Brit. J. Psychiat. (1969), 115, 1069–79

The Complete Analysis of a Data Matrix

By K. HOPE

Introduction

In the biological and social sciences the research worker is frequently faced with a two-dimensional, entity by variable, matrix of numbers. The matrix may contain scores of persons on tests, judgments of persons or objects by persons, measurements of physical characteristics of insects, economic statistics of areas, party votes in electoral districts, demographic or epidemiological or ecological counts of areas, and so on.

It is customary for those who analyse such matrices to assign a different ontological status to the two sets of arrays. In psychology for example persons are regarded as primary, hard, brute-facts; while tests are shifting, insubstantial, unreliable processes of mensuration. In economics a region is defined by tangible geographical boundary-marks and has a reality not achieved by monetary quantities, which are subject to random or systematic errors of estimation. But whatever the relative ontological standing of the entities and the variables within the field of study the research worker must be aware that, within the restricted field of the analysis, he is dealing not with persons or insects or regions but with numbers. If a variable is unreliable, then any particular value of that variable is unreliable, and it is unreliable for the person, insect or region of which it is a measure. To put the point another way: when a practitioner or technician applies his techniques of measurement to one particular entity, whether he is measuring a man's intelligence or the clover-content of a field, he is implicitly treating the entity as more basic or more important than the measuring processes which he employs. But when a research worker applies techniques of measurement to many entities he is as often as not doing so with the purpose of eliciting a structure of relations. He uses the entities to cast light on the variables and the variables to cast light on the entities. In the practical case the entity is primary, in the theoretical case neither entity nor variable is primary.

Any method of analysis which gives the appearance of omitting one set of arrays is likely to be misleading. Factor analysis, a method very frequently employed, appears to omit the entities since the results are reported in the form of projections of variables onto factors. Some factor analysts who are aware of this defect have sought to compensate by advocating the factor analysis of persons, which is as if a cook who had forgotten to lay the table were to compensate by popping the knives and forks in the cooking pot. In a factor analysis the entities are employed to illustrate the variables just as much as the variables are employed to measure the entities. The analysis should therefore reveal a reciprocity between entities and variables. In order to demonstrate this reciprocity it is necessary as a first step to draw a careful distinction between, on the one hand, the preparation of the data for analysis, and on the other the analysis proper. The first purpose of this paper is to draw this distinction and to show how the method of preparation should be determined by the nature of the analysis, just as the preparation of meat varies according to the method by which it is to be cooked.

Variables

The most fruitful demonstration of reciprocity between entities and variables is that of Burt (1940) and Slater (1958), who have explored the algebraic relations between the analyses of rows and columns of any matrix. It is not always realized that this demonstration applies to any matrix and not only to the typical centred, or centred and normalized, matrix

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press

of psychology (see for example the first matrix in Hope, 1966).

Many research workers are uncertain what it is they are analysing when they "extract the factors of a correlation matrix". Sometimes they speak as if they were operating in a variable space, and sometimes as if they were operating in an entity space. But what is a variable space or an entity space? Both these terms can refer only to the space* which is given by the score matrix or by some transformation of it. It cannot be too often repeated that the analysis of a matrix of covariances or correlations is only a useful short-cut to the analysis of the score matrix. If the so-called analysis of persons (in psychology) does not bear a simple relationship to the so-called analysis of tests then at least one of the analyses must be wrong.

The confusions which arise in so many analyses can be sorted out by breaking down the analysis into its logically discrete parts and by a consideration of the way in which each step can be made to conform both to the nature of the data and to the nature of the other steps. An important preliminary, particularly for psychologists, is a decision to forget the usual method of calculating a product–moment correlation coefficient, because this method confounds two quite distinct steps in the procedure.

The first step is the preparation of the data, and it must be carried out with an eye both to the requirements of the field of study and to the nature of the subsequent analysis. The raw data matrix S in Table I contains the hypothetical scores of twelve entities on four variables. Since very many analyses are self-contained in the sense that the mensuration is not carried out with reference to any norms or origins arising from preceding work, it is a common practice to centre the data for variables, that is to express each score on a variable as a deviation from the mean of all scores on that variable. Matrix M (Table II) is matrix S centred for variables. It is frequently desirable to eliminate differences between the

