RR-97-01

CITY-BLOCK SCALING: SMOOTHING STRATEGIES FOR AVOIDING LOCAL MINIMA

Patrick J.F. Groenen
Willem J. Heiser
Jacqueline J. Meulman

Department of Data Theory
Leiden University

City-Block Scaling: Smoothing Strategies for Avoiding Local Minima

P.J.F. Groenen[1] , W.J. Heiser, J.J. Meulman

Department of Data Theory, Leiden University,
P.O. Box 9555, 2300 RB Leiden, The Netherlands
(e-mail: groenen@rulfsw.fsw.leidenuniv.nl)

Abstract: Multidimensional scaling (MDS) with city-block distances suffers from many local minima if the Stress function is minimized. In fact, the problem can be viewed as a combinatorial problem, where finding the correct order of the coordinates on a dimension is crucial for attaining the minimum. Several strategies have been proposed for arriving at a global minimum of the Stress function. We pay particular attention to Pliner's (1996) smoothing strategy for unidimensional scaling, which smoothes the concave part of the Stress function. We discuss three extensions of this strategy to the multidimensional case with city-block distances. The first extension is shown to lead to problems because it yields a unidimensional solution. A second extension, proposed by Pliner (1986), and a third extension, distance smoothing introduced here, do not have this problem. Numerical experiments with the smoothing strategy have been limited to the unidimensional case. Therefore, we present a comparison study using real data, which shows that the smoothing strategy performs better than three other strategies considered.

1 Introduction

In multidimensional scaling (MDS) the objective is to represent dissimilarities between objects as distances between points in a low dimensional space. Apart from the Euclidean distance, the *city-block* (or L_1) distance is a popular choice. One of the properties of the city-block distance (not shared by the Euclidean distance) is dimensional additivity, that is, the total distance is a sum of the distances per dimension. For an overview of developments in the area of city-block distances, see Arabie (1991). The purpose of least squares MDS can be formalized mathematically as the minimization of the raw Stress function (Kruskal, 1964),

$$\sigma(\mathbf{X}) = \sum_{i<j} w_{ij} \left(\delta_{ij} - d_{ij}(\mathbf{X})\right)^2, \tag{1}$$

over the $n \times p$ matrix of coordinates $\mathbf{X}$ of n objects in p dimensions, where w_{ij} are nonnegative weights, δ_{ij} are nonnegative dissimilarities, and $d_{ij}(\mathbf{X})$

[1]Supported by The Netherlands Organization for Scientific Research (NWO) by grant nr. 030-56-403 for the 'PIONEER' project 'Subject Oriented Multivariate Analysis'.

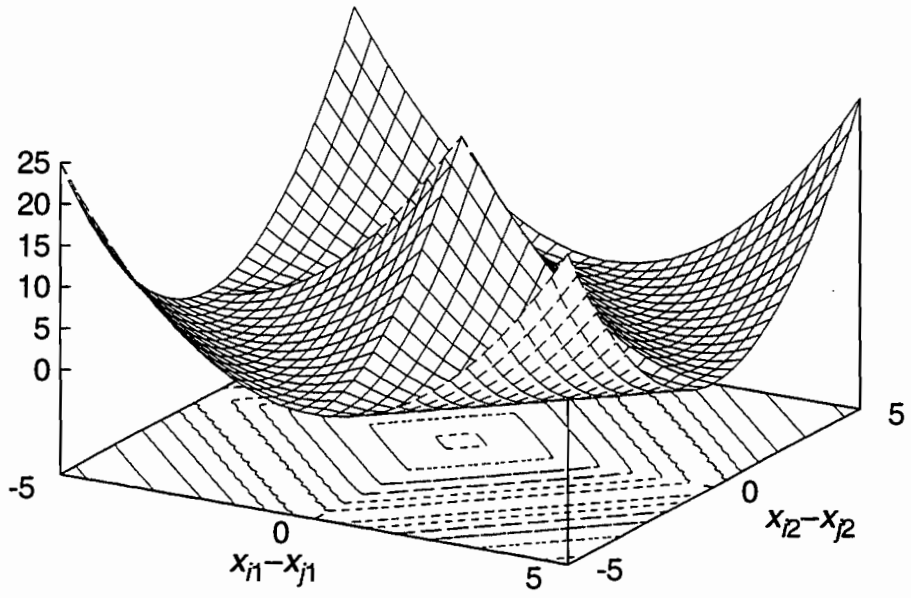


Figure 1: The contribution to $\sigma(\mathbf{X})$ of the term $(\delta_{ij} - d_{ij}(\mathbf{X}))^2$.

