Recent Developments in Statistics
J.R.Barra et al., editors
© North-Holland Publishing Company,(1977)

Invited Paper

APPLICATIONS OF CONVEX ANALYSIS TO MULTIDIMENSIONAL SCALING

Jan de Leeuw
Department of Data Theory
University of Leiden
Leiden, The Netherlands

In this paper we discuss the convergence of an algorithm for metric and nonmetric multidimensional scaling that is very similar to the C-matrix algorithm of Guttman. The paper improves some earlier results in two respects. In the first place the analysis is extended to cover general Minkovski metrics, in the second place a more elementary proof of convergence based on results of Robert is presented.

1: INTRODUCTION

In multidimensional scaling (MDS) problems the data consist of m nonnegative square matrices $\Delta_1, \Delta_2, \ldots, \Delta_m$ of order n, whose elements are interpreted as measures of dissimilarity between the n objects $o_1, o_2, \ldots, o_n$, measured at m replications $r_1, r_2, \ldots, r_m$. Thus δ_{ijk} is the dissimilarity between objects o_i and o_j at replication r_k. In a psychological context the objects are often called stimuli, and the replications are defined by the dissimilarity judgments of different subjects. Moreover we assume that m nonnegative square matrices $W_1, W_2, \ldots, W_m$ of order n are given, whose elements are interpreted as weights, i.e. w_{ijk} indicates the relative importance or precision of measurement δ_{ijk}.

Multidimensional scaling techniques represent the objects $o_1, o_2, \ldots, o_n$ as points $x_1, x_2, \ldots, x_n$ in a metric space $<\Omega,d>$ in such a way that the distances $d(x_i,x_j)$ are approximately equal to the dissimilarities δ_{ijk}. We sometimes write d_{ij} for $d(x_i,x_j)$.

In this paper we study representations of $O = \{o_1, o_2, \ldots, o_n\}$ in the space of all p-tuples of real numbers, in which the metric is defined by a norm $||.||$. Thus $d_{ij} = ||x_i - x_j||$. A representation of O is then an n x p matrix X, with row i representing o_i. We also define the notation $d_{ij}(X)$ for the distance between x_i and x_j.

The loss function we use in this paper to evaluate the badness-of-fit of a particular representation X is

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

Clearly $\sigma(X) \geq 0$, and $\sigma(X) = 0$ if and only if $d_{ij}(X) = \delta_{ijk}$ for all i,j,k with $w_{ijk} \neq 0$. If $w_{ijk} = 0$ then the value of $\sigma(X)$ does not depend on δ_{ijk}. This provides us with a simple device for handling *missing data*: if the observation corresponding with the triple i,j,k is missing, then we can choose δ_{ijk} arbitrarily, and set $w_{ijk} = 0$.

The first and most basic MDS problem we study in this paper is the minimization of $\sigma(X)$ over all n x p *configuration matrices* X. This is usually called *metric* MDS, to distinguish it from the more general *nonmetric* problem in which the δ_{ijk} are only partially known. Or, more precisely, in the MDS problem as we have defined it so far the δ_{ijk} have to be either completely known or completely unknown (missing), nonmetric MDS deals with various kinds of intermediate cases. In a later section of the paper we discuss straigthforward extensions of our techniques that cover nonmetric MDS.

2: PREVIOUS WORK

Algorithms for the minimization of $\sigma(X)$ have been proposed earlier by Kruskal (1964 a,b) and Guttman (1968). In fact both Kruskal and Guttman propose algorithms to solve the more general nonmetric scaling problems. In this general nonmetric case there are substantial differences between the two approaches, but if we specialize them to the metric MDS problem they become very similar. A detailed discussion and comparison of the algorithms and the corresponding computer programs is available in Lingoes and Roskam (1973). We only discuss the main ideas, and the major differences between the two approaches in the metric case.