TABLE I

Hypothetical raw scores of twelve entities (e) on four variables (v). Matrix S

	v_1	v_2	v_3	v_4
e_1	4	7	10	7
e_2	3	2	3	0
e_3	12	11	10	3
e_4	5	6	13	8
e_5	8	11	6	3
e_6	11	14	11	12
e_7	0	7	6	7
e_8	5	8	9	2
e_9	5	12	9	6
e_{10}	5	10	5	4
e_{11}	7	8	11	6
e_{12}	7	12	3	2
Means	6	9	8	5

TABLE II

Hypothetical scores centred for variables. Matrix M

	v_1	v_2	v_3	v_4
e_1	−2	−2	2	2
e_2	−3	−7	−5	−5
e_3	6	2	2	−2
e_4	−1	−3	5	3
e_5	2	2	−2	−2
e_6	5	5	3	7
e_7	−6	−2	−2	2
e_8	−1	−1	1	−3
e_9	−1	3	1	1
e_{10}	−1	1	−3	−1
e_{11}	1	−1	3	1
e_{12}	1	3	−5	−3
Means	0	0	0	0

* There has developed among psychologists a habit of referring to "the person-space" or "the test-space" of an investigation as if these were necessarily two different spaces. This use of the term space appears to be derived from set theory, in which the collectivity of all the elements which might enter into an analysis is known as the space of that analysis. In this sense space is equivalent to De Morgan's (1847) "universe of discourse". Thus the space of the astronomer is the set of all potentially observable heavenly bodies, and the space of the radio astronomer is similarly the set of all potentially-detectable radio-sources. But the space through which the astronaut travels is not a set or class of anything (unless it is the set of all possible relations between points) although it contains any number of spaces in the sense of universes of discourse. The astronaut's space differs from the extensional space of set theory in that it is charted rather than expressed in Venn diagrams, it has parts rather than subsets, and it is traversed rather than enumerated. When we speak of space in this article we are referring to the ordinary man's space rather than to the space of set theory.

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press

variances of variables because these differences are arbitrary and give an undue weight to variables which happen to have a high variance. This is achieved by normalizing each column of M, that is by dividing each element by the square root of the sum of squares of the column. For the sake of simplicity our example will not be normalized. In fact normalization is unnecessary because all four variables have the same variance.

The data have now been prepared and the analysis proper may begin at this point. The next step of the factor analyst is to extract factors, components, or latent vectors and by these terms he means vectors of variable loadings or saturations. In fact however a vector of variable loadings is only one possible mode of representation of a factor, the other mode being a vector of loadings of entities. These two sets of projections of components are shown in Tables III and IV. Algebraically they are equivalent ways of identifying the latent vector associated with each latent root λ. The equivalence can be simply expressed by the equation

$$M = P\Lambda^{-\frac{1}{2}}V'$$

where Λ is the diagonal matrix of latent roots (Table V).

TABLE III

Projections of variables onto components. Matrix V

	c_1	c_2	c_3	c_4
v_1	8	6	4	2
v_2	8	6	−4	−2
v_3	8	−6	4	−2
v_4	8	−6	−4	2
Sums of squares	256	144	64	16

TABLE IV

Projections of entities onto components. Matrix P

	c_1	c_2	c_3	c_4
e_1	0	−4	0	0
e_2	−10	0	2	2
e_3	4	4	4	0
e_4	2	−6	2	0
e_5	0	4	0	0
e_6	10	0	−2	2
e_7	−4	−4	−4	0
e_8	−2	0	2	−2
e_9	2	0	−2	−2
e_{10}	−2	2	−2	0
e_{11}	2	−2	2	0
e_{12}	−2	6	−2	0
Sums of squares	256	144	64	16

TABLE V

Diagonal matrix of latent roots. Matrix Λ

	c_1	c_2	c_3	c_4
c_1	256	0	0	0
c_2	0	144	0	0
c_3	0	0	64	0
c_4	0	0	0	16

There are two reasons why research workers choose to report V rather than P. The first is that V is almost invariably smaller than P, and the second is that the variables usually represent known or supposed dimensions, whereas no information about the entities other than that included in the analysis is available. Whether we take them separately or jointly these reasons are not sufficient grounds for supposing that we are analysing variables rather than entities, or that we are working in a variable space rather than an entity space. If we wished to give a less misleading name to the space we might choose a title from the field of inquiry, ee.g. the hostility space, or the foreign trade space. Alternatively we might refer to it as the component space since the components are the set of orthogonal axes to which both variables and entities are being referred. Although they are not the only possible set they are unique* in the sense that they are the set which satisfies the equations

$$(M'M - \lambda_i I)\, v_i = 0$$
$$(MM' - \lambda_i I)\, p_i = 0$$

where v_i is the i th column of V and p_i is the i th column of P.

