FITTING DISTANCES BY LEAST SQUARES

JAN DE LEEUW

CONTENTS

1. Introduction 1
2. The Loss Function 2
2.1. Reparametrization 4
2.2. Use of Homogeneity 5
2.3. Pictures of Stress 6
3. Local and Global Minima 7
3.1. The Gradient and Hessian of Stress 7
3.2. Majorization 9
3.3. Minima and Maxima of Stress 11
3.4. Special Cases 12
3.5. Destructive Generality 14
3.6. On the Planes 16
3.7. Inverse Scaling 16
4. A Majorization Algorithm for MDS 17
4.1. Majorization for STRESS 18
4.2. Speed of Convergence 19
4.3. Fixed-step Acceleration 20
4.4. Approximating Optimal Step-size 20
4.5. Steffenson Acceleration 21
4.6. Newton's Method 21
4.7. Negative Dissimilarities 21
4.8. More Type II Majorization 22
4.9. Global Minimization 23
References 23

1. INTRODUCTION

In this paper we review the problem of fitting *Euclidean distances* to data, using a least squares loss function. This problem must be distinguished from the problem of least squares fitting *squared Euclidean distances* to data, or

the problem of least squares fitting *scalar products* to data. A comparison of these three different approaches to multidimensional scaling is given by De Leeuw and Heiser de Leeuw and Heiser [1980b] and Meulman Meulman [1986]. Recent developments in squared distance scaling are reviewed in a nice paper by Glunt e.a. Glunt et al. [1990].

We shall prove some existence-type results and some convergence type results. They are both derived by using a constructive proof method, the *majorization method*. The method was introduced in 1977 in MDS by De Leeuw de Leeuw [1977]. A simplified treatment of the Euclidean case was published in 1978, with some extensions to individual differences scaling de Leeuw and Heiser [1977]. Individual differences scaling was subsequently recognized to be a special scaling problem with restrictions on the configuration, and a general approach to such restricted MDS problems was presented in de Leeuw and Heiser [1980a]. Speed of convergence of the algorithm was studied in de Leeuw [1988]. Other reviews, with some additional extensions, are Heiser [1990] and Mathar and Groenen [1991]. Although de Leeuw and Heiser [1980a] did use the majorization method to prove necessary conditions for an extremum, this has not really been followed up in the scaling and multivariate analysis literature.

In this paper we start with the basic results for the metric case, and then introduce various generalizations. Most of what we discuss is a review of results that have been published elsewhere, but some of it is new. I think the paper illustrates the remarkable power of the majorization method.

2. THE LOSS FUNCTION

Suppose X are the coordinates of n points in d dimensions. The $n \times d$ matrix X is cal led a *configuration.* We write $\mathbf{R}^{n\times d}$ for the space of configurations, and $\mathcal{R}^{n\times d}$ for the centered configurations (in which the columns of X sum to zero). The e_i are unit vectors of length n. Let $A_{ij} = (e_i - e_j)(e_i - e_j)'$. Then

$$d_{ij}^2(X) = \text{tr } X'A_{ij}X.$$

Suppose $\Delta = \{\delta_{ij}\}$ is a matrix of *dissimilarities* and $W = \{w_{ij}\}$ is a matrix of *weights.* The first basic problem we discuss in this paper is the following.

Problem P1: Find $X \in \mathcal{R}^{n\times d}$ in such a way that the loss function

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

is minimized.

Following Kruskal Kruskal [1964] we call $\sigma(X)$ the STRESS of a configuration.

We can suppose, without loss of generality, that dissimilarities and weights are symmetric and hollow (have zero diagonal). De Leeuw de Leeuw [1977] show how to partition STRESS in such a way that the asymmetric and diagonal parts end up in additive components that do not depend on the configuration. We can also suppose without loss of generality that half the weighted sum of squares of the dissimilarities is equal to one. We finally assume (at least for the time being) that the weights and dissimilarities are non-negative.

In order to make the problem more managable, we introduce some extra notation. First,

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

Then

$$\sigma(X) = 1 - \sum_{i=1}^n \sum_{j=1}^n w_{ij}\delta_{ij} d_{ij}(X) + \text{tr } X'VX.$$

Observe that $\text{rank}(V) \leq n-1$, with equality if and only if V is *irreducible*, i.e. if and only if V is not the direct sum of two or more matrices de Leeuw [1977].

The expression for STRESS can be simplified even more by defining *residuals*

$$r_{ij}(X) = \begin{cases} \frac{\delta_{ij}}{d_{ij}(X)} & \text{if } d_{ij}(X) > 0, \\ \text{arbitrary} & \text{if } d_{ij}(X) = 0. \end{cases}$$

Using $r_{ij}(X)$ we can now set

$$B(X) = \sum_{i=1}^n \sum_{j=1}^n w_{ij} r_{ij}(X) A_{ij},$$

and

$$\sigma(X) = 1 - \text{tr } X'B(X)X + \text{tr } X'VX.$$

Finally we define

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

which implies

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

This is the way the loss function is written in de Leeuw and Heiser [1980a]. Most of the results in the papers by De Leeuw, Heiser, Meulman, and Mathar are derived using this formalism. It is interesting in this formulation that both $\rho(\bullet)$ and $\eta(\bullet)$ are *norms*, i.e. homogeneous and nonnegative convex functions. The function $\eta^2(\bullet)$ is also convex, and consequently $\sigma(\bullet)$ is

the difference of two convex functions. Functions which are the difference of two convex functions are often called (somewhat unimaginatively) d.c. functions. They have been studied in considerable detail Hiriart-Urruty [1985], Shapiro and Yohim [1982].

Some additional manipulation brings the loss function into yet another interesting form. Define

$$\overline{X} = V^+ B(X) X,$$

as the *Guttman transform* of X. This is named after Guttman, who studied some properties of this transform in his important paper Guttman [1968]. Observe that although $B(X)$ is not uniquely defined when some of the $d_{ij}(X)$ are zero, the product $B(X)X$, and thus the Guttman transform, is unique. Using the Guttman transform $\rho(X) = \text{tr } X'V\overline{X}$, and thus

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

This is a representation we shall later use in the context of majorization theorems and algorithms.

2.1. **Reparametrization.** For some purposes it is useful to reparametrize the MDS problem. Suppose we have a number of basis configurations Z_p, and it is known (or required) that X is in the space spanned by the Z_p. Obviously this still covers the problem in which X is unrestricted, in which case we need $d(n-1)$ centered matrices for the basis. But this formalism can also be used to impose linear restrictions on the configuration. Examples are in de Leeuw and Heiser [1980a].

Let us write X in the form $X = \sum_{p=1}^t \theta_p Z_p$. Suppose, without loss of generality, that the Z_p are chosen in such a way that $\text{tr } Z_p' V Z_q = \delta^{pq}$. This can be enforced quite simply by applying the Gram-Schmidt process to the Z_p. Then $d_{ij}^2(\theta) = \text{tr } X'A_{ij}X = \theta' C_{ij}\theta$, where C_{ij} has elements $(C_{ij})_{pq} = \text{tr } Z_p' A_{ij} Z_q$ so that $\sum_{i=1}^n \sum_{j=1}^n w_{ij} C_{ij} = I$. If we renumber the index pairs (i, j) using $k = 1, \cdots, K$ we can write

$$B(\theta) = \sum_{k=1}^K w_k r_k(\theta) C_k,$$

where

$$r_k(\theta) = \begin{cases} \frac{\delta_k}{d_k(\theta)} & \text{if } d_k(\theta) > 0, \\ \text{arbitrary} & \text{if } d_k(\theta) = 0. \end{cases}$$

Then

$$\rho(\theta) = \sum_{k=1}^{K} w_k \delta_k \sqrt{\theta' C_k \theta} = \theta' B(\theta)\theta,$$

$$\eta^2(\theta) = \theta'(\sum_{k=1}^{K} w_k C_k)\theta = \theta'\theta.$$

Of course the Guttman transform in this context is just $\overline{\theta} = B(\theta)\theta$.

From the mathematical point of view the basic advantage of this new parametrization is its symmetry. We don't have to remember that some elements of X belong to the same row, while others don't. Or, to put it differently, we do not have to *vec* our matrices. Another advantage is also immediate. By adopting the parametrization we have eliminated the problem of rotational indeterminacy of MDS solutions. For an unrestricted MDS solution we do not use $d(n-1)$ but actually $d(n-1) - d(d-1)$ matrices Z_p. We can, for instance, set all upper-diagonal elements of the Z_p equal to zero. From the notational point of view we gain some additional simplicity by getting rid of the double index (i, j) and by not having to use the trace operator to define inner products of matrices.

