariance matrices. Principal component scores are computed using a procedure PROJECT and plotted using procedure MXPLOT.

If one has more variables than individuals or observations in the original data matrix ($m > n$), principal coordinates analysis may be used as a computational shortcut. (See the next section on "Principal Coordinates Analysis".) However, most principal coordinates analysis programs do not produce the loadings of the original variables, which may be desired as output.

# PRINCIPAL COORDINATES ANALYSIS

Principal coordinates analysis (PCORD) produces coordinates on axes in Euclidean space for a set of individuals or OTUs from data that consist of associations between all pairs of OTUs in the set. This association matrix from which the principal coordinates are calculated may be a matrix of inter-OTU distances, distance squared, other dissimilarity measures, or any of a wide variety of similarity coefficients [see Sneath and Sokal, 1973:248-249].

Starting with an initial matrix which represents either inter-OTU similarities or differences, PCORD produces scores on coordinate axes that summarize the inter-OTU distances. We will discuss first a principal coordinates analysis where one is given initially the Pythagorean or Euclidean distance. In this matrix, each off-diagonal element $d_{ij}$ represents the Pythagorean distance between individual OTUs i and j, and the diagonal elements, $d_{ii}$, the distance from an OTU to itself, is zero. (Note that the $d_{ij}$ are not divided by $m^{-1/2}$ as in the "average taxonomic distance" measure frequently used in numerical taxonomy [Sneath and Sokal, 1973:124]; however, the latter distances are proportional to $d_{ij}$.)

The distance matrix is transformed to an association matrix G where each:

$$g_{ij} = -.5d_{ij}^2 \qquad \{1\}$$

The diagonal elements of G, the $g_{ii}$, are still zero, so the sum of the diagonal elements, the trace of G, is zero. Therefore, if one found the eigenvalues and eigenvectors of this G matrix, the sum of the eigenvalues would be zero, which means that some positive and some negative eigenvalues would be produced. To insure that this G matrix has all eigenvalues greater than or equal to zero, G is

transformed to H using the relation:

$$h_{ij}=g_{ij}-g_{i\cdot}-g_{\cdot j}+g_{\cdot\cdot} \qquad \{2\}$$

The effect of this transformation is to preserve the interpoint distances as they were in G, but to express the position of the points in terms of the lengths of their vectors from the origin and the angles between those vectors. (The $g_{i\cdot}$ and $g_{\cdot j}$ are the means of the rows and of the columns, respectively, of G, and $g_{\cdot\cdot}$ is the grand mean of all of the $g_{ij}$'s.) Each diagonal element $h_{ii}$ of H now represents the squared distance of OTU i from the centroid of the data. The $h_{ij}$ when divided by the square roots of the corresponding $h_{ii}$ and $h_{jj}$ represent the cosine of the angle between OTU i and j.

The principal coordinates are the eigenvectors of H scaled to have lengths equal to the square root of their respective eigenvalues. The n elements of the first eigenvector are themselves the scores for each OTU on the first axis; the squared distances between scores (or projections) on this axis contribute most to the sum of all squared interpoint distances--all $d_{ij}^2$ in the original distance matrix. The scores for the OTUs on the second axis, given by the second eigenvector of H, make the second largest contribution orthogonal to the first axis, and so on. The scores on orthogonal axes are uncorrelated. The number of axes obtained will be less than or equal to n-1. If the sum of squared distances can be well represented in relatively few dimensions, then relatively few axes contribute a relatively large amount to the sum of squared distances. The proportion of the total of the squared distances explained by one or more principal coordinates will be the ratio of the sum of their corresponding eigenvalues to the sum of all of the eigenvalues.

As was stated before, the initial data for a

principal coordinates analysis may be a matrix of any sort of inter-OTU distance measure or similarity coefficient. Association indices--distances or similarities--are often calculated from an OTU by character data matrix, but may also be generated directly (without ever measuring separate characters for each OTU), such as in the measurement of immunological distances. Since principal coordinates are calculated without reference to any original variables or measures, principal coordinates analysis can be performed in these cases where only association matrices are available. One should note that if the initial matrix is a distance or difference matrix, it must be transformed to a similarity matrix, G, using equation (1) above. G is in turn mean-centered, producing H. The addition of a constant to G does not change the subsequent matrix H (see equation (2) above), and therefore does not change the interpoint distances.

The result of PCORD of an association matrix is a set of coordinates for the OTUs such that the distance squared between any two OTUs in this new coordinate system is a function of the value of the association between them in the original data set:

$$d_{ij}^2 = g_{ii} + g_{jj} - 2g_{ij} \qquad (3)$$

When the similarity measures of OTUs with themselves are one (which is true, for example, with correlation matrices and most similarity measures used in numerical taxonomy), equation (3) reduces to

$$d_{ij}^2 = 2(1 - g_{ij}) \qquad (4)$$

If the initial data matrix was a matrix of Pythagorean distances between points, then the distances between points in the final representation are the same as in the initial matrix. (Equation (4) is equivalent to equation (1) since the addition of a constant to G does

not the change the matrix H, obtained by mean-centering G, and therefore does not change the interpoint distances.) We now have a set of coordinate axes in which the first few axes explain a relatively large amount of the sum of interpoint distances.

When the initial association matrix is some other association measure, other than $d_{ij}$, the distances between points in the principal coordinates results will still be a function ({3} or {4}) of the association measure, but will no longer necessarily be the Pythagorean distance between the initial points. (This Pythagorean distance between initial points will be known only if initial coordinates were known. If only an association matrix is given, then "an original position" of the data points will not be known.) Principal coordinates analysis is therefore a method of constructing a metric representation of the relative position of points where the distances between points are proportional to the strength of the association, as measured by the association index used in constructing the original data matrix.

Gower [1967] has provided a simple approach to adding the principal coordinates scores of an additional OTU not used initially to calculate the principal coordinates. The new OTU can then be plotted together with the data already used in the analysis. The new point may also generate a new axis, and the coordinates for the OTUs on the new axis may also be determined. This is useful when new OTUs are added to a study in which the principal coordinates analysis is already completed, or for the addition by others of new data to already published plots. However, more of the original data are needed to apply this procedure than are usually given in the literature.

A comparison of principal coordinates analysis and

principal components analysis is useful for a clearer understanding of both methods. The most basic point of contrast is that principal component scores are computed by multiplying the original coordinates or measurements (or some transform of the original data such as the mean-centered data matrix B, or the standardized data, C) by coefficients that are the elements of the eigenvectors of the m x m association matrix between variables scaled to have length 1.0. In contrast, principal coordinate scores are determined solely from an OTU by OTU (n x n) association matrix, without the use of any original coordinates. However, a close relationship exists between the two methods. If PCORD is performed on a matrix of Pythagorean distances (a common association measure), then exactly the same coordinates and plots are obtained as if we had computed the principal component scores for OTUs on the principal axes determined from the sums-of-squares and cross-products matrix among variables, i.e. the variance-covariance matrix S multiplied by (n-1). Thus, the two methods--PCA, and PCORD on Pythagorean distance--used on a mean-centered data matrix give rise to the same set of interpoint distances, and are called "dual" to one another by Gower [1966a]. Also, a PCORD performed on the $d_{ij}$ matrix formed from C is dual to a principal components analysis of the correlation matrix, R = C'C/(n-1). The use of dichotomous variables in the "Simple Matching Coefficient" (SMC), a similarity coefficient between OTUs [Sneath and Sokal,1973:13,2] also leads to a result dual to PCA on the corresponding S matrix computed from an original 0/1 raw data matrix. As Gower points out: "Although conventional principal components analysis of (0/1) data may seem of dubious validity, it is exactly equivalent to assuming that the individuals are represented by points whose distance apart are proportional to" the square root of (1-SMC).

The dual properties of PCA and PCORD may be taken advantage of computationally when original data matrices are available. Only the smaller of the m x m and n x n association matrices need be computed, since the eigenvalues are the same. The eigenvectors of either may be obtained from the other [Jöreskog, Klovan and Reyment, 1976:110-111]. The main value of the principal coordinates procedure however, is not in these simpler dual cases, where one can make a choice whether to analyze the S or $d_{ij}$ matrix, but in those situations not paralleled in PCA in which the association matrix is computed using another similarity coefficient or distance measure, or in which the only data available is a similarity or distance matrix. In such analyses, PCORD is producing a representation of the OTUs in a multidimensional space, in which the OTUs are represented by points. The representation produced is one in which the distance between each pair of points is as nearly as possible proportional to the square root of the value of the association measure for that pair. If the associations cannot be completely represented in a Cartesian space, some of the eigenvalues will be negative, indicating that those axes have imaginary scores (with a negative variance along those axes). If the absolute values of such negative eigenvalues are small, then a large portion of the total interpoint associations are explained by the representation in a Cartesian space defined by the axes corresponding to the positive eigenvalues; the negative eigenvalues may be ignored. If, however, the absolute values of the negative eigenvalues are large, then the Cartesian or metric representation of the initial associations is a distorted summary and will therefore probably not be informative.

The directions of the principal coordinate axes are such that the distances between the projections of points onto the first axis make the single largest

contribution to the total of the interpoint distances, those onto the second axis the second largest contribution, and so forth. Thus, if the original associations are well represented by the final distances, a relative large amount of information about the associations between the OTUs is summarized in plots of the first few principal coordinates.

Principal coordinates analysis may also be used where comparisons are impossible for some characters on some OTU pairs, with the result that the association coefficient is computed from a reduced set of variables for those pairs. Thus principal coordinates analysis has certain advantages over PCA when the initial data set is incomplete. In PCA, there is no single obvious method to fill in missing values. However, the computation of a PCA requires a complete data set. Insertion of the mean value, or zero (if the data are mean-centered), for missing values pulls those OTUs with missing values nearer to the centroid of the points than is indicated by the actual data for those OTUs. For principal coordinates, in contrast, each value of the association index can be computed on the basis of only those variables for which data are present for both the individuals being compared. In this fashion, a moderate amount of data can be missing from the raw data matrix, and a complete association matrix still be computed, without any estimation of the missing data being required [Rohlf, 1972]. Negative eigenvalues will occur more frequently in the event of missing data.

Applications:

Principal coordinates analysis is a dimension-reducing technique. In studies where the initial data are associations--either similarities or

distances beween pairs of OTUs--this method provides coordinates that can be plotted for examination and that summarize the information about the association between pairs of points in the distances between pairs in the new coordinate system. Thus PCORD is an exploratory technique for examining the results of using any possible association index. (See Sneath and Sokal [1973:114-187] for a discussion of various measures of association.)

Gower [1971] has also defined a class of similarity coefficients which may be computed from raw data matrices and which combine dichotomous (presence-absence or alternate states) with nominal or qualitative characters (e.g. colors, alternative enzymes) as well as quantitative characters (measurements on any scales). See Holmes [1975], Berthou, Brower, and Reyment [1975], and Cook [1977] for examples. A weighting coefficient and missing comparisons are also included in his formula. Analysis of data sets computed using any member of this set of similarity coefficients will produce eigenvalues that are all greater than or equal to zero when there are no missing data. See Dodson [1976] for an example. Pilbeam [1969:13] states that "By experimentation, it was noted that for adequate comparison between two individuals, they should share values for at least two-thirds of the variates." This was based on 28 variables (for which the original data are given in an Appendix to his paper).

Gower [1966b, 1976] has recommended that principal coordinates analysis be used to display between-group differences using Mahalanobis $D^2$ as the association index and as a dimension-reduction technique instead of the more popularly used canonical variates analysis. [Reyment and Banfield, 1976, is an example of such an application.] The difference is discussed in some

detail in the section on "Discriminant Analysis".

PCORD is a metric form of multidimensional scaling (see the section on "Nonmetric Multidimensional Scaling") and may complement dendrogram-producing cluster analysis techniques, in that principal coordinates may be found for the same association matrix. Using PCORD one may estimate the dimensionality of the data, and produce plots of interpoint distances in a few dimensions. It is recommended that principal coordinates be used as a first step in nonmetric multidimensional scaling, to save time in that more costly technique (especially for large numbers of OTUs) [Gower, 1966a]. The results of applications of PCA, PCORD and nonmetric multidimensional scaling have been compared by Rohlf [1972] and Thorpe [1980], who come to rather different conclusions.

The dual relationship between principal coordinates and principal components analysis may be taken advantage of when PCA is the intended analysis and m, the number of variables in the study, is much larger than n, the number of OTUs. The distance-squared matrix will occupy less computer memory. The eigenvectors for the PCA analysis may also be obtained. (This fact does not seem to be taken advantage of in many studies; see Jöreskog, Klovan and Reyment [1976:110-111] for the appropriate formulae.)

Computational requirements:

A symmetrical association matrix. The association measure should be a metric or close to one, so that negative eigenvalues will be small.

Statistical assumptions:

When the PCORD is dual to PCA, i.e. the association measure is Pythagorean, then the tests used in PCA are equally appropriate if the statistical assumptions hold, since the results are equivalent. No tests or confidence interval procedures are appropriate for principal coordinates analysis when it is used primarily as an exploratory or ordination procedure.

Biological assumptions:

The distance or similarity measure used must satisfactorily measure the association between OTUs. If different characters are measured on different scales, the variables should be transformed in some way to the same scale or to scale-free variables [Gower, 1966a]. The choice of the association index involves assumptions about the nature of the structure of relationships one is looking for. For example, the use of the Pythagorean distance squared will search for linear relationships as in PCA. Another example would be the use of the correlation coefficient between OTUs as the index of association: this association index will emphasize 'shape' and ignore differences in 'size' that the distance matrix would emphasize when linear dimensions form the data. In other words, the effects of the association measure on the structures or pattern of points found should be considered in the interpretation of results.

As a dimension-reducing technique, the first few coordinates are often the only ones looked at in detail. This focus requires assumptions (as in PCA) that the structure one is interested in is satisfactorily summarized in the first few principal coordinates. This assumption can be quickly checked by

the computation of residuals to see if some interpoint distances are grossly underestimated by the first few principal coordinates. A minimum spanning tree superimposed on the principal coordinates plot (in two or three) dimensions, would connect nearest points together in the full space, and thus also indicate distortions or underestimation of some distances.

Computer packages and statistical programs:

A FORTRAN program PCOORD is listed in Blackith and Reyment [1971:171-185]. This program provides one form of Gower's index for dichotomous, qualitative and quantitative data.

Mather [1976:402-406] also gives a FORTRAN program PCORDA similar in design to that described in Blackith and Reyment, and claimed by the author to be more efficient.

A subroutine GOWER in NTSYS transforms any association matrix to a form suitable for principal coordinates analysis. A wide variety of association matrices are available in NTSYS.

A principal coordinates procedure may be developed in SAS using PROC MATRIX.

# OTHER PRINCIPAL COMPONENTS RELATED METHODS

The following two methods are related to principal components analysis, but are different enough to warrant special discussion. In each method the data matrix is scaled differently from a typical PCA. The goals of both of these methods are reduced-dimension displays of points for both OTUs and variables simultaneously.

## Correspondence analysis:

Correspondence analysis is a kind of principal components analysis on a specially scaled data matrix derived from tables of counts (i.e. contingency tables). Examples of such data arrays would be the numbers of fin rays for different species of fishes, or the abundance of m species at n localities, such as in ecological or biogeographical studies. The method combines some of the advantages of R-mode and Q-mode techniques (see "Factor Analysis" for definitions). The method has as its goal the production of loadings for both the OTUs and the variables on equivalent scales, which then may be plotted together. See Hill [1974] for a historical discussion and the relation of the method to several other methods.

The data are converted to proportions or estimates of probabilities, and then scaled by a form of simultaneous row and column standardization. The scaled matrix T is used to form association matrices T'T and TT'; these are then analyzed by standard eigenvalue and eigenvector routines available in most PCA or factor analysis procedures. The eigenvectors of TT' and T'T are scaled to have the same units to permit plotting the loadings on one scattergram. The formulae for scaling and the matrix operations are given in Jöreskog, Klovan and Reyment, [1976:107-113]. The

sum of the diagonal elements of the association matrix is related to the usual chi-squared statistic for independence in a contingency table. (If the hypothesis of independence is accepted then the matrix is nearly of rank one, and a unidimensional ordination will probably be sufficient for representing both OTUs and variables.) Stopping rules are required to determine the number of components to display or use, as in PCA and factor analysis.

The displays obtained from the analysis are especially interesting because points for both OTUs and variables are plotted. One may then interpret association of variables and OTUs in useful ways (discussed in Jöreskog, Klovan and Reyment [1976]). The method has been recently rediscovered and widely used in France, producing an extensive literature in French [see Jöreskog, Klovan and Reyment, 1976; David, Campiglio and Darling, 1974], including applications to systematics.

The data to be analyzed are most appropriately frequency data. Tables for discrete characters--fin rays for example, or nominal variables would lend themselves to this type of analysis. Hill [1974] discusses the application of the method to incidence data (presence/absence), appropriate transformations of continuous data, and ways of combining continuous and discrete data in the analysis to produce frequencies. As an ordination technique, correspondence analysis is called 'reciprocal averaging' [Hill, 1973, 1974]. The method has been applied to matrices of continuous measurements in systematic studies [for example Petit-Maire and Ponge, 1979]. It is not clear how the special scaling of the data affects the results in such applications, however, because it is difficult to understand the scaling geometrically. Jöreskog, Klovan and Reyment [1976] call the method a special

form of factor analysis, but this is in the context of considering PCA as a form of factor analysis. We wish to avoid this view (see discussion under "Factor Analysis" in METHODS).

The method is not available in any of the major packages. It may be programmed in very few steps using the procedure PROC MATRIX in SAS which has a simple call to an eigenvalue and eigenvector routine.

Biplot:

Biplot is a graphical display technique for modeling multivariate data sets. If the data can be sufficiently well approximated in two dimensions, two coordinates are determined for each OTU as well as for each variable. Both sets of coordinates are plotted on the same bivariate diagram.

The data matrix A is centered by subtracting the mean of all of the raw measurements, i.e. the grand mean, from each observation to produce the matrix B*. One then finds the first two vectors of variable loadings by a PCA of B*'B* (a variable-by-variable sums-of-squares and cross-products matrix of the transformed data) and the first two vectors of OTU loadings from a PCA of B*B*' (an individual-by-individual sums-of-squares and cross-products matrix of the transformed data). These loadings are then plotted as scores on the same plot. The sum of the first two eigenvalues (of either matrix--they both have the same eigenvalues) divided by the trace provides a goodness-of-fit statistic. A value of this statistic near one is considered a good approximation (it is >0.98 in most of the examples we have seen).

Gabriel [1971] explains biplot in detail, including its use for displaying clusters of OTUs, summarizing inter-OTU distances and for explaining correlations and variances among variables. A further discussion relating biplot to exploratory regression analysis is given in Bradu and Gabriel [1978]. References to discussions of weighted analysis and estimation of missing data using biplot are given in Bradu and Grine [1979].

Since the method is only advocated for those cases where a large proportion of the variance is explained by the first two eigenvalues, the use of biplot is appropriate for those situations where the researcher would feel that a bivariate plot, say of principal component scores, would summarize a great deal of the structure or relevant patterns in the data. The method can, however, clearly be extended to produce three or four axes with coordinates for both individuals and variables, but then one loses the advantage of an easily plotted summary. The main difference between biplot and two-component PCA is that the data are centered on the grand mean, and that the Q-mode loadings and the R-mode loadings are plotted on the same diagram. An equivalent way of describing the method is to think of it as a plot of the principal component scores, scaled to have variance equal to the eigenvalues, and the principal component loadings on the same diagram. Note that it is possible and useful, if there are not too many variables, to plot vectors corresponding to the original variables on any PCA plot of scores.

Bradu and Grine [1979] applied the biplot method to measurements of mammal-like reptile skulls. They present measurements for 30 variables on 26 skulls (including 7 holotypes) assigned to 14 species contained in 4 genera. 154 or 19.7% of the

measurements are missing, a not atypical situation in vertebrate paleontology. Only 3 specimens had complete suites of all 30 measurements available. A two-step biplot was fit. The missing values were estimated and then the full data matrix was analyzed. Raw data and log-transformed data were both examined. The relations between the variables could be explained by an isometric model (see section on "Size and Shape" in PURPOSES).

Computations can be performed in packages allowing the proper centering. A PCA can then be run on the specially scaled data matrix; the scores and loadings may then be plotted simultaneously. This should be possible in NTSYS, BMDP and SAS. A program is available from Gabriel [see Bradu and Grine, 1979].

# NONMETRIC MULTIDIMENSIONAL SCALING

The term multidimensional scaling refers to a variety of techniques, all of which produce dissimilarity/distance values in a few dimensions, making the structure or information in the matrix more obvious to the human observer. This is done largely by finding, for a matrix of proximity values between OTUs, an arrangement or a "plot" of points (each point representing an OTU) in a space of a few dimensions, and thus summarizing the structure in the original data. (This goal--dimension reduction--is discussed more extensively in the subsequent section reviewing purposes of analyses.) The arrangement of points, or the set of coordinates which describe that arrangement of the points in the space, is called the configuration for that analysis. The definition of multidimensional scaling is broad enough that it can be considered to include principal components and principal coordinates analyses.

A configuration is sought in this lower dimensional space that optimizes some stated relationship, $d = f(g)$, between the original associations (the $g_{ij}$'s between all pairs of points i and j), and the distances ($d_{ij}$'s) in the configuration. The goodness of fit of the $d_{ij}$'s to the $g_{ij}$'s is estimated by a measure of stress. If the relationship between the original associations and the distances between the points in the configuration depends on the numerical or metric properties of the g's, the form of multidimensional scaling that searches for the configuration optimizing that relationship is a form of metric multidimensional scaling. If, however, the relationship between g and d depends only on the relative ranks of the original associations, $g_{ij}$'s, the method is nonmetric multidimensional scaling

(NMDS). It is specifically this latter form of the method that is very often considered under the heading of "multidimensional scaling." [See Gnanadesikan, 1977:26-48, as well as Kruskal and Wish, 1978, for good introductions to NMDS.]

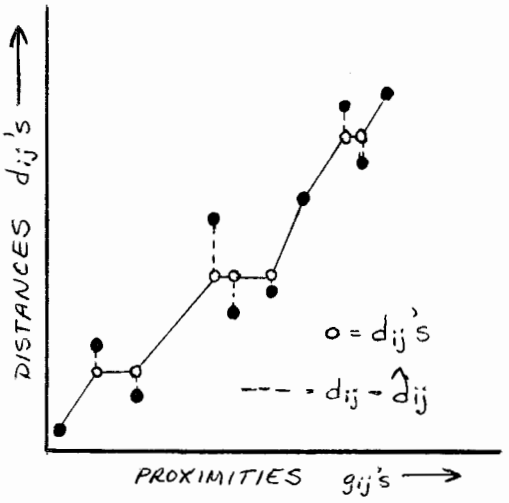
Some data gathered by systematists are already in the form of proximity values, such as strength of immunological response or the degree of DNA hybridization. Many data, however, are direct measurements on the OTUs individually, and unsuitable for NMDS in that form. A proximity matrix can be generated from such data by the calculation of any of a number of similarity coefficients or distance measures, such as the correlation coefficients between OTUs or the Euclidean distance between OTUs in the n-dimensional character space, to name just two common ones. (See Sneath and Sokal [1973:116-149] and Hartigan [1975] for discussion and references to similarity coefficients and distance measures.)

Next is chosen the number of dimensions, r, in which the resulting configuration is to be found. Usually, a series of analyses is made in 1, 2, 3, etc. dimensions, and the 'best' result chosen on the basis of a compromise between interpretability and good fit. (This will be discussed in more detail later.) It is important to note, however, that lower dimensional solutions are not merely projections of higher dimensional solutions. That is, the arrangement of the points relative to each other will not necessarily be the same in solutions of different dimensionality. Thus the choice of dimensionality in which the solution is sought will affect the final configuration. Further, the orientation of the axes themselves in any given solution is entirely arbitrary; there is no reason to expect the axes themselves to be interpretable directions. The configuration of points

in a low-dimensional space is the goal.

NMDS starts with an initial configuration, either randomly chosen or, preferably, obtained from either a previous solution in r-1 dimensions or a principal coordinates solution for r dimensions. All computational methods are iterative, moving the points relative to each other so that stress is reduced (that is, the correspondence between the $d_{ij}$ and the $g_{ij}$ is increased), and repeating this operation until stress does not decrease further. Because it is possible to reach local optima (as discussed in the INTRODUCTION--see page 45), and also because one wishes to minimize the number of iterations to save computing time, starting configurations are preferred that are likely to be in the neighborhood of the global optimum, such as the principal coordinates solution.

The correspondence between the original proximities and the distances in a configuration obtained from NMDS is usefully examined in a scatter plot often termed a Shepard diagram. The $d_{ij}$'s are plotted against the $g_{ij}$'s, so that each point represents a pair-wise comparison. If the configuration well represents the initial proximity values, the points will fall along a monotonically increasing or decreasing line (reflecting whether the proximity values g were similarities or distances), and the stress will be zero. Incongruences are evident as deviations from this monotone function.




Any of a number of measures of stress may be used. One of the most common measures is that given by Kruskal [1964a:9]:

$$S = \sqrt{\frac{\sum (d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}^2}}$$

where $\hat{d}_{ij} = f(g_{ij})$. For a monotonic function, the $\hat{d}_{ij}$ would be values that preserved monotonicity, as shown in the figure to the right. Thus, stress is measuring how far the configuration is from one that preserves monotonicity completely, and therefore is an indicator of how well the data are represented by the given configuration in the chosen number of dimensions. The value of stress obtained for the optimal configuration in a given dimensionality will decrease as one increases the number of dimensions for which a solution is found. An exception to this pattern indicates that a local optimum has been encountered, or that computation stopped before the iterative procedure converged to an optimum [see Kruskal and Wish, 1978:28,50].

The optimum number of dimensions is determined on the basis of several criteria. While stress will continue to decline with increasing dimensions, a noticeable change in the rate of decline often occurs; such an inflection point indicates the optimum dimensionality [Kruskal and Wish, 1978:53-56]. Clearly a 2- or at most a 3-dimensional solution is greatly preferrable since it can be easily comprehended. However, if more than 3 dimensions are required to adequately summarize the associations in the proximity matrix, the configuration may be subsequently rotated or new axes found for display, to produce 2- or 3-dimensional plots for display. PCORD on the $d_{ij}$ or PCA on the coordinates in the final configuration will produce the principal components of the configuration, so that the directions of maximum dispersion may be examined.

Another factor often cited as a criterion for choosing the dimensionality is the stability of the configuration. Kruskal and Wish [1978:34] give rules of thumb about the number of dimensions that can be

expected to give reasonably stable results (over repeated samples) given various sample sizes. If the goal includes making inferences about a population based on a sample of individuals, then one needs sufficient $g_{ij}$'s or pairs of points to provide information about directions of interest or trends within the population. See Kruskal and Wish's discussion, and references cited therein. In contrast, systematic applications frequently involve a need to summarize the data for only the given OTUs in fewer dimensions, and do not involve inferences about a larger statistical population. In that case, sample size is not a problem since the entire universe has been sampled. An example might be NMDS on the distances between centroids of 8 species when all the species of interest have been sampled. The choice of dimensionality is discussed further in Kruskal and Wish [1978] and in references cited therein.

A point which is worth emphasizing is that there are many computational procedures to compute solutions in NMDS as well as the varieties of metric multidimensional scaling. Some of this variation is in the measure of stress. Therefore, some clear statement of the computational procedures, at least the measure of stress employed, is usually necessary in reporting the results of NMDS. (A slightly technical review of methods and programs is available in Kruskal [1977].)

The results of NMDS have been compared with those of PCA and PCORD. The first set of studies, by Rohlf [1970, 1972], provide a useful comparison of these three ordination techniques. Rohlf [1972:279] has been subsequently cited in applications as support for a preference for NMDS over PCA and PCORD, and did indeed provide some useful recommendations; however, the statement that distances between near individuals or OTUs are shown more accurately in NMDS is incorrect.

As illustrated in Rohlf [1970:Fig.9], small distances appear to be represented more precisely by NMDS, but are, on the average, underestimated about the same amount as by PCA. The estimated distances in PCA (i.e. in a space of reduced dimensions) are of course always less than the original distances. In NMDS, there is no such constraint on the direction of error; Rohlf's example shows an overestimation of the larger distances. This, combined with the underestimation of the smaller distances, results in an exaggeration of the relative size of distances within the NMDS configuration. However, it is still true that NMDS provides a more accurate representation of the rank order of the smaller distances between points, as illustrated by Rohlf's figure, in that there is less variation in values of $d_{ij}$ for a given value of g in NMDS than in PCA. (NMDS is providing metric information only in that one obtains coordinates in Cartesian space from proximity values; it is not metric in the sense that the actual sizes of the $d_{ij}$ in the configuration have any real meaning in themselves.)

In contrast to Rohlf's recommendations, Thorpe [1980] finds NMDS less satisfactory than PCA or PCORD for representing the structure in a geographic variation study within one species. He suggests, in agreement with A. J. B. Anderson [1971:4], that NMDS may not show structure satisfactorily when there are outliers or major aggregations. Kruskal and Wish [1978:29-30] discuss the recognition of clustering from the Shepard diagram.

Applications:

The value of nonmetric multidimensional scaling lies in the production of low dimensional representations of association data without the constraint that the

represented relationships be linear, or that one have faith in the metric values of the original data or associations. For example, Fix and Lie-Injo [1975] choose NMDS because they have substantial faith in the accuracy of the rank order of their data, but less faith in the accuracy of the actual metric values.

NMDS is a useful ordination technique in applications in which one is not concerned that the distances between points be in correct proportion to each other but only that the rank order of the distances be maximally preserved--that is, that the relative sizes ("greater than" or "less than") be correctly shown. See discussion and references above. Thus NMDS provides a geometric summary of the relative nearness of OTUs, but is not a good method for looking for large or distinct clusters. [See Anderson, A. J. B., 1971, Skeel and Carbyn, 1977, and Thorpe, 1980, for examples and applications.]

Moss, Peterson and Atyeo [1977] provide an example of the use of NMDS in concert with metric ordination techniques, in this case PCA. The 3-dimensional representation from a PCA was used as the initial configuration, NMDS was done on taxonomic distances between OTUs, and then PCA was done on the NMDS configuration (in the space of reduced dimensionality) in order to align the axes along the major trends of variation. Drennan [1976] carries the point about indeterminancy of the specific directions of the axes one step further: not only are the NMDS axes meaningless, but one need not even be restricted to a linear direction. Drennan finds that his archeological sites order chronologically along a horseshoe-shaped curve, along which he can subsequently place undated sites.

Computational requirements:

None, beyond choice of an association measure if the original data are not associations.

Statistical assumptions:

If one is generalizing from one's sample to a population, then random sampling is required. Statistical approaches are still being developed for this method. Kruskal and Wish [1978:89-92] review statistical methods (based on Monte Carlo studies) for determining the appropriate number of dimensions. At present, the statistical approaches are limited in the range of situations in which they are applicable.

Biological assumptions:

The lack of structural assumptions in this method allows an exploratory use of NMDS remarkably free from biological assumptions. In any specific interpretation, the factors invoked to explain the configuration would be assumed to have been measured in some fashion by the original association measure.

Statistical packages and computer programs:

NTSYS has a widely used multidimensional scaling program. Other programs are referenced in Kruskal and Wish [1978] and Gnanadesikan [1977]. In some releases, SAS has a NMDS routine similar to ALSCAL [Kruskal and Wish, 1978:79].

# FACTOR ANALYSIS

Factor analysis is a multivariate technique used to describe a set of observed variables as a function of a set of hypothetical variables called factors. The technique originated in the attempt to find the factor (or factors) of "intelligence" using batteries of so-called "intelligence" tests and their scores as measured variables. Ideally all of the correlation or linear relationships among the measured variables should be explained by relatively few such factors. These factors, called common factors, account for the correlations and also account for a portion of the variance of each variable. The remaining variability is attributed to sources of variation unique to a given variable and random error or measurement error. The amount of each variable's variance explained by (or accounted for by) the common factors is called the communality for that variable, while the variance unique to each variable is called the uniqueness.