Up to this point the analysis has followed a well-mapped path though a path which is not trodden as frequently as it ought to be. In a recent monograph Slater (1964) has demonstrated the advantages of a geographical

* Factor analysts often gloss over this uniqueness in their urge to justify rotation of factors.

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press

or spherical co-ordinate representation of an analysis. In this procedure the components are identified with a set of orthogonal diameters of a sphere or hypersphere which has as many dimensions as there are components. The three-component case can be represented on a geographer's globe. The diameter emerging at 0° longitude on the equator may conveniently be taken to represent the first component; the north–south polar line may be taken to represent the third component; and the only remaining diameter which is at right angles to these two will represent the second component. Each variable can be envisaged as a diameter of the globe which is defined by the angles which it makes with the three reference diameters. If all variables have the same variance then each variable-diameter can be given visual representation by plotting either of the two points at which it meets the surface of the globe. In order to do this we calculate the longitude and latitude of each variable-point. The pattern of the distribution of points summarizes the structure of the field of study. Clustering of variables for example can be detected at a glance. If the central point of each of several clusters is taken to be the end-point of a factor-diameter then the orthogonality or non-orthogonality of factors can be determined without recourse to the construction of rotation matrices on the basis of arbitrary assumptions. Although only three dimensions can be presented in any one plot it will be possible to explore any number of sets of dimensions kaleidoscopically when facilities become available for the instantaneous three-dimensional projection and rotation of co-ordinate systems.

In a review of Slater's monograph (Hope, 1966) the writer has presented a very simple arithmetic example of a spherical co-ordinate analysis, together with a plot of the results. The artificial construction of the present example produces a plot in which the four variables are the four corner points of an elongated quincunx whose centre point is at 0°, 0°. The relations between principal axes and spherical co-ordinates are described in detail in chapters four and five of *Methods of Multivariate Analysis* (Hope, 1969).

Both variables and entities may be treated as lines or as points, though for most purposes it is appropriate to think of a variable as a line and an entity as a point. If an entity is thought of as a point then the square root of the variance of that entity measures the distance of the point from the origin. The origin is of course the centre of the globe, which is the centre of gravity of all the points in the analysis. As an example, the distance of the first entity from the origin may be calculated from either M or P as

$$((-2)^2+(-2)^2+2^2+2^2)^{\frac{1}{2}} =$$
$$= (0^2+(-4)^2+0^2+0^2)^{\frac{1}{2}} = 4$$

In reviewing Slater's method the present writer pointed out that the polar co-ordinates of entities may be calculated just as readily as the polar co-ordinates of variables.* The calculations may be performed on the rows of matrix P as on the rows of matrix V. But a literal mapping of the entity vectors is impracticable because they terminate at various distances from the origin (distance is calculated as in the above example) and they do not therefore all lie on a common (hyper)spherical surface. Although we have succeeded in producing a visual presentation of the relations among the variables, a presentation which allows the structure of the data to manifest itself without arbitrary manipulation by the analyst, we appear to be no nearer achieving a similar plot of the relations between persons. However, the following section demonstrates that a suitably chosen technique of numerical taxonomy will supply the deficiency.

Entities

Numerous taxonomic techniques have been suggested and many of them are to be found in Sokal and Sneath's textbook (1963). The technique to be described here was developed independently but it bears obvious similarities to methods invented by other workers.

Numerical taxonomy like factor analysis has been bedevilled by a failure to distinguish

* An entity vector will be a radius rather than a diameter and it cannot meet the surface of the sphere at more than one point. Two persons whose radii make a straight line are, within the terms of the analysis, opposites and the degree to which they are opposite is measured by the distance between them.

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press

between the preparation of the data matrix and its taxonomic analysis. Possible methods of preparation have been described in preceding sections. Before relating method of preparation to method of analysis it is necessary to describe the latter. This can be done most simply by an actual analysis carried out on a small data matrix which has been subjected to only a crude method of preparation.

TABLE VI

Hypothetical ranks of six entities (A–F) on three variables

	v_1	v_2	v_3
A	1	1	5
B	3	4	4
C	4	3	2
D	2	2	6
E	5	6	3
F	6	5	1

Table VI shows the ranks of six entities on three variables. We may calculate a rough-and-ready coefficient of disagreement between any pair of entities by adding up the moduli of the differences between their ranks on all the variables. Thus the disagreement between B and C is

$$|3-4|+|4-3|+|4-2| = 4$$

This is quite a low value since the minimum possible disagreement in the case of ranked data (with no ties) is equal to the number of variables, which is three. We could express the obtained level of disagreement as a coefficient of agreement by subtracting it from the maximum possible disagreement which is, for ranked data,

$$v(e-1) = 3(6-1) = 15$$

where v is the number of variables and e is the number of entities.