Kruskal proposes a gradient method of the form

$$X \leftarrow X - \alpha\nabla\sigma(X),$$

where $\nabla\sigma(X)$ is the *gradient* of σ at X, i.e. the n x p matrix of partial derivatives, and where $\alpha > 0$ is a *step-size*. Guttman on the other hand shows that the stationary equation $\nabla\sigma(X) = 0$ can be rewritten in the form $X - C(X)X = 0$, where C(X) is a square symmetric matrix valued function of X. He proposes the iterative process

$$X \leftarrow C(X)X.$$

By substituting Guttman's formula for the gradient in Kruskal's algorithm we find

$$X \leftarrow X - \alpha(X - C(X)X) = (1 - \alpha)X + \alpha C(X)X.$$

Thus Guttman's algorithm is a special case of Kruskal's with a constant step-size α equal to one. And Kruskal's algorithm can be interpreted as an over- or underrelaxed version of Guttman's algorithm. Interesting geometrical and mechanical interpretations of these algorithms have been discussed by Kruskal and Hart (1965), McGee (1966), and Gleason (1967).

There are two problems with the Kruskal-Guttman approach that specifically interest us. In the first place the distance function $d_{ij}(X)$ is typically not differentiable at all configurations X with $x_i = x_j$. This implies that a gradient method cannot be applied to $\sigma(X)$ without further specifications, it implies that the usual convergence theorems for gradient methods are invalid, and it also implies that local minimum points of $\sigma(X)$ need not satisfy the stationary equations. The second problem is that it has not been shown, for either Kruskal's "heuristic" or for Guttman's "constant" step-size procedure, that the resulting algorithms are indeed convergent. Kruskal (1969, 1971) has proved some partial results, and Guttman (1968) and Lingoes and Roskam (1973) have some heuristic arguments and some empirical results, but there is no complete convergence proof.

Until recently these problems have been ignored, or they have been "dissolved" by transforming the model and, through the model, the loss function. ALSCAL, for example, defines the loss on the squared distances and squared dissimilarities (Takane, Young, De Leeuw, 1976). Classical metric scaling methods apply both squaring and double centering to the dissimilarities, and then define the loss on the scalar products (Torgerson, 1958). These transformations do make the loss functions better behaved in some respects, but they do not really solve the problems with the Kruskal-Guttman approach, they merely transform them away. Moreover using transformations seems less direct, and does not generalize to other distance functions than the usual Euclidean one.

In this paper we derive a simple algorithm for directly minimizing $\sigma(X)$, that can easily proved to be convergent. Although the derivation of the algorithm does not use differentiation or stationary equations, it turns out that the algorithm is identical to Guttman's C-matrix method. One (modest) interpretation of the main result of this paper is that it provides a convergence proof for Guttman's algorithm. Another interpretation is that we show that the C-matrix method should not be interpreted as a gradient method. It is more natural to view it as a minimization method based on an analysis of the convexity properties of the distance function. In fact it may very well be better not to interprete Kruskal's algorithm as a gradient method, but as a relaxed version of the C-matrix method. This interpretation makes it possible, for example, to construct interesting optimal step-size procedures, that do not use heuristic arguments and several arbitrary parameters.

3: PROBLEM REDUCTION: PARTITIONING

In this section we reduce the MDS problem to a more simple form by partitioning the loss function into additive components. For this purpose we define

$$\underline{w}_{ij} = \frac{1}{m}\sum_{k=1}^{m} w_{ijk},$$

$$\underline{\delta}_{ij} = \sum_{k=1}^{m} w_{ijk}\delta_{ijk} \;/\; m\underline{w}_{ij},$$

$$w_{ij} = \tfrac{1}{2}(\underline{w}_{ij} + \underline{w}_{ji}),$$

$$\delta_{ij} = \frac{\underline{w}_{ij}\underline{\delta}_{ij} + \underline{w}_{ji}\underline{\delta}_{ji}}{\underline{w}_{ij} + \underline{w}_{ji}} \qquad \text{for } i \neq j,$$

$$w_{ii} = \delta_{ii} = 0.$$

As is customary in the analysis of variance we collect the components of the partitioning in a table.

