A GENERAL MODEL FOR NONMETRIC MULTIDIMENSIONAL SCALING[1]

Terry C. Gleason

The University of Michigan

June, 1967

[1] The work reported here was supported by Grant MH-10834 from the National Institute of Mental Health and by Grant GM-1231 from the National Institutes of Health. The author wishes to thank Drs. Dorwin Cartwright and Robyn M. Dawes for their many helpful comments on earlier drafts of this paper.

Michigan Mathematical Psychology Program : : MMPP 67 - 3

ABSTRACT

Nonmetric multidimensional scaling is a technique for finding a geometric representation of a proximity matrix using only the ordinal properties of the measures of proximity. A general model for finding such geometric representations is presented, and it is shown that current methods are special cases of the general model. It is also shown how the general model may be applied to conditional proximity matrices, a form of data not formerly amenable to such analysis.

A GENERAL MODEL FOR NONMETRIC MULTIDIMENSIONAL SCALING

There are many ways of approaching the problem of finding a geometric representation of a given data matrix. If the data can be interpreted as distances between points or as scalar products of vectors, and these measures are at the level of a ratio scale, then the traditional approach of factor analysis provides a means for recovering a configuration of points satisfying these distances. It is seldom, however, that ratio scale measurements are achieved without rather arbitrary assumptions. Consequently, in recent years much effort has been expended in attempts to find ways of recovering a geometric representation of a data matrix using only the ordinal properties of the data. This kind of approach, refered to here as nonmetric multidimensional scaling, had its origin in the work of Shepard (1962a, 1962b) and has been extended by Kruskal (1964a, 1964b) and Lingoes (1965a, 1965b, 1966). It is the intent of this paper to present a general model for nonmetric multidimensional scaling which incorporates each of the current methods as special cases. With the aid of the general model it is possible to compare these methods and to see how nonmetric multidimensional scaling may be extended to conditional proximity matrices, a kind of data not previously amenable to such analysis.

For the sake of clarity we shall adopt the following definitions and notational conventions. Let $X = \{x_1, x_2, \ldots, x_p\}$ be a set of p vectors (or points) in a space of n dimensions. The set X will be called a *configuration*. To denote a coordinate of a given vector we shall use double subscripts; thus $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$. In general the indices i and j will be used to denote points and the index k to denote dimensions. The

distance between tw points x_i and x_j will be denoted d_{ij}. For the purposes of this paper only Euclidean distances will be considered, hence,

$$(1) \qquad d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \; .$$

The purpose of nonmetric multidimensional scaling is to represent stimuli or individuals as a configuration of points in a space in such a way that some or all of the interpoint distances are monotonically related to measures of proximity among the stimuli or individuals. Hence the basic data for any scaling technique is a proximity matrix $S = [s_{ij}]$. There seems to be a wide range of measures that can be interpreted as proximities. For example, one might use correlations, judgments of similarity (Atteneave, 1950), confusions among competing responses (Dawes & Kramer, 1966) references between journals (Coombs, 1964), and preferential choice data, to name just a few. From the point of view of the model it is unimportant whether interpoint distances are a decreasing function of the proximities as in the case of similarities or an increasing function as in the case of dissimilarities. For convenience we shall assume for the remainder of this paper that if $s_{gh} < s_{ij}$, then we shall want to find points x_g, x_h, x_i, and x_j such that $d_{gh} < d_{ij}$.

Following Coombs (1964) we may distinguish four kinds of proximity matrices by means of two dichotomies: complete--conditional, and diagonal--off diagonal. If any two entries in S may be meaningfully compared, then S is said to be a *complete* proximity matrix. If meaningful comparisons may be made only within rows of S, then S is said to be a *conditional* proximity

matrix. For example, a correlation matrix is a complete proximity matrix. Given any two correlations we can meaningfully state which is the greater or whether they are equal. However, if for each row of the matrix we rank ordered the columns according to the magnitude of the correlations, say, giving the highest correlation the smallest rank and the algebraically smallest correlation the largest rank, then the matrix obtained by replacing the correlations with their ranks is a conditional proximity matrix. Meaningful comparisons between entries can no longer be made across rows but only within rows.