is the city-block distance between points i and j defined by

$$d_{ij}(\mathbf{X}) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|. \tag{2}$$

For city-block MDS the Stress function can be written as

$$\begin{aligned}
\sigma(\mathbf{X}) &= \sum_{i<j} w_{ij}\delta_{ij}^2 + \sum_{i<j} w_{ij} d_{ij}^2(\mathbf{X}) - 2\sum_{i<j} w_{ij}\delta_{ij} d_{ij}(\mathbf{X}) \\
&= \sum_{i<j} w_{ij}\delta_{ij}^2 + \sum_{i<j} w_{ij} \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \\
&\quad +2\sum_{i<j} w_{ij} \sum_{k<l} |x_{ik} - x_{jk}||x_{il} - x_{jl}| - 2\sum_{i<j} w_{ij}\delta_{ij} \sum_{k=1}^{p} |x_{ik} - x_{jk}| \\
&= \eta_\delta^2 + \eta_k^2(\mathbf{X}) + \eta_{k\neq l}^2(\mathbf{X}) - 2\rho(\mathbf{X}).
\end{aligned} \tag{3}$$

It has been noted by several authors (Heiser (1989), Arabie (1991), Hubert et al. (1992), Groenen and Heiser (1996)) that city-block MDS by minimization of (3) suffers from many local minima. The main concern of this paper is to develop and evaluate strategies that try to avoid local minima. To see why Stress has many local minima, consider the contribution of a single error term for objects i, j to $\sigma(\mathbf{X})$ in two dimensions with $w_{ij} = 1$, i.e., the residual $(\delta_{ij}^2 - [|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}|])^2$, see Figure 1. The sharp ridges give rise to the large number of local minima since they act as a barrier between four potential regions of attraction, due to discontinuities in the direction of change. Some of the best strategies proposed so far use combinatorial approaches (Heiser (1989), Hubert et al. (1992)), where finding the correct order of the coordinates on any dimension is crucial for attaining the overall

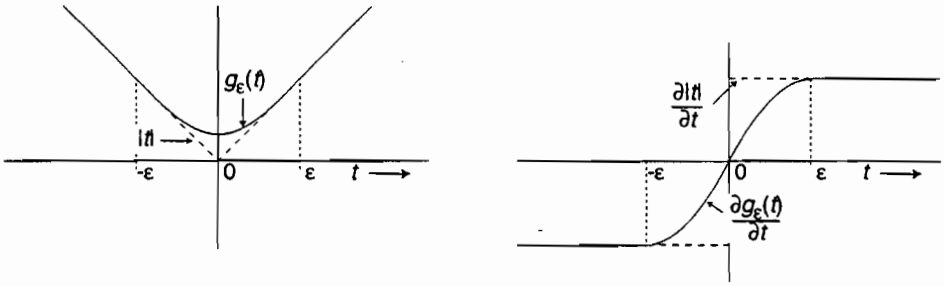


Figure 2: The left panel shows function $|t|$ (dashed line) and its smoothed version $g_\epsilon(t)$ (solid line) and the right panel shows the first derivative of these functions.

minimum. In this paper, we elaborate on a smoothing strategy proposed by Pliner (1986, 1996). The advantage of this approach is that continuous minimization methods can be used instead of combinatorial strategies, which become unpractical for large n. We propose a new smooth loss function that converges to Stress. So far, the smoothing strategy has not been compared to other strategies except for unidimensional scaling. We fill this gap by presenting a small comparison study that investigates the performance of the smoothing strategy relative to competing strategies.

2 The smoothing strategy

The smoothing strategy of Pliner (1986, 1996) smoothes the sharp ridges in the Stress function. It tries to avoid local minima by letting the smooth function gradually approach the original function, while removing the discontinuities in the gradient near the ridges. In unidimensional scaling, Pliner (1996) replaces $-|x_i - x_j|$ in $\rho(\mathbf{X})$ by $-g_\epsilon(x_i - x_j)$, see Figure 2. Here, $g_\epsilon(t)$ is defined by

$$g_\epsilon(t) = \begin{cases} t^2(3\epsilon - |t|)/3\epsilon^2 + \epsilon/3, & \text{if } |t| < \epsilon, \\ |t|, & \text{if } |t| \geq \epsilon. \end{cases} \tag{4}$$

Thus, distances smaller than ϵ are replaced by a smooth function, but large distances remain as they are. For unidimensional scaling, this smoothing strategy turns out to be very successful in reaching global minima (Pliner (1996)).

