P. R. Krishnaiah and L. N. Kanal, eds., *Handbook of Statistics, Vol. 2*
©North-Holland Publishing Company (1982) 285–316

13

Theory of Multidimensional Scaling

Jan de Leeuw and Willem Heiser

1. The multidimensional scaling problem

1.1. MDS: broad and narrow

It is difficult to give a precise definition of MDS, because some people use the term for a very specific class of techniques while others use it in a much more general sense. Consequently it makes sense to distinguish between MDS *in the broad sense* and MDS *in the narrow sense*. MDS in the broad sense includes various forms of cluster analysis and of linear multivariate analysis, MDS in the narrow sense represents dissimilarity data in a low-dimensional space.

People who prefer the broad-sense definition want to emphasize the close relationships of clustering and scaling techniques. Of course this does not imply that they are not aware of the important differences. Clustering techniques fit a non-dimensional discrete structure to dissimilarity data, narrow-sense MDS fits a continuous dimensional structure. But both types of technique can be formalized as representing distance-like data in a particular metric space by minimizing some kind of loss function. The difference, then, is in the choice of the target metric space, the structure of the two problems is very much alike. The paper pioneering this point of view is Hartigan's (1967). In fact Hartigan proceeds the other way around, he takes clustering as the starting point and lets clustering in the broad sense include narrow-sense MDS. The same starting point and the same order are more or less apparent in the important review papers of Cormack (1971) and Sibson (1972). An influential broad-sense paper that uses narrow-sense MDS as the starting point is that of Carroll (1976). In fact Carroll discusses techniques that explicitly combine aspects of clustering and narrow-sense MDS, and find mixed discrete/continuous representations. In a very recent paper Carroll and Arabie (1980) propose a useful taxonomy of MDS data and methods which is very broad indeed. Investigators who are less interested in formal similarities of methods and more interested in substantial differences in models have naturally emphasized the choice of the space, and consequently the differences between clustering and narrow-sense MDS. Of course detailed comparison of the two classes of techniques already presupposes a common framework. This is obvious

in the important papers of Shepard (1974) and Shepard and Arabie (1979), who present the two classes of methods essentially as complementing each other. If the emphasis shifts more towards a theory of similarity judgments, the discrete and continuous representations tend to become rivals. In a brilliant paper Tversky (1977) has attacked narrow-sense MDS as a realistic model for psychological similarity, and Sattath and Tversky (1977) have presented the free or additive tree as a more realistic alternative. Krumhansl (1978) has defended dimensional models.

We are not interested, in this paper, in psychological theories of similarity, i.e. in narrow-sense MDS as a miniature psychological theory. We are also not interested in discussing the many forms of cluster analysis and nondimensional scaling. Consequently we propose a definition of scaling which is quite broad (although not as broad as the one suggested by Carroll's taxonomy), and after proposing this definition we quickly specialize to MDS. From that point on there is no more need to distinguish between narrow and broad MDS. Our definition is inspired by the definition of Kruskal (1977), and by the discussion in Kruskal and Wish (1978), and Cliff (1973). We agree with these authors that MDS should be further classified, and that the most important distinctions are between metric and nonmetric MDS, and between two-way and three-way MDS. Other criteria are choice of metric, choice of loss function, and choice of algorithm, but these seem to be more technical and less essential.

1.2. Notation and definitions

In *metric two-way scaling* the data are a pair $\langle I, \delta \rangle$, with I a nonempty set and δ a real valued function on $I \times I$. The elements of I are called *objects* or *stimuli*, the number $\delta(i, j)$ is the *dissimilarity* between objects i and j. A second pair $\langle X, d \rangle$ is also available, with X another nonempty set and d a real valued function on $X \times X$. The elements of X are called *points*, the number $d(x, y)$ is the *distance* between points x and y. The metric two-way scaling problem is to construct a mapping ϕ of I into X in such a way that for all i and j in I the distance $d(\phi(i), \phi(j))$ is approximately equal to the dissimilarity $\delta(i, j)$.

In nonmetric two-way scaling the situation is more complicated. The data do not consist of a single real valued function δ on $I \times I$, but of a set Δ of real valued functions on $I \times I$. If we agree to call an arbitrary real valued function on $I \times I$ a *disparity*, then Δ is the set of all *admissible* disparities. The nonmetric two-way scaling problem is to find a mapping ϕ of I into X *and* an admissible disparity $\hat{d}$ in such a way that for all i and j in I the distance $d(\phi(i), \phi(j))$ is approximately equal to $\hat{d}(i, j)$.

In metric three-way scaling we use an additional set K. The elements of K are called *replications*, or *occasions*, or *individuals*. For each k in K we have a dissimilarity structure $\langle I, \delta_k \rangle$, and the metric three-way scaling problem is to find for each k a mapping ϕ_k from I into X such that $d(\phi_k(i), \phi_k(j))$ is approximately equal to $\delta_k(i, j)$ for all i and j in I. It will be obvious by now how to define the nonmetric three-way scaling problem: there is a set Δ_k of admissible disparities for each k, and we have to construct both the mappings ϕ_k and the disparities $\hat{d}_k$.

A number of additional comments are in order here. We have called $d(x, y)$ the distance of x and y, but we have not assumed that the function d satisfies the usual axioms for a metric. In the same way we have defined disparities as arbitrary functions on $I \times I$, while in most applications disparities (and dissimilarities) are distance-like too. For ease of reference we briefly mention the metric axioms here. For $\langle X, d \rangle$ we must have the following.

(M1) If $x \neq y$, then $d(x, y) > 0$, and $d(x, x) = 0$. (minimality)

(M2) $d(x, y) = d(y, x)$. (symmetry)

(M3) $d(x, y) + d(y, z) \geqslant d(x, z)$. (triangle inequality)

In most applications our target space $\langle X, d \rangle$ does satisfy these axioms, but in some (M1) must be replaced by

(M1′) $d(x, y) \geqslant 0$, and $d(x, x) = 0$. (weak minimality)

Moreover in many other applications we would like to drop symmetry. Tversky (1977) has argued, quite convincingly, that assuming symmetry is not very realistic in many psychological situations. In the same way $\langle I, \delta \rangle$ often satisfies (M1) and (M2), but sometimes only

(M1″) If $x \neq y$, then $d(x, y) > d(x, x) = d(y, y)$. (shifted minimality)

Because of the very many possibilities we do not impose a specific set of axioms about $\langle X, d \rangle$ and/or $\langle I, \delta \rangle$, we simply have to remember that they usually are distance-like.

The expression "approximately equal to" in our definitions has not been defined rigorously. As we mentioned previously the usual practice in scaling is to define a real valued *loss function*, and to construct the mapping of I into X in such a way that this loss function is minimized. In this sense a scaling problem is simply a minimization problem. One way in which scaling procedures differ is that they use different loss functions to fit the same structure. Another way in which they differ is that they use different algorithms to minimize the same loss function. We shall use these technical distinctions in the paper in our taxonomy of specific multidimensional scaling procedures.

It is now very simple to define MDS: a (p-dimensional) multidimensional scaling problem is a scaling problem in which X is $\mathbb{R}^p$, the space of all p-tuples of real numbers. Compare the definition given by Kruskal (1977, p. 296): "We define *multidimensional scaling* to mean any method for constructing a configuration of points in low-dimensional space from interpoint distances which have been corrupted by random error, or from rank order information about the corrupted distances." This is quite close to our definition. The most important difference is that Kruskal refers to 'corrupted distances' and 'random error'. We do not use these notions, because it seems unnecessary and maybe even harmful to commit ourselves to a more or less specific stochastic model in which we assume the existence of a 'true value.' This point is also discussed by Sibson

(1972). Moreover, not all types of deviations we meet in applications of MDS can reasonably be described as 'random error.' This is also emphasized by Guttman (1971).

We have now specified the space X, but not yet the metric d. The most familiar choice is, of course, the ordinary Euclidean metric. This has been used in at least 90% of all MDS applications, but already quite early in MDS history people were investigating alternatives and generalizations. In the first place the Euclidean metric is a member of the family of power metrics, Attneave (1950) found empirically that another power metric gave a better description of his data, and Restle (1959) proposed a qualitative theory of similarity which leads directly to Attneave's 'city block' metric. The power metrics themselves are special cases of general Minkovski metrics, they are also special cases of the general additive/difference metrics investigated axiomatically by Tversky and Krantz (1970). On the other hand Euclidean space is also a member of the class of spaces with a projective (or Cayley–Klein) metric, of which hyperbolic and elliptic space are other familiar examples. In Luneborg (1947) a theory of binocular visual space was discussed, based on the assumption of a hyperbolic metric, and this theory has always fascinated people who are active in MDS. It is clear, consequently, that choice of metric is another important criterion which can be used to classify MDS techniques.

1.3. History of MDS

Multidimensional scaling is quite old. Closely related methods have been used by surveyors and geographers since Gauss, Kruskal has discovered a crude MDS method used in systematic zoology by Boyden around 1930, and algorithms for the mapping of chromosomes from crossing-over frequencies can already be found in an interesting paper of Fisher (1922). The systematic development of MDS, however, has almost completely taken place in psychometrics.

The first important contribution to the theory of MDS, not to the methods, is probably Stumpf's (1880). He distinguishes four basic types of judgments which correspond with four of the eight types of data discussed in the book of Coombs (1964). Stumpf defines psychological distance explicitly as degree of dissimilarity, he argues that reliable judgments of dissimilarity are possible, indicates that judgments about very large and very small dissimilarities are not very reliable, mentions that small distances are often overestimated while large distances are underestimated, and argues that triadic comparisons are much easier than tetradic comparisons [124, pp. 56–65, pp. 122–123, pp. 128–133]. Stumpf's work did not lead to practical MDS procedures, and the same thing is true for later important theoretical work of Goldmeier (1937) and Landahl (1945).

The contributions to the method are due to the Thurstonian school. Richardson (1938) and Klingberg (1941) applied classical psychophysical data collection methods to pairs of stimuli, used Thurstone's 'law of comparative judgment' to transform the proportions to interval scale values, estimated an additive constant to convert the scale values to distance estimates, and constructed coordinates of

the points by using a theorem due to Young and Householder (1938). The Thurstonian methods were systematized by Torgerson in his thesis of 1951, the main results were published in [127]. Messick and Abelson (1956) contributed a better method to estimate the additive constant, and Torgerson summarizes the Thurstonian era in MDS in Chapter 11 of his book (1958).

The first criticisms of the Thurstonian approach are in an unpublished dissertation of Rowan in 1954. He pointed out that we can always choose the additive constant in such a way that the distances are Euclidean. Consequently the Thurstonian procedures tend to represent non-Euclidean relationships in Euclidean space, which is confusing, and makes it impossible to decide if the psychological distances are Euclidean or not. Rowan's work is discussed by Messick (1956). He points out that non-Euclidean data lead to a large additive constant, and a large additive constant leads to a large number of dimensions. As long as the Thurstonian procedures find that a small number of dimensions is sufficient to represent the distances, everything is all right. In the meantime a more interesting answer to Rowan's objections was in the making. In another unpublished dissertation Mellinger applied Torgerson's procedures to colour measurement. This was in 1956. He found six dimensions, while he expected to find only two or three. Helm (1960) replicated Mellinger's work, and found not less than twelve dimensions. But Helm made the important observation that this is mainly due to the fact that large distances tend to be underestimated. If he transformed the distances exponentially he merely found two dimensions, and if he transformed Mellinger's data exponentially he found three. Consequently a large number of dimensions is not necessarily due to non-Euclidean distances, it can also be due to a nonlinear relationship between dissimilarities and distances.