Let X be a set of points partitioned into two disjoint subsets $X_1 = \{x_1, \ldots, x_r\}$ and $X_2 = \{x_{r+1}, \ldots, x_p\}$. In Figure 1 we have schematically represented a general proximity matrix. The submatrices A and D are characterized by the fact that their entries are proximities between points of the same set, *i.e.*, the entries of A are a function of distances between points in X_1, and the entries of D reflect distances between points in X_2. Such matrices will be called *diagonal* proximity matrices. The proximities in matrices B and C represent distances between points drawn from different sets. The matrices will be called *off-diagonal* proximity matrices.

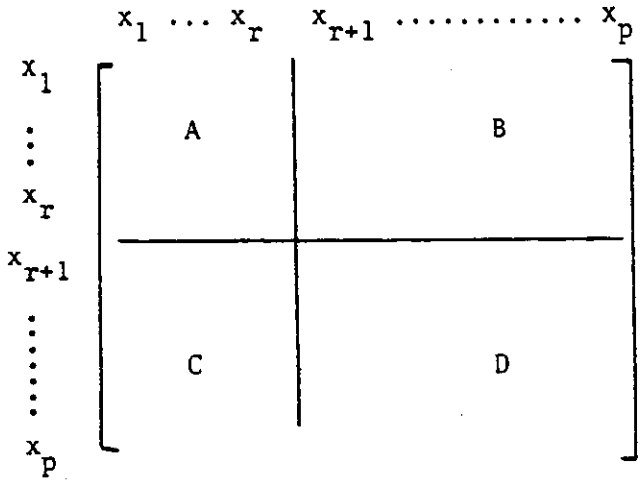

Fig. 1. A general proximity matrix consisting of diagonal matrices A and D and off-diagonal matrices B and C.

By combining these two dichotomies we obtain four kinds of proximity matrices. A *complete diagonal proximity matrix* contains proximities of pairs of points drawn from the same set and these proximities are comparable over the entire matrix. Examples of this kind of proximity matrix would include correlation matrices and judgments of similarity obtained by the method of paired comparisons. Since the entries of a complete diagonal proximity matrix are to be interpreted as some monotonic function of the distances between corresponding pairs of points, they must be symmetric, i.e., since $d_{ij} = d_{ji}$, it must be the case that $s_{ij} = s_{ji}$. Hence for convenience this kind of proximity matrix will be called simply a *symmetric proximity matrix.*

A *conditional diagonal proximity matrix* contains proximities of pairs of points drawn from the same set but these proximities are comparable only within rows of the matrix. This type of data is obtained, for example, when

one has a subject rank order a set of stimuli with respect to their similarity to a standard using, in turn, each of the stimuli in the set as the standard.

A *conditional off-diagonal proximity matrix* contains proximities of pairs of points drawn from different sets, and these proximities are comparable only within rows of the matrix. This kind of data arises, for example, when individuals (points of one set) indicate their preferences for objects (points from another set).

Finally a *complete off-diagonal proximity matrix* contains proximities of pairs of points drawn from different sets, and these proximities are comparable only within rows of the matrix. This kind of data arises, for example, when individuals (points of one set) indicate their preferences for objects (points from another set).

Finally a *complete off-diagonal proximity matrix* contains proximities of pairs of points drawn from different sets, and these proximities are comparable over the entire matrix. This kind of data is very rare, and therefore, will not be discussed further.

The fourfold classification of proximity matrices permits the placing of current scaling methods within a general framework. The techniques of Shepard (1962a, 1962b), Kruskal (1964a, 1964b), and Lingoes (1965b) are limited solely to symmetric proximity matrices--a very strong class of data. The method to be presented later in this paper as well as the work of Lingoes (1966) extends nonmetric multidimensional scaling to the much weaker, but more common, conditional proximity matrices (both diagonal and off-diagonal).

