The Majorization Approach to Multidimensional Scaling for Minkowski Distances

Patrick J.F. Groenen*, Rudolf Mathar** and Willem J. Heiser*

*Department of Data Theory, University of Leiden

P.O. Box 9555, 2300 RB Leiden, The Netherlands

e-mail: groenen@rulfsw.LeidenUniv.nl

e-mail: heiser@rulfsw.LeidenUniv.nl

**Institute of Statistics, Aachen University of Technology

Wüllnerstraße 3, D-5100 Aachen, Germany

e-mail: mathar@stochastik.rwth-aachen.de

Abstract

The majorization method for multidimensional scaling with Kruskal's STRESS was limited to Euclidean distances only. Here we extent the majorization algorithm to deal with Minkowski distances with $1 \leq p \leq 2$ and suggest an algorithm that is partially based on majorization for p outside this range. We give some convergence proofs and extent the zero distance theorem of De Leeuw (1984) to Minkowski distances with $p \geq 1$.

Keywords: multidimensional scaling, distance analysis, majorization, Minkowski distances.

1 Introduction

In least squares multidimensional scaling the objective is to represent non-negative dissimilarities between objects in a plot of low dimensionality where the interpoint distance should match the dissimilarities as closely as possible. Most frequently the Euclidean distance is used, but this need not be so. An important family of distance measures is formed by the Minkowski distances of which the Euclidean one is a special case. A relatively simple algorithm for least squares scaling using Euclidean distances is SMACOF of De Leeuw (1977) and De Leeuw and Heiser (1977), which generalizes Guttman's (1968) C-matrix method. An attractive feature of SMACOF is that it produces a monotone decreasing sequence of values of the loss function by using the concept of iterative majorization.

De Leeuw (1977) treats multidimensional scaling in the framework of convex analysis. He for the first time gave a convergence proof for Euclidean distances, and discussed extensions for general p, without giving an explicit algorithm. Using a similar approach,

Mathar and Groenen (1991) derive uniqueness results which lead to a convergence proof for general p. They also give interpretations in terms of directional derivatives and subgradient–projection methods.

Considering STRESS as a DC–function (difference of convex functions) Mathar and Meyer (1992) obtain subgradients for arbitrary p in quasi–quadratic form. This leads to an eigenvector problem which is solved by inverse iteration. Interestingly enough, in the Euclidean case this reduces to SMACOF again.

Here we extend the majorization approach to least squares scaling with Minkowski distances for $1 \leq p \leq 2$. For p outside this range algorithms that are partially based on majorization are developed. Kruskal's (1964a, 1964b) method also covers Minkowski distances, but works with a complicated step-size procedure, which makes convergence uncertain.

The STRESS function to be minimized over all configurations $\mathbf{X}$ may be written as

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

Here $\mathbf{X} = (x_{is})_{1\leq i\leq n, 1\leq s\leq k}$ is an $n \times k$ matrix with coordinates of n objects in k dimensions, the quantities w_{ij} are fixed non–negative weights, δ_{ij} are non–negative dissimilarities and

$$d_{ij}(\mathbf{X}) = \Big(\sum_{s=1}^{k} |x_{is} - x_{js}|^p\Big)^{\frac{1}{p}}, \quad 1 \leq p \leq \infty \tag{2}$$

denotes the Minkowski distance. We assume that the weight matrix is irreducible, i.e. there exists no partitioning of objects in disjoint subsets, such that $w_{ij} = 0$ whenever objects i and j are in different subsets. If the weight matrix is reducible then the problem can be decomposed in separate smaller multidimensional scaling problems, one for each subset. Some well known metrics and corresponding norms are obtained by proper choices of p, like the city-block metric (or Manhattan metric) for $p = 1$, the Euclidean one for $p = 2$ and the max norm for $p = \infty$. For a recent review article on Minkowski distances in multidimensional scaling we refer to Arabie (1991).

Our main objective is to minimize STRESS by the principle of majorization (see e.g. De Leeuw, 1988). Let us rewrite (1) as

$$\begin{aligned} \sigma^2(\mathbf{X}) \quad &= \quad \sum_{i<j}^{n} w_{ij}\delta_{ij}^2 + \sum_{i<j}^{n} w_{ij}d_{ij}^2(\mathbf{X}) - 2\sum_{i<j}^{n} w_{ij}\delta_{ij}d_{ij}(\mathbf{X}) \\ &= \quad \eta_\delta^2 + \eta^2(\mathbf{X}) - 2\rho(\mathbf{X}). \end{aligned} \tag{3}$$

We shall see that for $1 \leq p \leq 2$ a convergent algorithm for minimizing (3) can be obtained by using majorization. For $p < 1$ or $p > 2$ a convergent majorization algorithm is obtained by using an inner minimization step of a convex function. However, in the case of uni-dimensional scaling we run into a local minimum within a few steps. Here, more powerful combinatorial optimization methods are available, which are extensively discussed by Hubert and Arabie (1986).

The next section discusses the principle of minimizing a function by majorization. We indicate how $-\rho(\mathbf{X})$ and $\eta^2(\mathbf{X})$ can be majorized. A convergent algorithm for $1 \leq p \leq 2$ is derived and some convergence properties are given. Then we discuss two algorithms for p values outside this range. Next we show differentiability of STRESS at a local minimum for $p \geq 1$. Finally, as an illustration of our procedure we present a small example.

