P. R. Krishnaiah, ed., *Multivariate Analysis-V*

MULTIDIMENSIONAL SCALING WITH RESTRICTIONS ON THE CONFIGURATION

Jan DE LEEUW and Willem HEISER
Department of Data Theory, University of Leiden, Netherlands

A convergent algorithm model for multidimensional scaling with restrictions is described. The algorithm model is applied to accelerating MDS iterations, to fitting models similar to the ones used in the analysis of covariance structures, and to fitting individual differences models.

1. Introduction

It is difficult to define the precise methodological status of multidimensional scaling (MDS) techniques, and it is equally difficult to evaluate their usefulness. There are many applications of MDS, with results that are apparently regarded as useful by the people who have used the technique. Most of them, of course, have been in the social and behavioural sciences. Bibliographies covering most of these applications have been compiled by Bick et al [3] and by Nishisato [35]. Recently MDS has also been applied, apparently with success, in geography (Colledge and Rushton [11]), cartography (Gilbert [22]), genetics (Lalouel [32]), archeology (Kendall [27]), and biochemistry (Crippen [13, 14]).

If we study these applications carefully, it turns out that there are at least four different types. The first possibility is that a Euclidean spatial structure is known to exist, but is (partially) undetermined. We want to recover the structure 'optimally' by using MDS on the inexact or incomplete distance measurements. This type of application does not seem to occur in the behavioural sciences, but an example is Crippen's method for the conformation of molecules (Crippen [13, 14]). The second type of application arises if we know that a particular spatial structure exists (which is not necessarily Euclidean), and we want to compare an 'optimal' Euclidean map, based on dissimilarity data, with this known structure. In the behavioural sciences applications of type II occur in multidimensional psychophysics, more particularly in acoustics (an early example is Levelt et al. [33]). Most of the applications of MDS in geography and cartography

are of this type. An impressive example is the 'maps from marriages' paper by Kendall (28). The third type of application starts with a theory about dissimilarity data, which implies that they are related in a well-defined way to Euclidean distances. We want to test the theory and perhaps find the 'hidden structure'. Applications of this type occur in mathematical psychology. According to Shepard [41, p. 374–375] this is the type of application he had in mind when he started developing MDS as we know it now. The developments in measurement theory (Beals et al. [1]) and in statistical methods for MDS (Ramsay [37, 38]) are mainly relevant for applications of this type. The problems of archeological seriation (Kendall [27], Wilkinson [45]), and of genetic mapping of linkage groups (Lalouel [32]) are also of type III.

For the fourth type of application there is no known a priori structure, and there is no theory which even tells us that a spatial representation is appropriate. We merely want to find an 'optimal' spatial representation of our dissimilarity data, and we apply MDS because 'a picture is worth a thousand numbers'. More sophisticated arguments for this approach are based on invariance (we get the same picture from a wide variety of dissimilarity data), stability (the spatial representation has greater statistical reliability than the individual dissimilarity estimates), or communicability (cf. Shepard [41, p. 375–376]). Approximately 90% of the applications in the behavioural sciences have been of this fourth type. Measurement theory and traditional statistical theory are not very relevant here, because they require the formulation of a definite algebraic or stochastic model. MDS, as used in type IV, is a sophisticated technique for making plots. It is obvious that the four types of applications differ because they require different levels of external information (or theory). Type I requires more theory than type II, type II more than type III, and type III more than type IV, which seems to require no theory at all. In all cases, however, the external information is used only *after* the MDS analysis has been completed. It is used in the interpretation of the results, to relate them to the already existing body of theory. It is not incorporated directly into the analysis, because most current MDS programs do not have the capability to incorporate restrictions.

If we analyze the type IV applications in detail, however, it turns out that even there some kind of external information often is available, which is again used in the interpretation phase. The theory is, of course, much less specific than the theory used in type I applications. As a consequence it is often difficult to relate MDS results to external information, and the

interpretations can easily become far-fetched. In type I or type II applications the existing theory is so specific that interpretation of MDS results is unnecessary, they either make sense or they do not make sense. In type IV applications there is so much freedom that it is almost always possible to make some sort of sense out of MDS results. It follows from this discussion that it would be desirable to have MDS algorithms, which can incorporate all kinds of external information in the form of restrictions on the configuration. Such algorithms could be used in type I or type II situations, but there they would probably give the same results as unrestricted MDS methods. They will be especially useful in type III and type IV situations, in fact they tend to make type III applications out of type IV applications. We need 'confirmatory' MDS if we want to use our prior information efficiently.

2. Related work

For the very same reasons we discussed in the previous section a number of special methods have been proposed in the recent MDS literature which impose various kinds of restrictions on the solutions. In order to compare them with the method we are going to propose we need some terminology and notation. We follow Kruskal [31] as closely as possible. The data of a classical MDS problem are collected in a matrix $\Delta = \{\delta_{ij}\}$ of *dissimilarity measurements* between n objects. We want to find a *configuration matrix* X, of order $n \times p$, in such a way that $d_{ij}(X)$ is approximately equal to δ_{ij}, for all $i,j = 1,\ldots,n$. Here $d_{ij}(X)$ is the Euclidean distance between rows x_i and x_j, interpreted as *points* in p-dimensional space. Thus

$$d_{ij}^2(X) = (x_i - x_j)'(x_i - x_j).$$

'Approximately equal to' is usually defined in terms of some real-valued *loss function*, which measures departure from perfect fit. If this is the case, then the MDS problem can be formulated in a more specific way: we want to find X in such a way that the loss function is minimized.

A more general problem occurs if we have m dissimilarity matrices Δ_k. We then want to find configuration matrices X_k, in such a way that $d_{ij}(X_k)$ is approximately equal to δ_{ijk} for all $i,j = 1,\ldots,n$ and for all $k = 1,\ldots,m$. If the X_k are otherwise unrestricted the problem is equivalent to m separate

classical scaling problems, and consequently not of independent interest. It is therefore not surprising that the problem of imposing restrictions on MDS configurations started in the area of *individual differences scaling*.