THE GENERAL MODEL

It is useful to begin the presentation of a general model for non-metric multidimensional scaling techniques by first developing some geometrical facts. Let x_0 and x_1 be arbitrary but fixed points in n-space. A point x lying on the line passing through these points may be described parametrically by the equation,

$$x = (1 - \lambda)x_0 + x_1 \tag{2}$$

or alternately as

$$x = x_0 + \lambda(x_1 - x_0). \tag{3}$$

In this last representation, the point x_0 may be interpreted as a base point, the term $x_1 - x_0$ as a direction, and λ as a distance. Thus if $\lambda > 0$, we may say that x is at a distance λ in the direction of x_1 from x_0. Likewise if $\lambda < 0$, then x is at a distance $|\lambda|$ from x_0 in the direction away from x_1. The observations are illustrated for several values of λ in Figure 2.

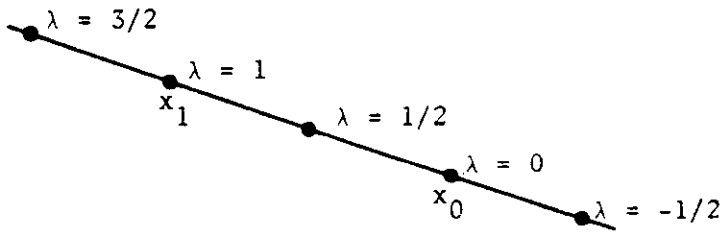

Fig. 2. Positions of the point $x = x_0 + \lambda(x_1 - x_0)$ for various values of λ.

Equation (3) is particularly helpful in stating, conceptually, the iterative methods used for calculating a configuration of points to fit a

given proximity matrix. Suppose we are given the following items: a set of proximities s_{ij}, a set of distances δ_{ij} which are monotonically related to the proximities and which represent distances between points in a configuration that adequately represent the data (the distances δ_{ij} will henceforth be called *appropriate distances*), and an arbitrary configuration of points X. For each pair $x_i, x_j \in X$ we may calculate the distance d_{ij}. If X is an appropriate configuration in the sense that its distances are monotonically related to the δ_{ij}, then the distances d_{ij} should be very close to or exactly the same as the appropriate distances. Generally, however, the monotonicity criterion will not be satisfied so that the points of X will have to be moved so as to improve the match between d_{ij} and δ_{ij}. Consider the ratio $\frac{\delta_{ij}}{d_{ij}}$. If d_{ij} is too large, that is the points x_i and x_j are too far apart, then $\frac{\delta_{ij}}{d_{ij}} < 1$. On the other hand, if d_{ij} is too small, $\frac{\delta_{ij}}{d_{ij}} > 1$. These facts suggest that to calculate a new position for the points x_i we need only substitute $(1 - \frac{\delta_{ij}}{d_{ij}})$ for λ in equation (3), that is

$$x = x_i + \left(1 - \frac{\delta_{ij}}{d_{ij}}\right)(x_j - x_i). \tag{4}$$

The point x in equation (4) represents a new estimate of the position of point x_i. Note that the distance of the point x from x_i is proportional to the discrepancy between d_{ij} and δ_{ij}. If $d_{ij} = \delta_{ij}$, then $x = x_i$. As d_{ij} becomes larger relative to δ_{ij}, x approaches more and more closely the point x_j. As d_{ij} becomes smaller relative to δ_{ij}, x moves farther and farther away from x_j.

For convenience let us adopt the convention

$$\lambda_{ij} = 1 - \frac{\delta_{ij}}{d_{ij}} . \tag{5}$$

We have seen from the above discussion that given a point x_i, we can get an estimate of a new position for x_i by applying equation (4) for each point x_j. To arrive at a new position for x_i using estimates from all the points x_j we simply take a weighted vector sum of each estimate, that is,

$$x_i' = x_i + \alpha \sum_{j=1}^{p} \lambda_{ij}(x_j - x_i), \tag{6}$$

where α is the weighting constant. Subsequently α will be called the *step size*.

