RR-97-02

GLOBAL OPTIMIZATION IN LEAST SQUARES MULTIDIMENSIONAL SCALING

P.J.F. Groenen
W.J. Heiser
J.J. Meulman

Department of Data Theory
Leiden University

Global Optimization in Least Squares Multidimensional Scaling by Distance Smoothing

P.J.F. Groenen* W.J. Heiser † J.J. Meulman*

October 28, 1997

Abstract

Least squares multidimensional scaling is known to have a serious problem of local minima, especially if one dimension is chosen, or if city-block distances are involved. One particular strategy, the smoothing strategy proposed by Pliner (1986, 1996), turns out to be quite successful in these cases. Here, we propose a slightly different approach, called distance smoothing. We extend distance smoothing for any Minkowski distance and show that the S-Stress loss function is a special case. In addition, we extend the majorization approach to multidimensional scaling to have a one-step update for Minkowski parameters larger than 2 and use the results for distance smoothing. We present simple ideas for finding quadratic majorizing functions. The performance of distance smoothing is investigated in several examples, including two simulation studies.

Keywords: multidimensional scaling, Minkowski distances, global optimization, smoothing, majorization.

*Department of Education, Data Theory Group, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands (e-mail: groenen@rulfsw.fsw.leidenuniv.nl). Supported by The Netherlands Organization for Scientific Research (NWO) by grant nr. 575-67-053 for the 'PIONEER' project 'Subject Oriented Multivariate Analysis' to the third author.

†Department of Psychology, Leiden University.

1 Introduction

The main purpose in least squares multidimensional scaling (MDS) is to represent dissimilarities between objects as distances between points in a low dimensional space. At the introduction of computational methods for MDS, it was realized that gradient based minimization methods can only guarantee local minima that need not be global minima. For example, Kruskal (1964) remarks: "If we seek a minimum by the method of steepest descent or by any other method of general use, there is nothing to prevent us from landing at a local minimum other than the true overall minimum."

For a review of the local minimum problem in MDS and the severity of the local minimum problem, see Groenen and Heiser (1996). Several solutions for obtaining a global minimum have been proposed with varying degree of success, e.g., multistart (Kruskal 1964), combinatorial methods for unidimensional scaling (Defays 1978; Hubert and Arabie 1986), combinatorial methods for city-block scaling (Hubert, Arabie, and Hesson-McInnis 1992; Heiser 1989), genetic algorithms (Mathar and Žilinskas 1993), simulated annealing (De Soete, Hubert, and Arabie 1988), and the tunneling method (Groenen 1993; Groenen and Heiser 1996). A different strategy, based on smoothing the loss function was applied successfully to unidimensional scaling (Pliner 1986, 1996) and to city-block MDS (Groenen, Heiser, and Meulman 1997). The good performance of the latter strategy is remarkable, since it shows that continuous minimization strategies (with a reasonable computational effort) can be applied successfully to minimization problems that are combinatorial in nature. The purpose of the present paper is to extend this distance-smoothing strategy to any metric Minkowski distance and investigate how well the strategy performs.

Let us describe formally least squares MDS. The objective is to minimize what is usually called Kruskal's raw Stress, defined as

$$\begin{aligned}\sigma^2(\mathbf{X}) &= \sum_{i<j}^{n} w_{ij}\left[\delta_{ij} - d_{ij}(\mathbf{X})\right]^2 \\ &= \sum_{i<j}^{n} w_{ij}\delta_{ij}^2 + \sum_{i<j}^{n} w_{ij}d_{ij}^2(\mathbf{X}) - 2\sum_{i<j}^{n} w_{ij}\delta_{ij}d_{ij}(\mathbf{X}) \\ &= \eta_\delta^2 + \eta^2(\mathbf{X}) - 2\rho(\mathbf{X}), \qquad (1)\end{aligned}$$

where the objects are represented in a p-dimensional space by points with coordinates given in the rows of the $n \times p$ matrix $\mathbf{X}$, δ_{ij} is a dissimilarity between objects i and j, w_{ij} is a given nonnegative weight, and the Minkowski distance $d_{ij}(\mathbf{X})$ is defined as

$$d_{ij}(\mathbf{X}) = \left(\sum_{s=1}^{p} |x_{is} - x_{js}|^q \right)^{1/q}, \tag{2}$$

with $q \geq 1$ the Minkowski parameter. Note that special cases of Minkowski distances include the city-block distance ($q = 1$), the Euclidean distance ($q = 2$), and the dominance distance ($q = \infty$).

To see why local minima occur, consider the following example with $n = 4$ using Euclidean distances in two dimensions. We investigate the Stress values for point 4, keeping point 1 fixed at (0, 0), point 2 at (5, 0), and point 3 at (2, −1). Suppose that the relevant dissimilarities are $\delta_{14} = 5, \delta_{24} = 3, \delta_{34} = 2$, so that the varying part of Stress can be written as

$$\begin{aligned} \sigma^2(x_{41}, x_{42}) \quad = \quad & (5 - [(x_{41} - 0)^2 + (x_{42} - 0)^2]^{1/2})^2 + \\ & (3 - [(x_{41} - 5)^2 + (x_{42} - 0)^2]^{1/2})^2 + \\ & (2 - [(x_{41} - 2)^2 + (x_{42} + 1)^2]^{1/2})^2. \end{aligned}$$

A graphical representation of the single error term $(5 - [(x_{41} - 0)^2 + (x_{42} - 0)^2]^{1/2})^2$ is given in Figure (1). If this were the only error term in Stress, then a nonunique global minimum of zero Stress can be obtained by placing point 4 anywhere on the circle at distance 5 of the origin. However, considering all three error terms in $\sigma(x_{41}, x_{42})$ simultaneously, see Figure 2, Stress has two local minima. Even though the single error terms have a simple shape (a peek and circular valley), their sum gives rise to the occurrence of multiple local minima.

This paper is organized as follows. First, we introduce distance smoothing for multidimensional scaling and the Huber function involved. Then, we discuss majorizing functions needed to extend the majorization algorithm for MDS of Groenen, Mathar, and Heiser (1995) to include a one-step update for Minkowski distances between Euclidean and dominance distance ($2 \leq q \leq \infty$). This extension is used in the majorization algorithm for minimizing the distance smoothing loss function. To see how well distance smoothing performs, a simulation study is performed on perfect and error perturbed data. In addition, distance smoothing is applied to some examples from the literature. Appendix A explains principles of iterative majorization to minimize a function and methods to find majorizing functions. Appendix B develops the majorizing algorithm for MDS and Appendix C presents the algorithm for distance smoothing.

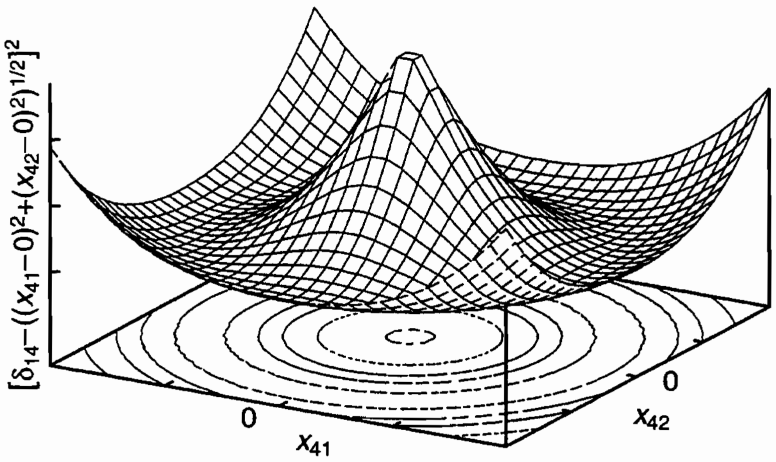

Figure 1: The surface of the error term for objects 1 and 4 of the Stress function, i.e., $(5-[(x_{41}-0)^2+(x_{42}-0)^2]^{1/2})^2$, in which the location of point 4 is varied while keeping point 1 fixed at (0, 0).

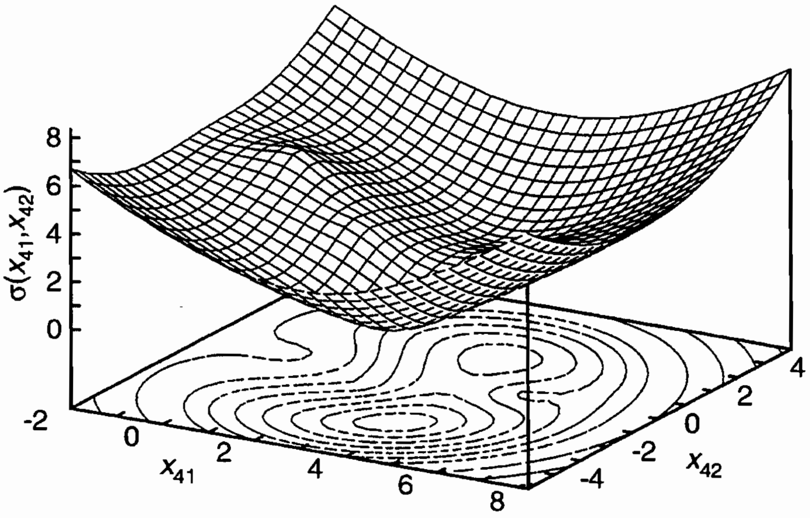

Figure 2: The surface of the Stress function $\sigma(x_{41}, x_{42})$ while varying the location of point 4 keeping the others fixed.

2 Distance Smoothing

It is well known that local minima occur in MDS. For unidimensional scaling, Pliner (1996) suggested the idea of smoothing the absolute difference $|x_{is} - x_{js}|$ by $g_\epsilon(x_{is} - x_{js})$, where $g_\epsilon(t)$ is defined by

$$g_\epsilon(t) = \begin{cases} t^2(3\epsilon - |t|)/3\epsilon^2 + \epsilon/3, & \text{if } |t| < \epsilon, \\ |t|, & \text{if } |t| \geq \epsilon. \end{cases} \tag{3}$$

The smoothness is controlled by the parameter ϵ. As ϵ approaches zero, $g_\epsilon(t)$ approaches $|t|$, but for large ϵ, $g_\epsilon(t)$ is considerably more smooth than $|t|$. The basic idea of smoothing is to replace the absolute values in $d_{ij}(\mathbf{X})$ by $g_\epsilon(t)$, and minimize Stress using this smoothed distance function with decreasing ϵ. Pliner (1996) was very successful in locating global minima in unidimensional scaling, which is infamous for its many local minima.