The goal of factor analysis is to explain the common variances or the correlations, in a number of factors r, less than the number of variables measured, m. The factor analysis model may be stated as:

$$C = FE' + V \qquad \{1\}$$

Here C is the n x m data matrix, mean-centered and standardized by variables. C may be expressed as the product of factor scores F and coefficients or factor loadings E', plus a matrix V of "unique" values--"residuals" or "errors". The factor scores F are values ("measurements") of the hypothetical variables (factors) for each individual. The factor scores matrix contains fewer columns than C: i.e., there are fewer factors than original variables. E' (or its transpose, E) is the matrix of coefficients or factor loadings by which F is weighted to explain the

common parts of the variables in C. V is an n x m matrix of unique values which are due to the unique factor for each original variable (including random error), and therefore are not explained by the common factors; the unique factors are uncorrelated with the common factors. Both the factor scores F and the factor loadings E must be estimated from the original data C.

Factor analysis begins with the production of an association matrix such as the correlation matrix or variance-covariance matrix. In order to avoid problems of scale, it is usually the correlation matrix R which is factored; all of the variables are then scale free. The factor model, given in {1} in terms of the scores on the original variables and scores on the resulting factors, may also be stated as:

$$R = EQE' + U \qquad \{2\}$$

Q is the matrix of covariances among the factors, e.g. Q=F'F. E is, as above in equation {1}, the matrix of factor loadings or factor pattern matrix, and U is the matrix of uniquenesses. When the factors themselves are uncorrelated (i.e. orthogonal), Q is the identity matrix, with off-diagonal elements equal to zero. The vectors of factor loadings are usually scaled to have length equal to the square root of the corresponding eigenvalue. The resulting factors have a variance of one, so the diagonal of Q consists of ones when this scaling is employed. The matrix U is defined to be a diagonal matrix; the unique part of each variable is uncorrelated with the unique part of all other variables. The factor structure is the matrix of correlations between the original variables and the factors, and is the product EQ.

Without further restrictions in the model, there is a built-in indeterminancy. Infinite sets of r factors

can be found which explain the correlations equally well. These sets of factors correspond to different rotations of the factors. Therefore, additional criteria have to be found to choose the specific solution or rotation for interpretation. The most common choice is to rotate to find some form of simple structure. Two kinds of rotations are considered: 1) rigid or orthogonal rotations where the factors remain uncorrelated; and 2) nonrigid or oblique rotations where the factors are correlated. For orthogonal solutions, the factor pattern and the factor structure matrices are identical since Q is the identity matrix. However, in oblique factor solutions Q is not the identity matrix. In that case, the elements of the factor pattern matrix E are not correlation coefficients, but may still be viewed as regression coefficients of the observed variables on the factors [Kim and Mueller, 1978b; Jöreskog, Klovan and Reyment, 1976:61].

Thurstone's criteria for simple structure are most often cited [Thurstone, 1947; Kim and Mueller, 1978b:30-32; Mulaik, 1972:219-221 and Rummel, 1970:380]. His suggestion was to rotate the original orthogonal factors to a configuration in which the new factors (no longer necessarily orthogonal) each have zero loadings for many variables, and each variable requires fewer than the total number of factors to explain its common variance. Mulaik [1972:221-224] gives a geometric description of simple structure. In general, the goal of rotation is usually to find factors which are related to clusters of interdependent variables. The simplest possible structure occurs if each variable has nonzero loadings on only one common factor, but this is very unlikely for real data. Various orthogonal rotations are available that provide for some aspects of simple structure. Examples of orthogonal rotation techniques are Varimax, Quartimax,

Equamax, etc. [Kim and Mueller, 1978b:34-37; Rummel, 1970:422; Mulaik, 1972:Chapter 10]. Oblique simple structure rotations are also available [Kim and Mueller, 1978b:37-41; Rummel, 1970:390; Mulaik, 1972:Chapter 11].

The following artificial example will show the effect of rotation on interpretation. The numbers are from Jöreskog, Klovan and Reyment [1976:62-67] but the example is cast in a different biological framework. The computations were checked using PROC FACTOR in SAS79.

Below in Table 1 is given a correlation matrix R for six morphological variables, presumably obtained from a large sample of organisms. Variables 1, 2, and 6 are from morphological complex I and 3, 4, and 5 from

| TABLE 1: | x1 | x2 | x3 | x4 | x5 | x6 |
|---|---|---|---|---|---|---|
| x1 | 1.000 | 0.720 | 0.378 | 0.324 | 0.270 | 0.270 |
| x2 | 0.720 | 1.000 | 0.336 | 0.288 | 0.240 | 0.240 |
| x3 | 0.378 | 0.336 | 1.000 | 0.420 | 0.350 | 0.126 |
| x4 | 0.324 | 0.288 | 0.420 | 1.000 | 0.300 | 0.108 |
| x5 | 0.270 | 0.240 | 0.350 | 0.300 | 1.000 | 0.090 |
| x6 | 0.270 | 0.240 | 0.126 | 0.108 | 0.090 | 1.000 |

morphological complex II. Note that some correlations are quite low, and the number of common factors is not obvious. A maximum likelihood factor analysis yields the following matrix of factor loadings E (the pattern matrix) and the associated communalities and uniquenesses (Table 2). Note that the communalities are the sums of squares of the rows of the pattern matrix, and that the communalities and the uniquenesses sum to one for each variable. The two orthogonal factors completely recover all of the correlations. Any correlation coefficient between any two variables

TABLE 2: Factor Pattern Matrix

| | Factor 1 | Factor 2 | Communalities | Uniquenesses |
|---|---|---|---|---|
| | --- E --- | | | |
| x1 | 0.889 | -0.138 | 0.81 | 0.19 |
| x2 | 0.791 | -0.122 | 0.64 | 0.36 |
| x3 | 0.501 | 0.489 | 0.49 | 0.51 |
| x4 | 0.429 | 0.419 | 0.36 | 0.64 |
| x5 | 0.358 | 0.349 | 0.25 | 0.75 |
| x6 | 0.296 | -0.046 | 0.09 | 0.91 |
| Variance explained | 2.067 | 0.573 | 2.64 | 3.36 |
| % Trace R | 34.4 | 9.6 | 45.0 | 55.0 |
| % Trace (R-U) | 76.4 | 23.6 | 100.0 | - |

may be obtained by taking the inner product of the corresponding rows of the factor pattern or loading matrix E. For example, the correlation between variables x2 and x3, is 0.791x0.501 +(-0.122)x0.489=0.337 (a slight round off-error). Thus the correlations in this artificially constructed example are completely accounted for by two factors. The first factor appears to be a "size" factor since all of the loadings have the same sign, though there are numerical differences among them. The second factor might be interpreted as a "shape" factor, contrasting the two complexes.

An orthogonal rotation, one of many possible, of about 9° produces new orthogonal factors and makes all loadings positive, as shown in Table 3. The communalities and uniquenesses are left unchanged by the rotation. The correlations are still recoverable from the two factors by the same multiplication process. The two factors might be interpreted as a "size" factor and a morphological complex II factor.

Next we will rotate the factors to an oblique

TABLE 3: Factor Pattern Matrix

| | Factor 1 | Factor 2 | Communalities | Uniquenesses |
|---|---|---|---|---|
| | -------- E* --------- | | | |
| x1 | 0.90 | 0 | 0.81 | 0.19 |
| x2 | 0.80 | 0 | 0.64 | 0.36 |
| x3 | 0.42 | 0.56 | 0.49 | 0.51 |
| x4 | 0.36 | 0.48 | 0.36 | 0.64 |
| x5 | 0.30 | 0.40 | 0.25 | 0.75 |
| x6 | 0.30 | 0 | 0.09 | 0.91 |
| Variance explained | 1.94 | 0.70 | 2.64 | 3.36 |
| % Trace R | 32 | 12 | 44 | 56 |
| % Trace (R-U) | 72.7 | 27.3 | 100 | - |

representation. This produces the pattern matrix in Table 4. (This result was reproduced using option PROMAX in SAS79 PROC FACTOR.) Tne elements of the factor pattern matrix $E^+$ are no longer correlations between the factors and the original variables, i.e., no longer equal to elements of the factor structure matrix, as they were for factor pattern matrices E and E*. The latter were equal to the corresponding factor structure matrices since those solutions were orthogonal. The communalites and uniquenesses are still the same. The factor structure matrix (given in Jöreskog, Klovan and Reyment [1976:65]) is obtained by multiplying the pattern matrix by the matrix of correlations between the factors. The correlation between these two factors is 0.60 and corresponds to an angle of about $53^{\circ}$ between the axes.

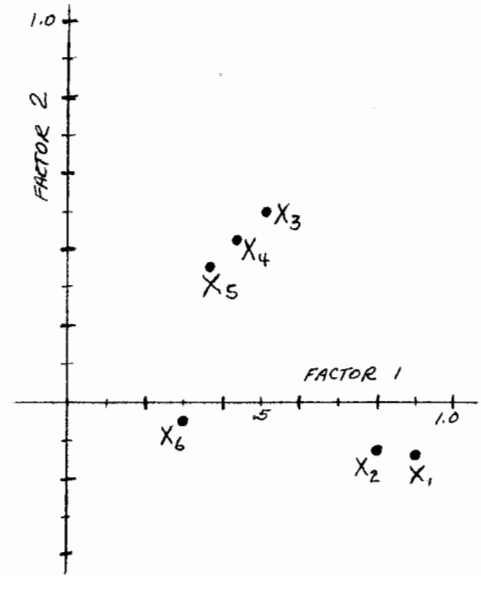
The original correlations among variables may still be recovered, but now from the product of the factor pattern and factor structure matrices, or from either matrix and the correlations between factors (matrix Q). The variance has been partitioned much more equably over the the two factors than in the arbitrary rigid

TABLE 4: Factor Pattern Matrix

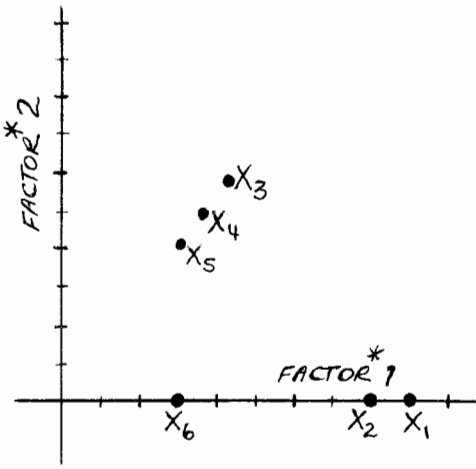
| | Factor 1 | Factor 2 | Communalities | Uniquenesses |
|---|---|---|---|---|
| | -------- $E^+$ --------- | | | |
| x1 | 0.90 | 0 | 0.81 | 0.19 |
| x2 | 0.80 | 0 | 0.64 | 0.36 |
| x3 | 0 | 0.70 | 0.49 | 0.51 |
| x4 | 0 | 0.60 | 0.36 | 0.64 |
| x5 | 0 | 0.50 | 0.25 | 0.75 |
| x6 | 0.30 | 0 | 0.09 | 0.91 |
| Variance explained | 1.54 | 1.10 | 2.64 | 3.36 |
| % Trace R | 26 | 18 | 44 | 56 |
| % Trace (R-U) | 59.1 | 40.9 | 100 | - |

rotation of the orthogonal factors. Simple structure solutions obtained from both rigid and oblique rotation procedures frequently have this property.

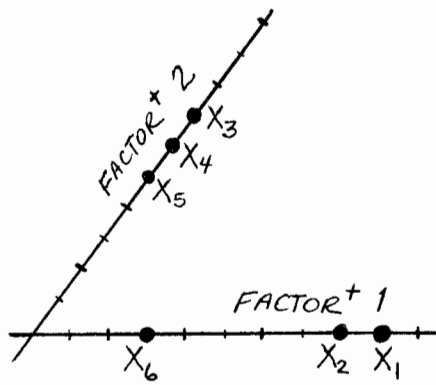
The two complexes have been completely resolved into two separate but correlated factors. This artificial example illustrates that more than one interpretation is possible, and perhaps desirable for the same data. A "size" component may be broken up by oblique rotations which reveal interdependent clusters of variables.




It is instructive to plot the factor pattern for each of the above steps; in such a plot the factor loadings for each variable are plotted on axes corresponding to the two factors, such that each point represents a variable. In the first figure, where the original maximum likelihood solution is given, all variables have positive projections on factor axis one, which represents the dominant factor. The rigid orthogonal rotation represented by the second figure produces zero loadings on the second factor axis for the variables of complex I; the second factor axis is




now in the direction of the variables in complex II. The third figure, which has axes at an angle of 53° (corresponding to the angle whose cosine is 0.60, the correlation between the factors), has the two factor axes coincident with the vectors representing the variables in the respective complexes. The lengths of the projections on the two factor axes in figure 3 are then just the square root of the communalities for the respective variables.




This example was constructed primarily to demonstrate the effects of rotation. Kim and Mueller [1978a:51-59] give a similar example in more detail. Real data will rarely give such perfect results: 1) with real data, a stopping rule will have to be determined for the number of factors, 2) there will be a residual matrix containing variation not explained by the model, and 3) the loadings will not all be positive or zero, but instead will have a range of values depending on the "separateness" of the clusters of the variables explained by the factors. Even more interesting was the effect of the choice of analytic procedure on the results. The results given above were not obtained when we used either non-iterated or iterated principal factor analysis--two of the more popular procedures in use. The question might then be raised, is one type of factor analysis more likely to give you the "right" answer than another? Asked in this general way, this question does not seem to us to be answerable (although some studies, e.g. Fisher [1973], have attempted to answer it); the "right" method depends on the problem. The varieties of factor analysis differ in their criteria for finding structure in the data, and therefore differ in what they are looking for and what they are likely to find. Thus a clear understanding of what each method is doing is desirable, to allow a proper match between problem, method, and interpretation. In factor analysis, this

understanding is difficult for most nonspecialists to achieve because of the complexity and diversity of factor analytic methods.

Part of the complexity of factor analysis arises from the diversity of options available at each step, and the proliferation of names for the variants of the method. First, there are a variety of ways to initially estimate the communalities to be placed along the diagonal of the correlation matrix. After this initial choice of communalities, the factors themselves may be extracted using a variety of methods. In non-iterative principal factor analysis (PFA), the initial values for the communalities remain unchanged. The results of PFA will therefore be dependent on the initial method of estimating communalities; different estimates will produce different results. In contrast, iterative principal factor analysis, PFA(iter) or MINRES of some authors, uses a starting estimate for the communalities as in PFA, but then re-estimates them from the factor loadings obtained in the analysis. The new estimates for the communalities are substituted for the diagonals of R and new loadings are computed. This is repeated until the solution is stable. Other factor analytic procedures will produce different final estimates for the communalities. The differences among the various factor analytic methods will result in differences in the factor loadings and scores produced by the various methods. The number of factors depends on the choice of any one of a variety of stopping rules. And after the set of factors is obtained there remain the options associated with rotation: oblique vs. orthogonal, and the criteria for choosing a particular rotation.

These options are surveyed well in the literature, briefly by Kim and Mueller [1968b] and in more detail by Cattell [1965a,b] and texts such as Rummel [1970],

Harman [1967], and the like. Indeed, many books have been devoted entirely to extensive discussions of factor analysis: the reader is directed to some of them for a more complete discussion than is possible in the short introduction presented here. An excellent place to start is with the two small paperbacks by Kim and Mueller [1978a,b], which include a clear glossary, examples of "setups" for factor analysis using SPSS, BMDP and SAS, and a useful bibliography. Another introductory paperback, with a geologic and paleontologic emphasis, is by Jöreskog, Klovan and Reyment [1976]. A good more comprehensive book is Rummel [1970]. Kim and Mueller [1978a,b] suggest other books of varying levels of difficulty.

It is worthwhile to contrast factor analysis with principal components analysis since there is much confusion and disagreement about the differences.

1) Both the factor analysis model and the PCA model may be written as equations (1) or (2) above. In factor analysis, matrix U is defined to be diagonal (unique variances on the diagonal and zero off-diagonals) and is of full rank; in PCA U is the residual matrix after removing r components and therefore has rank less than or equal to (m-r). In factor analysis, the matrix for which factors are found is R-U, the correlation matrix minus the matrix of uniquenesses, leaving the communalities in the diagonal. Thus only the common variance is explained by the factors. In principal components analysis the diagonal elements of R (or S) remain unaltered; thus the total variance and covariance in the data set is being explained by the principal components.

2) The goal of factor analysis is to find a parsimonious explanation for the correlations in a larger set of variables using a small set of factors;

added criteria must be given to determine communalities, find the number of factors, and rotate to the most interpretable solution out of an infinite number of solutions. A factor analysis solution produces correlated factors if oblique rotation is used. In contrast, the goal of PCA is not necessarily to explain correlations among the original variables using a reduced set of axes, but more often is to summarize the maximum amount of the variance contained in the original data using a reduced set of new orthogonal axes. Thus, the purpose of a PCA is usually to produce a space of reduced dimensionality. The principal components are a unique solution defined in terms of the proportion of total variance determined in the components displayed or retained. Principal components are always uncorrelated and orthogonal by definition because they correspond to principal axes of ellipsoids defined from the data.

3) Solutions from factor analyses are almost always rotated to simple structure using additional criteria. While the principal components solution may be rotated (and is by some practitioners), this destroys the ordered variance explanation property of the components, and they are therefore no longer principal components.

4) Principal component scores (matrix Z) are simply computed by multiplying the mean-centered or standardized data matrix, B or C, by the eigenvectors of S or R respectively. Factor scores (matrix F) must be estimated from the data and cannot be computed directly. Least squares regression is frequently used [JÖreskog, Klovan and Reyment, 1976:142-144; Rummel, 1970:Chapter 19]. However, other criteria are available [Kim and Mueller, 1978b:60-73; Mulaik, 1972: Chapter 13].

5) While both factor analysis and PCA may be done on any symmmetric association matrix, factor analysis is almost always done on a correlation matrix. In contrast, PCA is quite often done on a variance-covariance matrix as well as the correlation matrix (however see Statistical assumptions). This historical division may help one understand some of the contradictory definitions for principal factor analysis versus principal components analysis. Van de Geer [1971:137], for example, restricts PCA to an analysis of the variance-covariance matrix, and calls PCA of a correlation matrix principal factor analysis even when the diagonal elements are kept equal to one. This is contradictory to the distinction between the models we describe in (1) above and we strongly urge the reader to avoid such confusion. However, such diversity will be encountered and the reader should be forewarned.

Part of the confusion in differentiating between these methods arises from the fact that all of the major statistical packages available do principal component analysis through their factor analysis procedure. Therefore, in some of them, PCA may only be done on a correlation matrix. The terminology and labeling in the output from these routines also has a factor analysis orientation. Finally, one of the more widely used factor analytic procedures, principal factoring, corresponds to performing the operations of a principal components analysis on R-U (the correlation matrix minus the uniquenesses). The factors retained are then frequently rotated by one of the criteria available and then the result has been called for example, "Principal components (Varimax rotation)"--certainly a confusion of terms. In view of the clear differences between the two methods, we feel that it is most useful to keep the principal components analysis model, terminology and discussion separate from those of factor analysis.

Applications:

Considerable controversy exists about the usefulness of factor analysis. Most of the argument has been in the psychometric literature [see Blackith and Reyment, 1971:201-204 for references, some to systematic literature], and more recently in the physical anthropology literature [Kowalski, 1972 and references cited therein].

As an example of one opinion, Mulaik, in the introduction to his book [1972:xi-xiii], makes a persuasive statement that is neither pessimistic nor wholely optimistic. Although directed to psychologists, his point is general: "Factor analysis is not a method for discovering full-blown structural theories about a domain. ... factor analysis has been more profitably used when the researcher knew what [s]he was looking for." [Mulaik, 1972:xii] Mulaik therefore emphasizes confirmatory factor analysis in which prior hypotheses about the number of factors, or the nature of loadings may be proposed and tested. For example, hypotheses of isometry (corresponding to equal loadings on all variables) may be tested on a general "size" factor. Confirmatory factor analysis appears to us to have great potential usefulness in systematics as well as other fields, because it allows "custom-made" hypothesis-testing procedures. It is a logical outgrowth of clear and explicit statements about one's model: the assumptions being made and the parameters being estimated.

In contrast, many have considered exploratory factor analysis a profitable exercise; this approach will remain useful in beginning stages of many studies. Most of the explanation given above, and most analyses in the systematics literature are exploratory factor analyses. In this approach, the researcher is trying

to find relatively few factors which facilitate an interpretation of the data. No hypothesis is offered about the number of factors, nor the loadings, prior to the data analysis. See Stroud [1953] for an application to systematics. Rummel [1970:22-32 and elsewhere] and Cattell [1965a,b] give careful discussions of the philosophy and theory of exploratory factor analysis.

As a specific example, factor analysis has been thought useful in explaining or finding functionally related or correlated parts of organisms in systematic studies. Principal components often produces what is called a "size" component (frequently called a size factor) and orthogonal "shape" components (sometimes called shape factors) in which variables with large positive loadings are contrasted with those with large negative loadings. In contrast, factor analysis can yield simple factors (with loadings near 0 and 1) and is more versatile in that it can produce correlated factors that may be more interpretable biologically [Atchley, 1971; Gould, 1967]. Most rotations, however, break up the "size" component and combine it with other components to produce new factors [Benfer, 1975]. The new, rotated factors may be related to separate clusters of original variables which define "functional complexes". (Many systematists, however, feel more comfortable with a principal components solution, and for some applications it is indeed more appropriate.)

Factor analyses have been used in a variety of problems in systematics. We need to distinguish between two modes of factor analysis in applications. The discussion and explanation given above was in terms of R-mode factor analysis. In R-mode analysis it is the m x m matrix of correlations among variables that is factored, and the goal is relatively few factors that explain the correlations among the variables or

aspects of geographic variation. The factoring of binary data (0/1, presence/absence) from distributions of species, such as in ecological or distributional studies, is of questionable validity [Kim and Mueller, 1978b:73-75; Rummel, 1970:224-225]. However, factor scores may be plotted against geography as a way of expressing compactly the general trends in a large number of variables [e.g. Sokal and Rinkel, 1963; Sokal and Thomas, 1965:208-209].

Computational requirements:

A symmetrical matrix of associations, usually correlations are factored. However, some methods are based on covariances. Almost all models define the common factors and unique factors to be uncorrelated, and the unique factors to be uncorrelated with each other. The number of factors are determined by various cutoff rules, which usually must be chosen prior to the analysis (see Kim and Mueller [1978b] and Rummel [1970:Chapter 15] for an extensive discussion).

It is generally argued that the variables be continuous, rather than dichotomous or ordinal with a few categories; the interpretation of loadings on factors of binary data is difficult to justify [Kim and Mueller, 1978b:73-75; Rummel, 1970:224-225].

Statistical assumptions:

Tests of hypotheses on the numbers of factors or *a priori* vectors of loadings are available through maximum likelihood factor analysis (ML). The ML solution and tests are based on assumptions that the observations have a multivariate normal distribution [Kim and Mueller, 1978b; Morrison, 1967:264-268]. A

large sample goodness-of-fit chi-square statistic may be obtained to test for the number of factors [Kim and Mueller, 1978b; Morrison, 1967:269-270] and tests for other hypotheses are available [Kim and Mueller, 1978b; Sörbom and Jöreskog, 1976]. Canonical factor analysis (called RAO in SPSS) is essentially the same as ML factor analysis. The usual derivation of the maximum likelihood solution depends on the assumption of multivariate normality. While Jöreskog, Klovan and Reyment [1976:82] claim that maximum likelihood factor analysis is robust to violation of the assumption of multivariate normality, Kim and Mueller [1978b:77] state that the effects of violating this assumption are not clearly understood. However, Howe gives "an alternative approach" to factor analysis which leads to the same model (equation (2) above) and "with suitable assumptions" a maximum likelihood solution [fide Morrison, 1967:286-289] this approach does not depend on the multivariate normality assumption. Maximum likelihood factor analysis produces results that are invariant under changes in the scaling of the original variables.

It would seem to us that the most important criterion for acceptability of factor analysis as a useful tool in systematics would be the stability or repeatability of its results over different random samples of OTUs from the same population(s). For the size of samples usually studied, this must be determined by empirical sampling experiments, or computer simulation studies. Many such studies have been done in the social sciences [see Kim and and Mueller, 1978b]. It would be possible to use jackknife procedures in factor analysis (see INTRODUCTION for definition) but this would require a great deal more computation time. As an example, Sokal and Thomas [1965] found fairly good repeatability for factors summarizing morphological variation over localities,

but not for morphological variation among life stages for their organisms.

Biological assumptions:

The use of the correlation matrix as the symmetrical association matrix to be factored means that the method is examining linear relationships among variables (or OTUs in Q-mode factor analysis). (See the discussion in the section on "Principal Components".)

Most factor analysis results are not invariant under changes in scale. However, one attractive feature of maximum likelihood factor analysis is that the loadings for a covariance matrix analysis may be obtained simply from that of the correlation analysis by multiplying by scale factors. Therefore, it doesn't matter which association matrix is factored: unlike principal factor analysis and PCA, a ML factor analysis produces the same results under a change of scale. This is also true for some other models not considered here [Mulaik, 1972: Chapter 8].

For many systematic studies employing factor analysis as a kind of clustering technique, the assumptions of multivariate normality required for the tests will not be valid. If Q-mode factor analysis is employed, the tests that would be used in R-mode analysis will generally not be applicable because the assumptions of a "random sample" of characters for each OTU is seldom if ever valid, and the concept of a normal distribution of those characters in OTU space is not applicable. The scores of OTUs obtained from an R-mode analysis may be used in a subsequent clustering procedure; if clusters are thought to be present, the assumption of multivariate normality is clearly contradicted. Maximum likelihood factor analysis

procedures have been extended to allow comparison of factor structure among different populations [Kim and Mueller, 1978b; Sörbom and Jöreskog, 1976]; these have not been applied to systematic biology as far as we know.

Perhaps the most basic assumption of most applications of R-mode exploratory factor analysis is that the concept of simple structure is applicable to biological systems. One might question whether, for example, given modern genetic or ecological theory, one would expect to find a few definable factors that explain the patterns of morphological variation we see over ontogeny and phylogeny. We might hope that this is true, but can we expect it from our current hypotheses of development and evolution? Simple structure assumes not only fewer underlying factors than original variables, but also that only a few of these underlying factors affect or explain any given character. For a cautious discussion, see Sokal, Daly and Rohlf [1961:1116-1118].

Statistical packages and computer programs:

Most of the available packages (NTSYS, BMDP, SAS, SPSS and others) have a large number of options for factoring (generating an overwhelming number of combinations). However, some of the packages--SAS and SPSS--only allow factoring of correlation matrices. The major options available are summarized below in table form. Options usually have defaults, indicated by an *, which are used unless an alternate choice is made by the user. Not all of the defaults will work together, but generally an analysis will result if all of the defaults are used.

Principal components analysis is obtained in these

packages through PFA (non-iterative principal factoring) with unaltered diagonals. Brief descriptions for some of the options and methods are given at the end of the table. The reader is referred to the manuals or texts on factor analysis for fuller explanations.

| | BMDP | NTSYS | SAS79 | SPSS |
|---|---|---|---|---|
| 1. Matrices to be factored | R*,S, non-centered R,S | R,S, any | R | R |
| 2. Factor method | PFA*,ML,LJIFFY PFA(iter) | PFA,PFA(iter) (ALPHA,IMAGE ML-planned) | PFA*,ML PFA(iter) IMAGE | ALPHA,IMAGE PFA(iter)* RAO |
| 3. Communalities | Unalt* for PCA MCS* others Priors,Max | Unalt*,MCS Max Priors | Ones(PFA)* MCS,Max Priors | Ones,MCS,MAX Priors |
| 4. Number of factors | Mineigen(=1*) r | Mineigen(=1*) r | Mineigen(=1*) r,%trace | Mineigen(=1*) r |
| 5. Orthogonal | VARI*,QUARTI, EQMAX,ORTHOG None | VARIMAX (separate routine) | VARI,QUARTI EQMAX None* | VARI*,QUARTI EQMAX None |
| 6. Oblique rotations | DOBLI,ORTHOB DQUART | FUNCTNPLN PRIMENPL SPACEWARP | PROMAX | OBLIQUE |
| 7. *Plots* | *Loadings** Scores* | *(Separate* routines) | *Loadings* (Scores in separate routines) | *Loadings* (Scores in separate routiness) |
| 8. Special output | Partial corr Residual R Shaded R $R^{-1}$ Sorted loadings* $D^2$* | Eigenvectors scaled to length 1.0 | Eigenvectors of R scaled to length 1.0 | $R^{-1}$ |
| 9. Special features | Case weights, scores on other data | Choice of eigen routine | | Special missing data handling |
| 10. Maximum number of variables | 150 double prec. (adj. by user) | Set by user | 250 | 100 |

1. R=correlation matrix; S=variance-covariance matrix.
2. PFA--principal factoring (non-iterative)-see Manual.
   PFA(iter)--principal factoring (iterative)- see Manual, Kim and Mueller [1978b].

ML--maximum likelihood factoring--[Kim and Mueller, 1978b; Morrison, 1967].
ALPHA--[Kim and Mueller, 1978b; Mulaik, 1972:211-212].
IMAGE--[Kim and Mueller, 1978b].
RAO--essentially same as ML.
LJIFFY--[Kaiser, 1970].

3. Unalt--1.0 for R and variances for S.
MCS--multiple correlation coefficient between variable and all other variables.
Max--maximum correlation for variable with all others.
Priors--may input any values.

4. Mineigen--keeps all factors for eigenvalues larger than a stated minimum; default cutoff is 1.0.
r--one may choose the number of factors to keep and rotate.
%Trace--factors are extracted until a designated percentage of the trace is exceeded.

5. VARI--Varimax-[Kim and Mueller, 1978b; Mulaik, 1972:258-261].
QUARTI--Quartimax-[Kim and Mueller, 1978b; Mulaik, 1972:258-261].
EQMAX--Equamax--[Kim and Mueller, 1978b; Mulaik, 1972:262-263].
ORTHOG--[Harman, 1967:Chapter 12].
None--no rotation is requested.

6. PROMAX--[Kim and Mueller, 1978b; Rummel, 1970: Chapter 17].
DQUART--[Jennrich and Sampson, 1966].
ORTHOB--[Kaiser, 1970].
DOBLI--Doblimax-[Harman, 1967:334-341].
FUNCTNPLN--[Katz and Rohlf, 1974].
PRIMENPL--[Katz and Rohlf, 1975].
SPACEWARP--[Harris and Kaiser, 1964].

8. $D^2$--Mahalanobis d-square from each OTU to centroid for all data is computed, as well as for the r components retained, and for residuals; actually given as $D^2/m$ which is approximately chi-square/m [Hawkins, 1974].

ANDREWS PLOTS

This is a graphical method for systematically scanning an m-dimensional variable space in order to plot scores for OTUs as continuous functions of the directions chosen in the space. Andrews [1972] defined sets of coefficients that vary continuously, and that have properties that allow construction of *a priori* and *a posteriori* tests for comparing subsets of OTUs in directions of interest.

One way of looking at the procedure is to think of plotting the data on some initial arbitrary axis passing through the origin (or the centroid of the data); this axis could even be one of the original coordinate axes. Then the axis is rigidly rotated a small amount in some direction and new scores are computed. New scores are computed after every small angular displacement. If this is done systematically and thoroughly, then one scans the entire space--including directions corresponding to each of the original variables singly, and to any linear combination of the original variables, including canonical variates axes, principal component axes, etc. The scores of the OTUs on this scanning axis are plotted on the ordinate as a function of the displacement angle. This produces an essentially continuous line (when the size of the increment of the displacement angle becomes very small) for each OTU.

Andrews emphasized specific scanning paths that have nice mathematical properties in terms of finding confidence intervals or testing statistical hypotheses in particular directions (points on the path), or in directions chosen after an examination of the plot. He recommends transforming the variables to principal components for this purpose. The use of these paths as recommended by Andrews is limiting in that only a

relatively small part of the space about the first principal axis would be explored.

Other discussions have not limited this method to a principal components-oriented approach: Gnanadesikan [1977:209-210] suggests other weighting schemes that produce a more thorough scan of the space, but these do not have statistical properties that are as convenient. He also suggests trying a few different permutations of the coefficients. This will scan still more of the space. It is recommended for visual clarity that not too many OTUs be plotted in any one graph. Gnanadesikan gives some practical suggestions for representing large samples of OTUs.

Applications:

Applications are discussed in the section on "Exploration" in PURPOSES.

Computational requirements:

None.

Statistical assumptions:

For testing hypotheses or confidence intervals, it is necessary that the scores have the properties that satisfy the assumptions of the test statistic used. As the method generates univariate scores for any single test, then any appropriate univariate test may be used. For example, if the data have a multinormal distribution, then the scores in any direction will have a univariate normal distribution. If two or more groups of OTUs are present then t- or F-tests might be

appropriate for hypotheses of equality of means in the chosen direction. Andrews provides both a priori and a posteriori tests.