A general method, then, for calculating a configuration of points that satisfies a given proximity matrix consists of resolutions to the following problems: (a) selecting an appropriate initial configuration X_0, (b) estimating the appropriate distances between points from the proximities measures, that is, finding a function f such that $\delta_{ij} = f(s_{ij})$, (c) determining an appropriate constant α so that with repeated applications of (6) the successive configurations $X_1, X_2, \ldots, X_t$ converge, and finally, (d) constructing some measure of goodness of fit of the configuration to the data.

EXAMPLES OF THE GENERAL METHOD

It is instructive to compare this general method with the approaches of Shepard, Kruskal and Lingoes. We shall show in the following sections that each of the methods previously proposed may be seen as a special case of the above general outline and that each differs from the others only in the particular resolutions chosen for the four problems mentioned above.

Shepard's method, being the first to appear in the literature (1962a, 1962b) is now largely of historical interest. Its principal value here will be as a contrast to the methods of Kruskal and Lingoes. The first step is to standardize the proximity measures so that for all i and j, $0 \leq s_{ij} \leq 1$. This introduces no difficulties since the only property of the proximity measures that will be used in calculating a configuration of points to represent them will be their order relations. A configuration X with associated distances d_{ij} will represent a perfect fit to the data, if when the proximities s_{ij} and the distances d_{ij} are ranked low to high, the rank assigned s_{ij} is the same as the rank assigned d_{ij}. If X fails to fit the data, there will be, of course, some discrepancies in the two rank orders. Let $s(d_{ij})$ denote that proximity measure having the same rank as d_{ij}. Rather than attempting to estimate the appropriate distances δ_{ij}, Shepard turns the problem around and attempts to estimate the function f from distances to proximities by a comparison of the proximity measures. The quantity $s_{ij} - s(d_{ij})$ "amounts in effect to a comparison of each proximity measure with its corresponding distance after the scale of distance has been subjected to a nonlinear transformation that renders the distribution of distances identical with the dis-

tribution of proximity measures" (p. 134) . With this exception, Shepard's formula can be seen as essentially the same as that given as equation (6), i.e.,

$$(7) \quad x_i' = x_i + \alpha_1 \sum_{j=1}^{p} \left[\left(\frac{s_{ij} - s(d_{ij})}{d_{ij}} \right) (x_j - x_i) \right] + \sum_{j=1}^{p} B_j .$$

The factor $\sum_{j=1}^{p} B_j$ was introduced to reduce the dimensionality of the solution[2]. In general its contribution to the estimation of new coordinates for the point x_i is much less than the other factor in the equation (7) and hence it can be ignored for present purposes. The parameter α_1 is an empirically determined constant which serves to speed convergence.

As an initial configuration, Shepard suggested using a $p - 1$ dimensional simplex. After sufficient repeated applications of equation (7), the resulting configuration could be rotated to principal axes and an estimate of the appropriate dimensionality of the configuration obtained from this rotation. To evaluate the goodness of fit of the configuration to the data Shepard suggested the following measure.

$$(9) \quad S_1 = \frac{\sum_{j=1}^{p} \sum_{i=j+1}^{p} [s_{ij} - s(d_{ij})]^2}{\frac{p(p-1)}{2}}$$

Clearly when a perfect fit is achieved $s_{ij} = s(d_{ij})$ for all i and j, hence $S_1 = 0$.

---

[2] Actually $\sum_{j=1}^{p} B_j = \beta_1 \left(\frac{s_{ij} - \frac{\sum_{ij} s_{ij}}{\frac{p(p-1)}{2}}}{d_{ij}} \right) (x_j - x_i)$ where $\alpha_1 >> \beta_i$.

