Perspectives on the Application of Multivariate Statistics to Taxonomy

F. James Rohlf

Taxon, Vol. 20, No. 1. (Feb., 1971), pp. 85-90.

TAXON 20(1): 85–90. FEBRUARY 1971

PERSPECTIVES ON THE APPLICATION OF MULTIVARIATE STATISTICS TO TAXONOMY *

F. James Rohlf *

Summary

A brief outline is given of the principal types of multivariate statistical techniques which have found use in taxonomy. Techniques such as correlation, principal components, canonical correlation, and factor analyses are described for problems dealing with analysis of covariation within a single sample. Techniques such as canonical variate, cluster, multidimensional scaling, and network analyses are described for dealing with analyses of among sample variation. The purpose of this account is to give an intuitive understanding of what the various techniques have to offer to research in taxonomy.

Introduction

Since taxonomy is concerned with the classification of organisms based upon relationships (both cladistic and phenetic) inferred from characteristics of the *whole* organism, statistical analysis in this field must take into consideration the simultaneous covariation of many characters of the organism as possible. Thus taxonomy differs in an important way from fields such as physiology or biochemistry where investigations often are concerned with the effect of a certain combination of treatments upon a single variable of particular interest. In taxonomy there is often no special interest in the particular characters used. They are a means to an end, needed in order to compare samples of organisms taken from different localities or from what are believed to be different taxa. For these reasons, the techniques of multivariate statistics are of particular importance in taxonomy.

In the account given below I have outlined a variety of techniques which have found use in taxonomy. The account is purposely nonmathematical. Its intention is to give one a general intuitive feeling for the types of questions which can be answered using presently available multivariate techniques and to introduce some of the jargon of the field so that one can communicate the type of analysis desired to someone who can arrange for the actual computations to be performed (since most of the analysis require an enormous amount of arithmetic, the actual numerical computations will almost always have to be done on a highspeed digital computer).

Several texts are available dealing with the applications of multivariate statistics, e.g., Morrison (1967), Seal (1964), Cooley and Lohnes (1962), and Rao (1952). While these texts all have brief introductions to matrix algebra, the books Searle (1966) and Graybill (1969) should be consulted for a more complete understanding.

The account given below is divided into two main sections. The first one discusses techniques which analyze patterns of covariation found within a single sample and the second is concerned with analysis of variation between samples.

* Contribution number 29 from the Program in Ecology and Evolution, State University of New York at Stony Brook.

** Department of Ecology and Evolution, State University of New York at Stony Brook, Stony Brook, New York 11790.

Description of within sample variation

There are a number of ways in which the patterns of variation and covariation within a single sample can be described. If 3 or fewer characters have been used, then frequency distributions and scatter diagrams can give one a useful intuitive appreciation of the variation found. If many characters have been used, then one can still plot scatter diagrams for various combinations of the characters taken 2 or 3 at a time, but it is usually difficult to fully appreciate complex patterns of covariation. A conventional statistical description of the sample would involve the computation of the mean for each variable and the variance-covariance matrix (a symmetrical table containing the variances of each character down the main diagonal and covariances in the off diagonal cells). If the sample represents a random sample from a multivariate normal distribution then such statistics contain sufficient information for estimating various properties of the population from which the sample was drawn. If the population was not normally distributed, then other statistics must be computed. In univariate statistics, one can compute higher moments such as g_1 and g_2 to measure skewness and kurtosis (Sokal and Rohlf, 1969). The analogous matrices in multivariate statistics are difficult to interpret. For this reason most workers resort to graphical techniques in such situations.

It is difficult to test for goodness of fit of an observed sample to a multivariate normal distribution. One can test whether each character taken separately fits a univariate normal distribution. If even one character does not fit a normal distribution, then the entire suite of characters does not fit a multivariate normal distribution. However, even if they all fit it does not guarantee that the entire suite of characters is consistant with a multivariate normal, since there can be a variety of complex interactions among the characters. If one has very large samples, the p-dimensional space can be partitioned into a series of regions and compare the frequency of observations in each region with that which would be expected based on a multivariate normal distribution (using the sample means and covariances). With samples of the size usually employed in taxonomy, this is not practical unless only a very few characters are used. The only alternative is to perform some sort of multidimensional scaling analysis (ordination) which will enable one to reduce the dimensionality of the system which needs to be considered. That is, to construct a few axes which contain most of the information about the covariation among the observations found in the original characters. If 3 or fewer axes are sufficient, then one can examine scatter diagrams constructed by projecting the specimens onto these axes and then plotting them against one another. Techniques such as non-metric multidimensional scaling (Kruskal, 1964 and Rohlf, 1970) and principal components analysis (Seal, 1964; Rohlf, 1970; Jolicoeur and Mosiman, 1960) have been used in taxonomy.