2 Majorizing STRESS

Iterative majorization is a simple method for minimizing a function. It has been used in the field of multidimensional scaling among others by De Leeuw and Heiser (1980), Meulman (1986), De Leeuw (1988), Heiser (1988,1991). The main idea of iterative majorization is to operate on a simpler auxiliary function —the majorizing function— that is always larger than the original function, but touches the original function at a supporting point. Then the majorizing function is minimized, which can often be done in one step. The resulting configuration necessarily has a function value that is smaller than (or at most equal to) the function value of the supporting point. 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.

We describe the principle more formally. Let $\varphi(\mathbf{X})$ be a real valued function to be minimized over its domain $\mathcal{X}$. A function $\hat{\varphi}(\mathbf{X},\mathbf{Y})$, $\mathbf{X},\mathbf{Y}\in\mathcal{X}$, with the properties

$$\varphi(\mathbf{X})\leq\hat{\varphi}(\mathbf{X},\mathbf{Y})\quad\text{and}\quad\varphi(\mathbf{X})=\hat{\varphi}(\mathbf{X},\mathbf{X})\quad\text{for all }\mathbf{X},\mathbf{Y}\in\mathcal{X}$$

is called a majorizing function. Now, for fixed $\mathbf{Y}$ let $\mathbf{X}^+=\arg\min_{\mathbf{X}\in\mathcal{X}}\hat{\varphi}(\mathbf{X},\mathbf{Y})$ denote a minimum point. Then immediately we have the following chain of inequalities,

$$\varphi(\mathbf{X}^+)\leq\hat{\varphi}(\mathbf{X}^+,\mathbf{Y})\leq\hat{\varphi}(\mathbf{Y},\mathbf{Y})=\varphi(\mathbf{Y}),\tag{4}$$

which is named the *sandwich* inequality by De Leeuw (1992). Equality of (4) only occurs if $\mathbf{X}^+$ is also a stationary point of $\varphi(\mathbf{X})$. The majorization algorithm can be summarized as

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

Obviously, by (4) we obtain a decreasing sequence of function values. If the function is bounded from below this sequence converges. A necessary condition for a point $\mathbf{X}^*$ to be a minimizer of φ is that $\mathbf{X}^*$ minimizes $\hat{\varphi}(\cdot,\mathbf{X}^*)$ over $\mathcal{X}$. So, if $\varphi(\mathbf{X}^+)=\varphi(\mathbf{Y})$, this stationary condition is satisfied by $\mathbf{Y}$.

Other information has to be used to check if the point is a global minimum, a local minimum or even a saddle point. For finding the global minimum we could use the tunneling method of Groenen and Heiser (1991) that is also based heavily on majorization. Note that in general we cannot say much about the convergence behavior of the sequence of $\mathbf{X}^+$.

We shall now propose majorizing functions for the separate parts of $\sigma^2(\mathbf{X})$ in (3).

2.1 Majorization of $-\rho(\mathbf{X})$

The crossproduct term $-\rho(\mathbf{X})$ is majorized by applying Hölder's inequality to $-d_{ij}(\mathbf{X})$ for $p\geq 1$. If the denominator is positive it holds for any $\mathbf{X}$ and $\mathbf{Y}$ that

$$-d_{ij}(\mathbf{X})=-\Big(\sum_{s=1}^{k}|x_{is}-x_{js}|^p\Big)^{\frac{1}{p}}\quad\leq\quad-\frac{\sum_{s=1}^{k}|x_{is}-x_{js}||y_{is}-y_{js}|^{(p-1)}}{\Big(\sum_{s=1}^{k}|y_{is}-y_{js}|^p\Big)^{\frac{p-1}{p}}}$$

$$\leq -\frac{\sum_{s=1}^{k}(x_{is}-x_{js})|y_{is}-y_{js}|^{(p-1)}}{\left(\sum_{s=1}^{k}|y_{is}-y_{js}|^{p}\right)^{\frac{p-1}{p}}} \tag{5}$$

where the first inequality becomes an equality if $x_{is} = y_{is}$ for all i and s. If $d_{ij}(\mathbf{Y}) = 0$ we simply define the right hand side as 0 which preserves the validity of inequality (5). We furthermore use the conventions $0^0 = 1$ and $0 \cdot * = 0$. By using $-|y_{is}-y_{js}|^{(p-1)} \leq (y_{is}-y_{js})|y_{is}-y_{js}|^{(p-2)}$, multiplying both sides of (5) by $w_{ij}\delta_{ij}$ and summing over all i, j we obtain the following inequality.

$$-\rho(\mathbf{X}) \leq -\sum_{s=1}^{k}\mathbf{x}_s'\mathbf{B}_s(\mathbf{Y})\mathbf{y}_s = -\hat{\rho}(\mathbf{X},\mathbf{Y}), \tag{6}$$

where $\mathbf{x}_s$ and $\mathbf{y}_s$ denote the columns of $\mathbf{X}$ and $\mathbf{Y}$, respectively. $\mathbf{B}_s(\mathbf{Y})$ has off-diagonal elements

$$b_{ij}^{(s)} = -\frac{w_{ij}\delta_{ij}|y_{is}-y_{js}|^{(p-2)}}{d_{ij}^{p-1}(\mathbf{Y})}, \quad i \neq j, \tag{7}$$

if $d_{ij}(\mathbf{Y}) > 0$ and $b_{ij}^{(s)} = 0$ otherwise, and diagonal elements $b_{ii}^{(s)} = -\sum_{j\neq i} b_{ij}^{(s)}$. Obviously $\rho(\mathbf{X}) = \hat{\rho}(\mathbf{X},\mathbf{X})$ holds for all $\mathbf{X}$.