Kruskal's (1964a) approach to the scaling problem is essentially the same as that of Shepard, but it does have the desirable difference that Kruskal provides a rationale for many of the steps that Shepard had evolved on an intuitive basis. Whereas Shepard sought some monotonic function to relate s_{ij} and d_{ij}, Kruskal suggests calculating a best fitting monotonic function in a least squares sense. In particular, his method seeks values $\hat{d}_{ij}$ such that

$$\sum_{i=1}^{p} \sum_{j=i+1}^{p} (d_{ij} - \hat{d}_{ij})^2$$

is a minimum. He defines the following terms:

$$S^* = \sum_{i=1}^{p} \sum_{j=i+1}^{p} (d_{ij} - \hat{d}_{ij})^2, \tag{10}$$

$$T^* = \sum_{i=1}^{p} \sum_{j=i+1}^{p} d_{ij}^2, \tag{11}$$

$$S_2 = \sqrt{\frac{S^*}{T^*}}. \tag{12}$$

This last term, S_2, was called by Kruskal the *stress* of a configuration. Stress serves two functions in Kruskal's method. First it provides a measure of the goodness of fit of the configuration to the data. It is easy to see that S_2 approaches zero as the d_{ij}'s approach the $\hat{d}_{ij}$'s in much the same way that S_1 approached zero as the configuration approached a better fit in Shepard's method. However, unlike the calculations for S_1, Kruskal uses only the d_{ij}'s in calculating stress. This seems a much more reasonable procedure since we cannot be sure of the arithmetic properties of the s_{ij}'s having only required that they be measured on an ordinal scale. The second function provided by S_2 is its use in deciding how points should be moved

in calculating successive iterations. Kruskal suggests using the "method of steepest descent" in attempting to minimize S_2 as a function of the coordinates x_{ik} for each point x_i in the configuration. This is done by calculating the negative gradient for each coordinate, i.e., finding $-\frac{\partial S_2}{\partial x_{ik}}$ for each $i = 1, 2, \ldots, p$, $k = 1, 2, \ldots, n$. It is straightforward to show[3]

$$(13) \qquad -\frac{\partial S_2}{\partial x_{ik}} = \frac{1}{\sqrt{S^*T^*}} \sum_{j=1}^{p} \left(1 - \frac{\hat{d}_{ij}}{d_{ij}}\right)(x_{jk} - x_{ik}) .$$

Therefore a new estimate for the point x_i may be obtained by

$$(14) \qquad x_i' = x_i + \frac{\alpha_2}{\sqrt{S^*T^*}} \sum_{j=1}^{p} \left(1 - \frac{\hat{d}_{ij}}{d_{ij}}\right)(x_j - x_i)$$

where α_2 is a constant introduced to speed convergence. It is clear from equation (14) that Kruskal's iterative procedure is essentially that given in equation (6) differing only in the step size.

Kruskal's method differs from Shepard's in one other respect; in estimating the appropriate dimensionality of the configuration, Kruskal attempts solutions in n, n-1, n-2, etc. dimensions until the best stress value that can be obtained in a given number of dimensions becomes intolerably

---

[3] In this derivative T* has been treated as a constant function of x_{ik}. Given the general definition proposed by Kruskal this is not a necessity. But in the actual operation of his program he normalizes his configuration after each iteration so that

$$\sum_{i=1}^{p} \sum_{k=1}^{n} x_{ik}^2 = 1 \quad \text{and} \quad \sum_{i=1}^{p} x_{ik} = 0.$$

Under these constraints

$$T^* = \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} d_{ij}^2 = \frac{1}{2} \sum_{i=1}^{p} \sum_{k=1}^{p} \sum_{k=1}^{n} (x_{ik}^2 - 2x_{ik}x_{jk} + x_{jk}^2$$

$$= \sum_{i=1}^{p} \sum_{k=1}^{n} x_{ik}^2 - \sum_{k=1}^{n} \sum_{i=1}^{p} x_{ik} \sum_{j=1}^{p} x_{jk} = \sum_{i=1}^{p} \sum_{k=1}^{n} x_{ik}^2 = 1$$

justifying the derivative.

high. The appropriate dimensionality of the configuration can be estimated by looking at the pattern of stress values obtained from different dimensions.