The key paper in this area is Carroll and Chang [7]. They discuss restrictions of the form $X_k = YW_k$, with Y $n \times p$, and W_k $p \times p$ and diagonal. Moreover they present an elegant algorithm for fitting this model. The same model and algorithm were proposed, independently and in a different context, by Harshman [25]. We shall refer to this model as the INDSCAL model, using the name of the Carroll–Chang computer program. In the same paper Carroll and Chang also discuss (briefly) the more general model $X_k = YC_k$, with Y as before, and with C_k again $p \times p$, but not necessarily diagonal. This is the IDIOSCAL model, named after another Carroll-Chang computer program. In [26] Harshman proposes the PARAFAC-2 model, which is $X_k = YW_kZ'$, with Y as before, with W_k diagonal, and with Z another $p \times p$ matrix. We shall simply call this PARAFAC because in our context PARAFAC-1 is the same thing as INDSCAL. A general discussion of the algebraic properties of these models, their interrelationships, and of the corresponding algorithms, is given in Carroll and Wish [8], Kroonenberg and DeLeeuw [29], and De Leeuw and Pruzansky [20].

Existing MDS algorithms can be divided into two classes. In the first place there are the gradient algorithms. They include MDSCAL, KYST, SSA, and MINISSA. For MDSCAL and KYST we refer to Kruskal [30, 31], for SSA and MINISSA to Lingoes and Roskam [34]. And in the second place there are the alternating least squares algorithms. These methods divide the parameters of the problem into subsets, in such a way that minimizing the least squares loss function over each subset, with the other subsets fixed, is a comparatively simple problem. Kruskal [31, p. 309] calls such subproblems, and the parameter subsets which define them, 'nice'. The alternating least squares algorithms then cycles through its nice subproblems until convergence. Examples are INDSCAL and ALSCAL of Takane, Young, and De Leeuw [43].

The advantages of gradient methods in the MDS context have been reviewed quite thoroughly by Kruskal [31]. We mention two problems with the gradient method. The first one is that $d_{ij}(X)$ is not differentiable everywhere, if $x_i = x_j$ the partials do not exist. The second problem is the choice of step-size procedure in gradient methods. Kruskal [31, p. 315–319] has developed a step-size procedure based on some heuristic ideas, combined with a great deal of numerical experimentation. The method (also used in MINISSA) seems to work quite well in practice, but does not

guarantee monotone convergence of loss function values. Because the step-size is a complicated function of all previous gradients and function values, it seems extremely difficult to give a precise analysis of convergence behaviour of the procedure.

The disadvantages of ALS (i.e. alternating least squares) are of a completely different nature. ALS procedures always start by transforming the model (Kruskal [31, p. 309] would say 'by neglecting errors'), either to squared distances (as in ALSCAL) or to inner products (as in INDSCAL). The resulting loss functions are differentiable everywhere, and by definition ALS gives monotone convergence. The main problem with ALS is that transforming the model may be undesirable. In the first place the transformations only make sense in the case of Euclidean distances, and do not generalize to other scaling models. In the second place the transformations may affect the errors in undesirable ways. If large dissimilarities have the largest measurement or sampling errors, for example, then it is statistically unwise to apply unweighted least squares to the squared dissimilarities. In fact Ramsay [37, 38] suggests applying unweighted least squares to the logarithms of the dissimilarities, because this makes more sense from the error theoretical point of view. Of course ALS cannot be applied if we use logarithms. If the dissimilarity measurements are independent, then the double-centering transformation that converts to scalar products in the error-free case introduces dependencies. Again this makes the use of unweighted least squares problematical.

There is a third type of algorithm, which does not use transformations, uses weighted least squares to incorporate possible error-theories, leads to monotone convergence without step-size choices, and has no problems with differentiability. The method was originally derived in the unweighted case by Guttman [24]. He assumed differentiability, however, and interpreted the method as a gradient method with constant step-size. He did not show that the method worked, but empirical studies of Lingoes and Roskam [34] indicated that it seemed to converge in practice. De Leeuw [17] studied the method in the weighted least squares case, generalized it to nonmetric scaling, discussed non-Euclidean models, and gave the first formal convergence proof. An alternative derivation will be presented in the next section of this paper.

The forms of restricted scaling that have been proposed in the literature strongly depend on the type of algorithm that is used. Some constraints fit nicely into a gradient framework, others can easily be combined with ALS. In gradient algorithms, for example, it is very simple to fix some of the coordinates of the configuration at constant values. In fact this possibility

is already built into MINISSA. Bentler and Weeks [2] generalized this considerably. The coordinates x_{is} can be fixed, free, or proportional. By proportional we mean that the restrictions are of the form $x_{is} = u_{is}x_{jt}$, with the u_{is} known, and with x_{jt} a free coordinate. Bloxom [4] fits the general individual differences model $X_k = Y_k C_k$. The parameters in Y_k and C_k can be either fixed or free, and some free parameters can be restricted to be equal to others. This clearly has INDSCAL and IDIOSCAL as special cases. It seems to us that Bloxom's work cannot be generalized very much any more. More general restrictions can only be fitted into a gradient framework by using the much more complicated feasible directions method of nonlinear programming.

Alternating least squares algorithms usually solve subproblems which are linear regression problems. This is not true in all cases, the ALSCAL subproblems for example amount to solving cubic equations, but generally nice sets of variables enter linearly into the model equations. This implies that ALS is especially valuable in fitting multilinear models, and that the natural constraints that can be used with ALS are linear constraints. The combination of multilinear models (principal component analysis, three-mode component analysis, transformed Euclidean scaling) with linear constraints has recently been studied by Carroll, Green, and Carmone [9] and by Carroll and Pruzansky [10]. They show that incorporating linear restrictions is possible by transforming the input matrix, and by applying the CANDECOMP algorithm of Carroll and Chang [7] to this transformed matrix. In the MDS context the constraints imposed by Carroll et al are of the type $X_k = YCW_k$, with Y known, C and W_k unknown, and W_k diagonal for each k. More general constraints are possible, but these seem to be the most natural ones.

In the next section we show that our new algorithm combines the most convenient features of gradient and ALS methods. We also show that it can incorporate a very general class of constraints in a natural way.

