Brit. J. Psychiat. (1971), **119**, 121–7

Multivariate Statistical Methods and Classification Problems

By A. E. MAXWELL

Introduction

Multivariate statistical methods are being used increasingly in an effort to clarify problems in psychiatric classification. Numerous references to their application could be given, but it is sufficient here to mention recent publications by Kiloh and Garside (1963), Carney, Roth and Garside (1965) and by Kendell (1968) which have received especial attention (e.g. Eysenck, 1970), as they are concerned with the long-standing controversy about the classification of depressive illness. But there is some confusion, not least amongst the critics, about the particular roles which different multivariate techniques, notably factor analysis, can play in classification problems, and in this paper an attempt is made to clarify the situation. Almost as a trial of the reader's endurance this attempt necessitates some preliminary discussion of the psychometric concept of a 'unitary trait' and of the statistical rules involved in defining it. But after this problem has been disposed of attention is confined to a discussion of the main multivariate techniques in question—factor analysis, discriminant function and canonical variate analysis, and finally cluster analysis. The attitude adopted in the paper is admittedly purist, but when controversy arises, as in the case of the classification of depressive illnesses, it is well to be clear about the precise properties and purposes of the statistical models we employ if confusion is to be avoided.

Unitary Traits

The concept of a 'unitary trait' has played an important role in the measurement movement in psychology. At the quantitative level and in the multivariate context it had its origin in Spearman's postulation of a single general factor or 'unity of intellectual function', as he described it, running through all branches of intellectual activity; the implication being, as Galton had earlier suggested, that there existed a '*normal*' continuum along which people could be ordered with respect to mental capacity. Persistently, since Spearman's work at the beginning of the present century, psychologists have striven to construct scales to measure 'unitary traits', though not confining their attention to the cognitive field. Indeed 'scaling' in a wider context than that discussed here has become virtually a discipline in its own right; e.g. Torgerson, 1968; Kruskal, 1964; Phillips, 1966. In these endeavours and on the purely psychological side the emphasis is on selecting items which at face value 'measure the same thing' for inclusion in a scale. The degree to which this is achieved is then checked by a statistical examination of the matrix of intercorrelations of the items, which should approach closely to unit rank and thus be accounted for in terms of a single psychological factor.

Now apart from its unitary quality a desirable property of a scale is that it should discriminate well between people. To achieve good discrimination attention must be paid to the population on which the scale is to be used. If the items on a cognitive scale are scored '1' for *correct* and '0' for *incorrect*, or those on a scale containing psychiatric symptoms are scored '1' for *present* and '0' for *absent*, then it can be shown that the discriminative power of the scale is enhanced by choosing items which have an incidence of roughly 50 per cent in the population in question. In the case of scales for measuring cognitive traits the situation is relatively straightforward, as the items can be *deliberately* chosen to fit this population. But in a scale consisting of symptoms of a given illness the same freedom of choice is seldom present and the experimenter may feel obliged to include most of the items indicative of the illness and these will probably spread over the whole

https://doi.org/10.1192/bjp.119.549.121 Published online by Cambridge University Press

spectrum of symptom-incidence. In this case even if all the symptoms were measuring 'the same thing' the matrix of their intercorrelations in the population would not in general have unit rank. It would contain at least one additional factor—known as a difficulty factor—having loadings roughly proportional to the incidence of the symptoms. Difficulty factors can only be strictly avoided by using 50–50 items, though hopefully in other instances they will not account for too disturbing a percentage of the common variance.

Side by side with the desirability of using items which give good discrimination lies the need for the population sampled to be homogeneous with respect to the items in question. For suppose that the population were heterogeneous, then in an extreme case it might consist of two sub-populations in one of which all the items (*symptoms* say) were present while in the other they were absent. In random sampling from such a heterogeneous or multiple population each item would tend to have a 50 per cent incidence and the experimenter might be deceived. Clearly then the possibility of a multiple population must be kept in mind.

A rough method of checking that a population is homogeneous with respect to a set of dichotomously scored items is to arrange the items in order of incidence and the individuals in order of their total scores on the scale. The rectangular table showing the incidence of the items, when the population is homogeneous, will then tend to show two triangular displays with the unities in one triangle and the zeros in the other. But items (symptoms) are sometimes scored not dichotomously but on (say) *five* or *seven*-point subscales. In this situation we will expect the score on all subscales to be spread reasonably well over the points of the subscales when the population is homogeneous. Normality of spread would be ideal but is seldom achieved.

