PSYCHOMETRIKA—VOL. 61, NO. 3, 529–550
SEPTEMBER 1996

THE TUNNELING METHOD FOR GLOBAL OPTIMIZATION IN MULTIDIMENSIONAL SCALING

PATRICK J. F. GROENEN AND WILLEM J. HEISER

DEPARTMENT OF DATA THEORY, FACULTY OF SOCIAL AND BEHAVIOURAL SCIENCES, LEIDEN UNIVERSITY, THE NETHERLANDS

This paper focuses on the problem of local minima of the STRESS function. It turns out that unidimensional scaling is particularly prone to local minima, whereas full dimensional scaling with Euclidean distances has a local minimum that is global. For intermediate dimensionality with Euclidean distances it depends on the dissimilarities how severe the local minimum problem is. For city-block distances in any dimensionality many different local minima are found. A simulation experiment is presented that indicates under what conditions local minima can be expected. We introduce the tunneling method for global minimization, and adjust it for multidimensional scaling with general Minkowski distances. The tunneling method alternates a local search step, in which a local minimum is sought, with a tunneling step in which a different configuration is sought with the same STRESS as the previous local minimum. In this manner successively better local minima are obtained, and experimentation so far shows that the last one is often a global minimum.

Key words: multidimensional scaling, iterative majorization, global optimization, tunneling method.

1. Introduction

The aim of multidimensional scaling (MDS) is to represent dissimilarities between objects by interpoint distances in a low-dimensional space. This basic idea of MDS goes back to Torgerson (1958), Shepard (1962) and Kruskal (1964a, 1964b). There are different ways to implement MDS, but we shall use Kruskal's raw STRESS

$$\sigma(\mathbf{X}) = \sum_{i<j} w_{ij}(\delta_{ij} - \mathrm{d}_{ij}(\mathbf{X}))^2, \tag{1}$$

where $\mathbf{X}$ is an $n \times p$ coordinate matrix of the n objects in p dimensions, δ_{ij} is the dissimilarity between object i and object j, and w_{ij} are nonnegative fixed weights. The Minkowski distance between the coordinates in row i and row j is denoted by $\mathrm{d}_{ij}(\mathbf{X}) = (\sum_{s=1}^{p} |x_{is} - x_{js}|^q)^{1/q}$, with $q \geq 1$ the Minkowski parameter. Special cases are the *city-block* distance for $q = 1$ (for an overview, see Arabie, 1991), the *Euclidean* distance for $q = 2$, and the *dominance* distance for $q = \infty$. Different algorithms exist for minimizing (1) (or slight adaptations of it), like KYST of Kruskal, Young and Seery (1977). These algorithms are based on the method of steepest descent that makes use of the gradient of the STRESS function. This gradient is not defined if two points coincide, which happens regularly for $q = 1$. However, the majorization algorithm of de Leeuw and Heiser (1977), Heiser and de Leeuw (1977) and de Leeuw (1988) for the Euclidean distance case, and generalized by Groenen, Mathar, and Heiser (1995) to the Minkowski distance case handles zero distances without problems. Moreover, it can be proved that the majorization algorithm

This paper is based on the 1994 Psychometric Society's outstanding thesis award of the first author. The authors would like to thank Robert Tijssen of the CWTS Leiden for kindly making available the co-citation data of the Psychometric literature. This paper is an extended version of the paper presented at the Annual Meeting of the Psychometric Society at Champaign-Urbana, Illin., June 1994.

Requests for reprints should be sent to Patrick J. F. Groenen, Department of Data Theory, Faculty of Social and Behavioural Sciences, PO Box 9555, 2300 RB Leiden, THE NETHERLANDS.

0033-3123/96/0033-3123$00.75/0

converges to a stationary point. Therefore, we base our study on the iterative majorization algorithm. One of the difficult problems with (1) is that it may have many different local minima. This problem of local minima is the main topic of this paper. The paper is organized as follows. First, we consider two special cases of (1) for which the local minimum problem is different. For unidimensional scaling, many local minima can be expected, whereas full-dimensional scaling with Euclidean distances only has a global minimum. Then we present the tunneling method for finding a decreasing series of local minima. Several adaptations of the tunneling method are discussed to make it suitable for MDS. Then, the tunneling method is implemented by an algorithm based on iterative majorization. We continue with a simulation experiment in which we investigate the seriousness of local minima for STRESS with city-block and Euclidean distances and various combinations of p and n. Finally, we study the behavior of the tunneling method with some case studies. Our present study is confined to the metric MDS STRESS function (1). In principle, we could have extended our research to nonmetric MDS, but it appeared that the local minimum problem for the metric case is already difficult enough.

Before proceeding, two issues need attention. The first issue concerns the dependence of STRESS (1) on the normalization of δ_{ij}. This is undesirable because we want the STRESS value to be invariant under a normalization factor of δ_{ij}'s in the same way as a correlation coefficient is independent of the variance of the two variables involved. Let us decompose STRESS into

$$\sigma(\mathbf{X}) = \sum_{i<j} w_{ij}(\delta_{ij} - \mathrm{d}_{ij}(\mathbf{X}))^2 = \sum_{i<j} w_{ij}\delta_{ij}^2 + \sum_{i<j} w_{ij}\mathrm{d}_{ij}^2(\mathbf{X}) - 2\sum_{i<j} w_{ij}\delta_{ij}\mathrm{d}_{ij}(\mathbf{X})$$

$$= \eta_\delta^2 + \eta^2(\mathbf{X}) - 2\rho(\mathbf{X}). \tag{2}$$

Now, at a stationary point $\mathbf{X}^*$ it holds that $0 \leq \sigma(\mathbf{X}^*)/\eta_\delta^2 \leq 1$. The value of $\sigma(\mathbf{X}^*)/\eta_\delta^2$ can be seen as the proportion of unexplained variation in the dissimilarities (see, e.g., Commandeur, 1992), which can be seen as follows. Stationarity of $\mathbf{X}^*$ implies that for fixed $\mathbf{X}^*$ and free β $\sigma(\beta\mathbf{X}^*)$ must be stationary too. Since $\mathrm{d}_{ij}(\mathbf{X})$ is a positively homogenous function we have $\mathrm{d}_{ij}(\beta\mathbf{X}^*) = \beta\mathrm{d}_{ij}(\mathbf{X}^*)$, so that $\sigma(\beta\mathbf{X}^*) = \eta_\delta^2 + \beta^2\eta^2(\mathbf{X}^*) - 2\beta\rho(\mathbf{X}^*)$. The β^* that minimizes $\sigma(\beta\mathbf{X}^*)$ equals $\rho(\mathbf{X}^*)/\eta^2(\mathbf{X}^*)$ (see, e.g., Mathar & Groenen, 1991), so that $\sigma(\beta^*\mathbf{X}^*) = \eta_\delta^2 - \rho^2(\mathbf{X}^*)/\eta^2(\mathbf{X}^*)$. Dividing both sides by η_δ^2 gives $\sigma(\beta^*\mathbf{X}^*)/\eta_\delta^2 = 1 - \rho^2(\mathbf{X}^*)/(\eta_\delta^2\eta^2(\mathbf{X}^*))$. It is not difficult to show that $\rho^2(\mathbf{X}^*)/(\eta_\delta^2\eta^2(\mathbf{X}^*))$ is the squared Tucker's (1951) congruence coefficient between the (order $n(n-1)/2$) vectors of dissimilarities and distances, which is always between 0 and 1. A negative Tucker's coefficient does not arise due to the nonnegativity of distances and dissimilarities. In the sequel we use always $\sigma(\mathbf{X})/\eta_\delta^2$ for the STRESS function. The relation of Tucker's congruence coefficient and the minimization of STRESS was studied earlier by de Leeuw (1977).

The second preliminary issue is that we assume, without loss of generality, that the weight matrix $\mathbf{W} = \{w_{ij}\}$ is *irreducible*, that is, there exists no partitioning of objects into disjoint subsets, such that $w_{ij} = 0$ whenever objects i and j are in different subsets. If the weight matrix is reducible, then the MDS problem can be decomposed in separate smaller multidimensional scaling problems, one for each subset.

A number of results in this paper appeared, in greater detail, in Groenen (1993), which also contains discussions of related topics in MDS, such as combinatorial methods for unidimensional scaling, structured missing data designs in MDS, cluster differences scaling, and the extension of the majorization method to Minkowski distances. The main differences with the work of Groenen (1993) on the problem of local minima are: the extension of tunneling to Minkowski distances, the use of $\sigma(\mathbf{X})/\eta_\delta^2$ instead of $\sigma(\mathbf{X})$, and the introduction of a pole width parameter in the tunneling function.

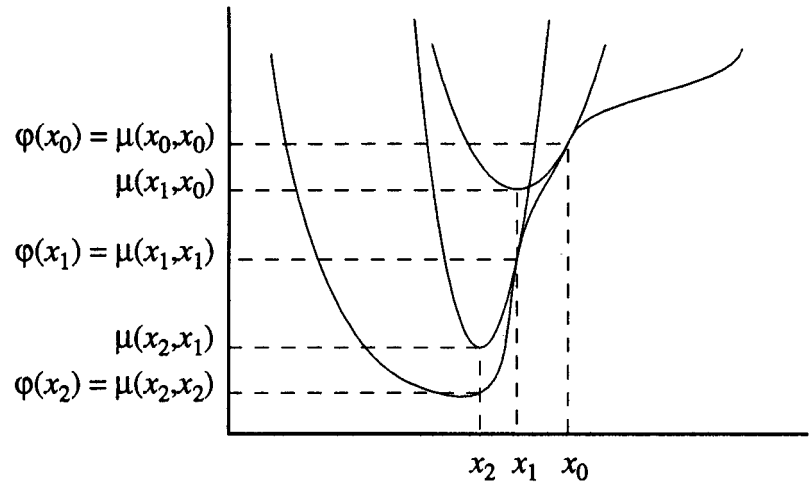


FIGURE 1.
Illustration of two steps of the iterative majorization method. The auxiliary function $\mu(x, x_0)$ is located above the original function $\varphi(x)$ and touches at the supporting point x_0. The minimum of the auxiliary function $\mu(x, x_0)$ is attained at x_1, where $\varphi(x_1)$ can never be larger than $\mu(x_1, x_0)$. These steps are repeated.

2. Local Minima for STRESS

In this section we introduce the principle of iterative majorization for minimizing a function. Next, the majorization algorithm SMACOF for MDS is discussed, which is needed to distinguish the two special cases of the STRESS function with respect to local minima, that is, unidimensional scaling ($p = 1$) and full-dimensional scaling ($p = n - 1$).

