CONVERGENCE OF THE MAJORIZATION ALGORITHM FOR MULTIDIMENSIONAL SCALING

Jan de Leeuw
Department of Data Theory FSW/RUL
Middelstegracht 4
2312 TW Leiden
Netherlands

1: Introduction

In this paper we study the convergence properties of a multidimensional scaling algorithm proposed by Guttman (1968), and studied earlier by De Leeuw (1977), De Leeuw and Heiser (1977, 1980). We unify and extend earlier qualitative results on convergence, and we give some new quantitative results. In a subsequent publication these quantitative results will be used to discuss various step-size procedures which have the purpose of accelerating the convergence.

The paper will be mainly about metric multidimensional scaling, towards the end we shall briefly discuss the nonmetric case. Results for nonmetric scaling are mainly simple extensions of the metric results, and in this sense metric scaling is more basic. Research in multidimensional scaling has moved in the direction of proposing more and more complicated models, with a very large number of parameters and sometimes even with severe discontinuities in the model. The implicit assumption in most of this work from an algorithmic point of view is that the gradient method is powerful enough to carry the burden of these very complicated models, because it can find down-hill directions everywhere. It seems to us that this implicit assumption is far too optimistic. Gradient-based methods have more or less failed in parametric mapping (Shepard and Carroll, 1966), in polynomial factor analysis (Carroll, 1972), in nonmetric unfolding (Heiser, 1981), and in the fitting of hierarchical or additive clusters to data (Carroll and Pruzansky, 1980, Arabie and Carroll, 1980). It is true that in some cases gradient-based methods have produced impressive results, even with these very complicated models, but in general they

must be combined with heuristics, very good initial estimates, interactive decisions along the way, multiple starts, and so on. Failure in this context must not be interpreted absolutely, it only means that for the complicated models it has not been possible to construct programs that can be used on a routine basis.

Development of multidimensional scaling programs has been largely pragmatic, perhaps even commercial. The emphasis has been on producing programs that work, at least in some cases, and comparatively little effort has been paid to theoretical problems associated with the loss functions and the algorithms that were used. Exceptions are the various Monte Carlo studies that have been published, but these have, by definition, only a suggestive and tentative nature. We think that a more theoretical study of loss functions and algorithms is long overdue. In fact we think that at the moment an in-deep study of some of the more simple models is more urgent than development of even more complicated models or sophisticated statistical superstructures.

2: Notation and terminology

We introduce the notation and terminology for metric multidimensional scaling. This is more or less standard (Kruskal and Wish, 1978). The data in a multidimensional scaling problem are collected in a symmetric nonnegative matrix $\Delta = \{\delta_{ij}\}$. The elements of Δ are called dissimilarities, δ_{ij} is the dissimilarity between objects i and j. There are n objects, thus Δ is of order n. We suppose that self-dissimilarities are zero,

as a consequence Δ is hollow (has a zero diagonal). The purpose of multidimensional scaling is to represent the objects as points in a low-dimensional Euclidean space, in such a way that the distance between points i and j is approximately equal to the dissimilarity of objects i and j. Thus x_i are n points in p-space, they are collected in the n x p matrix X, also called the configuration. The matrix D(X), with elements $d_{ij}(X)$, contains the Euclidean distances between the x_i. It follows that

$$d_{ij}^2(X) = (x_i - x_j)'(x_i - x_j) = \tag{1a}$$

$$= (e_i - e_j)'XX'(e_i - e_j) = \tag{1b}$$

$$= \text{tr } X'A_{ij}X. \tag{1c}$$

In (1b) the e_i are unit vectors (columns of the identity matrix of order n), in (1c) we have $A_{ij} = (e_i - e_j)(e_i - e_j)'$. In order to find out how successfull a representation is we compute the value of the loss function

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

where $W = \{w_{ij}\}$ is a symmetric, hollow, nonnegative matrix of weights. The purpose of metric multidimensional scaling can now be described more precisely: given weights, dissimilarities, and a dimensionality p, we want to find X that minimizes $\sigma(X)$.

2: Basic algorithm

The algorithm we discuss in this paper was first given by Guttman (1968). He derived it by setting the stationary equations for the minimization of $\sigma(X)$ equal to zero, and he observed that the algorithm

was a gradient method with constant step-size. Compare also Lingoes and Roskam (1973, p 8-10), Hartmann (1979, p 74-82), Borg (1981, p 88-92). In De Leeuw (1977) the very same algorithm was derived in a somewhat more general context from convex analysis as a subgradient method. It was observed that in the simple Euclidean case, which is the one we are interested in, the algorithm could be derived from the Cauchy-Schwartz inequality, without using differentiation or sub-differentiation. This is the derivation we present here, in a simplified version taken essentially from De Leeuw and Heiser (1980). In order to describe the algorithm we need some additional notation. In the first place

$$\eta^2(X) = \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} d_{ij}^2(X), \tag{3}$$

and

$$\rho(X) = \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \delta_{ij} d_{ij}(X). \tag{4}$$

If we assume, without loss of generality, that

$$\tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \delta_{ij}^2 = 1, \tag{5}$$

then

$$\sigma(X) = 1 - 2\rho(X) + \eta^2(X). \tag{6}$$

It is clear $\eta^2(X)$ is a convex quadratic function of X. From (1c) and (3) we have

$$\eta^2(X) = \text{tr } X'VX, \tag{7}$$

with

$$V = \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} A_{ij}. \tag{8}$$

In the sequel we assume that V is irreducible, which can be done without loss of generality (De Leeuw, 1977). In this context irreducibility is equivalent to the assumption that V has rank n - 1, or that the null space of V consists of the vectors with constant elements. Observe that from (8) we directly see that V is positive semi-definite.