In the meantime the important work of Attneave (1950) had been published. His study properly belongs to 'multidimensional psychophysics', in which the projections of the stimuli on the dimensions are given, and the question is merely how the subjects 'compute' the dissimilarities. In the notation of the previous section the mapping ϕ is given but the distance d is not, while in MDS the distance d is given and the mapping ϕ is not. In this paper Attneave pointed out that other distance measures may fit the data better than the Euclidean distance, and he also compared direct judgments of similarity with identification errors in paired associate learning. He found that the two measures of similarity had a nonlinear but monotonic relationship. In his 1955 dissertation Roger Shepard also studied errors in paired associates learning. The theoretical model is explained in [117], the mathematical treatment is in [116], and the experimental results are in [118]. Shepard found that to a fairly close approximation, distance is the negative logarithm of choice probability, which agrees with a choice theory analysis of similarity judgments by Luce (1961, 1963). Compare also [68]. On the other hand Shepard also found systematic deviations from this 'rational distance function', and concluded that the only thing everybody seemed to agree on was that the function was monotonic.

Also around 1950 Coombs began to develop his theory of data. The main components of this theory are the classification of data into the four basic

quadrants, the emphasis on the geometric interpretation of models and on the essentially qualitative nature of data. All these three aspects have been enormously influential, but the actual scaling procedures developed by Coombs and his associates have not been received with great enthusiasm. The main reason is that they were 'worksheet' methods involving many subjective choices, and that they gave nonmetric representations of nonmetric data. In the case of the one-dimensional unfolding model the derived ordered metric scale turned out to be close to an interval scale, but the multidimensional extensions of Bennett and Hays [6, 54] were much more problematical. The same thing is true for the method developed by Hays to analyze Q–IV–a (dis)similarity data, which is discussed in [24, Part V]. The methods of Hays and Bennett are the first nonmetric MDS methods. But, in the words of Coombs himself, "This method, however, new as it is, may already be superseded by a method that Roger Shepard has developed for analysis of similarities data" [24, p. 494]. Coombs aimed at Shepard's 1962 work.

Another important line of development can be found in the work of Louis Guttman. It is obvious from the book of Coombs (1964) that Guttman's work on the perfect scale, starting with [42] and culminating in [44], has been very influential. On the other hand Guttman's techniques which find metric representations from nonmetric data, discussed in [41, 43, 44], were not used a great deal. It seems that around 1960 Guttman had arrived at approximately the same point as Coombs and Shepard (compare for example [46]). Ultimately this resulted in [47], but the unpublished 1964 version of this paper was also widely circulated.

Although Coombs and his students had studied nonmetric MDS and although Guttman had proposed methods to quantify qualitative data, the real 'computational breakthrough' was due to Roger Shepard [119]. We have already seen that his earlier work on functions, relating dissimilarity to distance, pointed strongly in the direction of merely requiring a monotonic relationship without specifying a particular functional form. In his 1962 papers Shepard showed that an MDS technique could be constructed on the basis of this requirement but (perhaps even more important) he also showed that an efficient computer program could be constructed which implemented this technique. Moreover he showed with a large number of real and synthetic examples that his program could 'recover' metric information, and that the information in rank-ordered dissimilarities was usually sufficient to determine the configuration of points (cf. also [120]). This idea of 'getting something from nothing' appealed to a lot of people, and the idea that it could be obtained by simply pushing a button appealed to even more people. The worksheet methods of Coombs and Hays were quickly forgotten.

Another important effect of the Shepard 1962 papers was that they got Joseph Kruskal interested in MDS. He took the next important step, and in [73, 74] he introduced psychometricians to loss functions, monotone regression, and (gradient) minimization techniques. In these papers Kruskal puts Shepard's ideas, which still had a number of heuristic elements, on a firm footing. He showed, essentially, how any metric psychometric technique could be converted into a nonmetric one by using monotone regression in combination with a least squares

loss function, and he discussed how such functions could be minimized. Early systematizations of this approach are from Roskam (1968) and Young (1972). Torgerson (1965) reported closely related work that he had been doing, and made some thoughtful (and prophetic) comments about the usefulness of the new nonmetric procedures. Guttman (1968) contributed a long and complicated paper which introduced some useful notation and terminology, contributed some interesting mathematical insights, but, unfortunately, also a great deal of confusion. It is obvious now that Kruskal's discussion of his minimization method was rather too compact for most psychometricians at that time. The confusion is discussed, and also illustrated, in [87].

The main contribution to MDS since the Shepard–Kruskal 'computer revolution' is undoubtedly the paper by Carroll and Chang (1970) on three-way MDS. It follows the by now familiar pattern of presenting a not necessarily new model by presenting an efficient algorithm and an effective computer program, together with some convincing examples. A recent paper, following the same strategy, integrates two- and three-way metric and nonmetric MDS in a single program [125].

2. Multidimensional scaling models

2.1. The Euclidean distance model

2.1.1. Euclidean metric two-way scaling

We have already decided to study mappings of $\langle I, \delta \rangle$ into $\langle \mathbb{R}^p, d \rangle$, with $\mathbb{R}^p$ the space of all p-tuples of real numbers. The most familiar way to define a metric in $\mathbb{R}^p$ is as follows. For each x, y in $\mathbb{R}^p$ we define

$$d^2(x, y) = (x - y)'(x - y),$$

or, equivalently,

$$d(x, y) = \|x - y\|,$$

where $\|\cdot\|$ is the Euclidean norm. The metric two-way Euclidean MDS problem is to find a mapping ϕ of I into $\mathbb{R}^p$ such that $\delta(i, j)$ is approximately equal to $\|\phi(i) - \phi(j)\|$. In this section we are interested in the conditions under which this problem can be solved *exactly*, or, to put it differently, under which conditions $\langle I, \delta \rangle$ can be *imbedded* in $\langle \mathbb{R}^p, d \rangle$. This problem is also studied in classical distance geometry. There is some early work on the subject by Gauss, Dirichlet, and Hermite, but the first systematic contribution is the very first paper of Arthur Cayley (1841). Cayley's approach was generalized by Menger (1928) in a fundamental series of papers. Cayley and Menger used determinants to solve the imbedding problem, an alternative formulation in terms of quadratic forms was suggested by Fréchet (1935) and worked out by Schoenberg (1935). The same

result appears, apparently independently, in [141]. The contributions of Cayley, Menger, Fréchet, and Schoenberg are summarized and extended by Blumenthal (1938, 1953). In this paper we prefer the Schoenberg solution to the Menger solution, because it leads more directly to computational procedures.

We suppose throughout this section that $\langle I, \delta \rangle$ is a semi-metric space, by which we mean that δ satisfies both minimality and symmetry. For our first theorem we assume in addition that I is a finite set, say with n elements. We define the $n \times n$ matrix H with elements $h_{ij} = \delta^2(i, j)$, and the $n \times n$ matrix B with elements $b_{ij} = -\frac{1}{2}(h_{ij} - h_{i\cdot} - h_{\cdot j} + h_{\cdot\cdot})$, where dots replacing an index mean that we have averaged over that index. If J is the matrix which centers each n-vector, i.e. $J = I - \frac{1}{n}ee'$, then $B = -\frac{1}{2}JHJ$.

THEOREM 1. *The finite semimetric space $\langle I, \delta \rangle$ can be imbedded in Euclidean p-space if and only if B is positive semi-definite, and rank$(B) \leqslant p$.*

PROOF. Suppose $x_1, \ldots, x_n$ are p-vectors such that $\delta^2(i, j) = d^2(x_i, x_j)$. Define the n-vector a by $a_i = x_i'x_i$, and collect the x_i in the $n \times p$ matrix X. Then $H = ae' + ea' - 2XX'$, and consequently $B = -\frac{1}{2}JHJ = JXX'J$, which implies that B is positive semi-definite (from now on: psd) of $\text{rank}(B) = \text{rank}(JX) \leqslant p$. This proves necessity. Conversely if B is psd and $\text{rank}(B) \leqslant p$, then there is an $n \times p$ matrix X such that $B = XX'$. For this X we have $d^2(x_i, x_j) = b_{ii} + b_{jj} - 2b_{ij} = h_{ij}$.

In fact Schoenberg (1935) and Young and Householder (1938) prove a slightly different version of the theorem. They place the origin in one of the points while our version places the origin in the centroid of the points, which is considerably more elegant from a data analysis point of view. Our formulation is due to Tucker, Green, and Abelson (cf. [128, pp. 254–259]). Theorem 1 can be sharpened by defining $\langle I, \delta \rangle$ to be *irreducibly imbeddable* in Euclidean p-space if it is imbeddable, but not imbeddable in Euclidean $(p-1)$-space. A slight rewording of the proof shows that $\langle I, \delta \rangle$ is irreducibly imbeddable if and only if B is psd and $\text{rank}(B) = p$. A complete solution of the general imbedding problem, in which I can be infinite, was given by Menger (1928). We quote his result without proof; for an excellent discussion we refer to [10, Chapter IV].

THEOREM 2. *The semimetric space $\langle I, \delta \rangle$ can be imbedded in Euclidean p-space if and only if each subset of $p+3$ points can be imbedded in Euclidean p-space.*

The imbedding problem is the first fundamental problem of classical distance geometry. The second fundamental problem is the space problem. The imbedding problem gives conditions under which a semimetric space is metrically congruent to a subset of Euclidean p-space, the space problem gives conditions under which it is metrically congruent to Euclidean p-space itself, i.e. there must be a one–one correspondence between the two spaces. The first solution to the space problem was again given by Menger in 1928, a more elegant solution was found by Wilson

in 1932. The results are summarized in [10, Chapter V], a more recent summary is in [11]. Again we quote the main results without proof. Remember that in a semimetric space $\langle A, \mu \rangle$ the point b is *between* a and c if $\mu(a,b)+\mu(b,c)=\mu(a,c)$. A semimetric space is *convex* if for each pair a, c there is at least one $b \neq a, c$ between a and c, and it is *externally convex* if for each pair a, b there is at least one $c \neq a, b$ such that b is between a and c.

THEOREM 3. *The semimetric space $\langle I, \delta \rangle$ is congruent with Euclidean p-space if and only if it is complete, convex, externally convex, and irreducibly imbeddable in Euclidean p-space.*

To obtain Wilson's characterization we need an additional definition: a semimetric space has the Euclidean four-point property if each quadruple of points can be imbedded in Euclidean 3-space.

THEOREM 4. *The semimetric space $\langle I, \delta \rangle$ is congruent with Euclidean p-space if and only if it is complete, convex, externally convex, it satisfies the Euclidean four-point property, and each set of $p+1$ points defines a singular B-matrix (cf. Theorem 1).*

2.1.2. Euclidean non-metric two-way scaling

Again we suppose that I is finite, with n elements. The *additive constant* problem is to find $x_1, \ldots, x_n$ in $\mathbb{R}^p$ and a real number such that $\delta(i,j)+\alpha(1-\delta^{ij})$ is approximately equal to $d(x_i, x_j)$. Superscripted delta (δ^{ij}) is the Kronecker symbol.

THEOREM 5. *Suppose $\langle I, \delta \rangle$ is a semimetric space, and suppose I is finite with n elements. Then there is an α such that I with semimetric $\delta(i,j)+\alpha(1-\delta^{ij})$ can be imbedded in Euclidean $(n-1)$-space.*

PROOF. Define $H(\alpha)$ by the rule

$$h_{ij}(\alpha)=\delta^2(i,j)+2\alpha\delta(i,j)+\alpha^2(1-\delta^{ij}).$$

and define $B(\alpha)=-\frac{1}{2}JH(\alpha)J$. Clearly $B(\alpha)$ is of the form

$$B(\alpha)=B_0+2\alpha C_0+\tfrac{1}{2}\alpha^2 J.$$

This implies that there is an α_0 such that $b_{ij}(\alpha)<0$ for all $i \neq j$ if $\alpha > \alpha_0$. Because rows and columns of $B(\alpha)$ add up to zero a familiar matrix theorem [126] proves that $B(\alpha)$ is psd with rank $(B(\alpha))=n-1$ for all $\alpha > \alpha_0$. Theorem 1 now gives the required result.