Iterative majorization is a simple and attractive minimization method that generates a monotonically nonincreasing sequence of function values (see in the area of MDS, for example, de Leeuw & Heiser, 1977, 1980; de Leeuw, 1988; Meulman, 1986, 1992; Groenen, Mathar, & Heiser, 1995; Groenen, 1993; and Heiser, 1995). Here, this method is applied to minimize the objective function of the tunneling method. Iterative majorization is based upon replacing the original complicated function $\varphi(\mathbf{x})$ by an auxiliary function $\mu(\mathbf{x}, \mathbf{y})$, which has to meet the following requirements. First, the auxiliary function $\mu(\mathbf{x}, \mathbf{y})$ should be more simple to minimize than $\varphi(\mathbf{x})$. Secondly, the original function should always be below the auxiliary function, that is, $\varphi(\mathbf{x}) \leq \mu(\mathbf{x}, \mathbf{y})$. Thirdly, the auxiliary function should touch the surface at the so-called *supporting point* $\mathbf{y}$, that is, $\varphi(\mathbf{y}) = \mu(\mathbf{y}, \mathbf{y})$. If these three requirements are met, we call $\mu(\mathbf{x}, \mathbf{y})$ a *majorizing* function of $\varphi(\mathbf{x})$. Let the minimum of $\mu(\mathbf{x}, \mathbf{y})$ over $\mathbf{x}$ be attained at $\mathbf{x}^*$, for $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x}^*$ in the corresponding domain X. The requirements of the majorizing function imply the chain of inequalities

$$\varphi(\mathbf{x}^*) \leq \mu(\mathbf{x}^*, \mathbf{y}) \leq \mu(\mathbf{y}, \mathbf{y}) = \varphi(\mathbf{y}) \tag{3}$$

for all $\mathbf{x}, \mathbf{y} \in X$. A graphical representation of these inequalities is presented in Figure 1 for two successive iterations of iterative majorization of some function $\varphi(x)$, where x is a scalar, and where the successive values of x are indicated by x_0, x_1, and x_2.

The majorization algorithm is summarized by

1. $\mathbf{y} \leftarrow \mathbf{y}_0$, where $\mathbf{y}_0$ is a starting value.
2. Find $\mathbf{x}^+$ for which $\mu(\mathbf{x}^+, \mathbf{y}) = \min_{\mathbf{x}} \mu(\mathbf{x}, \mathbf{y})$.
3. If $\varphi(\mathbf{y}) - \varphi(\mathbf{x}^+) < \varepsilon$ then stop (ε being a small positive constant).
4. $\mathbf{y} \leftarrow \mathbf{x}^+$ and go to 2.

By Step 2 the majorization algorithm yields a nonincreasing sequence of function values. If the function $\varphi(\mathbf{x})$ is not bounded from below, or if there are not sufficient restrictions on

$\mathbf{x}$, then the stop criterion of Step 3 may never be met. In the sequel, this situation does not arise, because the function that is minimized is bounded below by zero. A more relaxed version of the algorithm is obtained by demanding in Step 2 merely that $\mu(\mathbf{x}^+, \mathbf{y}) \leq \mu(\mathbf{y}, \mathbf{y})$, instead of requiring that $\mathbf{x}^+$ is the minimum of $\mu(\mathbf{x}, \mathbf{y})$ over $\mathbf{x}$. This weaker form of majorization does not change inequality (3), so that a reduction of function values is retained and allows us to combine majorizing functions. A sufficient condition for a stationary point is that $\varphi(\mathbf{x}^+) = \varphi(\mathbf{y})$ and $\mathbf{x}^+ = \mathbf{y}$. A necessary condition for a point $\mathbf{x}^*$ to be a minimizer of $\varphi(\mathbf{x})$ is that $\mathbf{x}^*$ minimizes $\mu(\mathbf{x}, \mathbf{x}^*)$ over $\mathbf{x}$. So, if $\varphi(\mathbf{x}^+) = \varphi(\mathbf{y})$ and $\mathbf{x}^+ = \mathbf{y}$, this necessary condition is satisfied by $\mathbf{y}$. We distinguish two particularly useful special cases of majorization. The first one is majorization of a concave function, which we call *linear* majorization. Any concave function $\varphi(\mathbf{x})$ may, by the nature of concavity, be majorized by the linear function $\mu(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{b}(\mathbf{y}) + c(\mathbf{y})$ at any point $\mathbf{y}$. The second group of functions is characterized by a uniformly bounded Hessian. In that case, the function $\varphi(\mathbf{x})$ can be majorized by $\mu(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{A}(\mathbf{y})\mathbf{x} - \mathbf{x}'\mathbf{b}(\mathbf{y}) + c(\mathbf{y})$, with $\mathbf{A}(\mathbf{y})$ positive semidefinite. A bounded Hessian $\nabla^2\varphi(\mathbf{x})$ implies that the curvature (that is, the second derivative) of the function $\varphi(\mathbf{x})$ is always less than the curvature of some quadratic function, i.e., $\mathbf{x}'\nabla^2\varphi(\mathbf{x})\mathbf{x} \leq \mathbf{x}'\mathbf{A}(\mathbf{y})\mathbf{x}$. Therefore, we call this type of majorization *quadratic* majorization. The distinction between linear and quadratic majorization is due to de Leeuw (1993) who used the terms Type I and Type II majorization.

Minimization by iterative majorization has several advantages. Using linear and quadratic majorization has the advantage that the majorizing function is at most quadratic in $\mathbf{x}$. In turn, this property implies a one-step update without a line search, which is often necessary in steepest descent algorithms. Moreover, quadratic majorizating functions can use the relaxed update of de Leeuw and Heiser (1980), which usually results in an acceleration (at best halving the number of iterations), while retaining convergence. Under certain circumstances, Groenen (1993) proved linear convergence rate of a quadratic majorization algorithm. An important practical advantage of using iterative majorization is its diagnostic value during program development: if the loss function increases, then there must be an error.

These preliminaries give us enough material to introduce the majorizing function for the general Minkowski STRESS function, as given by Groenen, Mathar, and Heiser (1995) for $1 \leq q \leq 2$, that is,

$$\eta_\delta^2\sigma(\mathbf{X}) = \eta_\delta^2 + \eta^2(\mathbf{X}) - 2\rho(\mathbf{X}) = \eta_\delta^2 + \sum_{s=1}^{p} \mathbf{x}_s'\mathbf{A}_s(\mathbf{T}, \mathbf{X})\mathbf{x}_s - 2\sum_{s=1}^{p} \mathbf{x}_s'\mathbf{B}_s(\boldsymbol{\Delta}, \mathbf{X})\mathbf{x}_s$$

$$\leq \eta_\delta^2 + \sum_{s=1}^{p} \mathbf{x}_s'\mathbf{A}_s(\mathbf{T}, \mathbf{Y})\mathbf{x}_s - 2\sum_{s=1}^{p} \mathbf{x}_s'\mathbf{B}_s(\boldsymbol{\Delta}, \mathbf{Y})\mathbf{y}_s = \eta_\delta^2\mu_\sigma(\mathbf{X}, \mathbf{Y}), \tag{4}$$

where $\mathbf{T}$ is a matrix with $t_{ij} = 1$ and $t_{ii} = 0$ (needed later on), $\mathbf{x}_s$ is column s of $\mathbf{X}$, and matrices $\mathbf{A}_s(\mathbf{T}, \mathbf{Y}) = \{a_{ijs}\}$and $\mathbf{B}_s(\boldsymbol{\Delta}, \mathbf{Y}) = \{b_{ijs}\}$ are defined by

$$a_{ijs} = \begin{cases} -w_{ij}t_{ij}|y_{is} - y_{js}|q - 2/\mathrm{d}_{ij}^{q-2}(\mathbf{Y}) & \text{if } i \neq j \text{ and } |y_{is} - y_{js}| \neq 0 \\ -w_{ij}t_{ij}\varepsilon/\mathrm{d}_{ij}^{q-2}(\mathbf{Y}) & \text{if } i \neq j \text{ and } |y_{is} - y_{js}| = 0 \\ -\sum_{k \neq i} a_{iks} & \text{if } i = j \end{cases}$$

$$b_{ijs} = \begin{cases} -w_{ij}\delta_{ij}|y_{is} - y_{js}|q - 2/\mathrm{d}_{ij}^{q-1}(\mathbf{Y}) & \text{if } i \neq j \text{ and } \mathrm{d}_{ij}(\mathbf{Y}) \neq 0 \text{ and } |y_{is} - y_{js}| \neq 0 \\ 0 & \text{if } i \neq j \text{ and } \mathrm{d}_{ij}(\mathbf{Y}) = 0 \text{ or } |y_{is} - y_{js}| = 0 \\ -\sum_{k \neq i} b_{iks} & \text{if } i = j \end{cases} \tag{5}$$

for ε a small positive constant. The majorizing function $\mu_\sigma(\mathbf{X}, \mathbf{Y})$ is a quadratic function in $\mathbf{X}$. The inequalities (3) can be proved using the Hölder's inequality; $\mu_\sigma(\mathbf{X}, \mathbf{Y})$ is minimized for every dimension s by the update

$$\mathbf{x}_s^+ = \mathbf{A}_s(\mathbf{T}, \mathbf{Y})^- \mathbf{B}_s(\boldsymbol{\Delta}, \mathbf{Y})\mathbf{y}_s, \tag{6}$$

where $\mathbf{A}_s(\mathbf{T}, \mathbf{Y})^-$ is the Moore-Penrose inverse of $\mathbf{A}_s(\mathbf{T}, \mathbf{Y})$. Note that in all practical cases $\mathbf{A}_s(\mathbf{T}, \mathbf{Y})$ is of rank $n - 1$ with $\mathbf{1}$ in its null space. Therefore, we use the Moore-Penrose inverse $\mathbf{A}_s(\mathbf{T}, \mathbf{Y})^- = (\mathbf{A}_s(\mathbf{T}, \mathbf{Y}) + \mathbf{11}')^{-1} - n^{-2}\mathbf{11}'$. For the Euclidean distance, this update was given by de Leeuw and Heiser (1977), who called it the *Guttman transform*. We refer to algorithm (6) by the acronym SMACOF, which stands for Scaling by MAjorizing a COmplicated Function.

Having presented the general majorizing algorithm for MDS, we now switch to the first special case of unidimensional scaling. For *unidimensional scaling* the Minkowski parameter q is irrelevant. It has been noted by, amongst others, de Leeuw and Heiser (1977), Defays (1978), and Hubert and Arabie (1986) that minimizing the STRESS function with equal weights changes to a combinatorial problem when $p = 1$. Groenen (1993) extended this result to unequal weights. In the following, the complicated proof of Groenen is not reproduced, but an indication is given why unidimensional scaling is a combinatorial problem.

For $p = 1$ the distance between two points can be expressed as $d_{ij}(\mathbf{x}) = (x_i - x_j)$ sign $(x_i - x_j)$, where sign $(x_i - x_j) = 1$ for $x_i > x_j$, sign $(x_i - x_j) = 0$ for $x_i = x_j$, and sign $(x_i - x_j) = -1$ for $x_i < x_j$, where $\mathbf{x}$ denotes the $n \times 1$ column vector of coordinates. An important observation is that only the rankorder of $\mathbf{x}$ determines the sign $(x_i - x_j)$. In this case, STRESS can be expressed as

$$\sigma(\mathbf{x}) = \eta_\delta^2 + \mathbf{x}'\mathbf{V}\mathbf{x} - 2\sum_{i<j} w_{ij}\delta_{ij}(x_i - x_j) \text{ sign } (x_i - x_j), \tag{7}$$

where $\mathbf{V}$ is a matrix with off-diagonal elements $v_{ij} = -w_{ij}$ and diagonal elements $v_{ii} = -\sum_{j \neq i} v_{ij}$. This shows that the cross-product term of STRESS, $\rho(\mathbf{x})$, can be factored into a part that is linear in $\mathbf{x}$ and a part that depends only on the rankorder of the elements of $\mathbf{x}$. Let ψ denote the rankorder of the vector $\mathbf{x}$, thus $x_{\psi(1)} \leq x_{\psi(2)} \leq \ldots \leq x_{\psi(i)} \leq \ldots \leq x_{\psi(n)}$ with corresponding permutation matrix $\mathbf{P}$, so that $\mathbf{Px}$ is the vector with the elements ordered nondecreasingly.