2.2 Majorization of $\eta^2(\mathbf{X})$

Let us look at $\eta^2(\mathbf{X})$ and more specifically at its elements $d_{ij}^2(\mathbf{X})$. This is the square of the Minkowski distance which can be rewritten as

$$d_{ij}^2(\mathbf{X}) = \left(\sum_{s=1}^{k}|x_{is}-x_{js}|^{p}\right)^{\frac{2}{p}} = \left(\sum_{s=1}^{k}|x_{is}-x_{js}|^{2r}\right)^{\frac{1}{r}} \tag{8}$$

for $r = p/2$. Using Hölders inequality for $r \leq 1$, thus $p \leq 2$, gives

$$\left(\sum_{s=1}^{k}|x_{is}-x_{js}|^{2r}\right)^{\frac{1}{r}} \leq \frac{\sum_{s=1}^{k}(x_{is}-x_{js})^2|y_{is}-y_{js}|^{2(r-1)}}{\left(\sum_{s=1}^{k}|y_{is}-y_{js}|^{2r}\right)^{\frac{r-1}{r}}} \tag{9}$$

for all positive $|y_{is}-y_{js}|$ with equality if $x_{is} = y_{is}$ for all i and s. Since $\mathbf{x}'\mathbf{A}\mathbf{y} = \sum_{i<j} a_{ij}(x_i-x_j)(y_i-y_j)$ holds for any symmetric matrix $\mathbf{A}$ having off–diagonal elements $-a_{ij}$ and diagonal elements $\sum_{j\neq i} a_{ij}$, after multiplication of both sides of (9) by w_{ij} we get the following expression

$$\eta^2(\mathbf{X}) \leq \sum_{s=1}^{k}\mathbf{x}_s'\mathbf{A}_s(\mathbf{Y})\mathbf{x}_s = \hat{\eta}^2(\mathbf{X},\mathbf{Y}), \tag{10}$$

where $\mathbf{x}_s$ is the s^{th} column vector of $\mathbf{X}$ and $\mathbf{A}_s(\mathbf{Y})$ has off-diagonal elements

$$a_{ij}^{(s)} = -\frac{w_{ij}|y_{is}-y_{js}|^{p-2}}{d_{ij}^{p-2}(\mathbf{Y})}, \quad i \neq j, \tag{11}$$

and diagonal elements $a_{ii}^{(s)} = -\sum_{j \neq i} a_{ij}^{(s)}$. Using the same conventions as above we obtain $\eta^2(\mathbf{X}) = \hat{\eta}^2(\mathbf{X}, \mathbf{X})$ for any configuration $\mathbf{X}$.

If $y_{is} - y_{js} = 0$ for some i, j, s inequality (9) may not be true. In that case we simply replace $(y_{is} - y_{js})^2$ by some small positive constant ϵ, in much the same way as Heiser (1991) treats this when dealing with negative dissimilarities. The majorizing function remains larger than $\eta^2(\mathbf{X})$, but does not touch $\eta^2(\mathbf{X})$ for $\mathbf{X} = \mathbf{Y}$, although we may get arbitrarily close by letting ϵ approach to zero.

The majorization inequality (10) holds for any $p \leq 2$ and the majorizing function $\hat{\eta}^2(\mathbf{X}, \mathbf{Y})$ is a quadratic function in $\mathbf{X}$. For $p > 2$ all inequalities in this section have reversed sign.

3 The majorization algorithm for $1 \leq p \leq 2$

The two majorization inequalities can be used simultaneously for $1 \leq p \leq 2$. It results in an algorithm that produces a monotone decreasing series of function values and hence is convergent in this sense. Note that at best a local minimum is reached.

The majorization algorithm is derived from the stationary equation of the majorizing function, which is quadratic in $\mathbf{X}$. Thus we have

$$\begin{aligned} \sigma^2(\mathbf{X}) &\leq \eta_\delta^2 + \eta^2(\mathbf{X}, \mathbf{Y}) - 2\rho(\mathbf{X}, \mathbf{Y}) \\ &= \eta_\delta^2 + \sum_{s=1}^{k} \mathbf{x}_s' \mathbf{A}_s(\mathbf{Y}) \mathbf{x}_s - 2 \sum_{s=1}^{k} \mathbf{x}_s' \mathbf{B}_s(\mathbf{Y}) \mathbf{y}_s. \end{aligned} \tag{12}$$

Setting the gradient of the majorizing function (12) equal to zero implies for all s

$$\mathbf{x}_s = \mathbf{A}_s(\mathbf{Y})^- \mathbf{B}_s(\mathbf{Y}) \mathbf{y}_s, \tag{13}$$

where $\mathbf{A}_s(\mathbf{Y})^-$ is any generalized inverse of $\mathbf{A}_s(\mathbf{Y})$. The update may be computed simultaneously or dimensionwise. It can be shown that for $p = 2$ the proposed algorithm with update (13) simply reduces to the SMACOF algorithm of referred to earlier.