Again, for completeness, we formulate a second basic problem we shall study in this paper.

Problem P2: Find $\theta \in \mathcal{R}^t$ such that

$$\sigma(\theta) = \sum_{k=1}^{K} w_k (\delta_k - \sqrt{\theta' C_k \theta})^2,$$

is minimized.

The problem makes sense for any set of positive semi-definite matrices C_k. That problem P2 was derived from the Euclidean MDS problem is irrelevant for most of our results. Unless explicitly specified otherwise, we still assume that for all k we have $w_k > 0$ and $\delta_k \geq 0$, and we assume that $\sum_{k=1}^{K} w_k C_k = I$.

2.2. **Use of Homogeneity.** Actually, there is yet another reduction of the problem that is useful. We do not only eliminate the rotational indeterminacy, we can also get rid of a scale factor that tends to be in the way. Clearly, for $\alpha \geq 0$,

$$\sigma(\alpha\theta) = \sum_{k=1}^{K} w_k (\delta_k - \alpha\sqrt{\theta' C_k \theta})^2 = 1 - \alpha\rho(\theta) + \alpha^2\eta^2(\theta).$$

Thus

$$\min_{\alpha\geq 0}\sigma(\alpha\theta)=\begin{cases}1-\lambda^2(\theta) & \text{if } \rho(\theta)\geq 0,\\ 1 & \text{if } \rho(\theta)\leq 0,\end{cases}$$

where

$$\lambda^2(\theta)=\frac{\rho^2(\theta)}{\eta^2(\theta)},$$

and where the minimum is attained for

$$\hat\alpha=\begin{cases}\frac{\rho(\theta)}{\eta^2(\theta)} & \text{if } \rho(\theta)\geq 0,\\ 0 & \text{if } \rho(\theta)\leq 0.\end{cases}$$

Observe that under our assumptions, we always have $\rho(\theta)\geq 0$. Thus minimizing $\sigma(\bullet)$ amounts to the same thing as maximizing $\lambda(\bullet)$. Call this **Problem P3**. P3 in turn amounts to the same thing as maximizing $\rho(\theta)$ over all θ such that $\eta(\theta)=1$, which is also the same as maximizing $\rho(\theta)$ over all θ such that $\eta(\theta)\leq 1$. Call this **Problem P4**. The MDS problem was first formulated in the P3 or P4 form by de Leeuw [1977]. It leads to the problem of maximizing a norm over a convex set, or to the problem of maximizing the ratio of two norms. This makes it possible to use results of Robert [1967].

Thus another related problem we shall analyze in this paper is to maximize $\rho(\theta)$ over all θ on the unit sphere (or, equivalently, over all θ in the unit ball). As Robert and others have observed, this is a straightforward nonlinear generalization of an eigenvector-eigenvalue or Rayleigh-quotient problem. The only difference is that in the ordinary eigenvalue problem we maximize a quadratic norm over the unit sphere, while in the more complicated MDS problem a non-quadratic convex norm is maximized.

2.3. **Pictures of Stress.** Very few people have actually looked at STRESS in any detail. Usually the problem is just too high-dimensional to consider plotting, and the parametrization in terms of configurations also does not help. The new, symmetric parametrization is a bit more convenient in this respect.

We consider some simple examples. We have four points in two dimensions, with all weights equal to one, and all off-diagonal dissimilarities equal to $\frac{1}{6}\sqrt{6}$. This example has been analyzed in considerable detail before [de Leeuw, 1988, Groenen and Heiser, 1991].

In the first example we take Z_1 as an equilateral triangle, with centroid, and Z_2 as a square. We apply Gram-Schmidt to orthonormalize these matrices to $\tilde Z_1$ and $\tilde Z_2$, and we study $\sigma(\theta_1\tilde Z_1+\theta_2\tilde Z_2)$ This results in the following

matrices C_k.

$$\begin{bmatrix} \begin{pmatrix} +.2500 & -.0211 \\ -.0211 & +.0017 \end{pmatrix} & & \\ \begin{pmatrix} +.2500 & +.0788 \\ +.0788 & +.0647 \end{pmatrix} & \begin{pmatrix} +.2500 & -.0577 \\ -.0577 & +.0878 \end{pmatrix} & \\ \begin{pmatrix} +.0833 & -.0192 \\ -.0192 & +.2278 \end{pmatrix} & \begin{pmatrix} +.0833 & +.1172 \\ +.1172 & +.2048 \end{pmatrix} & \begin{pmatrix} +.0833 & -.0980 \\ -.0980 & +.4131 \end{pmatrix} \end{bmatrix}$$

Only the first matrix, the one corresponding with d_{12}, is singular, the others are non-singular.

If we look at the function from far away (Figure 1), it looks very much like a convex quadratic bowl.

INSERT FIGURE 1 ABOUT HERE

But for our purposes Figure 1 is more or less irrelevant, because we are really only interested in θ which are in the unit circle. If we look more closely, concentrating on the unit circle, we see a mountain on top of the origin, which is the only local maximum of STRESS. We also see, in Figure 2, valleys around the unit circle.

INSERT FIGURE 2 ABOUT HERE

But Figure 2 is still a bit misleading. The scale factor is still in the plot. Figure 3 looks at $\lambda(\zeta)$, for $\pi \leq \zeta \leq \pi$, and $\theta_1 = \sin(\zeta)$ and $\theta_2 = \cos(\zeta)$.

INSERT FIGURE 3 ABOUT HERE

It is clear in Figure 3 that there is a point on the circle where $\lambda(\bullet)$ is not differentiable. At that point the function has a local maximum. This corresponds with the small ridge in Figure 2, and it corresponds with the ray for which $d_{12}(X) = \theta' C_1 \theta = 0$. The ray is $\theta = \alpha(-.0841 - .9965)$, and obviously on that ray (as on any ray) STRESS is a convex quadratic in α.

3. Local and Global Minima

In the previous section we looked at STRESS in a global way. Now we shall pay more attention to the details, in particular to the locations where there are stationary values. Our tools are partly classical (first and second derivatives), but we also start using majorization to prove existence and characterization theorems.

3.1. **The Gradient and Hessian of Stress.**

One of the major advantages of the alternative parametrization in terms of θ is that formulas for the derivatives of the loss function, and results based on them, become considerably simpler.

First the gradient, at a point where the loss function is differentiable. Observe that we have differentiability if, and only if, $d_k(\theta) > 0$ for all k for which $w_k\delta_k > 0$. Simple computation gives

$$\mathcal{D}\sigma(\theta) = \theta - B(\theta)\theta = \theta - \overline{\theta}.$$

Necessary Condition Theorem: If the loss function is differentiable at θ, and has a stationary value there, then θ is equal to its Guttman transform $\overline{\theta}$.

Proof: Above. **Q.E.D.**

The second derivatives require a bit more computation. Define

$$H(\theta) = \sum_{k=1}^K w_k r_k(\theta)\Big(C_k - \frac{C_k\theta\theta'C_k}{\theta'C_k\theta}\Big).$$

Then

$$\begin{aligned}\mathcal{D}^2\rho(\theta) &= H(\theta),\\ \mathcal{D}^2\sigma(\theta) &= I - H(\theta).\end{aligned}$$

Second Derivative Theorem: Suppose STRESS is differentiable at θ. Then

(1) $H(\theta) \geq 0$.
(2) $H(\theta)\theta = 0$ for all θ.
(3) If θ is a local minimizer of STRESS, then $H(\theta) \leq I$. If the local minimum is strict, then $H(\theta) < I$.
(4) Let
$$r_+(\theta) = \max_{k=1}^K r_k(\theta).$$
Then $H(\theta) \leq r_+(\theta)I$.

Proof: $H(\theta)$ is a weighted sum of positive semi-definite matrices, and thus it is positive semi-definite. (2) is obvious. If θ is a local minimum, then $\mathcal{D}^2\sigma(\theta) \geq 0$, which means $H(\theta) \leq I$. The same argument applies for a strict (isolated) local minimum. We deduce (4) from $r_k(\theta) \leq r_+(\theta)$, and

$$C_k - \frac{C_k\theta\theta'C_k}{\theta'C_k\theta} \leq C_k.$$

Q.E.D.

Far Away Theorem: If $d_k(\theta) \geq \delta_k$ for all k, then $\mathcal{D}^2\sigma(\theta) \geq 0$.

Proof: The assumption in the theorem implies $r_+(\theta) \leq 1$, which implies $\mathcal{D}^2\sigma(\theta) \geq (1 - r_+(\theta))I \geq 0$. **Q.E.D.**

It is more difficult to formulate a similar, but opposite, theorem for small configurations. No matter how close θ is to the origin, we always have $\mathcal{D}^2\sigma(\theta)\theta = \theta$, i.e. the Hessian always has one eigenvalue equal to +1.