The process of (*a*) expressing the data as ranks and (*b*) calculating coefficients of disagreement constitute a rough-and-ready method of preparing the data. Although the coefficient of disagreement is clearly not an ideal statistic it may provide a useful quick approximation to the statistic described in the next section. In spite of its defects it does preserve the multidimensionality of the data in that no difference between two entities on a particular variable can contribute a negative quantity to the overall disagreement between them.

The taxonomic method to be employed will be demonstrated on the matrix of coefficients of disagreement given in Table VII. We wish

TABLE VII

Coefficients of disagreement between entities

	B	C	D	E	F
A	6	8	3	11	13
B		4	5	5	7
C			7	5	5
D				10	12
E					4

to see how these entities cluster. The lowest coefficient is for entities A and D and so we group these two together and calculate a vector of mean ranks for the group, viz.

Group 1 1·5 1·5 5·5

We now raise our criterion of disagreement by a step of say one, that is from three to four. We calculate a coefficient of disagreement between each ungrouped person and the vector of Group 1. None of these coefficients is as low as the criterion so we then scan Table VII to see whether any two ungrouped persons have a disagreement as low as four. Two pairs have, and so we group together B+C and E+F and calculate the vectors.

Group 2 3·5 3·5 3·0
Group 3 5·5 5·5 2·0

We now raise the criterion of disagreement by another step, to five. We compare each group vector with each other group vector to see whether any groups amalgamate at this level of the criterion. Groups 2 and 3 satisfy the criterion so we amalgamate them and calculate the vector of mean rankings of B+C +E+F. This gives us a new Group 2.

Group 2 4·5 4·5 2·5

There are no ungrouped persons left so we omit the comparisons between ungrouped persons and groups and also the comparisons between ungrouped persons. We continue

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press

to raise the criterion by steps of unity until all persons have joined a single group. The analysis may be represented by a dendrogram (Fig. 1) which gives an immediately comprehensible visual display of the clustering of persons. The relations between clusters on the tests may be represented in an equally simple manner. If we decide for example that the grouping at a criterion of five is of interest we may draw profiles of the two groups which exist at this level (Fig. 2). Since this is a profile diagram and not a graph the order of the variables may be re-arranged in the interests of simplicity. The scatter of points round the mean of a variable (which is a constant for ranked data without ties) is a measure of the contribution of that variable to the separation of groups.

Groups formed in this way could be subjected to discriminant function analysis or canonical analysis of discriminance, but the research worker must be wary of contamination between the two analyses. It is not enough simply to use a different set of variables, one must also be reasonably confident that the two analyses do not share correlated errors of measurement. Ideally after performing a taxonomic analysis on a set of variables A, the research worker should obtain a fresh sample of entities, allot each to a cluster on the basis of its similarity to the mean vectors of the clusters in the initial analysis, and then perform a canonical analysis of discriminance on the set of variables B. The relations between A and B might be subjected to a canonical correlation analysis.

Variables and Entities

We are now in a position to relate the method of numerical taxonomy illustrated in the immediately preceding section to the methods described in earlier sections. We left our twelve entities as points in a multidimensional space whose axes are components. The obvious measure of disagreement between two entities is the distance between them. The easiest way to conceptualize this distance is to refer each entity to the components, that is to employ the entries in matrix P (Table IV). The distance between the first two entities is the square root of

$$(0-(-10))^2+(-4-0)^2+$$
$$+(0-2)^2+(0-2)^2 = 124$$

It is not necessary to refer entities to components because distances can be calculated either from matrix M (Table II) or from matrix S (Table I). To calculate a distance from either of these matrices we require the sum of squares of each entity and the sum of products. For example, from matrix M