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 $|t| \geq \epsilon$. One such function — well known from robust statistics — is the Huber function (Huber, 1981), which is similar to (4) except that it is quadratic in t for $|t| < \epsilon$. Use of the Huber function in the present context is currently being studied.

The unidimensional smoothing strategy can be generalized to city-block MDS in several ways. Pliner (1996) seems to suggest that smoothing should only be applied to the terms of $\sigma(\mathbf{X})$ containing $-|x_{ik} - x_{jk}|$, i.e.,

$$\sigma_1(\mathbf{X}|\epsilon) = \eta_\delta^2 + \eta_k^2(\mathbf{X}) + \eta_{k\neq l}^2(\mathbf{X}) - 2\rho(\mathbf{X}|\epsilon), \tag{5}$$

where $\rho(\mathbf{X}|\epsilon) = \sum_{i<j} w_{ij}\delta_{ij} \sum_{k=1}^{p} g_\epsilon(x_{ik} - x_{jk})$. Thus, $-\sum_{k=1}^{p} g_\epsilon(x_{ik} - x_{jk})$ smoothes $-d_{ij}(\mathbf{X})$. An important property of (5) is that as ϵ approaches zero, $\sigma_1(\mathbf{X}|\epsilon)$ approaches the original Stress function $\sigma(\mathbf{X})$. However, $\sigma_1(\mathbf{X}|\epsilon)$ has the undesirable property that systematically unidimensional solutions are found whatever the chosen dimensionality. This property is undesirable because additional dimensions in all practical cases would decrease Stress. Therefore, $\sigma_1(\mathbf{X}|\epsilon)$ cannot be used as a smooth version of Stress for city-block MDS when the rank of $\mathbf{X}$ should be larger than 1.

A second smoothing loss function was proposed by Pliner (1986), i.e., he smoothes every factor $|x_{ik} - x_{jk}|$ by $g_\epsilon(x_{ik} - x_{jk})$ and leaves the quadratic terms $(x_{ik} - x_{jk})^2$ in $\eta_k^2(\mathbf{X})$ of $\sigma(\mathbf{X})$ as they are. This yields the smooth loss function

$$\sigma_2(\mathbf{X}|\epsilon) = \eta_\delta^2 + \eta_k^2(\mathbf{X}) + \eta_{k\neq l}^2(\mathbf{X}|\epsilon) - 2\rho(\mathbf{X}|\epsilon), \tag{6}$$

where $\eta_{k\neq l}^2(\mathbf{X}|\epsilon) = \sum_{k\neq l} \sum_{i<j} w_{ij} g_\epsilon(x_{ik} - x_{jk}) g_\epsilon(x_{il} - x_{jl})$. This loss function also approaches $\sigma(\mathbf{X})$ as $\epsilon \to 0$ and, in general, retains the rank of the initial configuration.

The smoothing loss function that we propose here is called *distance smoothing*. We replace all factors $|x_{ik} - x_{jk}|$ by $g_\epsilon(x_{ik} - x_{jk})$. Thus, we minimize

$$\begin{aligned} \sigma_3(\mathbf{X}|\epsilon) &= \sum_{i<j} w_{ij} \left(\delta_{ij} - \sum_k g_\epsilon(x_{ik} - x_{jk}) \right)^2 \\ &= \eta_\delta^2 + \eta_k^2(\mathbf{X}|\epsilon) + \eta_{k\neq l}^2(\mathbf{X}|\epsilon) - 2\rho(\mathbf{X}|\epsilon), \end{aligned} \tag{7}$$

thereby smoothing $\eta_k^2(\mathbf{X})$ as well by $\eta_k^2(\mathbf{X}|\epsilon) = \sum_k \sum_{i<j} w_{ij} g_\epsilon^2(x_{ik} - x_{jk})$. The advantage of this adaptation is that the loss function remains least squares, because it is the sum of squared differences of the dissimilarities and the smoothed distances. The minimum of $\sigma_3(\mathbf{X}|\epsilon)$ is not biased to a zero difference of the coordinates (in contrast to $\sigma_1(\mathbf{X}|\epsilon)$), and it has the property that it converges to $\sigma(\mathbf{X})$ as ϵ approaches zero. An even stronger property holds if we assume that all $|x_{ik} - x_{jk}|$ are strictly positive: then there exists an ϵ for which $\sigma_3(\mathbf{X}|\epsilon)$ reduces to $\sigma(\mathbf{X})$.