We can get around this problem by defining $\kappa_-(\theta)$ as the smallest nonzero eigenvalue of $H(\theta)$.

Near Zero Theorem: If $\lambda \leq \kappa_-(\theta)$ then $\xi'\mathcal{D}^2\sigma(\lambda\theta)\xi \leq 0$ for all ξ orthogonal to θ.

Proof: $H(\lambda\theta) = \frac{1}{\lambda}H(\theta)$. Thus if κ is a nonzero eigenvalue of $H(\theta)$, the corresponding nonzero eigenvalue of $\mathcal{D}^2\sigma(\theta)$ is $1 - \frac{\kappa}{\lambda}$. **Q.E.D.**

3.2. **Majorization.** Consider the problem of minimizing a function $\psi(\bullet)$ over a set Θ, which is a subset of $\mathbb{R}^p$. Suppose $\phi(\bullet, \bullet)$ is another function, defined on $\Theta \times \Xi$, where $\Theta \subseteq \Xi$, with the property that

$$\psi(\theta) \leq \phi(\theta, \xi) \text{ for all } \theta \in \Theta \text{ and } \xi \in \Xi,$$
$$\psi(\theta) = \phi(\theta, \theta) \text{ for all } \theta \in \Theta.$$

In that case we call $\phi(\bullet, \bullet)$ a *majorization function.*

Majorization functions are useful, because of the following simple result.

Necessity by Majorization Theorem: If $\hat{\theta}$ minimizes $\psi(\bullet)$ over Ω, then $\hat{\theta}$ also minimizes $\phi(\bullet, \hat{\theta})$ over Ω.

Proof: Suppose $\tilde{\theta} \neq \hat{\theta} \in \Omega$, and $\phi(\tilde{\theta}, \hat{\theta}) < \phi(\hat{\theta}, \hat{\theta})$. Then $\psi(\tilde{\theta}) \leq \phi(\tilde{\theta}, \hat{\theta}) < \phi(\hat{\theta}, \hat{\theta}) = \psi(\hat{\theta})$, which contradicts optimality of $\hat{\theta}$. **Q.E.D.**

The chain $\psi(\tilde{\theta}) \leq \phi(\tilde{\theta}, \hat{\theta}) < \phi(\hat{\theta}, \hat{\theta}) = \psi(\hat{\theta})$ which occurs in the proof of this theorem, is sometimes called the *Sandwich Inequality.* This is because we have two layers of $\phi(\bullet, \bullet)$ between two slices of $\psi(\bullet)$.

The concept of a majorization function is illustrated in Figures 4 and 5. In these figures we have plotted the function $\psi(\theta) = \theta^2 - \frac{1}{4}\theta^4$, and the majorization function $\phi(\theta, \xi) = \theta^2 + \frac{3}{4}\xi^4 - \xi^3\theta$. Figure 4 gives the majorization at $\xi = 1$, and Figure 5 gives it at $\xi = 2$.

INSERT FIGURES 4 AND 5 ABOUT HERE

We give two more general examples of majorization, which are particulary relevant for scaling applications. If $\psi(\bullet)$ is d.c., i.e. $\psi(\bullet) = \alpha(\bullet) - \delta(\bullet)$, with both $\alpha(\bullet)$ and $\beta(\bullet)$ convex, then we can use

$$\phi(\theta, \xi) = \alpha(\theta) - \beta(\xi) - (\theta - \xi)'\mathcal{D}\beta(\xi).$$

In this case the majorization function is convex in θ, and in the case in which $\alpha(\bullet)$ is quadratic, the majorization function is quadratic as well. Let us call this *Type I Majorization,* for ease of reference.

In the second example, suppose $\psi(\bullet)$ is twice-differentiable, and suppose there exists a $G > 0$ such that $\mathcal{D}^2\psi(\theta) \leq G$ for all θ. Then

$$\phi(\theta, \xi) = \psi(\xi) + (\theta - \xi)\mathcal{D}\psi(\xi) + (\theta - \xi)'G(\theta - \xi)$$

is a suitable majorization function, which is quadratic in θ. This is *Type II Majorization.*

The Necessity by Majorization Theorem tells us that if a function has a local minimum at a point, then the majorization function at that point also has a local minimum at the point. We can go a bit further by using (directional) derivatives Dem'yanov and Malozemov [1990].

We obviously have

$$\psi(\theta) = \min_{\xi \in \Xi} \phi(\theta, \xi).$$

Now suppose $\phi(\bullet, \bullet)$ is jointly continuous in its arguments, and suppose the partial derivative $\mathcal{D}_1\phi(\bullet, \bullet)$ exists everywhere. Also suppose Ξ is compact, or at least that the minimum can always be chosen in a compact subset of Ξ. Let

$$\mathcal{S}(\theta) = \{\xi \in \Xi \mid \phi(\theta, \xi) = \psi(\theta)\}.$$

Then

$$\nabla\psi(\theta, \zeta) = \min_{\xi \in \mathcal{S}(\theta)} \zeta' \mathcal{D}_1(\theta, \xi).$$

Here $\nabla\psi(\theta, \zeta)$ is the directional derivative, i.e. ζ is a unit-length direction vector, and

$$\nabla\psi(\theta, \zeta) = \lim_{\epsilon \downarrow 0} \frac{\psi(\theta + \epsilon\zeta) - \psi(\theta)}{\epsilon}.$$

If $\mathcal{S}(\theta)$ is a singleton, i.e. if $\psi(\theta) < \phi(\theta, \xi)$ for all $\xi \in \Xi$ with $\xi \neq \theta$, then

$$\mathcal{D}\psi(\theta) = \mathcal{D}_1\phi(\theta, \theta).$$

Thus, not only do the function values coincide at θ, but so do the derivatives.

We end the section by proposing a suitable majorization function for STRESS.

Stress Majorization Theorem: The function

$$\tau(\theta, \xi) = 1 - \eta^2(\overline{\xi}) + \eta^2(\theta - \overline{\xi})$$

majorizes $\sigma(\theta)$.

Proof: By Cauchy-Schwartz, $d_k(\theta)d_k(\xi) \geq \theta' C_k \xi$, which we can also write as

$$\delta_k d_k(\theta) \geq r_k(\xi)\theta' C_k \xi.$$

Thus

$$\sigma(\theta) \leq 1 - \theta' B(\xi)\xi + \eta^2(\theta) = 1 - \theta'\overline{\xi} + \eta^2(\theta),$$

and

$$\sigma(\theta) \leq 1 - \eta^2(\overline{\xi}) + \eta^2(\theta - \overline{\xi}).$$

If $\theta = \xi$ we have equality in Cauchy-Schwartz, and consequently everywhere else. **Q.E.D.**

This theorem was first proved in this form in de Leeuw and Heiser [1980a]. We improve it a little bit here.

Strict Stress Majorization Theorem: We have $\sigma(\theta) = \tau(\theta, \xi)$ if and only if $\xi = \theta + \mu$, where μ is such that $C_k\mu = 0$ for all k with $d_k(\theta) > 0$.

Proof: This is simple. If $d_k(\theta) > 0$ we have equality in Cauchy-Schwartz for index k if and only if $C_k(\theta - \xi) = 0$. If $d_k(\theta) = 0$ we have equality for arbitrary ξ. **Q.E.D.**

Of course this implies that $\mu = 0$ if all $d_k(\theta)$ are positive, and we have equality in that case only if $\theta = \xi$. It also implies that $d_k(\xi) = d_k(\theta)$ for all k for which $d_k(\theta) > 0$.

3.3. **Minima and Maxima of Stress.** Although STRESS is not differentiable at a point where some of the $\theta' C_k \theta$ are zero, it is always differentiable in all directions.

Directional Derivatives Theorem:

$$\nabla\sigma(\theta;\xi) = \xi'(\theta - \overline{\theta}) - \sum\sum\{w_k\delta_k d_k(\xi) \mid d_k(\theta) = 0\}.$$

Proof: Simply apply the formula

$$\nabla d_k(\theta;\xi) = \begin{cases} d_k^{-1}(\theta)\theta' C_k\xi & \text{if } d_k(\theta) > 0, \\ d_k(\xi) & \text{if } d_k(\theta) = 0. \end{cases}$$

Q.E.D.

The formula can also be derived from the formula for the directional derivative in terms of the partials of the majorizing functions. In particular, we have

$$\mathcal{D}_1\tau(\theta,\xi)) = \theta - \overline{\xi},$$

and

$$\overline{\theta + \mu} = \overline{\theta} + \sum\{w_k r_k(\mu) C_k\mu \mid d_k(\theta) = 0\},$$

if μ satisfies the conditions of the Strict Stress Majorization Theorem. The directional derivative can be used to provide a proof of the following result, which was first given by De Leeuw de Leeuw [1984].