$$SS_1 = 16$$
$$SS_2 = 108$$
$$SP = 0$$

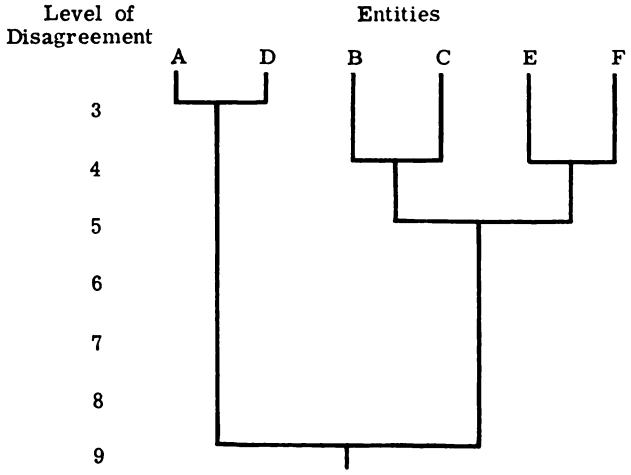


Fig. 1.—Dendrogram illustrating the similarities among six hypothetical entities.

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press

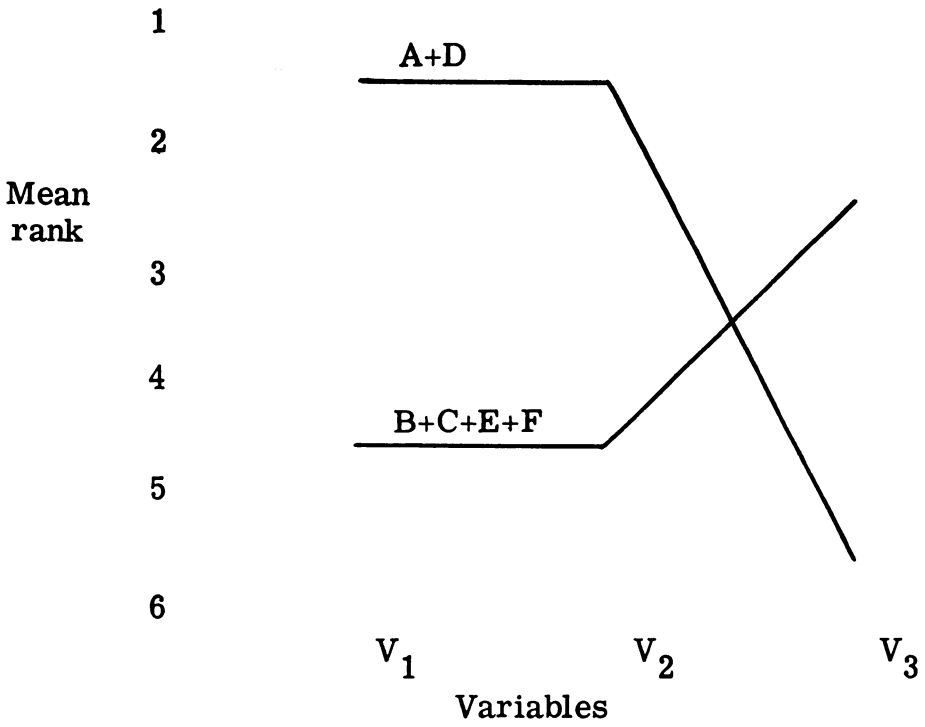


FIG. 2.—Profiles of two groups on three variables.

Distance is calculated as

$$(SS_1+SS_2-2SP)^{\frac{1}{2}} = (16+108-0)^{\frac{1}{2}} = \sqrt{124}$$

It is noteworthy that exactly the same formula applied to matrix S yields the same distance between persons one and two. The reason is that distance is invariant under change of origin. Centring S for variables simply means moving the origin to the centre of the cluster of persons. This in no way alters distances between persons. Distance is also invariant under rotation of axes. Although the analysis is multidimensional a distance between two points cannot be more than one-dimensional. Because distances are independent of the dimensionality of the space they can be calculated whether or not the covariance matrix for variables is nonsingular. In particular, they can be calculated even if there are more variables than entities. In some inquiries it will be desired to make the number of variables as large as possible* in order to ensure that all possible dimensions of the field have been explored.

The matrix of distances between pairs of entities replaces the matrix of rough-and-ready coefficients of disagreement as the starting-point of the taxonomic analysis. The analysis proceeds as before, with the means of taxonomic groups (whether on variables or on components) calculated in exactly the same way. At each level of the criterion the first stage of the analysis consists in the amalgamation of groups. Each group is compared with every other group in order to identify the two groups which have the least distance between them. These two groups are amalgamated provided that the distance between them is less than, or equal to, the criterion. Distances are recalculated to take account of this amalgamation. The lowest of the

* An undesirable consequence of this practice may be avoided by a modification which is described in the following section.