The function $\rho(X)$ is somewhat more complicated than $\eta^2(X)$. We know that $d_{ij}(X)$ is a convex and positively homogeneous function of X. Thus, by (4), the same thing is true for $\rho(X)$. It is convenient to write $\rho(X)$ as

$$\rho(X) = \text{tr } X'B(X)X, \tag{9}$$

with

$$B(X) = \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\delta_{ij}s_{ij}(X)A_{ij}, \tag{10}$$

where $s_{ij}(X) = 1/d_{ij}(X)$ if $d_{ij}(X) \neq 0$, and $s_{ij}(X) = 0$ otherwise. It is obvious that the difference between $\rho(X)$ and $\eta^2(X)$ is that V in (7) is a constant matrix, while B(X) in (9) is not. Thus $\eta^2(X)$ is quadratic in X, while $\rho(X)$ is not.

With the notation developed so far it is easy to explain the algorithm. Clearly $\nabla\eta^2(X) = 2VX$. Using the definitions it is also not difficult to see that $\nabla\rho(X) = B(X)X$, provided that $\rho(X)$ is differentiable at X. Thus $\nabla\sigma(X) = 2(VX - B(X)X)$. In De Leeuw and Heiser (1980) the Guttman transform of a configuration Y is defined as

$$\overline{Y} = V^{+}B(Y)Y, \tag{11}$$

with V^+ the Moore-Penrose inverse of V. Observe that the Guttman transform depends on weights and dissimilarities, and is consequently derived relative to a particular metric multidimensional scaling problem. Using the Guttman transform makes it possible to write $\nabla\sigma(X) = 2V(X - \overline{X})$, and thus $\nabla\sigma(X) = 0$ if and only if $X = \overline{X}$. This immediately suggests the algorithm $X_{k+1} = \overline{X_k}$, or more explicitly

$$X_{k+1} = V^+B(X_k)X_k. \tag{12}$$

Equivalently we can also write

$$X_{k+1} = X_k - \tfrac{1}{2}V^+\nabla\sigma(X_k), \tag{13}$$

which shows the gradient interpretation of the algorithm.

In this standard derivation of the algorithm we have made the provision that $\rho(X)$ had to be differentiable at X. This is true if and only if $d_{ij}(X) > 0$ for all i,j for which $w_{ij}\delta_{ij} > 0$. Let us agree to call a configuration <u>usable</u> if this condition is true. Thus if X is not usable, then $\sigma(X)$ is not differentiable at X, and interpretation (13) can not be used. Using definition (10) the iteration (12) can still be carried out. In De Leeuw (1977) and De Leeuw and Heiser (1980) it is shown that we do not need differentiability if we remember that by (6) our loss function is the difference of two convex functions. We use the concept of a <u>subdifferential</u> (Rockefellar, 1970, Clarke, 1975, 1981) to show that (13) can be written more generally as

$$X_{k+1} \in X_k - \tfrac{1}{2}V^+\partial\sigma(X_k), \tag{14}$$

with $\partial\sigma(X)$ the subdifferential of σ at X. For definitions and details we refer to the original publications. Although this generalization is elegant we do not use it in this paper, for two reasons. Although we

do not need differentiability for the qualitative study of convergence (i.e. for proving that the algorithm converges) we do need it for the quantitative study (i.e. for establishing the rate of convergence). And, in the second place, it was proved by De Leeuw (1984) that if σ has a local minimum at X, then X is usable. Since we are interested in the behaviour of our algorithm in the neighborhood of a local minimum anyway, this fortunate result shows that (14) is not strictly necessary.

Although we do not need the subdifferential in the definition of the algorithm we do use it in our convergence proofs. In the simple case we deal with in this paper this is not apparent, because the subdifferential inequality which is basic to the convergence proof is simply the Cauchy-Schwartz inequality. This is illustrated in the proof of the lemma, which is the foundation of our approach to multidimensional scaling.

Lemma 1: For all X and Y

$$\sigma(X) \leqslant 1 - \eta^2(\overline{Y}) + \eta^2(X - \overline{Y}). \tag{15a}$$

Moreover for all X

$$\sigma(X) = 1 - \eta^2(\overline{X}) + \eta^2(X - \overline{X}). \tag{15b}$$

Proof: By Cauchy-Schwartz

$$\text{tr } X'A_{ij}Y \leqslant \{\text{tr } X'A_{ij}X\}^{\frac{1}{2}}\{\text{tr } Y'A_{ij}Y\}^{\frac{1}{2}} = d_{ij}(X)d_{ij}(Y). \tag{16}$$

Multiply both sides by $w_{ij}\delta_{ij}s_{ij}(X)$ and add over all i,j. This gives, using (4) and (10),

$$\text{tr } X'B(Y)Y \leqslant \rho(X), \tag{17}$$

with equality if $Y = X$. We can also write (17) as

$$\rho(X) \geqslant \text{tr } X'V\overline{Y}; \tag{18}$$

and substitution in (6) gives

$$\sigma(X) \leqslant 1 - 2 \text{ tr } X'V\overline{Y} + \text{tr } X'VX. \tag{19}$$

We can write (19) as

$$\sigma(X) \leqslant 1 - \text{tr } \overline{Y}'V\overline{Y} + \text{tr } (X - \overline{Y})'V(X - \overline{Y}), \tag{20}$$

which is (15a). We had equality if $X = Y$, which is (15b).

Q.E.D.

