MULTIVARIATE ANALYSIS AS SPATIAL REPRESENTATIONS OF DISTANCES*

Joseph Woelfel

Michigan State University

East Lansing, Michigan

*I am grateful to James Gillham and George Barnett for assistance throughout this work, and to my colleagues at the University of Illinois Department of Sociology, where much of this work was done.

Rah Press

MULTIVARIATE ANALYSIS AS SPATIAL REPRESENTATION OF DISTANCES

Joseph Woelfel
Michigan State University

This paper begins by arguing that the multivariate data matrix with which all multivariate analysis begins may be seen as a vector space in which each variable is represented as a vector whose length is a function of its standard deviation and whose location relative to all other vectors is given by the angles (correlations) among the vectors. Different techniques of multivariate analysis (noteably correlation, multiple regression, path analysis, factor analysis and multidimensional scaling) are compared in terms of their ability to describe and make salient various different aspects of this underlying vector structure. The central argument is that classical multidimensional scaling, a technique seldom used by sociologists, offers clear-cut advantages over the other techniques for the description of the structure of aggregate (cultural as opposed to individual) data, particularly for the description of large scale cultural processes which take place over time. The paper argues that levels of descriptive accuracy and measurement reliability in the range of physical science research can be obtained by this technique.

The development of sophisticated tools of multivariate analysis in sociology has essentially been a process of carrying over techniques from other disciplines, principally psychology, economics, biology, and of course, statistics. While this has saved considerable work, it sometimes has the disadvantage of providing the sociologist with a fairly disjointed collection of techniques whose communalities are frequently overlooked or even unknown. The purpose of this paper is to point out the fundamental methematical and conceptual relationships among three powerful multivariate techniques: path analysis, factor analysis, and somewhat less well known to sociologists, multidimensional scaling.

In the most fundamental sense, all measurement, even in the social sciences, is the measurement of comparative distances. If we find, for example, that person A scores 5 on a political activism scale while person B scores 7, we have essentially established a distance of two scale units between A and B on the variable in question. When only one variable (such as the political scale just mentioned) is involved, only distances among persons can be ascertained, and very limited information is available. When more than one variable is available (say, politics and income) it becomes possible to establish distances not only among persons but also among the variables themselves. Thus, in a very imprecise way, if we find that the scores of a set of people are essentially similar on two variables (i.e., the values of scores of individuals on the two variables

are in some sense correlated) we can say that the variables are similar or "close" to each other. In the limiting case, if the scores of all individuals were exactly the same across two variables, i.e., if person i had an identical score on both variables, and that this was true for all n persons, then the two variables would be identical (coextensive). Thus, most multivariate techniques essentially attempt to establish relationships (distances) among variables on the basis of the scores of individuals on those variables.

The ordinary data collected in the typical multivariate analysis constitutes an nxv data matrix, where the columns $C_1, C_2 \ldots C_v$ represent concepts or variables and the rows $p_1, p_2 \ldots p_n$ represents persons. Thus, the typical data matrix D can be essentially seen

$$D = \begin{array}{c|cccc} & C_1 & C_2 & \cdots & C_v \\ p_1 & x_{11} & x_{12} & \cdots & x_{1c} \\ p_2 & x_{21} & x_{22} & \cdots & x_{2c} \\ \vdots & \vdots & \vdots & & \vdots \\ p_n & x_{n1} & x_{n2} & \cdots & x_{nc} \end{array}$$

to represent a vector space V_c where each person is uniquely represented as a vector in that space whose coordinates are given by his scores on all the variables. Similarly the transpose (D') of that matrix will be a vector space V_n wherein each variable is uniquely located in the space as a vector whose coordinates are given by its scores across all individuals. These vector spaces are very cumbersome, however, since these matrices are usually quite large and unwieldy.