Define $t_i = \sum_{j<i} w_{\psi(i)\psi(j)}\delta_{\psi(i)\psi(j)}$ and $u_i = \sum_{j>i} w_{\psi(i)\psi(j)}\delta_{\psi(i)\psi(j)}$, which are, respectively, the row sum up to the main diagonal and the row sum from the main diagonal of the matrix with elements $w_{\psi(i)\psi(j)}\delta_{\psi(i)\psi(j)}$. Using this notation, (7) can be written as

$$\sigma(\mathbf{x}) = \eta_\delta^2 + \mathbf{x}'\mathbf{V}\mathbf{x} - 2\mathbf{x}'\mathbf{P}'(\mathbf{t} - \mathbf{u}). \tag{8}$$

For a *given* rankorder ψ, (8) is quadratic in $\mathbf{x}$ and has its minimum when $\mathbf{x}$ is equal to the Guttman transform $\mathbf{V}^-\mathbf{P}'(\mathbf{t} - \mathbf{u})$. The Guttman transform of the majorization approach only uses the rankorder information of the previous configuration, since $\mathbf{P}$, $\mathbf{t}$, and $\mathbf{u}$ only depend on the permutation of $\mathbf{x}$. Therefore, SMACOF stops if the rankorder of $\mathbf{x}$ does not change, which usually happens in a few iterations. At this point, STRESS has a local minimum. Function (8) can also be expressed as

$$\sigma(\mathbf{x}) = \eta_\delta^2 + \|\mathbf{x} - \mathbf{V}^-\mathbf{P}'(\mathbf{t} - \mathbf{u})\|_\mathbf{V}^2 - \|\mathbf{t} - \mathbf{u}\|_{\mathbf{PV}^-\mathbf{P}'}^2, \tag{9}$$

where the term $t(\psi) = \|\mathbf{t} - \mathbf{u}\|_{\mathbf{PV}-\mathbf{P}'}^2$ is a function of the permutation only. Groenen (1993) proved that if $t(\psi)$ has a local maximum with respect to adjacent pairwise interchanges (that is, any local change of ψ, interchanging $\psi(i)$ and $\psi(i + 1)$, does not increase the value of $t(\psi)$), then $\sigma(\mathbf{x})$ has a local minimum for $\mathbf{x} = \mathbf{V}^-\mathbf{P}'(\mathbf{t} - \mathbf{u})$. Note that this is a stronger

formulation for a local minimum than we used for STRESS, since STRESS has a local minimum whenever the Guttman transform cannot change the order of $\mathbf{x}$. For the maximization of $t(\psi)$ several combinatorial strategies are possible, of which, if all $w_{ij} = 1$, the dynamic programming approach of Hubert and Arabie (1986) gives a global maximum for n not too large. Other combinatorial strategies are discussed and compared in Groenen (1993).

A second special case of minimizing STRESS appears in *full-dimensional scaling* with Euclidean distances, where the dimensionality is at most $p = n - 1$. In full-dimensional scaling there is only one minimum that is a global one, which is a result due to de Leeuw (1993). Let the squared distance $d_{ij}^2(\mathbf{X})$ be expressed as $\text{tr } \mathbf{X}'\mathbf{A}_{ij}\mathbf{X}$, where $\mathbf{A}_{ij} = (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)'$ and $\mathbf{e}_i$ is column i of the identity matrix. Then $\text{tr } \mathbf{X}'\mathbf{A}_{ij}\mathbf{X} = \text{tr } \mathbf{A}_{ij}\mathbf{X}\mathbf{X}' = \text{tr } \mathbf{A}_{ij}\mathbf{H}$ with $\mathbf{H}$ a symmetric positive semidefinite matrix of rank $n - 1$ and with $\mathbf{H1} = \mathbf{0}$. This notation allows us to express STRESS as

$$\sigma(\mathbf{H}) = \eta_\delta^2 + \eta^2(\mathbf{H}) - 2\rho(\mathbf{H})$$

$$= \eta_\delta^2 + \sum_{i<j} w_{ij} \text{ tr } \mathbf{A}_{ij}\mathbf{H} - 2 \sum_{i<j} w_{ij}\delta_{ij}(\text{tr } \mathbf{A}_{ij}\mathbf{H})^{1/2}. \tag{10}$$

The term $\Sigma_{i<j} w_{ij} \text{ tr } \mathbf{A}_{ij}\mathbf{H}$ is a linear function in $\mathbf{H}$. The second term takes minus the square root of the same linear function of $\mathbf{H}$, hence is a convex function in $\mathbf{H}$. The sum of a linear and a convex function is convex, thus $\sigma(\mathbf{H})$ is a convex function in $\mathbf{H}$. $\mathbf{H}$ is restricted to be in the convex cone of symmetric positive semidefinite matrices of rank $n - 1$, with $\mathbf{1}$ in its null space. Thus minimizing STRESS over $\mathbf{H}$ is minimizing a convex function over a convex set, which has a local minimum that is global. An extension of this result to other MDS loss functions can be found in Groenen, de Leeuw and Mathar (1996). Note that this result does not hold in case $\mathbf{H}$ is restricted to have $p < n - 1$, because the set of $\mathbf{H}$ restricted to have rank $p < n - 1$ is not convex. Although one would expect $\mathbf{H}$ to be of rank $n - 1$ at a minimum, this usually is not the case. In fact, numerical experiments suggest that, at the minimum, the rank of $\mathbf{H}$ does not exceed the number of positive eigenvalues in classical scaling. Critchley (1986) and Bailey and Gower (1990) proved this conjecture for S-STRESS, but a proof for STRESS is not known.

In this section we have introduced iterative majorization to minimize complex functions, presented the SMACOF algorithm for MDS, shown that unidimensional scaling has many local minima and that full-dimensional scaling with Euclidean distances has one minimum. In the next section we introduce the tunneling method for finding better local minima in the general situation $1 \leq p \leq n - 1$.

3. The Tunneling Method

The tunneling method consists of an iterative two-step procedure: in the first step, a local minimum is sought, and in the second step, another configuration is determined with exactly the same STRESS. It can be described by the following analogy. Suppose we wish to find the lowest spot in a selected area in the Alps. First, we pour some water and see where it stops: the local search. From this point, a global search is performed by digging tunnels horizontally until we come out of the mountain. There we pour water again, find out where it stops, and dig tunnels again. If we stay underground for a long time while digging the tunnel, we simply conclude that the last spot was in fact the lowest place in the area, the candidate global minimum.

The tunneling method for functions of more parameters is mainly due to Montalvo (1979), Gomez and Levy (1982), and Levy and Gomez (1985). Earlier, Vilkov, Zhidkov, and Shchedrin (1975) presented a one-parameter tunneling function. An important and

attractive feature of the tunneling algorithm is that successive local minima always have lower or equal function values. The tunneling step is the crux of the tunneling method. It is performed by minimization of a particular function, called the *tunneling function*. To find another configuration with the same STRESS, this function must have several characteristics. Some of these characteristics are met by the tunneling function originally defined by Gomez and Levy:

$$\tau_0(\mathbf{X}) = \frac{\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)}{\|\mathbf{X} - \mathbf{X}^*\|^2}, \tag{11}$$

where $\mathbf{X}^*$ is the local minimum configuration. The first characteristic that (11) exhibits, is that it has zero points for configurations with STRESS equal to $\sigma(\mathbf{X}^*)$. Secondly, these zero points are not necessarily the lowest possible value of the tunneling function. Thirdly, the factor $\|\mathbf{X}^* - \mathbf{X}\|^2$, also called the *pole* of the tunneling function, is used to create elevated values of the tunneling function near the local minimum configuration $\mathbf{X}^*$, so that a zero point at $\mathbf{X}^*$ is presumably excluded. Because the pole is a factor, it does not change the positions of the zero points different from $\mathbf{X}^*$. Note that $\tau_0(\mathbf{X})$ is not defined at the local minimum $\mathbf{X}^*$, so that any minimization algorithm should start from somewhere else.

The complex task of finding the global minimum of the STRESS function has been replaced by a similarly complex problem, that is, finding the zero points of the tunneling function by minimization. Fortunately, the latter problem has the distinct advantage of having an additional feature for a desired local minimum of the tunneling function: it should have STRESS smaller than or equal to $\sigma(\mathbf{X}^*)$. Clearly, if $\mathbf{X}^*$ is a unique global minimum it will be impossible to find a zero point of $\tau_0(\mathbf{X})$. For the moment, however, we assume that $\mathbf{X}^*$ is not a unique global minimum, so that zero points of the tunneling function do exist.

The tunneling function (11) has some major defects, some of which are solely a consequence of certain properties of the STRESS function. Therefore, we need a redefinition of the tunneling function that resolves these defects. The first problem involves changes of the tunneling function value when a configuration is rotated. This is inconsistent with the STRESS function, which is invariant under rotation. Therefore, the tunneling function has to be invariant under rotation too. The second problem involves the strength of the pole. If the pole is not strong enough, then the tunneling function has an undesired zero point at $\mathbf{X}^*$. For a slightly different tunneling function, Groenen (1993) showed analytically that some pole strength parameter must be included. A similar analytical result can be obtained for the third problem that we will call *attraction to the horizon*. It turns out that we must be careful not to end up with very large configurations that also yield tunneling function values close to zero, due to the denominator in (11). Finally, we want to be able to vary the area influenced by the pole, through so called *pole width* ω. We now present a redefinition of the tunneling function that solves these problems and still exhibits the characteristics initially set out as desirable:

$$\begin{aligned}\tau_1(\mathbf{X}) &= \eta_\delta^{-2\lambda}|\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)|^\lambda\left(1 + \frac{\omega}{\|\mathbf{D}(\mathbf{X}^*) - \mathbf{D}(\mathbf{X})\|_{\mathbf{W}}^2}\right)\\ &= \eta_\delta^{-2\lambda}|\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)|^\lambda\left(\frac{\|\mathbf{D}(\mathbf{X}^*) - \mathbf{D}(\mathbf{X})\|_{\mathbf{W}}^2 + \omega}{\|\mathbf{D}(\mathbf{X}^*) - \mathbf{D}(\mathbf{X})\|_{\mathbf{W}}^2}\right),\end{aligned} \tag{12}$$

where $\|\mathbf{D}(\mathbf{X}^*) - \mathbf{D}(\mathbf{X})\|_{\mathbf{W}}^2 = \sum_{i<j} w_{ij}(d_{ij}(\mathbf{X}^*) - d_{ij}(\mathbf{X}))^2$. The function $\tau_1(\mathbf{X})$ satisfies all requirements: the zero points (that is, points with STRESS equal to $\sigma(\mathbf{X}^*)$) do not change after multiplication, a rotation of $\mathbf{X}$ yields the same tunneling function value as $\tau_1(\mathbf{X})$, the influence of the pole can be made strong enough by adjusting the *pole strength* parameter

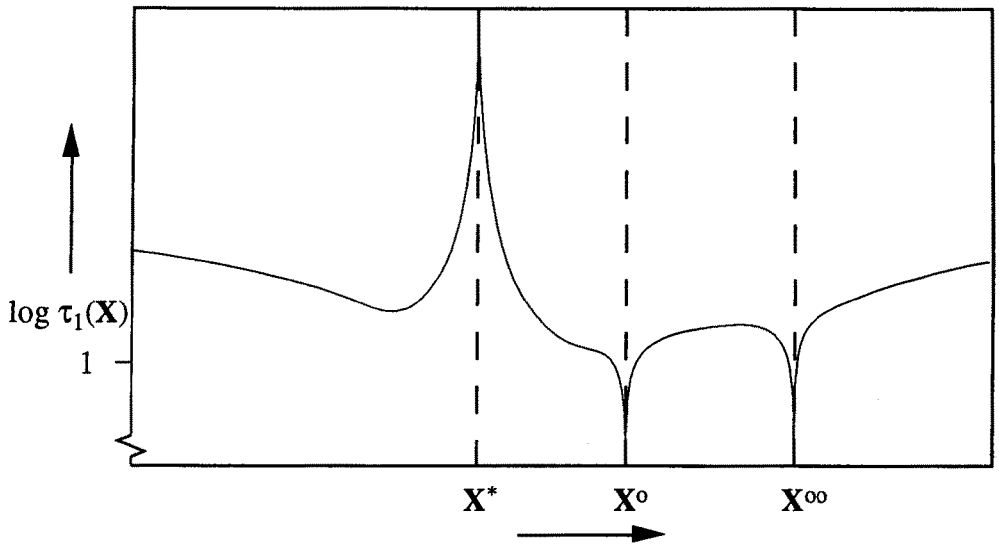


FIGURE 2.
A one-dimensional slice of the tunneling function $\tau_1(\mathbf{X})$.

