RR 87-17

Majorizing STRESS when some dissimilarities are negative

Willem J. Heiser

Majorizing STRESS when some dissimilarities are negative

Willem J. Heiser
Department of Data Theory
University of Leiden

December 1987

Abstract

The usual convergence proof of the SMACOF algorithm model for MDS depends on the assumption of nonnegativity of dissimilaritiy. A generalization of SMACOF is proposed to accomodate the situation in which some of the dissimilarities are negative. It involves the same inequality on which the LARAMP approach to the reciprocal location problem is based (Heiser, 1986, 1987). Three types of circumstances where some of the dissimilarities do become negative are outlined: nonmetric MDS with normalization on the variance, metric MDS with an additive constant, and MDS with City Block distance.

Majorizing STRESS when some dissimilarities are negative

Willem J. Heiser
Department of Data Theory
University of Leiden

1. Introduction

The problem of this report is how to ensure that an iterative multidimensional scaling (MDS) procedure converges monotonically to at least a local optimum when for any of a number of reasons (which will be discussed in more detail in section 7) some of the dissimilarities to be fitted are negative. De Leeuw (1977), and De Leeuw and Heiser (1977, 1980) have developed the so-called SMACOF algorithm model for MDS, with the specific aim of combining previous contributions of Kruskal (1964), Guttman (1968), and others into one least squares framework, in such a way that all procedures derived from the algorithm model would be monotonically convergent. In sections 2 and 3 this SMACOF work is reviewed in quite some detail, since we need to understand exactly why negative dissimilarities are a problem at all. Next, in section 4, the general idea of the proposed solution is explained, and further developed in sections 5 and 6. The generalized SMACOF algorithm can be applied under all further specifications of the MDS problem, such as unfolding, individual differences scaling, and various types of constraints.

2. The SMACOF inequalities

The SMACOF algorithm model is based upon *majorization* of the loss function (called Kruskal's STRESS) that has to be minimized, i.e.

$$\sigma(\mathbf{X}) \equiv \sum_i \sum_j w_{ij} [\delta_{ij} - d_{ij}(\mathbf{X})]^2 , \qquad (1)$$

where $\mathbf{X}$ is a *configuration* of n points in p dimensions, the w_{ij} (i=1,...,n; j=1,...,n) are known nonnegative quantities called *weights*, the δ_{ij} are the given *dissimilarities*, and $d_{ij}(\mathbf{X})$ is the *Euclidean distance* between point i and point j, obtained from the configuration as

$$d_{ij}(\mathbf{X}) \equiv \{ \sum_a (x_{ia} - x_{ja})^2 \}^{1/2} . \tag{2}$$

In order to keep our development sufficiently general, we do not assume symmetry for either dissimilarities or weights.

The loss function can be written as

$$\sigma(\mathbf{X}) = \sum_i \sum_j w_{ij} \delta^2_{ij} + \sum_i \sum_j w_{ij} d^2_{ij}(\mathbf{X}) - 2 \sum_i \sum_j w_{ij} \delta_{ij} d_{ij}(\mathbf{X}) . \tag{3}$$

In the problems to be considered in this report the first term of (3) is for the time being constant, the second one is quadratic in **X**, and the third term is the one to be minorized (since it appears with negative sign). The usual derivation (e.g., De Leeuw & Heiser, 1980) of the SMACOF algorithm starts from the Cauchy-Schwartz inequality, which states that, for any two vectors **u** and **z** with elements u_a and z_a we must have:

$$\{ \sum_a u^2_a \}^{1/2} \{ \sum_a z^2_a \}^{1/2} \geq \sum_a u_a z_a . \tag{4}$$

Substituting $u_a = x_{ia} - x_{ja}$ and $z_a = y_{ia} - y_{ja}$ we obtain, for any pair of configurations **X** and **Y**,

$$\begin{aligned} \{\sum_a (x_{ia} - x_{ja})^2_a\}^{1/2} \{\sum_a (y_{ia} - y_{ja})^2_a\}^{1/2} &\geq \sum_a (x_{ia} - x_{ja}) (y_{ia} - y_{ja}) \\ d_{ij}(\mathbf{X}) \, d_{ij}(\mathbf{Y}) &\geq \sum_a (x_{ia} - x_{ja}) (y_{ia} - y_{ja}) \end{aligned} \tag{5}$$