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press

new distances determines the next amalgamation. The process continues until no two groups are sufficiently close to one another to satisfy the criterion.

The second stage of the analysis consists in finding the lowest distance between any ungrouped entity and any group and allotting that entity to that group. Distances are recalculated to take account of this allotment. The lowest of the new distances determines the next allotment. The process continues until no entity is sufficiently close to any group to satisfy the criterion.

The third stage of the analysis consists in finding the lowest distance between any two ungrouped entities and grouping these two entities together. The process continues until no two ungrouped entities are sufficiently close together to satisfy the criterion.

The fourth stage of the analysis consists in a reiteration of the first stage, that is, the amalgamation of groups. This fourth stage is of only marginal value when the major groupings have become fairly well established. Sometimes, however, it does make an appreciable difference to the first steps of the analysis. There is no reason why the second and third stages should not be reiterated at each level of the criterion. But the benefits this would bring would be minimal. Any slight distortion occurring at a particular level of the criterion is automatically corrected at the next level.

If there are ties among the values of the distances then a slight distortion may result from an arbitrary decision to deal with one distance before another. The greater the number of variables, and the wider the range of scores on each variable, the less likely are there to be ties.

Figs. 3 and 4 contain the representations of the structure of the data matrix of the present example which emerge from the application of numerical taxonomy.

The taxonomic method described here is designed to identify clusters and to group them hierarchically though the hierarchies need not be symmetric. It tries to avoid wildly swinging vectors of group means and so it does not allow a member to drop out of a group when the vector for that group moves away from him. He remains there, as it were exercising his valence to bring it back towards him. If entities are distributed in, say, lines rather than clusters in the component space then the method of analysis described above would not be appropriate. But for very many

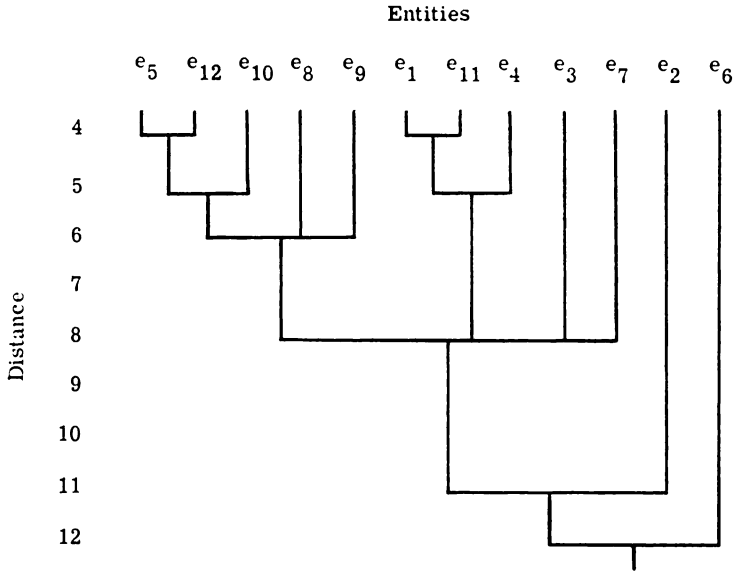


FIG. 3.—Dendrogram showing the clustering of twelve hypothetical entities.