THEOREM 6. *If $\langle I, \delta \rangle$ is a semimetric space, I is finite with n elements, and $\langle I, \delta \rangle$ cannot be imbedded in Euclidean $(n-1)$-space, then there is an α such that I with semimetric $\delta(i,j)+\alpha(1-\delta^{ij})$ can be imbedded in Euclidean $(n-2)$-space.*

PROOF. Consider $\lambda(\alpha)$, the minimum of $x'B(\alpha)x$ over all x satisfying $x'x=1$ and $x'e=0$. Clearly $\lambda(\alpha)$ is continuous. By hypothesis $\lambda(0)<0$, and the proof of Theorem 5 shows that $\lambda(\alpha)>0$ for $\alpha>\alpha_0$. Consequently $\lambda(\alpha)=0$ for some α between 0 and α_0.

These two theorems improve the (unpublished) results of Rowan we mentioned in Subsection 1.3. Another one-parameter family of transforms was used by Guttman (cf. [86]).

THEOREM 7. *Suppose $\langle I, \delta\rangle$ is a semimetric space, and suppose I is finite with n elements. Then there is an α such that I with the semimetric $[\{\delta^2(i,j)+\alpha(1-\delta^{ij})\}^{1/2}]$ can be imbedded in Euclidean $(n-2)$-space.*

PROOF. In this case $B(\alpha)=B_0+\frac{1}{2}\alpha J$, and $\lambda(\alpha)=\lambda(0)+\frac{1}{2}\alpha$. Thus if $\alpha=-2\lambda(0)$, then $\lambda(\alpha)=0$, and $B(\alpha)$ is psd of rank $n-2$.

The transformation considered in Theorem 7 is clearly monotone. Consequently the following theorem is an easy corrollary.

THEOREM 8. *Suppose $\langle I, \delta\rangle$ is a semimetric space, and suppose I is finite with n elements. Then there is a semimetric $\bar{\delta}$ such that $\bar{\delta}(i,j)\leqslant\bar{\delta}(i',j')$ if and only if $\delta(i,j)\leqslant\delta(i',j')$ for all $i\neq j$ and $i'\neq j'$, and such that $\langle I, \bar{\delta}\rangle$ can be imbedded in Euclidean $(n-2)$-space.*

It was already pointed out by Lingoes (1971) that the mapping constructed in the proof of Theorems 7 and 8 may lead to $\bar{\delta}(i,j)=0$ for some $i\neq j$. This means that in general the conclusion of Theorem 8 cannot be strengthened to $\bar{\delta}(i,j)\leqslant\bar{\delta}(i',j')$ if and only if $\delta(i,j)\leqslant\delta(i',j')$ for all i,j,i',j'. The precise conditions under which such a strengthening is possible were investigated by Holman (1972). His work is based partly on unpublished work of the distance geometer L. M. Kelly. In the first place a semimetric space $\langle I, \delta\rangle$ is called an *ultrametric space* if

$$\delta(i,j)\leqslant\max\{\delta(i,k),\delta(k,j)\}$$

for all i,j,k. This ultrametric inequality, which obviously implies the triangle inequality, is interesting from a data analysis point of view, because it is well known that a semimetric space can be imbedded in a hierarchical tree structure if and only if the semimetric satisfies the ultrametric inequality (cf. for example [62]). Hierarchical tree structures are very important in clustering and classification literature [25]. We now give the two theorems proved by Holman.

THEOREM 9. *Suppose $\langle I, \delta\rangle$ is an ultrametric space, and suppose I is finite with n elements. Then $\langle I, \delta\rangle$ can be imbedded in Euclidean $(n-1)$-space, but not in Euclidean $(n-2)$-space.*

Because the ultrametric inequality remains true if we transform the δ_{ij} monotonically, this also implies that no strictly monotone transformation of the δ_{ij} can be imbedded in Euclidean $(n-2)$-space.

THEOREM 10. *Suppose $\langle I, \delta\rangle$ is a semimetric space, and suppose I is finite with n elements. Then there is a semimetric $\bar{\delta}$ such that $\bar{\delta}(i, j) \leqslant \bar{\delta}(i', j')$ if and only if $\delta(i, j) \leqslant \delta(i', j')$ for all i, j, i', j' and such that $\langle I, \bar{\delta}\rangle$ can be imbedded in Euclidean $(n-2)$-space if and only if $\langle I, \delta\rangle$ is not an ultrametric space.*

For proofs of Theorems 9 and 10 we refer to Holman. Another interesting nonmetric MDS problem is the *missing data* problem. Suppose I is finite again, and suppose δ is defined only on a subset L of $I \times I$ (and is not defined on $\bar{L}$). We suppose that (i, i) is in L for all i, and that (i, j) is in L if and only if (j, i) is in L. Define a matrix H_0, which has elements (i, j) equal to $\delta^2(i, j)$ if (i, j) is in L, and equal to zero otherwise. For each (i, j) in $\bar{L}$ we define a matrix A_{ij}, which has elements (i, j) and (j, i) equal to $+1$, and all other elements equal to zero. Define

$$H(\theta) = H_0 + \left\{ \sum \theta_{ij} A_{ij} \mid (i, j) \in \bar{L} \right\},$$
$$B(\theta) = B_0 + \left\{ \sum \theta_{ij} T_{ij} \mid (i, j) \in \bar{L} \right\},$$

where $B_0 = -\frac{1}{2} J H_0 J$ and $T_{ij} = -\frac{1}{2} J A_{ij} J$ as usual. The following theorem is rather trivial but is given because it has computational consequences.

THEOREM 11. *Suppose $\langle I, \delta\rangle$ is a semimetric space with missing elements, suppose I is finite. Then $\langle I, \delta\rangle$ can be imbedded in Euclidean p-space if and only if there exist $\theta_{ij}, (i, j) \in \bar{L}, \theta_{ij} > 0$, such that $B(\theta)$ is psd and* $\operatorname{rank}(B(\theta)) \leqslant p$.

A special missing data problem is the *metric unfolding* problem. Here I is partitioned into the finite sets I_1 and I_2, and $L = I_1 \times I_2$. Because of the special structure of this problem a more precise analysis is possible than the one given in Theorem 11, for the results we refer to papers from Schönemann (1970), Gold (1973), and Heiser and De Leeuw (1978). A computationally more convenient version of Theorem 8 can also be formulated along the lines of Theorem 11. Define for each $i \neq j$ a matrix A_{ij} in which elements (i', j') and (j', i') are equal to $+1$ if $\delta(i', j') \geqslant \delta(i, j)$, and equal to zero otherwise. Define $T_{ij} = -\frac{1}{2} J A_{ij} J$ and $B(\theta) = \sum \theta_{ij} T_{ij}$.

THEOREM 12. *Suppose $\langle I, \delta\rangle$ is a semi-metric space, with I finite. Then there exists a semimetric $\bar{\delta}$ such that for all $i, j, i', j', \delta(i, j) \leqslant \delta(i', j')$ implies $\bar{\delta}(i, j) \leqslant \bar{\delta}(i', j')$ and such that $\langle I, \bar{\delta}\rangle$ can be imbedded in Euclidean p-space if and only if there exist nonnegative numbers θ_{ij} such that $B(\theta)$ is psd, and* $\operatorname{rank}(B(\theta)) \leqslant p$.

Up to now we have only analyzed nonmetric MDS in cases in which I was finite. It is clear, however, that we can combine the results in this section with

Theorem 2 from the previous subsection to find solutions to the general imbedding problem. The space problem is somewhat more complicated. The most interesting case is the one in which the elements of $I \times I$ are ordered. We want to study the conditions under which this order can be considered to be induced by a convex metric. This topic was studied by Beals and Krantz (1967), their treatment was improved by Krantz (1968), explained for psychologists by Beals, Krantz, and Tversky (1968), and generalized by Lew (1975). Minimality and symmetry can easily be defined in terms of the order, they are obviously necessary conditions for the representation of the order by any metric, convex or not. A number of 'technical' assumptions is also needed, which state that the space is continuous, without holes. They are usually untestable on empirical data. The most important assumption is based on an ordinal characterization of betweenness. Suppose i_1, i_2 and i_3 are three distinct objects in I. Following Beals and Krantz we define $\langle i_1 i_2 i_3 \rangle$ if and only if

(a) $(i_1, i_2') \leqslant (i_1, i_2)$ and $(i_1, i_3') \geqslant (i_1, i_3)$ imply $(i_2', i_3') \geqslant (i_2, i_3)$.

(b) if the conditions of (a) hold, and $(i_2', i_3') = (i_2, i_3)$, then $(i_1, i_2') = (i_1, i_2)$ and $(i_1, i_3') = (i_1, i_3)$.

Moreover we define $(i_1 i_2 i_3)$ if and only if both $\langle i_1 i_2 i_3 \rangle$ and $\langle i_3 i_2 i_1 \rangle$. The basic assumption is that if $0 < (i_1, i_2') < (i_1, i_3)$ then there is an i_2 in I such that $(i_1, i_2) = (i_1, i_2')$ and $(i_1 i_2 i_3)$. The conditions can be illustrated by drawing 'iso-similarity contours' as in [5]. Under the assumptions we have stated there exists a metric d on I such that

(a) $(i, j) \leqslant (i', j')$ if and only if $d(i, j) \leqslant d(i', j')$.

(b) $(i_1 i_2 i_3)$ if and only if $d(i_1, i_2) + d(i_2, i_3) = d(i_1, i_3)$.

To get to Euclidean space we can use Theorem 3 or 4. Completeness and external convexity can easily be defined in terms of the order relation and the betweenness relation. Because the convex metric constructed by Beals and Krantz is unique up to scale we can simply use it to test the Euclidean four point property (or the weaker versions of the property discussed by Blumenthal (1975)). A more direct approach is also possible. Blumenthal (1938, pp. 10–13) discusses an axiomatization of three-dimensional Euclidean geometry due to the Italian geometer M. Pieri. The single undefined elements are 'points', the single primitive relation 'i_1 is equally distant from i_2 and i_3'. This axiomatization, published in 1908, can easily be generalized to p-dimensional space.

Holman's Theorem 9 can be interpreted as a negative result, which shows that ultrametrics cannot be represented in low-dimensional Euclidean spaces. It is well known that ultrametric and tree distances are closely related to 'city block' or l_1-distances. In this sense the counter examples presented by Lew (1978) generalize Holman's results. Lew proves that the p-dimensional 'city block' spaces l_1^p cannot be imbedded into finite-dimensional Euclidean space, and that there is no monotone transform of the metric which makes such an imbedding possible. On the other hand it has been shown by Schoenberg (1937, 1938) that if $\delta(i, j)$ is the l_1-metric, then $\{\delta(i, j)\}^\gamma$, with $0 < \gamma \leqslant \frac{1}{2}$, can be imbedded into l_2, the natural infinite-dimensional generalization of Euclidean space. Lew (1978) also presents other interesting results based on Schoenberg's metric transform theory.

2.1.3. Euclidean three-way scaling

In our discussion of three-way scaling we restrict ourselves to the metric case with finite I. It will be clear from the previous sections how to extend at least some of the results to more general cases. In the major models that have been proposed for Euclidean three-way scaling we require that the mappings ϕ_k of I into $\mathbb{R}^p$ are of the form $\phi_k(i) = T_k x_i$, where $x_1, \ldots, x_n$ are elements of $\mathbb{R}^p$, and where $T_1, \ldots, T_m$ are $p \times p$ matrices. There are three special cases to be considered. We use names suggested by Carroll and Chang (1970), Carroll and Wish (1974), and Harshman (1972). The names are actually names of computer programs, but it is common practice to use them for the models as well. In IDIOSCAL the T_k are unrestricted, in PARAFAC the T_k must be of the form $T_k = W_k S$ with W_k diagonal and S unrestricted, and in INDSCAL we must have $T_k = W_k$ with W_k diagonal. As in the previous sections we define the matrices E_k with elements $\delta_k^2(i, j)$ and $B_k = -\frac{1}{2} J E_k J$. We also collect the x_i in the $n \times p$ matrix X, and we suppose (without loss of generality) that X is centered, i.e. $JX = X$. We study the conditions under which IDIOSCAL and INDSCAL can be fitted exactly. The results are due mainly to Schönemann (1972), but they have been simplified considerably by De Leeuw and Pruzansky (1978). Unfortunately it is not possible to give an equally satisfactory algebraic analysis of the PARAFAC model.