Now suppose $S(\oplus)$ denotes the index set of all (i,j) for which $d_{ij}(\mathbf{Y}) \neq 0$; and let $S(\mathsf{O})$ contain the remaining index pairs, i.e. all (i,j) for which $d_{ij}(\mathbf{Y}) = 0$. Thus we may write

$$\begin{aligned} \sum_i \sum_j w_{ij} \delta_{ij} d_{ij}(\mathbf{X}) = &\sum\sum_{(i,j) \in S(\oplus)} w_{ij} \delta_{ij} d_{ij}(\mathbf{X}) \\ &+ \sum\sum_{(i,j) \in S(\mathsf{O})} w_{ij} \delta_{ij} d_{ij}(\mathbf{X}) . \end{aligned} \tag{6}$$

For $(i,j) \in S(\oplus)$ we have, from (5) and the nonnegativity of w_{ij} and δ_{ij},

$$d_{ij}(\mathbf{X}) \geq \sum_a (1/d_{ij}(\mathbf{Y})) (x_{ia} - x_{ja}) (y_{ia} - y_{ja}) \tag{7a}$$

$$w_{ij} \delta_{ij} d_{ij}(\mathbf{X}) \geq \sum_a (w_{ij} \delta_{ij} / d_{ij}(\mathbf{Y})) (x_{ia} - x_{ja}) (y_{ia} - y_{ja}) . \tag{7b}$$

For $(i,j) \in S(\mathsf{O})$ we have, from the nonnegativity of $d_{ij}(\mathbf{X})$, w_{ij} and δ_{ij},

$$d_{ij}(\mathbf{X}) \geq 0 \tag{8a}$$

$$w_{ij}\,\delta_{ij}\,d_{ij}(\mathbf{X}) \geq 0\,. \tag{8b}$$

The inequalities (7b) and (8b) are the ones used in the SMACOF approach for minorizing the cross-product term on the right-hand side of (3). In particular, defining the matrix $\mathbf{C}(\mathbf{Y})$ with off-diagonal elements

$$c_{ij}(\mathbf{Y}) \equiv w_{ij}\,\delta_{ij}\,/d_{ij}(\mathbf{Y})\,, \qquad \text{for } (i,j) \in S(\oplus)\,, \tag{9a}$$

$$c_{ij}(\mathbf{Y}) \equiv 0\,, \qquad \text{for } (i,j) \in S(\mathsf{O})\,. \tag{9b}$$

we obtain

$$\Sigma_i\Sigma_j\, w_{ij}\,\delta_{ij}\,d_{ij}(\mathbf{X}) \geq \Sigma_i\Sigma_j\Sigma_a\, c_{ij}(\mathbf{Y})\,(x_{ia} - x_{ja})\,(y_{ia} - y_{ja})\,. \tag{10}$$

Before proceeding we notice that the inequalities (7b) and (8b) are only valid under the assumptions $\delta_{ij} \geq 0$ and $w_{ij} \geq 0$. It is the effect of dropping the first of these that we want to study in this paper.

3. Matrix formulation.

In this section the loss function (3) and the basic minorization result (10) are expressed directly in terms of matrix operations. This formulation enables us to state the STRESS diminishing property of each step in the SMACOF algorithm more easily. We first define

$$\eta^2(\Delta) \equiv \Sigma_i\Sigma_j\, w_{ij}\,\delta^2_{ij}\,, \tag{11a}$$

$$\eta^2(\mathbf{X}) \equiv \Sigma_i\Sigma_j\, w_{ij}\,d^2_{ij}(\mathbf{X})\,, \tag{11b}$$

$$\rho(\Delta,\mathbf{X}) \equiv \Sigma_i\Sigma_j\, w_{ij}\,\delta_{ij}\,d_{ij}(\mathbf{X})\,. \tag{11c}$$