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press

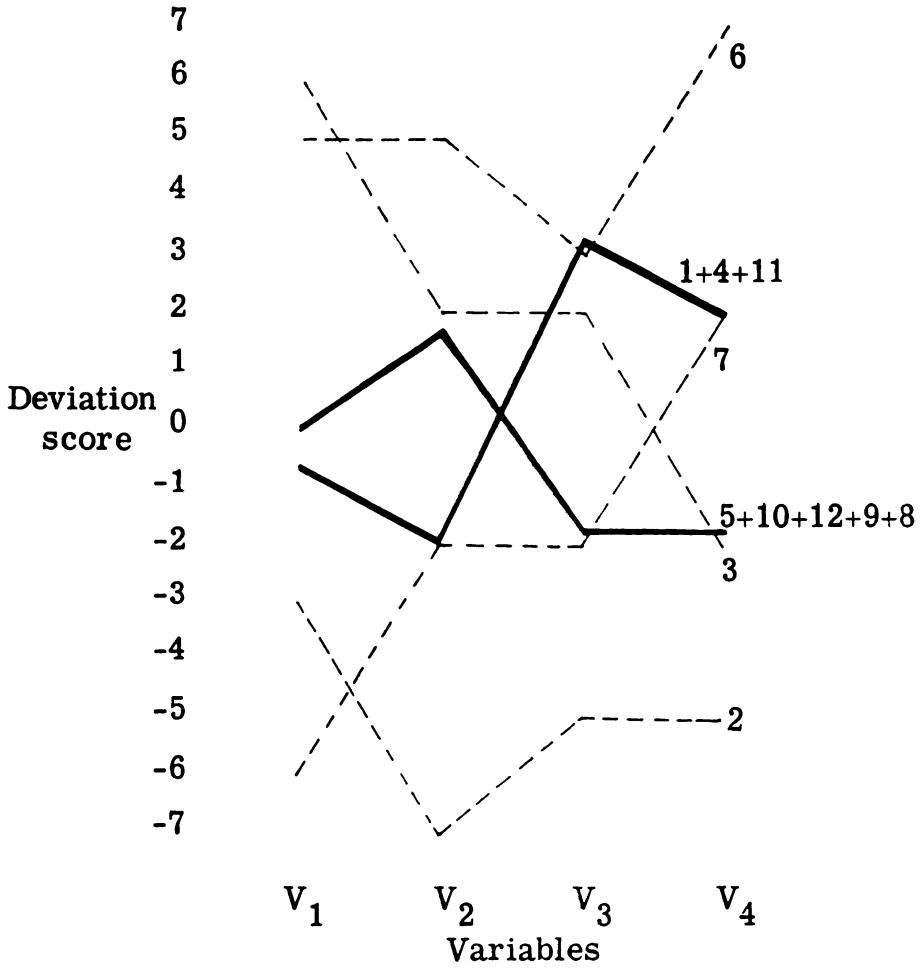


FIG. 4.—Profiles of two groups on four variables. Profiles of ungrouped entities are indicated by broken lines.

applications in the social and biological sciences it is in the writer's estimation a method with a high probability of success.

The taxonomic method and the method of polar or spherical co-ordinates between them yield visual displays of the relations between entities, the relations between variables, and the relations between entities and variables. If the methods of data preparation illustrated here have been employed then the research worker can be confident that each map or diagram is an aspect or projection of one and the same data matrix. The interpretation of the displays remains a matter of scientific judgment. Nevertheless, if he has prepared the data matrix in conformity with the requirements of the field of study the research worker knows that his interpretation is not subject to the distorting influence of artefacts introduced by the computing techniques employed.

REDUNDANCY

In the preceding section it was suggested that large numbers of variables might contribute to the taxonomic analysis. An objection to this recommendation is that the inclusion of a number of highly correlated variables will give undue weight to their common factor or factors.

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press

It may be that the research worker is as interested in a dimension measured by only two variables as he is in a dimension measured by fifteen. Various *ad hoc* procedures have been suggested to eliminate redundancy in the variables, but an obvious solution presents itself when taxonomic method is applied to a component space. In geometrical terms the research worker wishes to work in a (hyper)sphere in which every component, however tiny its latent root, has the same length as every other component. Such a space is simply achieved by dividing the columns of P by their sums of squares, which are latent roots, before calculating distances. The effect is to squash the ellipsoidal space into a spherical space, giving an equal weight to all independent dimensions in the calculation of distances between entities. Two formulae will illustrate the algebraic identities which ensue (cf. Hope, 1969). In these formulae subscript n indicates that a matrix has been normalized by columns. The weighting matrix by which we arrive at P_n is a modification of V, viz.

$$P_n = MV\Lambda^{-1}$$

By analogy with the first formula of this paper we may calculate

$$M_{new} = P_n V_n'$$

and this matrix (multiplied by a scalar of 48 in order to avoid fractions) is given in Table VIII. It can be seen that the correlation of every pair of variables in this matrix is zero. This is the effect we were aiming at—the elimination of redundancy. Although the taxonomic analysis can be carried out on M_{new} it may in some cases be less misleading to apply it to P_n since our vectors of mean scores of groups will not be comparable with similar vectors derived from a taxonomic analysis of M.

TABLE VIII

Matrix M_{new} multiplied by the scalar 48

	v_1	v_2	v_3	v_4
e_1	−8	−8	8	8
e_2	3	−33	−21	−9
e_3	26	2	10	−14
e_4	−3	−15	21	9
e_5	8	8	−8	−8
e_6	21	9	−3	33
e_7	−26	−2	−10	14
e_8	−9	3	15	−21
e_9	−15	21	9	−3
e_{10}	−5	7	−13	−1
e_{11}	5	−7	13	1
e_{12}	3	15	−21	−9

An obvious objection to the practice of giving equal weight to every component is the likelihood that some components are error components, or are measuring irrelevant dimensions. The research worker may be unable or unwilling to decide which components should be excluded from the analysis and he may demand a mechanical weighting procedure. One possible procedure would be to apply Bartlett's (1950, 1951) test for the heterogeneity of latent roots and to omit roots which do not display heterogeneity. This test is not, however, really appropriate to our problem. The hyperspace may be perfectly spherical and yield no heterogeneity even though no error is present. Sphericity and dimensionality are two quite distinct concepts.