Fortunately, almost all convergence theorems known from SMACOF still hold. Here we follow the proofs of De Leeuw (1988). The notation is simplified whenever it can be done without introducing ambiguity; at iteration m we write $\sigma_m^2 = \sigma^2(\mathbf{X}^m)$, $\rho_m = \rho(\mathbf{X}^m)$, $\eta_m^2 = \eta^2(\mathbf{X}^m)$, $\mathbf{B}_s^m = \mathbf{B}_s(\mathbf{X}^m)$, $\mathbf{A}_s^m = \mathbf{A}_s(\mathbf{X}^m)$. For convenience, we assume without loss of generality $\eta_\delta^2 = 1$. Observe that by applying the above ϵ–procedure, all $\mathbf{A}_s^m$ may be assumed to have rank $n - 1$ (cf. Mathar and Meyer, 1992)).

Since STRESS is the sum of squared differences, we may use the Cauchy–Schwarz inequality to obtain $\rho_m \leq \eta_m$. Because of $\mathbf{B}_s^m = \mathbf{A}_s^m \mathbf{A}_s^{m-} \mathbf{B}_s^m$, by Cauchy–Schwarz we get $\sum_s \mathbf{x}_s^{m\prime} \mathbf{B}_s^m \mathbf{x}_s^m \leq (\sum_s \mathbf{x}_s^{m\prime} \mathbf{A}_s^m \mathbf{x}_s^m)^{1/2} (\sum_s \mathbf{x}_s^{m+1\prime} \mathbf{A}_s^m \mathbf{x}_s^{m+1})^{1/2}$, or

$$\rho(\mathbf{X}^m) \leq \eta(\mathbf{X}^m) \hat{\eta}(\mathbf{X}^{m+1}, \mathbf{X}^m). \tag{14}$$

Furthermore, we have

$$\eta^2(\mathbf{X}^{m+1}) \leq \hat{\eta}^2(\mathbf{X}^{m+1}, \mathbf{X}^m) = \sum_s \mathbf{x}_s^{m\prime} \mathbf{B}_s^m \mathbf{A}_s^{m-} \mathbf{B}_s^m \mathbf{x}_s^m = \hat{\rho}(\mathbf{X}^{m+1}, \mathbf{X}^m) \leq \rho(\mathbf{X}^{m+1}), \tag{15}$$

which holds because of the majorization inequalities. Combining (14) and (15) gives

$$\eta(\mathbf{X}^m) \leq \frac{\rho(\mathbf{X}^m)}{\eta(\mathbf{X}^m)} \leq \hat{\eta}(\mathbf{X}^{m+1}, \mathbf{X}^m). \tag{16}$$

These inequalities let us form the following chain:

$$\eta_m^2 \leq \rho_m \leq \eta_m \hat{\eta}(\mathbf{X}^{m+1}, \mathbf{X}^m) \leq \hat{\eta}^2(\mathbf{X}^{m+1}, \mathbf{X}^m) = \hat{\rho}(\mathbf{X}^{m+1}, \mathbf{X}^m) \leq \rho_{m+1} \leq \eta_{m+1} \leq 1. \quad (17)$$

Further, define the measure of difference in squared distances as

$$\begin{aligned} \varepsilon_m^2 &= \hat{\eta}^2(\mathbf{X}^m - \mathbf{X}^{m+1}, \mathbf{X}^m) = \sum_s (\mathbf{x}_s^m - \mathbf{A}_s^{m-} \mathbf{B}_s^m \mathbf{x}_s^m)' \mathbf{A}_s^m (\mathbf{x}_s^m - \mathbf{A}_s^{m-} \mathbf{B}_s^m \mathbf{x}_s^m) \\ &= \eta^2(\mathbf{X}^m) + \hat{\eta}^2(\mathbf{X}^{m+1}, \mathbf{X}^m) - 2\rho(\mathbf{X}^m). \end{aligned} \quad (18)$$

These inequalities lead to the following observations:

1. $\rho_m \uparrow \rho_\infty$,
2. $\eta_m^2 \uparrow \eta_\infty^2 = \rho_\infty$,
3. $\sigma_m^2 \downarrow \sigma_\infty^2 = 1 - \rho_\infty$,
4. $\varepsilon_m^2 \rightarrow 0$.

The last assertion is proved by filling in the limiting values of η_m^2, $\hat{\eta}^2(\mathbf{X}^{m+1}, \mathbf{X}^m)$ and ρ_m. Since the metric of ε_m^2 depends on m, the convergence of ε_m^2 is hard to interpret with the exception of $p = 2$ when $\mathbf{A}_s^m$ is a matrix that is fixed and does not depend on $\mathbf{X}_m$. Although these observations are comforting, we do not have proofs about the convergence behavior of $\mathbf{X}^m$, except for the case $p = 2$, which are given by De Leeuw (1988).

For $p = 2$, the current algorithm has a linear convergence rate (see De Leeuw, 1988). Since we only use first order information, i.e. gradient information if STRESS is differentiable, this algorithm can be viewed as a steepest descent algorithm. Therefore, we expect that our majorization algorithm for $1 \leq p \leq 2$ has a linear convergence rate too.

Near a local minimum the STRESS function (like many functions) tends to behave like a quadratic function. Steepest descent algorithms are known to have orthogonal subsequent search directions near the local minimum, especially for almost quadratic functions. This so-called zigzag effect causes slow convergence. To avoid this undesirable behavior, De Leeuw and Heiser (1980) proposed to use a relaxed update $\mathbf{X}^+ = (1 - \alpha)\mathbf{X} + \alpha\overline{\mathbf{X}}$ with $0 \leq \alpha \leq 2$ and $\overline{\mathbf{X}}$ is the update as defined in (13). They reported that a fixed value of $\alpha = 2$ approximately halved the number of iterations. Here, we prove that the relaxed update also retains convergence for the general algorithm proposed above.