λ ($0 < \lambda < 1$), for large configurations minimizing $\tau_1(\mathbf{X})$ amounts to minimizing $|\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)|^\lambda$ because $1 + \omega\|\mathbf{D}(\mathbf{X}^*) - \mathbf{D}(\mathbf{X})\|_w^{-2}$ tends to one so that attraction to the horizon is avoided, and the area for which the pole is active can be extended by choosing the pole width parameter ω greater than one. A one-dimensional slice of $\tau_1(\mathbf{X})$ is given in Figure 2. A summary of all the properties of (12) is given in Table 1, where still an additional feature is shown for the tunneling function: the occurrence of *multiple* poles. If the tunneling function stops at $\mathbf{X}_1$ where $\sigma(\mathbf{X}_1) > \sigma(\mathbf{X}^*)$ then we add an additional pole to the tunneling function in order to avoid the unwanted stationary point $\mathbf{X}_1$.

For minimizing the tunneling function we use two important methods: *iterative majorization* and *parametric programming*, which keeps tunneling within the majorization framework of SMACOF.

4. Minimizing the Tunneling Function

In the previous section a tunneling function was developed that has desirable properties. This section discusses a zero finding algorithm of this function. Since the zero points are also the global minima of the tunneling function, we can use a minimization algorithm to find a zero point. The minimization method used here is a combination of parametric programming and iterative majorization. First, we show that parametric programming remains valid if iterative majorization is applied. Then we decompose the tunneling function into parts that can be majorized and present the general majorizing inequalities. These results are then combined to present an algorithm for the tunneling function with Minkowski distances with $1 \leq q \leq 2$.

The tunneling function $\tau_1(\mathbf{X})$ can be considered as a ratio of two functions of $\mathbf{X}$ and hence minimization of $\tau_1(\mathbf{X})$ can be seen as a fractional programming problem. A parametric programming algorithm to minimize such a function was proposed by Dinkelbach (1967). Dinkelbach's algorithm requires at each iteration that the absolute minimum over $\mathbf{X}$ of an auxiliary function can be obtained. This requirement is too strong for our purpose; Groenen and Heiser (1991) proved that to find a stationary point it suffices to find a lower value of the auxiliary function, not necessarily the absolute minimum. This can be seen as follows. Let the tunneling function $\tau_1(\mathbf{X})$ with one pole be given by $\mathrm{N}(\mathbf{X})(\omega + \mathrm{P}(\mathbf{X}))/\mathrm{P}(\mathbf{X})$

Table 1

The elements of the tunneling function $\tau_1(\mathbf{X})$ and their purpose.

	Purpose	Element
1.	Zero point if STRESS is equal to local minimum STRESS.	$\tau(\mathbf{X}) = \sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)$
2.	Zero points are the lowest tunneling function values.	$\tau(\mathbf{X}) = \lvert\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)\rvert$
3.	Avoid a zero point at $\mathbf{X}^*$ by erecting a pole.	$\tau(\mathbf{X}) = \lvert\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)\rvert/P(\mathbf{X})$
4.	Avoid a zero point at irrelevant transformations of $\mathbf{X}^*$.	$P(\mathbf{X}) = \lVert\mathbf{D}(\mathbf{X}^*) - \mathbf{D}(\mathbf{X})\rVert_{\mathbf{W}}^2$
5.	Ensure sufficiently strong pole (use pole strength parameter λ, $0<\lambda<1$).	$\tau(\mathbf{X}) = \lvert\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)\rvert^\lambda/P(\mathbf{X})$
6.	Avoid attraction to the horizon.	$\tau(\mathbf{X}) = \lvert\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)\rvert^\lambda(1 + 1/P(\mathbf{X}))$
7.	Extend working range of the pole (for $\omega > 1$)	$\tau(\mathbf{X}) = \lvert\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)\rvert^\lambda(1 + \omega/P(\mathbf{X}))$
8.	Have multiple poles to avoid different $\mathbf{X}_k$ (optional)	$\tau_1(\mathbf{X}) = \lvert\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)\rvert^\lambda \prod_{k=1}^r (1 + \omega/P_k(\mathbf{X}))^{1/r}$

with $N(\mathbf{X}) = |\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)|^\lambda$, or, for ease of notation, $\tau_1(\mathbf{X}) = M(\mathbf{X})/P(\mathbf{X})$, where $M(\mathbf{X}) = N(\mathbf{X})(\omega + P(\mathbf{X}))$, where we assume that $P(\mathbf{X}) > 0$. Suppose that we can find an $\mathbf{X}$ for a given $\mathbf{Y}$ for which

$$\tau_1(\mathbf{X}) = \frac{M(\mathbf{X})}{P(\mathbf{X})} \leq \frac{M(\mathbf{Y})}{P(\mathbf{Y})} = \tau_1(\mathbf{Y}) \tag{13}$$

holds. Multiplying both sides with $P(\mathbf{X})$ and using $P(\mathbf{X}) > 0$ gives the inequality

$$M(\mathbf{X}) - \frac{M(\mathbf{Y})}{P(\mathbf{Y})} P(\mathbf{X}) \leq 0 \text{ or } F(q, \mathbf{X}) = M(\mathbf{X}) - qP(\mathbf{X}) \leq 0. \tag{14}$$

Thus, as long as we are able to find an $\mathbf{X}$ for which $F(q, \mathbf{X}) < 0$ then we must have $\tau_1(\mathbf{X}) < \tau_1(\mathbf{Y})$. Therefore, using iterative majorization on $F(q, \mathbf{X})$ minimizes $\tau_1(\mathbf{X})$.

Instead of operating on $\tau_1(\mathbf{X})$, we operate on $\tau_1^{1/2}(\mathbf{X})$ which has the same zero points and the same stationary points, but is easier to handle in the majorizing framework. The derivation of one iteration for minimizing $\tau_1^{1/2}(\mathbf{X})$ consists of the following steps. For each majorizing function we only give the terms that are dependent of $\mathbf{X}$.

1. Use the fractional programming formulation to obtain $F(q, \mathbf{X})$.
2. Majorize the terms of $F(q, \mathbf{X})$, that is, $M(\mathbf{X}) = N(\mathbf{X})^{1/2}(\prod_{k=1}^r (P_k(\mathbf{X}) + \omega)^{1/r})^{1/2}$ and $-(\prod_{k=1}^r P_k(\mathbf{X})^{1/r})^{1/2}$.
3. Majorize $M(\mathbf{X})$ by the product of two functions. This yields a term with $N(\mathbf{X})$ and a term with $\prod_{k=1}^r (P_k(\mathbf{X}) + \omega)^{1/r}$.

3a. Majorize $N(\mathbf{X}) = |\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)|^\lambda$ by majorizing the root of a nonnegative function. This yields a term with $|\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)|$.

3b. Simply assume that $\sigma(\mathbf{X}) > \sigma(\mathbf{X}^*)$ so that $|\sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)| = \sigma(\mathbf{X}) - \sigma(\mathbf{X}^*)$ and check this assumption for every update. If the assumption is violated, then the

tunneling step stops because we have found a configuration with the same or a lower STRESS different from $\mathbf{X}^*$, which is the main purpose of the tunneling step.

3c. Use the regular majorizing inequalities of (4) to majorize $\sigma(\mathbf{X})$.

4. Majorize $\Pi_{k=1}^{r} (P_k(\mathbf{X}) + \omega)^{1/r}$ by majorizing the product of several functions. This yields terms with $P_k(\mathbf{X})$.

4a. Majorize $P_k(\mathbf{X})$ using exactly the same majorizing inequality as in SMACOF, where the place of δ_{ij} is taken by $d_{ij}(\mathbf{X}_k)$.

5. Majorize $-(\Pi_{k=1}^{r} (P_k(\mathbf{X})^{1/2})^{1/r})$ by majorizing minus the product of functions. The majorizing function has terms $P_k(\mathbf{X})$ and $-P_k(\mathbf{X})^{1/2}$.

5a. Majorizing $P_k(\mathbf{X})$ is identical to Step 4a.

5b. Majorize $-P_k(\mathbf{X})^{1/2}$ with the Cauchy-Schwarz inequality. This yields terms with $d_{ij}(\mathbf{X})$ and terms with $-d_{ij}(\mathbf{X})$.

5c. Majorize $d_{ij}(\mathbf{X})$ by quadratic majorization (see Heiser, 1991), which yields a majorizing function with $d_{ij}^2(\mathbf{X})$.

5d. Majorize $d_{ij}^2(\mathbf{X})$ with Hölder's inequality (which underlies SMACOF), which yields a quadratic function in $\mathbf{X}$.

5e. Majorize $-d_{ij}(\mathbf{X})$ with Hölder's inequality (which underlies SMACOF), which yields a linear function in $\mathbf{X}$.

We continue with presenting the majorizing inequalities, some of which are new. Since they are general majorizing inequalities, their application reaches beyond the minimization of the tunneling function and can readily be used for majorization of other functions as well. For notational convenience we represent a function by the scalar x.

To majorize the product of two functions (used in step 3), we use the basic inequality

$$\left(\frac{x_1}{y_1} - \frac{x_2}{y_2}\right)^2 \geq 0, \tag{15}$$

which is always true for $y_1, y_2 > 0$, because any square of a real argument is nonnegative. Rewriting this inequality gives

$$x_1 x_2 \leq \frac{1}{2}\frac{y_2}{y_1} x_1^2 + \frac{1}{2}\frac{y_1}{y_2} x_2^2, \tag{16}$$

which is a form of quadratic majorization. Furthermore, the inequality becomes a strict equality when x_1 equals y_1 and x_2 equals y_2. Majorization in Step 5c can also be explained by (16) if we substitute $d_{ij}(\mathbf{X})$ for x_1 and 1 for x_2. The positivity assumption can be too strong in Step 5c, where $d_{ij}(\mathbf{Y}) = 0$ can happen so that the right part of (16) becomes undefined. For such cases we use the suggestion of Heiser (1991) who replaces y_1 by a small positive constant ε. Although strictly speaking the majorization requirement of touching the original function at $\mathbf{Y}$ is lost, this hardly poses a problem if ε is chosen small enough. This adaptation is used implicitly whenever needed.