SOURCE	LOSS COMPONENT
Proper loss	$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\delta_{ij} - d_{ij}(X))^2.$
Symmetry	$\sum_{i}\sum_{\neq j} \{\underline{w}_{ij}\underline{\delta}^2_{ij} - w_{ij}\delta^2_{ij}\}.$
Hollowness	$\sum_{i=1}^{n} \underline{w}_{ii}\underline{\delta}^2_{ii}$.
Individual differences	$\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ijk}(\delta_{ijk} - \underline{\delta}_{ij})^2.$
Total loss	$\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ijk}(\delta_{ijk} - d_{ij}(X))^2.$

It is obvious that we minimize the total loss if we minimize the proper loss, and that the proper loss is more simple. In fact in defining the proper loss we can suppose without loss of generality that both the weigths and the dissimilarities are symmetric and hollow. The only assumption we make about the δ_{ij} is that they are nonnegative. The weigths are also assumed to be nonnegative, but we make the additional nondegeneracy assumption of irreducibility: we suppose that there is no partitioning of $\{1,2,\ldots,n\}$ such that $w_{ij} = 0$ whenever i and j belong to different members of the partition. Again this assumption causes no real loss of generality, because if all between-subset weigths are zero the MDS problem separates into a number of smaller problems corresponding with each of the subsets.

4: PROBLEM MANIPULATION: USE OF HOMOGENEITY

As we have shown in the previous section the metric MDS problem can be reformulated without loss of generality as the minimization of

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

over the n x p configuration matrices X. In this section we study a closely related maximization problem, that is in some respects more simple. For the discussion of this alternative problem we need the following definitions.

$$\rho(X) = \sum_{i<j}\sum w_{ij}\delta_{ij}d_{ij}(X),$$

$$\eta^2(X) = \sum_{i<j}\sum w_{ij}d^2_{ij}(X),$$

$$\eta^2_\delta = \sum_{i<j}\sum w_{ij}\delta^2_{ij},$$

and

$$\lambda(X) = \rho(X) \;/\; \eta(X)\eta_\delta.$$

Theorem 4.1: For all X we have $0 \leqslant \lambda(X) \leqslant 1$. Moreover $\lambda(X) = 1$ if and only if the dissimilarities δ_{ij} and the distances $d_{ij}(X)$ for which $w_{ij} \neq 0$ are proportional.

Proof: This follows directly from the Cauchy-Schwartz inequality applied to $\rho(X)$. //

Theorem 4.2: a) Suppose $\hat{X}$ minimizes $\sigma(X)$. Then $\hat{X}$ also maximizes $\lambda(X)$.
b) Suppose $\hat{X}$ maximizes $\lambda(X)$. Then $\{\rho(\hat{X})/\eta^2(\hat{X})\}.\hat{X}$ minimizes $\sigma(X)$.

Proof: Because $d_{ij}(\beta X) = \beta d_{ij}(X)$ for all X and all $\beta \geq 0$ we can reformulate the MDS problem as the minimization of

$$\sum_i\sum_j w_{ij}(\delta_{ij} - \beta d_{ij}(X))^2$$

over the n x p matrices X and over all $\beta \geqslant 0$. The minimum over β for fixed X is attained at

$$\hat{\beta} = \rho(X) \;/\; \eta^2(X),$$

and the value at the minimum is $\eta^2_\delta(1 - \lambda^2(X))$. The theorem follows from these computations. //

It follows from theorem 4.2 that we can solve the metric MDS problem by finding the configuration matrix that maximizes $\lambda(X)$.

5: THE EUCLIDEAN CASE

Suppose $d_{ij}(X)$ is Euclidean, i.e.

$$d^2_{ij}(X) = \sum_{s=1}^{p} (x_{is} - x_{js})^2.$$

In this case it is convenient to derive some matrix expressions for $\rho(X)$ and $\eta(X)$. Define the matrix valued function $B(X)$ by

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

$b_{ii}(X) = \sum w_{ij}\delta_{ij}s_{ij}(X)$.