Let us start by noting that STRESS may be written alternatively as

$$\sigma^2(\mathbf{X}) = \eta_\delta^2 + \hat{\eta}^2(\mathbf{X} - \overline{\mathbf{X}}, \mathbf{X}) - \hat{\eta}^2(\overline{\mathbf{X}}, \mathbf{X}). \quad (19)$$

Further, using $\sigma^2(\mathbf{X}^+) \leq \hat{\sigma}^2(\mathbf{X}^+, \mathbf{X})$ and some reformulation we have

$$\begin{aligned} \sigma^2(\mathbf{X}^+) &\leq \eta_\delta^2 + \hat{\eta}^2((1-\alpha)\mathbf{X} + \alpha\overline{\mathbf{X}}, \mathbf{X}) - 2\hat{\rho}((1-\alpha)\mathbf{X} + \alpha\overline{\mathbf{X}}, \mathbf{X}) \\ &= \eta_\delta^2 + (1-\alpha)^2 \hat{\eta}^2(\mathbf{X} - \overline{\mathbf{X}}, \mathbf{X}) - \hat{\eta}^2(\overline{\mathbf{X}}, \mathbf{X}) \\ &= \eta_\delta^2 + \hat{\eta}^2(\mathbf{X} - \overline{\mathbf{X}}, \mathbf{X}) - \hat{\eta}^2(\overline{\mathbf{X}}, \mathbf{X}) + \alpha(\alpha - 2)\hat{\eta}^2(\mathbf{X} - \overline{\mathbf{X}}, \mathbf{X}) \\ &= \sigma^2(\mathbf{X}) + \alpha(\alpha - 2)\hat{\eta}^2(\mathbf{X} - \overline{\mathbf{X}}, \mathbf{X}). \end{aligned} \quad (20)$$

So for $0 \leq \alpha \leq 2$, the relaxed update indeed reduces STRESS.

4 Minkowski distances with $p < 1$ or $p > 2$

De Leeuw (1977) discusses how STRESS for general Minkowski distances could be handled, but he indicates that generally there is a nontrivial inner optimization problem. Mathar and Groenen (1991) gave an algorithm for the maximization of a convex function over a convex set to which the problem of minimizing STRESS can be restated. Here, we elaborate on the majorization approach. Clearly, for p outside the range $[1, 2]$ the majorizing function (12) is not valid anymore. Nevertheless, we could still use it as a local quadratic approximation of STRESS and use the update defined by (13). Although the convergence results are no longer valid, we might end up with a solution satisfying the stationary equations. However, since we wish to retain convergence, we use the majorization inequalities, which suggest different approaches for $p < 1$ and p> 2.

4.1 Minkowski distances with $p > 2$

For any $p \geq 2$ we can now majorize STRESS by

$$\sigma^2(\mathbf{X}) \quad \leq \quad \eta_\delta^2 + \eta^2(\mathbf{X}) - 2\hat{\rho}(\mathbf{X}, \mathbf{Y}). \tag{21}$$

The squared Minkowski distance is a convex function, hence so is $\eta^2(\mathbf{X})$. Consequently, the majorizing function (21), which is the sum of a convex function and a linear function, is convex itself. The basic majorization algorithm can be applied. Observe that in step 2 the minimum of a convex function has to be computed which may be done by standard (convex) optimization techniques. The updates are given by

$$\mathbf{X}^{m+1} = \arg\min_{\mathbf{X}} \{\eta^2(\mathbf{X}) - 2\hat{\rho}(\mathbf{X}, \mathbf{X}^m)\}, \tag{22}$$

and because of (4) a decreasing sequence of function values $\sigma^2(\mathbf{X}^m)$ is obtained.

To derive convergence results, we need an expression for the stationary equation for a minimum point of (22). Note that if (21) is differentiable at $\mathbf{X}^{m+1}$, the gradients of $\eta^2(\mathbf{X})$ and $\hat{\rho}(\mathbf{X}, \mathbf{X}^m)$ at $\mathbf{X}^{m+1}$ are given by

$$2\mathbf{A}_s(\mathbf{X}^{m+1})\mathbf{x}_s^{m+1} \quad \text{and} \quad \mathbf{B}_s(\mathbf{X}^m)\mathbf{x}_s^m, \tag{23}$$

respectively. Thus the stationary equation for a minimum point of (22) is $\mathbf{A}(\mathbf{X}^{m+1})\mathbf{x}_s^{m+1} = \mathbf{B}(\mathbf{X}^m)\mathbf{x}_s^m$ for all s or equivalently

$$\mathbf{x}_s^{m+1} = \mathbf{A}_s(\mathbf{X}^{m+1})^- \mathbf{B}_s(\mathbf{X}^m)\mathbf{x}_s^m. \tag{24}$$

The convergence proof is based on (24) and follows the arguments in the previous section closely. Here too we have $\rho_m \leq \eta_m$. Since by Cauchy–Schwarz $\sum_s \mathbf{x}_s^{m\prime}\mathbf{B}_s^m\mathbf{x}_s^m \leq (\sum_s \mathbf{x}_s^{m\prime}\mathbf{A}_s^{m+1}\mathbf{x}_s^m)^{1/2}(\sum_s \mathbf{x}_s^{m+1\prime}\mathbf{A}_s^{m+1}\mathbf{x}_s^{m+1})^{1/2}$, or

$$\rho(\mathbf{X}^m) \leq \hat{\eta}(\mathbf{X}^m, \mathbf{X}^{m+1})\eta(\mathbf{X}^{m+1}) \leq \eta_m\eta_{m+1}. \tag{25}$$