Figure 1 provides a useful interpretation of lemma 1, and also shows the algorithmic implications. If we use the abbreviation $\omega_Y(X) = 1 - \eta^2(\overline{Y}) + \eta^2(X - \overline{Y})$, then for each Y the function ω_Y is quadratic in X. Moreover, by lemma 1, $\sigma(X) \leqslant \omega_Y(X)$ and $\sigma(Y) = \omega_Y(Y)$. Thus for each Y the function ω_Y majorizes the function σ, and they touch for $X = Y$. Moreover ω_Y is minimized over X by setting $X = \overline{Y}$, and thus we see that the Guttman transform decreases the loss function. We write this as

$$\sigma(\overline{Y}) \leqslant \omega_Y(\overline{Y}) < \omega_Y(Y) = \sigma(Y), \tag{21}$$

provided that $Y \neq \overline{Y}$. All this is illustrated in figure 1. Thus either $Y = \overline{Y}$, in which case $\nabla\sigma(Y) = 0$ and we have found a solution of the stationary equations, or $Y \neq \overline{Y}$ and we can decrease the loss by replacing Y by its Guttman transform. This constitutes the basic algorithm.

Insert figure 1 about here

4: Sequences generated by the algorithm

The algorithm (12) generates a sequence X_k, but also sequences of real numbers $\sigma_k = \sigma(X_k)$, $\rho_k = \rho(X_k)$, $\eta_k^2 = \eta^2(X_k)$. Also define

the sequence $\lambda_k = \rho(X_k)/\eta(X_k)$, and the sequence $\varepsilon_k^2 = \eta^2(X_k - \overline{X}_k)$. We first study these sequences of real numbers.

Theorem 1: a) $\rho_k \uparrow \rho_\infty$,

b) $\eta_k^2 \uparrow \eta_\infty^2 = \rho_\infty$,

c) $\lambda_k \uparrow \lambda_\infty = \eta_\infty$,

d) $\sigma_k \downarrow \sigma_\infty = 1 - \rho_\infty$,

e) $\varepsilon_k \to 0$.

Proof: From (3)(4)(5) we find, by using Cauchy-Schwartz, that $\rho(X) \leq \eta(X)$ for all X. Thus $\lambda_k \leq 1$ for all k. We can write (9) as $\rho(X) = \text{tr } X'V\overline{X}$, and again by Cauchy-Schwartz, $\rho(X) \leq \eta(X)\eta(\overline{X})$ for all X. Thus $\rho_k \leq \eta_k\eta_{k+1}$ and also $\lambda_k \leq \eta_{k+1}$. Now write (17) as $\rho(\overline{X}) \geq \text{tr } \overline{X}'B(X)X = \eta^2(\overline{X})$, which implies that $\rho_k \geq \eta_k^2$ and $\lambda_k \geq \eta_k$. Summarizing the results so far gives

$$\eta_k \leq \lambda_k \leq \eta_{k+1} \leq \lambda_{k+1} \leq 1, \tag{22a}$$

$$\eta_k^2 \leq \rho_k \leq \eta_k\eta_{k+1} \leq \eta_{k+1}^2 \leq \rho_{k+1} \leq 1. \tag{22b}$$

These chains are enough to prove parts (a)(b)(c). We have already proved the decrease of σ_k in (21), the limit value follows trivially from $\sigma_k = 1 - 2\rho_k + \eta_k^2$ together with $\rho_\infty = \eta_\infty^2$. This proves (d). For (e) we write $\varepsilon_k^2 = \eta_k^2 + \eta_{k+1}^2 - 2\rho_k$, and again use $\rho_\infty = \eta_\infty^2$.

Q.E.D.

Observe that the theorem does <u>not</u> state that ε_k decreases <u>monotonically</u> to zero. In fact it usually does not. The theorem also does not say that X_k converges. The sequence X_k is studied in a

separate theorem.

Theorem 2: Suppose $\mathbf{S}_\infty$ is the set of all accumulation points of the sequence X_k. Then:

a) S_∞ is nonempty,

b) if $X_\infty \in S_\infty$ then $\sigma(X_\infty) = \sigma_\infty$,

c) if $X_\infty \in S_\infty$ and X_∞ is usable then $X_\infty = \overline{X_\infty}$,

d) if S_∞ is not a singleton, then it is a continuum.

Proof: All X_k are column-centered. In the $p(n - 1)$ dimensional space of all column-centered $n \times p$ matrices the function η defines a norm. Because $\eta_k \leq 1$ for all k it follows that all X_k are in the unit ball of this normed space, consequently they have at least one accumulation point. This proves (a).

Suppose X_ℓ is a subsequence converging to X_∞. Then by continuity of σ sequence $\sigma(X_\ell)$ converges to $\sigma(X_\infty)$. But the only accumulation point of $\sigma(X_k)$ is σ_∞. This proves (b).

In the neighborhood of a usable configuration the Guttman transform is continuous. Thus, by the same argument that proved (b), we find that $\varepsilon(X_\ell) = \eta(X_\ell - \overline{X_\ell})$ converges to $\eta(X_\infty - \overline{X_\infty})$, which must be zero. This proves (c).

For result (d) we remember that a continuum is a closed set which cannot be written as the union of two or more disjoint closed sets. A proof of (d) is given by Ostrowski (1966, theorem 28.1).

Q.E.D.

Observe that theorem 2 does not say that X_k converges. This is

however quite irrelevant from a practical point of view. If we define κ-optimal configurations as those configurations for which $\eta(X - \overline{X}) < \kappa$, then for all $\kappa > 0$ the algorithm finds a κ-optimal configuration in a finite number of steps. This is all the convergence we ever need in practice. In theorem 2 there is a restriction in part (c), because we require that X_∞ is usable. This restriction can be removed easily by using subgradients (De Leeuw and Heiser, 1980). The condition of stationarity in this case is $0 \in \partial\sigma(X_\infty)$ or $X_\infty \in V^+\partial\rho(X_\infty)$. The generalization is again not very important in practice, because all local minima are usable (De Leeuw, 1984).