As a next step in reducing the complex data matrix to a comprehensible size, it is usual to premultiply the matrix D by its transpose to yield a new matrix:

$$D'D = S$$

where S (sometimes unfortunately called a "cross-products" matrix) is a matrix of inner or scalar products such that any entry

$$S_{ij} = p\ p\ \cos\alpha_{ij}$$

where p_i = the length of vector $\underline{i}$

p_j = the length of vector $\underline{j}$

α_{ij} = the angle included between vectors $\underline{i}$ and $\underline{j}$

Thus, any entry s_{ij} in the matrix S represents the distance from variable $\underline{i}$ to an origin for the space times the distance of variable $\underline{j}$ from that origin times the cosine of the angle between them. It is usually the case, however, that data have been collected on scales of different range for different variables (i.e., variable 1 may be measured on a 5-point likert scale; variable 2 may be age, etc.) and so the vector lengths are usually artifacts of the data collection procedure. Since the length of any vector p_i can be shown to be equal to $\sigma_i \sqrt{n}$, where σ_i = the standard deviation of $\underline{i}$ then any entry $s_{ij} = p_i p_j \cos\alpha_{ij} = n\sigma_i\sigma_j \cos\alpha_{ij}$. Dividing through by $n\sigma_i\sigma_j$ for every cell ij will thus standardize each vector to unit length. The resulting matrix, C, of course, is a standard correlation matrix where any entry $C_{ij} = \cos\alpha_{ij} = r_{ij}$.

While this new matrix C is more parsimonious than the original data matrix, clearly much information has already been lost (or, more aptly, was never really present) since the true vector lengths are unknown. Consequently, while the angles between variable vectors are known*, the distances between concepts are lost and cannot be recovered from this matrix. Since both factor analysis and path analysis usually begin with the correlation matrix, this loss is not trivial.

In spite of these losses, the data may not yet be expressed as parsimoniously as possible. It is very likely that not all the vectors present in the matrix C are necessary to represent all the information in the matrix--in fact, n points may always be scaled in at least n-1 space, and n-2 space if the data are monotone in form (Lingoes, 1971). This process, as all the multivariate techniques discussed here, involves the selection of some set of vectors smaller than the order of the matrix C in terms of which the data may be described.

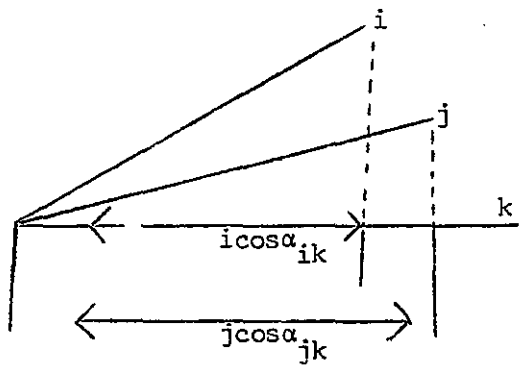
If we are interested in "explaining" or accounting for one or more of the variables in the matrix by means of some subset of the other variables (as in regression and path analysis), then the vector of that dependent variable is taken as a criterion. The variable is "explained" by measuring the net projections on the subset of independent vectors of that predictor vector. Since the predictor vectors cannot be guaranteed orthogonal to each other, the simple cosines of the correlation matrix will not suffice, however, even though--in the standardized case

---

*Standardizing the variables in this fashion also has the effect of establishing a common origin for all of the vectors at the centroid of the space, a fact which will be important later.

discussed here--the projection of any vector i on any vector j is given by the cosine of the angle α_{ij}. This is true since the projection of two correlated vectors i and j on a third vector k will overlap, as Figure 1 illustrates. Consequently, what is required is the set of partial projections; i.e., the proportion of the dependent vector K which is accounted for by the projection of j on k controlling for the projection of i on k, and vice versa. Consequently, these values (b's, β's, or partial correlations, depending upon the kind of analysis) are dependent on the angles between independent variable vectors. In the case of simple partial regression, this latter information is not conveniently available.