Statistical packages and computer programs:

The method is not available in the common packages. A program can be written in SAS79 using PROC MATRIX. PROC PLOT can then be used to produce plots for a limited number of increments.

# MULTIPLE REGRESSION

Multiple regression analysis is a technique for relating one observed variable to a set of m additional observed variables measured on each OTU. It may be used for: 1) prediction--for example, predicting body weight from a set of skeletal measurements, 2) describing or finding functional relationships--for example, relating measurements in a growth study to age, nutrition or environmental conditions, or relating morphological variables to geographic coordinates, or 3) removing the effects of one or more variables from a variable of interest--e.g., correcting for "size". A linear model of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + v_i \qquad \{1\}$$

is fitted to the data. $y_i$ is the value of the dependent or predicted variable for OTU i. $x_{i1}$ to $x_{im}$ are observed values for the independent or predictor variables for the same OTU. $\beta_0$ is the y-intercept, and $\beta_1, \beta_2, \ldots, \beta_m$ are unknown regression coefficients used to explain or predict a $y_i$ as a linear combination of the m $x_j$'s. $v_i$ is the error or residual variability--that part of $y_i$ not explained by dependence on the $x_j$'s. The model may be written in a compact matrix form as

$$Y = X\beta + V \qquad \{2\}$$

Y is a column vector of n values for the independent variable, X is an n x m+1 matrix of values for the independent variables (with a column of ones for "$x_0$"), $\beta$ is the column vector of m regression coefficients, and V is a column vector of residuals. The model in this form looks like the factor model in factor analysis. The difference is that in factor analysis only the analogue of Y (C in factor analysis) is observed and the analogues of X and $\beta$ (F and E) are estimated, while in regression analysis both Y and X

are observed and $\beta$ is the only vector of parameters estimated.

Many authors consider multiple regression a univariate technique, since there is only one Y or dependent variable. However, in some applications there may be m+1 random variables, since the values of X may be observed with sampling variability. (The implications of these alternatives--x's are fixed variables vs. x's are random variables--for hypothesis-testing, are discussed below under Statistical assumptions.) Just the inclusion of more than one random variable suffices as reason for some authors to consider multiple regression a multivariate technique. What is called "multivariate" multiple regression, in which there are p number of y's (dependent variables) and m x's (independent variables), produces exactly the same result as the p separate multiple regressions on the x variables, one for each y. This is not a method for relating a linear function of the y's to a linear function of the x's; the latter is achieved by canonical correlation.

The model for multiple regression is usually fit to the data by the least squares method; estimates $b_j$ of the parameters $\beta_j$ are determined to minimize the sum of squares of the residuals $v_i$. The derivation and formulae for doing this are given in most intermediate level statistical methods or biometrics texts [e.g. Snedecor and Cochran, 1967:Chapter 13], in books devoted to multiple regression [e.g. Draper and Smith, 1966], and in multivariate texts [e.g. Morrison, 1976:95-97]. Estimating the coefficients $b_0$, $b_1, \ldots, b_m$ and fitting the model to the data corresponds to finding the best fitting m-dimensional hyperplane or surface in an m+1 variable space. The least squares approach then minimizes the sum of squared distances parallel to the y-axis from the

observed points in m+1 space to the surface in m space. $b_0$ is the estimate of the y-intercept of the surface--the value of y for which all x's are zero. Each $b_j$ estimates the best fitting slope of the line formed by the intersection of the fitted surface and the $x_j y$ plane; that is the values of y when all x's except $x_j$ are equal to zero. The model may also be fit by setting $\beta_0=0$; in that case the best fitting surface is forced to pass through the origin of the space. In either case the surface passes through the means of all x's and y, i.e. the centroid of the data set. Generally it is better to leave $\beta_0$ in the model and test that it is zero [Snedecor and Cochran, 1967:166-167].

A measure of goodness of fit of the surface to the data is given by $R^2$, the multiple correlation coefficient squared. $R^2$ is also called the coefficient of determination. R is the correlation between the observed $y_i$'s and the predicted $\hat{y}_i$'s

$$\hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_m x_{im}. \qquad (3)$$

$R^2$ may also be interpreted as the proportion of the sum of squares of the $y_i$'s explained by the linear dependence on or association with the x's. $R^2=1$ means that the dependence is complete, and all of the y's fall on the fitted surface. $R^2=0$ means that y is uncorrelated with any of the x's (i.e. that none of the x's alone or in combination are any better predictors of y than is the mean of y). Since p+1 $\beta$'s are estimated, a perfect fit will be obtained with p+1 OTUs. Fewer OTUs will lead to an indeterminate solution, and more OTUs will allow estimation of the error or residual variance, and statistical inference if distributional assumptions are correct. Although $R^2$ is the most popular measure of goodness of fit of the regression model, useful information is frequently obtained from an examination of the residuals from the

surface of best fit. Hocking [1976:15-16] gives additional criteria for evaluating a regression model.

Several important problems arise in the use of multiple regression. One is how to choose the "best" subset of the x or predictor variables from a large number of available independent variables that are correlated with y. Another is the effect of highly correlated or nearly redundant x variables (i.e. determinants of the variance-covariance matrix of the x's near 0). This is the problem of so-called collinearities or multiple collinearities, which lead to unstable (i.e. difficult to reproduce under repeated sampling) and difficult-to-interpret vectors of b's, the coefficients. A very brief discussion of these problems is given here with references to appropriate literature. A comprehensive review has been given by Hocking [1976]. The problem of "outliers", which may play havoc with regression estimators and prediction, is discussed more extensively under PURPOSES. We only note here that possible solutions include "trimming" [Dixon and Massey, 1969:331] or using "robust" estimators (see section on "Robustness" in the INTRODUCTION), i.e. removing or giving lower weights to certain of the extreme values when fitting the model.

Stepwise regression has been a popular and widely available procedure for the selection of "best" subsets of variables. This is the case in which one has a large number of possible predictor variables and one wishes to reduce these to an optimal few. A variety of criteria are available for "stepping", i.e. sequentially adding or removing (or both) variables to the model. Four problems exist with all stepwise procedures: 1) they do not necessarily find the optimal subset, 2) the b's may be biased by the ad hoc selection procedure [Draper and Smith, 1966:81-85], 3) the over-all error rate of the procedure is complex and

not known, and 4) unwarranted significance is frequently placed on the order in which the variables enter the equation [Hocking, 1976:9]. The first problem may now be dealt with by doing all possible regressions (Hocking, 1977). For m variables, there are $2^m-1$ possible models--e.g. for 10 variables, there are 1023 models, or for 20 variables, 1,048,575 models. However, the programs, such as the one in SAS79, employ an algorithm that does not require searching among all possible models for each possible number r≤m of variables. (the algorithms are discussed in Hocking [1977]). The $R^2$ values for all m models with one variable are output, then $R^2$ for all models with two variables with larger $R^2$ than for one variable, and so on. This greatly reduces the number of models from which to choose. The coefficients can then be found for the models selected, and other computations performed on this restricted set.

The problem of collinearity has received much attention from statisticians recently but there is still not full agreement on the best procedures to use to deal with this problem [Chatterjee and Price, 1977:Chapter 7, and Hocking, 1976]. The problem arises when some of the x variables are highly correlated; the result of collinearity or high correlations among some x's is that the coefficients, the b's, for those x's will be much more sensitive to the same sampling variation in y than will the coefficients for x's with low correlations. However, it may be preferable in a functional model to get estimates of the relations between the dependent variable and the independent variables of interest; one might want to keep all the x's in the model rather than delete some of them as one might in a stepwise procedure. It can be shown that small eigenvalues of the variance-covariance matrix formed from the x's are indicators of collinearity

[Chatterjee and Price, 1977:166-167]. Because their reciprocals are employed, these eigenvalues strongly effect the values of the b's and their standard errors. Another measure of multicollinearity is the variance inflation factor (VIF). The VIF for each $x_j$ is $(1-R_j^2)^{-1}$ where $R_j^2$ is the multiple correlation coefficient squared of $x_j$ with the remaining m-1 x variables [op. cit.:182-183 for discussion of this measure and how it is used]. Two approaches or solutions to the problem of collinearity have been proposed:

1) The y's are regressed on the r<m principal components of $S_x$ with eigenvalues larger than a selected cutoff [op. cit.:Chapter 7]. (This method is available in BMDP and can be done in SAS through PROC MATRIX followed by one of the regression procedures.) The coefficients in the regression model are determined in terms of the principal components of the x variables. The b's for the original variables may be obtained by multiplying the regression coefficients for the principal components by the eigenvectors of $S_x$ for those components retained.

2) Ridge regression--a small constant is added to the variances on the diagonal of the variance-covariance matrix $S_x$. This increases the value of the smallest root and also stabilizes the estimates in a way similar to method (1) [op. cit.: Chapter 8 and Hocking, 1976]. This correction leads to biased but more precise estimates.

Applications:

Regression has been used for prediction or estimation: one might wish to estimate difficult-to-measure variables, such as brain volume,

from more easily obtained variables, such as linear skull measurements, or one might estimate values for missing data using those OTUs with complete sets of measurements to estimate the $\beta$'s for prediction of the missing y [e.g. Hoffman, Koeppl and Nadler, 1979:5 use the variable most highly correlated with the missing variable]. If one is estimating values, it is frequently desirable to find an optimal subset of the x variables and reduce the number of measurements required for a good prediction of the y value; this can be done using stepwise regression.

Patterns of geographical variation for single morphological variables, principal component scores or factor scores may be estimated through the use of multiple regression by using geographical coordinates or functions of coordinates as independent variables. This technique is called "trend surface analysis" [Davis, 1973:322-332]. The latitude and longitude (or other coordinates, such as elevation, may also be included) may be expanded into a polynomial or other series. The coefficients are estimated and the estimated y's are plotted on a base map of the study area [Marcus and Vandermeer, 1966]. This is a kind of smoothing of the geographically distributed data. $R^2$ and the mean squared error of the residuals are used as measures of goodness of fit. This problem fits situation (1) below, under Statistical assumptions, since the geographic coordinates are measured with negligible error. Appropriate hypotheses may be used to test for the degree of the polynomial giving an adequate fit to the data (a simpler model being preferred). This is an a posteriori procedure. Low order surfaces may describe clines, for example. An examination and plot of the residuals is useful for detection of outliers, or local populations requiring additional study.

The x's in a multiple regression analysis may be random variables, described in case (2) below. This would be the situation, for example, when using the method to estimate missing values. However, if the goal is to find a functional or descriptive relationship, such as in allometric studies, when no single variable of the m+1 random variables may be designated as dependent (that is any of the variables could appear on the left side of the model (1)), then a principal components formulation would be more appropriate [Jolicoeur, 1963; Kuhry and Marcus, 1977 for the bivariate case, and some other considerations; see section on "Size and Shape" in PURPOSES].

Analyses of variance may be viewed as a form of multiple regression in which the x's are dummy or indicator variables indicating the design variables [Kleinbaum and Kupper, 1978: Chapter 13]. Similarly analysis of covariance may be described as multiple regression in which some of the x's are indicators (0/1, e.g. 0 for male and 1 for female) and others are continuous variables. This latter model allows testing of additional hypotheses if the assumptions for the case with fixed x's are valid, including comparison of regression coefficients between sets of OTUs [op. cit.: Chapter 14]. A clever application of indicator variables in multiple regression is a test proposed to distinguish "phyletic gradualism" from "punctuated equilibrium" in stratigraphic sequences of fossil mammal samples [Bookstein, Gingerich and Kluge, 1978].

Two-group discriminant function analysis may be viewed as a special case of multiple regression, in which the dependent variable is dichotomous. For example in a discriminant analysis used to identify species 1 and species 2, the b's are proportional to the discriminant coefficients if the $y_i$'s for species 1 are set to one constant, and the $y_i$'s for species 2

are set to another constant. The actual discriminant coefficients are obtained if the first constant is set equal to $n_2/(n_1+n_2)$ and the second constant is set equal to $-n_1/(n_1+n_2)$, where $n_1$ and $n_2$ are the sample sizes of the two groups [Kleinbaum and Kupper, 1978:420]. Mahalanobis $D^2$ may be written as a function of $R^2$:

$$D^2 = \frac{(n_1+n_2)R^2}{n_1n_2(1-R^2)} \qquad \{4\}$$

One value of knowing these relations is that any multiple regression program may be used to perform an analysis of variance or a two-group discrimination [op. cit.: Chapter 22 for a discussion on the uses of regression for ANOVA and discrimination].

An approach to multiple regresion for a discrete dependent variable is given in Press [1972:265-272]. This method of analysis is called logit analysis or logistic regression (see METHODS section on "Multidimensional Contingency Tables").

Computational requirements:

Multiple regression requires that the data be partitioned into one dependent variable (the left side of the equation {1,2} and m independent variables (on the right side of the model equation). The number of OTUs must be n>m, and there should not be any linear dependencies among the x's. Another way of saying this is that the rank of the variance-covariance matrix of the x's should be m. (The rank will not be m+1 because of the column of ones for "$x_0$".)

Note that the model is a linear combination of the $\beta$'s. Any data transformation, either linear or nonlinear, of the x's or of y is allowed within this

model, since both y and the x's are observed values and are not being estimated.

Statistical assumptions:

Two different sets of assumptions may be invoked, corresponding to the two approaches in regression mentioned above: one set in which the x's are fixed-quantities without associated measurement error, and an alternate set in which the x's are random variables. We discuss the hypotheses and assumptions associated with each of these cases in more detail:

1) The x's are fixed quantities measured without error, for example time since birth in a growth study or geographic coordinates of a collecting locality; the errors $v_i$ are independent and normally distributed with mean zero and unknown variance. These two assumptions imply that the y's are normally distributed with expected values equal to $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im}$ and the same unknown variance as the v's.

If these assumptions are valid then one may test a series of pertinent hypotheses:

* all of the $\beta$'s=0: there is no functional or predictive relation between y and the x's.
* $\beta_0$=0: the true surface passes through the origin.
* a specific $\beta_j$=0 or subsets of $\beta_j$=0: certain x variables or groups of x variables do not add significantly to the model.

All of these tests are t- or F-tests and may be organized in the form of an analysis of variance [Snedecor and Cochran, 1967: Chapter 13 and Kleinbaum and Kupper, 1978: Chapter 10]. Most of the tests can be expressed in terms of $R^2$ or partitions of $R^2$.

Morrison [1976:114-116] gives the a posteriori test procedures for testing hypotheses about the x's that seem to contribute little to the relationship, based on an examination of the b's that are small. Similarly confidence intervals or bounds may be found for the prediction surface itself, for predictions of future y values for a new OTU for which the x's are known, for the $\beta$s, and for the unknown residual variance or mean squared errors [ibid.].

2) The independent variable y and the m x's are all random variables having a multivariate normal distribution. The same tests and confidence intervals as in case (1) are appropriate after a specific set of x's have been observed, which are then treated as fixed. One's results are therefore conditional on the x's observed, as well as on the validity of the same assumptions required for case (1).

There are tests for the hypothesis of independence of the residuals against a variety of alternatives. Serial correlations among the residuals indicate non-independence and are the basis of some of the tests [Chatterjee and Price, 1977:Chapter 6]. Runs tests and other non-parametric tests may also be used to check independence. Some of these tests are available in the various statistical packages. It is recommended that one examine observed residuals $(y-\hat{y})$ or standardized residuals, and that they be plotted against y or various of the x variables to visually assess their magnitude and pattern. See Gnanadesikan [1977:263-271] for additional graphical procedures for evaluating residuals.

Biological assumptions:

The partitioning of the initial data set into the

dependent variable y and the m predictor variables, $x_1$ through $x_m$, is the first step in a set of assumptions about the relationships of the variables. If the purpose is prediction of a value within a taxon, as when brain volume or body weight is predicted from linear measurements, the assumptions would be only that the sampling over individuals was adequate. However, the linear combination of measurements that optimally predicts brain volume for one species would not necessarily be that combination optimal for a different species, even a species from the same genus. Yet, if values of y were entirely lacking for some species, an assumption of homogeneity over species within a genus or even family might be necessary. And even when the population being sampled is adequately represented, the regression equation obtained will still be a better predictor of that sample's values than it will be on the average for subsequent random samples. One can guard against an assumption of too good a performance by cross-validation with other samples (see section on "Validity" in the INTRODUCTION). Another less adequate solution is to adjust the coefficient of determinaion so that one obtains a more realistic measure of the predictive value of the regression equation [Thorndike, 1978:162-166].

If the purpose of the regression analysis is to derive hypotheses about functional relationships between variables, then assumptions about biological processes are being invoked, and are not tested in any fashion by the multiple regression analysis. For example, if one wished to evaluate the relative potency of various environmental and resource factors affecting the reproductive success of some kind of animal, then interpretation of the coefficients obtained from multiple regression would invoke assumptions about the biological or ecological processes mediating the cause(s) (the environmental measures) and the effect

(increase or decrease in reproductive success). This analysis would not provide a way of distinguishing a case of causal dependence of one variable on another from a case of common dependence on an unrecorded third variable. This is, of course, true for any analysis evaluating correlation. One way of formulating causal models in correlation studies is in the framework of path analysis [Wright, 1954; Tukey, 1954; and Nie et al., 1975].

If the variables of interest in the model are highly correlated then the problem of multicollinearites might arise, which will lead to unstable and non-interpretable regression coefficients. Corrections for this problem are discussed above.

Both models described under statistical assumptions above include the assumptions, among others, that the error or residual terms are independent. The errors in the values of observed random variables in univariate statistics are independent whenever the sampling is random (see section on "Independence and random sampling" in the INTRODUCTION). In the multivariate case, however, random sampling is not sufficient: assuming independence of the error terms becomes an assumption that no major factor or common source of variation affecting the y's is omitted from some sort of representation in the set of predictor variables [Draper and Smith, 1966:81-85].

Statistical packages and computer programs:

This procedure is one of the most widely available in the general purpose statistical packages. There are many special purpose packages devoted to multiple regression. One or more of these will be available at most computer centers. Below is a brief outline of the

procedures available in BMDP, SAS and SPSS.

BMDP(77) has a series of multiple regression programs and some related procedures, these include:

P1R--Multiple Linear Regression
P2R--Stepwise Regression
P9R--All Possible Subsets Regression
P4R--Regression on Principal Components
P5R--Polynomial Regression

Regression-related techniques are found in PAM for estimating missing data, and there are several procedures for non-linear regression.

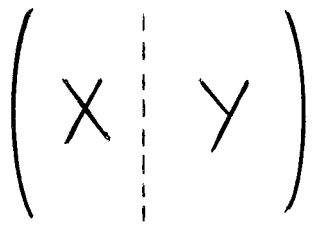
SAS has a series of multiple regression procedures, including:

PROC GLM--general linear model procedure, including general multiple linear regression
PROC RSQUARE--all possible regressions
PROC STEPWISE--stepwise multiple regression
PROC SYSREG--many variations of multiple linear regression including "seemingly unrelated regression" models

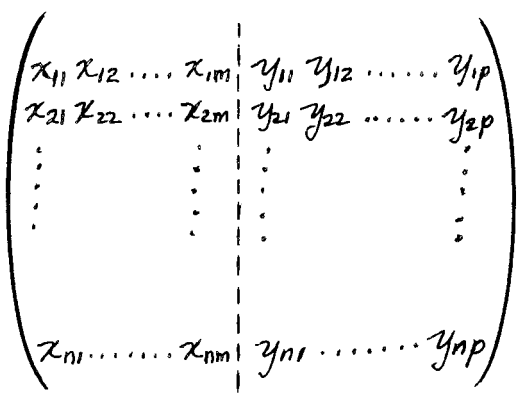
Procedures related to multiple regression are PROC AUTOREG for autoregression models (e.g. for time series or spatial autocorrelation analysis [Sokal and Oden, 1978a and b] and PROC NLIN for nonlinear regression. PROC PLOT may be used to plot residuals etc., and PROC MATRIX may be used to program one's own procedure for regression.

SPSS has a general regression procedure that includes various stepwise options. Chapter 21 of the SPSS manual [Nie et al., 1975] titled "Special Topics in General Linear Models" includes uses of the regression procedure for polynomial regression, analysis of variance using dummy variables and for path analysis models.

CANONICAL CORRELATION ANALYSIS




The association between two sets of variables may be studied by finding the canonical correlations between the two sets. For example, this method would describe the association between m morphological and p environmental variables measured over the same set of OTUs or individuals. The data matrix will then have n rows for OTUs and m+p columns for variables. If we define X as the submatrix corresponding to the m columns of one set of variables (e.g. the morphological variables) and Y as the p columns for the other set of variables (e.g. the environmental variables), then canonical correlation analysis (CCA) finds those linear combinations of the x's and of the y's that are maximally correlated subject to an orthogonality constraint. That is, CCA will find that combination of the y's that has maximum correlation with any linear combination of the x's, and also that linear combination of the x's which has maximum correlation with any linear combination of the y's. Out of the infinite number of linear combinations of x's and y's, CCA finds that particular pair most highly correlated with each other. This pair is the first pair of canonical variates. The second pair of canonical variates are that most highly correlated pair out of all possible linear combinations orthogonal to the first variates, and so forth, until m (if m<p) pairs of canonical variates have been found.




To start, the entire n x (m+p) matrix is mean-centered by variables. The linear combinations:

$$b_{11}x_{i1} + b_{21}x_{i2} + \ldots + b_{m1}x_{im} = w_{i1} \qquad \{1\}$$

$$c_{11}y_{i1} + c_{21}y_{i2} + \ldots + c_{p1}y_{ip} = z_{i1} \qquad \{2\}$$

define a pair of canonical variate scores for the ith individual on the first canonical variates. The correlation between $w_1$ and $z_1$ over all individuals

is the canonical correlation between the first pair of canonical variates. There are as many pairs of canonical variates and their corresponding canonical correlations as the lesser of m and p, here taken to be m (and the data sets can be redesignated to always make this true). The m canonical variates for the x's, i.e. the w's, are all uncorrelated with each other, as are all the z's--the m canonical variates of the y's. The correlation between $w_j$ and $z_k$ is zero for j not equal to k, and for j=k is the jth canonical correlation between the jth pair of canonical variates.

The m x m matrix B has as its columns the sets of coefficients for the m canonical variables for the X data, while the p x m matrix C has as its columns the coefficients for the m canonical variables for the Y data. The columns of the matrices B and C are the eigenvectors of the following products of matrices respectively:

$$S_x^{-1}S_{xy}S_y^{-1}S'_{xy} \quad \text{and} \quad S_y^{-1}S'_{xy}S_x^{-1}S_{xy} \qquad \{3\}$$

The canonical correlations are square roots of the corresponding eigenvalues, which are the same for both expressions in {3}; that is, each member of a given pair of canonical variates has associated with it the same eigenvalue. The variance-covariance matrix of the n x (m+p) data matrix is an (m+p) x (m+p) matrix. It can be partitioned into the m x m variance-covariance matrix $S_x$ of the x's, the p x p variance-covariance matrix $S_y$ of the y's, and the p x m and m x p matrices $S_{yx}$ and $S_{xy}$ (which are each other's transpose) of covariances between the x's and y's. The formulation {3} may also be given in terms of partitions of the overall correlation matrix, i.e. in terms of the matrices $R_x$, $R_y$ and $R_{xy}$ [Morrison, 1976:256-257]. Canonical correlation analysis has been extended to more than two sets of variables [Gnanadesikan, 1977:69-79].

$$\left(\begin{array}{c|c} S_x & S_{xy} \\ \hline S_{yx} = S'_{xy} & S_y \end{array}\right)$$

Interpretation of canonical correlation coefficients, and the corresponding coefficient vectors in B and C for the canonical variates z and w, is discussed in Levine [1977], and in detail in Cooley and Lohnes [1971:Chapter 6] who compare the technique to factor analysis. The contribution of the original variables to the canonical variates may be assessed with difficulty by examining the coefficient vectors, and more easily by examining the correlations between the original variables and the canonical variates--analogous to structure coefficients in factor analysis [Levine, 1977:18]. The formulae for the latter are given in Cooley and Lohnes [1971:170].

Cooley and Lohnes also discuss the calculation of the proportion of variance extracted from each data set and the redundancy of the two data sets. The redundancy of a data set with respect to a canonical correlation is the proportion of variance of the canonical variate for the data set [Cooley and Lohnes, 1971:170] times the canonical correlation coefficient squared. It may happen, for example, that for the X data the first canonical variate axis is near the first principal axis of X and therefore explains a substantial proportion of the variability in the X matrix. On the other hand, the corresponding canonical variate axis of the Y set may be near one of the component axes corresponding to a small eigenvalue of that set and therefore explains little of the variability of the Y set. The redundancy for the Y set will then be smaller than for the X set.

Cooley and Lohnes note that prior to the recognition of the idea of redundancy canonical correlations were frequently viewed as measures of the strength of the relationship between two sets of variables. The canonical correlations in fact only measure the degree of relationship between the canonical variates, the

corresponding w's and z's, which may not summarize important parts of their respective data sets. If the redundancy of each of the first canonical variates is low, then they are representing very little of the information or variability within their respective data sets, no matter how high the canonical correlation is between them. In such a situation, one might expect those canonical variates to be rather difficult to interpret; it is usually the major directions of variation that are most susceptible to explanation but there is not always reason to expect the directions of high variance within each of two sets of variables to necessarily be highly correlated between the sets.

Applications:

There has been limited application of canonical correlation analysis in systematic studies. The method has been used to relate morphological and environmental variables [Calhoon and Jameson, 1970, and Karr and James, 1975] and has been more widely used in ecological studies. Bivariate or three-way plots of the canonical variate scores within each set of variables may provide interpretable ordinations [Folse, in press]. Lavelle [1977] studied the canonical correlation between tooth and long bone size in primates.

Multiple regression may be viewed as a special form of canonical correlation analysis, in which p=1. The canonical correlation coefficient is then equal to the multiple correlation coefficient R (see section on "Multiple Regression").

Discriminant analysis and multivariate analysis of variance (MANOVA) may also be viewed as special forms of canonical correlation analysis. One set of

corresponding w's and z's, which may not summarize important parts of their respective data sets. If the redundancy of each of the first canonical variates is low, then they are representing very little of the information or variability within their respective data sets, no matter how high the canonical correlation is between them. In such a situation, one might expect those canonical variates to be rather difficult to interpret; it is usually the major directions of variation that are most susceptible to explanation but there is not always reason to expect the directions of high variance within each of two sets of variables to necessarily be highly correlated between the sets.

Applications:

There has been limited application of canonical correlation analysis in systematic studies. The method has been used to relate morphological and environmental variables [Calhoon and Jameson, 1970, and Karr and James, 1975] and has been more widely used in ecological studies. Bivariate or three-way plots of the canonical variate scores within each set of variables may provide interpretable ordinations [Folse, in press]. Lavelle [1977] studied the canonical correlation between tooth and long bone size in primates.

Multiple regression may be viewed as a special form of canonical correlation analysis, in which p=1. The canonical correlation coefficient is then equal to the multiple correlation coefficient R (see section on "Multiple Regression").

Discriminant analysis and multivariate analysis of variance (MANOVA) may also be viewed as special forms of canonical correlation analysis. One set of

variables represents the measured variables, e.g. linear measurements, for OTUs in different groups and the other set is called the "design matrix" and usually consists of 0's and 1's. For example, in k-group discriminant analysis, if there are $n_1$ OTUs from group 1, $n_2$ OTUs from group 2, etc., then the data matrix X has $n_1+n_2+...+n_k=n$ rows and m columns for the m variables measured on all of the n OTUs. Y is defined as an n x k-1 matrix of 0's and 1's. The first column of Y has 1's for those OTUs in group 1 and 0 everywhere else. The second column of Y has 1's for those OTUs in group 2 and 0 everywere else and so on through group (k-1). The kth group will have 0's for all k-1 columns. This is just one way of writing the design matrix for a one-way ANOVA [see Morrison, 1976, or Kleinbaum and Kupper, 1978:256-258]. The X and Y data sets are then used in a canonical correlation analysis. The coefficients in B for the x variables will define canonical variates whose scores will be proportional to those obtained in a multiple discriminant analysis. The test for the number of canonical correlation coefficients different from zero will correspond to the test for the number of canonical variates. The magnitude of the canonical correlations squared will be related to the overall ability of the canonical variates to discriminate the groups. Canonical correlation analysis and multiple discriminant analysis are solving the same problem, so that additional analogues are present. Similarly, since a multiple discriminant analysis can also be viewed as a form of multivariate analysis of variance, then in addition, because of the equivalence of canonical correlation and discriminant analysis, MANOVA provides another way of solving or viewing the same problem.

Computational requirements:

The n x (m+p) data matrix has m variables assigned to one set and p variables to the other. The variance-covariance matrices $S_x$ and $S_y$ must have inverses (that is they must both be nonsingular) as dictated by equations {1}. For actual measurements, singularity rarely occurs (or is "very unlikely") unless linear combinations of original variables are used, or if rows in one matrix are duplicated and reduce the number of different rows to a number less than m or p. This latter situation might arise if one had exactly the same values (e.g. day length, monthly mean precipitation, etc.) for each individual at a station, for example, and fewer such stations than variables. For community studies in which relative frequencies of species are reported, a linear dependency arises if, for example, each row of X adds to 100%. This is corrected by leaving out one species; which one doesn't matter.

Statistical assumptions:

If the x variables and y variables have a joint multinormal distribution, a large sample test of the hypotheses that the true canonical correlation is zero is applicable [Cooley and Lohnes, 1971:175]. The test statistic has an approximate chi-squared distribution. The largest canonical correlation is first tested. If the null hypothesis is accepted, there is no evidence for any canonical correlation between the two data sets. If the null hypothesis is rejected then the next canonical correlation may be tested and so on. This sequential set of tests gives a kind of stopping procedure for the number of canonical correlation coefficients to retain in a study.

Tests or confidence intervals for the coefficient vectors in B and C are unavailable as far as we know

(except in the case p=1, and then the problem is one of multiple regression--see the section on that method for tests). Large sample sizes seem to be required for reasonable faith that the results of an analysis are near the true values for the parameters. Some rules of thumb for required sample sizes indicate very large samples sizes are needed relative to the typical number of variables studied and may rule out the methodology as useful for many systematic studies [Thorndike, 1978 and references therein]. The method may be especially sensitive to outliers. Robust procedures are likely to be appropriate, and jackknife type procedures can be applied but require considerably more computer time (see sections in the INTRODUCTION on these topics). One method for validation of results is sample splitting. The canonical correlations and coefficient vectors are estimated from each half of the randomly split data and used to compute canonical variate scores for the other half; the two sets of scores and canonical correlations may then be compared.

Biological assumptions:

No distinction is necessarily made between predictor and criterion variables in this analysis. If one does hypothesize causal or functional relationships, then assumptions about processes are invoked, as discussed in the METHODS section on "Multiple Regression".

As with other canonical methods, the implicit assumption is that the important relationships among variables and between sets of variables are linear. If one has reason to suspect nonlinear relationships, one approach is to find appropriate transformations to enable analysis of such relationships by this method. Nonlinear canonical correlation analysis, analogous to nonlinear principal components, might be appropriate

but we are unaware of the availability of such procedures.

Statistical packages and computer programs:

BMDP procedure P6M gives the canonical correlations, coefficient vectors, structure coefficients, canonical variate scores and test statistics. Multiple correlation coefficients of each variable with those in the second set are also given. Data in the form of raw data, an overall variance-covariance matrix or a correlation matrix may be used as input. The data may be mean-centered or not. Canonical variables may be plotted against each other, or against original variables.

SAS79 does canonical correlation through procedure PROC CANCORR. Canonical correlations, coefficients for canonical variates and Bartlett's test statistic are given. The structure coefficients are given. Canonical variate scores are available and may be plotted through PROC PLOT.

SPSS accepts raw data or a correlation matrix as input to the procedure CANCORR. The canonical correlations and coefficients for the canonical variates are output, and canonical variate scores may be requested. Wilk's lambda is given for testing the null hypothesis of no linear association between the canonical variates; probabilites are given. There are several missing data options. Linearly dependent variables are automatically deleted, but the test statistics' degrees of freedom are not adjusted for this correction. When missing data cause some eigenvalues to be negative, the procedure still finds a solution to a problem which is "only an aprroximation of the problem asked for by the user".