The scores obtained by each individual on the constructed scale or test are geneally the sum (or perhaps a weighted sum) of his scores on the separate items, or subscales, of which the main scale is comprised. In general these total scores for a sample from a homogeneous population will tend to be normally distributed especially if the items are fairly numerous. This well-known fact follows from the *Central Limit Theorem* which, roughly phrased, says that 'the sum of a large number of random variables will be normally distributed if the variables are themselves normally distributed, but will also strongly approach normality almost regardless of the distributions of the variables summed'. Nevertheless one must remember that the normality of a distribution of scale scores still depends on the population for which the scale was constructed. For instance a cognitive test constructed so as to give good discrimination with 11-year old children would give a positively skewed distribution if administered to 8-year olds and would give a bimodal distribution if administered to a sample consisting of 8-year-olds and 15-year-olds in about equal proportions, the bimodality in the latter case being caused by the existence in the sample of two groups clustering at different points on the same scale.

It has already been noted that in some cases of scale construction the experimenter is free to choose his items to fulfil given requirements, while in other cases he may feel obliged to include all relevant items. In the latter case his items, even after the removal of possible spurious factors, will only exceptionally provide a unitary scale. More frequently the matrix of item intercorrelations will have a rank greater than unity, and several factors may be necessary to account for the observed covariation. But this fact of itself does not imply that the population being considered is other than homogeneous with respect to the variables being measured (see Section 3 and 5). But what if the population is heterogeneous with respect to the variables and contains two or more sub-populations clustering at different foci in the variable space? Such a situation if undetected would prove detrimental to a correlational analysis and would lead to quite spurious results. Methods for searching for evidence of subpopulations in a set of data are currently receiving wide attention and are discussed below in the section on Cluster Analysis.

Factor Analysis

There is also confusion about the role which factorial techniques can play in elucidating problems of classification. This confusion arises largely from an inadequate understanding of the properties of the factor model and of the conditions under which it may reasonably be employed. Perhaps the most important thing to note is that the model is linear and additive and in particular that the factors estimated from an analysis of the covariances or correlations between a set of variables are weighted linear functions of these variables. The variables themselves, taken in pairs, are assumed to be

https://doi.org/10.1192/bjp.119.549.121 Published online by Cambridge University Press

bivariate normal or, in the case of binary variables, to have incidences approaching 50 per cent (the 20 per cent–80 per cent range is generally satisfactory). Fulfilment of these assumptions is the experimenter's guard against spurious or biased correlations.

If the observed variables are normally distributed then the factors obtained from them, being weighted sums of them, may confidently be expected also to be normally distributed. Indeed the factors as weighted sums will tend strongly to normality, even if the variables themselves are not normally distributed, in view of the *central limit theorem*. Consequently, if an experimenter claims to have found a bimodal distribution of factor scores his analysis is doubly suspect—in the first instance because of the central limit theorem and in the second instance because of the assumptions under which correlations are estimated reliably.

To round off the discussion it may be worth mentioning that there is one situation in which a bimodal distribution of factors scores could legitimately arise. Suppose that we have samples from two populations for which the variance-covariance matrices for a set of variables are homogeneous but the means of the variables in the two populations are different. The correlations calculated separately for the two samples could now legitimately be averaged taking account if necessary of differences in sample size and a single factor analysis of the pooled results performed. Factor scores derived from such an analysis, if expressed in raw score units, might well lead to a bimodal distribution of factor scores if the means of the variables in one population were for the most part greater than those in the other. However, meaningful results, if raw scores were employed in the factor estimation equations, would only be obtained if all the observed variables had a common metric.

Enough has now been said to dispose of the idea of a bimodal distribution of factor scores but one further point of confusion is still outstanding. It has already been noted that a useful though not a sufficient criterion for postulating that a set of items measure a single unitary trait (factor) is that the matrix of their intercorrelations is of unit rank. But suppose these items were comprised of the accepted symptoms of a given illness they might not all be measures of 'the same thing' so that several factors were involved. It might then be erroneously thought that the plurality of factors would provide a basis for classifying the individuals in the sample.

To examine the problem let us assume for simplicity that all the correlations in the matrix are positive (though this assumption is not essential to the argument). Then in general the first factor would have loadings all with positive signs. Now it is a property of the arithmetic procedure of factor analysis, or more precisely of unrestricted factor analysis (Jöreskog, 1967), that the signs of the loadings on the second factor will be roughly half positive and half negative. In other words the second factor will divide the variables into two contrasting groups.