Zero Distances Theorem: If STRESS has a local minimum at $\hat{\theta}$, then $\hat{\theta}$ is stationary, and $d_k(\hat{\theta}) > 0$ for all k such that $w_k\delta_k > 0$. Thus STRESS is differentiable at a local minimum.

Proof: We must have $\nabla\sigma(\hat{\theta},\zeta) \geq 0$ for all ζ. This means we must also have

$$\nabla\sigma(\hat{\theta},\zeta) + \nabla\sigma(\hat{\theta},-\zeta) = -\sum\sum\{w_k\delta_k d_k(\xi) \mid d_k(\hat{\theta}) = 0\} \geq 0,$$

for all ζ. But this implies we must have $w_k\delta_k = 0$ for all k such that $d_k(\hat{\theta}) = 0$. **Q.E.D.**

Again we can prove this using majorization. We need a somewhat sharper result than the usual majorization, i.e. we need a slightly larger majorization function.

$$\sigma(\theta) \leq \tau_0(\theta,\xi) = \tau(\theta,\xi) + \sum\{w_k\delta_k\epsilon_k \mid d_k(\theta) = 0\},$$

where the ϵ_k are arbitrary positive numbers. By the Necessity by Majorization Theorem $\hat{\theta}$ must minimize $\tau_0(\bullet, \hat{\theta})$, and this can be done by minimizing each of the two components separately.

At the Origin Theorem: STRESS has a local maximum at the origin. STRESS has no other local maxima.

Proof: We have

$$\nabla\sigma(0,\xi) = -\sum_{k=1}^{K} w_k\delta_k d_k(\xi),$$

which is clearly non-positive. Now suppose we have a local maxima at $\theta \neq 0$. Then $\sigma(\theta+\epsilon\theta)$, seen as a function of ϵ, must have a local maximum for $\epsilon = 0$. But $\sigma(\theta+\epsilon\theta)$, is a convex quadratic, which does not have any maxima. **Q.E.D.**

The result above tells us that local maxima cannot exist away from the origin, because STRESS is a convex quadratic on any ray. This does not mean, of course, that we cannot have local maxima on other one-dimensional cross-sections (for instance, remember the circle in our small example).

3.4. **Special Cases.** Let us now look at some special cases, in which we can say more. In one-dimensional scaling we have

$$d_{ij}(x) = |x_i - x_j| = s_{ij}(x)(x_i - x_j),$$

with $s_{ij}(x) = \text{sign}(x_i - x_j)$. This means

$$\overline{x} = V^+ u(x),$$

where

$$u_i(x) = 2\sum_{j=1}^{n} w_{ij}\delta_{ij}s_{ij}(x).$$

Thus $\overline{x}$ only depends on the order of the x_i. There are, obviously, $n!$ orders, each order defines a cone $\mathcal{K}_\nu$, and a corresponding target vector z_ν. Thus we can find the global minimum by solving the $n!$ monotone regression problems

$$\min_{x\in\mathcal{K}_\nu} \eta^2(x - z_\nu),$$

and keeping the best one. Not all these $n!$ solutions define local minima, however.

One-dimensional Theorem: $x \in \mathcal{K}_\nu$ defines a local minimum of STRESS if and only if $x = z_\nu$, i.e. if and only if $z_\nu \in \mathcal{K}_\nu$.

Proof: By the Zero Distance Theorem we cannot have zero distances at a local minimum. Thus the projection on $\mathcal{K}_\nu$ cannot create equal distances, and thus z_ν must already be in the cone. **Q.E.D.**

Thus we do not actually have to carry out the projections in the monotone regression. We just cycle through the permutations and compute the z_ν. If it is in the correct order (i.e. in the same order that was used in generating it), then we save it (it is a local minimum). If we are done we have all local minima, and we select the one with the smallest loss value.

We have seen that the one-dimensional situation is quite different from the general one. We shall now show that full-dimensional scaling (i.e. with $d = n - 1$) is also special. We shall use a more convenient special-purpose parametrization. Let $H = XX'$. Then

$$d_{ij} = \sqrt{\text{tr } A_{ij}H} = \sqrt{h_{ii} + h_{jj} - 2h_{ij}}.$$

The metric MDS problem is to minimize $\sigma(H)$ over all H which are positive semi-definite, and have rank$(H) \leq d$. In the full-dimensional problem we drop the rank constraint.

Full Dimensional Theorem: The solution to the full dimensional scaling problem is unique: any local minimum is global.

Proof: We have to minimize $\sigma(H)$ over the convex set $H \geq 0$, i.e. H positive semi-definite. Now

$$\sum_{i=1}^n \sum_{j=1}^n d_{ij}^2(H) = h_{ii} + h_{jj} - 2h_{ij},$$

which is linear in H. Moreover

$$\sum_{i=1}^n \sum_{j=1}^n w_{ij}\delta_{ij}d_{ij}(H) = \sum_{i=1}^n \sum_{j=1}^n w_{ij}\delta_{ij}\sqrt{h_{ii} + h_{jj} - 2h_{ij}},$$

which is concave in H, because it is the weighted sum of square roots of positive linear functions. Thus $\sigma(H)$ is convex, and all local minima are global. **Q.E.D.**

There is another situation in which we can fairly easily find the global minimum of STRESS.

Simultaneous Diagonalization: Suppose all C_k can be diagonalized by a single orthonormal K. Then there are at most 2^t local minima, all with the same function value, which only differ in their sign pattern.

Proof: By applying K to all C_k we transform the problem to minimization of

$$\sigma(\theta) = 1 - \sum_{k=1}^K w_k\delta_k\sqrt{\theta'\Omega_k\theta} + \theta'\theta.$$

Here $\Omega_k = K'C_kK$ is diagonal. Define $\xi_p = \theta_p^2$. Then

$$\sigma(\theta) = 1 - \sum_{k=1}^K w_k\delta_k\sqrt{\sum_{p=1}^t \omega_{kp}\xi_p} + \sum_{p=1}^t \xi_p.$$

This is convex in ξ, with only one minimum. This single minimum translates back to at most 2^t local minima in θ. **Q.E.D.**

Although simultaneously diagonability cannot be expected to hold in practice, we see that to the extent to which it hold approximately we have an enormous number of local minima, which only differ in pattern of zeroes and in pattern of signs (for a given pattern of zeroes). We already go in step in the direction of this situation if we scale the problem such that $\sum_{k=1}^K w_kC_k = I$. We could go further and apply the simultaneous diagonalization method of De Leeuw and Pruzansky de Leeuw and Pruzansky [1978] to approach this condition even more.

The approach in the simultaneous diagonalization theorem can be extended. Unfortunately the result we arrive at is *almost* perfect, and because of the "almost" not really useful. Suppose the eigen-decompositions $C_k = K_k\Omega_kK_k'$ are used to define $\xi_k = K_k'\theta$. Then

$$\sigma(\xi_1, \cdots, \xi_K) = 1 - \sum_{k=1}^K w_k\delta_k\sqrt{\xi_k'\Omega_k\xi_k} + \sum_{k=1}^K w_k\xi_k'\Omega_k\xi_k.$$

We have to impose the linear restrictions

$$K_1\xi_1 = \cdots = K_K\xi_K.$$

By reformulating the problem in this way we have created a situation in which the loss function is convex in the parameters ξ_k^2, while the constraints are linear in the parameters ξ_k. Unfortunately this does not really seem to get us anywhere.

3.5. **Destructive Generality.** The Zero Distances Theorem is of great importance. Because of it we can more or less ignore the possibility of coinciding points, and the complications that arise because of that. But there are situations in which some of the crucial assumptions on which the theorem depends are not true any more. We could have, for instance, weights and/or dissimilarities that are negative. This destroys most of the convexity and majorization results we have so far. Fortunately we can save some constructive theorems from the rubble by combining Type I and Type II majorization. We present these results, which extend some earlier ones by Heiser Heiser [1990], in this section.