Cooley and Lohnes [1971:194-200] have listed a FORTRAN program for canonical correlation analysis. The input data are a correlation matrix of all of the correlations of the m+p variables. Canonical correlations, coefficient vectors, structure coefficients, redundancies, and proportion of variance extracted for each set of variables is given. Appropriate test statistics are also given.

Data obtained from discrete or categorical variables can be analyzed in the form of contingency tables. The 2 x 2 table is the simplest example of such data in which each specimen is classified and counted in its appropriate cell. For example, $n=n_{++}$ individuals might be classified:

| | Stripe present | Stripe absent | |
|---|---|---|---|
| Pupil round | $n_{11}$ | $n_{12}$ | $n_{1+}$ |
| Pupil oval | $n_{21}$ | $n_{22}$ | $n_{2+}$ |
| | $n_{+1}$ | $n_{+2}$ | $n_{++}$ |

A "+" subscript means summed over the values of that subscript, e.g. $n_{+1}=n_{11}+n_{21}$. The character on each margin may have more than two states; the resulting table will have dimensions equal to the number of rows times the number of columns. These tables are usually analyzed using chi-square tests for hypotheses of differences in proportions between rows or columns, or overall tests of independence [Sokal and Rohlf, 1969: Chapter 16]. Sokal and Rohlf also present an alternative G or log-likelihood ratio test for the same hypotheses.

This representation can be extended with more than two characters to multidimensional tables. There are as many margins or dimensions as there are variables of interest, although such tables cannot be simply presented. For example, five characters, 3 with 2 states and 2 with 3 states, produces a table with 72 cells (2x2x2x3x3=72), and 10 characters with 2 states each yields $2^{10}$ cells. Since the basic data are counts or frequencies, clearly large samples of specimens would be required to obtain many cells with counts of other than 0 or 1 if more than a very few characters are considered. A multiway table can always

be separated into a series of two-way tables, and separate chi-square tests done on them. However, it is difficult to partition and interpret the interactions of the variables in a multiway table using this type of chi-square analysis; the G test lends itself more to such partitioning [op. cit.:602-607].

The tables of counts may be converted to tables of relative frequencies by dividing each element by the total number of specimens. These frequencies are estimates of the probabilities with which each combination of character states will occur. Hypotheses about the equality of probabilities for rows or columns, or the independence of characters, may be stated in terms of products of the probabilities. For example, the null hypothesis of independence for the two-way table may be stated, $p_{11} = p_{1+}p_{+1}$. The probabilities being estimated in the entire two-way table are:

| | Stripe present | Stripe absent | |
|---|---|---|---|
| Pupil round | $p_{11}$ | $p_{12}$ | $p_{1+}$ |
| Pupil oval | $p_{21}$ | $p_{22}$ | $p_{2+}$ |
| | $p_{+1}$ | $p_{+2}$ | $p_{++}$ |

The hypotheses are tested by comparing two values: 1) the value for a cell, $p_{11}$ for example, predicted by the estimates of the marginal probabilities (here estimated by the marginal frequencies), with 2) the values or frequency observed for that cell.

The hypotheses can be rewritten in logarithmic form, which results in linear expressions. The example above would become: $\log p_{11} = \log p_{1+} + \log p_{+1}$. Linear models can be generated in this way to partition probabilities for the multiway contingency cells in terms of the character states or combinations of states. These so-called "log-linear models" have only

recently started to be incorporated into some of the statistical packages (separate log-linear programs are also available). The models look like analysis of variance models or linear models in general and can be used to test hypotheses and estimate parameters.

If one of the classification or marginal variables indicates taxon, then such tables could be used to test hypotheses about the relative frequencies of character states or their proportions in different taxa, or to develop discriminant functions based on discrete or categorical data. If one is interested in the relative proportions of character states or character state combinations for two taxa, the model may be written in terms of the "odds ratio" for two taxa [Colgan and Smith, 1978:158]; in this form the model is called the "logit model" or logistic model. (This name is derived from the relationship between this form of the model and the equation for the logistic curve [Bishop, Fienberg and Holland, 1975:358].)

The logistic model has been presented in the form of "logistic regression" analysis as an alternative to classical discriminant analysis [Press and Wilson, 1978]. The difference between the logistic formulation and the classical discriminant formulation is in the form in which the a posteriori probabilities of specimens coming from the two populations are presented. Logistic regression leads to the same discriminant procedure for normal distributions and homogeneity of covariance matrices as classical discriminant analysis. However, logistic regression may also be used for categorical variables, continuous variables, and mixtures of categorical and continuous variables [see Bishop, Fienberg and Holland, 1975:357-361 for a discussion of various models]. Press and Wilson [1978] claim that the logistic approach is relatively robust, "i.e., many types of

underlying assumptions lead to the same logistic formulation". The estimation procedure for the logistic regression coefficients is iterative; a classical discriminant function can serve as the starting point for iterations.

One attractive feature of the log-linear approach is that "logical zeros" can be adjusted for. A logical zero arises when cells of the table are empty because character combinations are impossible and known to be zero. (An example might be antler shape for females of certain species of cervids.) These are taken care of in the estimation procedure and computer algorithms which fit models and provide test statistics.

We know of no direct applications to systematics problems. Most of the applications in biology have been in ecology and animal behavior (see Fienberg, [1977] for the former, and Colgan and Smith, [1978] for the latter). Bishop, Fienberg and Holland [1975] is a comprehensive book on the subject with theory, a geometric development, and biological and other examples. This volume includes alternative approaches to treating multidimensional contingency tables [op. cit.: Chapter 10 for a comparison of methods].

A log-linear program called LOGLIN has been developed by Olivier and Neff [1976] and is available from those authors. PROC FUNCAT in SAS assumes a different estimation procedure than generally advocated by Bishop, Fienberg and Holland [1975], but includes the log-linear model (the differences are briefly discussed in the SAS79 manual [Barr, Goodnight and Sall, 1979:232-233]). Program P3F in BMDP fits a log-linear model to multiway frequency tables. An extensive discussion with examples is given in the BMDP77 manual [Dixon, 1977].

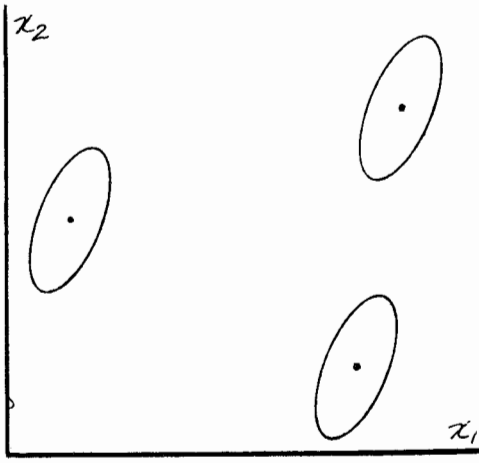
# DISCRIMINANT ANALYSIS

The term discriminant analysis refers to a variety of techniques that all have in common a partition of the n x m data matrix into groups of specimens (rows) by an *a priori* classification. The primary goal of discriminant analysis proper, i.e. classical discriminant analysis, is to develop a linear function of the characters, measured on an original sample partitioned into known groups; this function produces a score on the basis of which one can assign additional individuals to those *a priori* groups (using observed values of the same traits) with a minimal number of errors. Classical discriminant analysis procedures were derived by R. A. Fisher, under assumptions of multivariate normality and homogeneous variance-covariance structure among groups. Closely related to the goals of discrimination are techniques for displaying specimen scores in a space of reduced dimensions that maximizes the amount of variation among groups relative to that within groups. The vectors spanning this space are found through an eigenvector procedure, so the new variables formed from these eigenvectors and plotted in this space are called canonical variates or variables, and this form of analysis is sometimes called canonical variates analysis. Linear discriminant functions in a classical discriminant analysis can also be derived from the scores produced in a canonical variates analysis and are contained entirely with the canonical variates space. Multivariate analysis of variance (MANOVA) is a method mathematically and structurally closely related to discriminant analysis, but which emphasizes the testing of hypotheses of similarity and difference among the centroids of the *a priori* groups.

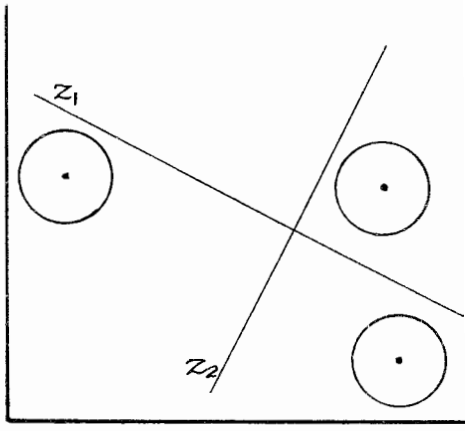
The distinctions just made, among classical discriminant analysis, canonical variates analysis, and

MANOVA, emphasize the differences in applications or goals of these variously named methods; the underlying mathematical structure of these three approaches is the same [Porebski, 1966]. However, this model is restrictive in several ways: only linear relationships among continuous characters are described, and multivariate normality and homogeneity of covariance structures are assumed. Alternate approaches have been developed that do not make either of the latter assumptions, produce nonlinear functions, are computed on discrete data, or are combinations of these conditions. However, the development of such approaches by statisticians has been largely directed at procedures for evaluation of the assignment of new individuals to groups--discriminant analysis proper. Statisticians have generally not been made aware of the interest of systematists in such alternate approaches for the description and low dimensional display of group differences and relationships. Therefore, most discussion of methods outside of the general underlying model is in the section below on discrimination in the restricted sense.

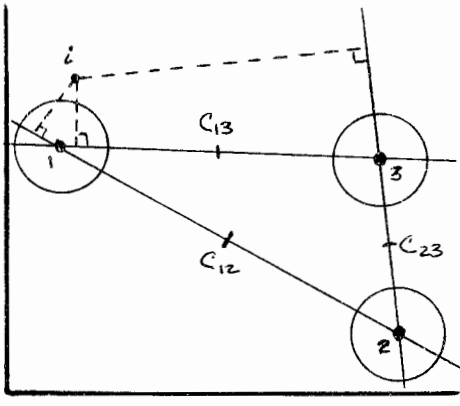
Before discussing the three orientations within discriminant analysis separately, we describe the common underlying model. In the general case, m variables are measured for each individual or specimen known to belong to one of the k groups. There are $n_1$ specimens in the first group, $n_2$ specimens in the second group, and so on to $n_k$ specimens in the kth or last group. An intuitive understanding can be gained by considering first a geometric description of what is done in the general case. Assuming multivariate normality within groups, the same sample sizes, and the same variance-covariance matrix within each group, the first step would be to think of a shearing and stretching transformation of the space such that the 95% concentration ellipsoids become spheroids. The

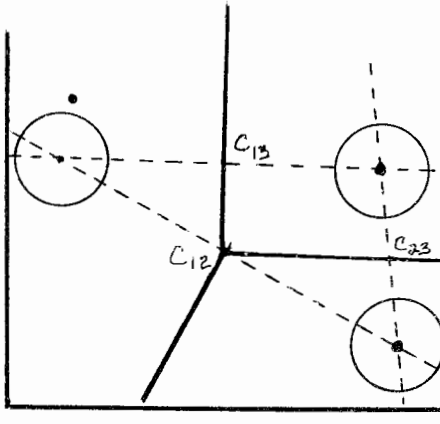


maximum dispersion among the groups would now be summarized by a principal components analysis of the centroids! The first principal component axis of centroids in this transformed space is the first canonical variate axis plotted in this space, and the second component axis is the second canonical variate axis, and so forth. And each individual would have a score on each canonical variate. This view also makes understandable the number of dimensions of the results: the centroids will be contained within a space of k-1 or m dimensions, whichever is fewer, just as in the case of the individuals in variable space in PCA.

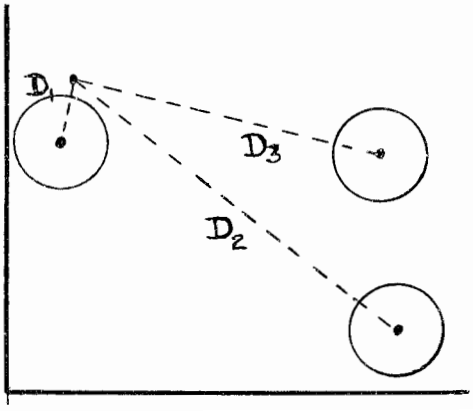



The discriminant function axes in this space are lines connecting the centroids. Consider a point representing an unknown and previously unidentified specimen (unidentified in the sense of not yet having been assigned to a group in the analysis). A cutoff point is designated on each discriminant function axis, usually at the midpoint between the centroids unless unequal a priori weights are assigned to the groups. The projections of the unidentified point will then fall on one side or the other; the point is assigned to that group which is determined by the classification rules [see Morrison, 1976:240]. A simpler picture results if one draws in, instead, the lines perpendicular to the discriminant function axes, intersecting the axes at the cutoff point. This divides the space into regions; the region into which a point falls determines its group assignment.




Another method for assignment of unknowns depends on the distance in the transformed space from the point to each centroid: the individual is assigned to that group to whose centroid it is closest. This distance, which is a simple Pythagorean distance in this space, is the Mahalanobis distance in both spaces, before and after transformation. (This will be demonstrated



shortly.) The classification functions available in several packages, and the determination of a posteriori probabilities for group assignment, are based on the Mahalanobis distance from the point to each centroid.

In this explanation of the geometry, we assumed within each group multivariate normality, equal sample sizes, and equal variance-covariance matricess, to simplify the description. If the populations from which the samples were drawn are not multivariate normal, or if one is only interested in the data at hand (and does not intend any inference about a population), the 95% concentration ellipsoids can no longer be estimated or are not relevant. However, the points are still there, forming an aggregation in m-dimensional space. If the variance-covariance structure varies among samples (even if the populations are normal), the initial hyperellipses will not be all the same size nor oriented in the same directions. The transformation of the space can still be done, but the resulting scatters will not have become spheroid, since the pooled variance-covariance matrix is used, thus transforming the space by an average of the covariance structures. Finally, if the sample sizes are unequal, the orientations of the canonical variate axes will be affected by the relative sizes of the samples in the different groups. Although statistical inferences will not be valid if these assumptions are violated, an analysis may still be useful for descriptive purposes (see page 157 below).

To look at a matrix representation of the procedure just described, let us go back to the k groups in the original m-dimensional space. The vector of means for variables is computed for each group. This defines the centroid for the group in the (k-1)-dimensional space (or the m-dimensional space if $m < k-1$). The among-groups variance-covariance matrix $S_A$ and pooled

within-groups variance-covariance matrix $S_W$ are computed as multivariate analogues of the among-groups mean square and within-groups mean square in a one-way analysis of variance. The canonical variates which summarize most of the variation among groups relative to the variation within groups are obtained from the eigenvectors of

$$S_A S_W^{-1} \qquad \{1\}$$

as weighting coefficients for the variables. Equation {1} is frequently given in terms of the sums-of-squares and cross-products matrices, i.e. the corresponding variance-covariance matrices multiplied by their degrees of freedom. The analyses are equivalent.

The eigenvectors of {1} define vectors of cosines of angles between the original variable axes and the canonical axes along which the canonical variate scores may be plotted (see the discussion of eigenvectors in "Principal Component Analysis", page 54). The scores of the individuals, plotted in two or three dimensions, give a visual demonstration of how good the separation is among the groups for the first two or three canonical variates, and thus give an indication of how well the discriminant functions will perform in assigning the original specimens to their proper groups. (For a large number of groups and variables, one may need to look at more than three axes.) While the canonical variate scores are uncorrelated for different axes, the canonical axes are usually not orthogonal in terms of the original data space. (The canonical variate axes, being axes for principal components among centroids, were orthogonal in that transformed space. If one transforms the data and axes back into the original space, the axes usually do not remain orthogonal.) The Euclidean distances between centroids in the canonical variates space have been named Mahalanobis distances, usually designated by D.

Mahalanobis distance squared may be defined for any two centroids and the pooled variance-covariance matrix as:

$$D^2=(\bar{x}_1-\bar{x}_2)'S_W^{-1}(\bar{x}_1-\bar{x}_2) \qquad \{2\}$$

where $(\bar{x}_1-\bar{x}_2)$ is the vector of differences in means of the variables for the two groups.

The linear discriminant function between two groups is a product of a discriminant coefficient vector and the vector of measurements for some individual i:

$$x_i'S_W^{-1}(\bar{x}_1-\bar{x}_2) \qquad \{3\}$$

The within-group variance-covariance matrix times the vector of differences in means between the groups is the discriminant coefficient vector. Assignment rules can be based on projections of points onto these discriminant function axes, or more simply, on the Mahalanobis distance squared between each individual and each of the k centroids. The specimen is assigned to that group to which its distance is smallest. In the output of many statistical packages, a posteriori probabilities are computed from the distances a new specimen is assigned to the group for which its a posteriori probability is greatest.

Computational requirements:

All of the discriminant analysis procedures require that the data have been assigned to a priori groups. This a priori classification of specimens in groups is based on variables other than those used in the analysis. The inverse of the $S_W$ matrix must exist, i.e. there must be more variables than the number of specimens minus the number of groups ($m > n-k$).

Statistical assumptions:

In classical discriminant analysis, the discriminant procedures will be optimal if the data have a multivariate normal distribution within each group. Statistical tests in canonical variates analysis, and tests and confidence intervals for Mahalanobis $D^2$, depend as well on the assumption of equality of variance-covariance matrices among groups. There are tests for equality of variance-covariance matrices but they are sensitive to normality assumptions, and their use is not highly recommended (see discussion on Homogeneity in the INTRODUCTION). The robustness of the two-sample test for equality of centroids has been studied, as well as the behavior of the discriminant analysis procedures for two groups when normality assumptions and equality of variance does not hold [Ito and Schull, 1964 and Lachenbruch, 1975].

Mahalanobis D is an estimate of the true Mahalanobis distance--the distance in the transformed, canonical variates space between the true means of the populations from which the samples are drawn. Mahalanobis $D^2$ is a biased estimate of this true distance squared in that it is on the average too large. It may be corrected for bias [Marcus, 1966 and Rao, 1952].

One will wish to know, for a discriminant analysis, how well the discriminant function performs. Note that "misclassifications" of individuals in the initial data, or other known individuals, are mistakes by the discriminant function; they cannnot logically be taken as mistakes in the a priori groupings (see page 17 in the INTRODUCTION). A summary of the assignments of specimens to groups may be given in terms of the number of correct and incorrect identifications in the form of a k x k contingency table. The probability of a correct assignment will be overestimated if such a table is based on the original data used to construct

the discriminant analysis. (The estimated error performance of classical discriminant analysis has been reviewed by Lachenbruch [1975].) A more realistic assessment of the probability of error will be obtained from another test data set whose a priori groups are also known, or if one has sufficient data, from a procedure of splitting the data into a training set and testing set. (Such a procedure is available in a number of the statistical packages.) However, when sample sizes are not large, the researcher is usually unwilling to withhold such data from the original analysis. The so-called "leave one out" or "jackknife" procedures (available in BMDP) produce better estimates of future behavior of the discriminant analysis (see page 31 in INTRODUCTION, on jackknife estimators).

The procedures and tests mentioned in this section are discussed in more detail in the separate sections below.

Discussion of the applications of discriminant analysis, and the biological assumptions involved in those applications, fall most logically into three divisions (based on the goal of the analysis) noted at the beginning of the section on this method. The variations that have been developed for various reasons (such as stepwise discriminant analysis, for finding maximally discriminating subsets of variables) are discussed within the appropriate sections.

1) Multivariate Analysis of Variance

In a taxonomic study, for example, in which the a priori groups of interest are populations or low level taxa, the null hypothesis that all of the groups have the same joint mean for the variables measured is a

multivariate extension of the usual one-way analysis of variance test for equality of k means [Sokal and Rohlf, 1969; Harris, 1975]. If the assumptions of multivariate normality and equality of variance-covariance matrices are valid, and if the null hypothesis is accepted, one concludes that the k groups of individuals represent random samples from one statistical population rather than from k populations with different centroids. If the null hypothesis is rejected, then one concludes there are some statistically significant differences among the populations, and that some of them have different mean vectors or centroids.

The multivariate tests for equality of two or more centroids (multivariate extensions of the two-sample t-test and k-sample one-way analysis of variance) have as assumptions that the observations are multivariate normally distributed with equal variance-covariance matrices. There is not full agreeement on the best test to use. [See Morrison, 1976:222-224, for a comparison of some of the those most frequently used.] Three of the tests proposed are functions of the eigenvalues of $S_A S_W^{-1}$ where $S_W$ is the variance-covariance pooled within groups and $S_A$ is the variance-covariance matrix among centroids. The test statistic usually presented in the statistical packages is a function of the product of the eigenvalues and is called Wilk's Lambda. However, there is some dissatisfaction with this statistic because it is possible for one to reject the null hypothesis of equality among all the centroids and subsequently have difficulty in finding the structure in the data that led to rejection [Harris, 1975:109-113]. Also, this statistic is asymptotic, which means that large samples are required for correct probability levels (see page 38 in INTRODUCTION).

Another statistic used to test the hypothesis of equality of the centroids (and also asymptotic) is the sum of the eigenvalues of $S_A S_W^{-1}$ or the trace of this matrix. This statistic is also a weighted sum of all of the Mahalanobis $D^2$ among the groups. A third statistic, strongly recommended by Harris [op. cit.] is based on the largest eigenvalue of $S_A S_W^{-1}$. In contrast to the previously mentioned tests, this largest root test is an exact test, so extremely large sample sizes are not required for correct probability levels. The largest root criterion is tabled in both Harris [1975] and Morrison [1976]. Both this test and Wilk's test may be used to determine the number of significant canonical variates. If the null hypothesis of equality of centroids is rejected, this is equivalent to identifying the first canonical variate as significant. Subsequent eigenvalues may then be tested in descending order of size.

The next problem is the assessment and interpretation of the nature of the differences among groups, if the null hypothesis of equality has been rejected. It is important to emphasize the difference between statistically significant differences and biologically interpretable differences. Very tiny differences will be statistically significant if one takes large enough samples. We must then have some measure of the size of the difference (e.g. Mahalanobis D), and some idea of what that difference means in terms of the problem and the biology of the organisms being studied.

There may be other null hypotheses of interest besides equality of means. Some are suggested here in terms of the hypotheses tested and the method of analysis: i) Is there the same amount of sexual dimorphism in the different groups? A two-way MANOVA would be appropriate to deal with this question,

looking at group differences, sex differences, and their interaction. ii) If the populations are distributed on some geographic or other gradient, is there a cline? The form of the cline may be tested by fitting polynomial regressions as a function of the distance in the form of linear contrasts or comparisons [Sokal and Rohlf, 1969:458-468; Morrsion, 1976:32-36, 197-204]. iii) Are groups of populations different on the average from other such groups that are separated by natural barriers? The differences would be examined by various forms of nested MANOVA, or again by linear contrasts. These hypotheses and many other possible hypotheses could frequently be a priori hypotheses since they would be suggested by one's knowledge of the organism and their environment.

The application of these kinds of tests are available in routines such as PROC GLM in SAS or other MANOVA procedures in other packages. They are analogous to the same kind of tests in univariate ANOVA, so that a knowledge of those procedures is a useful prerequisite for understanding MANOVA [Sokal and Rohlf, 1969; Snedecor and Cochran, 1967; and Kleinbaum and Kupper, 1978]. Most published applications seem to be limited to the one-way ANOVA null hypothesis of no difference among centroids. That is the only hypothesis tested in the discriminant analysis procedures in most of the packages.

Once the results of the analysis are in hand--the tables of means, the Mahalanobis distances, and the plots of the canonical variate scores for the individuals and centroids, then additional a posteriori hypotheses may be suggested. For example, comparing all pairs of groups, is analogous to the multiple comparisons problem in single variable analysis [Sokal and Rohlf, 1969:235-246]. However, comparisons in MANOVA are more complicated than in ANOVA as hypotheses

may be constructed involving both subsets of groups and subsets of variables. A survey discussion is given in Gabriel [1969]. A relatively simple technique for obtaining a conservative estimate of the overall probability associated with multiple tests of hypotheses can be obtained from the use of the Bonferroni method discussed in Morrison [1976:135-136], Harris [1975:98-101], and in "Statistical Inference" in PURPOSES in this manual. These references also give the general approach for any a priori or a posteriori comparison of interest, and Harris [1975:104] gives a table useful for setting up such tests.

A very general multiple comparison procedure for a posteriori comparisons over combinations of variables and groups based on the sums of squares simultaneous test procedure (SS-STP) has been developed by Gabriel [1969] (see Harris [1975:110] for a criticism of this procedure). Morrison [1976:197-204] and Harris [1975:103-106] give methods for finding confidence intervals for both planned comparisons and a posteriori comparisons after the null hypothesis of equality of group centroids is rejected in a MANOVA. For example, equality of western and eastern subgroups, or clines may be tested.

A confidence interval may also be obtained for the true $\Delta^2$ estimated by Mahalanobis $D^2$ using the non-central F-distribution. This is difficult to do from the published tables of non-central F; Bargmann [1970] has given an algorithm which can be programmed to determine confidence bounds for $\Delta^2$. Simultaneous confidence intervals for a matrix of $D^2$ values can be obtained by using Bonferroni inequalities, i.e. adjusting the individual confidence to $100(1-\alpha/c)$ where c is the number of $D^2$ values.

## 2) Canonical Variates Analysis

Canonical variates analysis (CVA) is the term widely used to refer to a classical discriminant analysis done with a goal of dimension reduction, usually with results expressed as a histogram or in bivariate plots. As pointed out above in the description of the general model underlying MANOVA, CVA, and classical discriminant analysis proper, assumptions of multivariate normality and equal variance-covariance structure within each group are necessary for statistical inference, or for the discriminant functions contained in the canonical variates space to be optimally discriminating. One may use the technique, however, with an exploratory intent, as a method of obtaining a low dimensional display of the groups in a space which has been "standardized" by the "average" variance-covariance matrix (weighted by sample sizes within groups).

The entire canonical variates solution for $k=3$ groups for $m>2$ variables can be displayed on a bivariate diagram, as also can the entire solution for $k>2$ groups and $m=2$ variables, since the dimensionality of the canonical variates space (and equation {1} above) is two in these cases. Three-dimensional "diagrams" or models may be constructed for $k=4$ groups and any number of variables, or for any number of groups and $m=3$ variables. For all larger m and k, a canonical variates analysis can be used as a dimension reduction procedure, since each canonical variate summarizes as much as possible of the variance among the centroids in the transformed hyperspace (see page 146) orthogonal to the previous canonical variates, just as principal components do among individuals in a character space. Thus the first canonical variate will summarize most among-groups variance, the second the next most, and so on. One hopes that the majority of

among-groups variance relative to pooled within-groups variance can be summarized by relatively few canonical variates, so that the resulting discriminant space will be low dimensional.

If there are $m>3$ variables and $k>4$ groups in the analysis, the distances among centroids or specimens can no longer be completely displayed in two or even three dimensions. The Mahalanobis $D^2$ between all pairs of centroids may be partitioned into two parts--one related to the axes displayed (or found to be significant; see "Statistical assumptions" below), and the other the residual part, computed for the remaining axes. If the residuals are small then the intergroup distances have been well represented by the low dimensional display.

Interpretation of canonical variate coefficients is a difficult problem. Since they are weights that are multiplied by the values for the specimen in order to compute the canonical variate scores, it is expected that one should be able to make meaningful statements about the importance of characters in the discriminant space displayed from their examination. However, they only indicate overall directions of difference among centroids (except in the two-sample case). Their relative importance depends on the magnitude of the standard deviation for the character they weight. Standardized coefficients, obtained by multiplying the canonical vector coefficients by the pooled standard deviation within groups for each variable, are unit free. They are the amount that the canonical variate score will change for each change of an original variable by one standard deviation.

The coefficient vectors are not orthogonal in the original m-dimensional space. In this respect they are like pattern coefficients in an oblique rotation of a

factor analysis solution. The correlation coefficients between the original variables and the canonical variates may also be found and are called structure coefficients [Cooley and Lohnes, 1971:247-248]. A difficulty with structure coefficients is that they have been defined in two ways: one in terms of the matrix $S_W$; the other in terms of the total sums-of-squares and cross-products matrix [ibid.; Bargmann, 1970].

Discriminant functions are defined and may be plotted in the canonical variates space. The discriminant functions are not the canonical variates nor the canonical variate axes, except in the special case of the two-group analysis. In a canonical variates analysis of two groups, the single canonical variate axis must be coincident with the line connecting the two centroids, the single discriminant function, since that will be the direction of greatest variation or distance between the centroids in the transformed space. In ($k \geq 3$)-group canonical variates analysis, the canonical variate axes will be in the directions of maximum dispersion among the centroids; these directions will very rarely be coincident with any of the lines connecting pairs of centroids--the discriminant functions.

Programs for discriminant function analysis in the statistical packages do not give the coefficients for the discriminant functions themselves (except when the canonical variate coefficients are given for the two-group case). What are usually given are the coefficients for the classification functions (the "discriminant functions" in SAS), which give a distance from an individual to a group centroid. The coefficients of these classification functions may be combined in pairs to form the discriminant coefficients of the discriminant functions, as originally defined.

The product of the vector of measurements for an individual and the vector of discriminant function coefficients is the score of that individual on that discriminant function.

One potential problem with scores for canonical variates is that the canonical coefficient vectors used to compute them are based on the $S_A$ matrix, which is a weighted sum of squared deviations of the group means from the grand mean. This means that each group's contribution to the pooled variability will be proportional to the sample size in each group. (Some computer packages allow a choice of weights.) But this means that the orientation of the canonical axes will be affected by the relative sizes of samples in the different groups. This weighting does not make much sense if the goal is to summarize intercentroid distances [Gower, 1966b and 1976]. Gower has suggested that the distances among centroids will be summarized better by a principal coordinates analysis of the k x k $D^2$ matrix between all pairs of groups. $D^2$ may also be corrected for bias before doing a principal coordinates analysis. It may be noted that the standard canonical variates analysis for displaying centroids is equivalent to a principal coordinates analysis of a matrix of weighted $D^2$ values, where $D^2$ for each pair of groups i and j would be multiplied by $n_i n_j/(n_i+n_j)$. (For example, with sample sizes varying from 5 to 25, this can inflate a $D^2$ for two samples of 25 five times relative to a $D^2$ for two samples of 5. In general this effect will inflate some $D^2$ values by a maximum of $n_{max}/n_{min}$.)

An alternative method would be to compute $S_A^*$ using unweighted sums of squares of the group means about the grand mean. For equal sample sizes over groups, the two procedures would be equivalent.

(Reyment and Banfield [1976] have applied Gower's suggested procedures to measurements of invertebrate fossils.) A useful way of examining the distortions of the true distances summarized in the reduced dimension plots is to superimpose a minimum spanning tree based on the D matrix. Those centroids which are connected, but are each closer to other centroids, must have a substantial distance component in some direction or directions not explained by the dimensions included in the plot.

Plots of the canonical variates scores are commonly used to display the scatter of individuals and the distance between the group means in a few dimensions. Equal frequency ellipses may be superimposed on the graphs to show the estimated bounds for the variability of the majority of each group (usually 95% or more). It is suggested that a separate bivariate ellipse be computed for each group separately for plotting on the bivariate canonical displays, as a way of viewing the projection of each group's multidimensional ellipsoid onto the two-dimensional plot. This provides a method of assessing the differences among the variance-covariance matrices (which were treated as homogeneous in the analysis) in the canonical space. The common practice of basing the ellipses (or circles if the axes are scaled to have standard deviation one) on the pooled variance-covariance matrix is not recommended, since such an ellipse or circle would not contain any information about the specific group.

If the groups are not homogeneous, i.e. their variance-covariance matrices are different, then the concentration ellipses will have different shapes. In this case concentration circles will not be appropriate, and the ellipses should be determined separately for each group. Confidence ellipses or circles may also be determined for the centroids and

displayed on the plots. Their area is a function of sample size, getting smaller as n increases. They should not be confused with the concentration ellipses however. The estimated concentration ellipses, on the other hand, converge to the true population concentration ellipses as n increases.

Another very useful graphical procedure is to plot projections of vectors along the axes in the original variable space on the plots of canonical variates. The vectors may be scaled to have length one in the original coordinate space or a length proportional to the standard deviation. The length of the original vectors must be stated for the plots to be interpretable. Jolicoeur [1959] uses this graphical display of the relation of the original variables to the canonical axes in his interpretation of the contribution of the original variables to the canonical variates produced (1 vs. 2 and 3 vs. 4). Their relative contribution can be determined from the graph without having to look at a table of values (they probably can be measured as precisely on the graphs by the reader as the amount of information they convey warrants).