A case in point is a factor analysis by Hamilton (1959) of a scale constructed to measure anxiety, which contained a number of 5-point subscales some of which measured 'psychic' aspects of anxiety while others measured 'somatic' aspects. The first factor was a general factor of anxiety on which all the scales had positive loadings. On the second factor the psychic subscales were contrasted with the somatic subscales by having loadings with opposite signs. Now, as already emphasized, factor scores estimated for either of these factors would have a distribution tending towards normality, so it could not be claimed that scores on the second factor would provide a basis for classifying the sample of patients into two clusters. The distribution of scores on this factor would simply show the vast majority of patients to be in the vicinity of the mean score for the factor which is roughly mid-way between the 'psychic' and the 'somatic' poles. Again someone might argue that if the factor axes were rotated to obtain orthogonal or oblique factors, one of which passed through the cluster of psychic subscales and the other through the cluster of somatic subscales these factors would form a basis for subdividing the sample into two distinct categories. Individuals with scores on one factor, it might be thought, would form a group of 'somatic anxiety states' and

https://doi.org/10.1192/bjp.119.549.121 Published online by Cambridge University Press

those on the other a group of 'psychic anxiety states'. But this is not so if the conditions for valid estimation of the correlations between variables are observed. Each patient will have a score on each factor and for each factor the majority of patients will again lie in the vicinity of the factor mean. In clinching the argument it is worth repeating (see also section 5) that if there *were* two distinct groups of patients in the sample the results of the analysis would *ipso facto* be suspect. Indeed it is an act of kindness of the gods that a proper use of factor analysis does not lead to a useful classification of patients. For if it did, an analysis which yielded p significant factors could lead to as many as 2^p classification categories and for even moderate values of p this would produce an overwhelming proliferation of categories.

To conclude this section it is necessary to mention an 'inverse' use of factor analysis sometimes referred to as Q-technique. In an ordinary factor analysis the tests, scales, items, etc., are treated as variables and the scores to be analysed are obtained from a sample of individuals. Then, as already seen, successive factors, by the pattern of the signs of their loadings, provide a method of subdividing the tests or items into contrasting subgroups. By a reversal of roles of tests and individuals in the factor process it is thought that a means is provided of dividing a sample of individuals into contrasting subgroups. In the new formulation the individuals are treated as variables and the population sampled is one of tests, items, etc. But with a population thus defined successive members of a sample would be unlikely to provide independent observations while the scope of the population, if it is to be reasonably homogeneous, is generally limited. In addition a satisfactory method of finding 'correlations' between people, except in the case where the variables are dichotomously scored, is hard to achieve without first reducing the scores on the variables to standardized form: even if this were done a change in sign of the scores on even one of the variables would alter the magnitudes of all the inter-person correlations. For these and other reasons, which would prove elaborate to describe in detail, the method has found few supporters. Even if the procedure could be satisfactorily carried through its results would be of limited value except possibly in situations in which individuals had loadings, after rotation, on one variable only. But such a result would be difficult to interpret except in the very rare situation in which intercorrelations between variables (tests) were negligible.

Discriminant Function Analysis

Discriminant function analysis was introduced by R. A. Fisher (1936) and provides a means of finding that weighted linear function of a set of variables which maximally discriminates between two populations. Later the idea was generalized by Rao (1948), and others, to enable more than two populations to be considered and in such situations the technique is generally referred to as *canonical variate analysis*. But it is important to appreciate that these forms of analysis do not provide tools for a direct attack on a classification problem, for they require existing classifications as their starting point. An example will illustrate how the procedures work. Suppose we have a random sample of retarded depressive patients and another of neurotic depressive patients. We could then obtain scores for the patients on a number of variables relevant to the illnesses, such as *anxiety*, *depression*, *irritability*, and so on. The algebra of the method then enables us to find that weighted function of the variables which gives maximal discrimination between the groups. Basic assumptions of the method are that the variables are normally distributed and linearily related in each population with equal variance-covariance matrices, but for discrimination to be possible at least some of the means must differ from one population to another. Here we may note that these requirements involve an inconsistency where dichotomously scored items are concerned, for if an item is to discriminate between two populations it must have a different incidence in each. Under these conditions the variance-covariance matrices of the items could not be the same in each population. Nevertheless discriminant function analysis works fairly well with dichotomously scored data if the samples are large and the item incidences are within the 20 per cent–80 per cent range, for within this range the variances of items do not differ a great deal.