Here

$$s_{ij}(X) = \begin{cases} d_{ij}^{-1}(X) & \text{if } d_{ij}(X) \neq 0, \\ 0 & \text{if } d_{ij}(X) = 0. \end{cases}$$

We also define the matrix V by

$v_{ij} = - w_{ij}$ if $i \neq j$,

$v_{ii} = \sum w_{ij}$.

Both B(X) and V are real symmetric matrices with nonpositive off-diagonal and nonnegative diagonal elements, whose rows and columns sum to zero. By a familiar matrix theorem they are consequently both positive semi-definite of rank not exceeding n - 1. Because V is irreducible by assumption we have in fact rank(V) = n - 1, and the null space of V is the set of all vectors with constant elements. (Taussky, 1949; also Varga, 1962, sections 1.4 and 1.5). If e is the n-vector with all elements equal to one, then the Moore-Penrose inverse of V is simply

$V^{+} = (V + \frac{1}{n} ee')^{-1} - \frac{1}{n} ee'$.

The following results can be verified easily.

Theorem 5.1: a) $\rho(X) = \text{tr } X'B(X)X$.
b) $\eta^2(X) = \text{tr } X'VX$.

We also define, for all pairs of configuration matrices,

$\mu(X,Y) = \text{tr } X'B(Y)Y$.

Theorem 5.2: $\mu(X,Y) \leqslant \rho(X)$ for all X,Y.

Proof: The Cauchy-Schwartz inequality implies

$$d_{ij}(X) \geqslant s_{ij}(Y) \sum_{s=1}^{p} (x_{is} - x_{js})(y_{is} - y_{js}).$$

If we multiply both sides with $w_{ij}\delta_{ij}$, sum all inequalities, and simplify, we find the inequality stated in the theorem. //

Using the notation developed in this section we can now define the B-matrix algorithm for Euclidean metric multidimensional scaling as the recursion

$X^{k+1} = V^{+}B(X^{k})X^{k}$.

The only difference between Guttman's C-matrix and our B-matrix is due to the fact that we have removed the homogeneity from the problem in section 4, this makes B more simple than C.

Theorem 5.3: a) The three sequences $\rho(X^k)$, $\eta(X^k)$, and $\lambda(X^k)$ are bounded and increasing. The limits are ρ_∞, $\eta_\infty = \rho_\infty^{\frac{1}{2}}$, and $\lambda_\infty = \rho_\infty / \eta_\delta\eta_\infty$.

b) The sequence X^k has convergent subsequences. If X_∞ is the limit of a convergent subsequence, then $\lambda(X_\infty) = \lambda_\infty$. Moreover X_∞ is a fixed point, i.e. $X_\infty = V^+B(X_\infty)X_\infty$, and if λ is differentiable at X , then $\nabla\lambda(X_\infty) = 0$.

c) $||X^{k+1} - X^k|| \rightarrow 0$.

Proof: From the Cauchy-Schwartz inequality

$$\rho(X^k) = \text{tr } X^kVX^{k+1} \leqslant \eta(X^k)\eta(X^{k+1}).$$

From theorem 5.2

$$\rho(X^{k+1}) \geqslant \text{tr } X^{k+1}B(X^k)X^k = \text{tr } X^{k+1}VX^{k+1} = \eta^2(X^{k+1}).$$

If we combine these inequalities we obtain

$$\eta(X^k) \leqslant \frac{\rho(X^k)}{\eta(X^k)} \leqslant \eta(X^{k+1}),$$

and

$$\rho(X^k) \leqslant \frac{\eta(X^k)}{\eta(X^{k+1})} . \rho(X^{k+1}) \leqslant \rho(X^{k+1}).$$