Now suppose that indeed weights and dissimilarities can be negative. We decompose them as differences of non-negative components. Write

$w_k = w_k^+ - w_k^-$, and $\delta_k = \delta_k^+ - \delta_k^-$. Then we see that

$$\sigma(\theta) = \sigma(0) - \sum_{k=1}^K u_k d_k(\theta) + \sum_{k=1}^K v_k d_k(\theta) + \theta' C^+ \theta - \theta' C^- \theta,$$

with

$$\begin{aligned} u_k &= w_k^+ \delta_k^+ + w_k^- \delta_k^-, \\ v_k &= w_k^+ \delta_k^- + w_k^- \delta_k^+, \end{aligned}$$

and

$$\begin{aligned} C^+ &= \sum_{k=1}^K w_k^+ C_k, \\ C^- &= \sum_{k=1}^K w_k^- C_k. \end{aligned}$$

The next chore, moving towards necessary conditions, is to construct a majorization function. Now obviously we can use

$$\sum_{k=1}^K u_k d_k(\theta) = \theta' B_u(\theta)\theta \geq \theta' B_u(\xi)\xi.$$

We do not touch the term $\theta' C^+ \theta$, and we use Type I majorization a second time in the form

$$\theta' C^- \theta \geq 2\theta' C^- \xi - \xi' C^- \xi.$$

This means that

$$\tau(\theta, \xi) = \sigma(0) - \theta' B_u(\xi)\xi + \theta' C^+ \theta - \theta' C^- \xi + \xi' C^- \xi + \sum_{k=1}^K v_k d_k(\theta)$$

majorizes $\sigma(\theta)$. Moreover $\tau(\bullet, \xi)$ is convex for each ξ.

But maybe we are not satisfied with a convex majorant, and we want a quadratic one. This means we still have to majorize the last term. Following Heiser [1990], we do this by using the arithmetic-geometric mean inequality, which is a special case of Type II majorization. First suppose $d_k(\xi) > 0$. Then

$$d_k(\theta) \leq \frac{1}{d_k(\xi)} \big(d_k^2(\theta) + d_k^2(\xi) \big).$$

This implies

$$\sum_{k=1}^K v_k d_k(\theta) \leq \theta' B_v(\xi)\theta + \xi' B_v(\xi)\xi.$$

Time to collect the results in a theorem.

General Stress Majorization Theorem : If $d_k(\xi) > 0$ for all k, then the function

$$\tau(\theta, \xi) = \sigma(0) - \theta' B_u(\xi)\xi + \theta' C^+\theta - \theta' C^-\xi + \xi' C^-\xi + \theta' B_v(\xi)\theta + \xi' B_v(\xi)\xi$$

majorizes STRESS.

Proof: Above. **Q.E.D.**

We now apply the Necessity by Majorization Theorem. To prevent uninteresting complications, we assume $C^+ > 0$.

General Necessary Condition Theorem: If $\hat{\theta}$ minimizes $\sigma(\bullet)$, then

$$\hat{\theta} = [B_v(\hat{\theta}) + C^+]^{-1}[B_u(\hat{\theta}) + C^-]\hat{\theta}.$$

Proof: Follows from the above. **Q.E.D.**

Observe that if weights and dissimilarities are positive, then $B_v(\bullet) = 0$ and $C^- = 0$, which means we recover our previous definition of stationarity.

3.6. **On the Planes.** Suppose $\mathcal{K} \subseteq \{1, \cdots, K\}$, and

$$\Theta_{\mathcal{K}} = \{\theta \mid \theta' C_k \theta = 0, \forall k \in \mathcal{K}\}.$$

It is clear that $\Theta_{\mathcal{K}}$ is non-empty. In fact if $\mathcal{L}_k$ is the subspace of all θ such that $C_k\theta = 0$, then $\Theta_{\mathcal{K}}$ is the intersection of the $\mathcal{L}_k$, with $k \in \mathcal{K}$.

3.7. **Inverse Scaling.** MDS constructs a mapping of dissimilarities to distances, defined by optimality in the least sqaures sense. This mapping is far from simple. We have seen that there are local minima, sharp ridges, and other irregularities. In order to understand the mapping a bit more completely, we now look at its inverse. Thus in stead of finding a configuration optimal for a given set of dissimilarities, we now look at all dissimilarities for which a given configuration is optimal. It turns out that this question is surprisingly simple to answer.

The key are the stationary equations. For simplicity we assume that $d_k(\theta) > 0$ for all k. Then we must have

$$\sum_{k=1}^{K} w_k \frac{\delta_k}{d_k(\theta)} C_k\theta = \theta.$$

The next theorem tells it all.

Inverse Scaling Theorem: Suppose θ is stationary. Then

$$\delta_k = d_k(\theta) + \sum_{r=1}^{R} u_{kr},$$

where the u_r are the linearly independent solutions of the homogeneous system

$$\sum_{k=1}^{K} w_k \frac{u_k}{d_k(\theta)} C_k \theta = 0.$$

Proof: In the formulation. **Q.E.D.**

Thus we have a simple constructive procedure to find the inverse of the MDS mapping. If we want to know which dissimilarities give the same solution, we take the solution, solve the homogeneous system, and construct the linear manifold of all solutions (which always includes the distances themselves). We could restrict our attention to non-negative solutions for the dissimilarities. Clearly if the distances are positive, there will always be a neighborhood for which the optimal dissimilarities are positive as well. If some distances are zero, it could happen that there is only one set of dissimilarities for which these distances are optimal.

All this cries for numerical examples, which I do not have (yet).

4. A Majorization Algorithm for MDS

We start with outlining the general principles of algorithm construction using the majorization approach. Suppose $\phi(\bullet, \bullet)$ majorizes $\psi(\bullet)$. The corresponding majorization algorithm sets

$$\theta^{(s+1)} = \operatorname{argmin} \phi(\theta, \theta^{(s)}).$$

We suppose, obviously, that the minimum exists. The majorization algorithm works, because of the following basic result.

Improvement Theorem: For all $s = 0, 1, \cdots$ we have

$$\psi(\theta^{(s+1)}) \leq \psi(\theta^{(s)}).$$

Proof: By the definition of the majorization function $\psi(\theta^{(s+1)}) \leq \phi(\theta^{(s+1)}, \theta^{(s)})$, and because of the minimizing property of $\theta^{(s+1)}$ also $\phi(\theta^{(s+1)}, \theta^{(s)}) \leq \phi(\theta^{(s)}, \theta^{(s)})$. Finally, again by definition, $\phi(\theta^{(s)}, \theta^{(s)}) = \psi(\theta^{(s)})$. Combining the three results proves the Improvement Theorem. **Q.E.D.**

Again the proof of the theorem relies on the Sandwich Inequality. Thus majorization algorithms minimize the majorization function to find the next update. In Figure 4 we have our current estimate $x = 1$, and we update to $x = \frac{1}{2}$. In Figure 5 we update $x = 2$ to $x = 4$. In general we update x to $x^+ = \frac{1}{2}x^3$. It is easy to see that in this example gives fast and stable convergence to the local minimum $x = 0$ if we start with $-\sqrt{2} < x^0 < +\sqrt{2}$. The algorithm stops, or stays at the same point, in the stationary points $0, +\sqrt{2}, -\sqrt{2}$. We have rapid divergence to $+\infty$ if $x^0 > +\sqrt{2}$, and to $-\infty$ if $x^0 < -\sqrt{2}$.

We can give a somewhat general result for Type I Majorization. If $\alpha(\theta) = \theta' G \theta$, then the algorithm becomes

$$\theta^{(s+1)} = G^{-1}\mathcal{D}\beta(\theta^{(s)}).$$

In Type II Majorization the algorithm becomes

$$\theta^{(s+1)} = \theta^{(s)} - G^{-1}\mathcal{D}\psi(\theta^{(s)}).$$

4.1. **Majorization for STRESS.** Now apply the principles discussed above to STRESS. In de Leeuw and Heiser [1980a] we find the following result.

Iteration Theorem: For all θ we have $\sigma(\overline{\theta}) \leq \sigma(\theta)$, with equality if and only if $\theta = \overline{\theta}$.

Proof: This is just the Improvement Theorem applied to MDS. **Q.E.D.**

It is clear now, I hope, what the algorithm is. We describe it formally.

(1) Start with a $\theta^{(0)}$. Set $s = 0$.
(2) Compute $\overline{\theta}^{(s)}$.
(3) If $\theta^{(s)} = \overline{\theta}^{(s)}$ go to stop.
(4) $\theta^{(s+1)} = \overline{\theta}^{(s)}$. Set $s = s + 1$. Go back.

Configurations which are equal to their Guttman transforms are called *stationary.* Clearly if the algorithm stops, it stops at a stationary configuration. If it does not stop, which we assume now, it generates an infinite sequence of $\theta^{(s)}$ and $\sigma^{(s)}$ values.

Algorithm Theorem: Suppose the algorithm generates an infinite sequence. Then

(1) There is a $\sigma_\infty \geq 0$, such that $\sigma^{(s)} \downarrow \sigma_\infty$.
(2) If θ_∞ is an accumulation point of $\theta^{(s)}$, then θ_∞ is stationary and $\sigma(\theta_\infty) = \sigma_\infty$.
(3) All $\theta^{(s)}$ with $s \geq 1$ are in the compact set $\{\theta \mid \eta(\theta) \leq 1\}$.
(4) $\eta(\theta^{(s)} - \theta^{(s+1)}) \to 0$.