3. Algorithm model

We want to minimize

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

over all X in Ω, a given subset of R^{np}, the space of all $n \times p$ matrices. The

w_{ij} are given non-negative weights, if $w_{ij}=0$ we treat δ_{ij} as missing. Without loss of generality we can assume that the dissimilarities are normalized by

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij}^2=1.$$

Then

$$\sigma(X)=1-2\rho(X)+\eta^2(X),$$

where

$$\rho(X)=\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij}d_{ij}(X), \quad \text{and} \quad \eta^2(X)=\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}d_{ij}^2(X).$$

The function η^2 is quadratic in X. If we define the matrix $V=\{v_{ij}\}$ by

$$v_{ij}=-(w_{ij}+w_{ji}) \quad \text{for } i\neq j, \qquad v_{ii}=\sum_{j\neq i}^{n}(w_{ij}+w_{ji}),$$

then

$$\eta^2(X)=\operatorname{tr}X'VX.$$

If $w_{ij}=1$ for all $i\neq j$, then $V=2(nI-ee')$. If X is centered, i.e. if the centroid of the n points x_i is the origin, then in this case $\eta^2(X)=2n\operatorname{tr}X'X$. In the general weighted case we assume that V is irreducible [17, p. 138], which implies that, if V^- is the Moore–Penrose inverse of V, then $X=VV^-X=V^-VX$ for all centered X. The function ρ causes the trouble with differentiability in gradient methods. Consequently we use another regularity property. Because the Euclidean distance is a convex and positively homogeneous function, the same thing is true for ρ. Observe that this result is true for much more general definitions of d_{ij}, in fact it remains true for $d_{ij}(X)=\mu(x_i-x_j)$, with μ any gauge. By using this fact we can extend at least some of our results to general Minkovski geometries, in which distance is not necessarily symmetric. For convex functions the notion of a gradient is replaced by that of a subgradient. The subgradient $\partial\rho(X)$ of ρ at X is a non-empty, convex, compact set, defined by $Y\in\partial\rho(X)$ if $\rho(Z)\geqslant\rho(X)+\operatorname{tr}Y'(Z-X)$ for all Z in R^{np}. If ρ is differentiable at X, with gradient $\nabla\rho(X)$, then the subgradient is the singleton $\{\nabla\rho(X)\}$. For proofs of these results we refer to Rockafellar [39, part V].

Theorem 1. *For all X, Y in R^{np} and for all $Z \in V^- \partial\rho(Y)$ we have*

$$\sigma(X) \leqslant 1 - \eta^2(Z) + \eta^2(X - Z).$$

For all X in R^{np} and for all $Z \in V^- \partial\rho(X)$ we have

$$\sigma(X) = 1 - \eta^2(Z) + \eta^2(X - Z).$$

Proof. Suppose $U \in \partial\rho(X)$. By definition $\rho(Z) \geqslant \rho(X) + \operatorname{tr} U'(Z - X)$ for all Z in R^{np}, and consequently also for all Z of the form αX, with $\alpha \geqslant 0$. Using the homogeneity of ρ gives $(1-\alpha)\rho(X) \leqslant (1-\alpha)\operatorname{tr} U'X$ for all $\alpha > 0$, which implies that $\rho(X) = \operatorname{tr} U'X$. Substitute this again in the definition of the subgradient. This gives $\rho(Z) \geqslant \operatorname{tr} U'Z$. If we substitute the last equation and inequality in the formula $\sigma(X) = 1 - 2\rho(X) + \eta^2(X)$, and simplify, we obtain the results stated in the theorem.

In the same way we can prove that $Z \in \partial\rho(X)$ implies that Z is centered. By definition we must have $\rho(X + e\mu') \geqslant \rho(X) + \operatorname{tr} Z'e\mu'$ for all p-vectors μ, i.e. for all translations. Because the distances are translation invariant this simplifies to $e'Z\mu \leqslant 0$ for all μ, which implies $e'Z = 0$.

Theorem 2. *If $Z \in V^- \partial\rho(X)$, then $\eta^2(Z) \leqslant 1$.*

Proof. Because the dissimilarities are normalized applying Cauchy-Schwartz to the definition of ρ gives $\rho(X) \leqslant \eta(X)$ for all X in R^{np}. The second part of Theorem 1 gives $\rho(X) = \operatorname{tr} Z'VX \leqslant \eta(Z)\eta(X)$. The first part gives $\rho(Z) \geqslant \operatorname{tr} Z'VZ = \eta^2(Z)$. If we combine the three inequalities, we obtain the chain

$$\frac{\rho(X)}{\eta(X)} \leqslant \eta(Z) \leqslant \frac{\rho(Z)}{\eta(Z)} \leqslant 1.$$

This is more than enough to prove the theorem.

We also define the metric projection P_Ω, in the metric defined by V, as

$$P_\Omega(X) = \left\{ Y \in \Omega \mid \eta^2(Y - X) = \min_{Z \in \Omega} \eta^2(Z - X) \right\}.$$

We assume that Ω is defined in such a way that $P_\Omega(X)$ is non-empty for all X in R^{np}. In the terminology used in nonlinear approximation theory we

assume that Ω is *proximinal*. It is sufficient for proximinality that Ω is either compact or closed and convex, but these conditions are by no means necessary. Using the metric projection we can now describe our basic algorithm model. We start with some X^0 in Ω, and in each iteration we find the update X^+ of our current solution by the rule

$$X^+ \in P_\Omega(V^- \partial\rho(X)).$$

There is one exception to this rule. If our current solution satisfies $X \in P_\Omega(V^- \partial\rho(X))$, then the algorithm stops. We call this an algorithm model because it is incomplete. We have not specified how to select an element from the subgradient, and we have not specified how we intend to solve the metric projection problem. It is obvious that the practical simplicity of our algorithm depends on the ease with which the metric projection problem can be solved. To use the terminology of Kruskal once again, some projection problems are nice while others are not so nice. We now prove a first convergence theorem for the case in which the algorithm generates an infinite sequence X^k, with corresponding loss function values $\sigma(X^k)$.

Theorem 3. *$\sigma(X^k)$ is a decreasing sequence, and consequently converges.*

Proof. Suppose $\overline{X} \in V^- \partial\rho(X)$. By part 1 of Theorem 1

$$\sigma(X^+) \leqslant 1 - \eta^2(\overline{X}) + \eta^2(X^+ - \overline{X}).$$

By assumption $X \notin P_\Omega(V^- \partial\rho(X))$, and thus, by the definition of the metric projection,

$$1 - \eta^2(\overline{X}) + \eta^2(X^+ - \overline{X}) < 1 - \eta^2(\overline{X}) + \eta^2(X - \overline{X}).$$

By part 2 of Theorem 1:

$$1 - \eta^2(\overline{X}) + \eta^2(X - \overline{X}) = \sigma(X).$$

If we combine the results we find $\sigma(X^+) < \sigma(X)$, and because the sequence is bounded below by zero it converges.