Groenen et al. (1997) applied this idea to city-block MDS, which also has many local minima. They introduced the term distance smoothing. Their results where also quite promising when compared to other strategies for city-block MDS. Here, we extend distance smoothing to any Minkowski distance (with $q \geq 1$) and provide a majorization algorithm yielding algorithm with monotone nonincreasing series of Stress values without a stepsize procedure. Pliner (1996) briefly suggests how to extend distance smoothing beyond unidimensional scaling but the suggested algorithm needs a stepsize procedure in every iteration to retain convergence.

Instead of $g_\epsilon(t)$, one might as well use other functions that have the property of being smooth if $|t| < \epsilon$, and approach $|t|$ for large $|t|$ (Hubert, personal communication). One such function is used in location theory, where

$$f_\epsilon(t) = (t^2 + \epsilon)^{1/2} \tag{4}$$

is used to smooth the distance (see, e.g., Francis and White 1974; Hubert and Busk 1976). A disadvantage of $f_\epsilon(t)$ in comparison with $g_\epsilon(t)$ is that $f_\epsilon(t) \neq |t|$ unless $\epsilon = 0$, whereas $g_\epsilon(t) = |t|$ for any $|t| \geq \epsilon$. Therefore, we shall not use $f_\epsilon(t)$ in the following.

A third alternative is derived from the Huber function — well known in robust statistics — (Huber 1981, p. 177), i.e.,

$$h_\epsilon(t) = \begin{cases} \frac{1}{2}t^2/\epsilon + \frac{1}{2}\epsilon, & \text{if } |t| < \epsilon, \\ |t|, & \text{if } |t| \geq \epsilon, \end{cases} \tag{5}$$

which differs from the Huber function by the constant term $\frac{1}{2}\epsilon$. The left panel of Figure 3 shows $|t|$ and the two smoothing functions $g_\epsilon(t)$ and $h_\epsilon(t)$.

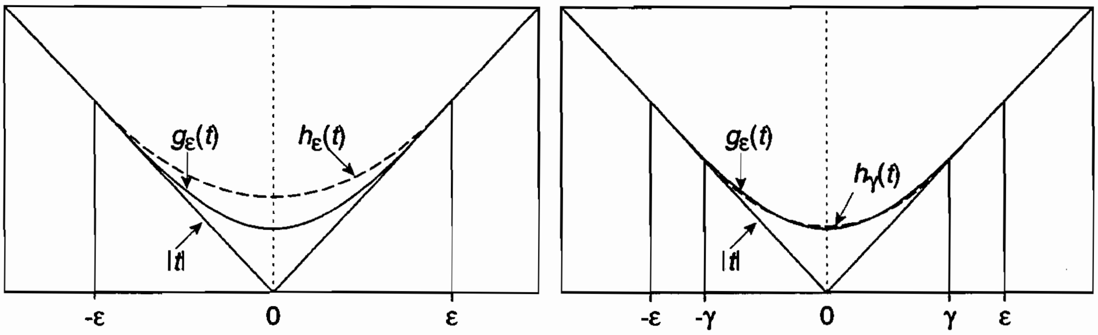

Figure 3: The left panel shows $|t|$ and the smoother functions $g_\epsilon(t)$ of Pliner (1996) and the Huber function $h_\epsilon(t)$. The right panel displays $h_\gamma(t)$ with $\gamma = .6922\epsilon$ so that $h_\gamma(t)$ resembles $g_\epsilon(t)$ as closely as possible.

We prefer the Huber smoother over $g_\epsilon(t)$, because (a) it is more simple ($h_\epsilon(t)$ is quadratic in t, whereas $g_\epsilon(t)$ is a third degree polynomial), (b) it is well known in the literature, (c) by proper choice of ϵ it can approximate $g_\epsilon(t)$ closely, and (d) we shall see later that the S-Stress loss function is a special case.

How should one choose ϵ in $h_\epsilon(t)$ such that it matches $g_\epsilon(t)$ as closely as possible? For the purpose of this comparison, let us replace ϵ in $h_\epsilon(t)$ by γ. To measure the difference between the two curves, we consider the squared area spanned between 0 and ϵ of the differences between the curves, i.e., $\tau(\gamma) = \int_0^\epsilon (h_\gamma(t) - g_\epsilon(t))^2 \mathrm{d}t$. Elementary calculus yields that

$$\tau(\gamma) = \frac{\gamma^5}{180\epsilon^2} - \frac{\gamma^4}{30\epsilon^2} + \frac{2\gamma^3}{15} - \frac{\epsilon\gamma^2}{9} + \frac{\epsilon^3}{63}, \tag{6}$$

which is minimized by choosing $\gamma = .692258/\epsilon$. Thus, for this choice of γ, the two smoothers are as close as possible, which is illustrated in the right panel of Figure 3.

The distance smoothing is obtained by replacing every absolute difference $|x_{is} - x_{js}|$ in (2) by the smoother $h_\epsilon(x_{is} - x_{js})$, i.e.,

$$d_{ij}(\mathbf{X}|\epsilon) = \left(\sum_{s=1}^{p} h_\epsilon^q(x_{is} - x_{js}) \right)^{1/q}. \tag{7}$$

The distance smoothing loss function becomes

$$\begin{aligned} \sigma_\epsilon^2(\mathbf{X}) &= \sum_{i<j}^{n} w_{ij} \left[\delta_{ij} - d_{ij}(\mathbf{X}|\epsilon) \right]^2 \\ &= \sum_{i<j}^{n} w_{ij} \delta_{ij}^2 + \sum_{i<j}^{n} w_{ij} d_{ij}^2(\mathbf{X}|\epsilon) - 2 \sum_{i<j}^{n} w_{ij} \delta_{ij} d_{ij}(\mathbf{X}|\epsilon) \\ &= \eta_\delta^2 + \eta_\epsilon^2(\mathbf{X}) - 2\rho_\epsilon(\mathbf{X}). \end{aligned} \tag{8}$$

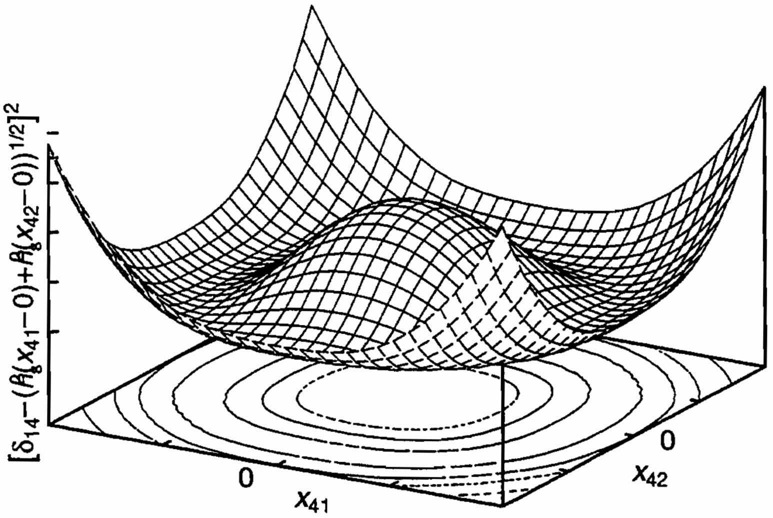

Figure 4: The surface of the error term for objects 1 and 4 of the distance smoothed loss function $\sigma_\epsilon(x_{41}, x_{42})$ in which the location of point 4 is varied while keeping point 1 fixed at (0, 0).

This function has the property that it approaches $\sigma^2(\mathbf{X})$ as ϵ tends to zero. Furthermore, for any $\epsilon > 0$, the gradient and Hessian are defined everywhere. For values of ϵ that are sufficiently large, $\sigma_\epsilon^2(\mathbf{X})$ is more smooth than $\sigma^2(\mathbf{X})$. This property can be seen when the single error term of $\sigma(x_{41}, x_{42})$ in Figure 1 is compared with the same error term of the smoothed function $\sigma_\epsilon(x_{41}, x_{42})$ in Figure 4. Another property shown in Figure 5 is that by increasing the smoothing parameter ϵ, $\sigma_\epsilon(x_{41}, x_{42})$ the two local minima melt into a single minimum.

The distance smoothing algorithm for MDS consists of the following steps:

1. Initialize: set $\epsilon \leftarrow \epsilon_0$, fix number of smoothing steps $r_{\max}$, set start configuration $\mathbf{X}_0$ to a random configuration.

2. For $r = 1$ to $r_{\max}$ do:

    - Reduce ϵ, i.e., $\epsilon \leftarrow \epsilon_0(r_{\max} - r + 1)/r_{\max}$.
    - Minimize $\sigma_\epsilon^2(\mathbf{X})$ using start configuration $\mathbf{X}_{r-1}$. Store the minimum configuration in $\mathbf{X}_r$.

3. Minimize $\sigma^2(\mathbf{X})$ using start configuration $\mathbf{X}_{r_{\max}}$.

Pliner (1996) recommended for unidimensional scaling using $g_\epsilon(t)$ to choose ϵ_0 as $2\max_{1\leq i\leq n} n^{-1}\sum_{j=1}^n \delta_{ij}$. Groenen et al. (1997) used this choice of ϵ_0 for city-block distance smoothing also with success. However, for

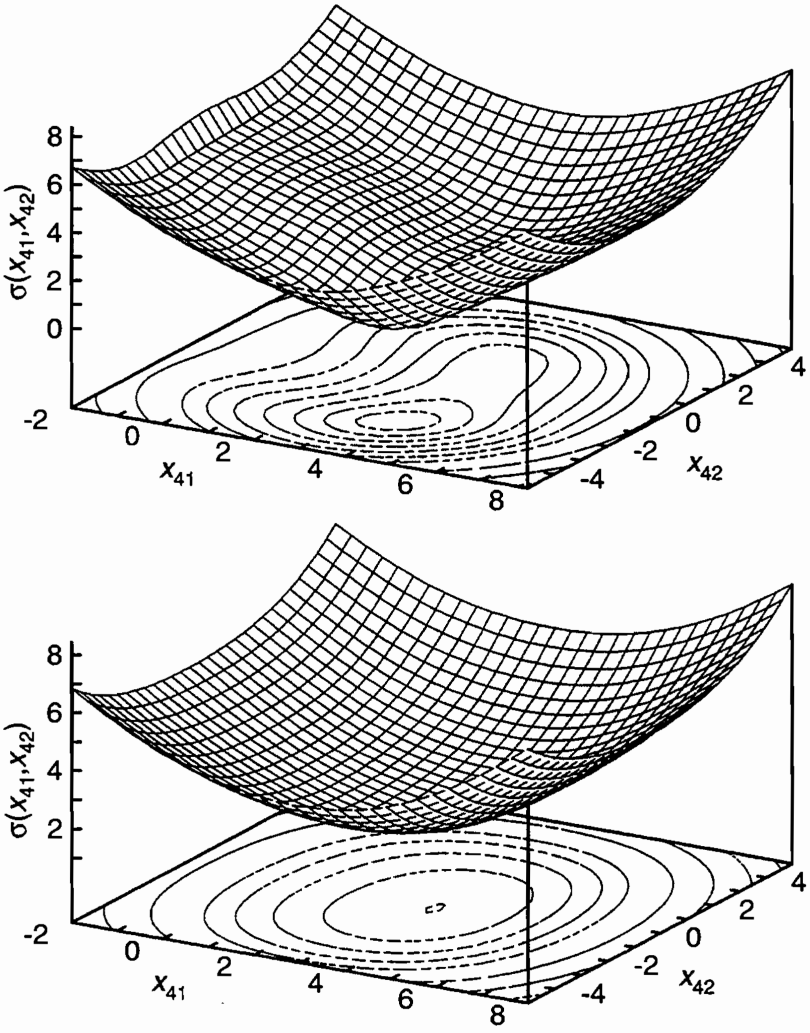

Figure 5: The surface of the distance smoothed loss function $\sigma_\epsilon(x_{41}, x_{42})$ while varying the location of point 4 keeping the others fixed. In the upper panel the smoothing parameter ϵ is set to 2 and in the lower panel to 5.