The dependence on δ_{ij} is introduced in the notation for greater clarity. Thus STRESS (3) can be re-expressed as

$$\sigma(\Delta,\mathbf{X}) = \eta^2(\Delta) + \eta^2(\mathbf{X}) - 2\,\rho(\Delta,\mathbf{X})\,. \tag{12}$$

Next we define the symmetric matrix $\mathbf{V}$ with elements

$$v_{ij} = -(w_{ij} + w_{ji}), \qquad \text{for } i \neq j, \tag{13a}$$

$$v_{ii} = \sum_{k \neq i} (w_{ik} + w_{ki}), \qquad \text{for } i = j. \tag{13b}$$

This definition allows us to write

$$\begin{aligned}
\eta^2(\mathbf{X}) &= \sum_i \sum_{j \neq i} w_{ij}\, d^2_{ij}(\mathbf{X}) , \\
&= \sum_i \sum_{j \neq i} \sum_a w_{ij}\, x^2_{ia} + \sum_j \sum_{i \neq j} \sum_a w_{ij}\, x^2_{ja} \\
&\qquad - 2 \sum_i \sum_{j \neq i} \sum_a w_{ij}\, x_{ia} x_{ja} \\
&= \sum_a \sum_i x^2_{ia} \{ \sum_{k \neq i} w_{ik} \} + \sum_a \sum_j x^2_{ja} \{ \sum_{k \neq j} w_{kj} \} \\
&\qquad - 2 \sum_a \sum_i \sum_{j \neq i} 1/2(w_{ij} + w_{ji})\, x_{ia} x_{ja} \\
&= \sum_a \sum_i \sum_{j=i} v_{ij}\, x_{ia} x_{ja} + \sum_a \sum_i \sum_{j \neq i} v_{ij}\, x_{ia} x_{ja} \\
&= \mathrm{tr}\, \mathbf{X}'\mathbf{V}\mathbf{X} .
\end{aligned} \tag{14}$$

Analogously, noting that

$$\begin{aligned}
\rho(\Delta, \mathbf{X}) &= \sum_i \sum_{j \neq i} w_{ij}\, \delta_{ij}\, d_{ij}(\mathbf{X}) . \\
&= \sum_i \sum_{j \neq i} c_{ij}(\mathbf{X})\, d^2_{ij}(\mathbf{X}) ,
\end{aligned} \tag{15}$$

we define the symmetric matrix $\mathbf{B}(\mathbf{X})$ with elements

$$b_{ij}(\mathbf{X}) = -\{c_{ij}(\mathbf{X}) + c_{ji}(\mathbf{X})\}, \qquad \text{for } i \neq j, \tag{16a}$$

$$b_{ii}(\mathbf{X}) = \sum_{k \neq i} \{c_{ik}(\mathbf{X}) + c_{ki}(\mathbf{X})\}, \qquad \text{for } i = j. \tag{16b}$$

This allows us to write, for the left-hand part of (10), expressed as in (15), and following the same steps as in the derivation of (14):

$$\rho(\Delta, \mathbf{X}) = \mathrm{tr}\, \mathbf{X}'\mathbf{B}(\mathbf{X})\mathbf{X} . \tag{17}$$

For the right-hand part of (10) we obtain

$$\begin{aligned}
\lambda(\Delta, \mathbf{X}, \mathbf{Y}) &= \sum_i \sum_j \sum_a c_{ij}(\mathbf{Y}) (x_{ia} - x_{ja}) (y_{ia} - y_{ja}) , \\
&= \sum_a \sum_i \sum_{j \neq i} c_{ij}(\mathbf{Y})\, x_{ia} y_{ia} - \sum_a \sum_i \sum_{j \neq i} c_{ij}(\mathbf{Y})\, x_{ia} y_{ja} \\
&\qquad + \sum_a \sum_i \sum_{j \neq i} c_{ij}(\mathbf{Y})\, x_{ja} y_{ja} - \sum_a \sum_i \sum_{j \neq i} c_{ij}(\mathbf{Y})\, x_{ja} y_{ia} \\
&= \sum_a \sum_i x_{ia} y_{ia} \{ \sum_{k \neq i} c_{ik}(\mathbf{Y}) \} - \sum_a \sum_i \sum_{j \neq i} c_{ij}(\mathbf{Y})\, x_{ia} y_{ja}
\end{aligned}$$