Theorem 3 is reassuring, but not entirely satisfactory because it tells us nothing about the behaviour of the sequence X^k. Before we study this in more detail we prove a simple necessary condition for a minimum of σ.

Theorem 4. *If X minimizes σ on Ω, then $X \in P_\Omega(V^- \partial\rho(X))$.*

Proof. Suppose $X \notin P_\Omega(V^- \partial\rho(X))$ and $Y \in P_\Omega(V^- \partial\rho(X))$. By exactly the same argument that proved Theorem 3 we show that $\sigma(Y) < \sigma(X)$.

From now on we call points satisfying the necessary condition of theorem 4 *desirable*. We already have the trivial result that if the algorithm stops, it stops at a desirable point. What happens if it does not stop? To study this we need some preliminary results.

Theorem 5. *All X^k are in the compact set $\{X \mid \eta(X) \leqslant 1 + \sqrt{\sigma(X^0)}\}$.*

Proof. From the proof of Theorem 3, using the same notation,

$$1 - \eta^2(\overline{X}) + \eta^2(X^+ - \overline{X}) < \sigma(X) \leqslant \sigma(X^0).$$

By using Theorem 2

$$\eta^2(X^+ - \overline{X}) \leqslant \sigma(X^0).$$

But, by Cauchy–Schwartz,

$$\big(\eta(X^+) - \eta(\overline{X})\big)^2 \leqslant \eta^2(X^+ - \overline{X}),$$

which implies, by using Theorem 2 again, that

$$\eta(X^+) \leqslant 1 + \sqrt{\sigma(X^0)}.$$

Theorem 6. *The point-to-set map $X \to P_\Omega(V^- \partial\rho(X))$ is closed.*

Proof. The subgradient map is closed (Rockafellar [39, p. 233]). The set of all $V^- \partial\rho(X)$, as X varies over R^{np}, is compact by Theorem 2. The map P_Ω is closed, and consequently the composition is closed (Zangwill [46]).

Theorem 7. *The sequence X^k has convergent subsequences. Each subsequential limit is a desirable point. All subsequential limits have the same value of σ.*

Proof. This follows from convergence Theorem A of Zangwill [46, p. 91].

The type of convergence described in Theorem 7 is still quite weak. We can prove a considerably stronger result by making the additional assumption that Ω is convex.

Theorem 8. *If Ω is convex, then $\eta^2(X^{k+1}-X^k)$ converges to zero.*

Proof. It follows from the proof of Theorem 3 that, if $\overline{X}\in V^-\partial\rho(X)$, $\eta^2(X-\overline{X})-\eta^2(X^+-\overline{X})=\eta^2(X^+-X)+2\,\mathrm{tr}(X^+-\overline{X})'V(X-X^+)$ converges to zero. If Ω is convex then, by convex approximation theory, the last term on the right is nonnegative. The first term on the right is positive, and thus both terms converge to zero.

The theorem does not say that the sequence X^k converges. It tells us, by a familiar theorem of Ostrowski, that either the sequence X^k converges or that the sequence has a continuum of accumulation points. Convergence follows if we make additional uniqueness or smoothness assumptions. From a practical point of view, however, the result of Theorem 8 is strong enough. If we define an ε-desirable point as a configuration X for which the distance between X and the set $P_\Omega(V^-\partial\rho(X))$ is less than ε, then the theorem tells us that the algorithm finds an ε-desirable point in a finite number of steps, no matter how small we choose the positive number ε. We also observe that the assumption that Ω is convex is not necessary for $\eta^2(X^{k+1}-X^k)\to 0$. It is often possible to prove this result from other uniqueness or smoothness assumptions.

There is a generalization of the algorithm which can be extremely important in those cases in which the metric projection problem is not nice. Suppose Q is a closed point-to-set map from $R^{np}\times\Omega$ into subsets of Ω. Suppose moreover that $Y\in Q(Z,X)$ implies that $\eta^2(Y-Z)\leqslant\eta^2(X-Z)$, with equality if and only if $X\in P_\Omega(Z)$. The new algorithm is $X^+\in Q(V^-\partial\rho(X),X)$. Convergence Theorems 3 and 7 remain true for this modified algorithm, which does not solve the metric projection problem completely, but merely takes a step in the right direction. In many examples the map Q can be defined by using alternating least squares.

There are some obvious relationships of our algorithm with alternating least squares theory. In each iteration we have to solve a least squares projection problem, if this subproblem is not easy to solve we can often use one or more ALS cycles to improve our current best solution. One interpretation of our algorithm is that the basic majorization result proved in theorem 1 makes it possible to apply the powerful machinery of ALS,

without first having to transform to squared distances or inner products. There are also some relationships with gradient theory. Suppose σ is differentiable at X. Then $\nabla\sigma(X)=2VX-2\nabla\rho(X)$, and thus the algorithm can be written as a gradient projection method of the form $X^+\in P_\Omega(X-\frac{1}{2}V^-\nabla\sigma(X))$. If we want to relate our algorithm to the earlier work of Guttman [24] we first have to specify how we select an element from the subgradient of ρ. Define the matrix $B(X)$ by

$$b_{ij}(X)=-(w_{ij}\delta_{ij}+w_{ji}\delta_{ji})s_{ij}(X) \quad \text{if } i\neq j,$$

with $s_{ij}(X)=1/d_{ij}(X)$ if $d_{ij}(X)\neq 0$, and $s_{ij}(X)=0$ otherwise. Moreover

$$b_{ii}(X)=-\sum_{j\neq i}^{n} b_{ij}(X).$$

By simple algebra we find $\rho(X)=\operatorname{tr}X'B(X)X$, and from the Cauchy–Schwarz inequality $B(X)X\in\partial\rho(X)$. Thus one possible specification of our algorithm is $X^+\in P_\Omega(V^-B(X)X)$. If ρ is differentiable at X, then $B(X)X=\nabla\rho(X)$. Guttman studied the case in which Ω is R^{np}, and in which the off-diagonal weights are all unity. In this case we find $X^+=(1/2n)B(X)X$, which is basically Guttman's correction matrix algorithm. Observe that the mapping B is not continuous at X if some of the $d_{ij}(X)$ are zero. This implies that the limits of convergent subsequences generated by the algorithm do not necessarily satisfy $X\in P_\Omega(V^-B(X)X)$, although they must always satisfy $X\in P_\Omega(V^-\partial\rho(X))$.