Proof: By the Iteration Theorem we have $\sigma^{(s+1)} < \sigma^{(s)}$, and thus we generate a decreasing sequence, bounded below by zero, which always converges. This proves part 1.

By the Sandwich Inequality

$$\sigma^{(s+1)} \leq \tau(\theta^{(s+1)}, \theta^{(s)}) = \min_{\xi} \tau(\xi, \theta^{(s)}) \leq \tau(\theta^{(s)}, \theta^{(s)}) = \sigma^{(s)}.$$

But

$$\tau(\theta^{(s+1)}, \theta^{(s)}) = 1 - \eta^2(\theta^{(s+1)}),$$

which proves part 3. Also

$$\tau(\theta^{(s)}, \theta^{(s)}) = 1 - \eta^2(\theta^{(s+1)}) + \eta^2(\theta^{(s)} - \theta^{(s+1)}).$$

It follows, again by the Sandwich Inequality, that $\eta^2(\theta^{(s)} - \theta^{(s+1)}) \rightarrow 0$, which is part 4. We have not proved part 2 yet, but it follows easily from the general convergence theory developed in Zangwill Zangwill [1969]. **Q.E.D.**

Part 4 of the Algorithm Theorem implies that either $\theta^{(s)}$ converges, or $\theta^{(s)}$ has a continuum of accumulation points, all stationary, and all with the same function value σ_∞ Ostrowski [1966]. Of course the algorithm given above is not practical, because it will never stop. We repair this by calling a configuration $\epsilon-$stationary if $\eta(\theta - \overline{\theta}) < \epsilon$. The theorem now says that the algorithm will stop at an $\epsilon-$stationary point after a finite number of iterations.

Guttman Guttman [1968] derived the algorithm by setting the gradient of stress equal to zero. He observed the convergence from any starting point, but did not prove it.

Gradient Theorem: If STRESS is differentiable at $\theta^{(s)}$ then the majorization algorithm can be written as

$$\theta^{(s+1)} = \theta^{(s)} - \mathcal{D}\sigma(\theta^{(s)}).$$

Proof: Follows directly from

$$\mathcal{D}\sigma(\theta) = \theta - \overline{\theta}.$$

Q.E.D.

4.2. **Speed of Convergence.** The iteration we study is a single-step functional iteration of the form $\theta^{s+1} = F(\theta^s)$, with F defined by the Guttman transform, i.e. $F(\theta) = \overline{\theta}$.

The theory of these single-step iterations Ortega and Rheinboldt [1970],Ostrowski [1966] tells us that the speed of convergence is determined by the size of the derivative of F. Thus, we compute the derivative of the Guttman transform. Of course this theory applies only if we converge to a point where all distances are positive.

Convergence Speed Theorem: Suppose θ is an accumulation point of the sequence generated by the algorithm. Suppose $\kappa_+(\theta)$ is the largest eigenvalue of $H(\theta)$. If $\kappa_+(\theta) < 1$ then the algorithm converges linearly to θ with rate $\kappa_+(\theta)$.

Proof: We know that $\mathcal{D}\rho(\theta) = \overline{\theta}$, and thus, using the results in the previous section,

$$\frac{\partial \overline{\theta}}{\partial \theta} = H(\theta).$$

The results now follows from the thorems in section 10.1 of Ortega and Rheinboldt [1970]. In fact, because $H(\theta)$ is real and symmetric, the rate of convergence is $\kappa_+(\theta)$ both in the Q_1 and the R_1 measure. **Q.E.D.**

4.3. **Fixed-step Acceleration.** The basic Guttman transform iteration can be accelerated without much extra cost. This was pointed out in de Leeuw and Heiser [1980a], and we discuss this result more extensively here. We also discuss Steffenson iteration and the Newton-Raphson method.

First consider a fixed-step version of the gradient method. Thus

$$\theta^+ = \theta - \alpha\mathcal{D}\sigma(\theta) = \theta - \alpha(\theta - \overline{\theta}).$$

Relaxation Theorem: If $0 < \alpha < 2$ the fixed step gradient method converges.

Proof: We have

$$\sigma(\theta^+) = \tau(\theta^+, \theta) = 1 - \eta^2(\overline{\theta}) + (1-\alpha)^2\eta^2(\theta - \overline{\theta}) < \tau(\theta, \theta) = \sigma(\theta).$$

This shows that the basic Sandwich Inequality continues to apply. **Q.E.D.**

Optimal Rate Theorem: The optimal step size is

$$\hat{\alpha} = \frac{2}{2 - \kappa_+}.$$

The corresponding convergence rate is

$$\hat{\kappa} = \frac{\kappa_+}{2 - \kappa_+}.$$

Proof: The rate of convergence at θ is $\kappa(\alpha) = \max[|(1-\alpha)+\alpha\kappa_+|, |(1-\alpha) + \alpha\kappa_-|]$. Since $\kappa_- = 0$, we have

$$\kappa(\alpha) = \begin{cases} (1-\alpha) + \alpha\kappa_+ & \text{if } 0 \leq \alpha \leq \frac{2}{2-\kappa_+}, \\ \alpha - 1 & \text{if } \frac{2}{2-\kappa_+} \leq \alpha \leq 2. \end{cases}$$

See Figure 6. The optimum follows by simple computation. **Q.E.D.**

INSERT FIGURE 6 ABOUT HERE

Now suppose $\kappa_+ = 1 - \epsilon$, with ϵ small. Then the optimal step-size is $\hat{\alpha} = 2 + O(\epsilon)$, and the convergence rate is $\hat{\kappa} = \kappa_+^2 + O(\epsilon^2)$. This show that in the case of very slow linear convergence we can approximately half the number of iterations by taking $\alpha = 2$. This is actually the step-size suggested by de Leeuw and Heiser [1980a], who report some numerical experiments to verify that indeed convergence is about twice as fast (with almost no additional computation in each step).

4.4. **Approximating Optimal Step-size.** Experiments with more complicated step-size procedures were reported in Stoop and de Leeuw [1983]. They can be compared with other one-parameter acceleration methods introduced into psychometrics by Ramsay Ramsay [1975]. We can tailor these methods to the MDS problem. Let us look at $\sigma(\theta - \alpha(\theta - \overline{\theta}))$ as a function of α. Clearly the value of the function at 0 is $\sigma(\theta)$, and the derivative at 0 is $-\eta^2(\theta - \overline{\theta})$. If we know the function at another value α, then we can fit

a quadratic with the same values and the same derivative at 0, and we can minimize the quadratic for the step size. The result is

$$\hat{\alpha} = \frac{\alpha^2\eta^2(\theta - \overline{\theta})}{\alpha\eta^2(\theta - \overline{\theta}) - (\sigma(\alpha) - \sigma(0))}.$$

Observe that if $\sigma(\alpha) > \sigma(0)$ then $\hat{\alpha} < \alpha$. Stoop and De Leeuw suggest to start with a generous overestimate of α, say $\alpha = 10$, and iterate the update formula until $\sigma(\alpha) < \sigma(0)$. This leads to at the most three iterations, and often gives big steps in the early stages. All in all, however, just using $\alpha = 2$ throughout seems more economical.

4.5. **Steffenson Acceleration.** We can also look at methods to obtain faster than linear convergence. The most obvious choice is Steffenson's acceleration method Nievergelt [1991], which still works on the basis of our simple majorization updates $\theta^+ = \overline{\theta}$. Define a double sequence of iterations $\theta^{(s,t)}$, with $s = 0, 1, \cdots$, and $t = 0, \cdots, K+1$. We set $\theta^{(s,t+1)} = \overline{\theta}^{(s,t)}$, for $t = 0, \cdots, K$. In order to find $\theta^{(s,0)}$ for $s = 1, 2, \cdots$ we compute

$$\theta^{(s+1,0)} = \theta^{(s,0)} + (I - \Gamma_s)^{-1}(\theta^{(s,1)} - \theta^{(s,0)}).$$

4.6. **Newton's Method.** Let us now look at Newton's method for the MDS problem. Using the results from the previous section we see that

$$\theta^{(s+1)} = \theta^{(s)} - (I - H(\theta^{(s)}))^{-1}(\theta^{(s)} - \overline{\theta}^{(s)}) = (I - H(\theta^{(s)}))^{-1}\overline{\theta}^{(s)}.$$

Under the usual conditions this is locally convergent, with a quadratic rate. From the expression for the Newton update we see that a suitable relaxed version of Newton's method is

$$\theta^{(s+1)} = (I - \alpha^{(s)}H(\theta^{(s)}))^{-1}\overline{\theta}^{(s)}.$$

If the step-size α^s is chosen to be equal to +1 throughout, we have Newton's method, if it is chosen as zero we have the majorization method. If we apply the Sandwich Inequality we see

$$\sigma(\theta^+) \leq \tau(\theta^+, \theta) = 1 - \eta^2(\overline{\theta}) + \eta^2((I - H(\theta))^{-1}H(\theta)\overline{\theta}).$$

4.7. **Negative Dissimilarities.** If dissimilarities (and/or weights) are negative, simple majorization does not work any more. We constructed a suitable majorization function in the previous chapter, and we look at the algorithm here. This case was also treated in detail by Heiser Heiser [1990]. It turns out that standard results on multifacility location problems Francis and Goldstein [1974] can be used in this case. The most interesting part of our previous

majorization analysis was the function

$$\omega(\theta) = \sum_{k=1}^{K} v_k d_k(\theta).$$

This is a weighted sum of Euclidean distances, which is exactly the objective function studied in multifacility location problems.