THEOREM 13. *The semimetric spaces* $\langle I, \delta_k \rangle$ *have an irreducible IDIOSCAL imbedding in Euclidean p-space if and only if*

(a) B_k *is psd,*

(b) $B_* = \sum B_k$ *has* rank p.

PROOF. We have to solve $B_k = X C_k X'$, with $C_k = T_k T_k'$. It follows directly that (a) is necessary. If $\text{rank}(B_*) = r < p$, then there is an IDIOSCAL imbedding in r-space, and consequently (b) is necessary for irreducibility. To prove sufficiency we identify the system (partially) by requiring that C_*, the sum of the C_k, is the identity matrix. Thus we have to solve $B_* = XX'$. Suppose $B_* = K\Lambda^2 K'$ is the canonical form of B_*, with Λ^2 the diagonal $p \times p$ matrix of eigenvalues. It follows that $X = K\Lambda L'$, with L square orthonormal but otherwise arbitrary. We now have to solve $B_k = K\Lambda L' C_k L \Lambda K'$ for C_k. Because $KK' B_k KK' = B_k$, for all k, the solution is simply $C_k = L\Lambda^{-1} K' B_k K \Lambda^{-1} L'$. Observe that indeed $C_* = I$.

THEOREM 14. *The semimetric spaces* $\langle I, \delta_k \rangle$ *have an irreducible INDSCAL imbedding in Euclidean p-space if and only if*

(a) B_k *is psd,*

(b) $\text{rank}(B_*) = p$,

(c) $B_k B_*^+ B_l = B_l B_*^+ B_k$ *for all* $k, l = 1, \ldots, m$.

PROOF. For the irreducible INDSCAL imbedding we can also require without loss of generality that $W_*^2 = I$. Moreover, we are only interested in X of full column rank p. By using exactly the same proof as in Theorem 13 we find that an INDSCAL solution exists if and only if L can be chosen in such a way that

$L\Lambda^{-1}K'B_kK\Lambda^{-1}L'$ is diagonal for each k, which is possible if and only if the $\Lambda^{-1}K'B_kK\Lambda^{-1}$ commute. This is true if and only if (c) is true.

It is possible in INDSCAL that solutions exist in which X is not of full column rank, and which are irreducible in the sense that no INDSCAL solution exists with a smaller value of p. In fact it is possible in some cases that INDSCAL solutions only exist for some $p > n$. This is one of the main reasons that INDSCAL is very interesting from a theoretical point of view; the other reason is that INDSCAL often gives unique solutions (up to a permutation of the dimensions), even if $p > n$. Thus there is no 'rotational indeterminacy'. We give a uniqueness theorem for the case in which X is of full column rank, the more general case is studied by Kruskal [75, 76].

THEOREM 15. *An irreducible INDSCAL imbedding in Euclidean p-space is unique if and only if for each $s \neq t$ there is a k such that $w_{ks}^2 \neq w_{kt}^2$.*

PROOF. The rotation matrix L in Theorem 14 is uniquely determined if and only if there is at least one linear combination of the $\Lambda^{-1}K'B_kK\Lambda^{-1}$ with different eigenvalues, which is true if and only if there is at least one linear combination of the W_k^2 with different diagonal values. This is equivalent to the condition in the theorem.

2.1.4. Asymmetries in Euclidean MDS

One of the main objections of Tversky (1977) against the use of metric dimensional models in psychology is that dissimilarity is very often not symmetric. This is not very convincing. In distance geometry it has been recognized from the start that the symmetry axiom may be too restrictive in some applications. Busemann (1955, p. 4) gives the familiar example that a distance downhill may be shorter than 'the same' distance uphill. Busemann's students Zaustinsky, Phadke, and Featherstone have generalized much of the theory of G-spaces to asymmetric metrics (cf. [14]).

In Euclidean MDS several people have proposed asymmetric modifications of the Euclidean distance. An early one by Kruskal (personal communication, 1973) is

$$d^2(\phi(i),\phi(j)) = \sum_{s=1}^{p} \left((x_{is} - x_{js}) + z_s\right)^2$$

where z is called the 'slide vector'. A related 'jet stream' model appears in [40] where Gower also has an interesting 'cyclone' model. Gower (1977), and Constantine and Gower (1978) also discuss MDS techniques which decompose a matrix in its symmetric and antisymmetric parts, and then compute the singular value decomposition of both parts. Baker (cf. [3]) has proposed

$$d^2(\phi(i),\phi(j)) = \sum_{s=1}^{p} w_{is}^2(x_{is} - x_{js})^2.$$

This ASYMSCAL model is incorporated as one of the options in the ALSCAL-4 computer program, which also has corresponding three-way asymmetric models [140].

Another idea, which is already quite old, is that dissimilarity is really symmetric (and even Euclidean), but it is contaminated by response bias or other uphill–downhill effects. We merely need a procedure to remove the response bias, and we can use the ordinary Euclidean model again. The procedures of Kruskal, Baker, and Gower are of the same type, but it is not immediately obvious how the corrections should be carried out. In the Shepard–Luce model for confusion matrices [88, 116] the corrections are very simple. We suppose that the confusion probabilities π_{ij} satisfy a model of the form $\pi_{ij} = \alpha_i \beta_j \eta_{ij}$, with η_{ij} symmetric. This is closely related to work in the contingency table literature on 'quasi-symmetry' [61] and on social mobility tables [49, Chapter VI]. It is also related to a recent proposal of Krumhansl (1978), who responds to Tversky's criticisms with the simple equation

$$\bar{d}(\phi(i), \phi(j)) = d(\phi(i), \phi(j)) + a\psi(i) + b\psi(j),$$

especially if we consider the fact that Shepard and Luce extend their model by supposing that $d(\phi(i), \phi(j)) = -\ln \eta_{ij}$. For recent extensions of the Shepard–Luce models we refer to work by Nakatani (1972), and Townsend (1978). Models of this form can be used to find maximum likelihood estimates of the quasi-symmetry parameters, and in the complete specification of the MDS-coordinates. De Leeuw and Heiser (1979) propose a variety of probability models and computational techniques for these 'discrete interaction matrices'.

2.1.5. Restricted Euclidean MDS

In restricted Euclidean MDS not all maps ϕ of I into $\mathbb{R}^p$ are feasible. If $x_i = \phi(i)$ and the x_i are collected in the $n \times p$ matrix X, then general restrictions can be written as $X \in \Omega$, with Ω a subset of $\mathbb{R}^{np}$, the space of all $n \times p$ matrices. Various types of linear and nonlinear restrictions are possible [7, 8, 33], but we discuss what seems to be the most important type of restrictions. There are some obvious relationships between MDS and principal component analysis (PCA). One way of formulating the relationship of the two techniques is that MDS fits distances to data, while PCA fits inner products. PCA has been extended to factor analysis and, more generally, to 'structural analysis' of covariance matrices (a recent review is given by Jöreskog, 1978). Similar extensions of MDS are possible, and probably useful [33].

The restrictions we discuss are of the form $x_i = Ty_i$, where T and the y_i may be specified in various different ways. Observe the relationship with three-way scaling where we used $\phi_k(i) = T_k y_i$. The first special case, important in multidimensional psychophysics, has y_i known and T unknown and unrestricted. The y_i can be collected in an $n \times q$ matrix Y, T is a $p \times q$ matrix. Without loss of generality we can require that $JY = Y$, and that $Y'Y = I$. We can now apply Theorem 1 to derive an imbedding theorem.

THEOREM 16. *Suppose* $\langle I, \delta \rangle$ *is a finite semimetric space,* Y *is an* $n \times q$ *matrix such that* $JY = Y$ *and* $Y'Y = I$. *Then there exists a* $p \times q$ *matrix* T *such that* $d(Ty_i, Ty_j) = \delta(i, j)$ *if and only if*

(a) $B = -\frac{1}{2}JEJ$ *is psd,*

(b) $\text{rank}(Y'BY) \leqslant p$,

(c) $\text{rank}(B + YY') = \text{rank}(B)$.

The proof is easy matrix algebra. In a principal component context this model has been discussed by Carroll, Green, and Carmone (1976). It has been extended to three-way PCA by Carroll and Pruzansky (1977). Various applications, in which for example the matrix Y is an ANOVA-type design matrix, are also discussed in these papers. A more complicated class of restrictions uses $x_i = Ty_i$ in combination with T diagonal. This can be used to fit simplexes and circumplexes. Cases in which X is partially restricted and partially free can be used to build MDS versions of common factor models.

If T is diagonal and Y is binary (zero–one) a special interpretation of the Euclidean distance model is possible. Suppose $P = \{1, \ldots, p\}$. If S is a subset of P and t is the p-vector with the diagonal elements of T, then we can define

$$\mu(S) = \sum \{t_s \mid s \in P\}.$$

If $S_1, \ldots, S_n$ are the subsets of P defined by

$$S_i = \{s \mid y_{is} = 1\},$$

then

$$d^2(x_i, x_j) = \mu(S_i \triangle S_j),$$

with $\triangle$ the symmetric difference. The subset model has been studied in distance geometry by Kelly (1968, also 1975 and the references given there), and in MDS by Shepard and Arabie (1979). For special systems of subsets we recover the additive tree models of Bunemann (1971), Cunningham (1978), Sattath and Tversky (1977). For even more special systems we find the hierarchical trees of Johnson (1967) and many, many others.

2.2. *Non-Euclidean models*

2.2.1. *Additive difference models*

In multidimensional psychophysics we suppose that the set of objects I has product structure, i.e. $I = I_1 \times \cdots \times I_p$, and $i \in I$ can be written as the p-tuple $(i_1, \ldots, i_p)$. In classical multidimensional psychophysics the I_s are sets of real numbers, and we suppose that

$$\delta(i, j) = \sum_{s=1}^{p} |i_s - j_s|,$$

[2], or

$$\delta^2(i,j)=\sum_{s=1}^{p}(i_s-j_s)^2,$$

[128]. This approach has been generalized by Tversky (1966), whose work is also discussed in [5], and is improved and generalized in [134]. Now the I_s do not have to be sets of real numbers anymore, they do not even have to be ordered sets. We suppose that there are real valued functions ψ_s on I_s, increasing functions $\chi_s:\mathbb{R}\to\mathbb{R}$, and an increasing function $F:\mathbb{R}\to\mathbb{R}$, such that

$$\delta(i,j)=F\left\{\sum_{s=1}^{p}\chi_s(|\psi_s(i_s)-\psi_s(j_s)|)\right\}.$$

This is called the additive difference model. Tversky gives necessary and sufficient conditions in terms of the dimensions of the product structure and the order relation on $I\times I$ which must be satisfied for an additive difference representation. It is also proved that the ψ_s are interval scales, and the χ_s are interval scales with a common unit. Of course an additive difference model does not necessarily define a metric. The additive difference representation is said to be compatible with a metric with additive segments if the representation satisfies the assumptions of Beals and Krantz (1967), i.e., if the order on $I\times I$ also defines a convex metric. Krantz and Tversky (1970) prove the very satisfactory result that compatibility proves that there is an $r\geqslant 1$ such that

$$\delta^r(i,j)=\sum_{s=1}^{p}|\psi_s(i_s)-\psi_s(j_s)|^r.$$

In other words, the only additive difference models compatible with a convex metric are the power metrics. Tests of the additive difference theory have been carried out by Tversky and Krantz (1969), Wender (1971), Krantz and Tversky (1975), and Schönemann (1978).