Because $\rho(X^k) \leqslant \eta_\delta^2$ and $\eta(X^k) \leqslant \eta_\delta$ it follows that both $\rho(X^k)$ and $\eta(X^k)$ are convergent increasing sequences, with limits, say, ρ_∞ and η_∞. Because $\lambda(X^k) \leqslant 1$ it follows that $\lambda(X^k)$ is another convergent increasing sequence with limit λ_∞. Moreover $\rho_\infty = \eta_\infty^2$, and $\lambda_\infty = \rho_\infty / \eta_\delta\eta_\infty$. It also follows that the sequence X^k lies in the compact set $\eta(X) \leqslant \eta_\delta$, and has convergent subsequences. There is equality in the basic chain of inequalities if and only if X is a fixed point. This implies that subsequential limits are fixed points. Finally

$$\text{tr } (X^{k+1} - X^k)'V(X^{k+1} - X^k) = \eta^2(X^{k+1}) + \eta^2(X^k) - 2\rho(X^k) \rightarrow 2(\eta_\infty^2 - \rho_\infty) = 0,$$

which implies part c of the theorem. //

A very similar result appears in Robert (1967). The general convergence theorems of Zangwill (1969) and Meyer (1976) are also relevant. Observe that theorem 5.3 does not say that X^k converges. This follows only if we make some rather arbitrary additional assumptions, for example that there is only a finite number of fixed points, or that one of the subsequential limits is an isolated fixed point. If X^k does not converge, it follows that the set of limit points is a continuum (a result due to Ostrowski, cf Daniel, 1971, section 6.3).

Theorem 5.3 is our basic convergence theorem for metric Euclidean MDS. It is quite satisfactory, and it has been proved by very elementary methods. In fact the proof only uses some elementary properties of sequences, and the Cauchy-Schwartz inequality.

6: NONEUCLIDEAN METRICS

Our method can be generalized to general Minkovski metrics, with the metric defined by a gauge ϕ. We shall use some elementary facts about gauges without proof. The proofs follow easily from the beautiful introduction to gauges and norms in Rockafellar (1970, chapter 15). Introductions to Minkovski geometry are given in Busemann and Kelley (1953), and Busemann (1955).

A gauge is a function $\phi:R^n \to R$ satisfying

G1: $\phi(x) \geq 0$.

G2: $\phi(x) = 0$ if and only if $x = 0$.

G3: $\phi(\mu x) = \mu\phi(x)$ for all $\mu \geq 0$.

G4: $\phi(x + y) \leq \phi(x) + \phi(y)$.

A gauge is a norm if we can replace G3 by the stronger

G5: $\phi(\mu x) = |\mu|\phi(x)$ for all μ.

A gauge defines a distance function by the rule

$d_{ij}(X) = \phi(x_i - x_j)$.

Unless the gauge is a norm this distance is not necessarily symmetric. With some minor modifications our results are also valid if we replace G2 by the weaker

G6: $\phi(0) = 0$.

In fact most of the results remain valid if we only assume G3 and G4, i.e. for all homogeneous convex functions. Thus gauges are relevant for our problem because they can be used to construct very general distance functions. But they are even more relevant because of the following result.

Theorem 6.1: Both $\rho(X)$ and $\eta(X)$ are gauges on the space of all n x p configuration matrices.

Proof: For both $\rho(X)$ and $\eta(X)$ property G1 is obvious. For $\eta(X)$ property G2 follows from irreducibility, for $\rho(X)$ property G6 is obvious, we could assume G2, but we never need it. Properties G3 and G4 follows from the fact that each $d_{ij}(X)$ is convex and homogeneous on the space of configuration matrices. //

The theorem shows that the metric MDS problem reduces to the maximization of a ratio of two gauges. Problems of that type have been studied by Robert (1967), Boyd (1974), Pham Dinh Tao (1975, 1976). Before we discuss their results and apply them to our problem we state some of the elementary facts about gauges. First define the polar of a gauge as the function $\phi^o:R^n \to R$ given by

$$\phi^o(y) = \max_x \frac{<x,y>}{\phi(x)} .$$

Here <.,.> denotes inner product.

Fact 6.2: a) The polar of a gauge is a gauge, the polar of a norm is a norm.

b) The polar of the polar of a gauge ϕ is the gauge ϕ.

c) The Euclidean norm $<x,x>^{\frac{1}{2}}$ is its own polar.

d) (Hölder's inequality). If ϕ and ϕ^o are polar gauges, then $<x,y> \leq \phi(x)\phi^o(y)$ for all x,y in R^n.