The conclusion from theorems 1 and 2 is that the sequences of loss and fit values generated by the algorithm converge monotonically. The difference between successive solutions converges to zero, which means that for all practical purposes the sequence of solutions converges. In a precise mathematical sense we have either convergence to a single point, or convergence to a continuum of stationary points, with all these stationary points having the same value of the loss function. We have not been able to exclude this last possibility, we also indicate that a natural candidate for such a continuum is available in any multidimensional scaling problem. If X_∞ is a stationary point, then for any p x p rotation matrix K the matrix $X_\infty K$ is also a stationary point, with the same loss function value. There is the possibility that X_k converges to a continuum of the form $S_\infty = \{X \mid X = X_\infty K\}$, in the sense that $\min \{\eta(X_k - X) \mid X \in S_\infty\}$ converges to zero, but X_k does not converge to a point in S_∞. Again it is clear that this is irrelevant from a practical point of view.

§: Derivatives of the Guttman-transform

A more detailed study into the convergence behaviour of the algorithm is possible if we determine the derivative of the Guttman transform. For a general discussion of the role of the derivative in one-step stationary iterative processes, such as our (12), we refer to Ostrowski (1966, chapter 22) or Ortega and Rheinboldt (1970, chapter 10). Throughout this section we suppose that X is usable. The derivative is considered as a linear operator mapping the $(n - 1)p$ dimensional space of column-centered configurations into itself. We can represent it by an $np \times np$ matrix in the computer, but here we prefer to give it as a linear map which associates with each centered configuration Y another centered configuration $\Gamma_X(Y)$. The map Γ_X is the derivative of the Guttman transform at the usable configuration X. Thus we have, for example,

$$\overline{X + Y} = \overline{X} + \Gamma_X(Y) + o(\eta(Y)), \tag{23}$$

showing the local linear approximation provided by the derivative. Observe that in (23) we have chosen η as the norm we are using in defining the derivative. This is not essential, of course, but it certainly is convenient.

We now give the formula for the derivative. Column s of $\Gamma_X(Y)$ is

$$\{\Gamma_X(Y)\}_s = V^+\{B(X)y_s - \sum_{t=1}^{p} H_{st}(X)y_t\}, \tag{24}$$

with

$$H_{st}(X) = \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{w_{ij}\delta_{ij}}{d_{ij}^3(X)} (x_{is} - x_{js})(x_{it} - x_{jt})A_{ij}. \tag{25}$$

Another useful way to write (24) is

$$\Gamma_X(Y) = V^+\{B(X)Y - U(X,Y)X\}, \tag{26}$$

with

$$U(X,Y) = \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{w_{ij}\delta_{ij}c_{ij}(X,Y)}{d_{ij}^3(X)}A_{ij}, \tag{27}$$

and $c_{ij}(X,Y) = \text{tr } X'A_{ij}Y$. Thus $U(X,X) = B(X)$.

Next we are interested in the eigenvalues and eigenvectors of Γ_X. Observe we have interpreted it as an operator on the $(n - 1)p$ dimensional space of centered configurations. It consequently has $(n - 1)p$ eigenvalues, not necessarily all of them different. If we consider Γ_X as an operator on the space of all np matrices, then it has p additional eigenvalues equal to zero. The corresponding eigensubspace contains the solutions of $\eta(Y) = 0$.

Before we proceed we have to single out one special case. If $p = 1$ then $H_{st}(X) = B(X)$ and thus $\Gamma_X(Y) = 0$. It is shown in De Leeuw and Heiser (1977) that the Guttman transform iterations converge in a finite number of steps if $p = 1$. This special case was already singled out by Guttman (1968). Defays (1978) and Heiser (1981) show that one-dimensional scaling is essentially a combinatorial problem. Because the result $\Gamma_X(Y) = 0$ stops all considerations having to do with rate of convergence, we assume from now on that $p > 1$.

Result 1. X is an eigenvector of Γ_X with eigenvalue zero. This follows directly from $U(X,X) = B(X)$, because (26) then says $\Gamma_X(X) = 0$.

Result 2. Γ_X has simple structure, i.e. $(n - 1)p$ linearly independent eigenvectors. This follows because $\Gamma_X(Y) = \lambda Y$ can be

written in the form $(\nabla^2\rho(X))Y = \lambda VY$, where $\nabla^2\rho$ is the operator corresponding to the second partials of ρ, which is consequently symmetric. Thus the eigenvalues are real, and the eigenvectors Y_s can be chosen such that $\text{tr } Y_s'VY_t = \delta^{st}$.

Result 3. The eigenvalues of Γ_X are non-negative. This follows from the representation in the previous result. Because ρ is convex the operator $\nabla^2\rho$ is positive semidefinite. A more direct proof is also possible. The eigenvalues of Γ_X are the stationary values of the Rayleigh-quotient $\text{tr } Y'V\Gamma_X(Y)/\eta^2(Y)$, whose numerator is

$$\text{tr } Y'V\Gamma_X(Y) = \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{w_{ij}\delta_{ij}}{d_{ij}(X)}\{d_{ij}^2(Y) - \frac{c_{ij}^2(X,Y)}{d_{ij}^2(X)}\}, \tag{28}$$

which is obviously nonnegative by Cauchy-Schwartz.

Result 4. If X is a local minimum of σ then all eigenvalues of Γ_X are less than or equal to one. This follows because if X is a local minimum then we must have $\nabla^2\sigma(X)$ positive semidefinite. This is true if and only if $I - \Gamma_X$ is positive semidefinite, i.e. if and only if all eigenvalues of Γ_X are less than or equal to one.