FIGURE 1:



Path analysis retrieves some of this information by selecting a subset of several of the vectors in the matrix in their turn as dependent variables, and consequently, some additional information about the angles among vectors is added, but considerable information is still lacking. To be sure, as correlations among residual variables and other data are added, additional information becomes available,

but the presentation of this data in terms of quantitative partial projections among correlated vectors in a space of unknown dimensionality can be immensly confusing. Furthermore, although the path diagrams accompanying path analysis appear to be pictorial, they are so only in a vague sense, since neither the lengths of the vectors nor the angles among them are rendered to scale in these drawings.

Since path analysis presents so fragmentary a picture of the latent structure of the data, considerable prior information about the theoretical relationships among the variables is absolutely requisite for utilization of the technique. In fact, since the path analytic model is so heavily identified, the path analyst is essentially constructing a mathematical vector structure of his own prior to the analysis and measuring the extent to which the data conform to that structure. Should the correspondence be poor, the range of possible alternative models the investigator may then attempt, even given all the information provided by path analysis, is usually overwhelming. Undoubtedly due to this fact, coupled with the imprecision of current sociological theory, it is safe to say that no truly satisfactory path model has yet been presented in the sociological literature, nor is one likely to be soon.

Factor analysis also attempts to describe the set of variable vectors in terms of a smaller set of vectors, but rather than selecting a subset of the original variable vectors as its criterion set, it constructs new vectors deliberately structured to be as convenient as possible. Essentially, factor analysis generates an orthonormal base which spans the vector space defined by the variable vectors. Since

the reference vectors (factors) of this base are orthogonal, the length of any vector is simply the sum of the squared loadings of that variable on each of the factors. In an important sense, factor analysis is a truly pictorial representation of the data, since a plot of the variables in the space spanned by the factors will represent the angles among variables graphically and to scale. Since the factor solution is under-identified, few constraints are imposed in the data, and the latent structure of the data is more clearly exposed. Unfortunately, since the data have been standardized, the vector lengths are unvailable, however, and while the factor space reveals directions and angles, it is not able to represent the distances among the variables, and so important information is still missing. Probably because only angles and consequently directions are represented in the factor space, almost all sociological uses of factor analysis have focused upon the identification of the "meanings" or "interpretations" of the directions in the space; e.g., one attempts to determine whether moving up a given factor increases one's political radicalism, etc. Consequently, the wholistic spatial properties of the factor space remain obscure, and attention is usually directed exclusively to the individual factors, one by one. This attempt to "name" the factors frequently make sense, but there is no reason why it must be so, and undoubtedly too much attention has been focused on the interpretation of factors.

Multidimensional scaling* while generally unfamiliar to most sociologists, can be seen as essentially an unstandardized factor analysis,

---

*This discussion will confine itself mainly to fully metric multidimensional scaling for aggregate data matrices, since this type of MDS is most likely the technique of choice for most sociological work.

with certain qualifications. As suggested earlier, a basic reason for the standardization of the scalar products matrix in both paths and factor analysis is that the vector lengths are generally artifacts of the scaling procedures utilized in the data collection process. Thus, the original "distances" in the raw data matrix are partly artifactual. It is possible, however, and in fact even easy, to obtain direct distance estimates among variables which are not artifacts, particularly in the sociological case. The simplest technique is to ask respondents to estimate distances among concepts directly after providing a suitable standard unit of measure; e.g., "If X and Y are u units apart, how far apart are a and b?" This type instrumentation has been little used by psychometricians, primarily since it requires a complex judgment from the respondent while providing virtually no structure for the scale, and thus such estimates tend to be highly unreliable for the individual case, and so many more reliable, although only ordinal, measures have been devised (Torgerson, 1957). The sociologist, however, since he is usually interested in determining the relationships among variables across many individuals, can easily obtain any level of reliability desired by averaging each distance estimate across all members of the sample and increasing his sample size for each distance estimate until satisfactory reliabilities are obtained. Since the mean distance will converge on the population true score as n becomes large, the sociologist may take advantage of the fact that the scale mentioned is a fully metric unbounded continuous ratio scale. Because all such distance estimates are averaged over the sample of respondents, the law

of large numbers assures that the precision of measurement attained is essentially a direct function of the sample size.