There are tests for the number of significant dimensions in the canonical space. These tests are often automatically produced in the output for the various discriminant analysis procedures. There is not full agreement on the best test for this purpose: Harris [1975:108-113] and Morrison [1976:222-224] discuss the alternatives; see also the earlier discussion in the section on MANOVA. The tests are all asymptotic tests (as long as $k>2$ and $m>2$), requiring large samples for determinations of the correct probability level, and thus for many studies with small samples will only be general indicators of the number of statistically significant axes. Interpretability of

the pattern of differences is also an important tool for evaluating the displays and the number of axes to retain and display.

Once the centroids are available as scores for the canonical variates (as they are in many of the outputs), it is useful to partition $D^2$ into a part explained by the dimensions retained and a part not represented by the displays or tables kept. This is simply done using the multivariate Pythagorean theorem on the mean scores. A scan of the residual $D^2$ matrix for large $D^2$ values (larger than the bias for example) will disclose groups that are in dimensions not summarized by the displays, or whose separation may be greater than that illustrated by the displays.

One common practice, probably used more in physical anthropology than elsewhere, is to plot individuals (e.g. "human" fossils) on the displays derived from samples of supposedly related taxa. This can be misleading. Even though the new individual may fall well within the 95% concentration ellipse of a specific group on the projection plotted, it may in fact be far from all groups in the display, perhaps in a direction perpendicular to the plot. This is easily checked by computing the distance from the new individual to all of the centroids included in the study. If the distance is small for some group then it is likely that, with respect to the variables measured, that individual could come from the statistical population for that group. The fact that $D^2$ has an approximate chi-square distribution can be used to estimate the probability that the unknown comes from a specific group. Rao (1965:492) has provided a test for the hypothesis that an individual comes from a population whose centroid lies on a line connecting two groups' centroids.

Much of the usefulness of canonical variates analysis lies in the production of plots of the groups in a space of reduced dimensionality. The calculations to produce the canonical variate axes require pooling of the variance-covariance matrices, which may be undesirable if the variance-covariance matrices are very different. When the variance-covariance matrices are heterogeneous among groups, quadratic discriminant functions may be found for discrimination between groups two at a time. The individual variance-covariance matrices for each of the groups are employed, with the result that there is no longer a single transformed space in which to plot all the groups. However, Mahalanobis $D^2$ can still be defined from each group with respect to another, as well as a $D^2$ between an individual and a group centroid, using the separate variance-covariance matrices of the groups. The asymmetrical matrix of distances so determined can be represented in few dimensions using principal coordinates or non-metrical multidimensional scaling. Since the distances will usually not be metrics when there is excessive heterogeneity in the data, one should be aware of the possibilty of obtaining negative eigenvalues (see section on "Principal Coordinates" in METHODS, above). However, if negative eigenvalues are obtained and they are small, then the representation in a Euclidean space will be an adequate summarization.

3) Discrimination

Classical discriminant analysis has the goal of the assignment of unknowns to predefined groups. This application is perhaps the least common in systematics, although the largest effort has been expended on its theory in the statistical literature. However, the identification of unknowns is a question of major

interest in many disciplines. If groups have been found to be different under a priori or a posteriori tests of hypotheses, then subsequent questions of interest might be about the amount of overlap between groups, the combination of characters which most accurately identify individuals to the groups identified as different, or the probability of error in identifying these new individuals to the defined groups. The first and last questions are of course intimately related: if there is no overlap (and unlikely to be) then there will be little or no error in assignment of new unknowns, provided they belong to one of the groups in the study.

In classical discriminant analysis, the most useful descriptive statistic for summarizing intergroup differences is Mahalanobis distance, D, or distance squared, $D^2$. The values of the statistic are usually given in the form of a k x k table comparing each group with all others. An examination of this table will indicate the separateness of some groups, or the clustering of others. In fact a cluster analysis of groups may be based on the D or $D^2$ matrix [Rao, 1952].

Most of the discriminant procedures available in the statistical packages compute the Mahalanobis distance from each individual to the centroid of each group: the individual is put in the group to which is closest. These distances or functions of them are usually given as "classification functions". New individuals may also be identified in this way, with respect to the groups defined in the analysis. However, the distance may be large between an individual and all groups under consideration. In that case the individual should not be assigned to a group, but should be placed in a category "other". (Unless such a category is included, the restrictive assumption is required that all

individuals belong to one of the k groups in the analysis; this would seem to be frequently unrealistic for many systematic applications.) The actual D or $D^2$ value for each individual can be scanned for excessively large values. $D^2$ is approximately chi-squared distributed with m degrees of freedom so that if all of the $D^2$ values for any given individual are larger than some cutoff ($X^2_{.99,m}$ or larger for example) than the individual can be assigned to the category "other". In this case, the "other" individuals will be outliers and may only show up on plots of canonical variates associated with the smaller eigenvalues or in the results of techniques such as Andrews plots. For example Andrews [1972] was able to distinguish an outlier fossil primate Proconsul africanus as belonging to neither "human" or "ape" groups by its oscillation among groups in the plot. The original authors had determined that Proconsul africanus was more ape-like from their plots of the first two canonical variates determined from eight measurements of the first permanent lower premolar.

The a posteriori probabilities of group membership display relative distance information in that they are the relative probabilities of obtaining the distance from a point to each centroid. The a posteriori probabilities always sum to one because of the construction of the equation defining them, and will therefore not be a good source of information about specimens which don't belong to the k groups.

Rao [1965] has provided a test statistic to test the hypothesis that a given OTU belongs to one of two groups in a discriminant analysis. We know of no application of this test.

If multivariate normality assumptions are valid than Mahalanobis D can be used to roughly estimate the

amount of overlap and error for medium to large samples. Since the probability of error of assignment may be determined by using $\Delta/2$ (the parameter estimated by D/2) as a normal variate, D/2 (for large samples) may be used as a rough guide to the amount of overlap between groups. For example a $\Delta$ value of 3.29 implies about 5% error of assignment. ($\Delta/2$=1.645 is the one-tailed cutoff point for 5% of the distribution in the tail of a normal variate with mean zero and standard deviation one, i.e. the normal distribution usually tabled.) D is actually a biased estimate of $\Delta$; the correction for the bias is given in Marcus [1969] and Rao [1952]. A better way to estimate the error is by actually finding the proportion of errors in assigning new known individuals to the groups using the empirically determined discriminant functions. This can be done by dividing the sample into two groups--one used for finding the functions, and the other for testing it. However, the researcher seldom has large enough samples to afford this procedure. Leave-one-out (or so-called "jackknife") procedures have been developed [Lachenbruch, 1975 and available in BMDP] which essentially determine the discriminant functions leaving one OTU out and then assign it to its closest group. The proportion of errors for each group then gives a better estimate of the probability of error for future assignments than one based only on the usual procedure using the initial data.

If some variables are not useful, or contribute little to the analysis this can be tested, either by using stepwise type procedures (see methods above) or by deleting characters from the discriminant function and testing the reduction in distance and discriminating ability Rao [1970] and Kshiragar [1972]. Variables which do not contribute to the discrimination may actually decrease the ability of the discriminator [Van Ness and Simpson, 1976]. Ridge type discriminant

procedures have been described when characters are very highly correlated within groups. This has been recently applied in the use of discriminant function analysis in a study of habitat utilization by woodrats [Cavallaro, Menke and Williams, in press]. There has been a lot of interest in identifying suitable habitat for various species in wildlife management using discriminant analysis and some of the newer techniques such as robust discriminant analysis are being applied [Harner and Whitmore, in press]. In order to reduce the number of variables in a discriminant analysis, step-wise discriminant analysis procedures have been developed to find subsets of "best" discriminating variables. This is analagous to step-wise regression. SPSS and BMDP both have step-wise discriminant analysis procedures. A wide variety of "stepping" criteria are available. If the default criterion in BMDP is used for example, the first variable added is the one which gives the maximum F in a one-way analysis of variance over groups. Since F is the ratio of mean square among groups to mean square within groups, it is equal to a univariate formulation of $S_A S_W^{-1}$ and yields the best single discriminating variable. The next variable included adds most to the discrimination contributed by the first variable included. On additional steps, a variable may be included if it significantly (significance level chosen by the user) adds to the discrimination. After a variable is added, one of the variables already in the analysis may be deleted if its deletion does not change the discrimination significantly. In this way a discriminant procedure may be developed based on relatively few of the variables recorded. As with all step-wise procedures, an optimal solution is not necessarily found. The performance of the variables included in the analysis can be assessed using the "leave-one-out" procedure.

All of the testing procedures described so far

require the assumptions of multivariate normality and equal variance-covariance matrices within groups. When the variance-covariance matrices are not equal but the populations are still multivariate normal (graphically, the 95% (or other probability) concentration ellipses have different orientations or size), the optimal discriminant function is the so-called "quadratic discriminant function". It involves a larger set of discriminant coefficients which are multiplied by values for the variables as well as their squares and cross-products for an individual specimen. A quadratic discriminant function, however, is not associated with a canonical variate for displaying the relationships of the groups and specimens, because a single, transformed space is not produced to permit plotting the data. Mahalanobis D-like statistics as descriptive statistics comparing two groups are not direct results of this procedure. Actually, even when the variance-covariance matrices are different in various groups, the discriminant functions approach based on assignment to the group with the smallest $D^2$ (computed using the variance-covariance matrix of that group) is still valid provided the data are reasonably multivariate normally distributed. When this assumption is not valid, the technique may still be used provided that the errors of assignment are small (from test data sets or leave-one-out type procedures). Besides classical discriminant analysis, which is optimal when normality and variance assumptions hold, a number of procedures have been developed for discrimination which are free of distributional assumptions, or are defined for discrete or mixtures of discrete and continuous data. Some of the methods have been developed in relation to the problem of pattern recognition [Chen, 1973]. We are unaware of any of these having been used in systematics, though some applications have appeared in the geological literature. None of these methods are associated with low dimensional plots, i.e.

require the assumptions of multivariate normality and equal variance-covariance matrices within groups. When the variance-covariance matrices are not equal but the populations are still multivariate normal (graphically, the 95% (or other probability) concentration ellipses have different orientations or size), the optimal discriminant function is the so-called "quadratic discriminant function". It involves a larger set of discriminant coefficients which are multiplied by values for the variables as well as their squares and cross-products for an individual specimen. A quadratic discriminant function, however, is not associated with a canonical variate for displaying the relationships of the groups and specimens, because a single, transformed space is not produced to permit plotting the data. Mahalanobis D-like statistics as descriptive statistics comparing two groups are not direct results of this procedure. Actually, even when the variance-covariance matrices are different in various groups, the discriminant functions approach based on assignment to the group with the smallest $D^2$ (computed using the variance-covariance matrix of that group) is still valid provided the data are reasonably multivariate normally distributed. When this assumption is not valid, the technique may still be used provided that the errors of assignment are small (from test data sets or leave-one-out type procedures). Besides classical discriminant analysis, which is optimal when normality and variance assumptions hold, a number of procedures have been developed for discrimination which are free of distributional assumptions, or are defined for discrete or mixtures of discrete and continuous data. Some of the methods have been developed in relation to the problem of pattern recognition [Chen, 1973]. We are unaware of any of these having been used in systematics, though some applications have appeared in the geological literature. None of these methods are associated with low dimensional plots, i.e.

canonical-variate-like procedures, which summarize intergroup distances. However, it is possible to define some kind of distance statistic which could then be analyzed using principal coordinates analysis or nonmetrical multidimensional scaling. Since all of the methods determine errors of assignment between pairs of populations, then one minus the estimated error probability for each pair of groups can be used as a kind of "distance" measure. A non-parametric discriminant analysis procedure developed by Fix and Hodges [1959] called nearest neighbor discrimination is available in SAS. The Euclidean distance squared (or optionally Mahalanobis distance squared defined in terms of the total variance-covariance matrix instead of $S_W$) is used to relate each specimen to every other specimen in the analysis. For each specimen to be assigned the distances to its k nearest neighbors are found, where k is chosen by the user. The specimen is put in the group which has the largest proportion of neighbors among the k nearest ones. There may be a tie, as the proportion may be the same for two or more groups. If the proportion of near neighbors for all groups is less than a threshold the specimen may be put into group "other". One may allow for a priori probabilities as in classical discriminant analysis. A contingency table summarizes assignments in the form of the number and proportion of correct and incorrect assignments for each group of specimens as in classical analysis.

A method developed from pattern recognition applications has been used for discrimination in geology [Howarth, 1973]. This approach, called adaptive pattern recognition, depends on dividing the data set into training and testing subsets. The training set is used to empirically "estimate" the probability distribution of each group. Each specimen of the test set is assigned to the group for which its

"probability" is greatest. The procedure is evaluated by examining the number of correct and incorrect assignments. It would seem that large data sets would be required, but Howarth [1973] obtained excellent results with relatively small sample sizes. This method has also been discussed by Habbema and Hermans [1977].

When the observed variables are not continuous but are counts or discrete variables with a range of values, and the assumption of multivariate normality is clearly violated, classical discriminant analysis may still perform satisfactorily. However, classical discriminant analysis is in theory unsuitable for binary (0/1) data or discrete data. Special techniques are available for these circumstances and for mixtures of continuous and discrete variables in Lachenbruch and Goldstein [1979]. However, systematists seldom have large enough sample sizes to use these techniques. A series of procedures for this type of data are also summarized in the book "Discrete Discrimination" by Goldstein and Dillon [1978]. When the data is a mix of discrete, 0/1 and continuous variables hybrid procedures may be used or a procedure called logistic regression may be used. This method is related to the log-linear model for contingency tables (see discussion under "Multiway Contingency Tables" in METHODS).

Summary of terminology and recommendations:

The term discriminant analysis is used for those procedures where the matrix data has been divided into groups of individuals on the basis of an a priori classification. The purpose of the analysis may be to set up assignment or identification rules for additional specimens whose group membership is unknown, or it may be used to describe the differences among

groups in terms of the relative success of identifying the correct group assignment of the original specimens. This is directly related to their distinctness, i.e. distance apart, relative to their within-group scatter.

A display of the specimens in a space of reduced dimensions which summarizes the ability to discriminate the specimens and centroids is obtained through plotting them on canonical variate axes. The new variables are called canonical variates and this form of the analysis is sometimes called canonical variates analysis. The coefficients which are used to compute the canonical variates scores are called canonical variate coefficients and the correlations between the original variables and the canonical variates have been called structure coefficients. (Some workers restrict discriminant analysis to two group discrimination and call k-group discrimination "multiple discriminant analysis". We see no reason for this distinction. It is analagous to distinguishing two-sample tests on means from k-sample tests in the analysis of variance, or using "multiple factor analysis" to distinguish models with two or more factors from factor analysis extracting one factor.)

Mahalanobis $D^2$ is the Euclidean distance squared in the canonical space when the canonical variates have all been scaled to have a variance or standard deviation of one within groups. This distance squared is also sometimes called the generalized distance (e.g. in SAS79).

The greatest confusion arises over the terminology related to the term "discriminant function". For example SPSS and Jolicoeur [1969] call the canonical variate coefficients "discriminant function coefficients" and BMDP calls them "coefficients for canonical variates". Originally "discriminant

functions" referred to functions (of the original variables) used to partition the multivariate space into regions for assignment of specimens to groups [Morrison, 1976:239-245]. This is the usage we wish to continue. For k groups there are k regions (not including "other") and k(k-1)/2 "discriminant functions" in this sense. SAS uses the term "discriminant function" in a different but related sense, to refer to a function (of the original variables) used to determine how far each specimen is from each centroid. BMDP and SPSS on the other hand call the same function a "classification function", which is the term we recommend. A priori probabilities may also be accounted for in functions of either definition, but drop out when the a priori probabilities are equal--the usual case in systematics. There are k such discriminant functions when there are k groups. The values of these classification functions are the Mahalanobis distances of the specimens from each of the group centroids; it is these distances squared that are used in the classification rules. Finally in the two-group case, and only in the two-group case, the canonical variate coefficients and the discriminant coefficients may be the same. Some researchers have combined the terminology and called the analysis "canonical analysis of discriminance"; this term does not seem useful to us.

Statistical packages and computer programs:

| | BMDP77 | SAS79 | SPSS (Release 8) |
|---|---|---|---|
| 1. Stepwise | Many stepping options | Not available | Many stepping options |
| 2. Covariance used for discrimination | Pooled | Pooled or separate | Pooled or separate |
| 3. Priors | 1/k*, specified | 1/k*, specified, prop. to $1/n_i$ | 1/k*, specified, prop. to $1n_i$ |
| 4. "Classification functions" | Yes | Yes (form depends on pooling) | Yes |
| 5. Canonical variate coefficients | Yes | No | Yes+ |
| 6. Centroids for cva | Yes | No | Yes+ |
| 7. Can. var. scores | Yes | No | Yes |
| 8. Classification table | Yes, also jackknife | Yes | Yes |
| 9. Mahalanobis $D^2$ for centroids# | pairwise F | yes if pooled in 2) above | pairwise F |
| 10. Test data or cross-valid. | Yes | Yes | No |
| 11. General tests available | Wilks Lambda | Het. of cov.-var., contrasts etc. in MANOVA in PROC GLM | Het. of cov. var., Wilks Lambda |
| 12. Non-parametric discrimination | | PROC NEIGHBOR | |
| 13. Logistic discrimination | procedure P3F | PROC FUNCAT | |
| 14. Plots | Histogram if k=2, can. var. scores | none | Histogram if k=2, 2 can. var. scores, territoral map |
| 15. Classification results | $D_i^2$ to each group, a posteriori probabilities | a posteriori probabilities | $D_i^2$ to each group; prob. based on $D^2$; a posteriori probabilities |

| | | | |
|---|---|---|---|
| 16. Eigenvalues and eigenvectors of $S_W^{-1}S_A$ | Yes | in PROC GLM | values only |
| 17. Special features | | | Varimax rotation, Missing data options; structure coefficients. |

* defaults

\+ Prior to release 8, SPSS scaled the canonical variates so that their variance over all cases was 1; now the variance is one within groups. This affects the canonical variates coefficients (standardized or not), the plots--for example the distances between centroids are now D units apart.

\# Mahalanobis $D^2$ is only available in SAS if the variance-covariance matrices are pooled (it may also be calculated from what is given in the output simply). $F_2$ is given in the other two packages.
$D^2$ may be obtained from:

$$D^2 = \frac{(n-k)m(n_1+n_2)}{(n-m-k+1)n_1n_2} F$$

# P U R P O S E S

## DATA SCREENING

There is a place for multivariate analysis at early stages in a study in addition to those analyses published to support taxonomic conclusions. One such early application is data screening: looking for errors in the data which may arise from measurement error, recording error, mis-identified specimens, etc. Screening also includes a search for outliers--values that are extreme for any reason including error. Dempster [1971:341] has called this general topic "data cleaning", i.e. the "search for outlying values or values otherwise known to be impossible". Univariate statistical screening and visual examination of bivariate scatter plots are widely and successfully used for this purpose. Multivariate methods that depend on interpoint distances have also been found useful for error and outlier detection. One looks for observations that "stick out" (Rohlf, 1975)--that is, observations that are peripheral to the mass of data and have excessive interpoint distances relative to the average interpoint distances. The search procedures will be most powerful when applied to homogeneous data sets--a group of specimens from one locality or of the same sex, for example.

The probability that an observation was drawn from the population represented by the sample can be estimated (at least for multivariate normal data). The Mahalanobis distance squared, $D^2$, of each point from the population centroid has a chi-square distribution with degrees of freedom equal to the number of variables m, if the true variance-covariance matrix for the population as well as the true population mean is

known. $D^2$ computed from the sample centroid (now estimating the variance-covariance matrix) will have approximately a chi-square distribution for large sample sizes. Values of $D^2$ may be compared to tabled chi-square values and the probability of exceeding the corresponding chi-square value determined and used to flag suspect OTUs.

The values of Mahalanobis distance squared computed by BMDP (as part of the principal components routine) are divided by the number of variables; i.e. a value for $D^2/m$ is given for each OTU. These values may be more easily interpreted since $D^2/m$ will be near 1.0 when the point is part of the cluster, and values for outliers will be much larger. ($D^2/m$ is approximately distributed as chi-square/m, tabled in Dixon and Massey [1969:466-467]. Probabilities may also be looked up in any table of the F-distribution where the observed $D^2/m$ is compared to the tabled value of F for m (numerator) and infinite (denominator) degrees of freedom.) $D_r^2/r$ is also part of the BMDP output, where $D_r^2$ is the Mahalanobis distance squared from the centroid of the data cloud in the r-dimensional space defined by the r principal components (chosen by whatever stopping rule is used). Similarly $D_{m-r}^2/(m-r)$ is given for OTUs in the m-r residual components space. These latter values may be used in a relatively powerful test for detecting errors, or outliers, since OTUs with erroneous values are expected to have high $D_{m-r}^2/(m-r)$ values. Hawkins [1974] has given a chi-square test for $D_{m-r}^2/(m-r)$ values of OTUs on the residual components. However for this test to approach approximately correct significance levels (an alpha equal to .05 was used), more than 50 OTUs for 5 variables and many more than 300 OTUs for 20 variables were required. With smaller sample sizes, the probabilities are underestimated and more points will be assigned to a tail of a distribution containing

points "significantly" far from the centroid than should in fact be assigned there. An alpha of .05 means that, on the average if one's data is multivariately normal, 5% of the data will appear significantly far from the centroid by chance alone. This will be true even if the sample size is large enough so that the probabilities are correct. Therefore the $D_m^2/m$ values given in the BMDP output can only serve as a rough guide to errors or outliers for individual OTUs. Most sample sizes available to systematists (especially when the need for homogeneous subsets is considered) will not be large enough to correctly determine significance probabilities for this test.

Rohlf [1975] has proposed a "generalized gap test" for detection of multivariate outliers based on an assumption of multivariate normality for the data. The test involves finding a minimum spanning tree (MST) based on the interpoint distances among OTUs. A visual examination of the lengths of the edges of the MST is recommended through use of a quantile plot [Gnanadesiken, 1977:196-207]. A test based on the largest squared length associated with the most peripheral observation is given. Rohlf gives a table of approximate percentage points for this test for sample sizes of 20-200. (See Rohlf [1975] for additional details and Warde and Norton [1977] for comments, and Rohlf [1977] for a reply.)

Perhaps as one motivation for searching for outliers, it might be noted that the outliers may be sticking out in dimensions containing little of the variation within the rest of the sample. Consider a three-dimensional example where all but two individuals lie in a thin pancake-shaped cluster in 3-space, while the two outliers are a large distance above the cluster. An adequate summary of the variation in the

total data would require three components or axes, while a good summary of the variation in the main body of the sample excluding the two outliers could be accomplished in two dimensions. The more usual situation probably involves a few outliers not so far away, so that the variance they contribute is explained, for example, by some of the smaller components in a principal components analysis. The conclusion, however, is that a data array with one or just a few outliers might fit in a lower dimensional space if the outliers are removed from the analysis.

Gnanadesikan [1977:260-266] has suggested that plots of the scores on the principal components corresponding to the small eigenvalues be examined for outliers. The last few components define homogenous, in the sense of small variance, variables which will produce tight clusters if there are no outliers. This procedure provides visual support for the test suggested by Hawkins [1974], described above.

Less formal methods of checking one's data for errors are frequently also a good idea. Even before examining bivariate plots and checking the range and other univariate descriptive statistics, merely listing one's data through a print procedure (such as PROC PRINT in SAS) will produce a list of error messages flagging some "typos", such as non-numeric characters or inconsistencies in formatting. One of us has also found that log-ratio diagrams [Simpson, 1941] have been very useful for flagging erroneous values, especially key-punch errors. These diagrams summarize shape by plotting the ratio between an individual's measurement and the value for an individual or taxon mean chosen as a standard for each variable considered. Erroneous values frequently do not follow the general pattern for a sample and "stick out" on the diagram. Particular methods such as this are a matter of personal choice;

however, it is in general a good practice to decide on some screening procedures to use routinely prior to submitting a data set to a complicated analysis. The familiarity with the data gained in the process will also facilitate choice and interpretation of subsequent analyses.

## DATA REDUCTION

Data reduction is motivated by two rather different objectives, with only a partial overlap in methodology to achieve these goals.

1) One may wish to reduce the number of variables or OTUs in a study prior to a major part of the analysis. For example, one might want to discard redundant variables to reduce the number of measurements that must be taken on subsequent specimens measured. Another intent could be to reduce the number of variables for an analysis in which the sample size relative to the number of variables is important (see discussion of "Sample size" in the INTRODUCTION) but no more specimens are available for measurement. Reduction of the number of OTUs or observations may be required prior to phyletic analyses, for example, since an algorithm may be costly and time-consuming when the number of OTUs is high.

2) Alternately, one's goal may be to produce a summary of one's data in a reduced number of dimensions, as a result of the analysis. All of the original variables are generally used in the analysis, while the results are summarized by a few new variables which are combinations of the original variables. Reasons for such summaries include the production of two- or three-dimensional plots for visual examination,

or the production of summary statistics for interpretation or even subsequent numerical analysis.

1) Reducing the number of original variables or OTUs:

Relatively few variables may be needed to display patterns of data or clearly differentiate among taxa, although measurements for many variables are frequently recorded in the initial stages of a study. However determination of which subset of the original variables is indeed sufficient can be a very difficult problem. For example, depending on the nature of the correlations among variables and the partitioning of variance and covariance among and within taxa, it generally cannot be predicted, on the basis of mere examination of the data matrix or variance-covariance matrices, which few or more variables will best summarize the differences. However, an effective method is to employ stepwise procedures for discriminant analysis, analogous to those used in regression to segregate a smaller group of variables that are potent discriminators among the a priori groups. When using stepwise procedures, one must always take into consideration that they may not find the most discriminating subset of variables, since reaching a local optimum is a possibility.

The magnitude of the loadings of variables on the canonical variates produced in a classical discriminant analysis have been sometimes used as a criterion for defining subsets of r variables intended to maximally discriminate among the groups. This, however is not recommended, since the linear combination of the r variables that maximally discriminates among groups does not necessarily contain the same coefficients for the same r variables loading maximally on the canonical variate(s) constructed from all m variables (see Bargmann [1970]).

Several statistics and statistical procedures can serve as guides to the retention of a subset of variables to be measured in a larger study based on a smaller sample pilot study. The choice partially depends on the goals of the study and how much one is willing to extrapolate from data that might seem to be "typical" for the complete study. For example, if the pilot study is on just one species--will the subset of measurements behave in the same way in a group of closely or more distantly related species? In a homogeneous data set, redundancy of variables may be determined by multiple correlations of each variable with all others. If the values for one variable are predicted with small error from all of the rest, then it might be removed from the larger study.

Factor analysis is another useful way to find how much the association among variables can be explained by a relatively few of the original variables. Although rotation to a simple structure solution in factor analysis is more often used to produce a few hypothetical variables summarizing the original variables for plotting or presenting results, the results of such a study may also provide a basis for choosing a subset of the original variables for further manipulation. If single variables or a small number of variables, highly correlated with one or a few factors, can adequately represent the data then that smaller subset may be retained. Sokal and Rinkel [1963] discuss the choice of single characters to represent factors and the use of simple combinations of original variables for data presentation. In a study of geographic variation over the eastern United States using 17 variables, they retain 3 factors for interpretation. They use the simple average of three different standardized variables to represent factor scores for each of the three factors in the plots of variation over geography. They also compute locality

means of these factor scores and use the SNK multiple comparison test [Sokal and Rohlf, 1969:239] together with the plots of their factor scores to define intraspecific populations. This approach requires establishing a criterion for selection of the variables to be used in computing such ad hoc scores; such criteria should always be stated explicitly in publication of the results.

If regression procedures are going to be used in the larger study then some form of stepwise procedures or all possible regressions will find subsets that may do well (however see some of the dangers of this procedure in the METHODS section on "Multiple Regression"). In multigroup studies, one may be interested in testing hypotheses using MANOVA, employing discriminant analysis for identification, or producing through canonical variates anlysis a display of among-group differences and overlap. In all these cases, stepwise procedures can be used to reduce the number of variables that accomplish the purpose desired. In fact some of the stepping criteria are designed in terms of the purpose of the study, though these seem to be seldom used. For example in BMDP, if one were interested in discrimination of sexes but were studying several geographic populations or species, then the choice of variables in stepping can be made to optimize the discrimination between sexes even though several species or populations are included as discrete groups in the analysis.

The techniques for reducing large numbers of OTUs to a more manageable number are functions of the nature of the study or seem to be ad hoc. In biological studies, the mean value of each character for the populations or species is frequently used, but the criteria for delimiting each population or species are often left unstated, vague, or arbitrary. And in phylogenetic

studies, the steps involved in evaluating within taxon variability relative to among variability are rarely given, although there such an evaluation is usually the basis for the delimitation of species.

In geographic studies involving large numbers of individuals at lower taxonomic levels, locality or quadrat means are frequently used to represent the OTUs [Kennedy and Schnell, 1978; Best, 1978]. This practice has the additional advantage of essentially removing the problem of missing data. While specimens may be missing individual measurements, there is usually enough information on each character to compute a reliable mean (in the sense of small standard error) for the locality or area. The appropriateness of this kind of summary can be tested by nested analysis of variance techniques [Sokal and Rohlf, 1969] on individual variables and by extension to m variables by nested multivariate analysis of variance. If several nested geographic levels are used this would be an appropriate exploratory technique for choosing the appropriate size of region for agglomeration. Uneven sampling and missing data would complicate this approach.

2) Dimension reduction:

Bivariate scattergrams have been an excellent technique for summarizing relationships between variables and disclosing patterns using variables two at a time. However, the number of scattergrams required to display all bivariate patterns increases roughly with the square of the number of variables, m. (The exact formula for this number is $m(m-1)/2$.) While 10 scattergrams are needed to show the patterns for 5 variables, 190 scattergrams are necessary for 20 variables. Therefore, one of the most useful purposes of multivariate analytic methods is in dimension

reduction--producing a fewer number of new abstract variables or dimensions to table and plot for examining relationships of OTUs and presenting results. These new variables are usually linear combinations of the entire set of orignal variables. PCA, principal coordinates, factor analysis, multidimensional scaling and related methods may all be used to reduce the number of dimensions in the data to a comprehensible few. Discriminant analysis may also have a similar goal.

PCA will summarize the largest portion of overall inter-OTU variability in the least number of dimensions. In graphs of the PCA scores for these dimensions, clusters representing taxa may be evident, allometric trends discerned, or heterogeneity may be exposed. However, since the principal components are linear functions of the original variables, then possible interesting non-linear functions of the orginal variables may not be discovered. Some form of data transformation or non-linear technique such as non-metrical multidimensional scaling or nonlinear PCA may be more appropriate [Gnanadesikan, 1977:26-62]. Principal coordinates analysis is similarly useful for dimension reduction in those cases where PCA is appropriate as it is dual to that technique (see discussion under "Principal Coordinates" in METHODS). When there are more variables then OTUs it is computationally more efficient to do principal coordinates analysis.

When a priori subsets of the OTUs are recognized, the intergroup differences, corrected for the intragroup variation and covariation may be displayed using the canonical variate scores from multiple discriminant analysis. This technique will summarize the largest percentage of intergroup differences relative to intragroup variation in the smallest number

of dimensions. Two or three canonical variates may expose the similarities and differences among the a priori clusters in a few plottable dimensions. The directions of the canonical variates are influenced by the relative sample sizes among the groups (see "Discriminant Analysis" in METHODS). Also heterogeneity among the variance-covariance matrices will be averaged out. However, the method does provide a few dimensions summarizing the largest percentage of among-groups variation.

The effectiveness of dimension reduction through PCA or CVA is usually assessed by the percent of total variance explained by the first two or three axes. Another very useful way to determine whether the distances reproduced on the reduced two- or three-dimensional plot are badly distorting the relative original distances in the original m space is to superimpose a minimum spanning tree on the reduced dimension plot. OTUs that are actually closest together in m space may each appear closer to other OTUs in the reduced dimension plot. They will reveal their true overall similarity by being connected together as nearest neighbors in the superimposed minimum spanning tree. This procedure is a sort of graphical analogue to the various stress measures or statistics employed in nonmetric multidimensional scaling to measure the goodness of fit of the Cartesian representation of the original distance data. Oxnard [1973] has also developed some additional graphical tools for revealing such distortions.