$q > 1$, distance smoothing works better if the ϵ_0 is chosen wider, i.e., $q^{1/2}.6922 \max_{1 \leq i \leq n}(\sum_{j=1}^{n} w_{ij})^{-1} \sum_{j=1}^{n} w_{ij}\delta_{ij}$.

A final remark concerns nonmetric MDS, or, more general, MDS with transformations of the proximities. In this case, we can proceed as in ordinary nonmetric MDS (e.g., see, Kruskal 1964, or, Borg and Groenen 1997): in (8) the δ_{ij}'s are replaced by $\hat{d}_{ij}$'s, where the $\hat{d}_{ij}$'s are least squares approximations to the distances, constrained in ordinal MDS to retain the order of the proximities and have a fixed sum of squares.

2.1 S-Stress as a Special Case of Distance Smoothing

Consider city-block MDS ($q = 1$). Let π_{ij}^2 be the squared dissimilarities, set $\delta_{ij} = \frac{1}{2}\epsilon p + \frac{1}{2}\pi_{ij}^2\epsilon^{-1}$, and choose ϵ large. Let us assume that $d_{ij}(\mathbf{X}|\epsilon) < \epsilon$ so that $d_{ij}(\mathbf{X}|\epsilon) = \frac{1}{2}\epsilon^{-1}\sum_s(x_{is} - x_{js})^2 + \frac{1}{2}\epsilon p$. This assumption is easily fulfilled by setting ϵ large enough, say $\epsilon = \max_{i<j}\delta_{ij}$, and doing a few minimization steps of $\sigma_\epsilon^2(\mathbf{X})$. Then, $\sigma_\epsilon^2(\mathbf{X})$ can be written as

$$\begin{aligned}
\sigma_\epsilon^2(\mathbf{X}) &= \sum_{i<j} w_{ij}\left[\delta_{ij} - d_{ij}(\mathbf{X}|\epsilon)\right]^2 \\
&= \sum_{i<j} w_{ij}\left[\tfrac{1}{2}\epsilon p + \tfrac{1}{2}\pi_{ij}^2\epsilon^{-1} - \tfrac{1}{2}\epsilon^{-1}\sum_s(x_{is} - x_{js})^2 - \tfrac{1}{2}\epsilon p\right]^2 \\
&= 1/(4\epsilon^2)\sum_{i<j} w_{ij}\left[\pi_{ij}^2 - \sum_s(x_{is} - x_{js})^2\right]^2, \qquad (9)
\end{aligned}$$

which is exactly $1/(4\epsilon^2)$ times the S-Stress loss function of Takane, Young, and De Leeuw (1977). This shows that S-Stress is a special case of distance smoothing.

3 Majorizing Functions for MDS with Minkowski Distances

To minimize (1) over $\mathbf{X}$, we elaborate on the majorization approach to multidimensional scaling (see, e.g. De Leeuw 1988), and in particular its extension for Minkowski distances proposed by Groenen et al. (1995). For more details on iterative majorization, we refer to De Leeuw (1994), Heiser (1995), or, for an introduction, to Borg and Groenen (1997). Some background and definitions of iterative majorization are discussed in Appendix A.

To majorize $\sigma^2(\mathbf{X})$ in (1) we apply linear majorization to $-\rho(\mathbf{X})$ and quadratic majorization to $\eta^2(\mathbf{X})$. The sum of these majorizing functions will majorize $\sigma^2(\mathbf{X})$.

3.1 Linear Majorization of $-d_{ij}(\mathbf{X})$

Groenen et al. (1995) majorized $-\rho(\mathbf{X})$ as follows. First, we notice that $-\rho(\mathbf{X})$ is a sum of $-d_{ij}(\mathbf{X})$ weighted by the nonnegative value $w_{ij}\delta_{ij}$. Using Hölder's inequality, Groenen et al. proved the linear majorizing inequality

$$-d_{ij}(\mathbf{X}) \leq -\sum_{s=1}^{p}(x_{is}-x_{js})(y_{is}-y_{js})b_{ijs}^{(1)}, \tag{10}$$

where

$$b_{ijs}^{(1)} = \begin{cases} \dfrac{|y_{is}-y_{js}|^{q-2}}{d_{ij}^{q-1}(\mathbf{Y})} & \text{if } |y_{is}-y_{js}| > 0 \text{ and } 1 \leq q < \infty, \\ 0 & \text{if } |y_{is}-y_{js}| = 0 \text{ and } 1 \leq q < \infty, \\ |y_{is}-y_{js}|^{-1} & \text{if } |y_{is}-y_{js}| > 0, q = \infty, \text{ and } s = s^*, \\ 0 & \text{if } (|y_{is}-y_{js}| = 0 \text{ or } s \neq s^*) \text{ and } q = \infty), \end{cases}$$

and s^* is defined for $q = \infty$ as the dimension on which the distance is maximal, i.e., $d_{ij}(\mathbf{Y}) = |y_{is}-y_{js}|$.

3.2 Quadratic Majorization of $d_{ij}^2(\mathbf{X})$

Here we concentrate on majorizing $d_{ij}^2(\mathbf{X})$. Groenen et al. (1995) proposed a quadratic majorizing inequality for the squared distance for $1 \leq q \leq 2$, i.e.,

$$d_{ij}^2(\mathbf{X}) \quad \leq \quad \sum_{s=1}^{p} a_{ijs}^{(1\leq q\leq 2)}(x_{is}-x_{js})^2, \tag{11}$$

where

$$a_{ijs}^{(1\leq q\leq 2)} = \frac{|y_{is}-y_{js}|^{q-2}}{d_{ij}^{q-2}(\mathbf{Y})}.$$

Note that if $d_{ij}(\mathbf{Y}) = 0$ or $|y_{is}-y_{js}| = 0$, (11) may become indetermined. If this is the case, we replace $|y_{is}-y_{js}|^{q-2}/d_{ij}^{q-2}(\mathbf{Y})$ by a small positive value. This substitution violates the equality condition of the majorizing function, but by making it small enough, it will usually not affect the behavior of the algorithm. We will assume this implicitly in the sequel, whenever needed.

For $q > 2$ the inequality sign in (11) is reversed, so that it cannot be used anymore as a quadratic majorizing function. In the sequel of this section we develop a new quadratic majorizing inequality for this case. We need an upperbound of the largest eigenvalues of the Hessian of $d_{ij}^2(\mathbf{X})$. For notational convenience we substitute $|x_{is}-x_{js}|$ by t_s so that the squared Minkowski distance becomes $d^2(\mathbf{t}) = (\sum_{s=1}^p t_s^q)^{2/q}$. The elements of the gradient are $\partial d^2(\mathbf{t})/\partial t_s = 2t_s^{q-1}/d^{q-2}(\mathbf{t})$. The Hessian $\mathbf{H} = \nabla^2 d^2(\mathbf{t})$ can be expressed as the difference $\mathbf{H} = 2(q-1)\mathbf{D} - 2(q-2)\mathbf{h}\mathbf{h}'$, where $\mathbf{D}$ is diagonal matrix with diagonal elements $(t_s/d(\mathbf{t}))^{q-2}$ and $\mathbf{h}$ has elements $(t_s/d(\mathbf{t}))^{q-1}$. First, we note that the normalization of $\mathbf{t}$ is irrelevant, because t_s is divided by $d(\mathbf{t})$ everywhere in $\mathbf{H}$. Thus, without loss of generality, we may assume $d(\mathbf{t}) = 1$, so that the diagonal elements of $\mathbf{D}$ simplify into t_s^{q-2} and the elements of $\mathbf{h}$ into t_s^{q-1}. We also assume for convenience that t_s is ordered decreasingly.

Let $\lambda(\mathbf{H})$ denote the largest eigenvalue of $\mathbf{H}$. An upper bound of the largest eigenvalue can be found as follows. Let $\mathbf{D}^{-1/2}$ be the diagonal matrix with elements $t_s^{1-q/2}$ if $t_s > 0$ and 0 otherwise. Then, $\mathbf{H}$ can be expressed as:

$$\mathbf{H} = \mathbf{D}^{1/2}[2(q-1)\mathbf{I} - 2(q-2)\mathbf{D}^{-1/2}\mathbf{h}\mathbf{h}'\mathbf{D}^{-1/2}]\mathbf{D}^{1/2} = \mathbf{D}^{1/2}\mathbf{V}\mathbf{D}^{1/2}.$$

Magnus and Neudecker (1988, p. 237) state that $\lambda(\mathbf{H}) \leq \lambda(\mathbf{D})\lambda(\mathbf{V})$. The largest eigenvalue of $\mathbf{D}$ is equal to one, which is obtained by choosing $t_1 = 1$ and the remaining $t_s = 0$ $(1 < s \leq p)$. The largest eigenvalue of $\mathbf{V}$ is equal to $2(q-1)$, since $\mathbf{D}^{-1/2}\mathbf{h}\mathbf{h}'\mathbf{D}^{-1/2}$ has eigenvalue 1, so that $\mathbf{V}$ has eigenvalues 2 for the eigenvector $\mathbf{D}^{-1/2}\mathbf{h}$ and $2(q-1)$ for the remaining eigenvectors. Thus, an upper bound for the largest eigenvalue of $\mathbf{H}$ is $\lambda = 2(q-1)$.