The discriminant function obtained in our example could now be employed to calculate a composite score for each patient and subsequently we could plot score-distributions separately for the two samples on a single axis. If the resulting distributions did not overlap, or if they overlapped only slightly, we would conclude that these two diagnostic groups represented separate categories of patients. Furthermore the results would enable

https://doi.org/10.1192/bjp.119.549.121 Published online by Cambridge University Press

us to check on the accuracy of the original classification, for a patient diagnosed as a retarded depressive whose composite score fell squarely into the neurotic distribution might on reconsideration be found to have been misclassified. Again a patient whose composite score fell well away from both distributions might be deemed not to belong to either category. In assessing the techniques we therefore conclude that even if they do not provide tools for a direct attack on classification they are of considerable value for checking the efficiency of classification categories already in use, or even of rejecting them. For example if two distributions of composite scores were found to overlap to the point of virtual coincidence it would then be necessary to search for further variables to try and improve the discrimination. But if such variables could not be found it would be reasonable to conclude that the assumption of two categories in the first instance was mistaken. A useful discussion of the conclusions which may be drawn from a discriminant function analysis is given by Hope (1969) in his review of Kendell's monograph (Kendell, 1968).

Cluster Analysis

When individuals in a sample have scores on each of p variates each individual may be thought of as being situated at a point in a p-dimensional space, the co-ordinates of the point being his scores on the variables. If the scores on each variable have a unimodal distribution they will tend to cluster around a single focus in the space, the co-ordinates of the focus being the medians of the variables. This type of distribution of points is what is implied when we speak of a population being homogeneous with respect to a set of variables. In addition, if the variables taken in pairs have bivariate normal distributions the overall distribution of points is said to be multivariate normal and this is the condition which must reasonably be met before a correlational-type analysis of the data, such as a factor analysis, is embarked upon. When the variables are continuous, with good dispersion, and the sample size is relatively large the homogeneity of the population can be adequately tested by examining each variable in turn to ensure that it is at least unimodal if not normally distributed. But if the variables are dichotomous or are simply ratings on scales consisting of a small number of points then these straightforward checks may prove inadequate. Consequently, if given sets of scores on such variables for a sample of individuals roughly similar, say patients broadly classified as 'neurotic', it is in general difficult to say in advance whether or not they come from a homogeneous population with respect to these variables.

The possibility always exists that in the multidimensional variable-space the patients may cluster at not one, but several foci. Under these conditions a correlational analysis of the data taken as a whole could be very misleading. For example, consider the simplest case in which there are just two clusters and two variables and let us suppose that for each cluster separately the variables are uncorrelated. Despite this zero correlation, a non-zero value between the two variables would be found if the clusters were reasonably separate and the data for the two were taken as a whole. A correlation coefficient thus found would be quite spurious.

Methods of cluster analysis aim at detecting, on the basis of scores on a set of variables, the existence of multiple clustering in a population. The detection of such clusters is important quite apart from the demands of correlational techniques, and in the case of psychiatry, for example, could have implications for the prognosis, allocation and treatment of patients. Many possible methods have been suggested (e.g. Zubin, 1938; Genrelli, 1963; Friedman and Rubin, 1967), and most of them are non-parametric in the sense that they do not involve distribution theory. Typically they use as their starting point a matrix of indices of similarity between pairs of individuals, a favourite index being euclidean distance. The matrix is then scanned using a computer and the individuals sorted into groups which meet some predefined criterion. Full details of current procedures are beyond the scope of this paper and at any rate are still in a fluid state. Several problems too remain to be solved regarding in particular the best starting points to use in

https://doi.org/10.1192/bjp.119.549.121 Published online by Cambridge University Press

the initial scanning process and the optimal number of groups to be assumed.

Parametric methods of cluster analysis are also under consideration. They usually come under the title 'method of mixtures', as the models involved assume a mixture of distributions of specified form. The problem then is to estimate the parameters of the models. After the estimation has been done the methods may be used for clustering purposes by obtaining for each member of a sample the probability of his belonging to each distribution and by assigning him to the distribution for which his probability is highest. Useful work in this field is reported by Wolfe (1965, 1967) but is too mathematical to be described fully here. However, a program provided by Wolfe has been implemented by Mr. B. Everitt in our Biometrics Unit for use on the University of London CDG 6600 Computer. The program is found to work especially satisfactorily when it is supplied with fairly good initial estimates of the parameters and these can be obtained by the use of several non-parametric procedures developed and programmed by Everitt.