We can study the conditions for equality in Hölder's inequality by introducing subdifferentials. Remember that a subgradient of a function ϕ at a point x is a vector y such that $\phi(z) \geq \phi(x) + <y,z - x>$ for all $z \in R^n$. The set of all subgradients of ϕ at a point x is the subdifferential of ϕ at x, and is written as $\partial\phi(x)$. Thus for each x the symbol $\partial\phi(x)$ stands for a subset of R^n, possibly empty. Again we mention some facts about subdifferentials, without proof. The proofs can be found in part V of Rockafellar (1970).

Fact 6.3: a) If ϕ is a finite convex function, then for each $x \in R^n$ the set $\partial\phi(x)$ is nonempty, convex, and compact.

b) If ϕ is differentiable at x with gradient $\nabla\phi(x)$, then $\partial\phi(x) = \{\nabla\phi(x)\}$.

c) The map $x \rightarrow \partial\phi(x)$ is closed, i.e. if $x_i \rightarrow x_\infty$, if $y_i \rightarrow y_\infty$, and if for each i also $y_i \in \partial\phi(x_i)$, then $y_\infty \in \partial\phi(x_\infty)$.

By combining the results of fact 6.2 and fact 6.3 we find the following results.

Fact 6.4: a) Suppose ϕ is a gauge. Then $y \in \partial\phi(x)$ if and only if $\phi(z) \geq <y,z>$ for all $z \in R^n$, and $\phi(x) = <y,x>$.

b) Suppose ϕ and ϕ^o are polar gauges. Then $x \in \partial\phi^o(y)$ if and only if $<x,y> = \phi^o(y)$ and $\phi(x) = 1$. Moreover $y \in \partial\phi(x)$ if and only if $<x,y> = \phi(x)$ and $\phi^o(y) = 1$.

c) Suppose ϕ and ϕ^o are polar gauges. Then $<x,y> = \phi(x)\phi^o(y)$ if and only if $x \in \phi(x)\partial\phi^o(y)$ if and only if $y \in \phi^o(y)\partial\phi(x)$.

Now consider the problem of maximizing the ratio

$$\lambda(x) = \frac{\phi(x)}{\psi(x)} ,$$

with both ϕ and ψ gauges. From the definitions of gauges and their polars we obtain the following result.

Theorem 6.5: Maximizing $\lambda(x)$ over R^n is equivalent to maximizing

$$\xi(x,y) \triangleq \frac{<x,y>}{\psi(x)\phi^o(y)}$$

over R^n x R^n, and this is equivalent to minimizing

$$\lambda^o(x) \triangleq \frac{\phi^o(x)}{\psi^o(x)}$$

over R^n.

By using Hölder's inequality we can derive the following necessary conditions for an extreme value.

Theorem 6.6: a) If $\hat{x},\hat{y}$ maximizes $\xi(x,y)$ then $\hat{y} \varepsilon \phi^{o}(\hat{y})\partial\phi(\hat{x})$ and $\hat{x} \varepsilon \psi(\hat{x})\partial\psi^{o}(\hat{y})$.

b) If $\hat{x}$ maximizes $\lambda(x)$ or minimizes $\lambda^{o}(x)$ then $\hat{x} \varepsilon \psi(\hat{x})\partial\psi^{o}.\partial\phi(\hat{x})$.

For maximizing $\lambda(x)$ the following algorithm was proposed by Robert (1967). We start with x^{0} such that $\psi(x^{0}) = 1$. Then define $y^{k} \varepsilon \partial\phi(x^{k})$ and $x^{k+1} \varepsilon \partial\psi^{o}(y^{k})$.

Theorem 6.7: a) The sequence $\lambda(x^{k})$ is increasing and convergent. The sequence $\lambda^{o}(y^{k})$ is decreasing and convergent. Both sequences converge to the same limit $\underline{\lambda}$.

b) All accumulation points of (x^{k},y^{k}) correspond with the same function value $\underline{\lambda} = \underline{\xi}$. Moreover all accumulation points satisfy the necessary conditions of theorem 6.6.