In Step 3a the root of a positive value is majorized. Observe that the function $f(x) = x^\lambda$ is concave for $x \geq 0$ and $0 < \lambda < 1$ due to its negative second derivative, so that linear majorization can be applied. The inequality

$$x^\lambda \leq (1 - \lambda) + \lambda x \tag{17}$$

is used which becomes an equality if $x = 1$ (Hardy, Littlewood, & Pólya, 1952). Replacing x by x/y yields the linear majorization inequality

$$x^\lambda \leq (1 - \lambda) y^\lambda + \lambda y^{\lambda - 1} x. \tag{18}$$

To majorize the product of many functions (Step 4) the inequality for the geometric mean and the arithmetic mean,

$$\prod_{k=1}^{r} x_k^{1/r} \leq \frac{1}{r} \sum_{k=1}^{r} x_k, \tag{19}$$

can be used; see, for example, Hardy et al. (1952). This inequality assumes $x_k \geq 0$ and equality occurs when all x_k are unity or zero. For r equal to 2, (19) reduces to (16). Replacing x_k by x_k/y_k and multiplying both sides by $\Pi_{k=1}^{r} y_k^{1/r}$ yields the linear majorizing inequality

$$\prod_{k=1}^{r} x_k^{1/r} \leq \frac{1}{r} \prod_{k=1}^{r} y_k^{1/r} \sum_{k=1}^{r} \frac{x_k}{y_k}. \tag{20}$$

In Step 5 we have to majorize minus a product of r functions. The geometric-arithmetic mean inequality (19) is used again to find a majorization inequality that is quadratic in x_k. The basic chain of inequalities is

$$0 \leq \left(1 - \frac{1}{r}\right)(\mathbf{x} - \mathbf{1})'(\mathbf{x} - \mathbf{1})$$

$$0 \leq \left(1 - \frac{1}{r}\right)\mathbf{x}'\mathbf{x} - 2\left(1 - \frac{1}{r}\right)\mathbf{x}'\mathbf{1} + \left(1 - \frac{1}{r}\right)\mathbf{1}'\mathbf{1}$$

$$0 \leq \prod_{k=1}^{r} x_k^{1/r} + \left(1 - \frac{1}{r}\right)\mathbf{x}'\mathbf{x} - \left(2 - \frac{1}{r}\right)\mathbf{x}'\mathbf{1} + (r - 1)$$

$$-\prod_{k=1}^{r} x_k^{1/r} \leq \left(1 - \frac{1}{r}\right)\mathbf{x}'\mathbf{x} - \left(2 - \frac{1}{r}\right)\mathbf{x}'\mathbf{1} + (r - 1), \tag{21}$$

which is a strict equality if $\mathbf{x}$ equals $\mathbf{1}$. Replacing x_k by x_k/y_k and multiplying both sides by $\Pi_{k=1}^{r} y_k^{1/r}$ gives the quadratic majorizing inequality

$$-\prod_{k=1}^{r} x_k^{1/r} \leq \prod_{k=1}^{r} y_k^{1/r}\left(\left(1 - \frac{1}{r}\right)\sum_{k=1}^{r} \frac{x_k^2}{y_k^2} - \left(2 - \frac{1}{r}\right)\sum_{k=1}^{r} \frac{x_k}{y_k} + (r - 1)\right), \tag{22}$$

with equality if $x_k = y_k$ for all k.

In Steps 5b and 5d we need to majorize minus the root of the sums of squares which can be linearly majorized by the Cauchy-Schwarz inequality

$$-\left(\sum_{k=1}^{r} x_k^2\right)^{1/2} \leq -\left(\sum_{k=1}^{r} y_k^2\right)^{-1/2} \sum_{k=1}^{r} x_k y_k, \tag{23}$$

where equality occurs if $x_k = y_k$ for all k.

Combining all these majorizing inequalities is a major administrative task, which is summarized here. Combining the majorizing inequalities for steps 2 to 5, multiplying both sides by $e = 2\mathrm{N}(\mathbf{Y})^{-1/2} \Pi_{k=1}^{r} [1 + \mathrm{P}_k(\mathbf{Y})]^{-1/(2r)}$, and by using $q = \tau_1^{1/2}(\mathbf{Y})$ we get

$$
\begin{aligned}
eF(q, \mathbf{X}) \leq & \\
& \lambda N(\mathbf{Y})^{-1/\lambda}\eta_\delta^{-2} && \sum_{i<j} w_{ij}[d_{ij}^2(\mathbf{X}) - 2\delta_{ij}d_{ij}(\mathbf{X})] \\
& + \frac{1}{r}\sum_{k=1}^{r}(\omega + P_k(\mathbf{Y}))^{-1} && \sum_{i<j} w_{ij}[d_{ij}^2(\mathbf{X}) - 2d_{ij}(\mathbf{X}_k^*)d_{ij}(\mathbf{X})] \\
& + 2\left(1 - \frac{1}{r}\right)\sum_{k=1}^{r} P_k^{-1}(\mathbf{Y}) && \sum_{i<j} w_{ij}[d_{ij}^2(\mathbf{X}) - 2d_{ij}(\mathbf{X}_k^*)d_{ij}(\mathbf{X})] \\
& + \left(2 - \frac{1}{r}\right)\sum_{k=1}^{r} P_k^{-1}(\mathbf{Y}) && \sum_{i<j} w_{ij}[d_{ij}^2(\mathbf{X})d_{ij}(\mathbf{X}_k^*)/d_{ij}(\mathbf{Y}) - 2d_{ij}(\mathbf{Y})d_{ij}(\mathbf{X})] + c.
\end{aligned}
\tag{24}
$$

Let $\alpha = \lambda N(\mathbf{Y})^{-1/\lambda}\eta_\delta^{-2}$, $\beta_k = 1/r\,(\omega + P_k(\mathbf{Y}))^{-1}$, $\phi_k = 2(1 - 1/r)\,P_k^{-1}(\mathbf{Y})$, and $\phi_k = (2 - 1/r)\,\Sigma_{k=1}^{r} P_k^{-1}(\mathbf{Y})$. If we majorize additionally $d_{ij}^2(\mathbf{X})$ and $-d_{ij}(\mathbf{X})$ (Steps 5d and 5e), then $eF(q, \mathbf{X})$ is majorized by

$$
eF(q, \mathbf{X}) \leq \sum_{s=1}^{p}[\mathbf{x}_s'\mathbf{A}_s(\mathbf{T}, \mathbf{Y})\mathbf{x}_s - 2\mathbf{x}_s'\mathbf{B}_s(\mathbf{U}, \mathbf{Y})\mathbf{y}_s] + c, \tag{25}
$$

where $t_{ij} = \alpha + \Sigma_{k=1}^{r}[\beta_k + \phi_k + \phi_k d_{ij}(\mathbf{X}_k^*)/d_{ij}(\mathbf{Y})]$ and $u_{ij} = \alpha\delta_{ij} + \Sigma_{k=1}^{r}[(\beta_k + \phi_k)d_{ij}(\mathbf{X}_k^*) + \phi_k d_{ij}(\mathbf{Y})]$. The right side of (25) is quadratic in $\mathbf{X}$, so that an update can be obtained in one step. Setting the gradient with respect to $\mathbf{x}_s$ equal to zero yields

$$
\nabla\left(\sum_{s=1}^{p}[\mathbf{x}_s'\mathbf{A}_s(\mathbf{T}, \mathbf{Y})\mathbf{x}_s - 2\mathbf{x}_s'\mathbf{B}_s(\mathbf{U}, \mathbf{Y})\mathbf{y}_s] + c\right) = 2\mathbf{A}_s(\mathbf{T}, \mathbf{Y})\mathbf{x}_s - 2\mathbf{B}_s(\mathbf{U}, \mathbf{Y})\mathbf{y}_s = \mathbf{0}, \tag{26}
$$

which implies that $\mathbf{A}_s(\mathbf{T}, \mathbf{Y})\mathbf{x}_s = \mathbf{B}_s(\mathbf{U}, \mathbf{Y})\mathbf{y}_s$, so that the update for decreasing $\tau_1(\mathbf{X})$ is given by

$$
\mathbf{x}_s = \mathbf{A}_s(\mathbf{T}, \mathbf{Y})^{-}\mathbf{B}_s(\mathbf{U}, \mathbf{Y})\mathbf{y}_s, \tag{27}
$$

where $\mathbf{A}_s(\mathbf{T}, \mathbf{Y})^{-}$ is the Moore-Penrose inverse. Iterating (27) decreases $\tau_1(\mathbf{X})$ until a stationary $\mathbf{X}$ is obtained.

5. Fine Tuning of the Tunneling Algorithm

The tunneling method needs a pole strength parameter and a pole width parameter. However, we do not know a priori which pole strength and pole width is best. To see how the tunneling algorithm performs under various choices of pole strength and pole width, we do a small experiment where these factors are varied systematically. From this tuning experiment, we can get a feeling for usable values of these parameters.

The tuning experiment has three factors: two different datasets of dissimilarities, various values of the pole strength parameter λ and different values of the pole width ω. The tunneling algorithm is started from a local minimum configuration that is known to be higher than the global minimum. To be more precise, for each combination of λ and ω the same starting configuration is used, which is the sum of the local minimum and a small random perturbation that remains the same over this experiment. To rate the efficiency of the tunneling algorithm we compare the number of iterations needed to end the tunneling step. If it takes more than 1000 iterations we stop the tunneling step and regard it as having failed.

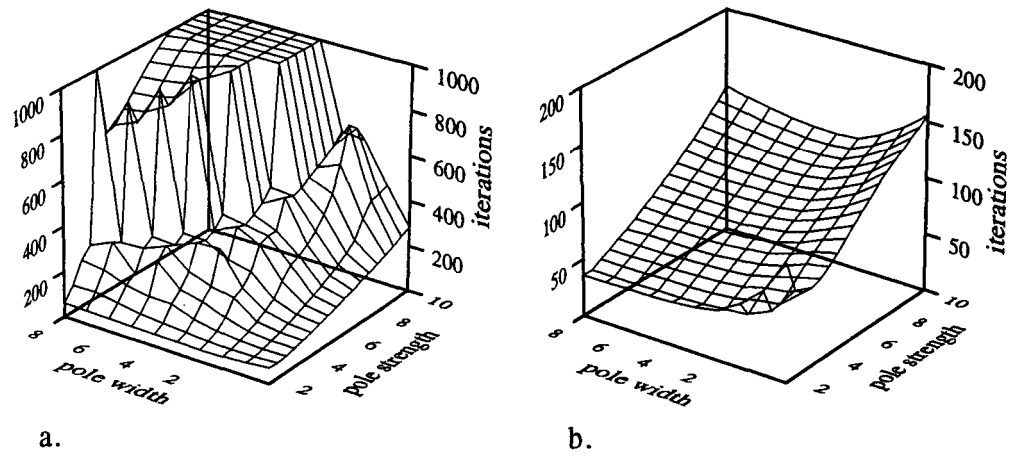


FIGURE 3.
Number of iterations needed to finish the tunneling step for different values of the pole strength parameter $1/\lambda$ and different normalizations of the dissimilarities η^2_δ. For every combination the tunneling step was stopped if 1000 iterations was exceeded. The dissimilarities in plot a. stem from the Mani collection, those in plot b. from a regular two-dimensional grid of 9 points.