If one is satisfied that the data are consistant with the multivariate normal distribution, then principal components analysis can serve as a particularly compact means to describe the variation found in ones sample. The first principal component indicates the direction in hyperspace in which the observations differ most (the relative magnitude of the first eigenvalue indicates the extent to which the observations vary in this direction. This direction often corresponds to variation in overall size of the specimens (but it can sometimes represent the directions in which polymorphs vary if the sample is heterogeneous). The other principal components are often more difficult to interpret but they usually correspond to various shape differences between specimens. These differences are usually expressed as contrasts (high positive coefficients for some characters and high negative coefficients for other characters). The particular contrasts which result from the analysis are a consequence of the structure of the correlations between the characters. The information given by a principal components analysis can also be used to construct equal frequency ellipses which enclose regions expected to enclose $(1-\alpha)$ 100% of the observations. An example of this construction for the 2-dimensional case is given in Sokal and Rohlf (1969).

If a major purpose of the analysis is the investigation of patterns of intercorrelations among the characters, it is often useful to perform a factor analysis with rotation to simple structure (Harmon, 1967). This type of analysis expresses basically the same information but displays the correlation structure present in a much simpler form. Here each axis (or factor) corresponds to a group of characters as indicated by high (in absolute value) correlations between each factor and a set of characters. Characters not belonging to a set should have correlations near zero. Examples of the use of factor analysis in systematics are: Rohlf and Sokal (1958), Gould and Garwood (1969). Some other examples are listed in Seal (1964).

When the suite of characters can be logically divided into two sets and the relationships (if any) between the two sets is of interest, one can employ canonical correlation analysis. This technique obtains that a linear combination of the characters from each set of characters is such that these two linear functions have the highest possible correlation. This largest correlation is called the canonical correlation and measures the extent to which relationships in one set of characters can be predicted by a knowledge of the other set of characters. For example, one could use this type of analysis to locate those features in the adult stage which can be predicted based upon a knowledge of the larval stage. Morrison (1967) gives an outline of the necessary computations.

Description of variation among samples

There are several approaches to the study of variation among samples. The "proper" approach depends upon the statistical model and the purposes of the analysis (i.e., the questions being asked).

The question most commonly asked is: "Are the samples homogeneous?" If each sample can be assumed to have been drawn from a multivariate normal distribution then we can use a generalization of Bartlet's test to test for homogeneity of the variance-covariance matrices (Seal, 1964; Reyment, 1969). If they are homogeneous then we can use the techniques of multivariate analysis of variance to test whether the means of the samples are significantly heterogeneous. Of course, we must remember that if the samples were drawn from different geographic regions of a species or from different species, then we *know* that the true means (and probably also the variances and covariances) are different in different statistical populations. What we are testing is whether we have sufficient evidence to demonstrate that such differences exist and to set confidence intervals on the magnitude of the differences. If the test of significance yields a significant result, then it usually will be of interest to isolate those characters whose differences between various samples were most important in contributing to the overall significance test (just as we would turn to either *a priori* or to multiple comparison tests in a similar situation in univariate anova). However, it is difficult to know how to fully break down the overall multivariate test in the most meaningful way. If one has designed the sampling so that one can test a variety of *a priori* hypotheses, then one is relatively well off. One can then partition the overall test into a series of tests reflecting differences due to time of year vs. locality vs. sex vs. food plant, etc. If, however, one simply has those samples which are available one must use some kind of *a posteriori* test. Several multivariate multiple comparisons tests have been devised. Gabriel and Sokal (1969) described a test which can be used for this purpose, but it has the disadvantage that it produces "too many" answers. The voluminous output of this procedure reflects the fact that there are a very large number of ways in which multivariate samples can differ from one another. The results of this type of test are usually expressed in terms of so-called maximum nonsignificant subsets. These sets have the property that the addition of any other sample or variable to the set would cause it to be significantly heterogeneous.