As for the initial configuration problem, Kruskal states that any configuration may be used.

The method developed by Lingoes (1965a), while proceeding from a quite different rational, is very similar in form to that of Kruskal. Given a configuration of points one may obtain a new configuration by the following formula

$$x'_{ik} = \frac{1}{p} \sum_{j=1}^{p} c_{ij} x_{jk} \tag{15}$$

where c_{ij} is given by

$$c_{ij} = \begin{cases} 1 + \sum_{h=1}^{p} \frac{d^*_{ih}}{d_{ih}} & \text{if } i = 1 \\ 1 - \frac{d^*_{ij}}{d_{ij}} & \text{if } i \neq j. \end{cases} \tag{16}$$

It is understood in (16) that $\frac{d^*_{ii}}{d_{ii}} = 0$. The values d^*_{ij} are Lingoes' estimates of the appropriate distance δ_{ij}. They are obtained by a procedure called the *principal of rank images* wherein the distances d_{ij} are themselves permuted so that as a set of numbers they are ranked from low to high. Given this permutation the number d^*_{ij} is that number given the same rank as s_{ij}. With successive iterations the distances d_{ij} converge on the values d^*_{ij} thereby approximating a monotonic function of the proximities s_{ij}. Again the rationale is the same as previous methods, only the function chosen to estimate δ_{ij} differs.

To put equation (15) into a form more closely representing those given above we apply (16) so that

$$x'_{ik} = \frac{1}{p}\left[\sum_{\substack{j=1\\ j\neq i}}^{p}\left(1 - \frac{d^*_{ij}}{d_{ij}}\right)x_{jk} + \left(1 + \sum_{j=1}^{p}\frac{d^*_{ij}}{d_{ij}}\right)x_{ik}\right]$$

$$= \frac{1}{p}\left[\sum_{\substack{j=1\\ j\neq i}}^{p} x_{jk} - \sum_{\substack{j=1\\ j\neq i}}^{p} x_{jk}\frac{d^*_{ij}}{d_{ij}} + x_{ik} + \sum_{j=1}^{p}\frac{d^*_{ij}}{d_{ij}}x_{ik}\right].$$

Using the fact that $\frac{d^*_{ii}}{d_{ii}} = 0$, hence

$$\sum_{\substack{j=1\\ j\neq i}}^{p}\frac{d^*_{ij}}{d_{ij}}x_{jk} = \sum_{j=1}^{p}\frac{d^*_{ij}}{d_{ij}}x_{jk},$$

we get

$$x'_{ik} = \frac{1}{p}\sum_{j=1}^{p}\left[x_{jk} - \frac{d^*_{ij}}{d_{ij}}(x_{jk} - x_{ik})\right]$$

$$= \frac{1}{p}\sum_{j=1}^{p}\left[x_{ik} + \left(1 - \frac{d^*_{ij}}{d_{ij}}\right)(x_{jk} - x_{ik})\right]$$

or alternatively,

$$x'_i = x_i + \frac{1}{p}\sum_{j=1}^{p}\left[\left(1 - \frac{d^*_{ij}}{d_{ij}}\right)(x_j - x_i)\right]. \tag{17}$$

We have thus reduced (15) to the same form as (6) with a step size of $\frac{1}{p}$.

Lingoes' criterion for goodness of fit is similar to Kruskal's

$$(18) \qquad S_3 = \frac{\sum_{i=1}^{p} \sum_{j=i+1}^{p} (d_{ij} - d^*_{ij})^2}{2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} d_{ij}^2} .$$

Assuming $d^*_{ij} = \hat{d}_{ij}$ for each i and j (which, of course, will not be true in general), then $S_3 = \frac{1}{2} S_2^2$.

