Brit. J. Psychiat. (1975), 126, 237–40

Multivariate Analysis: the Need for Data, and other Problems

By B. S. EVERITT

Summary

Multivariate analyses are an aid to, *not* a substitute for critical thinking in the area of data analysis. Meaningful results can only be produced by these methods if careful consideration is given to questions of sample size, variable type, variable distribution etc., and accusations of subjectivity in interpretation can only be overcome by replication.

The computer revolution has produced many problems for statisticians, not least of which is the ease with which experimenters may access packages of programs for multivariate analysis, and so bypass a 'difficult' (by which is meant one who will not do simply as he is told) statistician. Of course there are many abuses of univariate statistical methods. Here, however, the abuses are not likely to lead to such seriously misleading results as in the multivariate case.

Perhaps a major cause of the continuing misuse of statistical methods is the insistence of many journal editors in psychology and related areas, on articles being laced with multivariate analyses, and on encouraging the pedantic use of significance levels, i.e. the inevitable $p < ---$, as if such inclusions lent an air of respectability to their journal which it might not otherwise have had. Research workers in these fields would be better encouraged to devote more time to an initial screening of their data using simple graphical techniques, to ensure that their data are at least approximately suitable for more complicated multivariate analyses.

1. *Introduction*

In 1972, a brief article of mine was published in this *Journal* dealing with the problems of using the techniques of cluster analysis in practice. Since then it has occurred to me more and more frequently that warnings are needed not only about the relatively new technique of cluster analysis but also about more commonly used multivariate methods such as principal components, factor analysis, and discriminant function analysis. Such a belief has been compounded by my dealings with psychiatrists, psychologists and others, who seem to have come to regard multivariate analysis in particular, and statistics in general, as a universal remedy for 'poor' data, and, in some cases, as a substitute for almost no data at all. The purpose of this note, then, is to point out the difficulties in using various techniques of multivariate analysis in practice, and to suggest various ways of overcoming these difficulties.

2. *Sample size or the need for observations*

Multivariate data consists of a set of measurements (say p in number), made on a number (say N) of individuals. In many applications of multivariate techniques, such as factor analysis, reported in the literature, little attention seems to have been paid to the size of the ratio of N to p required to make any such analyses worth while. Many studies have used multivariate methods where N is of the same order of p; few investigations can be quoted where N is much larger than p. Of course, sample size is related to the 'economics' of data collection, the 'difficulty' of obtaining data in particular areas of research, etc., and so cannot be considered solely from a statistical viewpoint. However, there are various minimum requirements for multivariate analysis which might be made on statistical grounds, and which if not met with in practice should suggest using other simpler methods of analysis. For example, many multi-

https://doi.org/10.1192/bjp.126.3.237 Published online by Cambridge University Press

variate techniques, such as principal components analysis and factor analysis, begin with the computation of a correlation matrix between variables. This correlation matrix is an estimate of the population correlation matrix, and in a similar manner to the simple correlation coefficient has a sampling distribution which indicates its accuracy as an estimate. This distribution is extremely complicated and is of little help in obtaining any ideas as to how large N must be in relation to p before the population correlation matrix is reasonably accurately determined. However, various 'Monte Carlo' simulations can be done, and this author has found, for example, that if data are sampled from a population with zero correlations between its $p = 20$ variables, and a sample size of $N = 25$ is considered, then the estimated correlation matrices may differ widely for different samples, and correlations as high as $0{\cdot}7$ and as low as $-0{\cdot}7$ may be recorded. When the sample size is increased to one hundred, correlations may still be found as high as $0{\cdot}35$ and as low as $-0{\cdot}35$. Obviously, then, when p is of the same order as N, the sample correlation matrix is a poor estimate of the population matrix, and subsequently a principal components or factor analysis performed on the sample matrix would be at best unclear, but at worst positively misleading.

Simulations lead to various conclusions about the size of N required in relation to p which seem to be roughly in agreement with other suggestions made with respect to this problem. For example, Gorsuch (1974) recommends an absolute minimum ratio of five individuals to every variable, but not less than one hundred individuals for any analysis. This is perhaps rather too low a ratio of N to p, but might well be acceptable in an investigatory study for which the purpose was to generate hypotheses. In general, however, simulations show that perhaps ten individuals for each variable may be the kind of ratio to aim for, although even this may be rather optimistic, since Finney (1974), expresses doubts that the minimal N should increase less rapidly than p^2, perhaps even p^3 and even worse, perhaps 2^p. Such figures show that perhaps the majority of factor analyses, principal components analyses and the like reported in journals are based upon far too few observations for the number of variables considered. Experimenters planning to use these techniques should either obtain larger sample sizes or apply the analyses to particular subsets of their variables which are thought to be the most interesting or important. In general the conclusions of any analysis performed on data for which $N < 5p$ should be viewed with a certain scepticism.

3. *Types of variables*

In many investigations in psychiatry and in psychology, the variables recorded for each subject may be a mixture of *qualitative* and *quantitative* measurements. For example, the presence or absence of certain symptoms may be recorded, along with measurements such as height and weight. In many studies no consideration is given to the types of variables used or to their distributions; instead *all* are inserted into the factor analysis or principal components or cluster analysis 'machine' in the belief that in this 'best of all possible worlds' the results will be meaningful and interpretable. Unfortunately this is not a very realistic assumption. For example, Maxwell (1972), shows how spurious correlations can arise (leading to spurious 'factors') when the variables are scored in such a way that the great majority of the sample are given, say a zero score, i.e. when the distributions are highly skew. Therefore, before attempting any analysis, distributions of individual variables should be carefully examined and some consideration should be given to making *separate* analyses for different types of variables.