$$+ \sum_a \sum_j x_{ja} y_{ja} \{ \sum_{k \neq j} c_{kj}(\mathbf{Y}) \} - \sum_a \sum_i \sum_{j \neq i} c_{ji}(\mathbf{Y})\, x_{ia} y_{ja}$$
$$= \sum_a \sum_i \sum_{j=i} b_{ij}(\mathbf{Y})\, x_{ia} x_{ja} + \sum_a \sum_i \sum_{j \neq i} b_{ij}(\mathbf{Y})\, x_{ia} x_{ja}$$
$$= \text{tr}\ \mathbf{X}'\mathbf{B}(\mathbf{Y})\mathbf{Y} \ . \tag{18}$$

From (12), (14), and (17) the matrix formulation of STRESS becomes

$$\sigma(\Delta,\mathbf{X}) = \eta^2(\Delta) + \text{tr}\ \mathbf{X}'\mathbf{V}\mathbf{X} - 2\ \text{tr}\ \mathbf{X}'\mathbf{B}(\mathbf{X})\mathbf{X} \ , \tag{19}$$

and the majorizing function $\mu(\Delta,\mathbf{X},\mathbf{Y})$ used in the original SMACOF algorithm model is, using (18) as well,

$$\mu(\Delta,\mathbf{X},\mathbf{Y}) = \eta^2(\Delta) + \text{tr}\ \mathbf{X}'\mathbf{V}\mathbf{X} - 2\ \text{tr}\ \mathbf{X}'\mathbf{B}(\mathbf{Y})\mathbf{Y} \ . \tag{20}$$

It now follows easily from (10) that $\mu(\Delta,\mathbf{X},\mathbf{Y})$ has the properties

$$\sigma(\Delta,\mathbf{X}) \leq \mu(\Delta,\mathbf{X},\mathbf{Y}) \ , \tag{21a}$$
$$\sigma(\Delta,\mathbf{Y}) = \mu(\Delta,\mathbf{Y},\mathbf{Y}) \ , \tag{21b}$$

for all pairs of configurations $\mathbf{X}$ and $\mathbf{Y}$. So if we set $\mathbf{Y} = \mathbb{X}$, some known configuration of points, and define the update $\mathbf{X}^+$ of $\mathbb{X}$ as

$$\mathbf{X}^+ = \underset{\mathbf{X} \in \Omega}{\textit{argmin}}\ \ \mu(\Delta,\mathbf{X},\mathbb{X}) \ , \tag{22}$$

in which Ω denotes the feasible region defining restrictions on $\mathbf{X}$, then we are sure to obtain the sequence

$$\sigma(\Delta,\mathbb{X}) = \mu(\Delta,\mathbb{X},\mathbb{X}) \geq \mu(\Delta,\mathbf{X}^+,\mathbb{X}) \geq \sigma(\Delta,\mathbf{X}^+) \ . \tag{23}$$

The first inequality follows from (21b), the middle inequality follows from (22) (provided that $\mathbb{X}$ satisfies $\mathbb{X} \in \Omega$, a condition that should be kept in mind when developing procedures of this kind), and the last inequality follows from (21a). So the update always diminishes the value of STRESS, and the algorithm continues with $\mathbf{X}^+$ in the role of $\mathbb{X}$ until convergence. If $\mathbf{X}$ is unrestricted, problem (22) amounts to finding the unconstrained minimum of a quadratic function. Setting the partials equal to zero, we see that $\mathbf{X}^+$ must satisfy

$$\mathbf{V}\mathbf{X}^{+} = \mathbf{B}(\mathbb{X})\mathbb{X} \ . \tag{24}$$

This unrestricted update is called the *Guttman transform*, because Guttman (1968) was the first to propose using this type of update. So in the simplest MDS case it is sufficient to cycle through a series of Guttman transforms. In less simple cases the SMACOF algorithm model tells us to cycle through a series of constrained quadratic programming problems (22).