## EXPLORATION: LOOKING FOR STRUCTURE

A great deal of the multivariate statistics used in systematic biology may be summarized under the heading

of exploration. There is a considerable overlap with dimension reduction since the summary plots or tables that are produced in dimension-reducing techniques are frequently used primarily for exploratory analysis. Some examples of the kinds of structures or patterns being watched for include the number of taxa, the presence of trends, or clines over geography, or associations of characters constituting functional complexes separable from other such complexes.

Exploratory data analysis is rather difficult to describe succinctly. "Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques." The extensive use of dimension reduction and plotting is a result of the "recognition that the picture-examining eye is the best finder we have of the wholly unanticipated" [Tukey, 1980:23-24]. One important group of methods, then, are the various dimension-reducing and graphical approaches; generaly all original variables are kept, but weighted in various fashions in the summary statistics or plots.

Principal components analysis (PCA) is a widely used exploratory data analysis technique to expose structure in data through plots in a few dimensions. Also, it sometimes seems to be used as a confirmatory technique in the guise of exploration. Taxa which have been earlier proposed are looked for as clusters of points separated from other clusters of points; however, separation between such 'clusters' are exaggerated by polygons defining the sample limits. Part of the problem is that there is not a clear definition of a taxon in terms of multivariate structure. Features such as gaps between clusters versus relative continuity seem to be the kinds of criteria used at the species level, for example.

Exploration tools such as PCA, principal coordinate analysis and Andrews plotting (see METHODS) can be used to display data in order to look for such gaps. However, if one subsequently tests for mean differences between groups formed by arbitrarily partitioning the multivariate space on the basis of such plots, the result will of course appear to support the existence of such groups. There are other, better paths to tread through the analysis than this circular one. New data are required to confirm the hypotheses generated from the exploratory analysis. It is worthwhile to test a priori hypotheses suggested from earlier analyses, but the eventual taxonomic decisions are not determined by rejection or acceptance of statistical hypotheses--they are instead judgemental decisions. The taxonomic decisions and discussions should take statistical tests into consideration however. Searching for new groups is a valid use of principal components and other dimension-reducing techniques. An exploration of more than the first few components may be desirable to thoroughly search the multidimensional scatter, and superimposition of minimum spanning trees may expose additional dimensions of interest not summarized in the first few.

A potentially powerful technique for searching multidimensional space is the use of Andrews plots (see section on "Andrews plotting" in METHODS). Gnanadesikan [1977:207-217] summarizes the use of this technique and gives some alternative formulations to that of Andrews [1972]. Oxnard [1975] has used Andrews' plots to examine diversity and explore relationships among primates. Rather than use static axes determined by PCA, principal coordinate analysis or factoring, Andrews plotting finds scores on an axis which is allowed to systematically scan the multidimensional space. Scores on these axes are weighted combinations of the original variables or more

often the principal components. The weights may be thought of and scaled as direction cosines for new axes in the multidimensional space. The scores are displayed as continuous lines: one line in the graph represents each OTU. The direction of the axes in the space is summarized along the abscissa and the scores along the ordinate. Tight clusters of scores that stay together indicate homogeneous groups. Consistent separation of clusters indicate widely and consistently different groups. Directions in which clusters differ, but are themselves little variable, are directions of interest for discriminating groups and may suggest a posteriori hypotheses to test. A priori tests are also available [Andrews, 1972]. Andrews also gives confidence intervals and variances for the plots. OTUs or groups of OTUs that oscillate between other groups, or break up and come back together as different directions are explored represent heterogeneous groups, or outliers relative to the groups of taxa under examination.

Many exploratory techniques are used as ordination procedures. Ordination consists of putting OTUs into continuous sequences, usually in two- or three-dimensional space. PCA, principal coordinate analysis, factor analysis, nonmetrical multidimensional scaling, and canonical variates analysis may all be used for ordination. The results of an ordination are usually presented in the form of bivariate scattergrams or perspective drawings of three dimensional scatter (stereograms are sometimes given for the three-dimensional diagrams). In contrast, cluster analysis has as its goal partitioning of the OTUs into homogeneous subgroups--frequently presented in the form of a dendrogram. Networks and trees are used to connect neigboring OTUs. All three techniques may be used to explore the data matrix for structure. They may be used in concert--for example a dendrogram or

minimum spanning tree may be sumperimposed on an ordination in two or three dimensions; or clusters found by a clustering technique may be circled or distinguished by special plotting symbols. An extensive discussion comparing these three different approaches to data arrays may be found in Sneath and Sokal [1973:Chapter 5].

Finding interpretable associations among variables is a further exploratory goal. The metric multivariate techniques based on variance-covariance or correlation matrices among variables (e.g. PCA, factor analysis, canonical correlation and canonical variates analysis) all produce vectors of weighting coefficients which may be used to estimate scores for new variables for plotting, discrimination etc. Of these methods, only factor analysis has as one of its goals the discovery of factors (=hypothetical variables) chosen by criteria related to their potential interpretability. It seems tempting (and is advocated by some texts and manuals) to interpret and name principal components or canonical variates, to take the most widely used multivariate methods, in terms of the variables with relatively large (in absolute value) coefficients or loadings. It must be remembered that these coefficients have been found to maximize certain criteria (in the case of PCA--variance of components which are uncorrelated; in the case of CVA--variance among centroids relative to pooled with variance-covariance and uncorrelated variates). There is no a priori reason in a PCA to expect that directions which are orthogonal and are related to ordered variance explanation should find interpretable complexes or groups of functionally related variables.

If we wish to look for interpretable variable complexes, then a methodology designed to discover them should be used. Olson and Miller (1958) developed a

clustering approach for correlation matrices that was designed to find groups of associaged variables that can be identified with known functions. The analysis was done without computers and a series of extensive data examples were used to test the methodology. Their goal was to develop a method applicable to fossils where the associations between characters and functions were less well known or not known. (All of the original data are provided in their book.)

Exploratory factor analysis is a technique for examining associated sets of variables and finding clusters of variables or hypothetical new variables which may suggest hypotheses about relationships with function. These hypotheses may be subsequently tested with new data using confirmatory factor analysis. The partition of each variable into parts related to common factors and a unique portion, and the rotations available to search for "simple structure" come closer to the goals of Olson and Miller than other multivariate procedures that we know.

Maximum likelihood factor analysis has an added advantage over most other forms of factor analysis, in that either the variance-covariance matrix S or the correlation matrix R may be factored and simply transformed from one to the other (see section on "Factor Analysis" in METHODS) and goodness of fit tests are possible for the number of factors if the data have a multivariate normal distribution.

Interpretation of canonical variate coefficients is fraught with more difficulty then most (see "Discriminant Analysis" in METHODS). Bargmann [1970] discusses how the structure coefficients in canonical variates analysis may be used to select variables for further use. He points out, however, that the linear combination that produces maximum differences between

groups "is and remains an artificial variable, a mathematical construct". The variables loading on the second or other canonical variates should have their contributions to the first variate removed if interpretation is a goal. SPSS does provide for rotation of the "discriminant functions" using VARIMAX rotation. Multigroup factor analysis is an attractive area to explore for finding complexes of variables both within and among groups [Sörbom and Jöreskog, 1976].

## CLUSTER ANALYSIS, NUMERICAL CLADISTICS AND TREE ANALYSIS

The set of techniques mentioned briefly in this section are those developed for finding or constructing groupings, usually of individuals or OTUs. These techniques have been extensively used, discussed, and compared in publications on numerical taxonomy (NT) and outgrowths from that school. Although numerical taxonomy is sometimes identified with phenetics in some sort of limited sense, numerical cladistics is part of NT, and discussed in this section, because it is another set of clustering models and methods, distinguished only by the criteria for linkage of OTUs. The diversity of topics and methods that are included under the heading of this section is enormous, and very well covered in a number of books and review articles. Therefore, we will only note some philosophical and general methodological points, and some references to reviews that will provide an entry into the literature.

### Cluster analysis:

Cluster analyses are methods for analyzing a single

data array of n OTUs and m variables to find groups or clusters of OTUs or variables that are more similar within groups than among groups. Jardine and Sibson [1971:39-44] make a useful distinction between the cluster method or model, and the clustering algorithm(s) for achieving that model. The model is primarily a precise statement of the kind of representation desired, the basis for grouping, dividing, or structuring. On the other hand, there are the various algorithms that implement the method. For a particular method, one hopes to find an efficient algorithm.

Historically, the relationship between model or method and algorithm has not always been recognized. It has only been recently that it was realized that some different algorithms are actually performing the same methods. A distinction is usually made between algorithms (properly not methods) that are divisive and partition multivariate space into regions, and those that are agglomerative and form larger and larger groups in some form of hierarchy. Some methods, such as single linkage cluster analysis, can be implemented by both divisive and agglomerative algorithms, although the identity of the method was frequently not recognized at the time each algorithm was presented.

The distinction between model or method and algorithm should make obvious the need for a clear understanding of the model when a clustering procedure is chosen; the algorithm is less important except for practical reasons of time or cost. The problem with urging an awareness of this distinction is that the literature is largely unclear on this point; useful statements and explanations will usually be hard to find.

Another distinction to bear in mind is the

difference between clustering methods and exploratory procedures which facilitate the definition or observation of clusters. Thus, various dimension-reducing techniques, such as PCA, discriminant analysis or multidimensional scaling techniques, are not clustering techniques, because they do not find groups in initially unclustered data. Clusters may be discerned, for example, in a plot of PCA scores, but circling or otherwise defining clusters or groups on the basis of apparent clusters would be an ad hoc clustering procedure superimposed on the PCA results; PCA itself is an ordination technique, not a clustering technique. The component scores may, of course, be used as the data matrix for clustering, but the clustering itself is not part of PCA. Factor analysis is less clear, and in some ways intermediate: in an R-mode analysis, the positions of the OTUs in the reduced space defined by the factors may be used as a basis for ad hoc clustering, but clusters of OTUs are not produced by the factor analysis. However, in such an analysis, various criteria for rotation of the axes, for example, rotation to simple structure of Thurstone and others, produces a sort of clustering of variables, in which each factor would define a cluster of those variables with nonzero loadings on that factor. Similarly, a Q-mode factor analysis with some kinds of oblique rotation would result in a clustering of OTUs, in that each factor now represents a group of highly correlated OTUs.

A great deal of clustering, especially hierarchical clustering, has been done in taxonomy; the book by Sneath and Sokal [1973] is a comprehensive review of numerical taxonomy. The earlier edition [Sokal and Sneath, 1963] has a useful appendix which goes through an algorithm for the weighted pair group method (WPGM) step by step, starting from an n x m data matrix. The book by Hartigan [1975] discusses a number of

difference between clustering methods and exploratory procedures which facilitate the definition or observation of clusters. Thus, various dimension-reducing techniques, such as PCA, discriminant analysis or multidimensional scaling techniques, are not clustering techniques, because they do not find groups in initially unclustered data. Clusters may be discerned, for example, in a plot of PCA scores, but circling or otherwise defining clusters or groups on the basis of apparent clusters would be an ad hoc clustering procedure superimposed on the PCA results; PCA itself is an ordination technique, not a clustering technique. The component scores may, of course, be used as the data matrix for clustering, but the clustering itself is not part of PCA. Factor analysis is less clear, and in some ways intermediate: in an R-mode analysis, the positions of the OTUs in the reduced space defined by the factors may be used as a basis for ad hoc clustering, but clusters of OTUs are not produced by the factor analysis. However, in such an analysis, various criteria for rotation of the axes, for example, rotation to simple structure of Thurstone and others, produces a sort of clustering of variables, in which each factor would define a cluster of those variables with nonzero loadings on that factor. Similarly, a Q-mode factor analysis with some kinds of oblique rotation would result in a clustering of OTUs, in that each factor now represents a group of highly correlated OTUs.

A great deal of clustering, especially hierarchical clustering, has been done in taxonomy; the book by Sneath and Sokal [1973] is a comprehensive review of numerical taxonomy. The earlier edition [Sokal and Sneath, 1963] has a useful appendix which goes through an algorithm for the weighted pair group method (WPGM) step by step, starting from an n x m data matrix. The book by Hartigan [1975] discusses a number of

clustering techniques and gives FORTRAN programs for their implementation. Other books are Van Ryzin [1977] and Anderberg [1973]; Williams [1971] appears to be a useful review article.

NTSYS (Numerical Taxonomic System of Multivariate Statistical Programs) has a number of routines for clustering and provides a large number of association measures in addition to a number of multivariate statistical routines. It has been widely used for clustering and widely cited in the systematic literature.

SAS has a procedure called CLUSTER based on an algorithm suggested by Johnson [1967] which appears to be an average linkage cluster method.

BMDP(77) has a cluster procedure, P1M, for clustering variables with a choice of average linkage, single linkage, or complete linkage. P2M provides cluster analysis of cases which may be based on any of four distance measures and uses average linkage. A simultaneous clustering by cases and variables is given in P3M. Dendrograms are given for both P1M and P2M.

The book by Anderberg [1973] gives a set of FORTRAN IV programs for a wide variety of association measures and cluster analyses.

CLUSTAN is a clustering package with larger number of association measures and a number of both agglomerative and divisive clustering programs. Instructions are given in Wishart [1978]. The results of a cluster analysis are frequently presented in the form of a dendrogram, which is output by most of the clustering programs.

Numerical cladistics:

Numerical cladistics has as its goal finding a phylogenetic hypothesis or rooted tree from numerically coded characters or a distance matrix. Several approaches are summarized in Sneath and Sokal [1973:323-356]. The terminology used in the literature to describe the phylogenetic patterns is a little confused, but well summarized in Sneath and Sokal [1973:253-256, 325-327,332-333]. We will only point out here that a Wagner network is a form of tree in graph theory. A network is a directed tree or rooted tree. Cladograms are also discussed in Sneath and Sokal [1973:332-346].

Phylogenetic reconstruction is of major interest in systematic studies at all levels. The result of any clustering method depends on the similarity measure or coefficient employed and ultimately the objective criterion used to form clusters. Phylogenetic analysis is specifically directed at deducing genealogical relationships; therefore characters are evaluated in terms of hypotheses of primitive vs. derived states rather than overall similarity. The most popular criterion used for evaluating results is parsimony. A most parsimonious phylogenetic tree is one in which the number of evolutionary steps is minimal compared to other possible trees, and therefore it has the least number of reversals or homoplasies. Computer algorithms for finding Wagner networks and rooted Wagner networks are described in papers by Farris [1970 and 1972]. A program WAGNER by J. S. Farris finds Wagner networks from distance matrices. A direct solution to finding the shortest networks is not known; as in many iterative solutions, there is no guarantee that an optimal solution has been found. The total number of networks is too large to find all of them, and stepwise procedures are fraught with the usual

difficulties of finding local optima. One way of guarding against non-optimal solutions is to try several starting points, i.e. reorder the data and submit it again to the analysis.

Trees:

A useful multivariate technique is finding the shortest connected tree of n OTUs using n-1 links, i.e. linking all of the OTUs together with the shortest length tree. This has been called a shortest spanning tree [Sneath and Sokal, 1973:255-256], or a minimum spanning tree (MST), among other names. We prefer the latter because this seems to be the most common usage, and is the most euphonious. A half-dozen MST algorithms are referred to in Sneath and Sokal [1973:255]. The MST procedure is related to single linkage clustering [Hartigan, 1977:60]. A Wagner network is an MST with nodes or additional hypothetical taxonomic units (HTUs) placed between the OTUs.

An MST may be usefully superimposed on almost any kind of two- or three-dimensional projection of higher-dimensional multivariate data. Computed from the distance matrix among OTUs, it can be used to indicate the closest OTUs in the full dimensional space. Distortions in the reduced dimensional plot will be evident by OTUs that appear close together on the plot not being each other's nearest neighbors according to the MST. An MST may be superimposed on plots of principal components, principal coordinates, or canonical variates. Schnell [1970] plotted an MST on a three-dimensional perspective of OTUs using the first three principal components. Oxnard [1973] and Baker, Atchley and Mc Daniel [1972] superimpose MSTs on plots of canonical variate scores; the MST is computed from the matrix of among-group Mahalanobis $D^2$s.

NTSYS has an MST routine. Also, it is not difficult to program MST in PROC MATRIX in SAS.

## SIZE AND SHAPE

The analysis of size and shape has generated considerable discussion in the systematic literature [e.g., Mosimann and James, 1979; Gould, 1966; Corrucini, 1972, 1973, 1975b; Bookstein, 1977b, 1978; McMahon, 1973]. While size and shape may be studied in relation to functional hypotheses, or ecological roles of taxa, we will mainly discuss the problem of size correction or adjustment for comparison of OTUs and make some comments on the interpretation of "size" and "shape" components derived from output of multivariate analyses. A brief review of two special "shape" analysis techniques is also included. Part of the difficulty is: What is meant by size, and shape? Everyone knows what they are, but few have attempted to define them. Mosimann [1970] gives an axiomatic approach to the problem and Sprent [1972] reviews "The mathematics of size and shape". Gould [1966] discusses the biological problem, but mainly in a bivariate context.

In an analysis of only one variable, OTUs may be uniquely ordered. We can speak of size from smallest to largest along a a single axis. However, for multivariate data there is no unique ordering of points in the multidimensional space. A group of OTUs will only have the same shape, i.e. proportions, if they lie on a single ray or vector through the origin in the character space. In this case, these OTUs will be isometric, and the space will have only one dimension. If the OTUs fall on several vectors or rays through the origin then the definition of size is arbitrary. Any

single character (weight, cube root of weight, body length or skull length to name a few possibilities) may be considered as "size" characters, or "size" may be defined as a function of several characters. For example the first principal component score or the geometric mean of the linear dimensions for an OTU are possible compound size measures. If there is little scatter of the data about the size measure chosen, then the data can be "corrected" to a common size.

The first principal component of the variance-covariance matrix, the correlation matrix, or either of these after the data has been transformed (for example by a log transformation) is a popular choice for a "size" variable. A log transformation of linear, areal and volumetric variables puts these dimensions in the context of the allometric equation. If all of the coefficients (for linear dimensions measured in the same units) for the first principal component are of the same sign and have approximately the same value then the direction of greatest variation along the first principal component axis is correlated with the size of the individuals for each variable taken separately. The first principal component then is a kind of size component. The elements of the corresponding first eigenvector of a variance-covariance matrix should be divided by their corresponding standard deviations if the effects are to be evaluated in terms of standardized variables. For log-transformed data for linear dimensions the hypothesis of isometry may be tested [Morrison, 1976:295--the first eigenvector is compared to a vector with elements equal to $m^{-.5}$], or such a test can be done using confirmatory factor analysis. The ratio of the elements of the first eigenvector can be used as generalized allometric coefficients [Jolicoeur, 1963a; Corrucini, 1975b; Dodson, 1975].

If the first eigenvector is interpreted as a kind of "size" vector then all others must be "shape" components. The orthogonality imposed by PCA requires each subsequent vector to have both positive and negative values so that the inner product of each vector with the first "size" vector and every other vector will be zero. There is a temptation to interpret these as "shape vectors independent of size", but one must take into consideration the artificial constraint of orthogonality (in order to make their corresponding scores uncorrelated). "Shape" is rendered "independent", actually uncorrelated with "size" by the method, but this is an artefact of the constraints built into the method. Size-correlated shape changes are more to be expected, and even if the first component can be interpreted as a size component, the other components are more realistically mixtures of size and shape [Sprent, 1972]. (However, see Blackith and Reyment [1971] for stronger support for the use of PCA in analysis of size and shape than is presented here.) One of the most important points that Blackith and Reyment bring out in their discussions [op. cit.:29] is that one should not discard size in an interpretation of one's data.

Factor analysis offers a way of finding correlated or oblique factors which may be biologically interpretable and do not have the artificial constraints of PCA. However, as Sprent [1972] points, out the indeterminancy of factor analysis solutions (see section on "Factor analysis" in METHODS) may be one reason why they have not been used more for this purpose. Confirmatory factor analysis [Kim and Mueller, 1978b; Mulaik, 1972; and Jo"reskog and So"rbom, 1976] offers an approach in which a "size" factor may be chosen (or hypothesized) by the researcher and, additional correlated or uncorrelated "shape" factors fit to the data with a built-in

goodness of fit test. The test does depend on the assumption of multivariate normality, but may serve as an indication of fit in any case. Hopkins [1966] used a centroid factor analysis to study allometry of organ weights in the rat. Blackith and Reyment [1971:29] claim that the difference in the factor analysis solution from the PCA analysis in Hopkins' study was an artefact of the centroid method used. However, one of us has confirmed Hopkins' solution using maximum likelihood factor analysis.

The allometric equation has been found useful for describing the change in shape with size in many studies [Gould, 1966; Corrucini, 1972] and may be used as a basis for the "removal" of size effects in data Corrucini [1972]. Jolicoeur [1963a] has generalized the allometric equation for multivariate relations (discussed critically by Sprent [1972]). A logarithmic transformation (any base will do) of the original data changes the allometric relation (a power equation) to a linear equation in the logs of the variables.

The main purpose of the methods discussed above is the description of size in relation to shape. If on the other hand, one wishes to remove "size" effects, there are several techniques available. One is through the use of multivariate analysis of covariance, in which the size measure is the covariate. The degree of variation about the best fitting line for each group in a multigroup study as well as a test of hypothesis that the trend lines within groups have the same slope are available (see Snedecor and Cochran [1967: Chapter 14] for univariate analysis of covariance, and Morrison [1976:193-197] for multivariate analysis of covariance).

A warning is appropriate at this point concerning the use of ratios. Ratios as corrections for size are

most appropriate, as pointed out above, when there is an isometric relation among the variables. In this case, the line along which the data are scattered passes through the origin, and there is not much spread about the line (see Snedecor and Cochran [1967] for appropriate bivariate tests). When the relation between variables is linear, and the summary line does not pass through the origin then the analysis of covariance corrections are appropriate if the lines are parallel among groups. For example, in a discriminant analysis that compares several groups or taxa, one might wish to analyze the residuals after the first principal component is removed from the pooled within-groups covariance matrix, if the first PC seems to be a size-related variable. Other principal components, or other linear combinations may similarly be removed. These approaches will only be reasonable if the direction of the principal components or other chosen directions are similar for each group. A hypothesis for similar directions can be tested, but if the sample sizes are small for the groups, then the effect of the correction on the results would be difficult to assess. Reyment and Banfield [1976] and Gower [1976] have presented a technique along with examples for adjusting a discriminant analysis for trend(s), following an earlier suggestion of Burnaby [1966]. The method has been called "growth invariant discriminant functions" and is also discussed by Dunn [in press] as a form of data transformation. Using this method, any trend(s) or direction(s) of interest may be removed from the data before a discriminant analysis is performed. The use of the first principal component within groups has been suggested as a direction to remove from the pooled within-groups variance-covariance matrix if this is a strong size measure. However, if the among-group differences are in the same "size" directions as the principal components then this may not be a useful thing to do.

In any case, if there is a strong within-groups size trend it may have a relatively small effect in the discriminant space as the among-groups covariance matrix is multiplied by the inverse of the pooled within-groups covariance matrix in discriminant analysis; the eigenvalues of $S_W$ affect formulation {1} in "Discriminant Analysis" (in METHODS) as a function of their reciprocals. A simpler partition of distance into "size" and "shape" components had been suggested by Penrose [1954]. Bräuer [1979] is a recent application of Penrose's ideas.

When the relationships between variables is not linear then analysis of covariance using quadratic (or higher order) functions of the "size" variable(s) as covariates might be appropriate, provided that the non-linear trends are parallel in the multivariate space.

Mosimann's [1970] most important axiomatic approach to size and shape has been applied to multivariate analysis of bird variation along geographic gradients [Mosimann and James, 1979]. Organisms have the same shape in Mosimann's model if all of one individual's measurements are multiplied by one constant to obtain another's measurements (i.e. they are isometric). A convenient size measure is chosen (a single or compound measure); then a shape vector is a unit-free vector of measurements in which each of the original measurements is divided by the chosen arbitrary size measure (all in the same units). It is impossible to define a single shape vector that is independent of more than one size measure. Mosimann and James therefore define several size measures and study the geographic variation of shape relative to the different size measures. Mosimann also provides a test for the independence of size and shape based on the multiple correlation coefficient squared. Corrucini [1972, 1973, 1975]

removes the effect of size in his studies of primates by dividing each measurement by a size measure for the OTU, along the lines of Mosimann's suggestion. However he does not test for independence of size and shape, and does not attempt to look at shape in terms of different size measures as do Mosimann and James. Sprent [1972] offers some additional suggestions for studying size and shape outside of the form of the relationship between the two.

Many people have recognized basic shortcomings in using linear dimensions as descriptors of shape and shape change. Alternate approaches, however, have been difficult to formulate. A step in the right direction is the recording and analysis of Cartesian coordinates of points directly from a specimen rather than recording only the linear measurements among the points. (Recent examples include Benfer, 1975; Brower and Veinus, 1978; Hills and Brothwell, 1974.) Methods of analysis for such data are not widely developed, however. One early attempt which has fascinated nearly all natural historians at sometime in their development has been the transformation grids described by D'Arcy Thompson [1961]. These have also been frustrating because of their arbitrariness and lack of precision.

Two recent methods are available for more precise description of shape description of shape difference or change, and shape itself. These are Biorthogonal transformation grids and Fourier analysis respectively.

Biorthogonal transformation grids:

Bookstein [1977a;1977b;1978] has developed a special transformation grid for comparing two shapes. The transformation proposed by D'Arcy Thompson imposes a Cartesian coordinate system on one member of the pair

and then transforms the diagram keeping homologous points connected. This leads to an ambiguous asymmetrical transformation, depending on which shape is transformed. It also had never been successfully formulated mathematically, so the details of a transformation were frequently inaccurate. Bookstein's transformation is symmetrical and formulated precisely.

Consider a rectangle and a parallelogram. For these two figures, there is a unique pair of orthogonal axes on the rectangle, which remain orthogonal under transformation of the rectangle into the parallelogram, but change their length. The amounts of change in length are referred to as the dilations along the two axes. If the rectangle were now transformed into an irregular four-sided figure, there would no longer be a single pair of biorthogonal axes, but instead a pair with slightly different orientation for almost every infinitely small rectangle, i.e. almost every point, so one would have a continuous surface with infinite pairs of dilation values and an orientation for each biorthogonal pair. ("Almost every point" because singularities or undefined points in the surface can arise.) These surfaces are depicted by sampling the axes in a grid-like fashion, in which each axis drawn curves to show the orientation at each point it passes through, and intersects other axes always at right angles. The amount of dilation at selected points along these axes can be indicated next to the axes.

Warping, curvature, and rotation of parts of a figure are all effected on such a surface by differential dilation. Greater dilation along one axis will cause it to curve around parts that are undergoing less dilation. On biological figures, homologous points on the two figures are designated, and the change is assumed to be uniformly changing between these points.

then fitted to the data. If $y_i$ is the radius for the ith point, then the model is:

$$y_i = \beta_0 + \beta_{11}\cos\theta + \beta_{12}\sin\theta + \ldots\ldots + \beta_{m1}\cos m\theta + \beta_{m2}\sin m\theta + v_i \quad \{1\}$$

This is just a special case of linear regression with 2m regression coefficients in the model, for the m harmonics. Each pair of $\beta_j$'s, $\beta_{j1}$ and $\beta_{j2}$, gives the contribution of the jth harmonic to the overall $R^2$. The contribution for the jth harmonic is

$$R_j^2 = (b_{j1}^2 + b_{j2}^2) \quad \{2\}$$

Using regression theory (see "Multiple Regression" in METHODS) and tests, one can find the significant harmonics, or the stopping rule for the number of harmonics m to adequately represent the shape [Anderson, T. W., 1971:Chapter 4; Davis, 1973 for an introduction to the subject].

Two ways of presenting the results of a harmonic analysis are commonly used. The periodogram is a plot of $R_j^2$ (which is the amplitude of the jth harmonic) plotted against n/j, the period length of the jth harmonic; the other is a plot of $R_j^2$ against the frequency, i.e. the number of cycles in the n observations. The maximum number of harmonics that can be estimated is n/2. The goal of fitting the model is to sufficiently describe the shape with as few $b_j$s as possible. Residuals should be examined as an additional way of evaluating goodness of fit. If graphical equipment is available, outline data can be entered using a digitizing board. A computer graphics screen or hard copy plotter can then be used to compare the original shape to the fitted shape for each specimen.

Up to now the discussion has been in terms of describing a single specimen whose outline has been summarized in terms of a set of 2m regression coefficients. The 2m coefficients usually represent a

considerable data reduction of the original n points measured. The coefficients for each specimen can then be used as a vector of values for further multivariate analysis, for example PCA or canonical variates analysis.

Lestrel et al. [1977] has used the Fourier technique to compare the shape of the distal end of the femur among hominoids, using the coefficients for each specimen as input data to a canonical variates analysis. "Size" was removed from the analysis by adjusting each specimen to have the same area. They retained 14 coefficients. Lu [1965] applied harmonic analysis to the shape of the human face. Younker and Ehrlich [1977] have used "harmonic amplitudes" in a study of ostracod shape. The first six harmonics were used in a discriminant analysis of taxa at the species and genus level in two separate analyses. See also Kaesler and Waters [1972] and Waters [1977] for further examples in paleontology.

STATISTICAL INFERENCE:

We may consider two classes of hypotheses:

1) A posteriori hypotheses: those that are suggested by the results of the application of a method, usually an exploratory data analysis, or those suggested from the results of tests not considered initially in a priori analyses. An example of the former would be a test of the hypothesis of "isometry" for the first eigenvector from a principal components analysis in which the loadings or coefficients appear similar. Examples of the latter sort of a posteriori tests would be multiple comparison tests, after an analysis of variance or multivariate analysis of variance, to

determine homogeneous subsets of groups.

2) A priori hypotheses are stated before the application of the particular method. These hypotheses arise from earlier analyses, biological or physiological constraints, or models developed for a study. Examples are tests for specific allometric scaling [McMahon, 1975] used in a comparison of limb bone lengths and widths which had been done using PCA [Jolicoeur, 1963]. Another example would be the use of multivariate analysis to see if groups proposed by earlier workers were significantly different. Tests for sexual dimorphism would also generally be considered a priori analyses. The most extensive development of a priori testing procedures to present has been in MANOVA (see page 152 in METHODS). Also, confirmatory factor analysis offers opportunity for the development of extensive a priori tests.

However, most systematic studies are presently dominated by exploratory analyses and therefore many of the tests are a posteriori. Even when overall a priori tests for equality of centroids are used in canonical variates analysis or multiple discriminant analysis (actually in this guise a form of multivariate analysis of variance), deciding which groups may be agglomerated and which are separate is an a posteriori testing situation.

One consideration if very many tests are made (more often a problem in a posteriori testing), is that the actual overall significance level increases. For example, if one sets the probability of rejecting the null hypothesis (when it is in fact true) at 0.05 for each test and then performs a series of tests, the actual probability of incorrectly rejecting the null hypothesis sometime during the whole series of tests increases. For k independent tests, the overall

significance level becomes $1-(1-\alpha)^k$. Harris [1975:98-101] suggests a simple though somewhat conservative technique for setting a smaller overall significance level, based on the Bonferroni method [also see Morrison, 1976:33-34]. If one is studying, for example, 10 groups and reject the null hypothesis of group centroid equality, then there are 45 possible two-centroid comparisons to test for inequality. The general formula for k groups is k(k-1)/2 two-group comparisons. If one wanted an overall significance level of 0.05, one would compare all pairs using a (0.05/45) significance level for each comparison. This is a conservative formulation, since the overall significance level will be less than 0.05. In general for a significance level $\alpha$ and c comparisons, each comparison is done at a significance level ($\alpha$/c) to guarantee an overall significance level less than $\alpha$.

Statistical testing procedures have been developed within the context of most of the methods (see section on "Statistical assumptions" under each method). For example, within procedures such as principal components analysis or canonical variates analysis, one may test hypotheses about the number of significant axes, based on the relative size of the associated eigenvalue. However, the vast majority of tests developed to date require the assumption of multivariate normality, and if testing group differences, the assumption of equal variance-covariance matrices among populations from which the samples were drawn. In some circumstances, test procedures have been developed for only certain variants of a method--for example, most tests in PCA are for analyses based on the variance-covariance matrix rather than the correlation matrix.

As we noted in the Introduction, the the problem of checking these assumptions by tests of multivariate normality and tests of equality of covariance matrices

is a difficult one. Because of the problems in meeting these assumptions with biological samples, multivariate procedures have been more often used descriptively, in an exploratory rather than hypothesis-testing framework.