The description of the patterns of variation among samples in terms of sets (which may be partially overlapping) is usually not very convenient. Other techniques (which have less statistical rigor) have been developed to more conveniently express statistical relationships among the variables over the samples.

There are three main classes of techniques which are used to reveal the relationships among the samples in the p-dimensional space: multidimensional scaling, cluster analysis, and network analysis. These techniques all come under the heading of multivariate data analysis since their main purpose is to give insight into ones data and to place less stress on tests of significance. In biology these techniques are associated with the field of numerical taxonomy where they have been found very useful in elucidating taxonomic relationships.

Multidimensional scaling is used when one wishes to express relationships among the sample means in terms of their coordinates on a few specially constructed coordinate axes. The goal is to preserve as much of the information about interpoint distances as possible while reducing the number of variables to be considered from p down to k (where k is 1, 2, 3, or perhaps 4 at the most). If the variation within all of the samples is homogeneous (or at least the orientation of the scatter ellipsoids are similar) then one can validly compute a pooled within group variance-covariance matrix and then perform a canonical variates analysis (see Jolicoeur, 1959; Seal, 1964). In this type of analysis relationships among the samples are expressed relative to the average covariation found within the samples.

This type of analysis is also sometimes called a generalized discriminant analysis since in the special case where there are only two samples the canonical variable is the discriminant function. When there are more than two samples, the canonical variables constitute a set of linear combinations of the variables which best discriminate between the groups. They can be used to form a probabilistic identification scheme (Cooley and Lohnes, 1962).

If the within group variation is not homogeneous (particularly if the orientation of the scatter ellipsoids differ) then it is difficult to make use of the within sample information and one must base ones analyses on the among sample variation. For example, one can perform a principal components analysis on the among sample correlation matrix to obtain vectors which indicate the major trends of variation among groups. One can then project the standardized sample means onto these axes in order to be able to prepare a scatter diagram depicting the among group variation relative to the total amount of variation found among the samples (since the correlation matrix and standardized data were used).

When the samples correspond to higher taxa then one expects the within sample covariation to be heterogeneous. This is one reason why there is seldom any attempt to take within sample covariation into account in numerical taxonomy. Often the taxa being sampled are sufficiently distinct that only a few specimens are used to represent each taxon. This is a valid shortcut whenever the among sample variation is much larger than the within sample variation as would be expected, for example, when the samples correspond to different species sampled throughout a family. In such cases there seems little point in worrying about tests of significance — the species are obviously different from one another. What is uncertain is their relative degrees of overall similarity and the way in which this can be most simply expressed. Another alternative (which sometimes is capable of expressing the relationships in fewer dimensions) is non-metric multidimensional scaling (see Kruskal, 1964; Rohlf, 1970). If a sufficient amount of the among group variation can be expressed in k-dimensions ($k << p$) then one can visually look for patterns in the differences among the samples (results are mostly intuitive, few tests of significance are possible here, but one often gains considerable insight into ones data).

Cluster analysis sorts the samples into a series of sets. These sets may be mutually exclusive, hierarchic, or partially overlapping in various ways. Hierarchical clustering schemes have been used most commonly in taxonomy. Typically these techniques begin with a matrix of distances between sample means (computed in various ways) and a search for other points which are relatively close together and separated by gaps from other such groups. The distance coefficient can be computed in such a way as to take the within sample covariation into consideration if this is desired. The generalized distance D is one way in which this can be done (Rohlf, 1970; Seal, 1964; Rao, 1952). For data in which the relationships among the samples are hierarchic cluster analyses works rather efficiently. They tend to be somewhat unsatisfactory on data in which the distribution of points in the p-dimensional space form very elongated clusters or where there are many points which are intermediate between clusters. These techniques also do not reveal the fact that some clusters may be in between other clusters (see Rohlf, 1970 for a general discussion).

Network analyses express relationships in terms of a graph (in the sense of graph theory, Ore, 1963) which consists of vertices (corresponding to the samples) and edges (which are connections between vertices). The existence of an edge implies that the two vertices so connected share some relation between them (e.g., they are nearest neighbors in the p-dimensional space). The shortest simply connected network has been found to be useful in numerical taxonomy since it indicates in a convenient fashion the closest neighbor of each point. Kruskal (1956) and Prim (1957) give algorithms for constructing such networks. Jardine and Sibson (1968) have suggested the use of networks which are more than simply connected and thus more capable of summarizing multivariate relationships (and hence more complex to understand). An example of the use of a shortest connection network in taxonomy is given in Rohlf (1970).