The first example consists of data that originate from Robinson (1951), and were also analyzed by Hubert and Arabie (1986); the data come from the Mani collection of 8 archaeological deposits. The second data set consists of distances between 9 points in a regular two dimensional grid. We are focusing on the tunneling step that moves away from the local minimum configuration with STRESS .01220370 for the Mani collection (global minimum STRESS .00654532), and a local minimum with STRESS .05670459 for the regular grid data (global minimum STRESS 0). The pole strength parameter was varied in 18 steps from 1/1.5, 1/2, 1/2.5, . . . , 1/10. The pole width was varied from .5, 1, 2, 3, 4, 5, 6, 7, and 8. A three-dimensional graph of the results of both data sets is shown in Figure 3. In the plot the number of iterations needed to stop the tunneling step are given against different values of $1/\lambda$ and ω. We see that a large pole width and a large pole strength parameter increases the number of iterations needed to finish the tunneling step. For the Mani collection a pole strength value larger than 1/5 and a pole width larger than 4 results in many iterations in the tunneling step or even failure. In the grid example we see that a small pole strength and pole width performs not very good. These examples suggest the use of a moderate pole strength parameter, like 1/3, and a pole width of about $n/4$.

6. An Experiment on Local Minima

In this section we report a simulation experiment on the seriousness of the local minimum problem for dissimilarity matrices, using a known underlying configuration, for different dimensionality p, number of objects n, amounts of perturbation, and for Euclidean distances and city-block distances. Our hypothesis for Euclidean distances is that as p gets larger (with respect to n) less local minima occur. This hypothesis is based on the fact that for $p = 1$ many local minima occur, and for $p = n - 1$ no local minima occur. We also expect that the number of local minima grows as the amount of error imposed on the dissimilarities increases. For city-block distances we expect many local minima (Heiser, 1989; Hubert, Arabie, & Hesson-McInnis, 1992).

The simulation experiment was performed as follows. For the Euclidean distance, we generated a random configuration matrix for each combination of the following three factors:

1. size of n, that is, 5, 10, 20, and 40,
2. different values for the dimensionality p, that is, 1, 2, 3, 5, and 10,
3. different amount of error on the distance matrix, that is, 0%, 50%, 100%, and 250%.

The distance matrix of the random configuration matrix was perturbed by error and used as a dissimilarity matrix, where the amount of error varies according to the levels of Factor 3 above. We call such a dissimilarity matrix a gauge in the sequel. For each gauge, the SMACOF algorithm is started a hundred times from random $n \times p$ start configurations. We simply registered for each gauge which configuration yielded the lowest STRESS, and how often a local search ended in this candidate global minimum. In this way, we get an idea of the region of attraction of the candidate global minimum. Clearly, if no error is imposed the lowest local minimum has zero STRESS and hence is the global minimum. For city-block distances we used the same procedure for generating gauges, except that n was either 5 or 10.

We used multiplicative error from a log-normal distribution as proposed by Wagenaar and Padmos (1971) to perturb the distance matrix to obtain a gauge. Thus, true distances are generated and subsequently multiplied by errors of a log-normal distribution, which are positive so that no negative dissimilarities in our gauge occur. Other error distributions (like the normal distribution with zero mean as used by De Soete, Hubert & Arabie, 1988) can have the disadvantage of possibly introducing negative dissimilarities. Although the algorithm could be adapted to deal with negative dissimilarities (see Heiser, 1991), we do not wish to introduce this latent factor in our simulation study. The construction of the gauges can be summarized as follows. Given the dimensionality p and the number of objects n, a configuration matrix is constructed with uniformly distributed random coordinates in the interval [0, 1). The $a\%$ error gauge is formed by multiplying each dissimilarity by $e^{a\varepsilon/100}$, where ε is standard normally distributed error (mean 0, variance 1).

To find the number of different local minima we need to compare the local minimum configurations with each other. The comparison is complicated by the freedom of rotation. To decide whether two configurations are different, we compare two matrices of distances with the Tucker's congruence coefficient

$$\frac{\sum_{i<j} w_{ij} d_{ij}(\mathbf{X}_1) d_{ij}(\mathbf{X}_2)}{(\sum_{i<j} w_{ij} d_{ij}^2(\mathbf{X}_1))^{1/2} (\sum_{i<j} w_{ij} d_{ij}^2(\mathbf{X}_2))^{1/2}} \tag{28}$$

rather than the coordinates. Here, we regard two configurations equal if (28) is larger than $1 - 10^{-7}$.

The results of the experiment are presented in Tables 2 and 3. We have to interpret these tables with some care, since the experiment is limited in size, compares only two datasets, and may depend on the accuracy chosen. The most striking fact from Table 2 is that for most combinations all 100 random start configurations ended in the candidate global minimum. Apparently, the SMACOF algorithm is quite capable in reaching the global minimum. Also, we find many different local minima for $p = 1$ and only one for $p = 10$. This experiment lends support to the hypothesis that the local minimum problem is more severe for small dimensionality ($p = 2$, or $p = 3$) than for a larger dimensionality. As expected for the full-dimensional combinations of p and n ($n = 5, p = 5$; $n = 5, p = 10$; $n = 10, p = 10$) with Euclidean distances all searches led to the same minimum. For Euclidean distances there is a slight indication that the number of local minima increases as the error increases. For city-block distances many local minima occur over all combinations. It seems that the more objects there are, the more severe the local minimum problem. As the error increases, the number of local minima seem to decrease.

Table 2

Results of a simulation study on the number of local minima in MDS with Euclidean distances. Reported are the number of different local minima found after 100 random starts, and, in parentheses, the number of starts that ended in the lowest local minimum.

Dimensionality	Number of objects	Error level 0 %		50 %		100 %		250%	
1	5	36	(29)	10	(19)	4	(37)	1	(100)
	10	67	(32)	92	(2)	75	(2)	16	(7)
	20	65	(32)	100	(1)	95	(1)	86	(3)
	40	77	(16)	76	(5)	100	(1)	100	(1)
2	5	1	(100)	1	(100)	1	(100)	1	(100)
	10	7	(62)	1	(100)	1	(100)	1	(100)
	20	4	(92)	4	(78)	5	(66)	1	(100)
	40	7	(93)	7	(90)	26	(45)	18	(3)
3	5	2	(67)	1	(100)	1	(100)	1	(100)
	10	1	(100)	1	(100)	1	(100)	1	(100)
	20	1	(100)	1	(100)	1	(100)	1	(100)
	40	1	(100)	8	(38)	5	(63)	2	(56)
5	5	1	(100)	1	(100)	1	(100)	1	(100)
	10	1	(100)	1	(100)	1	(100)	1	(100)
	20	1	(100)	1	(100)	1	(100)	2	(99)
	40	1	(100)	4	(67)	1	(100)	1	(100)
10	5	1	(100)	1	(100)	1	(100)	1	(100)
	10	1	(100)	1	(100)	1	(100)	1	(100)
	20	1	(100)	1	(100)	1	(100)	1	(100)
	40	1	(100)	1	(100)	1	(100)	1	(100)

7. Performance of the Tunneling Method

In order to evaluate the tunneling method we consider four case studies: one artificial data set, and three empirical data sets. The tunneling method is compared against the multiple random start approach, or, in short, *multistart*. Multistart is a simple and widely used method to find a global minimum. It repeatedly starts with a random configuration, and retains the best out of k trials. Moreover, multistart is guaranteed to find a global minimum, albeit after infinitely many random starts, which is its weakness. Furthermore, multistart gives an indication of the *region of attraction* of a (global) minimum, which is the space from which all searches lead to the same minimum.

Case 1

The first small artificial example, studied extensively by de Leeuw (1988), concerns a 4×4 dissimilarity matrix, with all dissimilarities equal to $1/\sqrt{6}$. De Leeuw reports three stationary two-dimensional configurations $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$: $\mathbf{X}_1$ has four points equally spaced on a line with $\sigma(\mathbf{X}_1) = 0.16666667$, $\mathbf{X}_2$ has three points in the corners of an equilateral triangle and a point in the centroid with $\sigma(\mathbf{X}_2) = .06698730$, and $\mathbf{X}_3$ has four

Table 3

Results of the simulation study on the number of local minima in MDS with city-block distances. Reported are the number of different local minima found after 100 random starts, and, in parentheses, the number of starts that ended in the lowest local minimum.

Dimen-sionality	Number of objects	Error level 0 %		50 %		100 %		250%	
1	5	28	(31)	12	(16)	6	(30)	2	(92)
	10	76	(16)	89	(1)	47	(17)	20	(11)
2	5	71	(1)	31	(15)	21	(23)	3	(71)
	10	99	(1)	100	(1)	87	(2)	81	(1)
3	5	66	(1)	19	(32)	5	(70)	3	(77)
	10	100	(1)	99	(1)	100	(1)	59	(1)
5	5	72	(1)	5	(94)	6	(16)	3	(98)
	10	100	(1)	100	(1)	100	(1)	31	(20)
10	5	17	(17)	4	(97)	2	(99)	1	(100)

points in the corners of a square with $\sigma(\mathbf{X}_3) = .02859548$. Other stationary points, which are in fact saddle points, can be obtained by imposing objects on top of each other, so that they have zero distance.

We start the tunneling algorithm from $\mathbf{X}_1$, which is a stationary point for unidimensional scaling and a saddle point in two dimensions. The first objective for the tunneling algorithm is to find another configuration with a STRESS of 0.16666667, or lower. The tunneling algorithm was initialized with a starting configuration that is a sum of the unidimensional scaling solution and a random matrix. The latter is necessary, because we have to start tunneling from a different point than $\mathbf{X}_1$ (since $\tau_1(\mathbf{X}_1)$ is undefined) and because we have to increase the rank of the solution from one to two. The pole strength parameter λ was set to 1/3. After 45 iterations a solution was found with STRESS 0.12337749, which is lower than $\sigma(\mathbf{X}_1)$. A local search ended after 30 iterations in the square configuration $\mathbf{X}_3$, which is a global minimum of this example. Repeating the tunneling step results in finding a different square, which is a permutation (and/or rotation) of the square $\mathbf{X}_1$.

Case 2

Here the performance of the tunneling method is studied for the data gathered by Funk, Horowitz, Lipshitz, and Young (1974), also analyzed in Heiser (1991). They studied the perceived differences among thirteen ethnic subgroups of the American culture; 49 students had to rate the difference between all pairs of ethnic subgroups on a nine-point rating scale (1 = very similar, 9 = very different). The data consist of the average dissimilarity among those 49 respondents.