Result 5. Suppose $\lambda_{++}(X)$ is the maximum of the Rayleigh quotient $\text{tr } Y'B(X)Y/\eta^2(Y)$. Then all eigenvalues of Γ_X are less than or equal to $\lambda_{++}(X)$. This follows easily from (28), because obviously $c_{ij}^2(X,Y) \geq 0$, and thus $\text{tr } Y'V\Gamma_X(Y) \leq \text{tr } Y'B(X)Y$.

Result 6. If X is stationary, then Γ_X has $\frac{1}{2}p(p-1)$ eigenvalues equal to one. Choose $Y = XS$, with S anti-symmetric. Then $c_{ij}(X,Y) = \text{tr } X'A_{ij}XS = 0$, and thus from (26) we have $\Gamma_X(Y) = V^+B(X)XS = XS = Y$. The anti-symmetric matrix S can be chosen in $\frac{1}{2}p(p-1)$ linearly independent ways.

Result 7. If p is even, then $\lambda_+(\overline{X})$, the largest eigenvalue of $\Gamma_{\overline{X}}$, satisfies $\lambda_+(\overline{X}) \geq 1$. This follows if we take $Y = \overline{X}S$, with S anti-symmetric

as well as orthonormal, which is possible for all even p. Then tr $Y'V\Gamma_{\overline{X}}(Y) = \rho(\overline{X})$, and we know from the proof of theorem 1 that $\rho(\overline{X}) \gtrsim \eta^2(\overline{X})$. Thus for all X_k generated by our algorithm we also have $\lambda_+(X_k) \gtrsim 1$.

Result 8. If $\Gamma_X(Y) = \lambda Y$ and K is a rotation matrix, then $\Gamma_{XK}(YK) = \lambda YK$. This follows directly from (26).

Result 9. If $\Gamma_X(Y) = \lambda Y$, if X is stationary, and if $\lambda \neq 1$, then Y'VX is symmetric. This is because result 6 implies that tr Y'VXS = 0 for all antisymmetric S, which shows that Y'VX is symmetric.

6: A small example

The example we use has n = 4 objects, and has all dissimilarities δ_{ij}, with $i \neq j$, equal to $\frac{1}{6}\sqrt{6}$. The weights w_{ij} are all equal to one, and consequently normalization (5) is true.

Stationary point 1. Four points equally spaced on a line. This has σ equal to .1666666667. The matrix $V^+B(X) = \frac{1}{4}B(X)$ has eigenvalues 0, 1, 1.5, 1.8333 . The partials Γ_X depend on p. If p = 1 then $\Gamma_X = 0$, as we have seen. If p = 2, i.e. if there are four points on a line in two space, then $\Gamma_X = 0 + \frac{1}{4}B(X)$, and the eigenvalues are four zeroes plus the four eigenvalues of $\frac{1}{4}B(X)$. Thus for p = 2 the four points on a line are not a local minimum, while it is easy to see that for p = 1 they are. In fact for p = 1 they even give the global minimum. Observe that these results are true for all permutations of four equally spaced points. Thus we actually have a whole family of stationary points here.

Stationary point 2. Three points form an equilateral triangle, the fourth one is in the centroid of the triangle. Again this is actually a family of stationary points, all with σ = .06698729811. The eigenvalues

of $\frac{1}{4}B(X)$ are 0, 1, 1, 1.4641. For p = 2 the operator Γ_X has 8 eigenvalues. Three are equal to zero, three are equal to one, and two are equal to .2321. Remember that here we interprete Γ_X as an operator on R^8, if we let it operate on the centered configurations it has 6 eigenvalues because two zeroes are left out. It follows that the triangle plus centroid defines a local minimum but not an isolated one. There are at least three orthogonal directions Y for which (23) gives $\overline{X + \varepsilon Y} = X + \varepsilon Y + o(\varepsilon)$. Only one of them corresponds with the trivial (i.e. data-independent) unit eigenvalue of result 6 in the previous section. If p = 3 we have the eight eigenvalues mentioned earlier, plus the four eigenvalues of $\frac{1}{4}B(X)$, which shows that now X imbedded in three-space is not a local minimum any more.

Stationary point 3. Four points in the corners of a square. Loss is .02859547921. Eigenvalues of $\frac{1}{4}B(X)$ are 0, 1, 1, 1.1716. For p = 2 the six eigenvalues of Γ_X are zero, .4142, three times .5858, and one. The unit eigenvalue is trivial, there are no nontrivial unit eigenvalues. For the eigenvector with the trivial unit eigenvalue we have $X + \varepsilon Y = X(I + \varepsilon S)$ for S anti-symmetric. But $(I + \varepsilon S)'(I + \varepsilon S) = I + o(\varepsilon)$, and thus $(I + \varepsilon S)$ is a rotation matrix to the order we are interested in. This shows that the trivial unit eigenvalues correspond with rotations of the stationary solution X. The manifold of rotations of the square is isolated from other stationary values. Again the square is not a local minimum for p = 3, although it remains stationary.

Stationary point 4. Four points in the corners of a regular tetrahedron. Loss is zero. Eigenvalues of Γ_X are zero, three times $\frac{1}{2}$, two times $\frac{3}{4}$, and three times one. The unit eigenvalues correspond with the manifold of rotations. Clearly this solution is the global minimum.