The smoothing strategy for city-block MDS proposed in this paper is: (a) set initial value $\epsilon \leftarrow \epsilon_0$ and fix the number of smoothing steps $r_{\max}$, (b) for $r = 1$ to $r_{\max}$ do: minimize $\sigma_3(\mathbf{X}|\epsilon)$ and reduce ϵ, i.e., $\epsilon \leftarrow \epsilon - \epsilon_0/r$, (c) minimize $\sigma(\mathbf{X})$. As initial value for ϵ in smoothing for unidimensional scaling, Pliner (1996) recommends to choose $\epsilon_0 = 2\max_{1\leq i\leq n} n^{-1} \sum_{j=1}^{n} \delta_{ij}$ assuming all $w_{ij} = 1$. Since nonidentical weights are present in $\sigma_3(\mathbf{X}|\epsilon)$, we set ϵ_0 equal to $2\max_{1\leq i\leq n}(\sum_{j=1}^{n} w_{ij})^{-1} \sum_{j=1}^{n} w_{ij}\delta_{ij}$.

So far, we discussed smoothing for metric city-block MDS. Without much difficulty, smoothing can be extended to *nonmetric* city-block MDS as well. We can proceed as in ordinary nonmetric MDS (see Kruskal (1964), or, e.g., Borg and Groenen (1997)). In smoothing for ordinal city-block MDS one substitutes δ_{ij} by $\hat{d}_{ij}$ in (7), where the $\hat{d}_{ij}$'s are least squares approximates to the distances, constrained to retain the order of the data and have a fixed sum of squares.

3 Performance of the smoothing strategy

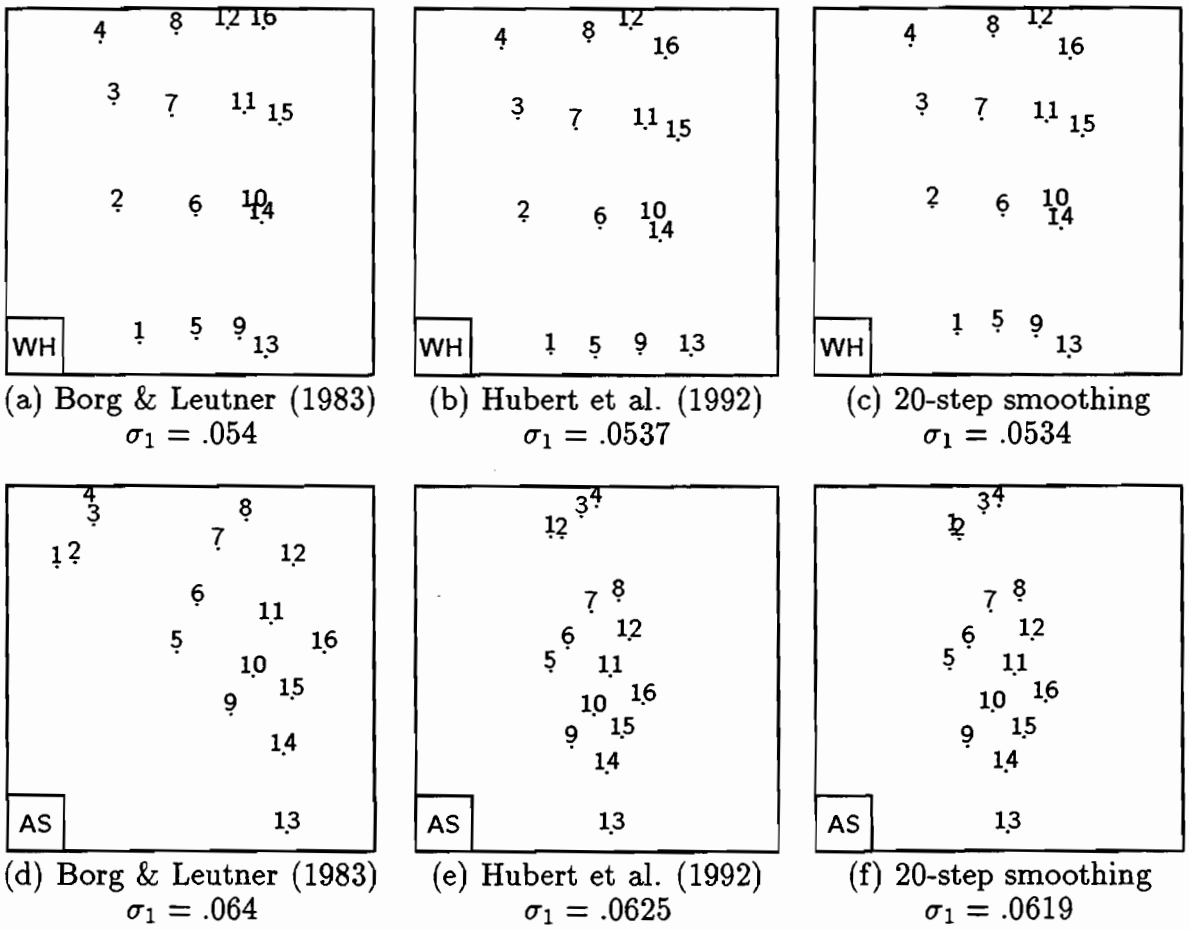
To test the performance of our distance smoothing algorithm, we compare it to three other methods: (a) the combinatorial strategy of Hubert et al. (1992), (b) the majorization approach of Groenen et al. (1995), here called "plain majorization", and (c) the MDS program in SYSTAT (Wilkinson (1988)). We used two data sets of Borg and Leutner (1983) on the perception of rectangles which were also analyzed by Hubert et al. (1992). Subjects rated the similarity of pairs of rectangles on a rating scale ranging from 0='equal, identical' to 9='extremely different'. The 16 rectangles varied in width and height (both in four levels). The data set (to which we refer as WH) contains the averaged similarity ratings of 21 subjects. A second set of rectangles was created by varying 'width + height' and 'width − height', thereby emphasizing the area and shape of the rectangles. The average over the similarity ratings of 21 other subjects make up the second data set (the AS data). The data are treated ordinally by the primary approach to ties, which implies that in each step of our smoothing algorithm the proximities are optimally transformed in a least squares way by monotone regression. The figures for SYSTAT and the combinatorial strategy are copied from Hubert et al. (1992). The value of Stress reported is Kruskal's Stress-1 which can be shown to be equal to $(\sigma(\mathbf{X})/\eta_\delta^2)^{1/2}$ if we allow $\mathbf{X}$ to be optimally dilated (see, Borg and Groenen (1997)). To minimize $\sigma_3(\mathbf{X}|\epsilon)$ we have used majorization. Details of this algorithm will be presented in a forthcoming paper (Groenen, Heiser, and Meulman, in preparation).