# B I B L I O G R A P H Y

The notation in the right margin is to be interpreted as follows:

* = good readable discussion; if in combination with T, refers to portions.
T = Technical, frequently cited for documentation.
A = additional applications not cited in the text.
C = cited for completeness; not examined by us.
P = manual for a computer package, or a book which contains computer programs.

Several journals regularly publish articles on multivariate statistical methodology that are likely to be of interest or use to systematists. Some of these are: American Statistician; Applied Statstics; Biometrics; International Association of Mathematical Geology, Journal; Journal of the American Statistical Association; Journal of Multivariate Statistics; Multivariate Behavioral Research; Psychometrika; Technometrics.


Afifi, A. A. and S. P. Azen. 1972. Statistical Analysis: A Computer Oriented Approach. New York: Academic Press.

Afifi, A. A. and R. M. Elashoff. 1969. Missing observations in multivariate statistics. Journal of the American Statistical Association 64:337-358,359-365. T

Amick, D. J. and H. J. Walberg, eds. 1975. Introductory Multivariate Analysis. Berkeley, California: McCutchan Publishing Company.

Anderberg, M. R. 1973. Cluster Analysis for Applications. New York: Academic Press. C

Anderson, A. J. B. 1971. Numerical examination of multivariate soil samples. International Association of Mathematical Geology, Journal 3:1-14.

Anderson, T. W. 1958. An Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons. [Classical theoretical foundation for many methods] T

Anderson, T. W. 1971. The Statistical Analysis of Time Series. New York: John Wiley & Sons. T

Andrews, D. F. 1972. Plots of high-dimensional data. Biometrics 28:125-136. *,T

Andrews, P. and D. B. Williams. 1973. The use of principal components analysis in physical anthropology. American Journal of Physical Anthropology 39(2):291-303. *

Ashton, E. H., M. J. R. Healy and S. Lipton. 1957. The descriptive use of discriminant functions in physical anthropology. *

Proceedings of the Royal Society of London, B.145:552-72.

Atchley, W. R. 1971. Components of sexual dimorphism in Chironomus larvae (Diptera: Chironomidae). American Naturalist 105:455-466. [oblique factor analysis in systematics reprinted in Bryant and Atchley, 1975]

Atchley, W. R. and E. H. Bryant., eds. 1975. Multivariate Statistical Methods: Among-groups Covariation. Stroudsberg, Pennsylvania: Dowden, Hutchinson & Ross (Distributed by Halsted Press, a Division of John Wiley & Sons). [collected papers, reprinted; includes Gabriel, 1969; Porebski, 1966; Jolicoeur, 1959; Rohlf, 1970, 1971 and 1972; Kruskal, 1964a; Burnaby, 1966] *,T

Baker, R. J., W. R. Atchley and V. R. McDaniel. 1972. Karyology and morphometrics of Peter's tent-making bat, Uroderma bilobatum Peters (Chiroptera, Phyllostomatidae). Systematic Zoology 21(4):414-429. [reproduced in Atchley and Bryant, 1975]

Bargmann, R. E. 1970. Interpretation and use of a generalized discriminant function. pp. 35-60 in Bose et al., eds. 1970. *,T

Barr, A. J., J. H. Goodnight and J. P. Sall. 1979. SAS User's Guide. 1979 ed. Raleigh, North Carolina: SAS Institute Inc. [new procedures are constantly appearing; there is a 1980 supplement which includes nonmetric multidimensional scaling] P

Beale, E. M. L. and R. J. A. Little. 1975. Missing values in multivariate analysis. Journal of the Royal Statistical Society B,37(1):129-145. C

Benfer, R. A. 1975. Morphometric analysis of Cartesian coordinates of the human skull. American Journal of Physical Anthropology 42:371-382. [uses X,Y,Z coordinates instead of linear measurements; PCA analysis and then rotates to find clusters of variables]

Berthou, P. Y., J. C. Brower and R. A. Reyment. 1975. Morphometrical study of Choffat's vascoceratids from Portugal. Bulletin of the Geological Institutions of the University of Uppsala, New Series 6:73-83. [use of PCORD and canonical variates, actual pictures of specimens on plots]

Best, T. L. 1978. Variation in kangaroo rats (genus Dipodomys) of the Heermanni group in Baja California, Mexico. Journal of Mammalogy 59(1):160-175. [PCA and cluster analysis on centroids of samples; minimum spanning tree; use of NTSYS]

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Massachusetts: The MIT Press. [Log-linear model with examples in biology] *,T

Bissell, A. F. and R. A. Ferguson. 1975. The jackknife--toy, tool or two-edged weapon? The Statistician 24(2):79-100. *

Blackith, R. E. and R. A. Reyment. 1971. Multivariate Morphometrics. London: Academic Press.

Bogan, M. A. 1974. Identification of Myotis californicus and M. leibii in southwestern North America. Proceedings of the Biological Society of Washington 87(7):49-56. [use of two group discriminant analysis and canonical variates] *

Bogan, M. A. 1978. A new species of Myotis from the Islas Tres Marias, Nayarit, Mexico, with comments on variation in Myotis nigricans. Journal of Mammalogy 59(3):519-530. [discriminant analysis; canonical variates plots with vectors for orignal variables plotted]

Bookstein, F. L. 1977a. Orthogenesis of the hominids: an exploration using biorthogonal grids. Science 197:901-904. *

Bookstein, F. L. 1977b. The study of shape transformation after D'Arcy Thompson. Mathematical Biosciences 34:177-219. *,T

Bookstein, F. L. 1978. The Measurement of Biological Shape and Shape Change. Levin, S. ed. Lecture Notes in Biomathematics, Vol. 24. Berlin: Springer-Verlag. *,T

Bookstein, F. L., P. D. Gingerich and A. G. Kluge. 1978. Hierarchical linear modeling of the tempo and mode of evolution. Paleobiology 4(2):120-134. [example of use of dummy variables in a regression model]

Bose, R. C., I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao and K. J. C. Smith, eds. 1970. Essays in Probability and Statistics Chapel Hill: The University of North Carolina Press. T

Bradu, D. and K. R. Gabriel. 1978. The biplot as a diagnostic tool for models of two-way tables. Technometrics, 20:47-68. T

Bradu, D. and F. E. Grine. 1979. Multivariate analysis of Diademodontine crania from South Africa and Zambia. South African Journal of Science 75:441-448. [example of biplot]

Brand, L. R. and R. E. Ryckman. 1969. Biosystematics of Peromyscus eremicus, P. guardia, and P. interparietalus. Journal of Mammalogy 50(3):501-513. [example of CVA] A

Bräuer, G. 1979. Some remarks on the interpretation of Penrose's size and shape components. Craniometrical similarities of African groups. Journal of Human Evolution 8:759-765.

Brower, J. C. and J. Veinus. 1974. The statistical zap versus the shotgun approach. International Association of Mathematical Geology, Journal 6(4):311-332. C

Brower, J. C. and J. Veinus. 1978. Multivariate analysis of allometry using point coordinates. Journal of Paleontology 52(5):1037-1053.

Bryant, E. H. and W. R. Atchley, eds. 1975. Multivariate Statistical Methods: Within-groups Covariation. Stroudsberg, Pennsylvania: Dowden, Hutchinson & Ross (Distributed by Halsted Press, a Division of John Wiley & Sons). [collected papers, reprinted; includes Rohlf, 1971; Rao, 1964; Jolicoeur and Mosimann, 1960; Cattell, 1965a; Jennrich and Sampson, 1966; Stroud, 1953 and Atchley, 1971] *,T

Bryant, E. H. and C. R. Turner. 1979. Comparative morphometric adaptation of the housefly and the face fly in the United States. Evolution 33(4):759-770.

Burnaby, T. P. 1966. Growth invariant discriminant functions and generalized distances. Biometrics 22(1):96-110. [reproduced in Atchley and Bryant, 1975] *,T

Cacoullos, T., ed. 1973. Discriminant Analysis and Applications. New York: Academic Press. T

Calhoon, R. E. and D. L. Jameson. 1970. Canonical correlation between variation in weather and variation in size in the Pacific tree frog, Hyla regilla, in Southern California. Copeia 1970(1):124-134. *

Carleton, M. D. and R. E. Eshelman. 1979. A synopsis of fossil grasshopper mice, genus Onychomys, and their relationships to Recent species. Claude W. Hibbard Memorial Volume 7, Papers on Paleontology No. 21, Museum of Paleontology, The University of Michigan, Ann Arbor, Michigan, 63 pp. [PCA and cluster analysis; use of MIDAS; minimum spanning tree; relates PCA scores to time scale]

Carlson, D. S. 1976. Patterns of morphological variation in the human midface and upper face. pp. 277-279 in J. A. McNamara Jr., ed., Factors Affecting the Growth of the Midface. Ann Arbor, Michigan: Center for Human Growth and Development. [PCA, terminology mixed up but clever way to plot components in terms of morphology] A

Castillo-Munoz, R. and R. J. Howarth. 1976. Application of the empirical discriminant function to regional geochemical data from the United Kingdom. Geological Society of America Bulletin 87:1567-1581. A

Cattell, R. B. 1965a. Factor analysis: An introduction to essentials. I. The purpose and underlying models. Biometrics 21:190-215. *,T

Cattell, R. B. 1965b. Factor analysis: An introduction to essentials. II. The role of factor analysis in research.

Biometrics 21:405-435.

Cavallaro, J. I., J. W. Menke and W. A. Williams. in press. paper presented at "Workshop: Use of Multivariate Statistics in Studies of Wildlife Habitat.", 23-25 April 1980, Burlington, Vermont. Proceedings in preparation, D. E. Capen, ed. [adjustment for multicollinearity in discriminant analysis]

Chaddna, R. L. and L. F. Marcus. 1968. An empirical comparison of distance statistics for populations with unequal covariance matrices. Biometrics 24(3):683-694. T

Chatterjee, S. and B. Price. 1977. Regression Analysis by Example. New York: John Wiley & Sons. *

Chen, Chi-hau. 1973. Statistical Pattern Recognition. Rochelle Park, New Jersey: Hayden Book Company.

Colgan, P. W. and J. T. Smith. 1978. Multidimensional contingency table analysis. Chapter 6, pp. 145-174 in Colgan, ed. 1978. [good introduction to the subject] *

Colgan, P. W., ed. 1978. Quantitative Ethology. New York: John Wiley & Sons. [collection of papers, many on multivariate methods]

Constandse-Westerman, T. S. 1972. Coefficients of Biological Distance. New York: Humanities Press. C

Cook, P. L. 1977. The genus Tremogasterina Canu (Bryozoa, Cheilostomata). Bulletin British Museum natural History (Zoology) 32(5):103-165.[use of PCORD]

Cook, R. C. and G. E. Lord. 1978. Identification of stocks of Bristol Bay Sockeye salmon, Oncorhynchus nerka, by evaluating scale patterns with a polynomial discriminant method. Fishery Bulletin 76(2):415-423. A

Cooley, W. W. and P. R. Lohnes. 1971. Multivariate Data Analysis New York, John Wiley & Sons. *,T,P

Corruccini, R. S. 1972. Allometry correction in taximetrics. Systematic Zoology 21:375-383.

Corruccini, R. S. 1973. Size and shape in similarity coefficients based on metric characters. American Journal of Physical Anthropology 38:743-754.

Corruccini, R. S. 1975a. Morphometric assessment of australopithecine postcranial affinities. Systematic Zoology 24(2):226-233. [size correction]

Corruccini, R. S. 1975b. Multivariate analysis in biological anthropology: some considerations. Journal of Human *

Evolution 4:1-19.

Corruccini, R. S. 1978. Primate skeletal allometry and hominoid evolution. Evolution 32(4):752-758.

Corruccini, R. S., R. L. Ciochon and H. Mc Henry. 1976. The postcranium of Miocene hominoids: were dryopithecines merely "dental apes"? Primates 17(2):205-223.

Cox, D. R. and N. J. H. Small. 1978. Testing multivariate normality. Biometrika 65(2):263-272.

Csuti, B. A. 1979. Patterns of adaptation and variation in the Great Basin kangaroo rat (Dipodomys microps). University of California Publications in Zoology 111:1-71. [use of discrimination function analysis] A

David, M., C. Campiglio, and R. Darling. 1974. Progress in R- and Q-mode analysis: Correspondence analysis and its application to the study of geological processes. Canadian Journal of Earth Sciences 11:131-146.

Davis, J. C. 1973. Statistics and Data Analysis in Geology. New York: John Wiley & Sons. [good summary of matrix algebra; chapter on multivariate statistics; includes FORTRAN programs for some of the procedures] *,P

Delaney, M. J. and M. J. R. Healy. 1964. Variation in the long-tailed field-mouse etc., II. Simultaneous examination of all characters. Proceedings of the Royal Society of London B,161:200-207. A

Dempster, A. P. 1964. Tests for the equality of two covariance matrices in relation to a best linear discriminator analysis. Annals of Mathematical Statistics 35:190-199. C

Dempster, A. P. 1969. Elements of Continuous Multivariate Analysis. Reading, Massachusetts: Addison-Wesley. T

Dempster, A. P. 1971. An overview of multivariate data analysis. Journal of Multivariate Statistics 1:316-346.

Dempster, A. P., N. M. Laird and D. B. Rubin. 1976. Maximum likelihood from incomplete data via the EM algorithm. Research Reports S-38, NS-320. Cambridge, Massachusetts: Department of Statistics, Harvard University. T

Dipillo, P. J. 1976. The application of bias to discriminant analysis. Communications in Statistics A5(9):834-844.

Dixon, W. J. 1977. BMDP-77. Biomedical computer programs P-series. Los Angeles, University of California Press. [a newsletter gives updates, there are supplementary descriptions and a shorter users guide] P

Dixon, W. J. and F. J. Massey, Jr. 1969. Introduction to Statistical Methods. 3rd ed., New York: McGraw-Hill.

Dodson, P. 1975. Functional and ecological significance of relative growth in Alligator. Journal of Zoology, London 175:315-355. [example of PCA for study of allometry] *

Dodson, P. 1976. Quantitative aspects of relative growth and sexual dimorphism in Protoceratops. Journal of Paleontology 50(5):929-940.[uses PCORD]

Doveton, J. H. 1979. Numerical methods for the reconstruction of fossil material in three dimensions. Geological Magazine 116(3):215-226. A

Draper, N. R. and H. Smith. 1966. Applied Regression Analysis. New York: John Wiley & Sons. [popular reference on multiple regression] *

Drennan, R. D. 1976. A refinement of chronological seriation using nonmetric multidimensional scaling. American Antiquity 41(3):290-302. A

Dunn, J. E. in press. paper presented at "Workshop: Use of Multivariate Statistics in Studies of Wildlife Habitat." 23-25 April 1980, Burlington, Vermont. Proceedings in preparation, D. E. Capin, ed. [up to date discussion of transformations in multivariate analysis]

Eger, J. L. and R. L. Peterson. 1979. Distribution and systematic relationship of Tadarida bivittata and Tadarida ansorgei (Chiroptera: Molossidae). Canadian Journal of Zoology, 57:1887-1895. [PCA on correlations with superimposed minimum spanning tree; discriminant analysis between species] A

Elder, W. H. and C. M. Hayden. 1977. Use of discriminant function in taxonomic determination of canids from Missouri. Journal of Mammalogy 58(1):17-24. A

Eldredge, N. 1968. Convergence between two Pennsylvanian gastropod species: a multivariate mathematical approach. Journal of Paleontology 42(1):186-196.

Enslein, K., A. Ralston and H. S. Wilf, eds. 1977. Statistical Methods for Digital Computers, Vol. III. New York: John Wiley & Sons. [explains algorithms for multivariate procedures] T

Everitt, B. 1978. Graphical techniques for multivariate data. London: Heinemann Educational Books. [fide review in Technometrics 21(1):134 for 1979] C

Everitt, B. S. 1979. A Monte Carlo investigation of the robustness of Hotelling's one- and two-sample $T^2$ tests. Journal of the American Statistical Association 74:48-51. T

Farris, J. S. 1969. A successive approximations approach to character weighting. Systematic Zoology 18(4):374-385. *,T

Farris, J. S. 1970. Methods for computing Wagner trees. Systematic Zoology 19(1):83-92.

Farris, J. S. 1972. Estimating phylogenetic trees from distance matrices. The American Naturalist 106:645-668.

Farris, J. S., A. G. Kluge and M. J. Eckardt. 1970. A numerical approach to phylogenetic systematics. Systematic Zoology 19(2):172-189.

Fasham, M. J. R. 1977. A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines, and coenoplanes. Ecology 58(3):551-561. A

Fienberg, S. 1977. The Analysis of Cross-Classified Categorical Data. Cambridge: MIT Press. *

Fisher, D. R. 1973. A comparison of various techniques of multiple factor analysis applied to biosystematic data. Science Bulletin, The University of Kansas 50(3):127-162.

Fix, A. G. and L. E. Lie-Injo. 1975. Genetic microdifferentiation in the Semai Senoi of Malaysia. American Journal of Physical Anthropology 43(1):47-55.

Fix, E. and J. L. Hodges. 1959. Discriminatory analysis: Nonparametric discrimination: consistency properties. Report No. 4, Project No. 21-49-004, School of Aviation Medicine, Randolph Air Force Base, Texas. C

Folse, L. J., Jr. in press. paper presented at "Workshop: Use of Multivariate Statistics in Studies of Wildlife Habitat.", 23-25 April 1980, Burlington, Vermont. Proceedings in preparation, D. E. Capen, ed. [example of canonical correlation]

Frane, J. W. 1976. Some simple procedures for handling missing data in multivariate analysis. Psychometrika 41(3):409-415. *

Gabriel, K. R. 1969. A comparison of some methods of simultaneous inference in MANOVA. pp. 67-86 in Krishnaiah, 1969. T

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58(3):453-467.

Gabriel, K. R. and S. Zamir. 1979. Lower approximation of matrices by least squares with any choice of weights. Technometrics 21(4):489-498 [additional biplot reference] T

Gnent, A. W. 1979. Some considerations governing the selection of appropriate statistical procedures I. Questions of numerical scale and research interest. The Biologist 61(2):59-73. *

Giles, E. 1960. Multivariate analysis of Pleistocene and Recent coyotes (Canis latrans) from California. University of California Publications in Geological Sciences 36(8):369-390. [use of Mahalanobis $D^2$ to compare subspecies] *

Gnanadesikan, R. 1977. Methods for Statistical Data Analysis of Multivariate Observations. New York: John Wiley & Sons. [the most recent comprehensive treatment; extremely useful]

Gnanadesikan, R. and J. R. Kettenring. 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28:81-124. T

Goldstein, M. and W. R. Dillon. 1978. Discrete Discriminant Analysis. New York: John Wiley & Sons. T,P

Good, P., ed. 1979. Randomization Vol. 1. 10367 Paw Paw Lake Dr., Mattawan, Michigan 49071. [new journal on randomization tests; write editor for further information]

Gould, S. J. 1966. Allometry and size in ontogeny and phylogeny. Biological Reviews 41:587-640.

Gould, S. J. 1967. Evolutionary patterns in pelycosaurian reptiles: a factor analytic study. Evolution 21:385-401. [R-mode factor analysis with oblique rotation; Q-mode using non-centered R and oblique rotation] *

Gould, S. J. 1969. An evolutionary microcosm: Pleistocene and Recent history of the land snail P. (Poecilozonites) in Bermuda. Bulletin of the Museum of Comparative Zoology 138(7):407-531.

Gould, S. J. and R. A. Garwood. 1969. Levels of integration in mammalian dentitions: an analysis of correlations in Nesophontes micrus (Insectivora) and Oryzomys couesi (Rodentia). Evolution 23(2):276-300. [compares Olson and Miller, 1958 technique with factor analysis].

Gould, S. J. and R. F. Johnston. 1972. Geographic variation. Annual Review of Ecology and Systematics 3:457-498.

Gower, J. C. 1966a. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53(3&4):325-338. [Classical paper on principal coordinates analysis--well worth reading] *,T

Gower, J. C. 1966b. A Q-technique for the calculation of canonical variates. Biometrika 53(3&4):588-590. *,T

Gower, J. C. 1968. Adding a point to vector diagrams in multivariate analysis. Biometrika 55:582-585.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857-870.

Gower, J. C. 1976. Growth-free canonical variates and generalized inverses. Bulletin of the Geological Institutes of the University of Uppsala, New Series 7:1-10. T

Green, P. E. and J. D. Carroll. 1976. Mathematical Tools for Applied Multivariate Analysis. New York: Academic Press. [development of matrix algebra and geometric relations] *

Habbema, J. D. F. and J. Hermans. 1977. Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics 19(4):487-493.

Harman, H. H. 1967. Modern Factor Analysis, 2nd Ed. Chicago: The University of Chicago Press.

Harner, E. J. and R. C. Whitmore. in press. paper presented at "Workshop: Use of Multivariate Statistics in Studies of Wildlife Habitat.", 23-25 April 1980, Burlington, Vermont. Proceedings in preparation, D. E. Capen, ed. [paper on robust discrimination]

Harris, C. W. and H. F. Kaiser. 1964. Oblique factor analytic solutions by orthogonal transformations. Psychometrika, 29(4):347-362. [reference for SPACEWARP rotation available in NTSYS]

Harris, R. J. 1975. A Primer of Multivariate Statistics. New York: Academic Press. [First chapter gives good overview of multivariate statistics. Social sciences perspective.]

Hartigan, J. A. 1975. Clustering algorithms. New York: John Wiley & Sons. T,P

Hartigan, J. A. 1977. Distribution problems in clustering. pp. 45-71 in Van Ryzin, ed. 1977.

Hawkins, D. M. 1974. The detection of errors in multivariate data using principal components. Journal of the American Statistical Association 69:340-344. T

Hellack, J. J. and G. D. Schnell. 1977. Phenetic analysis of the subfamily Cardinalinae using external and skeletal characters. The Wilson Bulletin 89(1):130-148. [cluster and PCA; superimposed MST] *

Hill, M. O. 1974. Correspondence analysis: a neglected multivariate method. Applied Statistics 23:340-354. T

Hill, M. O. and A. J. E. Smith. 1976. Principal component analysis of taxonomic data with multi-state characters. Taxon 25(2/3):249-255. [shows PCA on discrete data leads to correspondence analysis]

Hills, M. and D. R. Brothwell. 1974. The use of large numbers of variables to measure the shape of a restricted area of bone. Journal of Archaeological Science 1:135-150.

Hocking, R. R. 1976. The analysis and selection of variables in linear regression. Biometrics 32:1-49. *,T

Hocking, R. R. 1977. Selection of the best subset of regression variables. pp. 39-57 in Enslein, Ralston and Wilf, eds. 1977. T

Hodges, J. L. 1955. Discriminatory Analysis. 1. Survey of Discriminatory Analysis. Project #21-49-004, Report #1, School of Aviation Medicine, United States Air Force, Randolph Air Force Base, Texas. [excellent introductory review and development] *

Hoffman, R. S., J. W. Koeppl and C. F. Nadler. 1979. The relationships of the Amphiberigian marmots (Mammalia: Sciuridae). Occasional Papers of the Museum of Natural History, Lawrence Kansas, The University of Kansas No. 83; pp.1-56. [Uses variety of methods in concert--PCA, canonical variates, missing data replaced by regresson; use of BMDP and NTSYS] *

Hohn, M. E. 1978. Stratigraphic correlation by principal components: effects of missing data. Journal of Geology 86:524-532. A

Holmes, J. M. C. 1975. A comparison of numerical taxonomic techniques using measurements on the genera Gammarus and Marinogammarus (Amphipoda). Biological Journal of the Linnean Society 7(3):183-214.[comparison of PCORD, PCA, clustering and canonical variates as ordination procedures]

Hope, K. 1968. Methods of Multivariate Analysis. London: University of London Press. [not available in U.S.; clear introduction to methodology with examples] *

Hopkins, J. W. 1966. Some considerations in multivariate allometry. Biometrics 22:747-760.

Horst, P. 1965. Factor Analysis of Data Matrices. New York: Holt, Rinehart and Winston. P

Howarth, R. J. 1971. An empirical discriminant method applied to sedimentary-rock classification from major-element geochemistry. International Association of Mathematical Geology, Journal 3(1):51-60.

Howarth, R. J. 1973. The pattern recognition problem in applied geochemistry. pp. 259-273 in Exploration Geochemistry 1972, M. J. Jones London: Institution of Mining and Metallurgy.

Howells, W. W. 1957. Cranial variation in man. Papers of the Peabody Museum of Archaeology and Ethnology 67:1-259. A

Hull, C. H. and N. H. Nie. 1979. SPSS Update. New Procedures and Facilities for Releases 7 and 8. New York:McGraw-Hill. [some changes to discriminant procedures] P

Imbrie, J. and T. H. Van Andel. 1964. Vector analysis of heavy-mineral data. Geological Society of America Bulletin 75:1131-1156. *

Ito, K. 1969. On the effect of heteroscedasticity and nonnormality upon some multivariate test procedures. pp. 87-120 in Krishnaiah, 1969. T

Ito, K. and W. J. Schull. 1964. On the robustness of the $T_o^2$ test in multivariate analysis of variance when variance-covariance matrices are not equal. Biometrika, 51:71-82. T

Jackson, J. E. and G. S. Mudholkar. 1979. Control procedures for residuals associated with principal component analysis. Technometrics 21(3):341-349.

Jacobshagen, B. 1979. Morphometric studies in the taxonomy of the orang-utan (Pongo pygmaeus, L. 1760).Folia primatologica 32:(1-2):29-34.

Jamison, P. L. and S. L. Zegura. 1973. A univariate and multivariate examination of measurement error in anthropometry. American Journal of Physical Anthropology 40:197-203.

Jardine, N. and R. Sibson. 1971. Mathematical Taxonomy. London: John Wiley & Sons. *,T

Jenkins, P. D. 1976. Variation in eurasian shrews of the genus Crocidura (Insectivora: Soricidae). Bulletin of the British Museum (Natural History) 30(7):271-309. [large matrix of D among species] A

Jennrich, R. I. 1977a. Stepwise regression. pp. 58-75 in Enslein, Ralston and Wilf, Eds. 1977. T

Jennrich, R. I. 1977b. Stepwise discriminant analysis. pp. 76-95. in Enslein, Ralston and Wilf, eds. 1977. T

Jennrich, R. I. and P. F. Sampson. 1966. Rotation for simple loadings. Psychometrika 31(3):313-323. C

Johnson, S. C. 1967. Hierarchical clustering schemes. Psychometrika C

32:241-254.

Johnston, R. F. 1973. Evolution in the house sparrow. IV. Replicate studies in phenetic variation. Systematic Zoology 22:219-226. [principal component analysis at low taxonomic level for large samples]

Jolicoeur, P. 1959. Multivariate geographical variation in the wolf *
Canis lupus L. Evolution 13(3):283-299. [reproduced in Atchley and Bryant, 1975]

Jolicoeur, P. 1963a. The degree of robustness in Martes americana. *
Growth, 27:1-27. [tests a priori allometric hypothesis of relation of lengths and widths of limb bones]

Jolicoeur, P. 1963b. The multivariate generalization of the T
allometry equation. Biometrics 19:497-499.

Jolicoeur, P. and J. E. Mosimann. 1960. Size and shape variation in *
the painted turtle: A principal component analysis. Growth 24(4):339-354. [Reprinted in Bryant and Atchley, 1975]

Jöreskog, K. G., J. E. Klovan, and R. A. Reyment. 1976. Geological *
Factor Analysis. Methods in Geomathematics 1. Amsterdam: Elsevier.

Kaesler, R. L. and J. A. Waters. 1972. Fourier analysis of the ostracode margin. Geological Society of America Bulletin 83:1169-1178.

Kaesler, R. L., P. S. Mulvany and L. S. Kornicker. 1977. Delimitation of the Antartic convergence by cluster analysis and ordination of benthic Myodocopid Ostracoda. Saalfeiden, Sixth International Ostracod Symposium, pp.235-244. [use of multidimensional scaling]

Kaiser, H. F. 1970. A second generation little jiffy. Psychometrika C
35:401-415.

Karr, J. R. and F. C. James. 1975. Eco-morphological configurations A
and convergent evolution in species communities. pp. 258-291 in Cody, M. L. and J. M. Diamond, eds., Ecology and Evolution of Communities. Cambridge, Massachusetts: The Belknap Press. [canonical correlation analysis of relation between environment and morphology]

Katz, J. O. and F. J. Rohlf. 1974. Functionplane--a new approach to T
simple structure rotation. Psychometrika 39(1):37-51. [method available in NTSYS as FUNCTNPLN]

Katz, J. O. and F. J. Rohlf. 1975. Primary product functionplane: T
an oblique rotation to simple structure. Multivariate Behavioral Research 10:219-232. [available in NTSYS as PRIMENPL - also see erratum 10:509-511 of same journal]

Kay, R. F. 1975. The functional adaptations of primate molar teeth. American Journal of Physical Anthropology 43(2):195-215. A

Kempthorne, O., T. A. Bancroft, J. W. Gowen and J. L. Lush, eds. 1954. Statistics and Mathematics in Biology. Ames, Iowa: The Iowa State College Press. [short useful articles on multivariate statistics with applications; includes Tukey, 1954, Wright, 1954 and Rao, 1954] *,T

Kendall, M. G. 1957. A Course in Multivariate Analysis. London: Charles Griffins. [widely cited, e.g. Cooley and Lohnes, 1971] C

Kennedy, M. L. and G. D. Schnell. 1978. Geographic variation and sexual dimorphism in Ord's kangaroo rat, Dipodomys ordii. Journal of Mammology 59(1):45-59.[PCA on area means for large samples using NTSYS; minimum spanning tree; geographical plots of PCA scores] *

Kim, J. and C. W. Mueller. 1978a. Introduction to Factor Analysis. What It Is and How To Do It. No. 13 in Series: Quantitative Applications in the Social Sciences. Beverly Hills, Sage Publications. [Excellent short introductory paperback. Includes comprehensive glossary, and a guide to the literature. Simple setups for use of computer packages-- SPSS, BMD, SAS and OSIRIS] *

Kim, J. and C. W. Mueller. 1978b. Factor Analysis. Statistical Methods and Practical Issues. No. 14 in Series: Quantitative Applications in the Social Sciences. Beverly Hills, Sage Publications. [Same biblio. and glossary as 1978a. Good concise introductory coverage of factor methods, number of factors, rotation, etc.] *

Kleinbaum, D. G. and L. L. Kupper. 1978. Applied Regression Analysis and Other Multivariable Methods. North Scituate, Massachusetts: Duxbury Press. *

Kowalski, C. J. 1972. A commentary on the use of multivariate statistical methods in anthropomorphic research. American Journal of Physical Anthropology, 36:119-132. *

Krishnaiah, P. R., ed. 1966. Multivariate Analysis. New York: Academic Press. [Proceedings of an International Symposium] T

Krishnaiah, P. R., ed. 1969. Multivariate Analysis-II. New York: Academic Press. [Proceedings of the Second International Symposium on Multivariate Analysis] T

Krishnaiah, P. R., ed. 1973. Multivariate Analysis-III. New York: Academic Press. Proceedings of the 3rd International Symposium on Multivariate Analysis; technical statements on state of the art] T

Krishnaiah, P. R., ed. 1977. Multivariate Analysis-IV. Amsterdam: T

North-Holland. [Proceedings of the Fourth International Symposium; technical state of the art papers]

Kruskal, J. B. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1-27. T

Kruskal, J. B. 1964b. Nonmetric multidimensional scaling: a numerical method Psychometrika 29:28-42. T

Kruskal, J. B. 1977. Multidimensional scaling and other methods for discovering structure. pp. 296-339. in Enslein, Ralston and Wilf, eds. 1977. T

Kruskal, J. B. and M. Wish. 1978. Multidimensional Scaling. No. 11 in Series: Quantitative Applications in the Social Sciences, Beverly Hills, Sage Publications. [Excellent paperback introduction to the subject. Includes guide to available programs.] *

Kshirsagar, A. M. 1972. Multivariate Analysis. New York: Marcell Dekker. T

Kuhry, B. and L. F. Marcus. 1977. Bivariate linear models in biometry. Systematic Zoology 26(2):201-209.