4. Unrestricted scaling

It is convenient to introduce the *Guttman-transform* $\overline{X}$ of a configuration matrix X as $\overline{X}=V^-B(X)X$. If Ω is R^{np}, then the basic algorithm is $X^+=\overline{X}$. The chain in the proof of Theorem 3 becomes

$$\sigma(X^+)\leqslant 1-\eta^2(\overline{X})+\eta^2(X^+-\overline{X})=1-\eta^2(\overline{X})$$

$$\leqslant 1-\eta^2(\overline{X})+\eta^2(X-\overline{X})=\sigma(X).$$

This immediately implies that $\eta^2(X^+-X)$ converges to zero. If $\sigma(X^k)$ decreases to σ_∞, then $\eta^2(X^k)$ increases to $\eta^2_\infty=1-\sigma_\infty$. Moreover from the

proof of theorem 2 we find that $\rho(X^k)$ increases to $\rho_\infty = \eta_\infty^2$, and that $\lambda(X^k) = \rho(X^k)/\eta(X^k)$ increases to $\lambda_\infty = \eta_\infty$.

As an example of a more general algorithm using the map $Q(\bar{X}, X)$ we have $X^+ = \bar{X} + \alpha(X - \bar{X})$. We find $\sigma(X^+) \leqslant \sigma(X) + (\alpha^2 - 1)\eta^2(X - \bar{X})$, and thus we have convergence if $-1 < \alpha < +1$. Moreover $\eta^2(X^+ - X) = (1-\alpha)2\eta^2(X - \bar{X})$ also converges to zero in this case. The reasons for choosing $\alpha \neq 0$ are as follows. Very slow linear convergence is typical for MDS algorithms, and precise solutions are impossible if we do not apply some sort of acceleration device. A more precise analysis of the convergence rate of our MDS algorithms will be given elsewhere. For the moment we merely observe that using $\alpha = \varepsilon_0 - 1$, with ε_0 a very small positive number, preserves global convergence and approximately halves the number of iterations required to obtain a given precision, at no extra cost. The explanation is, roughly, as follows. The basic algorithm $X^+ = \bar{X}$ converges linearly with rate $\kappa_0 = 1 - \varepsilon_1$, with ε_1 another small positive number. If we use any $\alpha \neq 0$ we obtain the rate $\kappa_\alpha = \kappa_0 + \alpha(1 - \kappa_0)$, and if $\alpha = \varepsilon_0 - 1$ then this is equal to $(1-\varepsilon_1)^2 + \varepsilon_1(\varepsilon_0 - \varepsilon_1)$, which is approximately $(1-\varepsilon_1)^2$. Thus the convergence rate is squared, the number of iterations is halved.

A further generalization of our basic algorithm can be used to derive versions of the optimal gradient, memory gradient, and conjugate gradient methods. In the unweighted case we have at a point where σ is differentiable that $X^+ = \bar{X} + \alpha(X - \bar{X})$ can also be written as $X^+ = X - \mu\nabla\sigma(X)$ with $\mu = (1-\alpha)/4n$. If $\alpha = 0$ then $\mu = 1/4n$, if $\alpha = -1$ then $\mu = 1/2n$. We can also interpret $\sigma(X - \mu\nabla\sigma(X))$ as a function of μ and minimize it over μ with our algorithm. To see how this is done consider the more general problem of minimizing $\sigma(X)$ over all X of the form $Y_0 + \sum \mu_s Y_s$, where the Y_s, including Y_0, are fixed matrices. This is a problem with linear restrictions. The solution by projection can be computed from $\hat{\mu} = C^{-1}b$, where C contains the elements $\operatorname{tr} Y_s' V Y_t$, and b the elements $\operatorname{tr} Y_s' V(\bar{X} - Y_0)$. We then use $\hat{\mu}$ to compute X^+, compute its Guttman-transform, and repeat the procedure until convergence. Observe that by using a relaxation parameter again, we can also set $\hat{\mu} = 2C^{-1}b$, and still obtain convergence.

In Table 1 we have collected the results of a number of experiments with these step-size techniques. There are five different data sets.

*K*1: Nine Dutch political parties, data collected by Van der Kamp, initial configuration Kruskal's *L*-shape (cf. [30]).

*K*2: Same data, initial configuration Young–Householder–Torgerson.

*F*1: Thirteen ethnic groups, from Funk et al. [21], initial *L*.

*F*2: Same data, YHT initial configuration.

D: Ten points, $\delta_{ij} = 1$ for all $i \neq j$, initial *L*.

Table 1
Speed of convergence of various step-size procedures

Example	No it	loss	rate
$K1, \alpha=0$	99	0.0444477	0.945592
$K1, \alpha=-1$	57	0.0444380	0.888642
$K1, \alpha=\text{opt}$	47	0.0444365	0.855814
$K2, \alpha=0$	62	0.0446179	0.912394
$K2, \alpha=-1$	35	0.0446135	0.852053
$K2, \alpha=\text{opt}$	32	0.0446115	0.831023
$F1, \alpha=0$	201	0.0655642	0.940073
$F1, \alpha=-1$	111	0.0655561	0.876544
$F1, \alpha=\text{opt}$	107	0.0655556	0.872367
$F2, \alpha=0$	67	0.0602583	0.713715
$F2, \alpha=-1$	36	0.0602560	0.507205
$F2, \alpha=\text{opt}$	35	0.0602564	0.527138
$D, \alpha=0$	123	0.0111065	0.927797
$D, \alpha=-1$	72	0.0111058	0.853318
$D, \alpha=\text{opt}$	59	0.0111057	0.812373

All analyses were in two dimensions, and unweighted. We iterated until $\sigma-\sigma^{+}<10^{-6}$. For each example there were three runs, one with $\alpha=0$, one with $\alpha=-1$, and one which uses inner iterations for the optimal step until $|\alpha^{+}-\alpha|<10^{-3}$. The table lists the number of major iterations, the minimum value of the loss function, and an empirical estimate of the linear convergence rate $(\sigma^{+}-\sigma^{++})/(\sigma-\sigma^{+})$. The examples show that indeed $\alpha=-1$ approximately halves the number of iterations and approximately squares the convergence rate of $\alpha=0$. Using the optimal step-size gives a slightly better convergence rate, and fewer major iterations. But because inner iterations are about as expensive as major iterations, computing the optimal α is not worthwhile (the argument is the same as in [31, p. 315–316]). Even an 'ideal' step-size routine which finds the optimal α in a single inner iteration would be as expensive as using $\alpha=0$ throughout. The step-size routine we use needs on the average between five and fifteen inner iterations, which makes the complete algorithm about five times as expensive as the one with $\alpha=0$. Thus $\alpha=-1$ is by far the most efficient choice, and we have adopted this in our FORTRAN computer program SMACOF1 for metric MDS. The examples also illustrate the completely different point that using a good start can reduce the number of iterations considerably, and that different starts can lead to different local optima.