The last inequality follows from (10) with reversed inequality sign for $p > 2$. Finally we need

$$\eta^2(\mathbf{X}^{m+1}) = \hat{\rho}(\mathbf{X}^{m+1}, \mathbf{X}^m) \leq \rho(\mathbf{X}^{m+1}), \tag{26}$$

which follows from majorization (6). (25) and (26) yield that $\eta_m \leq \rho_m/\eta_m \leq \eta_{m+1}$ and $\eta_m \leq 1$. Altogether we have the chain

$$\eta_m^2 \leq \rho_m \leq \eta_{m+1}\eta_m \leq \eta_{m+1}^2 \leq \rho_{m+1} \leq \eta_{m+1} \leq 1, \tag{27}$$

that proves the convergence properties of ρ_m, η_m^2 and σ_m^2. We can obtain a converging sequence of ε_m^2 too by defining its metric to be $\mathbf{A}_s^{m+1}$ for each dimension s. A relaxed version of this algorithm can be obtained by searching in step 2 for a configuration which has lower value of the majorizing function in (21), but not necessarily the minimum value. For this relaxed procedure STRESS decreases iteratively, but we were not able to confirm the convergence result of ρ_m and η_m^2.

Note that for $p > 2$ zero distances are not replaced by ϵ in $\mathbf{A}_s(\mathbf{X})$. This is valid since for $p > 2$ Hölders inequality (9) has opposite sign, and simply states that $d_{ij}^2(\mathbf{X}) \geq 0$, which is clearly true.

4.2 Minkowski distances with $p < 1$

There is no metric interpretation for $d_{ij}(\mathbf{X})$ when $p < 1$, since the triangle inequality does not hold anymore, but the method can be applied along the same lines. For $p < 1$ the role of $\eta^2(\mathbf{X})$ and $\rho(\mathbf{X})$ is reversed. We can majorize STRESS by

$$\sigma^2(\mathbf{X}) \quad \leq \quad \eta_\delta^2 + \hat{\eta}^2(\mathbf{X}, \mathbf{Y}) - 2\rho(\mathbf{X}). \tag{28}$$

The Minkowski distance for $p < 1$ is a concave function in $\mathbf{X}$, which makes $-\rho(\mathbf{X})$ convex. Being the sum of two convex functions, the majorizing function (28) is convex too. The partial majorization algorithm for $p < 1$ consists of two steps: 1. compute the majorizing function (28), 2. compute the minimum of (28) for fixed $\mathbf{Y}$ and return to 1 unless convergence occurred. The solution of step 2 is a global minimum since the right hand side of (28) is a convex function, although this need not be a unique solution. It can be computed by using a standard (convex) optimization technique. Furthermore, it can be proved that (28) is differentiable at a global minimum.

It seems much harder to derive convergence results on the sequence of η_m^2 and ρ_m. Trivially, the sequence of σ_m^2 converges, due to majorization.

5 Differentiability at a local minimum

For Euclidean distances De Leeuw (1984) proved that STRESS is always differentiable at a local minimum for *usable* data. He calls data usable if $w_{ij}\delta_{ij} > 0$ holds for all pairs i, j. Here we extend this result to Minkowski distances with $p \geq 1$.

The proof follows the one of De Leeuw (1984) closely. Although STRESS is not differentiable at points $\mathbf{X}$ where some $d_{ij}(\mathbf{X})$ are zero, it has directional derivatives in all directions at any point. The directional derivative of STRESS at $\mathbf{X}$ in direction $\mathbf{Y}$ is defined by

$$\nabla\sigma^2(\mathbf{X};\mathbf{Y}) = \lim_{\varepsilon\downarrow 0} \frac{\sigma^2(\mathbf{X}+\varepsilon\mathbf{Y}) - \sigma^2(\mathbf{X})}{\varepsilon}. \tag{29}$$

The directional derivative in direction $\mathbf{y}$ of a function $f(\mathbf{x})$ with gradient $g(\mathbf{x})$ equals $g(\mathbf{x})'\mathbf{y}$ if $g(\mathbf{x})$ exists at $\mathbf{x}$. The directional derivatives of the (squared) Minkowski distances

are given by

$$\nabla d_{ij}(\mathbf{X};\mathbf{Y}) = \begin{cases} d_{ij}^{1-p}(\mathbf{X})\sum_{s=1}^{k}(y_{is}-y_{js})|x_{is}-x_{js}|^{(p-1)}, & \text{if } d_{ij}(\mathbf{X})\neq 0, \\ d_{ij}(\mathbf{Y}), & \text{if } d_{ij}(\mathbf{X})=0, \end{cases}$$
$$\nabla d_{ij}^2(\mathbf{X};\mathbf{Y}) = \begin{cases} 2d_{ij}^{1-\frac{p}{2}}(\mathbf{X})\sum_{s=1}^{k}(y_{is}-y_{js})|x_{is}-x_{js}|^{(p-1)}, & \text{if } d_{ij}(\mathbf{X})\neq 0, \\ 0, & \text{if } d_{ij}(\mathbf{X})=0. \end{cases} \tag{30}$$