Comprehension of multivariate relationships is difficult. This difficulty is not helped by the fact that classical multivariate statistical techniques tend to result in a single number which is used for tests of significance. Such statistics are often difficult to interpret in terms of the particular samples and variables under investigation. For this reason more emphasis has been placed in the last few years upon a variety of graphical techniques which allow one to visualize many parameters of the sample simultaneously. The account given above is an attempt to give one a brief intuitive introduction to the types of techniques which are apt to be found useful in taxonomy. A number of workers are attempting to develop new mathematical tools which will allow a simple but efficient graphical summarization of multivariate relationships. Such developments, if successful, could have a large impact upon taxonomic methodology.

References

Cooley, W. W. and P. R. Lohnes 1962 – Multivariate procedures for the behavioral sciences. Wiley: New York 211 pp.

Gabriel, K. R. and R. R. Sokal 1969 – A new statistical approach to geographic variation analysis. Systematic Zool. 18: 259–278.

Graybill, F. A. 1969 – Introduction to matrices with applications in Statistics. Wadsworth: Belmont, Calif. 372 pp.

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: 276–300.

Harmon, H. H. 1967 – Modern factor analysis. Chicago, 470 pp.

Jardine, N. and R. Sibson 1968 – The construction of hierarchic and non-hierarchic classifications. Computer Jour. 11: 177–184.

Jolicoeur, P. 1959 – Multivariate geographical variation in the wolf, *Canis Lupus* L. Evolution, 13: 283–299.

Jolicoeur, P. and J. E. Mosimann 1960 – Size and shape variation in the painted turtle. A principal component analysis. Growth, 24: 339–354.

Kruskal, J. B. 1956 – On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Amer. Math. Soc., 7: 48–50.

Kruskal, J. B. 1964 – Non-metric multidimensional scaling. Psychometrica, 29: 1–27.

Morrison, D. F. 1967 – Multivariate statistical methods. McGraw-Hill: New York, 338 pp.

Ore, O. 1963 – Graphs and their uses. Random House: New York, 131 pp.

Prim, R. C. 1957 – Shortest connection networks and some generalizations. Bell System Technical Jour. 36: 1389–1401.

Rao, C. R. 1952 – Advanced statistical methods in biometrical research. Wiley: New York, 390 pp.

Reyment, R. A. 1969 – Biometrical techniques in systematics. In Systematic Biology. Publ. 1692 National Academy of Sciences. pp. 541–594.

Rohlf, F. J. 1970 – Adaptive hierarchical clustering schemes. Systematic Zool., 19: 58–82.

Rohlf, F. J. and R. R. Sokal 1958 – The description of taxonomic relationships by factor analysis. Systematic Zool., 11: 1–16.

Seal, H. 1964 – Multivariate statistical analysis for biologists. Wiley: New York, 207 pp.

Searle, S. R. 1966 – Matrix algebra for the biological sciences. Wiley: New York, 296 pp.

Sokal, R. R. and F. J. Rohlf 1969 – Biometry. Freeman: San Francisco, 776 pp.

Perspectives on the Application of Multivariate Statistics to Taxonomy
F. James Rohlf
Taxon, Vol. 20, No. 1. (Feb., 1971), pp. 85-90.
Stable URL:
http://links.jstor.org/sici?sici=0040-0262%28197102%2920%3A1%3C85%3APOTAOM%3E2.0.CO%3B2-E

References

Levels of Integration in Mammalian Dentitions: An Analysis of Correlations in Nesophontes micrus (Insectivora) and Oryzomys couesi (Rodentia)
Stephen Jay Gould; Robert A. Garwood
Evolution, Vol. 23, No. 2. (Jun., 1969), pp. 276-300.
Stable URL:
http://links.jstor.org/sici?sici=0014-3820%28196906%2923%3A2%3C276%3ALOIIMD%3E2.0.CO%3B2-I

Multivariate Geographical Variation in the Wolf Canis lupus L.
Pierre Jolicoeur
Evolution, Vol. 13, No. 3. (Sep., 1959), pp. 283-299.
Stable URL:
http://links.jstor.org/sici?sici=0014-3820%28195909%2913%3A3%3C283%3AMGVITW%3E2.0.CO%3B2-W