Multistart with 1000 random starts yielded about 850 different local minima. The histogram in Figure 4 shows the distribution of the STRESS of the local minima. The distribution is bimodal with the lowest STRESS value .06025. Starting from the worst local minimum with STRESS 0.08603658 and with $\lambda = 1/3$ and $\omega = 10$, the tunneling algorithm found a candidate global minimum with STRESS 0.06130282 in nine steps, which is larger

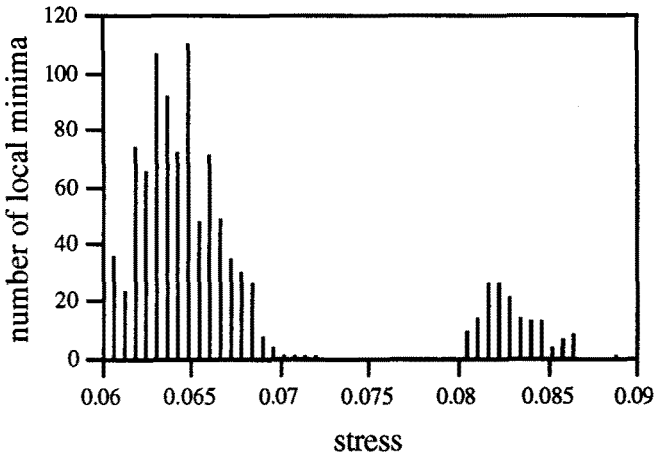


FIGURE 4.
Histogram of the number of local minima of the Funk data.

than the one found by multistart. With $\lambda = 1/3$ and $\omega = 13$ the tunneling method ended with a candidate global minimum with STRESS 0.06065357976. With $\lambda = 1/3$ and $\omega = 2$ the tunneling method ended with a candidate global minimum with STRESS 0.08146279. We found that the tunneling method finds better local minima easily in the early stages. However, in the final stage the tunneling step takes more time or may even fail. Occasionally, the tunneling step is finished rapidly, leading to a slightly better local minimum. For such cases it seems that the local search ended too early, which is repaired by the next tunneling step. This example with many local minima shows that the tunneling method behaves quite differently for various values of the pole width. Here, the tunneling method is able to find a local minimum close, but not equal to the global minimum.

Case 3

Here the performance of the tunneling method for city-block distances is studied. The data used are reported by Green, Carmone, and Smith (1989, Appendix B, p. 382 ff.). 38 students judged each pair out of 10 different cola brands on their similarity on a nine point rating scale (1 = not similar at all, 9 = very similar). The dissimilarities were accumulated over the subjects and are reported in Groenen, Mathar, and Heiser (1995).

Multistart with 100 random starts yielded 100 different local minima, which confirms the severe local minimum problem for city-block distances. The worst local minimum found by multistart has STRESS 0.21343222, and the best has STRESS 0.04042472. Starting the tunneling method with $\omega = 10$ and $\lambda = 1/3$ from the worst local minimum yielded a candidate global minimum at 0.14688168 in 9 tunneling steps. Redoing the tunneling method with ω increased to 100 yielded a candidate global minimum STRESS of 0.07095518 in 5 tunneling steps, whereas $\omega = 1000$ led to a STRESS of 0.03531523 in 6 tunneling steps. This result shows that the tunneling method is able to outperform multistart, but its success depends on the pole width parameter ω. Apparently, for MDS with city-block distances the local minimum problem is very severe.

Case 4

The second example with a real data set consists of citations between journals in the psychometric literature. Mapping of citation matrices has been applied by, amongst others, Weeks and Bentler (1982) and Tijssen (1992). The basic idea is that articles in a scientific journal in a specific field tend to cite articles in other journals in the same field. In this way

a map can be obtained to find groups of journals that are cognitively linked. The citation matrix between 28 journals is given in Table 4. Since the entries in Table 4 are similarity measures we need to transform them into dissimilarities. We use the gravity model

$$d_{ij}^2(\mathbf{X}) = \frac{m_{i+}m_{+j}}{n_{ij}}, \tag{29}$$

with n_{ij} the number of citations between journal i and j, and m_{i+} the sum of the citations to journal i. Equation (29) models the gravitational force between large masses in physics, like moon and earth, to their distance. Thus, the force n_{ij} between two journals is inversely related to the square of their distance, given their masses m_{i+} and m_{+j}. These masses correct for the total number of citations of a journal. Clearly, if a journal is cited very often the probability of another journal citing the first one is larger. This model has been used in a similar context by Zielman (1991), Groenen (1993) and Heiser and Groenen (1994). Model (29) translates easily into the framework of STRESS as

$$\sigma^2(\mathbf{X}) = \sum_{i<j} w_{ij}\left(\left(\frac{m_{i+}m_{+j}}{n_{ij}}\right)^{1/2} - d_{ij}(\mathbf{X})\right)^2. \tag{30}$$

Setting $\delta_{ij} = (m_{i+}m_{+j}/n_{ij})^{1/2}$ translates the gravity model back to STRESS. If n_{ij} is zero, δ_{ij} is not defined and w_{ij} is set to zero, otherwise w_{ij} is set to one. Table 4 gives the number of citations in journal i to journal j which may differ from the number of citations in journal j to journal i. Here, we are interested in the symmetric part of the data only, so Table 4 is symmetrized before applying the gravity model. Self-citations n_{ii} were excluded, since they do not contribute to the structure of citations among journals.

To assess the local minimum problem for the citation data we did multistart with 100 random starts and compared their solutions. We obtained 11 different local minima of which the lowest (σ = .13532209) attracted 9% and the worst (σ = .14316019) attracted 1%. The nine best local minima are given in Figure 5 and the lowest local minimum in Figure 6. To compare the local minima, Psychometrika is placed in the origin, "Perc Mot Sk" positive on the first axis and zero on the second axis, and "Appl Psyc M" free on the first axis and positive on the second axis. In all solutions, the statistical journals are located on the left hand side, whereas the more psychologically oriented journals are located on the right-hand side, with Psychometrika in between. Vertically, there is an important difference between journals in the field of educational measurement near the top, and the area of mathematical psychology at the bottom. The differences between the local minima are small and mostly concern interchanging the order within the groups $\{e, f, k\}$, $\{s, h\}$ and $\{j, 2\}$.

The tunneling algorithm was started from the worst local minimum and yielded after one tunneling step Figure 5b. Further tunneling did not yield the best configuration found by multistart. The reason for the tunneling method to fail could be that the difference in STRESS values is too small, so that the tunneling function is not steep enough for leading the configuration towards the candidate global minimum.

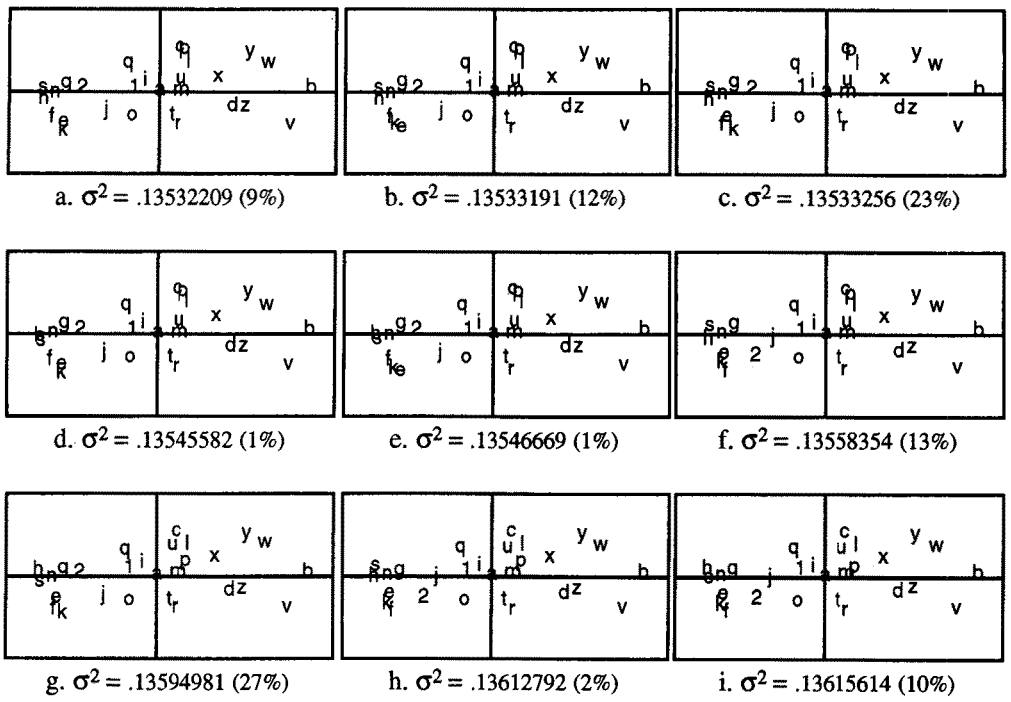
8. Conclusions and Discussion

This paper has presented several results on the local minimum problem for MDS. Four cases can be distinguished: (a) unidimensional scaling for which the local minimum problem is particularly severe, (b) full-dimensional scaling with Euclidean distances which has no local minimum problem, (c) MDS with dimensionality from 2 to $n - 2$ using Euclidean distances for which our numerical experiment suggests that the number of local minima decreases with increasing dimensionality, and (d) MDS with city-block distances for which many local minima were found. Next, the tunneling method was introduced,

Table 4

Citation matrix between journals in the psychometric literature summed over 1991 and 1992. The rows are the cited journals and the columns are the citing journals. Reproduced with kind permission of CWTS, Leiden University.

	journal	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	1	2
a	Psychometri	383	38	161	45	9	4	10	4	181	28	11	80	26	28	60	48	51	21	11	25	146	21	59	46	41	10	33	38
b	Perc Mot Sk	1	902	0	1	0	0	0	0	0	0	0	8	3	0	0	0	0	3	0	0	1	75	80	96	83	17	0	0
c	Appl Psyc M	32	1	94	2	0	0	0	0	26	0	0	49	6	2	0	51	7	0	0	2	20	0	16	20	8	0	4	1
d	Ann R Psych	9	15	1	82	0	0	0	0	3	0	0	8	1	0	1	2	1	14	0	0	3	25	17	52	21	27	1	0
e	Ann Statist	30	0	2	2	943	36	93	171	11	13	90	5	0	422	6	0	6	10	126	4	15	0	0	7	0	0	12	17
f	Appl Stat	17	0	1	2	14	144	96	32	10	11	42	2	0	72	6	0	2	4	30	1	2	0	1	1	1	0	3	45
g	Biometrics	24	4	8	8	32	35	641	121	18	7	63	6	9	203	10	5	7	2	51	6	4	1	3	11	3	0	5	42
h	Biometrika	65	2	6	4	222	86	416	470	34	26	96	21	5	358	10	1	33	7	185	7	19	1	2	33	3	5	9	83
i	Br J Math S	23	1	20	10	1	2	1	1	84	2	2	21	9	5	1	6	14	5	1	9	33	8	12	22	0	2	20	7
j	Chem Intell	0	0	0	0	0	1	0	0	0	195	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
k	Comput Stat	15	0	1	1	2	2	4	3	3	0	17	0	0	4	4	0	1	0	8	0	1	0	0	1	0	0	0	2
l	Educ Psyc M	6	37	25	8	0	0	1	0	13	0	2	205	7	0	2	20	6	1	0	0	37	0	56	20	82	0	6	0
m	Exp Aging R	0	3	0	12	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	1	2	3	6	2	0	0	0
n	J Am Stat A	76	4	12	6	325	87	367	261	48	33	154	26	10	793	17	5	41	7	142	3	26	2	1	48	4	6	41	92
o	J Classif	14	0	0	52	0	0	1	1	1	0	3	0	0	1	35	0	0	0	0	2	2	0	0	1	0	0	2	0
p	J Educ Meas	20	6	47	1	0	0	1	0	7	0	0	33	0	1	1	103	16	0	0	0	2	0	3	6	12	1	0	2
q	J Educ Stat	9	1	6	0	1	0	1	0	10	0	0	3	1	5	0	7	12	0	0	0	1	0	0	2	0	0	1	0
r	J Math Psyc	32	5	11	19	3	1	1	0	10	0	0	1	0	0	7	0	1	143	0	95	5	24	1	10	0	26	0	3
s	J Roy Sta B	30	0	6	1	131	63	163	200	26	18	71	0	0	250	8	2	15	8	204	2	5	1	0	11	0	1	6	32
t	Math Soc Sc	0	0	0	6	0	0	0	0	1	0	0	0	0	0	1	0	0	13	0	45	0	0	0	1	0	1	0	0
u	Multiv Be R	12	10	17	5	0	0	1	0	33	0	0	24	14	1	0	0	2	0	0	0	104	0	55	18	21	1	21	2
v	Perc Psych	16	110	3	84	0	0	0	0	4	0	0	0	0	0	1	0	1	35	0	13	0	1229	12	73	2	55	0	0
w	Pers Indiv	0	21	0	10	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	7	0	857	20	73	0	0	0
x	Psychol B	13	139	43	87	1	0	3	0	60	0	3	126	45	9	3	16	39	25	0	2	115	71	152	492	146	73	47	4
y	Psychol Rep	0	139	6	8	0	0	1	0	0	0	0	29	2	0	0	1	1	0	0	0	8	0	161	83	628	2	0	0
z	Psychol Rev	28	118	13	167	0	0	0	0	7	3	0	9	8	2	5	4	0	106	1	29	8	271	60	231	74	266	4	1
1	Sociol Meth	4	0	1	0	0	0	0	0	35	0	0	4	5	7	0	0	3	0	0	0	10	0	1	8	1	0	40	6
2	Statisticia	1	0	0	0	3	8	11	5	5	1	2	0	0	7	0	0	1	0	9	0	3	0	0	0	0	0	0	33