7: Rate of convergence

The discussion in sections 4 and 5 shows that at least part of the difficulty with proving actual convergence of our iterations comes from the rotational indeterminacy of multidimensional scaling. Because of this rotational indeterminacy Γ_X has at least $\frac{1}{2}p(p-1)$ unit eigenvalues at a stationary point X. If we eliminate rotational indeterminacy, we eliminate these difficulties. First we call a regular stationary point X isolated if Γ_X has exactly $\frac{1}{2}p(p-1)$ unit eigenvalues. We call it an isolated local minimum if all eigenvalues are less than or equal to one. In the small example in the previous section the four points in the corner of a square are an isolated local minimum for p = 2. For p = 3 it is neither isolated, nor a local minimum. The equilateral triangle, with centroid, is a local minimum for p = 2, but not an isolated local minimum. The regular tetrahedron is an isolated local minimum for p = 3. At an isolated local minimum we use the symbol κ for the largest eigenvalue less than one. We call it the level of the isolated local minimum.

Theorem 3: If the accumulations points of X_k are all isolated local minima with the same level κ, then $\varepsilon_{k+1}/\varepsilon_k \to \kappa$.

Proof: We want to apply the general theorems of Ostrowki and Ortega and Rheiboldt referred to above. First we eliminate rotational indeterminacy by defining a new sequence X_k^o. For each k the configuration X_k^o is a rotation of X_k, moreover it is a specific rotation which identifies the configuration uniquely in the manifold of rotations. We can rotate to principal components, for example, with some special

provision for equal eigenvalues. Because X_k^o is a rotation of X_k, the sequences σ_k, ρ_k, η_k, λ_k generated by this modified algorithm are exactly the same. So is $\varepsilon_k^2 = \eta_{k+1}^2 + \eta_k^2 - 2\rho_k$, although now $\varepsilon_k \neq \eta(X_{k+1}^o - X_k^o)$. The transformation $\overset{o}{X}$, which maps X_k^o into X_{k+1}^o, has a derivative Γ_X^o at a stationary point with exactly the same eigenvalues as Γ_X, except for the $\frac{1}{2}p(p-1)$ unit eigenvalues, which are replaced by zeroes. Thus $\kappa < 1$ is actually the largest eigenvalue of Γ_X^o, which means that X_k^o converges linearly, with rate κ. It is clear, by the way, that the theorem remains true if we merely suppose that one of the accumulation points is an isolated local minimum. Because of theorem 2 in that case they all are.

Q.E.D.

If the stationary point is a non-isolated local minimum, or not even a local minimum, then theorem 3 does not say anything about the rate of convergence. This does not seem to be a very important restriction of generality in practice, because it seems difficult to get our algorithm to converge to a non-isolated local optimum. We illustrate this with our small example.

The equilateral triangle, with centroid, is a non-isolated local minimum for $p = 2$. Start the iterations from a small perturbation of this stationary point. With a very close start ($\sigma = .0669873151$) we have convergence to the stationary value in 10-decimal precision within 5 iterations. The ratio $\varepsilon_{k+1}/\varepsilon_k$ continues to increase, however, although extremely slowly. It is .99 at iteration 8, and .999 at iteration 10. We have stopped the process at iteration 30, at which point we are still equally close to the stationary value, and the ratio is still increasing,

albeit exceptionally slowly. We have restarted the iterations somewhat further away from the stationary point (σ = .0669895385). After 10 iterations σ is down to .0669877606 and ε^2 is 7 x 10^{-9}. The ratio $\varepsilon^2_{k+1}/\varepsilon^2_k$ is .9739919946. At iteration 50 we have σ = .0669873813, ε^2 = 36 x 10^{-10}, and $\varepsilon^2_{k+1}/\varepsilon^2_k$ = .9926593514. Around iteration 60 the value of σ drops below .066987298 (equilateral triangle with centroid) and ε^2 begins to rise, causing a ratio $\varepsilon_{k+1}/\varepsilon_k$ larger than one. This continues for a very long time. At iteration 200, for instance, we have σ = .0669757275 and ε^2 = 3413 x 10^{-10}. The ratio of successive epsilons is still larger than one. This continues until iteration 250. In the meantime ε^2 has increased to .0026904972, and σ , which started dropping rapidly at iteration 225, is down to .0368738809. Convergence now becomes rapid, and within 20 iterations the configuration converges to the four corners of the square, which is an isolated local minimum (in fact the global minimum) for p = 2. At iteration 270 we have σ = .0285954792, $\varepsilon^2 < 10^{-10}$, and $\varepsilon^2_{k+1}/\varepsilon^2_k$ = .3431684733, which is for all practical purposes equal to κ^2. Thus we have started close to a non-isolated local minimum. The algorithm has a great deal of difficulty to get away from it, but ultimately succeeds. With a restart even further away (σ = .0675512622) the algorithm has difficulty escaping only until iteration 30. Then ε^2 decreases rapidly again, and we have convergence to the square in 55 iterations. It is now not difficult to conjecture that in the first start, in which we seemed to converge on the equilateral triangle, we merely did not continue long enough. After hundreds or perhaps thousands of iterations we would converge on the square again, if we continued.

8: Nonmetric scaling

As indicated in De Leeuw (1977) we can think of nonmetric scaling algorithms in two different ways. In the first place as alternating least squares methods, which alternate one gradient step (or Guttman transform) with a monotone regression step. Of course it is possible, and perhaps sometimes advisable, to perform more 'metric scaling steps' between monotone regressions. But of course these metric scaling steps have a rate of convergence which is described by our results above. In this interpretation the basic multidimensional scaling loss function is interpreted as a function of two sets of variables, the configuration X and the disparities $\hat{D}$.