5. Linear restrictions

We have already analyzed one particular example of using linear restrictions on X in the previous section. A more interesting example from a practical point of view is $X = YC$, with Y a given $n \times q$ matrix, and C an unknown $q \times p$ matrix. This can also be fitted with the gradient method of Bloxom [4], and in an inner product framework by the ALS method of Carroll et al [9]. Our algorithm simply gives

$$X^+ = YC^+ = Y(Y'VY)^{-1}Y'V\overline{X} = Y(Y'VY)^{-1}Y'B(X)YC,$$

provided that the inverse exists. It is clear that we can require without loss of generality that $Y'VY = I$, which simplifies matters even more. In applications Y can be an ANOVA-type design matrix, it can also be a matrix with real valued measurements on q 'independent' variables. The last case occurs, for example, in multidimensional psychophysics. Here Y contains physical characteristics of the stimuli, such as frequency characteristics of synthetically generated vowels.

In many special cases C can be restricted to be diagonal. In this case the update is, supposing $y_s'Vy_s = 1$ for all $s = 1, \ldots, q$,

$$c_s^+ = y_s'V\bar{x}_s = c_s y_s'B(X)y_s.$$

Observe that if $c_s \geqslant 0$, then $c_s^+ \geqslant 0$. And if $c_s = 0$ then $c_s^+ = 0$. The model with C diagonal can also be fitted with the gradient method of Bentler and Weeks (2). A particularly interesting class of examples has both a diagonal C and a binary Y. Here the *squared* distance has a set theoretical interpretation. If H_i is the set $\{s | y_{is} = 1\}$, and the measure of a subset of $\{1, 2, \ldots, q\}$ is defined as $m(H) = \sum \{c_s^2 | s \in H\}$, then

$$d_{ij}^2(X) = m(H_i \Delta H_j),$$

with Δ the symmetric difference. Thus in this case the squared Euclidean distance is a metric too (cf. the discussion of the ADCLUS model by Shepard [41, p. 414–417]). In some cases a graph theoretical interpretation is possible. If we are fitting a simplex, for example, we use the matrix Y defined by $y_{is} = 1$ if $i \geqslant s$, and $y_{is} = 0$ otherwise. Call this matrix Y_S. In this case $d_{ij}^2(X)$ is the path length distance in a simple chain, i.e. $d_{ij}^2(X)$ can be represented as a Euclidean distance in one dimension, while $d_{ij}(X)$ itself needs $n-1$ dimensions. This was already observed by Guttman [23]. Another possibility is fitting a circumplex, with width (k_1, k_2), where

$1 \leqslant k_1 \leqslant k_2 \leqslant n$. Then $y_{is}=1$ if $k_1 \leqslant |i-s| \leqslant k_2$, and $y_{is}=0$ otherwise. We call a matrix of this class Y_C. This makes $d_{ij}^2(X)$ the path length distance in a circular graph, consisting of a single cycle. In the same way we can construct for each tree a binary matrix Y such that $d_{ij}^2(YC)$ is the path length distance in the tree [6]. In the particular case $Y=I$ the tree is a star. In factor analytic terminology $Y=I$ can also be interpreted as no common, only unique variance [2, p. 140].

The theory generalizes without any further complications to mixed models in which parts are unrestricted and other parts are restricted. A particularly attractive class of models is $X=|X_1|YC_1|C_2|$ in which X_1 is $n \times q$ and unrestricted, Y is either Y_S or one of the Y_C, and C_1 and C_2 are both square and diagonal. The models can be coded by a three digit number $q_1q_2q_3$. Here q_1 is the number of unrestricted dimensions, if $q_1=0$ the X_1 part is missing from the model. If $q_2=0$ the YC_1 part is missing, if $q_2=1$ then $Y=Y_S$, if $q_2=2$ then Y is one of the Y_C. If $q_3=0$ the C_2 part is missing, if $q_3=1$ it is not missing. Thus $q01$ is the q-dimensional MDS model with uniquenesses, also discussed by Bentler and Weeks [2], model 010 is a simplex, 011 is a quasi-simplex, and so on.

In many types of applications unrestricted MDS finds semi-circular, parabolic or horse-shoe shaped configurations in two dimensions. By using the simplex and circumplex we can often find more parsimonious representations. This is illustrated by the results in Table 2. We applied the computerprogram SMACOF2, which fits these three-digit models, to a number of dissimilarity matrices.

Table 2
Minimum loss for various models and various examples

Model	*K*	*S*	*L*	*C*	*DP*	*DK*	*DV*
100	0.4167	0.3493	0.3270				
200	0.2214	0.1370	0.1772	0.1075	0.1068	0.1404	0.0784
010	0.2016	0.1029	0.1990	0.0681	0.0836	0.1755	0.0864
020				0.2051	0.2391	0.0985	0.2211
001	0.1588	0.2125		0.3181			
101	0.0640	0.0430					
201	0.0543			0.0388			
011	0.1001	0.0513	0.1933	0.0561	0.0666	0.1314	0.0816
021				0.1245	0.1841	0.0829	0.2090
110	0.1466	0.0536	0.1644				
210	0.0848	0.0221		0.0535			
111		0.0185					

K: Nine Dutch political parties, data collected by Van der Kamp in 1968, averaged similarity ratings, subjects 100 students.
S: Six Swedish political parties, data taken from Sjöberg [42].
L: 15 musical intervals, taken from Levelt et al [33].
C: Nine number ability tests, data from Coombs [12].
DP: 13 Dutch political parties, data from De Gruyter [15]. Data averaged over students who vote PvdA (social democrat).
DK: Same data, but averaged over students who vote KVP (christian democrat).
DV: Same data, but averaged over students who vote VVD (liberal party).