Lingoes has a unique approach to estimating the initial configuration and its dimensionality. If the proximities are ranked from smallest to largest, then the first dimension of the configuration may be approximated by the distance represented by the largest rank. To get successive orthogonal dimensions one assumes "that all distances are identical and equal to $p(p - 1)/2$, the largest possible distance. Now all that remains to be done is to form the ratios of the given ranks to the largest rank, calculate the c-matrix [i.e., the matrix whose entries are c_{ij} (see expression (16) above)], and solve for the eigenvalues and eigenvectors of C" (p. 7). Except for the vector with the largest root, the remaining eigenvectors may be used as a starting configuration. Lingoes suggests that starting with this configuration will escape problems of local minima to which Shepard's and Kruskal's methods are subject, but no proof is offered. Moreover, it is not clear from his discussion why the c-matrix so constructed can be reasonably considered a matrix of scalar products.

The above discussion has shown how similar existing methods of nonmetric multidimensional scaling are. The differences lie in their resolutions of the problems of estimating an initial configuration, choosing an appropriate

step size to speed convergence, and estimating the function which monotonically relates distances to proximity measures. Despite much experience with each of these algorithms, there is as yet no proofs available for the convergence of these procedures. Undoubtedly all these programs are subject to problems of local minima, that is, there exist configurations which cannot be improved upon by small movements of the points, but which have poorer fits to the data than other configurations. This is particularly true, for example, when all points are given the same coordinates. In such a case $S_2 = S_3 = 0$. At present the best approach to the problem of local minima is to attempt several solutions from different starting configurations.

It is worth mentioning that there is another possible approach to the problem of estimating a reasonable starting configuration and its appropriate dimensionality that has not been suggested by the methods discussed above. Coombs and Kao (1960) conjectured that if one correlates all pairs of rows of an off-diagonal proximity matrix, the eigenvectors of the resulting correlation matrix will provide an approximation to the configuration of points represented by the matrix. In their conjecture they indicated that this configuration should overestimate the appropriate dimensionality of the data by one dimension. Ross and Cliff (1964) proved a similar theorem for a correlation matrix obtained by correlating squared distances. Their result indicates that the appropriate dimensionality of the configuration will be overestimated by two dimensions. The author has experimented with a number of empirical and artificial examples and has found that estimating an initial configuration from the eigenvectors of

of the correlation matrix is a satisfactory point for analyzing both diagonal and off-diagonal proximity matrices. If one uses those eigenvalues which account for 99% of the trace of the correlation matrix (which is, of course the number of rows in the proximity matrix), the method generally overestimates the appropriate dimensionality of the configuration by at most one or two dimensions, and, on occasion, the estimated number of dimensions is correct.

A GENERAL METHOD FOR CONDITIONAL PROXIMITY MATRICES

The methods so far discussed have been concerned with analyzing symmetric proximity matrices. We now consider the case of data described by a conditional diagonal proximity matrix or even more generally by an off-diagonal conditional proximity matrix. The approach is essentially the same as that given for the symmetric case in the sense that equation (6) is used for calculating successive configurations. It differs, however, in the method by which the appropriate distances δ_{ij} are estimated. Since the entries in a conditional proximity matrix may only be compared meaningfully by rows, it is necessary to find a best fitting monotonic function relating distances in the configuration with proximities for <u>each row separately</u>. Let $\hat{d}_{ij}$ be the value corresponding to d_{ij} obtained by fitting a monotonic function to the $i^{\underline{th}}$ row of the proximity matrix. A new configuration X' may be calculated from X by the formula

$$x_i' = x_i + \frac{\alpha_3}{q} \sum_{i=1}^{q} \left(1 - \frac{\hat{d}_{ij}}{d_{ij}} \right) (x_j - x_i) , \tag{19}$$

where α_3 is a constant and q is the number of columns in the data matrix (in the case of an off-diagonal proximity matrix q may be different from p) or else is p depending on whether x_i represents a column or a row in the matrix. A measure of goodness of fit analogous to Kruskal's stress may be defined by

$$S_4 = \sqrt{\frac{\sum_{i=1}^{p} \sum_{j=1}^{q} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i=1}^{p} \sum_{j=1}^{q} d_{ij}^2}} . \tag{20}$$