2.2.2. Minkovski geometry

Power metrics are already mentioned by Torgerson (1958, p. 294), but they did not become popular in psychometrics until Kruskal (1964), Shepard (1964) and Cross (1965). Imbedding of finite semimetric spaces in Euclidean space is investigated by transforming squared distances to scalar products. There is no scalar product associated with power metrics, and, as a consequence, there is no imbedding theory for finite sets. There are some results for the city block case in [35], but Eisler supposes that the order in which the points project on the dimensions is known.

There are more results in the infinite case. We have already discussed the elegant characterization of the power metrics by Krantz and Tversky, but this

supposes that the objects have a product structure. If only the metric, or an order on $I \times I$, is given, we have to follow a different route. We can use results of Andalafte and Blumenthal (1964) that characterize Banach spaces in the class of complete convex metric spaces, by using ordinal properties of the metric only. A Minkovski space is a finite dimensional Banach space in which power metrics can be characterized by using homogeneity. Another possibility is to characterize Minkovski spaces in the class of straight G-spaces, as in [13, pp. 144–163], by using this theory of parallels. In both cases we simply have to add some qualitative axioms to the ones given by Beals and Krantz (1967), the additional axioms are 'testable' and not 'technical' in the sense of Beals, Krantz, and Tversky (1968). Both Anadalafte and Blumenthal (1964) and Busemann (1955) list additional simple qualitative properties which characterize Euclidean space in the class of Minkovski spaces.

2.2.3. Classical non-Euclidean geometries

We have already seen earlier in this paper that Luneborg's theory of binocular vision has inspired psychometricians (for example Indow, 1975, with references) to look at hyperbolic space. For the classical non-Euclidean spaces far more interesting results are available then for Minkovski spaces. In fact Blumenthal (1953) concentrates throughout his book on imbedding semimetric spaces in hyperbolic, elliptic, spherical, and Euclidean spaces. He uses analogues of the Cayley–Menger determinant, but Schoenberg (1935) gives a quadratic form result for spherical space, and Valentine (1969) gives a quadratic form result for hyperbolic space. The relationships between the elementary spaces are studied, using matrices and quadratic forms, in [114]. In MDS Indow and his collaborators have used ad hoc methods to find imbeddings in hyperbolic space, Pieszko (1975) has constructed a method to imbed in Riemannian spaces of non-constant curvature. Lindman and Caelli (1978) criticized Pieszko's work and proposed a different algorithm. We give versions of the theorems of Valentine and Schoenberg, which make them maximally like the Euclidean Theorem 1. We write $S_{p,\rho}$ for the p-dimensional spherical space of radius ρ.

THEOREM 17. *The finite semimetric space* $\langle I, \delta \rangle$ *can be imbedded in* $S_{p,\rho}$ *if and only if*

(a) $\delta(i, j) \leqslant \pi\rho$,

(b) *the matrix* B *with elements* $b_{ij} = \cos \delta(i, j)/\rho$ *is psd, and*

(c) $\text{rank}(B) \leqslant p + 1$.

For p-dimensional hyperbolic space $H_{p,\rho}$ the situation is even more like Euclidean space. We define the matrix H with elements $h_{ij} = \cosh \delta(i, j)/\rho^2$, and the matrix B by $b_{ij} = 1 - (h_{ij} h_{..}/h_{i.} h_{.j})$, where dots are averages again.

THEOREM 18. *The finite semimetric space* $\langle I, \delta \rangle$ *can be imbedded in* $H_{p,\rho}$ *if and only if* B *is positive semi-definite, and* $\text{rank}(B) \leqslant p$.

Congruence orders of $S_{p,\rho}$ and $H_{p,\rho}$, i.e. the analogues of Theorem 2, are studied in Blumenthal (1953). The theorem is also true in these spaces, they also have congruence order $p+3$. For elliptic space the situation is more complicated. For imbedding results and congruence order results we refer to Blumenthal [10, Chapters IX–XI], and the more recent review of Seidel [115]. Because of the developments of non-Euclidean geometry at the end of the previous century and in the beginning of this century we know many ways to solve the space problem for these geometries. Not all of them are useful for our purposes, however. Busemann discusses many metric characterizations. The most convenient one is given in the following Theorem [13, p. 331].

THEOREM 19. *If each bisector $B(a, a')$ (i.e. the locus $xa = xa'$) of a* G*-space R contains with any two points x, y at least one segment $T(x, y)$, then the space is Euclidean, hyperbolic, or spherical of dimension greater than* 1.

Again it is easy to add the flatness of the bisectors to the axioms of Beals and Krantz (1967) for G-spaces.

3. Multidimensional scaling algorithms

3.1. Least squares on the scalar products

3.1.1. Metric two-way MDS

Theorem 1 suggests that a reasonable way to solve the metric Euclidean two-way scaling problem is to minimize the loss function

$$\sigma_1(X) = \operatorname{tr}(B - XX')^2$$

over all X in $\mathbb{R}^{np}$. Suppose $\lambda_1 \geqslant \cdots \geqslant \lambda_n$ are the ordered eigenvalues of B, let $\bar{\lambda}_s = \min(0, \lambda_s)$, and let k_s be an eigenvector corresponding with λ_s; the k_s corresponding with equal eigenvalues are chosen to be orthogonal.

THEOREM 20

$$\min\{\sigma_1(X) \mid X \in \mathbb{R}^{np}\} = \sum_{s=1}^{p} \bar{\lambda}_s^2 + \sum_{s=p+1}^{n} \lambda_s^2.$$

Moreover, the minimum is attained if we set column s of X equal to

$$k_s(\lambda_s - \bar{\lambda}_s)^{1/2}, \quad s = 1, \ldots, p.$$

A proof is given, for example, by Keller (1958). This theorem justifies the classical scaling method explained most extensively by Torgerson (1958). Observe that the last columns of X are equal to zero if B does not have p positive eigenvalues. This metric scaling procedure has three major advantages compared with other competing ones. In the first place we know how to compute eigenvectors and eigenvalues precisely and efficiently. In the second place the solutions are nested in the sense that the solution for q dimensions is contained in the solution for p dimensions if $q < p$. And finally, we are sure that we find the global minimum of the loss function, and not merely a local minimum. The disadvantage is that the procedure of computing B only makes sense in the Euclidean case. We can use Theorems 17 and 18 to construct metric scaling procedures for spherical and hyperbolic geometry, but they use another definition of B. If we cannot transform to scalar products, as in the Minkovski case, then these procedures cannot be used. Another disadvantage in the Euclidean and hyperbolic case is that the elements of B are not independent if the elements of E are independent, which makes unweighted least squares look bad. Moreover, the method loses much of its computational appeal if there are missing data, while this does not bother other methods.

Carroll, Green, and Carmone (1976) have already pointed out that the simple scaling procedure can be generalized if we have linear restrictions of the form $X = YT'$, with $Y'Y = I$ and Y known. The loss function $\sigma_1(X)$ becomes $\sigma_1(S) = \text{tr}(B - YSY')^2$, with $S = T'T$. If we define $\bar{S} = Y'BY$, then we can write $\sigma_1(S) = \text{tr}(B - Y\bar{S}Y')^2 + \text{tr}(\bar{S} - S)^2$ which we can minimize by minimizing $\text{tr}(\bar{S} - T'T)^2$ over T, using the least squares result of Keller (1958) again, as we did in Theorem 20. The asymmetric 'slide vector' model can also be fitted by using the same matrix methods.

3.1.2. Nonmetric two-way MDS

Theorem 5 suggests an interesting way to solve the additive constant problem [106]. We have to minimize

$$\sigma_1(X, \alpha) = \text{tr}\{B(\alpha) - XX'\}^2$$

over all X in $\mathbb{R}^{np}$, and over α. Define the minimum of $\sigma_1(X, \alpha)$ over X for fixed α as $\zeta(\alpha)$. Then, in the same way as in Theorem 20,

$$\zeta(\alpha) = \sum_{s=1}^{p} \bar{\lambda}_s^2(\alpha) + \sum_{s=p+1}^{n} \lambda_s^2(\alpha).$$

This is a function of the single real parameter α, which can be minimized efficiently in a number of ways. It is clear that the approach generalizes without further complications to any problem in which we have a one-parameter family of matrices $B(\alpha)$. By using the theory of lambda-matrices [81] we can in all cases construct efficient algorithms. This fact was used by Critchley (1978) in his

alternative approach to MDS. In our notation he does not optimize a criterion which depends on the choice of the dimensionality p, but he suggests to maximize the variance of the $\lambda(\alpha)$. This usually makes a good fit in a low dimensionality possible. The approach in this section can obviously be generalized to multiparameter problems. Theorems 11 and 12 suggest defining $\sigma_1(X,\theta)$ and $\zeta(\theta)$. Nonmetric Euclidean MDS can be formulated as minimization of these two functions. In unpublished research De Leeuw, Takane, and Young have applied alternating least squares to $\sigma_1(X,\theta)$. Each iteration has two steps. In step 1 we minimize $\sigma_1(X,\theta)$ over X with θ fixed at its current value. This is a partial eigen-problem. In later iterations a very good initial estimate is available, and we consequently use simultaneous iteration [105] to compute the vectors. In step 2 we minimize $\sigma_1(X,\theta)$ over θ with X fixed at its current value. This is a linear least squares problem. Often the vector θ is restricted to be nonnegative. We use a fast iterative quadratic programming routine using the good initial estimate. The complete algorithm, called INDISCAL, is extremely fast, especially if θ has only a few elements (few missing data, or many ties in ordinal data). If we want a very precise solution we can use Newton's method in the final iterations. Alternatively we can also use Critchley's criterion in multiparameter problems.

The major disadvantage of these procedures is that they may converge to non-global minima. Moreover, it is usually the case that $B(\theta)$ or $B(\alpha)$ has some negative eigenvalues at the optimum, which may be undesirable in some applications. An alternative approach is to require that $B(\theta)$ is psd, and to maximize the sum of the first p eigenvalues $\lambda(\theta)$. In missing data problems or ordinal MDS problems this amounts to maximizing a convex function on a convex set. In [29] some ways of solving this problem have been suggested. By approximating the convex set by a sequence of convex polyhedra we can approximate the global minimum of the loss function in this case (but the algorithm is very expensive). In ordinal MDS (Theorem 12) the matrix $B(\theta)$ is of the form $\Sigma\theta_p T_p$. Now minimizing $\zeta(\theta)$ is easy enough, we simply set $\theta = 0$. Thus in this case the problem must be *normalized*. If we minimize $\zeta(\theta)$, we require $\operatorname{tr} B^2(\theta) = 1$. If we maximize the sum of the first p eigenvalues, we require $\operatorname{tr} B(\theta) = 1$.

The metric unfolding algorithm of Schönemann (1970) also belongs in this section. It is based on an algebraic analysis of the metric unfolding model, which has been clarified further by Gold (1973). Unfortunately, numerical experiments of Heiser and De Leeuw (1978, 1979) indicate that Schönemann's algorithm does not work very well. As a matter of fact, even the best metric unfolding methods do not work very well. Nonmetric unfolding methods do not work at all.