Table 2 lists the minimum loss function values for various models. In this table we use the square root of the loss, which is more convenient. For *K*, *S*, and *C* the simplex 010 is considerably better than two-dimensional MDS 200, despite the fact that 010 has only n-1 parameters while 200 has $2n-3$ parameters. For *K* and *S* the quasi-simplex 011 is again considerably better, and the scaling models with uniqueness 101 and 201 also perform well. For *L* we would also prefer 010 to 200. In the data *DP*, *DK*, and *DV* we expect that people from the middle will think that extremist left and extremist right are quite alike, while people from the moderate left and moderate right will think this to a lesser degree. This would imply that the *DK*-data are more circumplex-like, while the *DP* and *DV*-data must be more simplex-like. We do find something like this in Table 2.

6. Nonlinear restrictions

There are many models possible which impose nonlinear restrictions. We only discuss some examples, but do not work them out in detail. In nonlinear problems computing the metric projection may not be simple. Thus we may have to use alternating least squares inner iterations.

For the first example we discuss inner iterations are unnecessary. The restrictions are $x_s \in K_s$, with K_s a given convex cone. The subproblem is to solve p cone projection problems, one for each dimension. If K_s is the cone of all vectors with a given ordering of the elements, for example, then the subproblems are monotone regression problems. This problem was also studied by Noma and Johnson [36], who used a completely different algorithm. More generally the requirements $x_s \in K_s$ can be used to scale dimensions which are defined nominally, ordinally, or numerically. These extensions of MDS are closely related to recent extensions of principal

component analysis such as PRINCIPALS [44] or HOMALS [16].

In the second example (inspired by Borg [5, p. 638]) we want to restrict the representation of a subset of the objects in such a way that they are on a straight line in two-space, or, more generally, on a q-dimensional linear manifold in p-space. The corresponding projection problem is finding the best rank q approximation to a submatrix, which is easy to solve by computing the singular value decomposition. More generally it may be interesting to require that some or all of the objects are on a nonlinear manifold of some sort. In Levelt et al. [33], for example, it would make sense to require that the tone intervals are on a quadratic manifold, in colour studies we can require that the objects are on a circle. This generally requires more complicated iterative techniques than the linear case.

The third example we discuss is $X = YC$, with Y binary and C diagonal as before, but now with both Y and C unknown. This is the ADCLUS model of Arabee and Shepard (discussed in [41]). We have to use ALS here. Computing the optimal C for fixed Y has been explained in the previous section, and computing the optimal binary Y for fixed C is also quite simple. After one or more inner ALS cycles we compute a new Guttman-transform, and start a new major iteration. The major problem with the algorithm is not the amount of computation, but the fact that there are so many desirable points. This is due to the discreteness of the restrictions, and cannot be helped.

7. Individual differences

The theory of Section 13 generalizes in an obvious way to individual differences scaling. For each iteration we have to minimize a function of the form

$$\sum_{k=1}^{m} \operatorname{tr}(Z_k - \bar{X}_k)' V_k (Z_k - \bar{X}_k)$$

over the Z_k, where $\bar{X}_k$ are the Guttman-transforms of the X_k from the previous iteration, i.e. $\bar{X}_k = V_k^- B_k(X_k) X_k$. In the general weighted case we need ALS techniques to fit the individual differences models discussed in Section 2. This is true in particular if we also want to impose restrictions of the form discussed by Bloxom [4]. If the weights are all equal the computations simplify considerably. We discuss the three most important models in some detail, and show how the subproblems can be solved.

For IDIOSCAL we define the $n \times mp$ supermatrix $\overline{X}$, containing the current $\overline{X}_k$, and the $p \times mp$ supermatrix C, containing the C_k. The problem can be rewritten as minimization of $\text{tr}(\overline{X} - YC)'(\overline{X} - YC)$, which means that we can set Y equal to the normalized eigenvectors corresponding with the p largest eigenvalues of $\overline{X}\overline{X}'$, and set $C_k = Y'\overline{X}_k$. For INDSCAL we can proceed by fitting one dimension at a time. Define the $n \times m$ matrices $\overline{X}_s$, which contain column s of each of the $\overline{X}_k$, and define the vectors c_s, which contain diagonal element s of each of the C_k. Then we must minimize, for each s separately, $\text{tr}(\overline{X}_s - y_s c_s')'(\overline{X}_s - y_s c_s')$. Thus we can set y_s equal to the normalized eigenvector corresponding with the largest eigenvalue of $\overline{X}_s\overline{X}_s'$, and we set $c_s' = y_s'\overline{X}_s$.

For PARAFAC we have to minimize, in each subproblem,

$$\sum_{k=1}^{m} tr(\overline{X}_k - YC_kZ')'(\overline{X}_k - YC_kZ').$$

But this is exactly the problem solved by the CANDECOMP algorithm of Carroll and Chang [7]. From the general theory we know that it suffices to perform just one CANDECOMP cycle before computing new Guttman-transforms.

8. Nonmetric scaling

All the techniques discussed in this paper are for metric MDS, in which the dissimilarities are either completely known or completely unknown (if δ_{ij} is unknown we set $w_{ij} = 0$). If the dissimilarities are partially known, for example up to a linear transformation or up to a monotone transformation, then we have to apply nonmetric MDS. The loss is now a function of the configuration *and* of the *disparities*, which are admissible transformations of the dissimilarities. De Leeuw [17] shows that in the unrestricted case the basic algorithm $X^+ = V^- B(X)X$ is still convergent, provided we replace the normalized dissimilarities by the normalized disparities in each iteration. The disparities are computed, in the classical case, by monotone regression. This result remains true in the case in which there are restrictions on the configuration. De Leeuw and Heiser [19] study more general partitioned loss function for nonmetric MDS, and show that the basic algorithm can still be used, but that we must not only change the dissimilarities in each iteration but possibly also the weights. Again this

result remains true in the restricted case. Consequently for nonmetric MDS with restrictions only the computation of the Guttman-transform becomes more complicated, the metric projection problem remains exactly the same.

References

[1] Beals, R., Krantz, D. H., and Tversky, A. (1968). Foundations of multidimensional scaling. *Psychol. Rev.* **75**, 127–142.

[2] Bentler, P. M., and Weeks, D. G. (1978). Restricted multidimensional scaling models. *J. Math. Psychol.* **17**, 138–151.