It is also possible to view the loss function as a function of X alone. This is the original definition of stress as proposed by Kruskal (1964a, 1964b). If $\sigma(X,\hat{D})$ is the 'stress' used in the first approach, then Kruskal's stress is the minimum of $\sigma(X,\hat{D})$ over all feasible disparities $\hat{D}$. Thus $\hat{D}$ is 'projected out', and the remaining function depends on X only. This elementary fact has caused a great deal of confusion in the early days of multidimensional scaling. The confusion was made even bigger by the fact that the derivative of $\sigma(X)$ is the same as the partial derivative of $\sigma(X,\hat{D})$ with respect to X, evaluated at the optimal $\hat{D}(X)$. Thus $\sigma(X) = \sigma(X,D(X))$, but also $\nabla\sigma(X) = \nabla_X\sigma(X,D(X))$. This last result is due to Kruskal (1971), who used it to show that $\sigma(X)$ is differentiable (whenever $\sigma(X,\hat{D})$ is differentiable). It is also used by De Leeuw (1977) to show that $X_{k+1} = V^+\hat{B}(X_k)X_k$, with $\hat{B}(X)$ defined as in (10) but with $\hat{D}(X)$ substituted for Δ, is still a convergent subgradient algorithm. Thus our qualitative convergence results remain true without modification, both in the

alternating least squares and in the (sub)gradient interpretation.

Unfortunately the transformation $X \rightarrow \hat{D}(X)$ is generally not differentiable. In fact to find $\hat{D}(X)$ we have to project $D(X)$ orthogonally on a polyhedral convex cone, which implies that the transformation in piecewise linear in the distances. The linear pieces are joined in a continuous, but not in a smooth way. If we have convergence of the nonmetric scaling algorithm to a point where the cone-projection is locally the same linear map, then our convergence results apply. In general, however, we cannot exclude the possibility that the convergence is to a point on the boundary of two regions with different projection maps. These linear maps corresponds, by the way, to the partitioning into blocks found by the monotone regression algorithm. In this case our results do not apply, and they must be adapted.

9: Summary and conclusions

We have shown in this paper that the basic majorization algorithm for multidimensional scaling converges to a stationary point, if convergence is defined in terms of the asymptotic regularity of the generated sequence, i.e. in terms of the fact that the distance between two consecutive members of the sequence converges to zero. We have also shown that if one of the accumulation point of the sequence is an isolated local minimum, then convergence is linear. This seems to be all that is needed for practical applications. It follows from our small numerical example, that it is possible that actual computer programs stop at saddle points which are not local minima, and that in the neighborhood of such saddle points convergence may look sublinear. Our experience with many practical examples indicates

that the level of isolated local minima in multidimensional scaling is very often close to unity. Thus although convergence is theoretically linear, it can be extremely slow. Thus it becomes very important, at least in some cases, to look for ways to speed up linear convergence, or even for ways to attain superlinear convergence. These acceleration devices will be investigated in subsequent publications. Simple ways to speed up linear convergence were already investigated by De Leeuw and Heiser (1980), more complicated ones were studied by Stoop and De Leeuw (1983). Of course the formulas derived in section 12 can be used quite easily to derive the exact form of Newton's method for multidimensional scaling, but our numerical experience so far suggests that Newton's method must also be used with a great deal of care in this context. Our main conclusion is that the majorization method is reliable and very simple, but that it is generally slow, and sometimes intolerably slow. It seems to us that an additional conclusion is that one should always study the second derivatives of the loss function at the stopping point of the algorithm. This indicates if we have stopped at a local minimum or not, it also indicates how much improvement we can expect in various directions. This improvement information can be used either to try to make one or more Newton-steps, or to derive information on the stability of the solution.

REFERENCES

Arabie, P., and Carroll, J.D. (1980), "MAPCLUS: A Mathematical Programming Approach to Fitting the ADCLUS Model," Psychometrika, 45, 211-235.

Borg, I., (1981), Anwendungsorientierte Multidimensionale Skalierung. Berlin; Springer.

Carroll, J.D., (1972) "Individual differences and multidimensional scaling," in Multidimensional Scaling: Theory and Applications in the Behavioural Sciences, Vol I, Theory, eds. R.N. Shepard, A.K. Romney and S. Nerlove, New York: Seminar Press.

Carroll,J.D., and Pruzansky, S. (1980), "Discrete and Hybrid Scaling Models," in Similarity and Choice, eds. E.D. Lantermann and H. Feger, Bern: Hans Huber.

Clarke, F.H. (1975), "Generalized Gradients and Applications," Transactions of the American Mathematical Society, 205, 247-262.

Clarke, F.H. (1981), "Generalized Gradients of Lipschitz Functionals," Advances in Mathematics, 40, 52-67.

Defays, D. (1978), "A Short Note on a Method of Seriation," British Journal of Mathematical and Statistical Psychology, 31, 49-53.

De Leeuw, J. (1977), "Applications of Convex Analysis to Multidimensional Scaling," in Recent Developments in Statistics, eds. J.R. Barra, F. Brodeau, G. Romier and B. Van Cutsem, Amsterdam: North Holland Publishing Company.

De Leeuw, J. (1984), "Differentiability of Kruskal's Stress at a Local Minimum," Psychometrika, 49, 111-113.

De Leeuw, J. and Heiser, W.J. (1977) "Convergence of Correction Matrix Algorithms for Multidimensional Scaling," in Geometric Representations of Relational data, ed. J.C. Lingoes, Ann Arbor: Mathesis Press.

De Leeuw, J. and Heiser, W.J. (1980), "Multidimensional Scaling with Restrictions on the Configuration," in Multivariate Analysis, Vol V, ed. P.R. Krishnaiah, Amsterdam: North Holland Publishing Company.

Guttman, L. (1968), "A General Nonmetric Technique for finding the Smallest Coordinate Space for a Configuration of Points," Psychometrika, 33, 469-506.