Numerical experimentation yields an even lower upper bound for $q > 2$, i.e., $\lambda(\mathbf{H}) \leq (q-1)2^{2/q}$, where equality occurs whenever $t_1 = t_2$ and $t_3 = \ldots = t_p = 0$. However, we were not able to find a proof for this proposition.

Combining the results above with those on quadratic majorization at the end of Appendix A, we get a quadratic majorizing function for the squared Minkowski distance with $q > 2$, i.e.,

$$d_{ij}^2(\mathbf{X}) \leq a^{(q>2)}\sum_{s=1}^p(x_{is}-x_{js})^2 - 2\sum_{s=1}^p(x_{is}-x_{js})(y_{is}-y_{js})b_{ijs}^{(q>2)} + c_{ij}^{(q>2)}, \quad (12)$$

where

$$\begin{aligned} a^{(q>2)} &= \lambda/2 \\ b_{ijs}^{(q>2)} &= \begin{cases} a^{(q>2)} - |y_{is}-y_{js}|^{q-2}/d_{ij}^{q-2}(\mathbf{Y}) & \text{if } |y_{is}-y_{js}| > 0 \\ 0 & \text{if } |y_{is}-y_{js}| = 0 \end{cases} \\ c_{ij}^{(q>2)} &= a^{(q>2)}\sum_{s=1}^p(y_{is}-y_{js})^2 - d_{ij}^2(\mathbf{Y}). \end{aligned}$$

For the dominance distance, $q = \infty$, λ also becomes ∞, which is clearly not desirable. For this case, a better quadratic majorizing function can be found. Let t_s be defined as above, where the elements are ordered decreasingly. This means that $d^2(\mathbf{t}) = (\sum_s t_s^\infty)^{2/\infty} = (\max_s t_s)^2 = t_1^2$. To find a quadratic majorizing function $f(\mathbf{t},\mathbf{u}) = a(\mathbf{u})\mathbf{t}'\mathbf{t} - 2\mathbf{t}'\mathbf{b}(\mathbf{u}) + c(\mathbf{u})$, we need to meet four conditions:

1. $d^2(\mathbf{u}) = f(\mathbf{u},\mathbf{u})$,
2. $\nabla d^2(\mathbf{u}) = \nabla f(\mathbf{u},\mathbf{u})$,
3. $d^2(\mathbf{Pu}) = f(\mathbf{Pu},\mathbf{u})$, where $\mathbf{P}$ is a permutation matrix that interchanges u_1 and u_2, so that $\mathbf{Pu}$ has elements $u_2, u_1, u_3, u_4, \ldots, u_p$, and
4. $\nabla d^2(\mathbf{Pu}) = \nabla f(\mathbf{Pu},\mathbf{u})$.

These conditions imply that $f(\mathbf{t},\mathbf{u})$ should touch $d^2(\mathbf{t})$ at $\mathbf{u}$ and at $\mathbf{Pu}$. Any $f(\mathbf{t},\mathbf{u})$ satisfying these four conditions has $a \geq 1$ and, hence, is a majorizing function of $d(\mathbf{t})$. Conditions (2) and (4) yield

$$\begin{bmatrix} 2u_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 2au_1 - 2b_1 \\ 2au_2 - 2b_2 \\ 2au_3 - 2b_3 \\ \vdots \\ 2au_p - 2b_p \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ 2u_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 2au_2 - 2b_1 \\ 2au_1 - 2b_2 \\ 2au_3 - 2b_3 \\ \vdots \\ 2au_p - 2b_p \end{bmatrix}.$$

This system of equalities is solved by $a = u_1/(u_1 - u_2)$, $b_1 = b_2 = au_2$, and $b_s = au_s$ for $s > 2$. Since $u_1 > u_2$, a is greater than 1, so that the majorizing function has a greater second derivative than $d^2(\mathbf{t})$. If $u_1 = u_2$, then a is not defined. For such cases, we add a small value ε to u_1. By choosing ε small enough, convergence is retained for all practical purposes.

Let ϕ_s be an index for pair i,j that orders the values $|y_{is} - y_{js}|$ decreasingly, so that $|y_{i\phi_1} - y_{j\phi_1}| \leq |y_{i\phi_2} - y_{j\phi_2}| \leq \ldots \leq |y_{i\phi_p} - y_{j\phi_p}|$. The majorizing function for $q = \infty$ becomes

$$d_{ij}^2(\mathbf{X}) \leq a_{ij}^{(q=\infty)} \sum_s (x_{is} - x_{js})^2 - 2\sum_s (x_{is} - x_{js})(y_{is} - y_{js}) b_{ijs}^{(q=\infty)} + c_{ij}^{(q=\infty)}, \quad (13)$$

where

$$a_{ij}^{(q=\infty)} = \begin{cases} \dfrac{|y_{i\phi_1} - y_{j\phi_1}|}{|y_{i\phi_1} - y_{j\phi_1}| - |y_{i\phi_2} - y_{j\phi_2}|} & \text{if } |y_{i\phi_1} - y_{j\phi_1}| - |y_{i\phi_2} - y_{j\phi_2}| > \varepsilon, \\ \dfrac{|y_{i\phi_1} - y_{j\phi_1}| + \varepsilon}{\varepsilon} & \text{if } |y_{i\phi_1} - y_{j\phi_1}| - |y_{i\phi_2} - y_{j\phi_2}| \leq \varepsilon, \end{cases}$$

$$b_{ijs}^{(q=\infty)} = \begin{cases} a_{ij}^{(q=\infty)} & \text{if } s \neq \phi_1, \\ a_{ij}^{(q=\infty)} \dfrac{|y_{i\phi_2} - y_{j\phi_2}|}{|y_{i\phi_1} - y_{j\phi_1}|} & \text{if } s = \phi_1, \end{cases}$$

$$c_{ij}^{(q=\infty)} = \sum_s (2b_{ijs}^{(q=\infty)} - a_{ij}^{(q=\infty)})(y_{is} - y_{js})^2 + d_{ij}^2(\mathbf{Y}).$$

Appendix B combines the majorizing functions discussed here and presents the majorizing algorithm for minimizing Stress.

in t that majorizes $h_\epsilon^2(t)$ such that the conditions Q1 to Q3 in Appendix A are satisfied. This yields

$$h_\epsilon^2(x_{is}-x_{js}) \leq a^{(h^2)}(x_{is}-x_{js})^2 - 2(x_{is}-x_{js})(y_{is}-y_{js})b_{ijs}^{(h^2)} + c_{ijs}^{(h^2)}, \quad (15)$$

where

$$\begin{aligned} a^{(h^2)} &= \kappa, \\ b_{ijs}^{(h^2)} &= \begin{cases} a^{(h^2)} - \frac{(y_{is}-y_{js})^2}{2\epsilon^2} - \frac{1}{2} & \text{if } |y_{is}-y_{js}| < \epsilon, \\ a^{(h^2)} - 1 & \text{if } |y_{is}-y_{js}| \geq \epsilon, \end{cases} \\ c_{ijs}^{(h^2)} &= h_\epsilon^2(y_{is}-y_{js}) - a^{(h^2)}(y_{is}-y_{js})^2 + 2(y_{is}-y_{js})^2 b_{ijs}^{(h^2)}. \end{aligned}$$

Appendix C combines the majorizing functions to obtain a majorization algorithm for minimizing $\sigma_\epsilon^2(\mathbf{X})$.

5 Numerical Experiments

To test the performance of distance smoothing, the method was applied to several data sets, where the following factors are varied: (a) the dimensionality, (b) the Minkowski parameter q, and (c) the minimization method. Our main interest here was to study how often the method is capable in detecting the global minimum. Note that we report the Kruskal's Stress-1 values, σ, which are equal to $(\sigma^2(\mathbf{X})/\eta_\delta^2)^{1/2}$ at a local minimum if we allow the configuration to be optimally dilated (Borg and Groenen 1997, p. 201). The minimization methods we used are: (a) distance smoothing, (b) SMACOF (Scaling by MAjorizing a COmplicated Function) of De Leeuw and Heiser (1977) for $q = 2$, Groenen et al. (1995) for $1 \leq q \leq 2$, and Section 3 for $q > 2$, and (c) KYST of Kruskal, Young, and Seery (1978).

The algorithms were specified as follows. We used 20 smoothing steps for distance smoothing. The relaxed update was used. Every smoothing step was terminated whenever the number of iterations exceeded 1000 or two subsequent values of $\sigma^2(\mathbf{X})/\eta_\delta^2$ did not change more than 10^{-6}. SMACOF and KYST also had a maximum of 1000 iterations or where terminated if Stress differed less than 10^{-8} for SMACOF, or the ratio of subsequent Stress values was between 1 and .999999 for KYST.

We report two simulation studies and two experiments on empirical data.

5.1 Perfect Distance Data

The first experiment involved the recovery of perfect distance data varying dimensionality ($p = 1, 2$, or 3) and Minkowski parameter (q = 1, 2, 3, 4, 5,

4 Majorization Functions for Distance Smoothing

Below we derive majorizing functions needed for minimizing $\sigma^2_\epsilon(\mathbf{X})$. We use the majorizing inequalities of the previous section by substituting $h_\epsilon(x_{is}-x_{js})$ for $|x_{is}-x_{js}|$, $h_\epsilon(y_{is}-y_{js})$ for $|y_{is}-y_{js}|$, $d_{ij}(\mathbf{X}|\epsilon)$ for $d_{ij}(\mathbf{X})$, and $d_{ij}(\mathbf{Y}|\epsilon)$ for $d_{ij}(\mathbf{Y})$. This substitution leads to functions in $h^2_\epsilon(x_{ik}-x_{jk})$ and $-h_\epsilon(x_{ik}-x_{jk})$. Thus, to majorize $\sigma^2_\epsilon(\mathbf{X})$ we only have to derive majorizing inequalities for $h^2_\epsilon(x_{ik}-x_{jk})$ and $-h_\epsilon(x_{ik}-x_{jk})$.