4. *Detection of outliers*

The presence of 'outliers', that is observations widely different from the others in the sample, may have a serious affect on any multivariate analysis performed on the data. For example such observations may lead to poor estimates of correlations in the correlation matrix to be subjected to principal components or factor analysis. An example is given by Gnadesikan and Wilk, 1969. Some attempt should therefore

https://doi.org/10.1192/bjp.126.3.237 Published online by Cambridge University Press

be made to discover such observations prior to analysis. Various methods have been given for the detection of outliers, for example Kruskal (1960), and Anscombe (1960). In many cases a careful examination may show that outliers were produced as a result of mispunching or misrecording. In these cases the observations can be corrected before analysis. If, however, 'real' outliers are present, careful consideration must be given to what such a result infers, and to whether the observation should be removed before analysis. The *automatic* removal of 'real' outliers is not to be encouraged, because in certain situations such observations may have a high information content.

5. *Interpretation of results*

The interpretation of the results of a multivariate analysis is generally a difficult task, made more difficult in many cases because of problems discussed in previous sections such as inadequate sample size. Perhaps one of the most difficult areas is the interpretation of 'factors' produced from a principal components or factor analysis and the interpretation of 'clusters' produced by one of the numerous techniques of cluster analysis.

The subjective element in the naming of factors in well known and is perhaps one of the reasons that factor analysis was (perhaps still is) the black sheep of multivariate analysis. Many different methods of rotation of factors are available, and different methods may lead to different final factors. The experienced 'factor interpreter' can often come up with plausible explanations for 'unexpected' loadings, and even in cases where loadings appear that were not expected on the basis of hypothesis, he may reason, *a posteriori*, that the loadings 'make sense'. Horn (1967), demonstrates, by analysing some artificially generated data, that because of the high subjective element present in interpreting factors quite 'meaningful' factors may be derived from data containing no such structure. Also, as already mentioned, if N is small in relation to p one should be wary about concluding that 'high' factor loadings are defining 'real' and 'meaningful' factors.

The problems of the techniques of cluster analysis have been discussed elsewhere (Everitt, 1972; Everitt, 1974), and here I shall mention only briefly the problem of interpreting 'clusters'. The first problem is that different techniques will in general produce different clusters, and so it is advisable to apply several different methods to one's data and accept for interpretation only those clusters produced by all methods. Another technique for validating clusters is to assess their 'predictive' ability with respect to variables not used in producing them. For example, Paykel (1972), after performing a cluster analysis of a sample of depressed patients, assessed the predictive value of the clusters with respect to outcome on treatment with amitriptyline.

The main method of overcoming subjectivity of interpretation in multivariate analysis is by replication, that is by collecting further samples and repeating the analyses. The analysis of the first sample may possibly be used to generate hypotheses, which may then be tested on later samples. Of course, if one's sample is large enough such replication could be achieved by dividing the sample randomly into two and analysing each half separately. What should be avoided is using one sample to both generate and test the hypothesis. As remarked elsewhere (Williams and Dale, 1965), 'generation of the hypothesis cannot be used as its own evidence'.

References

Anscombe, F. J. (1960) Rejection of outliers. *Technometrics*, **2**, 123–46.

Everitt, B. S. (1972) Cluster analysis: a brief discussion of some of the problems. *Brit. J. Psychiatry*, **120**, 143–5.

—— (1974) *Cluster Analysis*. Heinemann Educational Books.

Finney, D. J. (1974) Problems, data and inference: The address of the President. *J. Roy. Stats. Soc.*, Series A, **137**, 1–22.

Gnanadesikan, R. & Wilk, M. B. (1969) Data analytic methods in multivariate statistical analysis. In *Multivariate Analysis* (ed. P. R. Krishnaian), pp. 593–638. New York: Academic Press.

Gorsuch, R. L. (1974) *Factor Analysis*. W. B. Saunders Co.

Horn, J. L. (1967) On subjectivity in factor analysis. *Educational and Psychological Measurement*, **27**, 811–20.

Kruskal, W. H. (1960) Some remarks on wild observations. *Technometrics*, **2**, 1–3.

https://doi.org/10.1192/bjp.126.3.237 Published online by Cambridge University Press

MAXWELL, A. E. (1972) Difficulties in a dimensional description of symptomatology. *Brit. J. Psychiat.*, **121**, 19–26.

PAYKEL, E. S. (1972) Depressive typologies and response to amitriptyline. *Brit. J. Psychiatry*, **120**, 147–56.

WILLIAMS, W. T. & DALE, M. B. (1965) Fundamental problems in numerical taxonomy. In *Advances in Botanical Research* (ed. R. D. Preston). London: Academic Press.

A synopsis of this paper was published in the November 1974 *Journal*.

B. S. Everitt, B.Sc., M.Sc., *Lecturer, Biometrics Units Institute of Psychiatry, De Crespigny Park, London, SE5 8AF*

(*Received 1 July 1974*)

https://doi.org/10.1192/bjp.126.3.237 Published online by Cambridge University Press