Hartmann, W. (1979), Geometrische Modelle zur Analyse empirischer Daten, Berlin: Akademie Verlag.

Heiser, W.J. (1981), Unfolding Analysis of Proximity Data, Unpublished Doctoral Dissertation, University of Leiden.

Kruskal, J.B. (1964a), "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," Psychometrika, 29, 1-28.

Kruskal, J.B. (1964b), "Nonmetric Multidimensional Scaling: a Numerical Method," Psychometrika, 29, 115-129.

Kruskal, J.B. (1971), "Monotone Regression: Continuity and Differentiability Properties," Psychometrika, 36. 57-62.

Kruskal, J.B. and Wish, M. (1978), Multidimensional Scaling, Beverley Hills: Sage Publications.

Lingoes, J.C. and Roskam, E.E. "A Mathematical and Empirical Comparison of two Multidimensional Scaling Algorithms," Psychometrika, 38, monograph supplement.

Ortega, J.M. and Rheinboldt, W.C. (1970), Iterative Solution of Nonlinear Equations in Several Variables, New York: Academic Press.

Ostrowski, A.M. (1966), Solution of Equations and Systems of Equations, New York: Academic Press.

Rockafellar, R.T. (1970), Convex Analysis, Princeton: Princeton University Press.

Shepard, R.N. and Carroll, J.D. (1966), "Parametric Representation

of Nonlinear Data Structures," in Multivariate Analysis, Vol I, ed. P.R. Krishnaiah, New York: Academic Press.

Stoop, I. and De Leeuw, J. The Stepsize in Multidimensional Scaling Algorithms, Paper presented at the Third European Meeting of the Psychometric Society, Jouy-en-Josas, France, july5-8, 1983.

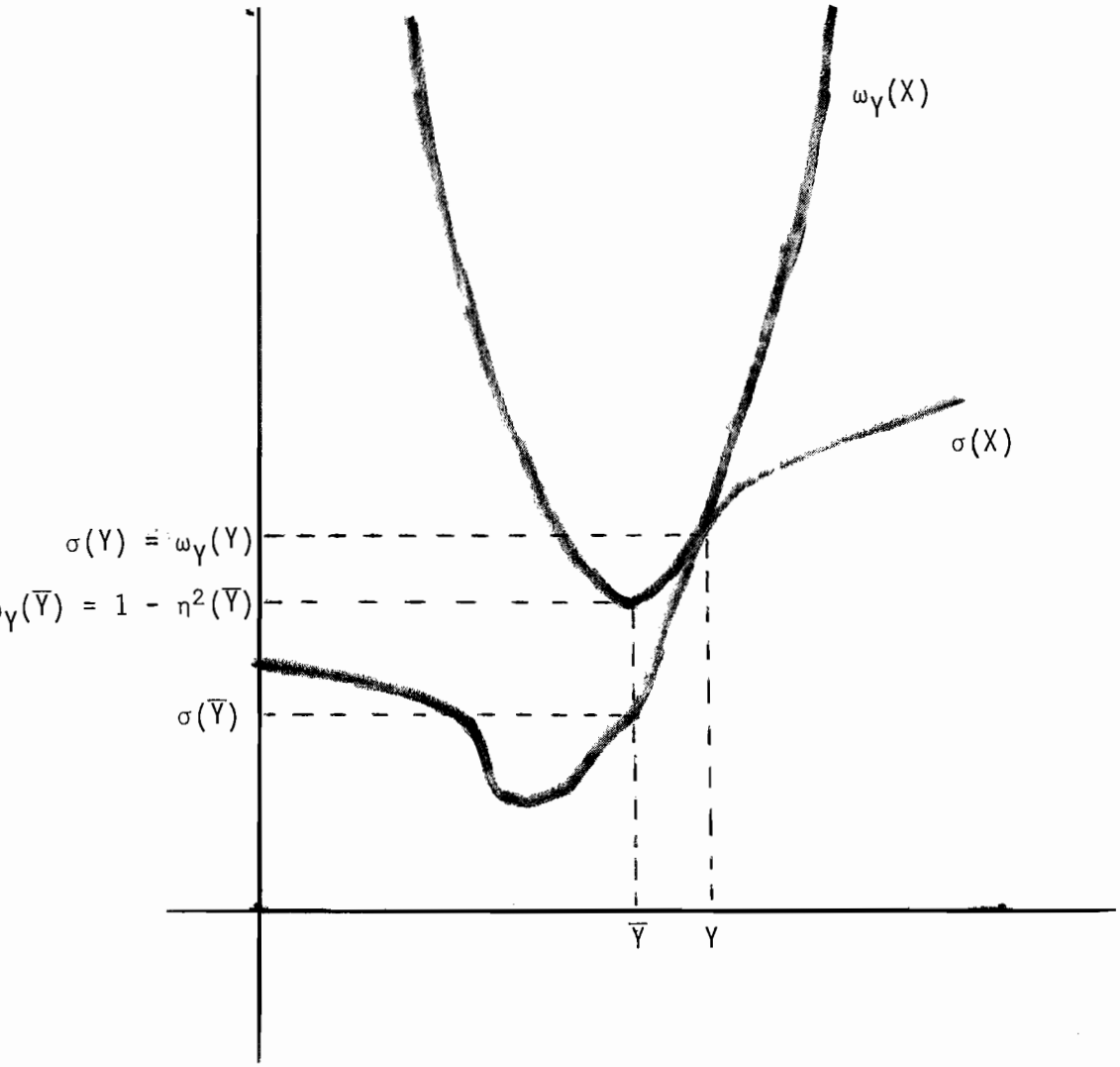

figure 1: majorization of loss by quadratics