Consider $-h_\epsilon(x_{ik}-x_{jk})$, or, after substitution of $t = x_{ik}-x_{jk}$, $-h_\epsilon(t)$. If $|t| \geq \epsilon$ then $h_\epsilon(t) = |t|$. Applying the Cauchy-Schwarz inequality for one term only gives

$$-|t| \leq -\frac{tu}{|u|}$$

(Kiers and Groenen 1996). Note that in (5) $|u| > \epsilon > 0$, so that division by zero cannot occur. On the other hand, if $|t| < \epsilon$, then we have to majorize $-h_\epsilon(t) = -\frac{1}{2}t^2/\epsilon - \frac{1}{2}\epsilon$. This function is concave in t and, thus, can be linearly majorized. The majorizing function is derived from the inequality $(t-u)^2 \geq 0$, or, equivalently, $-t^2 \leq u^2 - 2tu$. Using these majorizing inequalities and multiplying with $h_\epsilon(y_{ik}-y_{jk})$, we obtain the result

$$-h_\epsilon(x_{is}-x_{js})h_\epsilon(y_{ik}-y_{jk}) \leq -(x_{is}-x_{js})(y_{is}-y_{js})b^{(-h)}_{ijs} + c^{(-h)}_{ijs}, \qquad (14)$$

where

$$b^{(-h)}_{ijs} = \begin{cases} 1 & \text{if } |y_{is}-y_{js}| \geq \epsilon, \\ h_\epsilon(y_{is}-y_{js})/\epsilon & \text{if } |y_{is}-y_{js}| < \epsilon, \end{cases}$$

$$c^{(-h)}_{ijs} = \begin{cases} 0 & \text{if } |y_{is}-y_{js}| \geq \epsilon, \\ h_\epsilon(y_{is}-y_{js})[(y_{ik}-y_{jk})^2/\epsilon - h_\epsilon(y_{is}-y_{js})] & \text{if } |y_{is}-y_{js}| < \epsilon. \end{cases}$$

What remains to be done is to find a (quadratic) majorizing inequality for $h^2_\epsilon(t)$. Since $h_\epsilon(t)$ is convex in t on the interval $-\epsilon < t < \epsilon$, so is its square. Convexity implies that on this interval $h^2_\epsilon(t)$ has a positive second derivative. Because at $|t| = \epsilon$ the functions $[\frac{1}{2}t^2/\epsilon + \frac{1}{2}\epsilon]^2$ and t^2 have the same function value and the same first derivative, there must exist an upper bound κ of the second derivative $h^2_\epsilon(t)$. This implies that on the interval $-\epsilon < t < \epsilon$ the curvature of $h^2_\epsilon(t)$ never becomes larger than κ. It can be verified that $\kappa = 4$ at $t = \epsilon$. Outside this interval, the second derivative of $h^2_\epsilon(t) = t^2$ equals 2, so that the maximum second derivative of $h^2_\epsilon(t)$ over the entire interval equals $\max(\kappa, 2) = \kappa$. Therefore, there exists a quadratic function

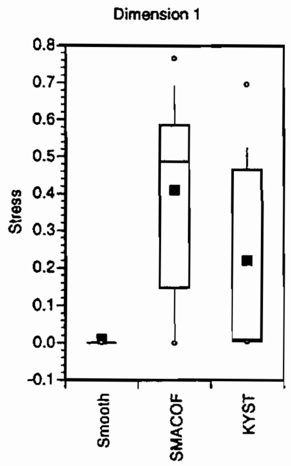

Figure 6: Distribution of Stress value of 100 random starts for perfect distance data in one dimension.

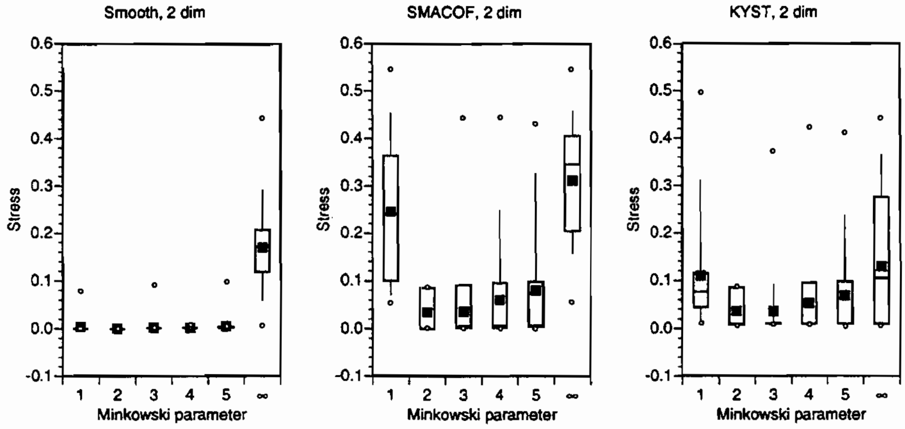

Figure 7: Distribution of Stress value of 100 random starts for perfect distance data in two dimensions.

and ∞). Note that the Minkowski parameter is irrelevant for unidimensional scaling. For each combination of p and q a random configuration of ten points was determined and their distances served as the dissimilarity matrix. Then, for each of the three minimization methods the Stress values of the local minima was recorded for 100 random starts. Figures 6, 7, and 8 give the distribution of the Stress values in 1, 2, and 3 dimensions. The distribution of the values are presented in boxplots, where the ends of the box mark the 25th and 75th percentile, the end points of the lines mark the 10th and 90th percentile, the line in the box the median, the square marks the mean, and the open circles the extremes.

In one dimension, SMACOF and KYST only rarely succeeded in finding the zero Stress solution, whereas distance smoothing always found the global

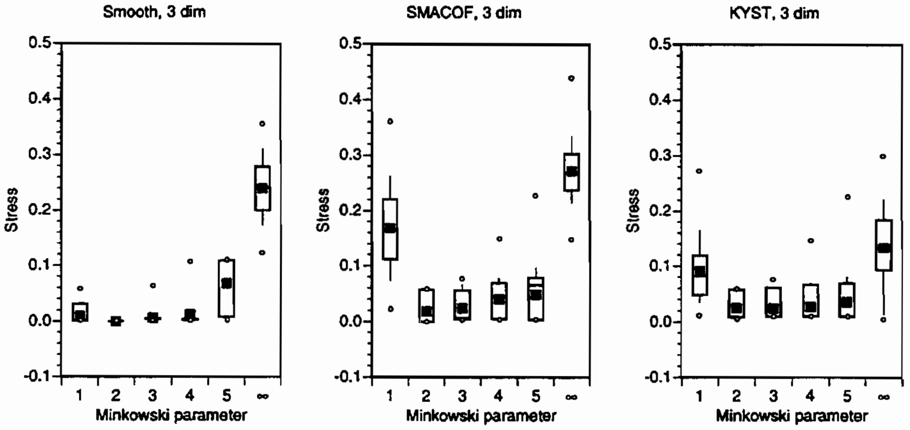

Figure 8: Distribution of Stress value of 100 random starts for perfect distance data in three dimensions.

minimum. These results are completely in line with those found by Pliner (1996) and with the fact that unidimensional scaling is a combinatorial problem (see, e.g., De Leeuw and Heiser 1977; Defays 1978; Hubert and Arabie 1986; Groenen and Heiser 1996), which makes it hard to find proper local minima for gradient based minimization methods as SMACOF and KYST.

In two dimensions and for all q except $q = \infty$, almost all runs of distance smoothing yielded a zero Stress local minimum. SMACOF and KYST had more difficulty in finding the global minimum, especially for $q = 1, 4$, and 5. KYST performed remarkably well for $q = 3$. For $q = 2$, SMACOF and KYST found two different Stress values, i.e., either 0 or 0.085.

In three dimensions, distance smoothing still performed well, with the exception of $q = 1$ where about 30% of the solutions yielded nonzero Stress values and $q = 5$ where about 60% yielded nonzero Stress. SMACOF and KYST had difficulty in finding zero Stress solutions, specifically for $q = 1$.

In two and three dimensions, all three methods had difficulty in reconstructing the zero Stress solution if the dominance distance is used. However, in this case KYST performed systematically better than distance smoothing and SMACOF.

5.2 Error Perturbed Distance Data

In real applications, it is not very likely to come across perfect distance data. Therefore, we have set up a simulation experiment involving error perturbed distances to investigate how distance smoothing performs. Apart from the three minimization methods we varied the following factors: $n = 10, 20$; $p = 2, 3$; $q = 1, 2, \infty$; error $= 15\%$, and 30%. This design produces 24 different dissimilarity matrices for which 100 random starts were done for each of the three minimization methods yielding a total of 7200 runs. To obtain an error perturbed distance matrix $\mathbf{\Delta}$ we started by generating a configuration $\mathbf{X}$ of randomly distributed points within distance one of the origin. Then, the dissimilarities were computed by

$$\delta_{ij} = \left(\sum_{s=1}^{p} |x_{is}^{(e)} - x_{js}^{(e)}|^q \right)^{1/q} \tag{16}$$

(Ramsay 1969), where $x_{is}^{(e)} = x_{is} + N(0, e)$, where $N(0, e)$ denotes the normal distribution with mean zero and variance e. One of the advantages of this method is that the dissimilarities are always positive while there is an underlying true configuration.

Table 1 reports the average Stress and the standard deviation for each cell in the design. Distance smoothing seemed to perform very well for city-block distances, much better than SMACOF and KYST. Also for Euclidean distances, the average Stress was lowest for distance smoothing with a standard error of almost zero, indicating that distance smoothing almost always located the global minimum. For the dominance distance, distance smoothing performed worse than KYST: the average Stress for distance smoothing was systematically higher than for KYST. SMACOF showed the worst performance of the three methods in this case.

An analysis of variance was done to find the significant effects in the design. We chose all those effects which were able to explain more than 5% of the variance: we included the main effects and two two-way interactions (method by q and q by p). The results are reported in Table 2. This model explains about 87% of the total variance. We see that the method does make a difference, and in particular the methods differed significantly for varying Minkowski distances.

To see which method was best capable in locating the global minimum, we compare the minimal Stress values found in 100 random starts in Table 3. For $q = 1$, distance smoothing found the lowest minimum in all cases, KYST found it only once, and SMACOF never found it. For $q = 2$, all methods found the same lowest Stress. For $q = \infty$, KYST always found the lowest Stress. Using a rational startconfiguration obtained by classical scaling, the three methods behaved as we saw before (see Table 4). Distance smoothing performed best for $q = 1$. For $q = 2$, the three methods almost always found the same Stress. For $q = \infty$, KYST performed better than distance smoothing and SMACOF.