Summary and Discussion

In this paper the roles which several widely used multivariate techniques—factor analysis, canonical variate analysis and cluster analysis—can play in the problem of classification of individuals are appraised. Of these techniques cluster analysis is seen to be the only one which enables a direct attack on the problem to be made. Unfortunately, fully efficient methods have yet to be invented. Those at present available are known to work well in cases in which a heterogeneous population contains reasonably distinct clusters, and they provide a satisfactory means of identifying these.

Discriminant function and canonical variate analyses are seen not to be classificatory devices in their own right, as they require as their starting point an existing classification, but they are useful tools for checking on the validity of that classification. Factor analysis as ordinarily used, that is with tests, items, scale scores, etc., as variables, does provide means of separating the variables into contrasting subgroups. But it does not enable the individuals in the sample to be classified in any clear-cut and acceptable manner. This is so because the scores for the members of the sample on each of a set of factors (rotated or unrotated) are by expectation normally distributed so that the majority of the sample falls in the vicinity of the mean of each factor distribution, and all individuals have scores on all factors. For a given factor, rotated let us say to represent some readily recognizable feature or aspect of behaviour, the most we can say is that individuals lying in one tail of its distribution would tend to display this feature in a marked degree while those in the other would display it minimally. But this would not imply a *categorical* difference between the individuals: as a group they are normally distributed along a continuum where the factor is concerned. Finally, if the scores of the individuals on two factors (rotated or unrotated) are plotted using orthogonal axes, then when the factors are uncorrelated the contours of density of the points will be circles and when they are correlated the contours will be ellipses.

References

Carney, M. W. P., Roth, M., and Garside, R. F. (1965). 'The diagnosis of depressive syndromes and the prediction of E.C.T. response.' *Brit. J. Psychiat.*, **111**, 659–74.

Eysenck, H. J. (1970). 'The classification of depressive illness.' *Brit. J. Psychiat.*, **117**, 241–50.

Fisher, R. A. (1936). 'The use of multiple measurements in taxonomic problems.' *Ann. Eugen. (Lond.)*, **7**, 179–84.

Friedman, H. P., and Rubin, J. (1967). 'On some invariant criteria for grouping data.' *J. Amer. statist. Assoc.* **62**, 1159–78.

Genrelli, J. A. (1963). 'A method for detecting subgroups in a population and specifying their membership.' *J. Psychol.*, **55**, 457–68.

Hamilton, M. (1959). 'The assessment of anxiety states by rating.' *Brit. J. med. Psychol.*, **32**, 50–5.

Hope, K. (1969). 'Review of *The Classification of Depressive Illnesses*, by R. E. Kendell.' *Brit. J. Psychiat.*, **115**, 731–41.

Jöreskog, K. G. (1967). 'Some contributions to maximum likelihood factor analysis.' *Psychometrika*, **32**, 443–82.

Kendell, R. E. (1968). *The Classification of Depressive Illnesses*, Institute of Psychiatry Maudsley Monograph No. 18. London: Oxford University Press.

https://doi.org/10.1192/bjp.119.549.121 Published online by Cambridge University Press

KILOH, L. G., and GARSIDE, R. F. (1963). 'The independence of neurotic depression and endogenous depression.' *Brit. J. Psychiat.*, **109**, 451–63.

KRUSKAL, J. B. (1964). 'Nonmetric multidimensional scaling: A numerical method.' *Psychometrika*, **29**, 115–29.

PHILLIPS, J. P. N. (1966). 'On a certain type of partial higher-ordered metric scaling.' *Brit. J. math and stat. Psychol.*, **19**, 77–86.

RAO, C. R. (1948). 'The utilization of multiple measurements in problems of biological classification.' *J. Roy. statist. Soc. B.*, **10**, 159–93.

TORGERSON, W. S. (1968). 'Scaling.' In *International Encyclopedia of the Social Sciences*, **14**, 25–39. Macmillan Inc.

WOLFE, J. H. (1965). 'A computer program for the maximum likelihood analysis of types.' *Technical Bulletin* **65–15**. San Diego: Bureau of Naval Personnel.

—— (1967). NORMIX: computation methods for estimating the parameters of multivariate normal mixture of distributions. *Research Memorandum SRM* **68–2**. San Diego: Bureau of Naval Personnel.

ZUBIN, J. (1938). A technique for measuring likemindedness. *J. abn. soc. Psychol.*, **33**, 508–16.

A. E. Maxwell, M.A., M.Ed., Ph.D., *Professor of Psychological Statistics, Institute of Psychiatry (University of London), De Crespigny Park, London S.E.5*

(*Received 14 October 1970*)

https://doi.org/10.1192/bjp.119.549.121 Published online by Cambridge University Press