There are two basic methods, or classes of methods, to solve multifacility location problems. The first one is Weiszfeld's method. This was first described for a simpler location problem in Weiszfeld [1937], and it was first described rigorously in two beautiful papers [Kuhn, 1967, 1973]. The convergence rate of Weiszfeld's method was studied in detail by Katz [1974], but the convergence properties of the algorithm are still not completely resolved [Chandrasekaran and Tamir, 1989]. For our purposes, the most interesting paper on Weiszfeld's method is Vosz and Eckhardt [1980], where a general class of quadratic majorization methods is discussed with Weiszfeld's method as a special case. This class is what we have refrerred to earlier as the Type II majorization mathods.

The majorization for the location problem discussed in Vosz and Eckhardt [1980], turns out to be identical to the one used in Heiser [1990]. A more general version was derived above. It leads to the following algorithm

Theorem 4.1 (Generalized Majorization Theorem)**.** *If $d_k(\theta^{(s)}) > 0$ for all k with $w_k\delta_k^- > 0$ then the algorithm*

$$\theta^{(s+1)} = [C^+ + B_v(\theta^{(s)})]^{-1}[B_u(\theta^{(s)}) + C^-]\theta^{(s)}$$

is a convergent majorization algorithm.

Proof. This is just the majorization result in the General Necessary Condition Theorem. □

The remaining problem is what to do if $d_k(\xi) = 0$. The problem is assumed away in Vosz and Eckhardt [1980], but it is discussed by Heiser. We give a slightly different discussion, based on the algorithms for multifacility location proposed by Calamai and A.R.Conn [1980, 1982], and by Overton [1983].

4.8. **More Type II Majorization.** The technique used in the previous section can be extended quite easily to an even more general problem. Suppose we want to minimize

$$\sigma(\theta) = \sum_{k=1}^{K} \phi(\delta_k - d_k(\theta)),$$

where ϕ has a uniformly bounded second derivative.

We use the inequality

$$\begin{aligned}&\phi((\delta_k - d_k(\xi)) - (d_k(\theta) - d_k(\xi)))\\ \leq&\phi(\delta_k - d_k(\xi)) - \phi'(\delta_k - d_k(\xi))(d_k(\theta) - d_k(\xi)) + G(d_k(\theta) - d_k(\xi))^2\end{aligned}$$

to construct the majorization function. In the same way we can handle the problem of minimizing $\sigma = \phi(\delta - d(\theta))$, where ϕ is a function with a uniform bound on the Hessian.

The majorization technique in this section can be used for various robust versions of least squares scaling, using Huber and Biweight methods Verboon [1990].

4.9. **Global Minimization.**

REFERENCES

P.H. Calamai and A.R.Conn. A stable algorithm for solving the multifacility location problem involving Euclidean distances. *SIAM Journal for Scientific and Statistical Computing*, 1:512–526, 1980.

P.H. Calamai and A.R.Conn. A second-order method for solving the continuous multifacility location problem. In G.A. Watson, editor, *Numerical Analysis*, Berlin, Germany, 1982. Springer-Verlag. Lecture Notes in Mathematics, 912.

R. Chandrasekaran and A. Tamir. Open questions concerning Weiszfeld's algorithm for the Fermat-Weber location problem. *Mathematical Programming*, 44:293–295, 1989.

Jan de Leeuw. Applications of convex analysis to multidimensional scaling. In J.R. Barra, F. Brodeau, G. Romier, and B. van Cutsem, editors, *Recent developments in statistics*, pages 133–145, Amsterdam, The Netherlands, 1977. North Holland Publishing Company.

Jan de Leeuw. Differentiability of Kruskal's Stress at a local minimum. *Psychometrika*, 49:111–113, 1984.

Jan de Leeuw. Convergence of the majorization method for multidimensional scaling. *Journal of Classification*, 5:163–180, 1988.

Jan de Leeuw and Willem J. Heiser. Convergence of correction matrix algorithms for multidimensional scaling. In J.C. Lingoes, editor, *Geometric representations of relational data*, pages 735–752. Mathesis Press, Ann Arbor, Michigan, 1977.

Jan de Leeuw and Willem J. Heiser. Multidimensional scaling with restrictions on the configuration. In P.R. Krishnaiah, editor, *Multivariate Analysis, volume V*, pages 501–522, Amsterdam, The Netherlands, 1980a. North Holland Publishing Company.

Jan de Leeuw and Willem J. Heiser. Theory of multidimensional scaling. In P.R. Krishnaiah, editor, *Handbook of statistics, volume II*. North Holland Publishing Company, Amsterdam, The Netherlands, 1980b.

Jan de Leeuw and Sandra Pruzansky. A new computational method to fit the weighted Euclidean distance model. *Psychometrika*, 43:479–490, 1978.

V.F. Dem'yanov and V.N. Malozemov. *Introduction to minimax*. Dover, New York, New York, 1990.

R.L. Francis and J.M. Goldstein. Location theory: a selective biography. *Operations Research*, 22:400–410, 1974.

W. Glunt, T.L. Hayden, S. Hong, and J. Wells. An alternating projection algorithm for computing the nearest Euclidean distance matrix. *SIAM Journal of Matrix Analysis and Applications*, 11:589–600, 1990.

Patrick Groenen and Willem Heiser. An improved tunneling function for finding a decreasing series of local minima in MDS. Technical Report RR-91-06, Department of Data Theory, University of Leiden, Leiden, The Netherlands, 1991.

Louis Guttman. A general nonmetric technique for fitting the smallest coordinate space for a configuration of points. *Psychometrika*, 33:469–506, 1968.

Willem J. Heiser. A generalized majorization method for least squares multidimensional scaling of pseudodistances that may be negative. *Psychometrika*, 56:7–27, 1990.

J.-B. Hiriart-Urruty. Generalized differentiability, duality abd optimization for problems dealing with differences of convex functions. In *Lecture Notes in economics and mathematical systems, no. 256*, pages 37–70. Springer Verlag, Berlin, Germany, 1985.

I. Norman Katz. Local convergence in Fermat's problem. *Mathematical Programming*, 6:89–104, 1974.

Joseph B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

Harold W. Kuhn. On a pair of dual nonlinear programs. In J. Abadie, editor, *Methods of nonlinear programming*, chapter 3, pages 37–54. North Holland Publishing Company, Amsterdam, The Netherlands, 1967.

Harold W. Kuhn. A note on Fermat's problem. *Mathematical Programming*, 4:98–107, 1973.

Rudolf Mathar and Patrick Groenen. Algorithms in convex analysis applied to multidimensional scaling. Unpublished Manuscript, 1991.

Jacqueline Meulman. *A distance approach to nonlinear multivariate analysis*. DSWO Press, Leiden, The Netherlands, 1986.

Yves Nievergelt. Aitken's and Steffensen's accelerations in several variables. *Numerische Mathematik*, 59:295–310, 1991.

James Ortega and Werner Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, New York, New York, 1970.

Alexander M. Ostrowski. *Solutions of equations and systems of equations*. Academic Press, New York, New York, 1966.

Michael L. Overton. A quadratically convergent method for minimizing a sum of Euclidean norms. *Mathematical Programming*, 27:34–63, 1983.

James O. Ramsay. Solving implicit equations in psychometric data analysis. *Psychometrika*, 40:337–360, 1975.

F. Robert. Calcul du rapport maximal de deux normes sur $\mathcal{R}^n$. *R.I.R.O.*, 1: 97–118, 1967.

A. Shapiro and Y. Yohim. On functions, representable as a difference of two convex functions, and necessary conditions in constrained optimization. Technical report, Ben-Gurion University of the Negev, Beer-Sheva, Israel, 1982.