[3] Bick, W., Bauer, H., Mueller, P. J., and Gieseke, O. (1977). Multidimensional scaling and clustering techniques (theory and applications in the social sciences). A Bibliography. Institut für angewandte Sozialforschung. Universität zu Köln.

[4] Bloxom, B. (1978). Constrained multidimensional scaling in N spaces. *Psychometrika* **43** 397–408.

[5] Borg, I. (1977). Geometric representations of individual differences, In *Geometric Representations of Relational Data*. Mathesis Press. Ann Arbor.

[6] Bunemann, P. (1971). The recovery of trees from measures of dissimilarity. In *Mathematics in the archeological and historical sciences*. University of Edinburgh Press, Edinburgh.

[7] 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.

[8] Carroll, J. D., and Wish, M. (1974). Models and methods for three-way multidimensional scaling. In *Contemporary Developments in Mathematical Psychology*. Freeman, San Francisco.

[9] Carroll, J. D., Green, P. E., and Carmone, F. J. (1976). CANDELINC (*CAN*onical *DE*composition with *LIN*ear Constraints): a new method for multidimensional analysis with constrained solutions. Paper presented at *International Congress of Psychology*, Paris, France.

[10] Carroll, J. D., and Pruzansky, S. (1977). MULTILINC: MULTIWAY CANDELINC (*CAN*onical *DE*composition with *LIN*ear Constraints). Paper presented at *American Psychological Association* meeting, San Francisco.

[11] Colledge, R. G., and Rushton, G. (1972). Multidimensional scaling: review and geographical applications. *Geographic technical papers series*, no. 10. Association of American geographers. Washington.

[12] Coombs, C. H. (1941). A factorial study of number ability. *Psychometrika* **6**, 161–189.

[13] Crippen, G. M. (1977). A novel approach to calculation of conformation: distance geometry. *J. Computational Physics* 24, 96–107.

[14] Crippen, G. M. (1978). Rapid calculation of coordinates from distance measures. *J. Computational Physics* **26**, 449–452.

[15] De Gruyter, D. N. M. (1967). The cognitive structure of Dutch political parties in 1967. *Report EO1-67*. Psychological Institute. University of Leiden. The Netherlands.

[16] De Leeuw, J. (1976). HOMALS. Paper presented at *Psychometric Society meeting*, Murray Hill, N. J.

[17] De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In

Progress in Statistics. North Holland Publishing Company, Amsterdam.

[18] De Leeuw, J., Young, F. W., and Takane, Y. (1976). Additive structure in qualitative data: an alternating least squares method with optimal scaling features. Psychometrika **41**, 471–503.

[19] 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.

[20] De Leeuw, J., and Pruzansky, S. (1978). A new computational method to fit the weighted Euclidean distance model. *Psychometrika* **43** 479–490.

[21] Funk, S., Horowitz, A., and Young, F. W. (1976). The perceived structure of American ethnic groups: the use of multidimensional scaling in stereotype research. *Sociometry* 39, 116–130.

[22] Gilbert, E. N. (1974). Distortion in maps. *SIAM Review* **16**, 47–62.

[23] Guttman, L. (1955). An additive metric from all the principal components of a perfect scale. *British J. Math. Statist. Psychol.* **8**, 17–24.

[24] Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika* **33**, 469–506.

[25] Harshman, R. A. (1970). Foundations of the PARAFAC procedure: models and conditions for an explanatory multi-modal factor analysis. *Unpublished thesis*, UCLA.

[26] Harshman, R. A. (1972). PARAFAC2: mathematical and technical notes. *Working papers in phonetics*, no. 22, UCLA.

[27] Kendall, D. G. (1971). Construction of maps from "odd" bits of information. *Nature* **231**, 158–159.

[28] Kendall, D. G. (1971). Maps from marriages: an application of nonmetric multidimensional scaling to parish register data. In *Mathematics in the archeological and historical sciences*. University of Edinburgh Press, Edinburgh.

[29] Kroonenberg, P., and De Leeuw, J. (1977). TUCKALS2: a principal component analysis of three mode data. *RN 01-77*. Department of Data Theory. University of Leiden. The Netherlands.

[30] Kruskal, J. B. (1964). Nonmetric multidimensional scaling. *Psychometrika* **29**, 1–27 and 115–129.

[31] Kruskal, J. B. (1977). Multidimensional scaling and other methods for discovering structure. In *Mathematical methods for digital computers*, vol III, New York, Wiley.

[32] Lalouel, J. M. (1977). Linkage mapping from pairwise recombination data. *Heredity* **38**, 61–77.

[33] Levelt, W. J. M., Van de Geer, J. P., and Plomp, R. (1966). Triadic comparisons of musical intervals. *British J. Math. Statist. Psychol.* **19**, 163–179.

[34] Lingoes, J. C., and Roskam, E. E. (1973). A mathematical and empirical analysis of two mutlidimensional scaling algorithms, *Psychometrika*, **38**. Monograph supplement.

[35] Nishisato, S. (1978). Multidimensional scaling: a historical sketch and bibliography. Ontario institute for studies in education. Toronto.

[36] Noma, E., and Johnson, J. (1977). Constrained nonmetric multidimensional scaling configurations. *Tech. Rep.* **60**. Human Performance Center, University of Michigan. Ann Arbor.

[37] Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika* **42**, 241–266.

[38] Ramsay, J. O. (1978). Confidence regions for multidimensional scaling analysis. *Psychometrika* **43**, 145–160.

[39] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton.
[40] Schönemann, P. H. (1972). An algebraic solution for a class-of subjective metrics models. *Psychometrika* **37**, 441–451.
[41] Shepard, R. N. (1974). Representation of structure in similarity data: problems and prospects. *Psychometrika* **39**, 373–421.
[42] Sjöberg, L. (1975). Choice frequency and similarity. *Psych. Rep.* **23**, University of Göteborg. Sweden.
[43] Takane, Y., Young, F. W., and De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika* **42**, 7–67.
[44] Takane, Y., Young, F. W., and De Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features. *Psychometrika* **43**, 279–282.
[45] Wilkinson, E. M. (1970). Techniques of data analysis: seriation theory. Unpublished dissertation, Cambridge University, G. B.
[46] Zangwill, W. I. (1969). *Nonlinear Programming: a Unified Approach*. Prentice Hall, Englewood Cliffs, NJ.