3.1.3. Three-way MDS

The theory in Subsection 2.1.3 suggests that we minimize

$$\sigma_1(X;C_k) = \sum_{k=1}^{m} \operatorname{tr}(B_k - XC_kX')^2$$

over X and $C_1,\ldots,C_m$. In the IDIOSCAL case there are no further restrictions on

the C_k, in the INDSCAL case we require them to be diagonal. In [17,50] an alternating least squares algorithm was proposed for the INDSCAL model. A slightly different ALS method was proposed and implemented for the IDIOSCAL model by Kronenberg and De Leeuw (1978). The algorithms have two substeps in each iteration: in the first substep we minimize $\sigma_1(X; C_k)$ over the C_k with X fixed at its current value, in the second substep we minimize $\sigma_1(X; C_k)$ over X with the C_k fixed at their current values. The first subproblem is simple because we can solve it for each k separately and because XC_kX' is linear in C_k. The second problem is less simple; Carroll and Chang propose a general ALS trick which we shall call *splitting*. Instead of minimizing $\sigma_1(X; C_k)$ we minimize

$$\sigma_1(X, Y; C_k) = \sum \text{tr}(B_k - XC_kY')^2$$

over both X and Y. Because the B_k are symmetric and the C_k are symmetric too, we expect X and Y to converge to the same value. Splitting can be used to generalize ALS from multivariate multilinear problems to multivariate polynomial problems in many cases but the precise conditions under which splitting works have not been established. In the three-way case it is easy to show that using splitting is closely related to using the Gauss–Newton method. In the Gauss–Newton method we minimize the approximation

$$\sum \text{tr}\{B_k - (XC_kX' + XC_k\Delta' + \Delta C_kX')\}^2$$

over Δ, and then set the new X equal to the old X plus Δ. This gives the same iterates as ALS applied to $\sigma_1(X, Y; C_k)$. Of course convergence of the ALS procedures is no problem, they converge more or less by definition. Often convergence is painfully slow, however. Ramsay (1973) has shown that we can usually accellerate convergence considerably by choosing a suitable relaxation factor in the ALS iterations. Another disadvantage of ALS in this context is that the procedure may converge to a C_k which is not psd, or to a W_k which is not nonnegative.

We can compute very good initial estimates for our iterative procedures by using Theorems 13 and 14. The proof of Theorem 13 gives us IDIOSCAL estimates of X and C_k, Theorem 14 says that we can find INDSCAL estimates if we diagonalize these C_k. A number of these two-step procedures has been proposed. The first one in [112], the most straightforward one in [34]. This last paper also has the necessary references. Another two-step procedure proposed by De Leeuw is used to construct the initial configuration in ALSCAL. It is described in [139]. It is clear that we can construct nonmetric versions of three-way MDS by combining the results in this section with those from the previous section. Carroll and Chang (unpublished) have experimented with nonmetric INDSCAL, called NINDSCAL, while Richard Sands (in press) has a nonmetric version of IDIOSCAL which is comparable to ALSCAL.

3.2. Least squares on the squared distances

3.2.1. Two-way MDS

For metric two-way MDS we can also consider the loss function

$$\sigma_2(X)=\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{h_{ij}-d_{ij}^2(X)\right\}^2,$$

with $h_{ij}=\delta^2(i,j)$, as before, and $d_{ij}^2(X)=(x_i-x_j)'(x_i-x_j)$. This loss function was proposed by Obenchain (1971) and Hayashi (1974), but efficient algorithms to minimize functions like this in several different MDS situations were proposed by Takane, Young, and De Leeuw (1977). The current version of the ALSCAL algorithm (see [140]) can handle all kinds of metric/nonmetric two/three way data structures, using the basic two-step alternating least squares methodology of Young, De Leeuw, and Takane (in press). The interesting problem is how we must minimize $\sigma_2(X)$ over X. In ALSCAL a single coordinate is changed at the time, the other coordinates are fixed at current values, and we cycle through the coordinates. Of course σ_2 is a quartic in each coordinate, we minimize over the coordinate by solving a cubic. The procedure may not be very appealing at first sight but is surprisingly efficient.

The loss function $\sigma_2(X)$ is called SSTRESS by Takane et al., the loss function $\sigma_1(X)$ is called STRAIN by Carroll. There is an interesting relationship between STRAIN and SSTRESS which explains why the initial configuration routines for ALSCAL work as well as they do.

THEOREM 21. *If X is centered, then* $\sigma_2(X)\geqslant 4\sigma_1(X)$.

PROOF. Define $\sigma(X,b)=\operatorname{tr}\{U(b)-XX'\}^2$, where $U(b)$ has elements

$$u_{ij}(b)=-\tfrac{1}{2}(h_{ij}-b_i-b_j).$$

Then $\sigma_1(x)=\min\{\sigma(X,b)|b\}$, while $\sigma_2(X)=4\sigma(X,a)$, $a_i=\Sigma x_{is}^2$.

If we combine Theorems 20 and 21, we obtain a lower bound for

$$\min\{\sigma_2(X)|X\in\mathbb{R}^{np}\}.$$

Similar bounds can be obtained for nonmetric and three-way versions of SSTRESS in terms of STRAIN.

In nonmetric two-way scaling we have to minimize $\sigma_2(X,\hat{d})$ over X in $\mathbb{R}^{np}$ and over all admissible disparities. In the ordinal case we have to use normalization requirements again to prevent certain forms of degeneracy. Discussions on how to normalize SSTRESS are in [138] and in [140]. In the simplest case (for ordinal two-way MDS) we require that the sum of the fourth powers of the disparities is unity (the loss function is the sum of squares of the differences between squared disparities and squared distances).

THEOREM 22. *Suppose that a matrix with all off-diagonal disparities equal is admissible. Suppose in addition that n and p are such that an $n \times p$ matrix Y exists with $\Sigma y_{is} = 0$ for each s, $\Sigma y_{is}^2 = n/p$ for each s, $\Sigma y_{is} y_{it} = 0$ for all $s \neq t$, and $\Sigma y_{is}^2 = 1$ for all i. Then*

$$\min\{\sigma_2(X, \hat{d}) \mid X, \hat{d}\} \leq 1 - \left(\frac{p}{p+1}\right)\left(\frac{n}{n-1}\right).$$

PROOF. Simply substitute constant $\hat{d}$, suitably normalized, and Y in the formula for SSTRESS.

The assumptions in Theorem 22 are quite realistic. Constant disparities are admissible in all ordinal MDS problems, a matrix Y with the required properties exists if $p = 1$ and n is even, and also if $p = 2$ (regular polygon). The theorem is important because it tells us that even if there is no structure at all in the data, we can still find fairly low SSTRESS by distributing clumps of points regularly on a sphere. Observe that Y satisfies the conditions of the theorem for $n = 10$ and $p = 2$ if we have ten points equally spaced on a circle, but also if we have five groups of two points equally spaced on a circle. In fact, in many applications we have found that ALSCAL tends to make clumps on a sphere.

Using squared distance loss has one important advantage over inner product loss. We have seen that the double centering operation $B = -\frac{1}{2}JHJ$ can introduce statistical dependencies, but it also complicates the regression problems in the optimal scaling step. In the ordinal case, for example, minimizing $\sigma_1(X, \theta)$ for fixed X over $\theta \geq 0$ is a quadratic programming problem, minimizing $\sigma_2(X, \hat{d})$ over the admissible $\hat{d}$ for fixed X is also a quadratic programming problem but it has simpler structure and consequently the more efficient monotone regression algorithm can be used. A disadvantage is that using squared distances is more complicated in the metric case, and Theorem 22 shows that using squared distances may bias towards regular spherical configurations. For inner product loss the corresponding upper bound is $1 - p/(n-1)$, which is attained for all Y satisfying the less stringent conditions $\Sigma y_{is} y_{it} = \delta^{st}$.

3.2.2. Three-way MDS

The only three-way MDS program based on squared distance loss is ALSCAL. We do not discuss the algorithm here because the principles are obvious from the previous section. There are two substeps, the first one is the optimal scaling step, it finds new disparities for a given configuration, the second one changes the configuration by the cubic equation algorithm of ALSCAL and the weights for the individuals by linear regression techniques. There is an interesting modification which fits naturally into the ALSCAL framework, although it has not been implemented in the current versions. We have seen that the inner product algorithms can give estimates of $C_k = T_k T_k'$ that are not psd. The same thing is true for ALSCAL, but if we minimize the loss over T_k instead of over C_k we do

not have this problem, and the minimization can be carried out 'one variable at a time' by using the cubic equation solver again. This has the additional advantage that we can easily incorporate rank restrictions on the C_k. If we require that the T_k are $p \times 1$, for example, we can fit the 'personal compensatory model' mentioned by Coombs (1964, p. 199), and by Roskam (1968, Chapter IV).

An important question in constructing MDS loss functions is how they should be normalized. This is discussed in general terms in [78], and for ALSCAL in [125] and [140]. McCallum (1977) studies the effect of different normalizations in a three-way situation empirically. Other Monte Carlo studies have been carried out by McCallum and Cornelius (1977) who study metric recovery by ALSCAL, and by McCallum (unpublished) who compares ALSCAL and INDSCAL recovery (in terms of mean squared error). It seems that metric INDSCAL often gives better results than nonmetric ALSCAL, even for nonmetric data. This may be due to the difference in metric/nonmetric, but also to the difference between scalar product and squared distance loss. Takane, Young, and De Leeuw (1977) compare CPU-times of ALSCAL and INDSCAL. The fact that ALSCAL is much faster seems to be due almost completely to the better initial configuration, cf. [34].

3.3. Least squares on the distances

3.3.1. Metric two-way MDS

The most familiar MDS programs are based on the loss function

$$\sigma_3(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\big(\delta_{ij} - d_{ij}(X)\big)^2$$

where we have written δ_{ij} for $\delta(i, j)$ and where we have introduced nonnegative *weights* w_{ij}. For $w_{ij} \equiv 1$ this loss function is STRESS, introduced by Kruskal [73, 74]. The Guttman–Lingoes–Roskam programs are also based on loss functions of this form. Using σ_3 seems somewhat more direct than using σ_2 or σ_1, moreover, both σ_2 and σ_1 do not make much sense if the distances are non-Euclidean. A possible disadvantage of σ_3 is that it is somewhat less smooth, and that computer programs that minimize σ_3 usually converge more slowly than programs minimizing σ_1 or σ_2. Moreover, the classical Young–Householder–Torgerson starting point works better for σ_2 and σ_1, which has possibly some consequences for the frequency of local minima. There are as yet, however, no detailed comparisons of the three types of loss functions.

We have introduced the weights w_{ij} for various reasons. If there is information about the variability of the δ_{ij}, we usually prefer weighted least squares for statistical reasons, if there is a large number of independent identically distributed replications, then weighted least squares gives efficient estimates and the minimum of σ_3 has a chi-square distribution. Another reason for using weights is that we can compare STRESS and SSTRESS more easily. It is obvious that if

$\delta_{ij} \approx d_{ij}(X)$ and if we choose $w_{ij} = 4\delta_{ij}^2$, then $\sigma_3(X) \approx \sigma_2(X)$. Thus, if a good fit is possible, we can imitate the behaviour of σ_2 by using σ_3 with suitable weights Ramsey (1977, 1978) has proposed the loss function

$$\sigma_4(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\ln \delta_{ij} - \ln d_{ij}(X)\right)^2,$$

which makes sense for log-normally distributed dissimilarities. Again, if $\delta_{ij} \approx d_{ij}(X)$ and if we choose $w_{ij} = 1/\delta_{ij}^2$, we find $\sigma_3(X) \approx \sigma_4(X)$.