For each of the two data sets, 100 random starts were fed into our smoothing algorithm, both with 5 and 20 smoothing steps. The minimization of the smoothing function $\sigma_3(\mathbf{X}|\epsilon)$ was stopped whenever the decrease in loss was smaller than 10^{-5}. The stopping criterion in the final minimization of $\sigma(\mathbf{X})$ was set to 10^{-8}.

The results are summarized in Table 1. These results show that the double smoothing strategy with 20 smoothing steps is overall the best strategy. Second best are the 5 step smoothing strategy and the combinatorial method. The worst two methods are plain majorization and SYSTAT. The maximum values of Stress of the smoothing strategies (and from the first quartile upwards for the plain majorization approach) are very high because the transformation obtained by ordinal MDS is degenerated towards equal values for almost all proximities. Therefore, a Shepard diagram (that plots the original proximity values against the transformed values along with the residuals) displays an almost horizontal line. In comparison to 20-step smoothing, the 5-step strategy gives slightly higher Stress values in the summary statistics. For the WH data, 47% had Stress-1 of .0534, and 42% of .0541 using the 20-step smoothing approach, whereas the 5-step version had 33% of .0534, and 84% was smaller than .0565. For the AS data, the 20-step approach had 8% Stress-1 of .0619, and 72% smaller than .0700, whereas the 5-step approach found in 3% of the cases Stress-1 of .0626, and 57% smaller than .0700. These results indicate that the smoothing strategies give good local

Table 1: Summary statistics of Stress-1 values for ordinal city-block MDS on the rectangle data sets WH and AS for 100 multiple random starts. The figures for SYSTAT and the combinatorial strategy (indicated by an '*') are copied from Hubert et al. (1992).