4. How to proceed when some of the dissimilarities are negative

We have seen that the STRESS diminishing property (23) of the SMACOF update, in particular also of the Guttman transform, depends critically on (21a), which follows from (10), which in turn is based on (7b) and (8b), which are only true if δ_{ij} is nonnegative. Some of the circumstances under which δ_{ij} may be or become negative will be discussed in a short while. Two of them arise when $\sigma(\Delta,\mathbf{X})$ is optimized over $\Delta \in \Gamma$ as well, where Γ denotes a region (frequently a convex cone) of feasible dissimilarity quantifications. We then work with an extended sequence, starting with some initial quantification $\mathbb{\Delta}$, and proceeding as

$$\begin{aligned}\sigma(\mathbb{\Delta},\mathbb{X}) \geq \\ \geq \sigma(\Delta^{+},\mathbb{X}) = \mu(\Delta^{+},\mathbb{X},\mathbb{X}) \geq \mu(\Delta^{+},\mathbf{X}^{+},\mathbb{X}) \geq \sigma(\Delta^{+},\mathbf{X}^{+}) \geq \\ \geq \sigma(\Delta^{++},\mathbf{X}^{+}) \geq \ldots\end{aligned} \tag{25}$$

So in order to keep the extended process convergent, it is imperative to generalize the SMACOF algorithm model so as to accomodate cases in which elements of Δ^{+} have become negative.

Let us register the elements of Δ which are negative in the index set $S(N)$, and express them explicitly as $\delta_{ij} = - |\delta_{ij}|$, for $(i,j) \in S(N)$. Then the remaining index pairs are collected in $S(P)$, and STRESS (12) changes into

$$\sigma(\Delta,\mathbf{X}) = \eta^2(\Delta) + \eta^2(\mathbf{X}) - 2\,\rho_P(\Delta,\mathbf{X}) + 2\,\rho_N(\Delta,\mathbf{X})\,, \tag{26}$$

where the cross product term $\rho(\Delta,\mathbf{X})$ is split into the two components

$$\rho_P(\Delta,\mathbf{X}) \equiv \sum\sum_{(i,j)\in S(P)} w_{ij}\,\delta_{ij}\,d_{ij}(\mathbf{X})\,, \tag{27a}$$

$$\rho_N(\Delta,\mathbf{X}) \equiv \sum\sum_{(i,j)\in S(N)} w_{ij}\,|\delta_{ij}|\,d_{ij}(\mathbf{X})\,. \tag{27b}$$

Clearly, whereas $\rho_P(\Delta,\mathbf{X})$, like $\rho(\Delta,\mathbf{X})$ before, must be *minorized* because it appears with negative sign in STRESS, the term $\rho_N(\Delta,\mathbf{X})$ must be *majorized* in order to majorize the entire function. The partitioning of the index pairs into $S(N)$ and $S(P)$ is independent from the partitioning into $S(\mathsf{O})$ and $S(\oplus)$, so there are four combinations to consider. When $(i,j) \in S(P)$ we simply stick to the previous definitions (9a) and (9b); for $(i,j) \in S(N)$ we will first consider the case that we also have $(i,j) \in S(\oplus)$, in which $d_{ij}(\mathbf{Y})$ is *usable* - a term introduced by De Leeuw (1984a) - and next the case $(i,j) \in S(\mathsf{O})$ in which the known distance is *not* usable, i.e. $d_{ij}(\mathbf{Y}) = 0$. By applying the majorization result formulated in Heiser (1986) - also see Heiser (1987a, 1987b) - we are able to adjust the matrix formulation of STRESS (19) and the corresponding majorizing function (20), and thus to generalize the SMACOF algorithm model so that the convergence results described in the previous section remain valid.