For fixed $\mathbf{X}$, let $P = \{(i,j) \mid i<j,\ d_{ij}(\mathbf{X})>0\}$ and $Q = \{(i,j) \mid i<j,\ d_{ij}(\mathbf{X})=0\}$. Then the directional derivative of STRESS can be written as

$$\begin{aligned} \nabla\sigma^2(\mathbf{X};\mathbf{Y}) &= \sum_{i<j} w_{ij}\nabla d_{ij}^2(\mathbf{X};\mathbf{Y}) - 2\sum_{i<j} w_{ij}\delta_{ij}\nabla d_{ij}(\mathbf{X};\mathbf{Y}) \\ &= 2\sum_{(i,j)\in P}\frac{w_{ij}\sum_{s=1}^{k}(y_{is}-y_{js})|x_{is}-x_{js}|^{(p-1)}}{d_{ij}^{\frac{p}{2}-1}(\mathbf{X})} \\ &\quad -2\sum_{(i,j)\in P}\frac{w_{ij}\delta_{ij}\sum_{s=1}^{k}(y_{is}-y_{js})|x_{is}-x_{js}|^{(p-1)}}{d_{ij}^{p-1}(\mathbf{X})} \\ &\quad -2\sum_{(i,j)\in Q} w_{ij}\delta_{ij}d_{ij}(\mathbf{Y}). \end{aligned} \tag{31}$$

If $\mathbf{X}$ is a local minimum, the directional derivative in all directions is nonnegative, which implies that for any $\mathbf{Y}$ we must have

$$\nabla\sigma^2(\mathbf{X};\mathbf{Y}) + \nabla\sigma^2(\mathbf{X};-\mathbf{Y}) = -4\sum_{(i,j)\in Q} w_{ij}\delta_{ij}d_{ij}(\mathbf{Y}) \geq 0. \tag{32}$$

Now choose $\mathbf{Y}$ such that $d_{ij}(\mathbf{Y})>0$ for all i,j. Whenever $w_{ij}\delta_{ij}>0$ it follows that $d_{ij}(\mathbf{X})>0$. If $w_{ij}\delta_{ij}>0$ for all i,j then $Q=\emptyset$. Hence, for usable data the STRESS function with Minkowski distances $p\geq 1$ is always differentiable at a local minimum $\mathbf{X}$, and has $d_{ij}(\mathbf{X})>0$ for all i,j. A zero distance can only occur at a local minimum if and only if $w_{ij}\delta_{ij}=0$.

6 An example

To give an illustration of our algorithm we present a small example. Green, Carmone and Smith (1989) reported preferences of 38 students for 10 varieties of cola. Every pair of cola's was judged on their similarity on a nine value rating scale. The dissimilarities were accumulated over the subjects and the result is reported in Table 1. The cola data were analyzed in 2 dimensions with four different values of p, i.e. 1, 1.33, 1.66 and 2. For each value of p we used 25 different starting configurations. The iterations stopped whenever STRESS changed less than 10^{-8} in two subsequent iterations. Furthermore, we repeated the experiment using the relaxed update (that is $\alpha=2$). The results of both experiments are given in Table 2. The third column in the table gives the best STRESS values given p. The best fit was obtained for $p=1.33$. Inspection of the configuration corresponding to the best STRESS value suggested a diet non-diet dimension and a cola non-cola dimension. However, there was an inconsistency of the position of Diet 7-up on the diet non-diet dimension. Therefore, we positioned Diet 7-up in between Diet Slice and Diet Pepsi on the first dimension and restarted the algorithm. After 272 iterations

Table 1: The dissimilarities between 10 cola's reported by Green et al. (1989) accumulated over 38 judges.

	Pepsi	Coke	Classic Coke	Diet Pepsi	Diet Slice	Diet 7-Up	dr. Pepper	Slice	7-Up	Tab
Pepsi	0									
Coke	127	0								
Classic Coke	169	143	0							
Diet Pepsi	204	235	243	0						
Diet Slice	309	318	326	285	0					
Diet 7-Up	320	322	327	288	155	0				
dr. Pepper	286	256	258	259	312	306	0			
Slice	317	318	318	312	131	164	300	0		
7-Up	321	318	318	317	170	136	295	132	0	
Tab	238	231	242	194	285	281	256	291	297	0

Table 2: The results of scaling of the cola data of Green et al. (1989) for different values of p.

	Average STRESS	Average no of iterations	Lowest STRESS $\sigma(\mathbf{X}^*)$	No of iterations of $\sigma(\mathbf{X}^*)$
normal update				
$p = 1$	0.11713930	233.36	0.04785617	696
$p = 1.33$	0.05160195	251.96	0.03199579	284
$p = 1.66$	0.04086771	360.44	0.03491206	758
$p = 2$	0.04145104	145.92	0.03678052	141
relaxed update				
$p = 1$	0.11787413	100.40	0.04193646	232
$p = 1.33$	0.05483608	149.60	0.03425142	178
$p = 1.66$	0.04111878	243.32	0.03467676	350
$p = 2$	0.04070994	92.08	0.03685458	95