Table 1: Average Stress $\bar{\sigma}$ over 100 random starts (and the standard deviation s.d. σ) for error perturbed distances (15% and 30%) varying dimensionality p (two and three), number of objects n (10 and 20), and Minkowski distance (city-block, Euclidean, and dominance).

			smooth		SMACOF		KYST	
p	n	$e\%$	$\bar{\sigma}$	s.d. σ	$\bar{\sigma}$	s.d. σ	$\bar{\sigma}$	s.d. σ
$q = 1$								
2	10	15	.212	.001	.330	.062	.283	.049
		30	.272	.000	.400	.057	.346	.042
2	20	15	.335	.000	.426	.031	.405	.028
		30	.396	.005	.469	.022	.450	.019
3	10	15	.139	.005	.249	.053	.205	.033
		30	.170	.004	.260	.041	.220	.030
3	20	15	.256	.002	.326	.022	.309	.017
		30	.275	.002	.348	.020	.335	.019
$q = 2$								
2	10	15	.235	.000	.244	.016	.243	.014
		30	.222	.000	.245	.033	.244	.033
2	20	15	.320	.000	.331	.015	.327	.012
		30	.330	.002	.336	.013	.335	.013
3	10	15	.159	.000	.164	.008	.164	.008
		30	.205	.000	.205	.000	.205	.000
3	20	15	.258	.000	.262	.005	.262	.005
		30	.316	.001	.318	.003	.318	.003
$q = \infty$								
2	10	15	.292	.045	.338	.060	.280	.047
		30	.349	.031	.387	.044	.333	.031
2	20	15	.390	.016	.420	.023	.378	.016
		30	.424	.017	.454	.024	.414	.020
3	10	15	.215	.042	.255	.053	.185	.040
		30	.214	.038	.241	.043	.189	.038
3	20	15	.316	.023	.355	.026	.290	.026
		30	.350	.017	.379	.024	.324	.019

Table 2: Analysis of variance of the error perturbed distance experiment.

Source of variation	SS	DF	MS	F	Sig of F
n	18.43	1	18.43	19800.73	.000
e	1.73	1	1.73	1859.89	.000
Method	2.55	2	1.28	1370.98	.000
q	5.30	2	2.65	2847.04	.000
p	12.09	1	12.09	12990.29	.000
Method by q	2.37	4	.59	637.44	.000
p by q	1.06	2	.53	571.52	.000
(Model)	43.53	13	3.35	3598.44	.000
Residual	6.69	7186	.00		
(Total)	50.22	7199	.01		

Table 3: Minimum values of Stress over 100 random starts of the same experiment reported in Table 1.

			$q = 1$			$q = 2$			$q = \infty$		
p	n	e%	smooth	SMACOF	KYST	smooth	SMACOF	KYST	smooth	SMACOF	KYST
2	10	15	.208	.209	.212	.235	.235	.235	.205	.223	.203
		30	.272	.297	.272	.222	.222	.222	.294	.308	.289
2	20	15	.335	.342	.337	.319	.319	.319	.357	.375	.346
		30	.391	.425	.401	.328	.328	.328	.389	.400	.374
3	10	15	.137	.161	.150	.159	.159	.159	.139	.146	.120
		30	.167	.173	.172	.205	.205	.205	.119	.153	.117
3	20	15	.251	.285	.271	.258	.258	.258	.269	.281	.246
		30	.268	.303	.298	.315	.315	.315	.313	.316	.281

Table 4: Stress values obtained by using classical scaling as start configuration.

			$q = 1$			$q = 2$			$q = \infty$		
p	n	$e\%$	smooth	SMACOF	KYST	smooth	SMACOF	KYST	smooth	SMACOF	KYST
2	10	15	.212	.209	.215	.235	.235	.235	.249	.238	.233
		30	.272	.287	.287	.222	.222	.222	.306	.301	.296
2	20	15	.335	.391	.387	.319	.319	.319	.368	.368	.356
		30	.402	.429	.424	.328	.331	.331	.389	.398	.380
3	10	15	.137	.155	.166	.159	.159	.159	.203	.200	.186
		30	.167	.189	.185	.205	.205	.205	.199	.224	.119
3	20	15	.256	.270	.264	.258	.258	.258	.284	.296	.265
		30	.275	.292	.284	.317	.317	.317	.311	.335	.292

Table 5: Best Stress values of 10 random starts of MDS on cola data of Green et al. (1989) in two dimensions.

	Distance		
q	Smoothing	SMACOF	KYST
1.00	.169437	.218761	.192295
1.33	.177194	.178874	.179519
1.66	.185316	.186848	.186276
2.00	.191782	.191782	.191978

5.3 Cola Data

The next data set concerns the cola data of Green, Carmone, and Smith (1989) (also used by Groenen et al. 1995), who reported preferences of 38 students for 10 varieties of cola. Every pair of colas was judged on their similarity on a nine point rating scale and the dissimilarities were accumulated over the subjects. Table 5 reports the lowest Stress values found by SMACOF, KYST, and distance smoothing using 10 random starts. In all cases distance smoothing found the best solution. The strategy of doing 10 random starts of distance smoothing seems to be sufficient to locate (candidate) global minima.

5.4 Similarity Among Ethnic Subgroups Data

Groenen and Heiser (1996) studied the occurrence of local minima in a data set of Funk, Horowitz, Lipshitz, and Young (1974) on perceived differences among thirteen ethnic subgroups of the American culture. The data consist of the average dissimilarity among 49 respondents who rated the difference between all pairs of ethnic subgroups on a nine-point rating scale (1 = very similar, 9 = very different). Using Euclidean distances, Groenen and Heiser (1996) found many different local minima using multistart with 1000 random starts. The lowest Stress value of 0.24546 occurred in 2.8% of the random starts. (Note that Groenen and Heiser (1996) reported squared Stress values.) Out of ten random starts, distance smoothing found the same minimum five times (50%). This example indicates that the region of attraction to the global minimum is greatly enlarged by using distance smoothing.

6 Discussion and Conclusion

In this paper we discussed how distance smoothing can be applied to MDS with Minkowski distances. The Huber function was proposed for smoothing the distances. The S-Stress loss function turned out to be a special case of the loss function used by distance smoothing. We have extended the majorization algorithm to minimize Stress with any Minkowski distance (with $q \geq 1$) by quadratic majorization. These results were used to develop a majorization algorithm for minimizing the distance smoothing loss function.

Numerical experiments on several data sets showed that distance smoothing is very well capable in locating a global minimum for small and moderate Minkowski distances, especially for city-block and Euclidean distances. For high q, notably for dominance distances, distance smoothing did not perform any better than a gradient method like KYST. For perfect distance data, distance smoothing almost always found the zero Stress solution for Minkowski parameter between 1 and 5, in contrast to two competitive methods, SMACOF and KYST, that often ended in nonzero local minima. Distance smoothing on perfect distance data in three dimensions did not always find the global minimum. For nonperfect data, distance smoothing outperforms SMACOF and KYST for city-block and Euclidean distances and either located the same candidate global minimum as SMACOF and KYST or found lower global minima. However, for dominance distance, distance smoothing did not perform better than KYST. Two examples using empirical data show that the strategy of ten random starts of distance smoothing gives a (very) high probability of finding the global minimum.

Distance smoothing for the dominance distance did not perform up to our expectations. Preliminary experimentation with an adaptation suggests that the performance could possibly be improved upon, but the results are too premature to justify inclusion in this paper.

Distance smoothing could be extended to incorporate constraints on the configuration in confirmatory MDS (see De Leeuw and Heiser 1980). This would allow to apply distance smoothing three-way extensions of MDS, such as individual differences models. It remains to be investigated under what constraints distance smoothing retains its good performance.

We conclude that distance smoothing works fine for unidimensional scaling and the two important cases of city-block and Euclidean MDS, but that it requires adjustments to deal with dominance distances. To find a good candidate global minimum, 10 random starts of distance smoothing should be enough unless the dominance distance is used.

References

BORG, I., and GROENEN, P. J. F. (1997), *Modern multidimensional scaling: Theory and applications*, New York: Springer.

DEFAYS, D. (1978), "A short note on a method of seriation," *British Journal of Mathematical and Statistical Psychology*, *3*, 49–53.

DE LEEUW, J. (1988), "Convergence of the majorization method for multidimensional scaling," *Journal of Classification*, *5*, 163–180.

DE LEEUW, J. (1993), "Fitting distances by least squares," Technical Report, No. 130, Los Angeles, California: Interdivisonal Program in Statistics, UCLA.

DE LEEUW, J. (1994), "Block relaxation algorithms in statistics," in *Information systems and data analysis*, Eds., H.-H. Bock, W. Lenski, and M. M. Richter, Berlin: Springer, 308–324.

DE LEEUW, J., and HEISER, W. J. (1977), "Convergence of correction-matrix algorithms for multidimensional scaling," in *Geometric representations of relational data*, Eds., J. C. Lingoes, E. E. Roskam, and I. Borg, Ann Arbor, MI: Mathesis Press, 735–752.

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, The Netherlands: North Holland Publishing Company, 501–522.

DE SOETE, G., HUBERT, L., and ARABIE, P. (1988), "On the use of simulated annealing for combinatorial data analysis," in *Data, expert knowledge and decisions*, Eds., W. Gaul and M. Schader, Berlin: Springer, 329–340.

FRANCIS, R. L., and WHITE, J. A. (1974), *Facility layout and location*, Englewood Cliffs, NJ: Prentice-Hall.

FUNK, S. G., HOROWITZ, A. D., LIPSHITZ, R., and YOUNG, F. W. (1974), "The perceived structure of american ethnic groups: The use of multidimensional scaling in stereotype research," *Personality and Social Psychology Bulletin*, *1*, 66–68.

GREEN, P. E., CARMONE, F. J. J., and SMITH, S. (1989), *Multidimensional scaling, concepts and applications*, Boston: Allyn and Bacon.

GROENEN, P. J. F. (1993), *The majorization approach to multidimensional scaling: Some problems and extensions*, Leiden, The Netherlands: DSWO Press.

GROENEN, P. J. F., and HEISER, W. J. (1996), "The tunneling method for global optimization in multidimensional scaling," *Psychometrika*, *61*, 529–550.

GROENEN, P. J. F., HEISER, W. J., and MEULMAN, J. J. (1997), "City-block scaling: Smoothing strategies for avoiding local minima," Technical Report, No. RR-97-01, Leiden: Department of Data Theory.

GROENEN, P. J. F., MATHAR, R., and HEISER, W. J. (1995), "The majorization approach to multidimensional scaling for Minkowski distances," *Journal of Classification*, *12*, 3–19.

HEISER, W. J. (1989), "The city-block model for three-way multidimensional scaling," in *Multiway data analysis*, Eds., R. Coppi and S. Bolasco, Amsterdam: Elsevier Science, 395–404.

HEISER, W. J. (1995), "Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis," in *Recent advances in descriptive multivariate analysis*, Ed., W. J. Krzanowski, Oxford: Oxford University Press, 157–189.

HUBER, P. J. (1981), *Robust statistics*, New York: Wiley.

HUBERT, L. J., and ARABIE, P. (1986), "Unidimensional scaling and combinatorial optimization," in *Multidimensional data analysis*, Eds., J. De Leeuw, W. J. Heiser, J. J. Meulman, and F. Critchley, Leiden, The Netherlands: DSWO-Press, 181–196.