Proof: From our facts about gauges

$\phi^{o}(y^{k}) = 1,$

$\psi(x^{k+1}) = 1,$

$<x^{k},y^{k}> = \phi(x^{k}),$

$<x^{k+1},y^{k}> = \psi^{o}(y^{k}).$

By applying Hölder's inequality

$\phi(x^{k}) = <x^{k},y^{k}> \leqslant \psi^{o}(y^{k})\psi(x^{k}) = \psi^{o}(y^{k}),$

$\psi^{o}(y^{k}) = <x^{k+1},y^{k}> \leqslant \phi(x^{k+1})\phi^{o}(y^{k}) = \phi(x^{k+1}),$

and thus

$\lambda(x^{k}) \leqslant 1/\lambda^{o}(y^{k}) \leqslant \lambda(x^{k+1}),$

which implies part a. Because the subdifferentials are closed and the iterations remain in a compact set we can apply the general convergence theorems of Zangwill (1969) to get b. //

If we compare 6.7 and 5.3 we see that 6.7 has no part c, and is consequently weaker than 5.3. It is possible to prove that $||x^{k+1} - x^{k}|| \rightarrow 0$ in this more general context too, but we need additional assumptions. One of the more natural ones is that ϕ or ψ^{o} or both are differentiable at all stationary points, other possibilities are discussed by Meyer (1976). A far more important difference between the algorithms of sections 5 and 6 is that in most cases the function ψ^{o} cannot be computed in closed form. The same thing is true for the subdifferential $\partial\psi^{o}$. This means that we must compute x^{k+1} by maximizing $<x,y^{k}>$ over $\{x \mid \psi(x) = 1\}$. This is a convex programming problem, which cannot be solved in a finite number of steps in general. Consequently we need a version of theorem 6.7 in which this convex programming problem is truncated after a finite number of steps. Zangwill's general convergence theory shows how this truncating should be done.

The simplifications in section 5 are possible, because the gauge ψ is ellipsoidal in this case. If both ψ and ϕ are ellipsoidal, then Robert has pointed out that the algorithm reduces to the power method for solving a generalized eigenvalue problem. Compare also Pham Dinh Tao (1976). It is of some interest that Guttman already pointed out that his C-matrix method for MDS looked like a sort of generalized power method. The analysis in this paper shows what the exact relationships are.

7: NONMETRIC MDS

In the simplest forms of nonmetric MDS we must minimize

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

over all n x p configuration matrices X and over all n x n disparity matrices Δ. The disparity matrices must be chosen from a known convex cone Γ, the metric MDS problem is the special case in which Γ is a ray, the additive constant problem is the special case in which Γ is a two-dimensional subspace. We briefly indicate the modifications needed to apply our ideas to nonmetric MDS in this simple form. More complicated partitioned loss functions, with more complicated normalizations, will be discussed in subsequent publications.

By using the homogeneity of the distance function as in section 4 we can show that the nonmetric MDS problem is equivalent to the maximization of

$$\lambda(X,\Delta) = \frac{\rho(X,\Delta)}{\eta(X)\eta(\Delta)},$$

with

$$\rho(X,\Delta) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij}d_{ij}(X),$$

$$\eta^2(\Delta) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij}^2,$$

and $\eta(X)$ as before. If we define

$$\rho(X) = \max_{\Delta\,\varepsilon\,\Gamma} \frac{\rho(X,\Delta)}{\eta(\Delta)},$$

then $\rho(X)$ is a homogeneous convex function, in fact a gauge. Thus we have a problem of the familiar form, a ratio of gauges must be maximized, and the algorithm of section 6 can be applied.

If the distance function is Euclidean, the analysis of section 5 can be used. The only difference with metric MDS is in the definition of $\rho(X)$, in the nonmetric case we have to compute the optimum Δ for given X in order to compute $\rho(X)$. We can compute the optimum $\hat{\Delta}(X)$ as the unique minimizer of

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

over the cone Γ. After solving this regression problem we can normalize the solution such that $\eta(\hat{\Delta}(X)) = 1$, but this is not strictly necessary. The B-matrix algorithm for nonmetric Euclidean MDS is defined as the recursion

$$X^{k+1} = V^{+}B(X^{k},\hat{\Delta}(X^{k}))X^{k},$$

with $B(X,\hat{\Delta}(X))$ defined as $B(X)$, but with $\hat{\delta}_{ij}(X)$ substituted for δ_{ij}. The same inequalities and equations can be derived as in 5.1, 5.2, and the proof of 5.3.