Procedurally, this data collection technique yields a three-dimensional concepts x concepts x persons matrix which is averaged across the n persons into a two-dimensional concept x concepts square symmetric matrix D where any entry d_{ij} represents the average distance between concepts i and j as seen by the respondents. This matrix D is transformed routinely into a scalar products matrix B, although generally it is the practice of investigators to 'double-center' this matrix by establishing an origin for the space at the centroid of the distribution. This can be done simply during the construction of the scalar products matrix, and the transformation is given by the equation:

$$b^{*}_{ij} = 1/2 \left(\frac{\sum_{i=1}^{n} d^2_{ij}}{n} + \frac{\sum_{j=1}^{n} d^2_{ij}}{n} - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d^2_{ij}}{n} - d^2_{ij} \right)$$

which is a straightforward linear transformation which sacrifices none of the information present in the original matrix D (Torgerson, 1958).

This new centroid scalar products matrix is such that any entry:

$$b^{*}_{ij} = p_i p_j \cos\alpha_{ij}$$

but it is important to recall that the vector lengths p_i and p_j are not artifacts in this case, and so standardization of this matrix is neither necessary nor desirable. Consequently, when this matrix B^{*} is reduced to its base by routine factorization (i.e., the application of any standard eigen routine, such as principal axis or jacobi), the result is

a

a factor matrix, F, whose columns F_1, $F_2 \ldots F_k$ are orthogonal vectors with their origin at the centroid of the vector space spanned by F, and where any entry F_{ij} represents the projection (loading) of the ith variable of the jth factor. This matrix has the further properties such that:

$$p_i = \sqrt{\sum_{f=1}^{k} a_{if}^2}$$

where a_{if} = the loading of the ith variable on the fth factor. That is, the square root of the sum of squared projections of the ith variable across all the k factors equals the length of the vector of the ith variable, and, of central concern:

$$d_{ij} = \sqrt{\sum_{f=1}^{k} (a_{if} - a_{jf})^2}$$

This last expression shows that the original distance matrix can be completely recovered from the factor matrix with no loss of information. It is even possible, based on the strength of two additional but plausible assumptions, to recover still further information as follows:

Almost all scaling techniques, whether uni- or multidimensional, commonly share a single starting assumption; that is, that concepts may be represented as points on a continuum or in a space. This assumption, however, is almost certainly overly rigid in almost all circumstances. What is more likely is that concepts or variables

being scaled are representable more accurately by intervals on a scale or regions in a space. The color spectrum, for example, does not represent colors as points on a scale, but intervals. Moreover, some colors occupy larger intervals than others; yellow, for example, occupies a smaller interval of the color spectrum than blue. Furthermore, when respondents are asked to estimate the distances between such concepts, it is likely that the distance between the near boundaries of the regions will be reported. As Figure 2 illustrates, these (reported) surface-to-surface distances are related to the center-to-center distances by the expression:

$$\hat{d}_{ij} = d_{ij} + r_i + r_j$$

where $\hat{d}_{ij}$ = center-to-center distance

d_{ij} = (reported) surface-to-surface distance

r_i = the radius of concept i

r_j = the radius of concept j

It may be argued, then, that all original distance estimates are systematically too small by a variable amount. Furthermore, attempts to fit these truncated distances into a real space will be thwarted. By definition, a real space is one in which any three points i, j, and k must satisfy the relation*.

$$d_{ij} + d_{ik} \geq d_{jk}$$

$$d_{ij} + d_{jk} \geq d_{ik}$$

$$d_{ik} + d_{jk} \geq d_{ij}$$

---

*If one of these expressions is satisfied as an equality, the points are collinear; if all are satisfied as equalities, the points are coterminus.