Strategy	Minimum	1st Quartile	Median	3rd Quartile	Maximum
WH data					
plain majorization	.0903	.3594	.3671	.3750	.4001
SYSTAT*	.0665	.1545	.1748	.2992	.3758
combinatorial*	.0537	.0666	.0701	.1368	.1803
smoothing 5 steps	.0534	.0534	.0565	.0565	.3302
smoothing 20 steps	.0534	.0534	.0541	.0541	.3235
AS data					
plain majorization	.1069	.3568	.3660	.3757	.4001
SYSTAT*	.0702	.0783	.1252	.3617	.3804
combinatorial*	.0625	.0699	.0754	.0860	.1078
smoothing 5 steps	.0626	.0668	.0693	.0844	.3294
smoothing 20 steps	.0619	.0644	.0666	.0741	.3294



(a) Borg & Leutner (1983) $\sigma_1 = .054$ (b) Hubert et al. (1992) $\sigma_1 = .0537$ (c) 20-step smoothing $\sigma_1 = .0534$

(d) Borg & Leutner (1983) $\sigma_1 = .064$ (e) Hubert et al. (1992) $\sigma_1 = .0625$ (f) 20-step smoothing $\sigma_1 = .0619$

Figure 3: Solutions obtained for WH data (upper panels) and AS data (lower panels) by (a) Borg and Leutner (1983) (left panels), (b) Hubert et al. (1992) (middle panels), and (c) the 20-step smoothing strategy (right panels).

minima in more than half of the searches. Moreover, as the number of steps is increased, the frequency of finding the global minimum also increases.

The (best) configurations found by Borg and Leutner (1983), Hubert et al. (1992), and the 20-step smoothing strategy for the WH data are shown in Figure 3 (upper panels). All three solutions for the WH data reconstruct the grid like structure used for generating the rectangles. The difference between the solutions is either in the positioning of points 12 and 16, or in that of points 1, 5, 9, and 13. For the AS data in Figure 3 (lower panels), the solution found by Borg and Leutner (1983) is considerably different from the solutions found by the combinatorial and the smoothing strategies. Hubert et al. (1992) state that height seems to be much more important than width. Note that the combinatorial solution in panel (e) of Figure 3 only differs from the smoothing solution in panel (f) in the location of point 1.

4 Discussion and conclusions

This paper shows that the combinatorial problem of least-squares city-block MDS can be solved by a smoothing strategy, which is a continuous minimization problem by nature. We have considered three extensions of the basic smoothing strategy suggested by Pliner (1996) for unidimensional scaling. One extension that only smoothes the concave part of the Stress function systematically yields rank-one solutions. However, the distance smoothing strategy proposed in this paper, which smoothes all the absolute value terms in the Stress function, and the strategy of Pliner (1986) give technically correct results.

A comparison study of two real data sets suggests that the distance smoothing strategy gives somewhat better results than the combinatorial strategy of Hubert et al. (1992), and much better results than gradient based methods such as the plain majorization approach of Groenen et al. (1995) and city-block MDS in SYSTAT. Moreover, as the number of smoothing steps is increased, the probability of finding a global minimum also increases.

The distance smoothing strategy can be extended to MDS with Euclidean distances as well. In this way, one would get a method that, hopefully, finds global minima with a much larger probability. Constraints on the configuration in confirmatory MDS as proposed by De Leeuw and Heiser (1980) can be implemented in the majorizing algorithm of smoothed city-block MDS without much difficulty.

In this paper we have regarded the distance smoothing strategy as an approach to find the global minimum in city-block MDS. However, the smoothing function of the absolute value, $g_\epsilon(x_{is} - x_{js})$, can also be viewed as being part of a model. In such a model, differences larger than ϵ are treated as they are, but smaller differences are made somewhat larger. The extreme of a zero difference is transformed by the smoother into a value of $\epsilon/3$. This model approach could be applied if all dissimilarities are by their nature larger than this value.

References

ARABIE, P. (1991): Was Euclid an unnecessarily sophisticated psychologist? *Psychometrika, 56*, 567–587.

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

BORG, I. and LEUTNER, D. (1983): Dimensional models for the perception of rectangles. *Perception and Psychophysics, 34*, 257–269.

DE LEEUW, J. and HEISER, W.J. (1980): Multidimensional scaling with restrictions on the configuration. In P.R. Krishnaiah (eds.): *Multivariate analysis, V.* North Holland Publishing Company, Amsterdam, 501–522.

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. (in preparation): Global optimization in least squares multidimensional scaling: A smoothing approach. Working paper, Department of Data Theory, Leiden, The Netherlands.

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: R. Coppi and S. Bolasco (eds.): *Multiway data analysis.* Elsevier Science, Amsterdam, 395–404.

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

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

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

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.

WILKINSON, L. (1988): SYSTAT: The system for statistics. SYSTAT Inc., Evanston, IL.