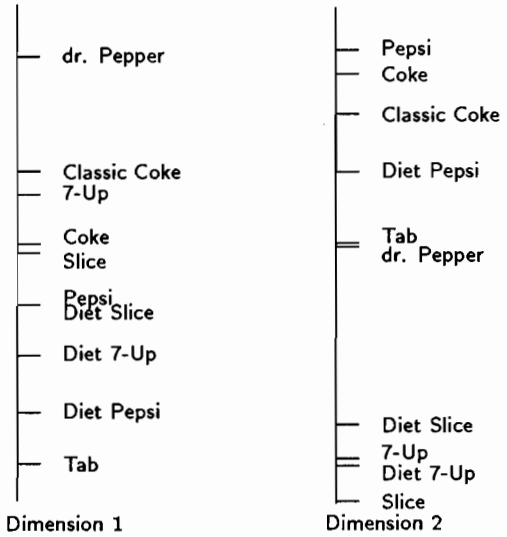

Figure 1: The configuration of the best solution of the cola data, with $p = 1.33$ and having STRESS 0.03175500 .

a configuration was reached with lower STRESS value 0.03175500. This configuration is given in Figure 1. On purpose we do not display a two-dimensional representation, since it may lead to interpretations based on Euclidean distances. The first dimension appears to be a diet non-diet dimension. There is a large difference between dr. Pepper and Tab, which is absent in the second dimension. The other beverages are separated in the second dimension into the group Pepsi, Coke, Classic Coke and Diet Pepsi and the other group of (Diet) Slice and (Diet) 7-Up. This dimension could be interpreted as a Cola non-Cola dimension.

The second column in Table 2 suggests that the relaxed update is indeed faster than the normal update, as was expected. The average STRESS of 25 runs was generally larger for the relaxed update than the normal update, except for $p = 2$.

7 Concluding remarks

The majorization approach can be easily extended to deal with non-metric multidimensional scaling. In that case, the procedure can be incorporated in an alternating least squares algorithm where the majorization step finds a better configuration and the optimal scaling step yields optimal pseudo- distances, given the distances. For details we refer to Kruskal (1977) and De Leeuw and Heiser (1977).

The current algorithm gives an alternative for the approach which accommodates the possibility of negative dissimilarities discussed by Heiser (1991). Negative dissimilarities occur (among other situations) when city-block distances are fitted dimensionwise as was done in Heiser (1989).

Some open questions with regard to the current algorithm remain. We expect our algorithm to have a linear converge rate, as is the case when $p = 2$ (see De Leeuw, 1988). Further research has to be done on this topic. Another interesting topic is how to incorporate constraints on the configuration, as discussed by De Leeuw and Heiser (1980), in the majorization algorithm for STRESS with general Minkowski distance. The local minimum problem also remains to be investigated.

References

Arabie, P. (1991), Was Euclid an unnecessarily sophisticated psychologist? *Psychometrika, 56*, 567-587.

De Leeuw, J. (1977), Applications of convex analysis to multidimensional scaling. In: J. Barra et al. (Eds.), *Recent developments in statistics*, 133-145, Amsterdam: North-Holland.

De Leeuw, J. (1984), Differentiability of Kruskal's Stress at a local minimum. *Psychometrika, 49*, 111-113.

De Leeuw, J. (1988), Convergence of the majorization method for multidimensional scaling. *Journal of Classification, 5*, 163-180.

De Leeuw, J. (1992), *Fitting distances by least squares.* Unpublished report.

De Leeuw, J., Heiser, W.J. (1977), Convergence of correction-matrix algorithms for multidimensional scaling. In: J.C. Lingoes (Ed.), *Geometric representations of relational data*, 735-751, Ann Arbor (Michigan): Mathesis Press.

De Leeuw, J., Heiser, W.J. (1980), Multidimensional scaling with restrictions on the configuration. In: P.R. Krishnaiah (Ed.), *Multivariate analysis V*, 501-522, Amsterdam: North-Holland.

Green, P.E., Carmone, F.J. Jr, Smith, S.M. (1989), *Multidimensional scaling, concepts and applications.* Boston: Allyn and Bacon.

Groenen, P.J.F., Heiser, W.J. (1991), *An improved tunneling function for finding a decreasing series of local minima.* Internal report RR-91-06, Leiden: Department of Data Theory.

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

Heiser, W.J. (1988), Multidimensional scaling with least absolute residuals. In: H.H. Bock (Ed.), *Classification and related methods of data analysis*, 455-462, Amsterdam: North-Holland.

Heiser, W.J. (1989), The city-block model for three-way multidimensional scaling. In: R. Coppi and S. Bolasco (Eds.), *Multiway Data Analysis*, 395-404, Amsterdam: North-Holland.

Heiser, W.J. (1991), A generalized majorization method for least squares multidimensional scaling of pseudo distances that may be negative. *Psychometrika, 56*, 7-27.

Hubert, L.J., Arabie, P. (1986), Unidimensional scaling and combinatorial optimization. In: J. De Leeuw, W.J. Heiser, J. Meulman and F. Critchley (Eds.), *Multidimensional Data Analysis*, 181-196, Leiden: DSWO Press.

Kruskal, J.B. (1964a), Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika, 29*,1-27.

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

Kruskal, J.B. (1977), Multidimensional scaling and other methods for discovering structure. In: K. Enslein, A. Ralston and H.S. Wilf (Eds.), *Statistical methods for digital computers, Vol III*, 296-339, New York: Wiley.

Mathar, R., Groenen, P.J.F. (1991), Algorithms in convex analysis applied to multidimensional scaling. In: E. Diday and Y. Lechevallier (Eds.), *Symbolic-numeric data analysis and learning*, 45-56, New York: Nova Science Publishers.

Mathar, R., Meyer, R. (1992), *Algorithms in convex analysis to fit l_p-distance matrices.* Unpublished manuscript.

Meulman, J.J. (1986), *A distance approach to nonlinear multivariate analysis*, Leiden: DSWO Press.