The algorithms for minimizing $\sigma_3(X)$ proposed by Kruskal [73, 74], Roskam (1968), Guttman (1968), Lingoes and Roskam (1973) are gradient methods. They are consequently of the form

$$X^{(\tau+1)} = X^{(\tau)} - \alpha_\tau \nabla \sigma_3(X^{(\tau)}),$$

where index τ is the iteration number, $\nabla\sigma_3$ is the gradient, and $\alpha_\tau > 0$ is the step-size. Kruskal (1977) discusses in detail how he chooses his step-sizes in MDSCAL and KYST, the same approach with some minor modifications is adopted in the MINISSA programs of Lingoes and Roskam (1973). KYST [80] also fits the other power metrics, but for powers other than two there are both computational and interpretational difficulties. The city-block (power = 1) and sup-metric (power = ∞) are easy to interpret by very difficult to fit because of the serious discontinuities of the gradient and the multitude of local minima. The intermediate cases are easier to fit but difficult to interpret. We prefer a somewhat different approach to step-size. Consider the Euclidean case first. The cross product term

$$\rho(X) = \sum \sum w_{ij} \delta_{ij} d_{ij}(X)$$

in the definition of $\sigma_3(X)$ is a homogeneous convex function, the term $\eta^2(X) = \sum \sum d_{ij}^2(X)$ is quadratic and can be written as $\eta^2(X) = \operatorname{tr} X'VX$ for some V. If ρ is differentiable at X, then $\nabla\sigma_3(X) = 2VX - 2\nabla\rho(X)$, which suggests the algorithm

$$X^{(\tau+1)} = V^+ \nabla\rho(X^{(\tau)}),$$

with V^+ a generalized inverse of V. If ρ is *not* differentiable at X, which happens only if $x_i = x_j$ for some $i \neq j$, then we can use the subgradient $\partial\rho(X)$ instead of the gradient $\nabla\rho(X)$, and use the algorithm

$$X^{(\tau+1)} \in V^+ \partial\rho(X^{(\tau)}).$$

De Leeuw and Heiser (1980) proved the following global convergence theorem.

THEOREM 23. *Consider the algorithm* $\overline{X}^{(\tau)} \in V^+ \partial\rho(X^{(\tau)})$ *and*

$$X^{(\tau+1)} = \alpha_\tau \overline{X}^{(\tau)} + (1 - \alpha_\tau) X^{(\tau)}$$

with $0<\epsilon_1<\alpha_\tau<2-\epsilon_2<2$. *Then* $\sigma_3(X^{(\tau)})$ *is a decreasing, and thus convergent, sequence. Moreover*

$$\operatorname{tr}(X^{(\tau+1)}-X^{(\tau)})'V(X^{(\tau+1)}-X^{(\tau)})$$

converges to zero.

In [30] there is a similar, but less general, result for general Minkovski metrics. In that case the computations in an iteration are also considerably more complicated.

3.3.2. MDS with restrictions

Consider general restrictions of the form $X\in\Omega$, with Ω a given subset of $\mathbb{R}^{np}$. Gradient methods can be used without much trouble if the restrictions are simple (fix some parameters at constant values, restrict others to be equal). This has been discussed by Bentler and Weeks (1977) and by Bloxom (1978). For complicated sets of restrictions the gradient methods get into trouble and must be replaced by one of the more complicated feasible direction methods of nonlinear programming. The approach based on convex analysis generalizes quite easily. De Leeuw and Heiser (1980) proposed the convergent algorithm

$$X^{(\tau+1)}\in P_\Omega\big(V^+\partial\rho(X^{(\tau)})\big),$$

where P_Ω is the metric projection on Ω in the metric defined by V. Three-way MDS methods are special cases of this general model, because we can use an $nm\times nm$ supermatrix of weights W, with only the m diagonal submatrices of order n nonzero. The configurations can be collected in an $nm\times p$ supermatrix X whose m submatrices of order $n\times p$ must satisfy the restrictions $X_k=YT_k$.

3.3.3. Nonmetric MDS

Nonmetric versions of all algorithms discussed in the previous sections can be easily constructed by using the general optimal scaling approach which alternates disparity adjustment and parameter (or configuration) adjustment. In the convex analysis approach we can moreover use the fact that the maximum of $\rho(X)$ over normalized disparities is the pointwise maximum of convex functions and consequently also convex. Thus if we impose suitable normalization requirements, the same theory applies as in the metric case (cf. [31]).

References

[1] Andalafte, E. Z. and Blumenthal, L. M. (1964). Metric characterizations of Banach and Euclidean spaces. *Fund. Math.* **55**, 23–55.

[2] Attneave, F. (1950). Dimensions of similarity. *Amer. J. Psychol.* **63**, 516–556.

[3] Baker, R. F., Young, F. W. and Takane, Y. (1979). An asymmetric Euclidean model: an alternating least squares method with optimal scaling features. *Psychometrika*, to.appear.

[4] Beals, R. and Krantz, D. H. (1967). Metrics and geodesics induced by order relations. *Math. Z.* **101**, 285–298.
[5] Beals, R., Krantz, D. H. and Tversky, A. (1968). Foundations of multidimensional scaling. *Psychol. Rev.* **75**, 127–142.
[6] Bennett, J. F. and Hays, W. L. (1960). Multidimensional unfolding, determining the dimensionality of ranked preference data. *Psychometrika* **25**, 27–43.
[7] Bentler, P. M. and Weeks, D. G. (1978). Restricted multidimensional scaling. *J. Math. Psychol.* **17**, 138–151.
[8] Bloxom, B. (1978). Constrained multidimensional scaling in N-spaces. *Psychometrika* **43**, 397–408.
[9] Blumenthal, L. M. (1938). Distance geometries. *Univ. Missouri Studies* **13** (2).
[10] Blumenthal, L. M. (1953). *Theory and Applications of Distance Geometry*. Clarendon Press, Oxford.
[11] Blumenthal, L. M. (1975). Four point properties and norm postulates. In: L. M. Kelly, ed., *The Geometry of Metric and Linear Spaces. Lecture Notes in Mathematics* **490**. Springer, Berlin.
[12] Bunemann, P. (1971). The recovery of trees from measures of dissimilarity. In: R. F. Hodson, D. G. Kendall and A. Taitu, eds., *Mathematics in the Archeological and Historical Sciences*. University of Edinburgh Press, Edinburgh.
[13] Busemann, H. (1955). *The Geometry of Geodesics*. Academic Press, New York.
[14] Busemann, H. (1970). *Recent Synthetic Differential Geometry*. Springer, Berlin.
[15] Carroll, J. D. (1976). Spatial, non-spatial and hybrid models for scaling. *Psychometrika* **41**, 439–463.
[16] Carroll, J. D. and Arabie, P. (1980). Multidimensional scaling. *Ann. Rev. Psychol.* **31**, 607–649.
[17] Carroll, J. D. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart–Young' decomposition. *Psychometrika* **35**, 283–319.
[18] Carroll, J. D., Green, P. E. and Carmone, F. J. (1976). CANDELINC: a new method for multidimensional analysis with constrained solutions. Paper presented at *International Congress of Psychology*, Paris.
[19] Carroll, J. D. and Pruzansky, S. (1977). MULTILINC: multiway CANDELINC. Paper presented at *American Psychological Association Meeting*, San Francisco.
[20] Carroll, J. D. and Wish, M. (1974). Models and methods for three-way multidimensional scaling. In: *Contemporary Developments in Mathematical Psychology*. Freeman, San Francisco.
[21] Cayley, A. (1841). On a theorem in the geometry of position. *Cambridge Math. J.* **2**, 267–271.
[22] Cliff, N. (1973). Scaling. *Ann. Rev. Psychol.* **24**, 473–506.
[23] Constantine, A. G. and Gower, J. C. (1978). Graphical representation of asymmetric matrices. *Appl. Statist.* **27**, 297–304.
[24] Coombs, C. H. (1964). *A Theory of Data*. Wiley, New York.
[25] Cormack, R. M. (1971). A review of classification. *J. Roy. Statist. Soc. Ser A.* **134**, 321–367.
[26] Critchley, F. (1978). Multidimensional scaling: a critique and an alternative. In: L. C. A. Corsten and J. Hermans, eds., *COMPSTAT 1978*. Physika Verlag, Vienna.
[27] Cross, D. V. (1965). Metric properties of multidimensional stimulus generalization. In: J. R. Barra et al., eds., *Stimulus Generalization*. Stanford University Press, Stanford.
[28] Cunningham, J. P. (1978). Free trees and bidirectional trees as representations of psychological distance. *J. Math. Psychol.* **17**, 165–188.
[29] De Leeuw, J. (1970). The Euclidean distance model. Tech. Rept. RN 02-70. Department of Datatheory, University of Leiden.
[30] De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In: J. C. Lingoes, ed., *Progress in Statistics*. North-Holland, Amsterdam.
[31] De Leeuw, J. and Heiser, W. (1977). Convergence of correction matrix algorithms for multidimensional scaling. In: *Geometric Representations of Relational Data*. Mathesis Press, Ann Arbor.
[32] De Leeuw, J. and Heiser, W. (1979). Maximum likelihood multidimensional scaling of interaction data. Department of Datatheory, University of Leiden.

[33] De Leeuw, J. and Heiser, W. (1980). Multidimensional scaling with restrictions on the configuration. In: P. R. Krishnaiah, ed., *Multivariate Analysis, Vol. V.* North–Holland, Amsterdam.
[34] De Leeuw, J. and Pruzansky, S. (1978). A new computational method to fit the weighted Euclidean model. *Psychometrika* **43**, 479–490.
[35] Eisler, H. (1973). The algebraic and statistical tractability of the city block metric. *Brit. J. Math. Statist. Psychol.* **26**, 212–218.
[36] Fisher, R. A. (1922). The systematic location of genes by means of cross-over ratios. *American Naturalist* **56**, 406–411.
[37] Fréchet, M. (1935). Sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert. *Ann. Math.* **36**, 705–718.
[38] Gold, E. M. (1973). Metric unfolding: data requirements for unique solution and clarification of Schönemann's algorithm. *Psychometrika* **38**, 555–569.
[39] Goldmeier, E. (1937). Über Ähnlichkeit bei gesehenen Figuren. *Psychol. Forschung* **21**, 146–208.
[40] Gower, J. C. (1977). The analysis of asymmetry and orthogonality. In: J. R. Barra et al., eds., *Progress in Statistics.* North–Holland, Amsterdam.
[41] Guttman, L. (1941). The quantification of a class of attributes: a theory and method of scale construction. In: P. Horst, ed., *The Prediction of Personal Adjustment.* Social Science Research Council, New York.
[42] Guttman, L. (1944). A basis for scaling qualitative data. *Amer. Sociol. Rev.* **9**, 139–150.
[43] Guttman, L. (1946). An approach for quantifying paired comparisons and rank order. *Ann. Math. Statist.* **17**, 144–163.
[44] Guttman, L. (1950). The principal components of scale analysis. In: S. A. Stouffer, ed., *Measurement and Prediction.* Princeton University Press, Princeton.
[45] Guttman, L. (1957). Introduction to facet design and analysis. Paper presented at *Fifteenth Int. Congress Psychol.*, Brussels.
[46] Guttman, L. (1959). Metricizing rank-ordered or unordered data for a linear factor analysis. *Sankhya* **21**, 257–268.
[47] Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika* **33**, 469–506.
[48] Guttman, L. (1971). Measurement as structural theory. *Psychometrika* **36**, 329–347.
[49] Haberman, S. (1974). *The Analysis of Frequency Data.* University of Chicago Press, Chicago.
[50] Harshman, R. A. (1970). Foundations of the PARAFAC procedure: models and conditions for an explanatory multi-modal factor analysis. Department of Phonetics, UCLA.
[51] Harshman, R. A. (1972). PARAFAC2: mathematical and technical notes. Working papers in phonetics No. 22, UCLA.
[52] Hartigan, J. A. (1967). Representation of similarity matrices by trees. *J. Amer. Statist. Assoc.* **62**, 1140–1158.
[53] Hayashi, C. (1974). Minimum dimension analysis MDA: one of the methods of multidimensional quantification. *Behaviormetrika* **1**, 1–24.
[54] Hays, W. L. and Bennett, J. F. (1961). Multidimensional unfolding: determining configuration from complete rank order preference data. *Psychometrika* **26**, 221–238.
[55] Heiser, W. and De Leeuw, J. (1977). How to use SMACOF-I. Department of Datatheory, University of Leiden.
[56] Heiser, W. and De Leeuw, J. (1979). Metric multidimensional unfolding. *MDN, Bulletin VVS* **4**, 26–50.
[57] Heiser, W. And De Leeuw, J. (1979). How to use SMACOF-III. Department of Datatheory, University of Leiden.
[58] Helm, C. E. (1959). A multidimensional ratio scaling analysis of color relations. E.T.S., Princeton.
[59] Holman, E. W. (1972). The relation between hierarchical and Euclidean models for psychological distances. *Psychometrika* **37**, 417–423.
[60] Indow, T. (1975). An application of MDS to study binocular visual space. In: *US–Japan seminar on MDS.* La Jolla.