Theorem 7.1: Parts a,b,c of theorem 5.3 are also true for the nonmetric Euclidean B-matrix algorithm.

In the nonmetric case the differences between our B-matrix method and Guttman's C-matrix method are larger than in the metric case. One important reason is that Guttman uses rank images, while our convexity approach forces us to use monotone regression estimates of the δ_{ij}. I have not been able to find a rigorously defined optimization problem in which rank images can be used. This does not mean, of course, that we cannot use rank images in the earlier iterations of an MDS algorithm. In the earlier iterations we can do anything we please. As in TORSCA we can use the semi-nonmetric Young-Householder process, or as in MINISSA we can use rank images. We only have to switch to monotone regression and the B-matrix algorithm if things are getting out of hand (if the loss starts to increase, for example).

REFERENCES

Busemann, H. The geometry of geodesics. New York, Academic Press, 1955.

Busemann, H. and Kelly, P. Projective geometry and projective metrics. New York, Academic Press, 1953.

Daniel, J.W. The approximate minimization of functionals. Englewood Cliffs, Prentice Hall, Inc, 1971.

Gleason, T.C. A general model for nonmetric multidimensional scaling. Michigan mathematical psychology program, 1967, 3.

Guttman, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika, 1968, 33, 469 - 506.

Kruskal, J.B. Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. Psychometrika, 1964, 29, 1-28.

Kruskal, J.B. Nonmetric multidimensional scaling: a numerical method.
Psychometrika, 1964, 29, 115-129.

Kruskal, J.B. A new convergence condition for methods of ascent.
Unpublished memo, Bell Labs, Murray Hill, 1967.

Kruskal, J.B. Monotone regression: continuity and differentiability properties.
Psychometrika, 1971, 36, 57-62.

Kruskal, J.B. and Hart, R.E. A geometric interpretation of diagnostic data from a digital machine.
Bell System Technical J., 1966, 45, 1299-1338.

Lingoes, J.C. and Roskam, E.E. A mathematical and empirical analysis of two multidimensional scaling algorithms.
Psychometrika, 1973, 38, monograph supplement.

McGee, V.E. The multidimensional analysis of 'elastic' distances.
Brit. J. Math. Statist. Psychol., 1966, 19, 181-196.

Meyer, R.R. Sufficient conditions for the convergence of monotonic mathematical programming algorithms.
J. Comp. System Sciences, 1976, 12, 108-121.

Pham Dinh Tao Eléments homoduaux d'une matrice A relatifs à un couple des normes (ϕ,ψ). Applications au calcul de $S_{\phi\psi}(A)$.
Séminaire d'analyse numérique, Grenoble, 1975, no 236.

Pham Dinh Tao Calcul du maximum d'une forme quadratique définie positive sur la boule unité de ψ_∞.
Séminaire d'analyse numérique, Grenoble, 1976, no 247.

Robert, F. Calcul du rapport maximal de deux normes sur R^n.
R.I.R.O., 1967, 1, 97-118.

Rockafellar, R.T. Convex analysis.
Princeton, Princeton University Press, 1970.

Takane, Y, Young, F.W., and De Leeuw, J. Nonmetric individual differences multidimensional scaling.
Psychometrika, 1976, 41, in press.

Taussky, O. A recurring theorem on determinants.
Amer. Math. Monthly, 1949, 56, 672-676.

Torgerson, W. Theory and methods of scaling.
New York, Wiley, 1958.

Varga,R.S. Matrix iterative analysis.
Englewood Cliffs, Prentice Hall, 1962.

Zangwill, W.I. Nonlinear programming: a unified approach.
Englewood Cliffs, Prentice Hall, 1969.