5. Majorization in case the known distance is usable

Throughout this section we assume that the known distance $d_{ij}(\mathbf{Y})$ is usable, i.e. $d_{ij}(\mathbf{Y}) > 0$. Thus although $\mathbf{Y}$ may vary over the same region as $\mathbf{X}$, i.e. Ω, we first restrict attention to that part of Ω that excludes zero distance. For purposes of algorithm construction $\mathbf{Y}$ is the (temporarily) "known" configuration, perhaps better be called the *supporting configuration*, as it defines the

place where the loss function and the majorizing function touch each other (21b). The next section will then develop an approximate majorization result that can be applied when $d_{ij}(\mathbf{Y})$ is not usable.

The key observation is that the inequality

$$[\, d_{ij}(\mathbf{Y}) - d_{ij}(\mathbf{X}) \,]^2 \geq 0 \tag{28}$$

is true for any $\mathbf{X}$ and $\mathbf{Y}$. From (28) it follows that we must have

$$2\, d_{ij}(\mathbf{Y})\, d_{ij}(\mathbf{X}) \leq d^2_{ij}(\mathbf{Y}) + d^2_{ij}(\mathbf{X}) , \tag{29a}$$

$$2\, w_{ij}\, |\delta_{ij}|\, d_{ij}(\mathbf{X}) \leq w_{ij}\, |\delta_{ij}|\, d_{ij}(\mathbf{Y}) + \{ w_{ij}\, |\delta_{ij}| / d_{ij}(\mathbf{Y}) \}\, d^2_{ij}(\mathbf{X}) . \tag{29b}$$

We can go from (29a) to (29b) because $d_{ij}(\mathbf{Y})$ is usable and $w_{ij}\, |\delta_{ij}|$ is nonnegative. Note that we have equality when $\mathbf{X} = \mathbf{Y}$, as required for the majorization method to work. Inequality (29b) is our basic result for generalizing the SMACOF algorithm model (it was first developed, albeit in a different context, in Heiser, 1986). The convex function on the left-hand side is majorized by a quadratic function on the right-hand side. Because $\mathbf{Y}$ is considered to be known, we need not worry about the first part of that piece of the majorizing function, and the second part is of the same form as the components of $\eta^2(\mathbf{X})$. So up to a constant we can keep the majorizing function in the same general shape (20), i.e. for the negative dissimilarity case we use

$$\underline{\mu}(\Delta,\mathbf{X},\mathbf{Y}) = c + \eta^2(\Delta) + \mathrm{tr}\, \mathbf{X}'\underline{\mathbf{V}}(\mathbf{Y})\mathbf{X} - 2\, \mathrm{tr}\, \mathbf{X}'\underline{\mathbf{B}}(\mathbf{Y})\mathbf{Y} , \tag{30}$$

where c is the sum of $w_{ij}\, |\delta_{ij}|\, d_{ij}(\mathbf{Y})$ over all $(i,j) \in S(N) \cap S(\oplus)$, and where the off-diagonal elements of $\underline{\mathbf{V}}(\mathbf{Y})$ and $\underline{\mathbf{B}}(\mathbf{Y})$ are obtained, like those of $\mathbf{V}$ and $\mathbf{B}(\mathbf{Y})$ in (13a) and (16a), as the symmetrized off-diagonal elements of the new matrices $\underline{\mathbf{W}}(\mathbf{Y})$ and $\underline{\mathbf{C}}(\mathbf{Y})$, respectively, defined as

$$\underline{w}_{ij}(\mathbf{Y}) = w_{ij} \qquad \text{for } (i,j) \in S(P) \cap S(\oplus), \tag{31a}$$

$$\underline{w}_{ij}(\mathbf{Y}) = \{ w_{ij}\, (d_{ij}(\mathbf{Y}) + |\delta_{ij}|) \} / d_{ij}(\mathbf{Y}) \qquad \text{for } (i,j) \in S(N) \cap S(\oplus), \tag{31b}$$

$$\underline{c}_{ij}(\mathbf{Y}) = w_{ij}\delta_{ij} / d_{ij}(\mathbf{Y}) \qquad \text{for } (i,j) \in S(P) \cap S(\oplus), \tag{32a}$$

$$\underline{c}_{ij}(\mathbf{Y}) = 0 \qquad \text{for } (i,j) \in S(N) \cap S(\oplus). \tag{32b}$$