HUBERT, L. J., ARABIE, P., and HESSON-MCINNIS, M. (1992), "Multidimensional scaling in the city-block metric: A combinatorial approach," *Journal of Classification*, *9*, 211–236.

HUBERT, L. J., and BUSK, P. (1976), "Normative location theory: Placement in continuous space," *Journal of Mathematical Psychology*, *14*, 187–210.

KIERS, H. A. L., and GROENEN, P. J. F. (1996), "A monotonicaly convergent algorithm for orthogonal congruence rotation," *Psychometrika*, *61*, 375–389.

KRUSKAL, J. B. (1964), "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, *29*, 115–129.

KRUSKAL, J. B., YOUNG, F. W., and SEERY, J. B. (1978), "How to use KYST-2, a very flexible program to do multidimensional scaling and unfolding," Technical Report, Murray Hill, NJ: Bell Labs.

MAGNUS, J. R., and NEUDECKER, H. (1988), *Matrix differential calculus with appilications in statistics and econometrics*, Chichester: Wiley.

MATHAR, R., and ŽILINSKAS, A. (1993), "On global optimization in two dimensional scaling," *Acta Aplicandae Mathematica*, *33*, 109–118.

PLINER, V. (1986), "The problem of multidimensional metric scaling," *Automation and Remote Control*, *47*, 560–567.

PLINER, V. (1996), "Metric, unidimensional scaling and global optimization," *Journal of Classification*, *13*, 3–18.

RAMSAY, J. O. (1969), "Some statistical considerations in multidimensional scaling," *Psychometrika*, *34*, 167–182.

TAKANE, Y., YOUNG, F. W., and DE LEEUW, J. (1977), "Nonmetric individual differences multidimensional scaling: An alternating least-squares method with optimal scaling features," *Psychometrika*, *42*, 7–67.

A Iterative Majorization

The main idea of iterative majorization is to operate on a simpler auxiliary function —the *majorizing function* $\mu(\mathbf{x},\mathbf{y})$— whose value is always larger than that of the original function, $\phi(\mathbf{x}) \leq \mu(\mathbf{x},\mathbf{y})$, but touches the original function at a *supporting point* $\mathbf{y}$, $\phi(\mathbf{y}) = \mu(\mathbf{y},\mathbf{y})$. Then the majorizing function is minimized, which often can be done in one step. The resulting configuration, $\mathbf{x}^+$, necessarily has a function value that is smaller than (or equal to) the function value at the supporting point, $\phi(\mathbf{x}^+) \leq \mu(\mathbf{x}^+,\mathbf{y})$. Therefore,

$$\phi(\mathbf{x}^+) \leq \mu(\mathbf{x}^+,\mathbf{y}) \leq \mu(\mathbf{y},\mathbf{y}) = \phi(\mathbf{y}). \tag{17}$$

This new configuration becomes the supporting point of the next majorizing function, and so on. We iterate over this process until convergence occurs due to a lower bound of the function or due to constraints. This brings us to one of the main advantages of majorization over many traditional minimization methods, which is that a converging sequence of function values is obtained without a stepsize procedure that may be computationally expensive and unreliable.

A useful property for constructing majorizing functions is: if $\mu_1(\mathbf{x},\mathbf{y})$ majorizes $\phi_1(\mathbf{x})$ and $\mu_2(\mathbf{x},\mathbf{y})$ majorizes $\phi_2(\mathbf{x})$ then $\phi_1(\mathbf{x})+\phi_2(\mathbf{x})$ is majorized by $\mu_1(\mathbf{x},\mathbf{y}) + \mu_2(\mathbf{x},\mathbf{y})$.

De Leeuw (1993) proposed to distinguish two sorts of majorization: (1) *linear* majorization of convex function, i.e., $\phi(\mathbf{x}) \leq \mu(\mathbf{x},\mathbf{y}) = \mathbf{x}'\mathbf{b}(\mathbf{y}) + c(\mathbf{y})$, and (2) *quadratic* majorization of a function with a bounded second derivative (Hessian), i.e., $\phi(\mathbf{x}) \leq \mu(\mathbf{x},\mathbf{y}) = \mathbf{x}'\mathbf{A}(\mathbf{y})\mathbf{x}-2\mathbf{x}'\mathbf{b}(\mathbf{y})+c(\mathbf{y})$. To obtain a linear majorizing function, we have to satisfy three conditions:

L1: $\phi(\mathbf{x})$ must be concave.

L2: ϕ has to touch μ at the supporting point $\mathbf{y}$, i.e., $\phi(\mathbf{y}) = \mu(\mathbf{y},\mathbf{y}) = \mathbf{y}'\mathbf{b}(\mathbf{y}) + c(\mathbf{y})$.

L3: ϕ and μ have the same tangent at $\mathbf{y}$ if the gradient $\nabla\phi(\mathbf{x}) = \partial\phi(\mathbf{x})/\partial\mathbf{x}$ exists at $\mathbf{y}$, i.e., $\nabla\phi(\mathbf{x}) = \nabla\mu(\mathbf{x},\mathbf{y}) = \mathbf{b}(\mathbf{y})$.

Note that concavity of $\phi(\mathbf{x})$ ensures that the linear function $\mu(\mathbf{x},\mathbf{y})$ is always larger than (or equal to) $\phi(\mathbf{x})$. These three requirements are fulfilled by choosing $\mathbf{b}(\mathbf{y}) = \nabla\phi(\mathbf{x})$ and $c(\mathbf{y}) = \phi(\mathbf{y}) - \mathbf{y}'\mathbf{b}(\mathbf{y})$.

For quadratic majorization, it is assumed that $\phi(\mathbf{x})$ is twice differentiable over its domain. To obtain a quadratic majorizing function, we have to satisfy the conditions:

Q1: The matrix of the second derivatives of $\phi(\mathbf{x})$, $\nabla^2\phi(\mathbf{x})$, must be bounded, i.e., $\mathbf{x}'\nabla^2\phi(\mathbf{x})\mathbf{x} \leq \frac{1}{2}\mathbf{x}'\mathbf{A}(\mathbf{y})\mathbf{x}$ for all $\mathbf{x}$.

Q2: ϕ has to touch μ at the supporting point $\mathbf{y}$, i.e., $\phi(\mathbf{y}) = \mu(\mathbf{y},\mathbf{y}) = \mathbf{y}'\mathbf{A}(\mathbf{y})\mathbf{y} - \mathbf{y}'\mathbf{b}(\mathbf{y}) + c(\mathbf{y})$.

Q3: ϕ and μ have the same tangent at $\mathbf{y}$, i.e., $\nabla\phi(\mathbf{x}) = \nabla\mu(\mathbf{x},\mathbf{y}) = 2\mathbf{A}(\mathbf{y})\mathbf{y} - 2\mathbf{b}(\mathbf{y})$.

Given an $\mathbf{A}(\mathbf{y})$ that satisfies condition Q1, the other two conditions are satisfied by choosing $\mathbf{b}(\mathbf{y}) = \mathbf{A}(\mathbf{y})\mathbf{y} - \frac{1}{2}\nabla\phi(\mathbf{y})$ and $c(\mathbf{y}) = \phi(\mathbf{y}) + \mathbf{y}'\mathbf{A}(\mathbf{y})\mathbf{y} - \mathbf{y}'\nabla\phi(\mathbf{y})$. It can be hard to find a matrix $\mathbf{A}(\mathbf{y})$ that satisfies condition Q1. However, in the algorithmic sections we provide some explicit strategies for finding $\mathbf{A}(\mathbf{y})$.

B A Quadratic Majorization Algorithm for MDS with Minkowski Distances

Here we combine the majorizing functions for $d_{ij}^2(\mathbf{X})$ and $d_{ij}^2(\mathbf{X})$ from Section 3 to obtain a majorizing function for Stress.

Multiply (11), (12), and (13) for majorizing $d_{ij}^2(\mathbf{X})$ by w_{ij}, multiply (10) by $2w_{ij}\delta_{ij}$, and sum over all i,j. This operation gives

$$\begin{aligned}
\sigma^2(\mathbf{X}) &\leq \mu(\mathbf{X};\mathbf{Y}) = \eta_\delta^2 + \sum_s \sum_{i<j} |a_{ijs}|(x_{is} - x_{js})^2 \\
&\quad -2\sum_s \sum_{i<j} |b_{ijs}|(x_{is} - x_{js})(y_{is} - y_{js}) + c \\
&= \eta_\delta^2 + \sum_s \mathbf{x}_s'\mathbf{A}_s(\mathbf{Y})\mathbf{x}_s - 2\sum_s \mathbf{x}_s'\mathbf{B}_s(\mathbf{Y})\mathbf{y}_s + c(\mathbf{Y}), \qquad (18)
\end{aligned}$$

where $\mathbf{x}_s$ denotes column s of $\mathbf{X}$, $\mathbf{A}_s(\mathbf{Y})$ has elements

$$a_{ijs} = \begin{cases} -w_{ij}a_{ijs}^{(1\leq q\leq 2)} & \text{if } i \neq j \text{ and } 1 \leq q \leq 2, \\ -w_{ij}a^{(q>2)} & \text{if } i \neq j \text{ and } q > 2, \\ -w_{ij}a_{ij}^{(q=\infty)} & \text{if } i \neq j \text{ and } q = \infty, \\ -\sum_{j\neq i} a_{ijs} & \text{if } i = j, \end{cases}$$

$\mathbf{B}_s(\mathbf{Y})$ has elements

$$b_{ijs} = \begin{cases} -w_{ij}\delta_{ij}b_{ijs}^{(1)} & \text{if } i \neq j \text{ and } 1 \leq q \leq 2, \\ -w_{ij}[\delta_{ij}b_{ijs}^{(1)} + b_{ijs}^{(q>2)}] & \text{if } i \neq j \text{ and } q > 2, \\ -w_{ij}[\delta_{ij}b_{ijs}^{(1)} + b_{ijs}^{(q=\infty)}] & \text{if } i \neq j \text{ and } q = \infty, \\ -\sum_{j\neq i} b_{ijs} & \text{if } i = j, \end{cases}$$

and $c(\mathbf{Y})$ is defined as

$$c(\mathbf{Y}) \;=\; \begin{cases} 0 & \text{if } i \neq j \text{ and } 1 \leq q \leq 2, \\ \sum_{i<j} w_{ij} c_{ij}^{(q>2)} & \text{if } i \neq j \text{ and } q > 2, \\ \sum_{i<j} w_{ij} c_{ij}^{(q=\infty)} & \text{if } i \neq j \text{ and } q = \infty. \end{cases}$$

Since the right-hand-part of (18) majorizes (1), we have equality if $\mathbf{Y} = \mathbf{X}$.