Kullback, S. 1968. Information Theory and Statistics. New York: Dover Publications. T

Lachenbruch, P. A. 1975. Discriminant Analysis. New York: Hafner Press. *,T

Lachenbruch, P. A. and M. Goldstein. 1979. Discriminant analysis. Biometrics 35:69-85. [Current review]. *,T

Lavelle, L. B. 1977. Relationship between tooth and long bone size. American Journal of Physical Anthropology 46:423-425. [use of canonical correlation]

Lestrel, P. E., W. H. Kimbel, F. W. Prior and M. L. Fleischmann. 1977. Size and shape of the hominoid distal femur: Fourier analysis. American Journal of Physical Anthropology 46:281-290. *

Levine, M. S. 1977. Canonical Analysis and Factor Comparison. No. 6 in Series: Quantitative Applications in the Social Sciences. Beverly Hills, California: Sage Publications [short paperback on canonical correlation and factor comparisons]

Lu, K. H. 1965. Harmonic analysis of the human face. Biometrics 21:491-505. *

Manaster, B. J. 1979. Locomotor adaptations within the Cercopithecus genus: a multivariate approach. American Journal of Physical A

Anthropology 50:169-182.

Marcus, L. F. 1969. Measurement of selection using distance statistics in the prehistoric orang-utan Pongo pygmaeus palaeosumatrensis. Evolution 23(2):301-307.

Marcus, L. F. and J. H. Vandermeer. 1966. Regional trends in geographic variation. Systematic Zoology 15(1):1-13.

Mather, P. M. 1976. Computational Methods of Multivariate Analysis in Physical Geography. New York: John Wiley & Sons. [includes computer programs for many of the methods including PCA, principal coordinates and nometricl multidimensional scaling]

McHenry, H. M. and R. S. Corruccini. 1975. Multivariate analysis of early hominid pelvic bones. American Journal of Physical Anthropology 43:263-270. [Uses PCA and principal coordinates; size removed; interesting but confusing] A

McMahon, T. 1973. Size and shape in biology. Science 179:1201-1204. *

McMahon, T. 1975. Allometry and biomechanics: limb bones in adult ungulates. The American Naturalist 108:547-563. A

Menozzi, P., A. Piazza and L. Cavalli-Sforza. 1978. Synthetic maps of human gene frequencies in Europeans. Science 201(1):786-792.

Mickevich, M. F. and M. S. Johnson. 1976. Congruence between morphological and allozyme data in evolutionary inference and character evolution. Systematic Zoology 25:260-270. *

Morrison, D. F. 1976. Multivariate Statistical Methods, 2nd. ed.. New York: McGraw-Hill. *,T

Mosimann, J. E. 1970. Size allometry: size and shape variables with characterizations of the lognormal and generalized gamma distributions. Journal of the American Statistical Association 65:930-945. T

Mosimann, J. E. and F. C. James. 1979. New statistical methods for allometry with application to Florida red-winged blackbirds. Evolution 33:444-459. *

Moss, W. W., P. C. Peterson and W. T. Atyeo. 1977. A multivariate assessment of phenetic relationships within the feather mite familty Eustathiidae (Acari). Systematic Zoology 26(4):386-409.

Mosteller, F. and J. W. Tukey. 1977. Data Analysis and Regression. Reading, Massachusetts: Addison-Wesley. *,T

Mulaik, S. A. 1972. The Foundations of Factor Analysis. New York: *,T

McGraw-Hill. [comprehensive book with theory and good discussion (especially of confirmatory factor analysis)]

Neff, N. A. and G. R. Smith. 1979. Multivariate analysis of hybrid fishes. Systematic Zoology 28(2):176-196. *

Nie, N. H., C. H. Hull, K. Steinbrenner and D. H. Bent. 1975. SPSS. Statistical Package for the Social Sciences, 2nd. ed. New York: McGraw-Hill.

Olson, E. C. 1964. Morphological integration and the meaning of characters in classification systems. pp. 123-156 in Phenetic and Phylogenetic Classification. Systematics Association Publication Number 6, London. *

Olson, E. C. and R. L. Miller. 1958. Morphological Integration. Chicago: The University of Chicago Press. *

Overall, J. E. and C. J. Klett. 1972. Applied Multivariate Analysis. New York: McGraw-Hill.

Oxnard, C. E. 1969. Mathematics, shape and function: a study in primate anatomy. American Scientist 57:75-96. *

Oxnard, C. 1973. Form and Pattern in Human Evolution: Some Mathematical, Physical and Engineering Approaches. Chicago: The University of Chicago Press. [Some good explanations of methods with primate examples] *

Oxnard, C. 1975. Uniqueness and Diversity in Human Evolution. Chicago: University of Chicago Press. [Andrews plots used and experimented with] *

Oxnard, C. E. 1978. One biologist's view of morphometrics. Annual Review of Ecology and Systematics 9:219-241. *

Pearson, E. S. and N. W. Please. 1975. Relation between the shape of population distribution and the robustness of four simple test statistics. Biometrika 62(2):223-241. *

Penrose, L. S. 1954. Distance, size and shape. Annnals of Eugenics 18:337-343. C

Petit-Maire, N. and J. F. Ponge. 1979. Primate cranium morphology through ontogenesis and phylogenesis, factorial analysis of global variation. Journal of Human Evolution 8:233-234. [use of correspondance analysis]

Pilbeam, D. R. 1969. Tertiary Pongidae of East Africa: Evolutionary relationships and taxonomy. Bulletin Peabody Museum of Natural History Yale University 31:1-185.[use of PCORD; some missing data]

Pilbeam, D. and S. J. Gould. 1974. Size and scaling in human A

evolution. Science 186:892-901.

Pimentel, R. A. 1979. Morphometrics: The Multivariate Analysis of Biological Data. Dubuque, Iowa: Kendall/Hunt. [useful approach to many methods useful to systematists; however there are many errors (some indicated on an erratum sheet), the examples are not well chosen, and the terminology is careless--one hopes that a careful editing of a future edition will make this important attempt more useful]

Pizzimenti, John J. 1975. Evolution of the prairie dog genus Cynomys. Occasional papers of the Museum of Natural History, The University of Kansas, Lawrence Kansas No. 39, pp. 1-73. [estimates missing data using regression, PCA in both NTSYS and BMD on both R and S matrix with comparison; minimum spanning tree] *

Platt, T. and K. L. Dennan. 1975. Spectral analysis in ecology. Annual Review of Ecology and Systematics 6:189-210. A

Porebski, O. R. 1966. On the interrelated nature of the multivariate statistics used in discriminatory analysis. British Journal of Mathematical and Statistical Psychology 19(2):197-214. [reproduced in Atchley and Bryant, 1975] *,T

Press, S. J. 1972. Applied Multivariate Analysis. New York: Holt, Rinehart and Winston.

Press, S. J. and S. Wilson. 1978. Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association 73:699-705. *,T

Rak, Y. Z. 1977. Metric description and analysis of cranial contours. Kroeber Anthropological Society Papers 50:13-20. A

Ralston, A. and H. S. Wilf. 1964. Mathematical Methods for Digital Computers. New York: John Wiley & Sons. T

Rao, C. R. 1952. Advanced Statistical Methods in Biometric Research. New York: John Wiley & Sons. *,T

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhya, Ser. A, 26:329-358. [reprinted in Bryant and Atchley, 1975] *,T

Rao, C. R. 1965. Linear Statistical Inference and Its Applications. New York: John Wiley & Sons. T

Rao, C. R. 1970. Inference on discriminant function coefficients. pp. 587-602 in Bose et al. 1970. T

Rao, C. R. 1972. Recent trends of research work in multivariate analysis. Biometrics 28:3-22. *,T

Rempe, U. and Weber, E. E. 1972. An illustration of the principal ideas of MANOVA. Biometrics 28:235-238. T

Rencher, A. C. and F. C. Pun. 1980. Inflation of $R^2$ in best subset regression. Technometrics 22(1):49-53. C

Reyment, R. A. 1969. A multivariate paleontological growth problem. Biometrics 25(1):1-8. [tests for orientations of principal component axes]

Reyment, R. A. 1971. Multivariate normality in morphometric analysis. International Association for Mathematical Geology, Journal 3(4):357-368.

Reyment, R. and C. F. Banfield. 1976. Growth-free canonical variates applied to fossil foraminifers. Bulletin of the Geological Institutes of the University of Uppsala, New Series 7;11-21.

Rightmire, G. P. 1976. Multidimensional scaling and the analysis of human biological diversity in subsaharan Africa. American Journal of Physical Anthropology 44:445-451.

Robinson, J. W. and R. S. Hoffmann. 1975. Geographical and interspecific cranial variation in big-eared ground squirrels (Spermophilus): a multivariate study. Systematic Zoology 24:79-88. [Compares PCA results to canonical variates (calls discrimination); states assumptions for tests; good discussion] *

Rohlf, F. J. 1970. Adaptive hierarchical clustering schemes. Systematic Zoology 19:58-82. [reproduced in Atchley and Bryant, 1975] *,T

Rohlf, F. J. 1971. Perspectives on the application of multivariate statistics to taxonomy. Taxon 20(1):85-90. [reproduced in Bryant and Atchley,1975 and Atchley and Bryant, 1975; good review] *

Rohlf, F. J. 1972. An empirical comparison of three ordination techniques in numerical taxonomy. Systematic Zoology 21: 271-280. [comparison of MDS, PCA and principal coordinates; reproduced in Atchley and Bryant, 1975] *

Rohlf, F. J. 1975. Generalization of the gap test for the detection of multivariate outliers. Biometrics 31:93-101.

Rohlf, F. J. 1977. [author's reply to Warde and Norton, 1977]. Biometrics, 33(4):763-764.

Rohlf, F. J. and R. R. Sokal. 1962. The description of taxonomic relationships by factor analysis. Systematic Zoology 11(1):1-16. *

Rummel, R. J. 1970. Applied Factor Analysis. Evanston, Illinois: *

Northwestern University Press.

Scheffe, H. 1959. The Analysis of Variance. New York: John Wiley & Sons. T

Schnell, G. D. 1970. A phenetic study of the Suborder Lari (Aves) I. Methods and results of principal component analyses. Systematic Zoology 19(1):35-57. [example of MST used with PCA] *

Seal, H. 1964. Multivariate Statistical Analysis for Biologists. New York: John Wiley & Sons. *,T

Searle, S. R. 1966. Matrix Algebra for Biological Sciences. New York, John Wiley & Sons. C

Simpson, G. G. 1941. Large Pleistocene felids of North America. American Museum Novitates no. 1136, 27 pp. *

Skeel, M. A. and L. N. Carbyn. 1977. The morphological relationship of gray wolves (Canis lupus) in national parks of central Canada. Canadian Journal of Zoology 55(4):737-747. [use of nonmetrical multidimensional scaling, PCA and discriminant analysis] A

Sneath, P. H. A. and R. R. Sokal. 1973. Numerical Taxonomy. San Francisco: W. H. Freeman. *

Snedecor, G. W. and W. G. Cochran. 1967. Statistical Methods, 6th ed. Ames, Iowa: The Iowa State University Press. [good practical book for ANOVA and multiple regression] *

Sokal, R. R. 1965. Statistical methods in systematics. Biological Reviews 40:337-391. [mainly univariate] *

Sokal, R. R. and N. L. Oden. 1978. Spatial autocorrelation in biology: 1. Methodology. Biological Journal of the Linnean Society 10:199-228. A

Sokal, R. R. and N. L. Oden. 1978. Spatial autocorrelation in biology: 2. Some biological implications and four applications of evolutionary and ecological interest. Biological Journal of the Linnean Society 10(2):229-249.

Sokal, R. R. and R. C. Rinkel. 1963. Geographic variation of alate Pemphigus populitransversus in eastern North America. University of Kansas Science Bulletin 64(10):467-507. [example of plots of factor scores] *

Sokal, R. R. and F. J. Rohlf. 1969. Biometrics. San Francisco: W. H. Freeman. *

Sokal, R. R. and P. H. A. Sneath. 1963. Principles of Numerical Taxonomy. San Francisco: W. H. Freeman. *

Sokal, R. R. and P. A. Thomas. 1965. Geographic variation of Pemphigus populi-transversus in eastern North America: stem mothers and new data on alates. University of Kansas Science Bulletin 46(5):201-252. [use of factor analysis and plotting of factor scores] *

Sokal, R. R., H. V. Daly, and F. J. Rohlf. 1961. Factor analytical procedures in a biological model. University of Kansas Science Bulletin 42(10):1099-1121. *

Sörbom, D. and K. G. Jöreskog. 1976. COFAMM: Confirmatory Factor Analysis with Modal Modification: User's Guide. Chicago: National Educational Resources. P

Spain, A. V., G. E. Heinsohn, H. Marsh and R. L. Correll. 1976. Sexual dimorphism and other sources of variation in a sample of dugong skulls from north Queensland. Australian Journal of Zoology 24:491-497. [uses GENSTAT; PCA on ratios to condylo-premaxillary length to remove "size"; uses residuals from one-way ANOVA to remove sex differences] A

Spicer, R. A. and C. R. Hill. 1979. Principal components and correspondence analyses of quantitative data from a Jurassic plant bed. Review of Paleobotany and Palynology 28:273-299. A

Stroud, C. P. 1953. An application of factor analysis to the systematics of Kalotermes. Systematic Zoology 2(2):76-92. [discussion of simple structure; reprinted in Bryant and Atchley, 1975]

Tatsuoka, M. M. 1971. Multivariate Analysis: Techniques for Educational and Psychological Research. New York: John Wiley & Sons.

Thompson, D'A. W. 1961. On Growth and Form (orig. 1917, 1942), ed. J. T. Bonner. Cambridge: The University Press. C

Thorndike, R. M. 1978. Correlational Procedures for Research. New York: Gardner Press (Distributed by Halsted Press, a Division of John Wiley & Sons.)

Thorpe, R. S. 1976. Biometric analysis of geographic variation and racial affinities. Biological Reviews 51:407-452. *

Thorpe, R. S. 1980. A comparative study of ordination techniques in numerical taxonomy in relation to racial variation in the ringed snake Natrix natrix (L.). Biological Journal of the Linnean Society 13:7-40. [Compares PCA, PCORD and non-metrical multidimensional scaling for R-mode and Q-mode techniques] *

Thurstone, L. L. 1947. Multiple Factor Analysis. Chicago: University of Chicago Press. C

Timm, N. H. 1975. Multivariate Analysis with Applications in Education and Psychology. Monterrey, California: Brooks/Cole Publishing Company.

Torgerson, W. S. 1958. Theory and Methods of Scaling. New York: John Wiley & Sons. C

Tukey, J. W. 1954. Causation, rgression and path analysis. pp. 35-66 in Kempthorne et al., eds., 1954. *,T

Tukey, J. W. 1977. Exploratory Data Analysis. Reading, Massachusetts: Addison-Wesley. *

Tukey, J. W. 1980. We need both exploratory and confirmatory. The American Statistician 34(1):23-25 [pithy article well worth reading] *

Van de Geer, J. P. 1971. Introduction to Multivariate Analysis for the Social Sciences. San Francisco: W. H. Freeman. *,T

Van Ness, J. 1979. On the effects of dimension in discriminant analysis for unequal covariance populations. Technometrics 21(1):111-127. *

Van Ness, J. W. and C. Simpson. 1976. On the effects of dimension in discriminant analysis. Technometrics 18(2):175-187. *

Van Ryzin, J., ed. 1977. Classification and Clustering. New York: Academic Press. *,T

Van Valen, L. 1974. Multivariate structural statistics in natural history. Journal of Theoretical Biology 45:235-247.

Van Valen, L. 1978. The statistics of variation. Evolutionary Theory 4:33-43. *

Vark, G. N. van. 1974. The investigation of human cremated skeletal material by multivariate statistical methods. I. Methodology. Ossa 1:63-95. *,T

Vark, G. N. van. 1976. A critical evaluation of the application of multivariate statistical methods to the study of human populations from their skeletal remains. Homo 27(2):94-114. *,T

Veldman, D. J. 1967. Fortran Programming for the Behavioral Sciences. New York: Holt, Rinehart, and Winston. P

Wallace, J. T. and R. S. Bader. 1967. Factor analysis in morphometric traits of the house mouse. Systematic Zoology 16(2):144-148. *

Warde, W. D. and J. M. Norton 1977. On Rohlf's generalization of the gap test for the detection of multivariate outliers. Biometrics 33:762-763. [See Rohlf, 1975 and Rohlf, 1977]

Waterman, T. H. and H. J. Morowitz, eds. 1965. Theoretical and Mathematical Biology. New York: Blaisdell. [includes articles on multivariate statistical applications by M. S. Bartlett and R. E. Blackith]

Waters, J. A. 1977. Quantification of shape by use of Fourier analysis: the Mississippian blastoid genus Pentremites. Paleobiology 3:288-299.

Wnite, J. W. and R. F. Gunst. 1979. Latent root regression: Large sample analysis. Technometrics 21(4):481-488. T

Wichern, D. W. and G. A. Churchill. A comparison of ridge estimators. Technometrics 20(3):301-311. T

Wilk, M. B. and R. Gnanadesikan. 1961. Graphical analysis of multi-response data using ordered distances. Proceedings of the National Academy of Sciences 47(8):1209-1212. T

Willan, A. R. and D. G. Watts. 1978. Meaningful multicollinearity measures. Technometrics 20(4):407-412. T

Williams, W. T. 1971. Principles of clustering. Annual Review of Ecology and Systematics 2:203-326. C

Williamson, M. H. 1978. The ordination of incidence data. Journal of Ecology 66:911-920. [correction for principal coordinates of 0,1 data to avoid Kendall's horseshoe]

Wishart, D. 1978. CLUSTAN user manual. Edinburgh: Edinburgh University. C

Wolpoff, M. H. 1976. Multivariate discrimination, tooth measurements and early hominid taxonomy. Journal of Human Evolution 5:339-344.

Wright, S. 1954. The interpretation of multivariate systems. pp. 11-33 in Kempthorne et al., eds., 1954. [Discussion of path analysis] *,T

Younker, J. L. and R. Ehrlich. 1977. Fourier biometrics: harmonic amplitudes as multivariate shape descriptors. Systematic Zoology 26:336-342.

# PUBLICATION OF THE RESULTS OF MULTIVARIATE STUDIES

by Michael A. Bogan

Several factors govern both the style and content of published multivariate studies. Prior to discussing some of these factors, it may be useful to reiterate why we publish our results at all. Rummel [1970] has suggested the following: 1) To allow critical evaluation of the research so that its substantive conclusions can be given proper weight and compared with other studies, 2) To allow others to independently confirm or refute the results, i.e. to allow replication, 3) To advance knowledge through the publication of findings, and 4) To encourage similar efforts in allied areas or to suggest alternative designs or hypotheses. Multivariate analyses are one of many tools available to researchers to use in scientific studies. However, due to their complex nature, especially to the uninitiated, it is important that enough information be published to enable the readers to evaluate the methods used, and if they choose, to replicate the study on the same or a new data set.

Other sections of this manual have addressed the overall design and use of multivariate analyses, so I will restrict myself to a few general comments on the publication of properly designed multivariate studies. Probably the most often-asked question is "How much of my data and methods should be published?". The usual answer to this question is that sufficient information must be presented to clearly outline your approach and methods and to enable others to understand and utilize your methods. However, the more practical answer is that the mode of publication will greatly influence, if not actually determine what will be published. This consideration renders absolute guidelines difficult to formulate. I will suggest some minimal standards that seem to me to be useful or appropriate.

First of all, it will rarely be possible to publish (all of) the raw or original data. Possible exceptions to this may be publications of a monographic nature. It is necessary to inform the reader where a copy of the original data may be obtained. In systematic studies, this also may be true of lists of specimens examined. Thus, the Methods section of the paper should include a statement that the original data (and a list of specimens examined) are on file and give instructions on how to obtain copies. Such material may consist of typed lists or tables, copies of original data sheets, magnetic tape records, or punched cards. These data should be provided free or at a nominal charge to interested parties. Depending on space limitations in the journal, these comments also may apply to long lists of characters or coded traits used in systematic studies. Some of this material also may be presented in Appendices to the paper.

Obviously, some parts or subsets of the data will be published, often in summary form (e.g. means and variances for example) as tables or figures. Some types of intermediate results or multivariate statistics, such as correlation or similarity-distance matrices, fall into this category. The decision on publishing such material should be based on the importance of these data to the results and conclusions to be presented. In addition, the editors and reviewers will comment on the appropriateness of including such information.

Where space limitations restrict the amount of such data to be published, the data should be available from the author. Whenever statistical data are presented to which tests of significance may be applied, appropriate information should be included to support the tests of significance (p values, n, degrees of freedom).

Although the publication of intermediate or exploratory results on the data may be possible, any procedure that was used in preliminary stages to sort or explore the data should be briefly outlined. This will allow the reader to follow the development of the study.

Those portions of the data or results that are used in formulating conclusions should be presented. In the case of multivariate studies, this might include character loadings, correlation coefficients, dendrograms, and plots of OTUs on axes. Publication of such results allows the reader to evaluate the author's conclusions. Available measures of distortion (e.g. cophenetic correlation coefficients) should be given.

All statistical methods used should be clearly and succinctly explained, or a published reference given, in the Methods section of the paper. In cases where a "canned" computer package has been used, a citation to the published documentation may suffice. Examples of this would be the use, in a standard fashion, of programs from BMD-BMDP, SAS, SPSS; all of these have published reference manuals (see Appendix II). Dave Schmidly has provided information on some of these programs. New and unique computer multivariate programs, developed for specific purposes and not accompanied by published documentation, should be explained. Special features and computational details should be presented, and if necessary, be accompanied by mathematical formulae. Some comment should be made regarding the availability of the programs and where they may be obtained. If the programs have been presented in published form, but not used previously in a practical application, the original reference should be cited.

It is also important to clearly explain and reference any procedures used to standardize or transform the data prior to subjecting it to multivariate analyses. In some cases it will be possible to cite a reference for the use of such procedures, but in the case of a new technique, the procedure should be explained and the underlying assumptions given.

Finally, recent issues of the journal to which you wish to submit the article should serve as a general guide on how much data and background documentation to included in the paper. When in doubt, I think that the best guideline is to include that material that you feel is necessary to allow your audience to understand your study. Then the editorial process of refereed journals can assist you in refining what will ultimately be published. Some additional suggestions can be found in Sneath and Sokal [1973], the Fourth Edition of the Council of Biology Editors Style Manual [1978], and in the Guidelines for Manuscripts published as a supplement to Volume 60, No. 3, of the Journal of Mammalogy [1979].

# STATISTICAL PACKAGES AND COMPUTER PROGRAMS

with the help of David J. Schmidly and Mark D. Engstrom

Most large computer centers will have available one of the three major commercially vended statistical packages--BMD (and/or BMDP), SPSS, or SAS. These all have many of the multivariate routines other packages, e.g. OMNITAB and P-STAT, which have been widely distributed, as well as packages which have been developed at local centers which have less wide distribution, but may still be available at your computer center (e.g. MIDAS from the University of Michigan). NTSYS is a multivariate package of special interest to systematists. We give a partial list of the source of these packages and some of their characteristics in Table 2 at the end of this Appendix.

There are other sources of programs besides the packages mentioned above. Published multivariate programs have appeared in some of the texts on multivariate statistics and programming. These have usually been written in FORTRAN and can be adapted to your local computer with little trouble [e.g. programs in Cooley and Lohnes, 1971; Blackith and Reyment, 1971]. Books with such programs are indicated by a "P" in the right margin of entries in the Bibliography. Some books which do not have programs themselves, but have appendices or other guides to available programs (note the date of publication of the book as these lists quickly become out of date) are Gnanadesikan [1977], Harris [1975], Kim and Mueller [1978a,b], Kruskal and Wish [1978] and Press [1972]. There are also commercially vended programs for specific purposes; for example, the major confirmatory factor analysis programs are only available in this way (from National Educational Resources, Inc., P. O. Box A3650, Chicago, Illinois 60690). Check with others at your institution: someone, especially in the social sciences, may have already purchased them. Some examples are MULTIVARIANCE, COFAMM, and LISREL. Also check with your computer center about the availability of other potential programs of interest. Someone in psychology, sociology, economics, physical education or statistics (among others) may have already written or gotten hold of the program or programs that may be useful for your application.

Most computers come with some kind of statistical subroutine package. Frequently these are in the form of FORTRAN subroutines which can be assembled into useable programs with minimal additional programming. IMSL (International Mathematical Subroutine Library) is the major, vended, up-to-date FORTRAN and PL/I subroutine library available at most large ceneters. It includes almost all of the ingredients for assembling programs for multivariate methods (e.g. matrix operations such as addition, multiplication, inverse and eigenvalue-eigenvector determination).

Finally, if you are so inclined, there is probably no better way to find out what is going on in a method than to program it yourself. This is becoming easier all the time. For example, using PROC MATRIX in SAS, one can write short routines in matrix notation that will perform essentially all of the analyses described in this manual. APL (A Programming Language--mainly available on IBM computers) is an iteractive matrix- or array-oriented language, and if available at

your center, can be a p[illegible] quick way to program multivariate data analyses.

Once one has the manu[illegible]ecies of the programs, examples of output and whatever other [illegible]entation is available, then the fun begins. Programs use dif[illegible] conventions for input (e.g. some allow alphabetic group names, [illegible] do not), accept data in different formats (e.g. they may be [illegible]ted to 80 columns per line), and have limits as to the number [illegible]ables etc. These requirements should be thoroughly investigated [illegible] your data is entered on cards or typed into a machine. With [illegible] trouble, the data can always be reformatted in the computer. Output conventions are even more varied: a "classification functio[illegible] one package for discriminant analysis is called a "discriminant [illegible]" in another. Vectors of coefficients may be scaled [illegible]ently in the various outputs for the same procedure. Even the [illegible] packages may change their conventions from one release to another [illegible]. SPSS changed their scaling of coefficients in canonical [illegible]ates analysis in Release 8, the most recent release. This also [illegible] scaling of the various plots produced.

Check carefully with your local computer center, read the documentation carefully, run a small, clearly understood example (e.g. from one of the textbooks), and make sure you understand each and every item of the output that you are going to use, especially with respect to nomenclature and scaling. Check the answers for accuracy (see the section on "Errors and accuracy" in the INTRODUCTION) and try several routines if you have the patience. Report any errors to your center. All of the packages and programs have errors ("bugs"). Most of the important ones have been removed--but they still show up. Check carefully that you have up-to-date documentation appropriate for the release of the program or package available at your center.

When you have decided which multivariate procedures are most appropriate for analyzing your data, select packages having programs that follow these procedures and provide the needed output. It is important not to let the programs available in a particular package dictate or limit the analysis performed. Generally, no single statistical package will have all of the programs available which are needed to analyze the data set. Even when a package contains a program for a needed method, the output from that program may not contain all of the relevant results in a useful form. The options available for each procedure, as well as the results chosen from the output, vary widely from package to package. Consequently, most investigators use a consortium of several packages, choosing the most desirable programs in each package for their specific problem. As a general rule, if one of the three large multipurpose packages (BMD-BMDP, SPSS or SAS) is available, it can be used in conjunction with NTSYS (which is easy and inexpensive to obtain) to provide most of the multivariate techniques one might wish to use.

The usefulness of local [illegible]ariate expertise seems to vary from location to location, as does the amount of help available on specific programs. Many institutions have personnel trained to help researchers with their programming and computing problems. It is a

good idea to identify these people and know where and to whom to go with any problems concerning the use of statistical packages and programs. In addition, for the major packages, short courses or tutorials are offered to teach users how to become familiar with the fundamentals of the package, both statistically and mechanically (e.g., video-cassettes are available at some centers for auto-tutorials on SPSS).

University courses and/or persons with expertise concerning multivariate methods and specifically, their applications in biology, are often difficult to find. One of the best ways to improve your understanding of multivariate methodology is to attend one or more of the seminars, short-courses, or workshops which are currently offered at several places in the country. Most of these study sessions are user oriented. They concentrate on the basic mathematical and geometrical structures of the techniques and on providing a basis for understanding assumptions, limitations and common mistakes. Many of these conferences focus on understanding a variety of statistical techniques and the interpretation of the output from particular packaged programs. Participants normally need only to understand classical univariate statistical analysis, hypothesis testing, algebraic manipulations, and have had some contact with matrix algebra in order to find one of these conferences useful. See Table 1, below, for a list of some of these.

The following are a few brief comments on the major packages, and Table 2 summarizes some of their features.

BMD and BMDP--The Biomedical Computer Programs have been in use for over 15 years. BMDP (the latest version is BMDP79) supercedes previous BMD releases and most of the BMD series. Among the most desirable features of these packages are: 1) their streamlined program writeups--no need to read many pages of introductory material, 2) the graphical displays available in many programs, and 3) the complete input and output examples furnished in the manual as well as references and details of computational procedures and program abstracts. BMD and BMDP are written in FORTRAN which makes them useable on many computers.

SPSS--The Statistical Package for the Social Sciences is widely used and is currently installed at over 1500 sites in more than 60 countries. Among its most desirable features are its simple English language and research-oriented control syntax as well as its comprehensive and lucid documentation. The manual includes introductory statistical presentations, statistical references, and varied applications with full examples. Many users find the manual clear, informative and well documented, and easier to use than the BMD and BMDP manuals. SPSS is written in FORTRAN which makes it widely useable on many computers. There are small versions available for some mini-computers.

SAS--Among the Statistical Analysis System's main assets are its simple, free-format language and its complete library of mathematical, statistical, and trigonometric functions. The manual is simple and

Table 1

List of Some Short Courses on Multivariate Statistics or Computer Packages

| Course | Sponsoring Organization | Duration |
|---|---|---|
| Short Course on Multivariate Statistics | Department of Mathematics and Statistics<br>University of Pittsburgh, Penn. 15260 | 3 days |
| Multivariate Research and the Statistical Package for the Social Sciences | University of Colorado<br>Boulder, Colorado | 5 days |
| SAS Statistics Course Data Analysis | SAS Institute, Inc.<br>Box 8000<br>Cary, North Carolina | 2 days |
| Multivariate Analysis | Institute for Professional Education<br>Suite 303<br>1515 North Court House Road<br>Arlington, Virginia 22201 | 5 days |
| Clustering and Numerical Classification | Control Data Corporation | 2 days |
| BMDP-79 Summer Tutorial | Tutorial Registration<br>Health Sciences Computing Facility<br>University of California<br>Los Angeles, Calfifornia 90024 | 2 days |
| SAS | Sas Institute, Inc. | 2 days |

easy to understand and contains excellent examples of program input-output (the current version is SAS79). The manual contains little explanation of statistical methodology but numerous references to this type of information are given following the explanation of each procedure. SAS is not as widely used as are the SPSS and BMD-BMDP packages, but the number of investigators using SAS is increasing each year. One of the available options in the 1979 version of SAS is the BMDP procedure, which calls any one of the BMDP programs to analyze data in an SAS data set and prints the results. A convert procedure is also available which will convert BMDP and SPSS system files to SAS data sets. SAS is written in PLI and only runs on IBM 360/370 computers.

NTSYS--The main focus of the Numerical Taxonomy System of Multivariate Statistical Programs is on applications in the area of multivariate statistical analysis with emphasis in the field of numerical taxonomy. This is among the most popular and widely used packages in systematics. Compared to BMD-BMDP, SPSS and SAS, it has the advantage of having a number of related programs associated with numerical taxonomy. The cluster analysis routines have been widely used because of the variety of similarity coefficients and clustering algorithms available.

Table 2

| Statistical Procedures Available | BMDP | SPSS | SAS | NT-SYS |
|---|---|---|---|---|
| Cluster Analysis | X | | X | X |
| Dicriminant Analysis and Canonical Variates | X | X | X | |
| Canonical Correlation | X | X | X | |
| Multiple Regression | X | X | X | |
| Principal Components | X | X | X | X |
| Factor Analysis | X | X | X | X |
| Nonmetrical Multidimensional Scaling | | | X | X |
| Loglinear Model and Logistic Regression | X | | X | |

Addresses

| Package | Availability | Manual |
|---|---|---|
| BMD-BMDP | Health Science Computing Facility<br>University of California<br>Los Angeles, California 90024 | University of California Press<br>2223 Fulton Street<br>Berkeley, California 94720 |
| SPSS | SPSS, Inc.<br>Suite 3300<br>444 North Michigan Avenue<br>Chicago, Illinois 60613 | McGraw-Hill<br>Publishers |
| SAS | SAS Institute, Inc.<br>P. O. Box 10066<br>Raleigh, North Carolina 27605 | SAS Institute, Inc.<br>P. O. Box 10066<br>Raleigh, North Carolina 27605 |
| NTSYS | Dr. F. James Rohlf<br>Department of Ecology and Evolution<br>The State University of New York<br>Stony Brook, New York 11794 | Copy provided with program tapes |