a. $\sigma^2 = .13532209$ (9%) b. $\sigma^2 = .13533191$ (12%) c. $\sigma^2 = .13533256$ (23%)

d. $\sigma^2 = .13545582$ (1%) e. $\sigma^2 = .13546669$ (1%) f. $\sigma^2 = .13558354$ (13%)

g. $\sigma^2 = .13594981$ (27%) h. $\sigma^2 = .13612792$ (2%) i. $\sigma^2 = .13615614$ (10%)

FIGURE 5.
The nine best local minima out of eleven obtained with multistart on the citation data. In parentheses the percentage of starts ending in that local minimum is given.

which aims at finding an ever decreasing series of local minima. The method iterates between a local search and a tunneling step in which a configuration is sought with equal or less STRESS than the previous local minimum. The proposed tunneling function was adapted partly to make it suitable for MDS and partly to make it better behaved. Several problems have been solved so that the tunneling function does not lead to trivial solutions. The function can handle more than one pole, which is needed to avoid unwanted stationary points in the tunneling function. Furthermore, we presented a minimization algorithm

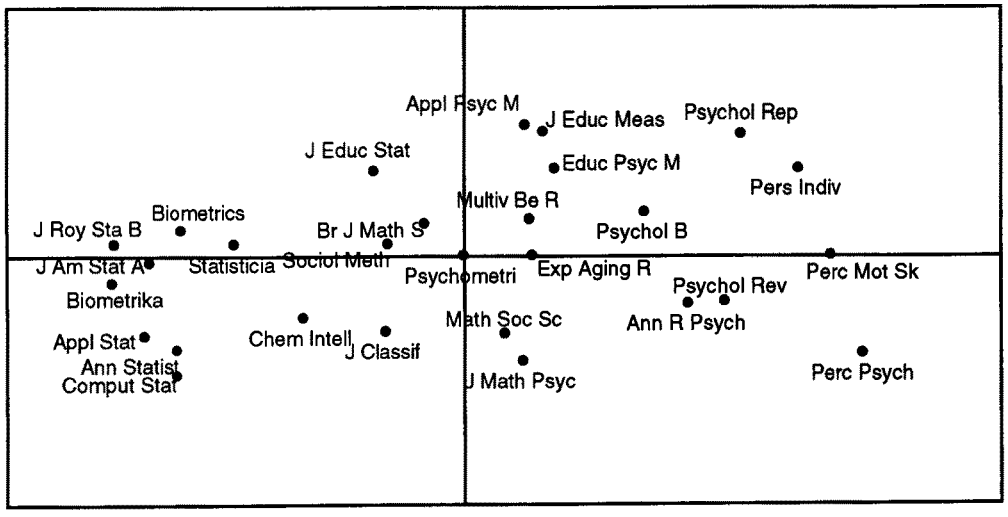


FIGURE 6.
The candidate global minimum of the citation data. This scatter plot is the same as Figure 5a.

for the tunneling function based on iterative majorization and we extended the parametric programming algorithm to majorization. Several new majorization inequalities were found. A tuning study showed that a pole strength parameter of 1/2 or 1/3 and the pole width parameter of $n/4$ yields satisfactory results.

In four case studies tunneling was compared with multistart. It turns out that tunneling performs adequately in most cases. However, sometimes the tunneling method is not able to find a better minimum than multistart. We remark that the tunneling method is able to find better local minima systematically. Therefore, it is a serious contender of multistart for finding a global minimum. However, the method is not guaranteed to find a global minimum and may stop too early. The effectiveness and speed of the tunneling method is greatly influenced by the pole width parameter and the pole strength parameter. For some combinations, we obtain good results, for others the method is slow or fails. It seems that the performance of the tunneling method is related to the dataset itself; if the region of attraction of a global minimum is large, then the tunneling method finds the global minimum relatively easily.

In this paper we have limited ourselves to metric MDS. Without any complication the tunneling method could be adapted for nonmetric MDS, where a step is included for finding optimal disparities. However, this changes the minimization problem of STRESS and influences the occurrence of local minima. Therefore, a separate study for local minima in nonmetric MDS is needed, where the performance of the tunneling method needs to be reconsidered.

References

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

Bailey, R. A., & Gower, J. C. (1990). Approximating a symmetric matrix. *Psychometrika, 55,* 665–675.

Commandeur, J. J. F. (1992). *Missing data in the distance approach to Principal Component Analysis* (Research Rep. No. RR-92-07). Leiden: Department of Data Theory.

Critchley, F. (1986). Dimensionality theorems in MDS and HCA. In E. Diday et al. (Eds), *Data analysis and informatics, Vol. 4* (pp. 45–70). Amsterdam: North-Holland.

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

de Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In J. R. Barra, F. Brodeau, G. Romier, & B. van Cutsem (Eds.), *Recent developments in statistics* (pp. 133–145). Amsterdam: North-Holland.

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.* Unpublished manuscript.

de Leeuw, J., & Heiser, W. J. (1977). Convergence of correction matrix algorithms for multidimensional scaling. In J. C. Lingoes, E. Roskam, & I. Borg (Eds.), *Geometric representations of relational data* (pp. 735–752). Ann Arbor: Mathesis Press.

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

De Soete, G., Hubert, L., & Arabie, P. (1988). On the use of simulated annealing for combinatorial data analysis. In W. Gaul & M. Schader (Eds.), *Data, expert, knowledge and decisions* (pp. 329–340). Berlin: Springer-Verlag.

Dinkelbach, W. (1967). On nonlinear fractional programming. *Management Science, 13,* 492–498.

Funk, S. G., Horowitz, A. D., Lipshitz, R., & 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.

Gomez, S., & Levy, A. V. (1982). The tunneling method for solving the constrained global optimization problem with non-connected feasible regions. In J. P. Hennart (Ed.), *Lecture notes in mathematics, 909* (pp. 34–47). Berlin: Springer-Verlag.

Green, P. E., Carmone, F. J. Jr., & Smith, S. M. (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: DSWO Press.

Groenen, P. J. F., & Heiser, W. J. (1991). *An improved tunneling function for finding a decreasing series of local minima* (Research Rep. No. RR-91-06). Leiden: Department of Data Theory.

Groenen, P. J. F., de Leeuw, J., & Mathar, R. (1996). Least squares multidimensional scaling with transformed distances. In W. Gaul & D. Pfeifer (Eds.), *Studies in classification, data analysis, and knowledge organization* (pp. 177–185). Berlin: Springer.

Groenen, P. J. F., Mathar, R., & Heiser, W. J. (1995). The majorization approach to multidimensional scaling for Minkowski distances. *Journal of Classification, 12*, 3–19.

Hardy, G. H., Littlewood, J. E. & Pólya, G. (1952). *Inequalities* (2nd ed.). Cambridge: University Press.

Heiser, W. J. (1989). The city-block model for three-way multidimensional scaling. In R. Coppi & S. Bolasco (Eds.), *Multiway data analysis* (pp. 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.

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

Heiser, W. J., & de Leeuw, J. (1977). *How to use SMACOF-I* (Research Rep. No. UG-86-02). Leiden: Department of Data Theory.

Heiser, W. J., & Groenen, P. J. F. (1994). *Cluster differences scaling with a within clusters loss component and a fuzzy successive approximation strategy to avoid local minima* (Research Rep. No. RR-94-03). Leiden: Department of Data Theory.

Hubert, L. J., & Arabie, P. (1986). Unidimensional scaling and combinatorial optimization. In J. de Leeuw, W. J. Heiser, J. Meulman & F. Critchley (Eds.), *Multidimensional data analysis* (pp. 181–196). Leiden: DSWO Press.

Hubert, L. J., Arabie, P., & Hesson-McInnis, M. (1992). Multidimensional scaling in the city-block metric: A combinatorial approach. *Journal of Classification, 9*, 211–236.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1–28.

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

Kruskal, J. B., Young, F. W., & Seery, J. B. (1977). *How to use KYST-2, a very flexible program to do multidimensional scaling and unfolding*. Murray Hill, NJ: AT&T Bell Laboratories.

Levy, A. V., & Gomez, S. (1985). The tunneling method applied to global optimization. In P. T. Boggs, R. H. Byrd, & R. B. Schnabel (Eds.), *Numerical optimization 1984* (pp. 213–244). Philadelphia: SIAM.

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

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

Meulman, J. J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations. *Psychometrika, 57*, 539–565.

Montalvo, A. (1979). *Development of a new algorithm for the global minimization of functions* Unpublished doctoral dissertation, Universidad Nacional Autonoma de Mexico.

Robinson, W. S. (1951). A method for chronologically ordering archaeological deposits. *American Antiquity, 16*, 293–301.

Shepard, R. N. (1962). Analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika, 27*, 125–140.

Tijssen, R. J. W. (1992). *Cartography of science: Scientometric mapping with multidimensional scaling methods*. Leiden: DSWO Press.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Tucker, W. S. (1951). *A method for synthesis of factor analysis studies* (Personel Research Section Rep. No. 984). Washington DC: Department of the Army.

Vilkov, A. V., Zhidkov, N. P., & Shchedrin, B. M. (1975). A method of finding the global minimum of a function of one variable. *USSR Computational Mathematics and Mathematical Physics, 15*, 1040–1042.

Wagenaar, W. A., & Padmos, P. (1971). Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. *British Journal of Mathematical and Statistical Psychology, 24*, 101–110.

Weeks, D. G., & Bentler, P. M. (1982). Restricted multidimensional scaling models for asymmetric proximities. *Psychometrika, 47*, 201–208.

Zielman, B. (1991). *Three-way scaling of asymmetric proximities* (Research Rep. No. RR-91-01). Leiden: Department of Data Theory.

Manuscript received 2/9/95
Final version received 4/3/95