A solution might be found by forcing the analyst to define what he means by relevant dimensions, or rather by asking him to indicate a system of relations which contains all the relevant, and none of the irrelevant, aspects of his data. He might, for example, take two independent observations of the value of every variable in every entity, and he might specify that only variance common to the two sets of observations was to be employed in the definition of the space of the complete analysis.

Let us suppose that observer X and observer Y have each measured the same set of variables in the sample of entities. Then we may calculate the multivariate regression of X's measures on Y's measures and delete from X's covariance matrix that part which cannot be accounted for by regression. (We may also choose to delete that part which is not associated with a significant canonical root.) An analogous adjustment is made to Y's covariance matrix. The latent roots and vectors of the two adjusted matrices may be extracted. (Each set of normalized vectors will be transformable into the set of normalized vectors associated with the canonical roots.) We may calculate the position of each entity on each set of dimensions and treat these

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press

scores as the basic data of the complete analysis. Although the dimensions will be orthogonal within sets they will be correlated between sets. The components of the conjoined sets of dimensions X and Y will not, in general, be identical with the components of either of the two sets taken individually. If we wish to weight a dimension according to the extent to which it is predictable* from the alternative set of dimensions then we make the sum of squares of the scores on each dimension equal to, or proportional to, the latent root of that dimension.† If on the other hand we decide to give equal weight to all dimensions of the conjoined set of dimensions then we extract the components of the conjoined set, calculate columns of component scores, and normalize each column.

In some studies these complications may be unnecessary. It may suffice to average the measurements of the two (or more) observers, perhaps after centring and standardizing each variable for each observer. This simple procedure will be satisfactory if it can be assumed that the two observers are equally competent in their measurement of each variable. It has the advantage of not introducing the element of capitalization on chance associations which occurs in the application of regression methods. It cannot be employed if there has been no replication of measurements. The regression method may be applicable where there has been no replication provided that the research worker is willing to specify the two sets of variables whose common dimensions he wishes to employ.

* The suggested technique is in fact a generalization of Guttman's image analysis, in which a variable's communality is defined as that part of its variance which is predictable from all the other variables in the analysis.

† If T is the dispersion matrix of set X and θ is that part of T which is predictable from set Y then the latent roots in question are the lambdas of the equation

$$|\theta - \lambda_i(T - \theta)| = 0$$

Summary

The writer demonstrates and illustrates the relations between spherical co-ordinate representation and numerical taxonomy. He shows how both methods can be employed to produce a comprehensive visual display of the structure of any two-dimensional data matrix.

References

Bartlett, M. S. (1950). "Tests of significance in factor analysis." *Brit. J. Psychol., Stat. Sec.*, 3, 77–85.

—— (1951). "The effect of standardization on a chi-squared approximation in factor analysis." *Biometrika*, 38, 337–344.

Burt, C. (1940). *The Factors of the Mind.* London: University of London Press.

De Morgan, A. (1847). *Formal Logic.* London: Taylor and Walton.

Hope, K. (1966). "Cos and Cosmos." *Brit. J. Psychiat.*, 112, 1155–1163.

—— (1969). *Methods of Multivariate Analysis.* London: University of London Press.

Slater, P. (1958). "The general relationship between test factors and person factors: application to preference matrices." *Nature*, 181, 1225.

—— (1964). *The Principal Components of a Repertory Grid.* London: Andrews.

Sokal, R. R., and Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy.* San Francisco: Freeman.

K. Hope, M.A., Ph.D., B.A., *M.R.C. Unit for Research on the Epidemiology of Psychiatric Illness, Edinburgh University Department of Psychiatry, Royal Edinburgh Hospital, Morningside Park, Edinburgh 10; now Lecturer in Methods of Social Research in the University of Oxford, and Fellow of Nuffield College*

(*Received 6 March, 1968, originally submitted 9 November, 1966*)

https://doi.org/10.1192/bjp.115.526.1069 Published online by Cambridge University Press