The minimum of $\mu(\mathbf{X};\mathbf{Y})$ update is obtained by setting the gradient of the majorizing function $\mu(\mathbf{X};\mathbf{Y})$ equal to zero and solve for $\mathbf{x}_s$, i.e.,

$$\mathbf{x}_s = \mathbf{A}_s(\mathbf{Y})^{-}\mathbf{B}_s(\mathbf{Y})\mathbf{y}_s \tag{19}$$

for $s = 1, \ldots, p$, where $\mathbf{A}_s(\mathbf{Y})^{-}$ is any generalized inverse of $\mathbf{A}_s(\mathbf{Y})$. A convenient generalized inverse for $\mathbf{A}_s(\mathbf{Y})$ is the Moore-Penrose inverse, which is, in this case, equal to $(\mathbf{A}_s(\mathbf{Y}) + \mathbf{1}\mathbf{1}')^{-1} - n^{-2}\mathbf{1}\mathbf{1}'$ with $\mathbf{1}$ an n vector of ones.

The majorization algorithm can be summarized as

1. $\mathbf{Y} \leftarrow \mathbf{Y}_0$.
2. Find $\mathbf{X}^{+}$ by (19) for which $\mu(\mathbf{X}^{+}, \mathbf{Y}) = \min_{\mathbf{X}} \mu(\mathbf{X}, \mathbf{Y})$.
3. If $\sigma^2(\mathbf{Y}) - \sigma^2(\mathbf{X}^{+}) < \varepsilon$ then stop. (ε a small positive constant.)
4. $\mathbf{Y} \leftarrow \mathbf{X}^{+}$ and go to 2.

Note that the convergence results in Groenen et al. (1995) still hold. De Leeuw and Heiser (1980) and Heiser (1995) have shown that using the update $2\mathbf{X}^{+} - \mathbf{Y}$, the so-called relaxed update, in step 2 may half the number of iterations without destroying convergence.

C A Majorizing Algorithm for Distance Smoothing

To find a majorizing function for $\sigma_\epsilon^2(\mathbf{X})$, we use the results from the previous section, where $|y_{is} - y_{js}|$ is substituted $h_\epsilon(y_{is} - y_{js})$ and $d_{ij}(\mathbf{Y})$ by $d_{ij}(\mathbf{Y}|\epsilon)$. Then, $(x_{is} - x_{js})^2$ is substituted by $h_\epsilon^2(y_{is} - y_{js})$, which is majorized by (15), and $-(x_{is} - x_{js})$ is substituted by $-h_\epsilon(y_{is} - y_{js})$, which is majorized by (14). This yields

$$\begin{aligned} \sigma_\epsilon^2(\mathbf{X}) \;&\leq\; \mu_\epsilon(\mathbf{X};\mathbf{Y}) = \eta_\delta^2 + \sum_s \sum_{i<j} |a_{ijs}^{(\epsilon)}|(x_{is} - x_{js})^2 \\ &\qquad -2\sum_s \sum_{i<j} |b_{ijs}^{(\epsilon)}|(x_{is} - x_{js})(y_{is} - y_{js}) + c^{(\epsilon)} \\ &= \eta_\delta^2 + \sum_s \mathbf{x}_s'\mathbf{A}_s^{(\epsilon)}(\mathbf{Y})\mathbf{x}_s - 2\sum_s \mathbf{x}_s'\mathbf{B}_s^{(\epsilon)}(\mathbf{Y})\mathbf{y}_s + c^{(\epsilon)}(\mathbf{Y}). \end{aligned} \tag{20}$$

The matrix $\mathbf{A}_s^{(\epsilon)}(\mathbf{Y})$ has elements

$$a_{ijs}^{(\epsilon)} = \begin{cases} -w_{ij}a^{(h^2)}a_{ijs}^{(1\leq q\leq 2|\epsilon)} & \text{if } i \neq j \text{ and } 1 \leq q \leq 2, \\ -w_{ij}a^{(h^2)}a^{(q>2|\epsilon)} & \text{if } i \neq j \text{ and } q > 2, \\ -w_{ij}a^{(h^2)}a_{ij}^{(q=\infty|\epsilon)} & \text{if } i \neq j \text{ and } q = \infty, \\ -\sum_{j\neq i} a_{ijs}^{(\epsilon)} & \text{if } i = j, \end{cases}$$

where $a_{ijs}^{(1\leq q\leq 2|\epsilon)}$ equals $h_\epsilon^{q-2}(y_{is}-y_{js})/d_{ij}^{q-2}(\mathbf{Y}|\epsilon)$, $a^{(q>2|\epsilon)}$ equals λ, and $a_{ij}^{(q=\infty|\epsilon)}$ equals

$$\begin{cases} \dfrac{h_\epsilon(y_{i\phi_1} - y_{j\phi_1})}{h_\epsilon(y_{i\phi_1} - y_{j\phi_1}) - h_\epsilon(y_{i\phi_2} - y_{j\phi_2})} & \text{if } h_\epsilon(y_{i\phi_1} - y_{j\phi_1}) - h_\epsilon(y_{i\phi_2} - y_{j\phi_2}) > \varepsilon, \\ \dfrac{h_\epsilon(y_{i\phi_1} - y_{j\phi_1}) + \varepsilon}{\varepsilon} & \text{if } h_\epsilon(y_{i\phi_1} - y_{j\phi_1}) - h_\epsilon(y_{i\phi_2} - y_{j\phi_2}) \leq \varepsilon, \end{cases}$$

with ϕ_s ordering $h_\epsilon(y_{is} - y_{js})$ decreasingly over $s = 1, \ldots, p$. The matrix $\mathbf{B}_s^{(\epsilon)}(\mathbf{Y})$ has elements $b_{ijs}^{(\epsilon)}$ equal to

$$\begin{cases} -w_{ij}[\delta_{ij}b_{ijs}^{(-h)}b_{ijs}^{(1|\epsilon)} + a_{ijs}^{(1\leq q\leq 2|\epsilon)}b_{ijs}^{(h^2)}] & \text{if } i \neq j \text{ and } 1 \leq q \leq 2, \\ -w_{ij}[\delta_{ij}b_{ijs}^{(-h)}b_{ijs}^{(1|\epsilon)} + b_{ijs}^{(-h)}b_{ijs}^{(q>2|\epsilon)} + a^{(q>2|\epsilon)}b_{ijs}^{(h^2)}] & \text{if } i \neq j \text{ and } q > 2, \\ -w_{ij}[\delta_{ij}b_{ijs}^{(-h)}b_{ijs}^{(1|\epsilon)} + b_{ijs}^{(-h)}b_{ijs}^{(q=\infty|\epsilon)} + a_{ij}^{(q=\infty|\epsilon)}b_{ijs}^{(h^2)}] & \text{if } i \neq j \text{ and } q = \infty, \\ -\sum_{j\neq i} b_{ijs}^{(\epsilon)} & \text{if } i = j, \end{cases}$$

where

$$b_{ijs}^{(1|\epsilon)} = \begin{cases} \dfrac{h_\epsilon^{q-2}(y_{is} - y_{js})}{d_{ij}^{q-1}(\mathbf{Y}|\epsilon)} & \text{if } 1 \leq q < \infty, \\ h_\epsilon^{-1}(y_{is} - y_{js}) & \text{if } q = \infty \text{ and } s = \phi_1, \\ 0 & \text{if } q = \infty \text{ and } s \neq \phi_1, \end{cases}$$

$$b_{ijs}^{(q>2|\epsilon)} = a^{(q>2|\epsilon)} - \frac{h_\epsilon^{q-2}(y_{is} - y_{js})}{d_{ij}^{q-2}(\mathbf{Y}|\epsilon)},$$

$$b_{ijs}^{(q=\infty|\epsilon)} = \begin{cases} a_{ij}^{(q=\infty|\epsilon)} & \text{if } s \neq \phi_1, \\ a_{ij}^{(q=\infty|\epsilon)}\dfrac{h_\epsilon(y_{i\phi_2} - y_{j\phi_2})}{h_\epsilon(y_{i\phi_1} - y_{j\phi_1})} & \text{if } s = \phi_1. \end{cases}$$

For $1 \leq q \leq 2$, the constant $c(\mathbf{Y})$ is defined by

$$\sum_{s,i<j} w_{ij}[a_{ijs}^{(1\leq q\leq 2|\epsilon)}c_{ijs}^{(h^2)} + \delta_{ij}b_{ijs}^{(1|\epsilon)}c_{ijs}^{(-h)}],$$

for $2 < q < \infty$ by

$$\sum_{i<j} w_{ij}c_{ij}^{(q>2|\epsilon)} + \sum_{s,i<j} w_{ij}[a^{(q>2|\epsilon)}c_{ijs}^{(h^2)} + b_{ijs}^{(q>2|\epsilon)}c_{ijs}^{(-h)} + \delta_{ij}b_{ijs}^{(1|\epsilon)}c_{ijs}^{(-h)}],$$

for $q = \infty$ by

$$\sum_{i<j} w_{ij}[c_{ij}^{(q=\infty)} + \sum_{s,i<j} w_{ij}[a_{ij}^{(q=\infty|\epsilon)} c_{ijs}^{(h^2)} + b_{ijs}^{(q=\infty|\epsilon)} c_{ijs}^{(-h)} + \delta_{ij} b_{ijs}^{(1|\epsilon)} c_{ijs}^{(-h)}],$$

where

$$\begin{aligned} c_{ij}^{(q>2|\epsilon)} &= a^{(q>2|\epsilon)} \sum_s h_\epsilon^2(y_{is} - y_{js}) - d_{ij}^2(\mathbf{Y}|\epsilon), \\ c_{ij}^{(q=\infty|\epsilon)} &= \sum_s (2b^{(q=\infty|\epsilon)} - a^{(q=\infty|\epsilon)}) h_\epsilon^2(y_{is} - y_{js}) + d_{ij}^2(\mathbf{Y}|\epsilon). \end{aligned}$$

Since the right hand part of $\mu_\epsilon(\mathbf{X}; \mathbf{Y})$ majorizes $\sigma_\epsilon^2(\mathbf{X})$, we have equality if $\mathbf{Y} = \mathbf{X}$.

The minimum of $\mu_\epsilon(\mathbf{X}; \mathbf{Y})$ is obtained by setting its gradient equal to zero and solve for $\mathbf{x}_s$, i.e.,

$$\mathbf{x}_s = \mathbf{A}_s^{(\epsilon)}(\mathbf{Y})^{-} \mathbf{B}_s^{(\epsilon)}(\mathbf{Y}) \mathbf{y}_s \text{ for } s = 1, \ldots, p. \tag{21}$$