By fitting a function to each row separately we have added degrees of freedom to the data and therefore ensured that the configuration's fit to the data will appear better than that which might be obtained with symmetric proximities. Whereas $S_2 < .01$ would be judged as a very good fit with symmetric data, S_4 must be approximately .001 for conditional diagonal proximity matrices and approximately .0001 for conditional off-diagonal matrices before a good fit to the data has been achieved.

A word should be said about the parameter α_3. Experience has shown that when $\alpha_3 = 1$, convergence is much too slow. On the other hand, no good rule for calculating α_3 has been formulated from the inspection of many matrices subjected to the algorithm. In practice it has been found to be satisfactory if the value of α_3 is gradually increased from iteration to iteration until the rate of decrease of S_4 begins to fall off. As subsequent iterations are calculated α_3 is kept as large as possible without allowing S_4 to increase.

AN OUTLINE OF THE GENERAL ALGORITHM

The general algorithm which has been programmed for the IBM 7090 has the following outline.

1. The data matrix is read in along with relevant parameters describing the data and options to be taken by the program.

2. If a starting configuration is to be supplied, it is read in. Otherwise the program calculates the initial configuration using the correlational approach described on pages 16 & 17 above. In the case of an off-diagonal matrix the coordinates of the column points are estimated by setting them equal to those of the row points to which they are closest as suggested[4] by the data matrix.

3. Distances from each row point to each column point are calculated using the Euclidean formula

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}.$$

4. For each row point the best fitting monotonic function is calculated for distances to the column points[5].

5. The quantity S_4 is calculated to evaluate the configuration. If sufficiently low, the procedure is terminated. Otherwise the program continues.

6. A new configuration is calculated using (10) and the program returns to step 3.

If the program exceeds a preset number of iterations or convergence becomes too slow, calculations are terminated, the configuration is rotated

---

[4] The word "suggested" is used intentionally since we are in the awkward position of comparing proximity measures across rows.

[5] For an efficient algorithm to calculate the best fitting monotonic function see Kruskal (1964b, p. 127).

REFERENCES

Atteneave, F. Dimensions of similarity. *Amer. J. Psychol.*, 1950, 63, 516-556.

Coombs, C. H. *A Theory of Data.* New York: Wiley, 1964.

Coombs, C. H. & Kao, R. C. On the connection between factor analysis and multidimensional unfolding. *Psychometrika*, 1960, 25, 219-231.

Dawes, R. M. & Kramer, E. A proximity analysis of vocally expressed emotion. *Perceptual & Motor Skills*, 1966, 22, 571-574.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964, 29, 1-27. (a)

Kruskal, J. B. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 1964, 29, 115-130. (b)

Lingoes, J. C. New computer developments in pattern analysis and nonmetric techniques. Unpublished manuscript, 1965. (a)

Lingoes, J. C. An IBM-7090 program for Guttman-Lingoes smallest space analysis--I. *Behav. Sci.*, 1965, 10, 183-184. (b)

Lingoes, J. C. An IBM-7090 program for Guttman-Lingoes smallest space analysis--RII. *Behav. Sci.*, 1966, 11, 322.

Ross, J. & Cliff, N. A generalization of the interpoint distance model. *Psychometrika*, 1964, 29, 167-176.

Shepard, R. H. The analysis of proximities: multidimensional scaling with an unknown distance function I. *Psychometrika*, 1962, 27, 125-140. (a)

Shepard, R. H. The analysis of proximities: multidimensional scaling with an unknown distance function II. *Psychometrika*, 1962, 27, 219-246. (b)