Ineke Stoop and Jan de Leeuw. The step-size in multidimensional scaling algorithms. Presented at the Third European Psychometric Society Meeting in Jouy-en-Josas, France, 1983.

Peter Verboon. Majorization with iteratively reweighted least squares: a general approach to optimize a class of resistant loss functions. Technical Report RR-90-07, Department of Data Theory, University of Leiden, Leiden, The Netherlands, 1990.

H. Vosz and U. Eckhardt. Linear convergence of generalized Weiszfeld's method. *Computing*, 25:243–251, 1980.

E. Weiszfeld. Sur le point par lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematics Journal*, 43:355–386, 1937.

Willard I. Zangwill. *Nonlinear Programming: a unified approach*. Prentice-Hall, Englewood Cliffs, New Jersey, 1969.

Figure 1. STRESS from far away.

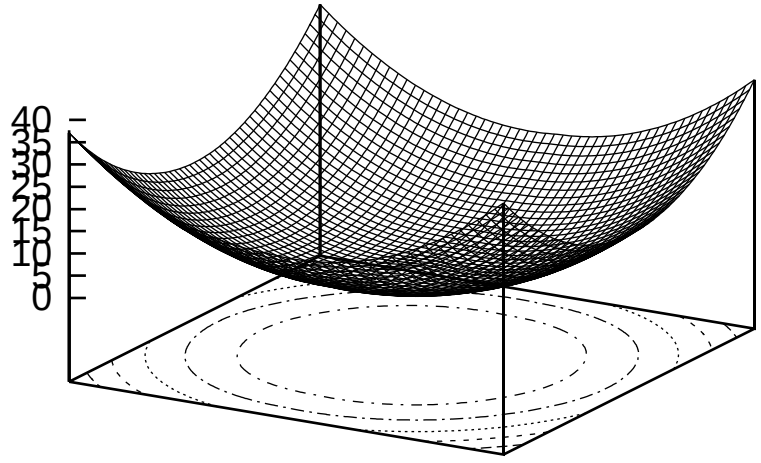

Figure 2. STRESS from up close.

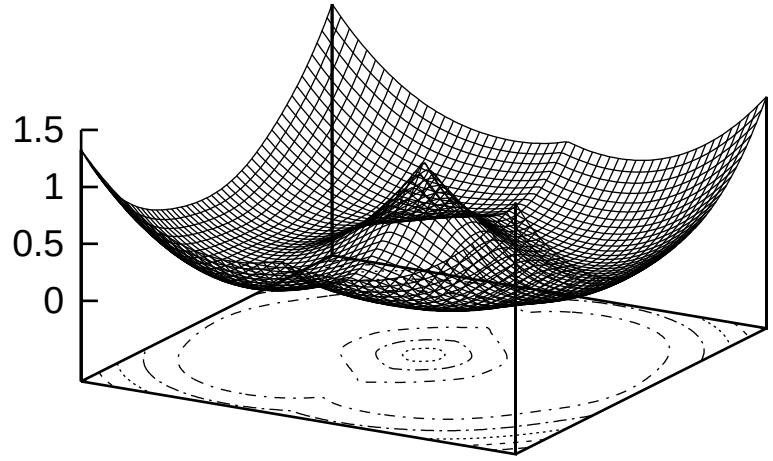

Figure 3. STRESS on the unit circle.

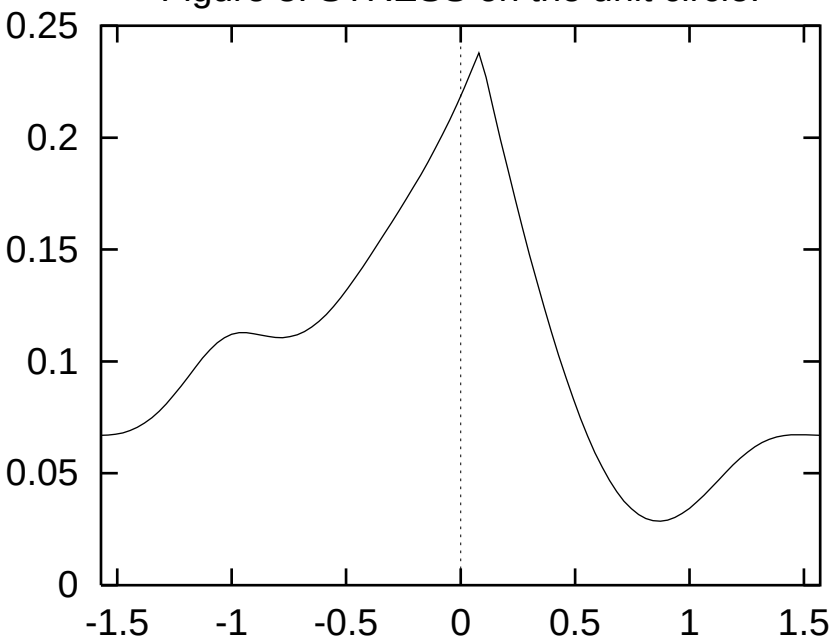

Figure 4. Majorization at 1.

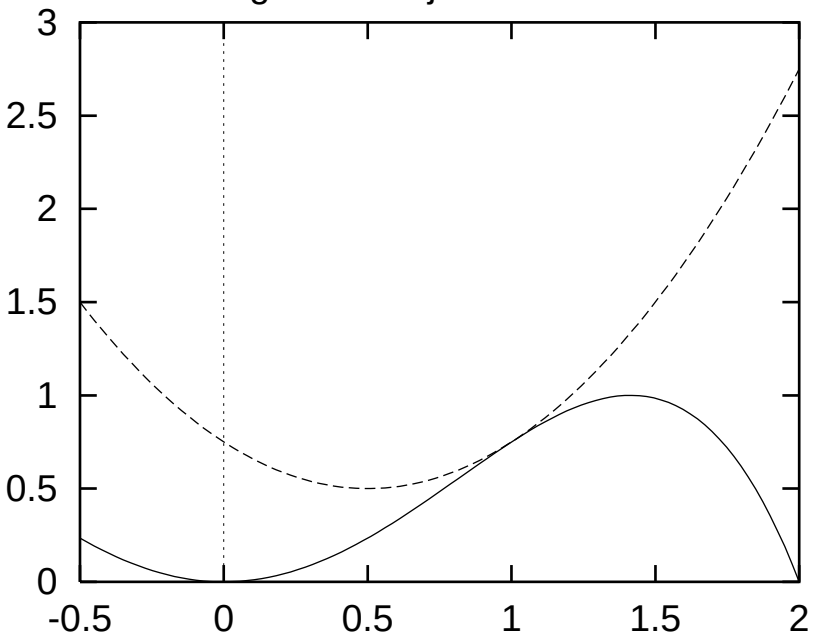

Figure 5. Majorization at 2

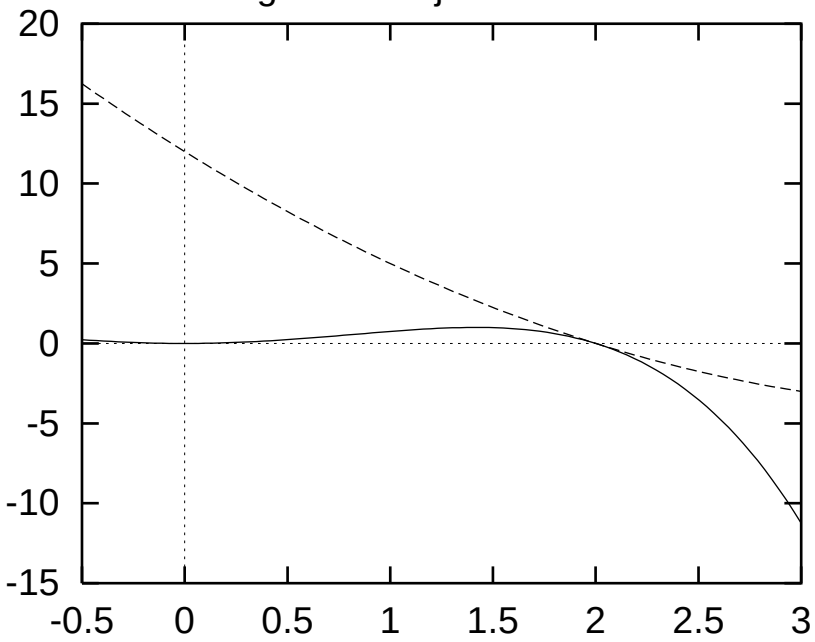

Figure 6. Optimal step-size.