[61] Ireland, C. T., Ku, H. H. and Kullback, S. (1969). Symmetry and marginal homogeneity in an $r \times r$ contingency table. *J. Amer. Statist. Assoc.* **64**, 1323–1341.
[62] Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika* **32**, 241–254.
[63] Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika* **43**, 443–477.
[64] Keller, J. B. (1962). Factorization of matrices by least squares. *Biometrika* **49**, 239–242.
[65] Kelly, J. B. (1968). Products of zero–one matrices. *Can. J. Math.* **20**, 298–329.
[66] Kelly, J. B. (1975). Hypermetric spaces. In: L. M. Kelly, ed., *The Geometry of Metric and Linear Spaces. Lecture Notes in Mathematics* **490**. Springer, Berlin.
[67] Klingberg, F. L. (1941). Studies in measurement of the relations between sovereign states. *Psychometrika* **6**, 335–352.
[68] Krantz, D. H. (1967). Rational distance functions for multidimensional scaling. *J. Math. Psychol.* **4**, 226–245.
[69] Krantz, D. H. (1968). A survey of measurement theory. In: G. B. Dantzig and A. F. Veinott, eds., *Mathematics of the Decision Sciences*. American Mathematical Society, Providence.
[70] Krantz, D. H. and Tversky, A. (1975). Similarity of rectangles: an analysis of subjective dimensions. *J. Math. Psychol.* **12**, 4–34.
[71] Kroonenberg, P. M. and De Leeuw, J. (1977). TUCKALS2: a principal component analysis of three mode data. Tech. Rept. RN 01-77. Department of Datatheory, University of Leiden.
[72] Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: the interrelationship between similarity and spatial density. *Psychol. Rev.* **85**, 445–463.
[73] Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**, 1–27.
[74] Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika* **29**, 28–42.
[75] Kruskal, J. B. (1976). More factors than subjects, tests, and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika* **41**, 281–293.
[76] Kruskal, J. B. (1977). Trilinear decomposition of three-way arrays: rank and uniqueness in arithmetic complexity and in statistical models. *Linear Algebra Appl.* **18**, 95–138.
[77] Kruskal, J. B. (1977). Multidimensional scaling and other methods for discovering structure. In: *Statistical Methods for Digital Computers*. Wiley, New York.
[78] Kruskal, J. B. and Carroll, J. D. (1969). Geometric models and badness of fit functions. In: P. R. Krishnaiah, ed., *Multivariate Analysis, Vol. II*. Academic Press, New York.
[79] Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Sage Publications, Beverly Hills.
[80] Kruskal, J. B., Young, F. W. and Seery, J. B. (1977). How to use KYST-2, a very flexible program to do multidimensional scaling and unfolding. Bell Laboratories, Murray Hill.
[81] Lancaster, P. (1977). A review of numerical methods for eigenvalue problems nonlinear in the parameter. In: *Numerik und Anwendungen von Eigenwertaufgaben und Verzweigungsproblemen. Internat. Ser. Numer. Math.* **38**. Birkhauser, Basel.
[82] Landahl, H. D. (1945). Neural mechanisms for the concepts of difference and similarity. *Bull. Math. Biophysics* **7**, 83–88.
[83] Lew, J. S. (1975). Preorder relations and pseudoconvex metrics. *Amer. J. Math.* **97**, 344–363.
[84] Lew, J. S. (1978). Some counterexamples in multidimensional scaling. *J. Math. Psychol.* **17**, 247–254.
[85] Lindman, H. and Caelli, T. (1978). Constant curvature Riemannian scaling. *J. Math. Psychol.* **17**, 89–109.
[86] Lingoes, J. C. (1971). Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika* **36**, 195–203.
[87] Lingoes, J. C. and Roskam, E. E. (1973). A mathematical and empirical analysis of two multidimensional scaling algorithms. *Psychometrika* **38**, monograph supplement.
[88] Luce, R. D. (1961). A choice theory analysis of similarity judgements. *Psychometrika* **26**, 151–163.
[89] Luce, R. D. (1963). Detection and recognition. In: R. D. Luce, R. R. Bush and E. Galanter, eds., *Handbook of Mathematical Psychology, Vol. I*. Wiley, New York.

[90] Luneborg, R. K. (1947). *Mathematical Analysis of Binocular Vision*. Princeton University Press, Princeton.
[91] MacCallum, R. C. (1977). Effects of conditionality on INDSCAL and ALSCAL weights. *Psychometrika* **42**, 297–305.
[92] MacCallum, R. C. and Cornelius III, E. T. (1977). A Monte Carlo investigation of recovery of structure by ALSCAL. *Psychometrika* **42**, 401–428.
[93] Menger, K. (1928). Untersuchungen über allgemeine Metrik. *Math. Ann.* **100**, 75–163.
[94] Messick, S. J. (1956). Some recent theoretical developments in multidimensional scaling. *Ed. Psychol. Meas.* **16**, 82–100.
[95] Messick, S. J. and Abelson, R. P. (1956). The additive constant problem in multidimensional scaling. *Psychometrika* **21**, 1–15.
[96] Nakatani, L. H. (1972). Confusion-choice model for multidimensional psychophysics. *J. Math. Psychol.* **9**, 104–127.
[97] Obenchain, R. L. (1971). Squared distance scaling as an alternative to principal components analysis. Bell Laboratories, Holmdell.
[98] Pieszko, H. (1975). Multidimensional scaling in Riemannian space. *J. Math. Psychol.* **12**, 449–477.
[99] Ramsay, J. O. (1975). Solving implicit equations in psychometric data analysis. *Psychometrika* **40**, 337–360.
[100] Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika* **42**, 241–266.
[101] Ramsay, J. O. (1978). Confidence regions for multidimensional scaling analysis. *Psychometrika* **43**, 145–160.
[102] Restle, F. (1959). A metric and an ordering on sets. *Psychometrika* **24**, 207–220.
[103] Richardson, M. W. (1938). Multidimensional psychophysics. *Psychol. Bull.* **35**, 659–660.
[104] Roskam, E. E. (1968). Metric analysis of ordinal data in psychology. VAM, Voorschoten, The Netherlands.
[105] Rutishauser, H. (1970). Simultaneous iteration method for symmetric matrices. *Numer. Math.* **16**, 205–223.
[106] Saito, T. (1978). An alternative procedure to the additive constant problem in metric multidimensional scaling. *Psychometrika* **43**, 193–201.
[107] Sattath, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika* **42**, 319–345.
[108] Schoenberg, I. J. (1935). Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert." *Ann. Math.* **38**, 724–732.
[109] Schoenberg, I. J. (1937). On certain metric spaces arising from Euclidean space by a change of metric and their imbedding in Hilbert space. *Ann. Math.* **40**, 787–793.
[110] Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.* **44**, 522–536.
[111] Schönemann, P. H. (1970). On metric multidimensional unfolding. *Psychometrika* **35**, 349–366.
[112] Schönemann, P. H. (1972). An algebraic solution for a class of subjective metrics models. *Psychometrika* **37**, 441–451.
[113] Schönemann, P. H. (1977). Similarity of rectangles. *J. Math. Psychol.* **16**, 161–165.
[114] Seidel, J. J. (1955). Angles and distances in n-dimensional Euclidean and non-Euclidean geometry. Parts I, II, III. *Indag. Math.* **17**, 329–335, 336–340, 535–541.
[115] Seidel, J. J. (1975). Metric problems in elliptic geometry. In: L. M. Kelly, ed., *The Geometry of Metric and Linear Spaces. Lecture Notes in Mathematics* **490**. Springer, Berlin.
[116] Shepard, R. N. (1957). Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika* **22**, 325–345.
[117] Shepard, R. N. (1958). Stimulus and response generalization: tests of a model relating generalization to distance in psychological space. *J. Exp. Psychol.* **55**, 509–523.
[118] Shepard, R. N. (1958). Stimulus and response generalization: deduction of the generalization gradient from a trace model. *Psychol. Rev.* **65**, 242–256.
[119] Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function, Parts I, II. *Psychometrika* **27**, 125–140, 219–246.

[120] Shepard, R. N. (1966). Metric structures in ordinal data. *J. Math. Psychol.* **3**, 287–315.
[121] Shepard, R. N. (1974). Representation of structure in similarity data: problems and prospects. *Psychometrika* **39**, 373–421.
[122] Shepard, R. N. and Arabie, P. (1979). Additive clustering: representation of similarities as combinations of discrete overlapping properties. *Psychol. Rev.* **86**, 87–123.
[123] Sibson, R. (1972). Order-invariant methods for data analysis. *J. Roy. Statist. Soc. Ser. B* **34**, 311–349.
[124] Stumpf, C. (1880). *Tonpsychologie, Vol. I and II.* Teubner, Leipzig.
[125] Takane, Y., Young, F. W. and De Leeuw, J. (1977). Nonmetric individual differences in multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika* **42**, 7–67.
[126] Taussky, O. (1949). A recurring theorem on determinants. *Amer. Math. Monthly* **56**, 672–676.
[127] Torgerson, W. (1952). Multidimensional scaling I—theory and methods. *Psychometrika* **17**, 401–419.
[128] Torgerson, W. (1958). *Theory and Methods of Scaling.* Wiley, New York.
[129] Torgerson, W. (1965). Multidimensional scaling of similarity. *Psychometrika* **30**, 379–393.
[130] Townsend, J. T. (1978). A clarification of some current multiplicative confusion models. *J. Math. Psychol.* **18**, 25–38.
[131] Tversky, A. (1966). The dimensional representation and the metric structure of similarity data. Michigan Math. Psychol. Program.
[132] Tversky, A. (1977). Features of similarity. *Psychol. Rev.* **84**, 327–352.
[133] Tversky, A. and Krantz, D. H. (1969). Similarity of schematic faces: a test of interdimensional additivity. *Perception and Psychophysics* **5**, 124–128.
[134] Tversky, A. and Krantz, D. H. (1970). The dimensional representation and the metric structure of similarity data. *J. Math. Psychol.* **7**, 572–596.
[135] Valentine, J. E. (1969). Hyperbolic spaces and quadratic forms. *Proc. Amer. Math. Soc.* **37**, 607–610.
[136] Wender, K. (1971). A test of independence of dimensions in multidimensional scaling. *Perception and Psychophysics* **10**, 30–32.
[137] Young, F. W. (1972). A model for polynomial conjoint analysis algorithms. In: R. N. Shepard, A. K. Romney and S. B. Nerlove, eds., *Multidimensional Scaling: Theory and Applications in the Social Sciences, Vol. I.* Seminar Press, New York.
[138] Young, F. W., De Leeuw, J. and Takane, Y. (1980). Quantifying qualitative data. In: E. Lantermann and H. Feger, eds., *Similarity and Choice.* Huber, Bern.
[139] Young, F. W., Takane, Y. and Lewyckyj, R. (1978). Three notes on ALSCAL. *Psychometrika* **43**, 433–435.
[140] Young, F. W. and Lewyckyj, R. (1979). ALSCAL-4 User's guide. Data analysis and theory associates, Carrboro, NC.
[141] Young, G. and Householder, A. S. (1938). discussion of a set of points in terms of their mutual distances. *Psychometrika* **3**, 19–22.