When the point assumption is violated, as in the matrix D, however, attempts to represent the distances among the surfaces of the hyperspheres as distances among points will generally fail to satisfy the "triangle inequalities" constraints described above. Such a matrix will not be positive, and factorization will yield negative eigenroots signifying the projections of at least some of the variable vectors into imaginary space, that is, a space in which distances may be negative. Since we have attributed this failure of the real space assumption to a shortening of the distances in the space by a function of the sizes of the concepts scaled, what is called for is a reduction of the imaginary space to zero by an expansion of the real space. This can be done conveniently by subtracting the largest negative eigenroot (i.e., the smallest root algebraically or λmin) from every entry of the diagonal of the centroid scalar products matrix B^*, since λmin equals the sum of the squared projections of all the concepts scaled on the largest negative factor and hence represents the sequared vector length of the longest imaginary factor, while the diagonal entries of the matrix B^* represent the lengths of the vectors of all the scaled concepts in the space. This operation:

$$\hat{B} = B^* - I\lambda\text{min}$$

will yield an adjusted scalar products matrix $\hat{B}$ which is just positive semidefinite (i.e., contains no negative latent roots). Since the off diagonal cells of $\hat{B}$ are the same as those of B^*, and since they further represent:

$$\hat{B}_{ij} = p_i p_j \cos\alpha_{ij}$$

where p_i and p_j have been increased, this operation reduces $\cos\alpha_{ij}$, thus increasing α_{ij} and consequently every distance d_{ij} in the original distance matrix D will be increased by a function of the cosine of the angle α_{ij}. If the original distance matrix D is subtracted from the distance matrix $\hat{D}$ corresponding to the matrix $\hat{B}$, the resulting matrix R can be seen to be a matrix of sums of radii corresponding to the scalar equation:

$$\hat{d}_{ij} - d_{ij} = r_i + r_j$$

or, in matrix form:

$$\hat{D} - D = R$$

This matrix R is overidentified and easily solved for the individual radii.

The advantages of such a technique are dramatic. First, it enables fully continuous true ratio scaling of any level of precision desired (accuracy equivalent to typical physical science measures is not unrealistic). Secondly, no information contained in the data need be lost, and in fact much latent information is uncovered. Third, the solution arrived at is fully graphic and, particularly when the dimensionality of the resulting space is three or fewer, as is very frequently the case, even visual. Of perhaps even greater importance, given the application of a suitable rotation and translation routine*, is the clearout advantage of metric multidimensional

*Although rotation is not discussed here, several satisfactory techniques and software are available. See Cliff, 1969.

scaling for studies involving time-ordered observations. Given a series of observations over a set of known time periods, by the simple subtraction of coordinates over time, motions through the spatial manifold over time may be expressed as velocities, as given by:

$$V_i = \frac{d_i}{\Delta t} = \frac{\sqrt{\sum_{f=1}^{K} (a_{if} - b_{if})^2}}{t_2 - t_1}$$

where a_{if} = the loading of the ith variable measured at time one on the fth factor of the t_1 space

b_{if} = the loading of the ith variable measured at time two on the fth factor of the t_2 space

d_i = the distance variable i has moved across time

Δt = the interval of time between measures

V = velocity

and, given multiple time periods, as accellerations:

$$\overline{A}_i = \frac{\Delta V_i}{\Delta t}$$

The chief focus of attention of the investigator, of course, can then be put on the lawfulness of these movements through the space over time. Should these movements prove lawful in their behavior, that is, if they could be shown to move in some clearly patterned fashion, theoretical development could proceed very swiftly due to the very high level of measurement precision obtainable.

It would seem, in the light of these considerations, particularly the plausibility of its principle assumptions, the highly quantitative mathematical theoretical structure suited to these techniques, and its

consequently very high level of measurement precision, along with its graphic qualities and particular suitability for time-ordered study of aggregate cultural variables, techniques of multidimensional scaling--particularly the metric or "classical" model described in this paper, deserve greatly increased attention by mathematically oriented sociologists and theorists.

References

Cliff, Norman. Analytic Rotation to a Functional Relationship. Psychometrika, 27, 283-296.

_____. Orthogonal Rotation to Congruence, Psychometrika, 31, 33-42.

Torgerson, Warren. Theory and Method of Scaling, New Yor: John Wiley & Sons, 1958.