The diagonal elements of $\underline{\mathbf{V}}(\mathbf{Y})$ and $\underline{\mathbf{B}}(\mathbf{Y})$ are again such that their rows and columns sum to zero (as with $\mathbf{V}$ and $\mathbf{B}(\mathbf{Y})$ in equations (13b) and (16b), respectively). From (26) and (29b) it now follows that the new function $\underline{\mu}(\Delta,\mathbf{X},\mathbf{Y})$ in (30) truely majorizes $\sigma(\Delta,\mathbf{X})$, so all convergence results of section 3 remain valid when using $\underline{\mathbf{V}}(\mathbf{Y})$ and $\underline{\mathbf{B}}(\mathbf{Y})$ in the generalized SMACOF algorithm. For instance, the generalized Guttman transform (the unrestricted minimizer of the majorizing function) must satisfy

$$\underline{\mathbf{V}}(\mathbb{X})\mathbf{X}^{+} = \underline{\mathbf{B}}(\mathbb{X})\mathbb{X} \ , \tag{33}$$

in analogy with (24), and for restricted updates we have to replace $\mu(\Delta,\mathbf{X},\mathbb{X})$ with $\underline{\mu}(\Delta,\mathbf{X},\mathbb{X})$ in subproblem (22).

6. Approximate majorization in case the known distance is not usable

When the known distance is not usable, i.e. when $d_{ij}(\mathbf{Y}) = 0$, the standard SMACOF approach can still rely on the inequality $d_{ij}(\mathbf{X}) \geq 0$; in the negative dissimilarity case this possibility is no longer available since we need an upper bound, not a lower bound. Two strategies for handling this case have been described in Heiser (1987a); we will use one of these here, adding one slight modification.

The problem is not so much to find an auxiliary function that is always above the original function, but it is to find one that becomes equal to the original one when $\mathbf{X} = \mathbf{Y}$. Thus it will not be sufficient to use an inequality like $d_{ij}(\mathbf{X}) \leq k$, with k a "large" constant, since then we remain too far above the original function. The function $w_{ij}\,|\delta_{ij}|\,d_{ij}(\mathbf{X})$ is piecewise linear if one moves the points along anyone particular direction of p-dimensional space, and has a sharp "edge" at $\mathbf{X} = \mathbf{Y}$, i.e. where $d_{ij}(\mathbf{X}) = d_{ij}(\mathbf{Y}) = 0$. This implies that, although it is not hard to *minorize* it in this neighbourhood (in fact, this is done with a global *linear* function in the ordinary SMACOF approach), it is extremely hard, if not impossible, to *majorize* it there with any continuous function whatsoever. Our approach therefore will be to drop the requirement of *exact* equality of the two

functions at $\mathbf{X} = \mathbf{Y}$, and to be satisfied with *approximate* equality, i.e. with staying above the original function at a distance of at most ε, with ε some arbitrarily small positive number.

Consider the inequality

$$[\,(\varepsilon / |\delta_{ij}|) - d_{ij}(\mathbf{X})\,]^2 \geq 0\,. \tag{34}$$

whenever it occurs that $d_{ij}(\mathbf{Y}) = 0$. In the same way as (29b) was derived from (28) we now find that

$$2\, w_{ij}\, |\delta_{ij}|\, d_{ij}(\mathbf{X}) \leq w_{ij}\, \varepsilon + \{w_{ij}\, |\delta_{ij}|^2 / \varepsilon\}\, d^2_{ij}(\mathbf{X})\,. \tag{35}$$

Clearly, the majorizing function $\underline{\mu}(\Delta,\mathbf{X},\mathbf{Y})$ as defined in (30) needs the following adjustments for $(i,j) \in S(O)$: c is the sum of $w_{ij}\, |\delta_{ij}|\, d_{ij}(\mathbf{Y})$ over all $(i,j) \in S(N) \cap S(\oplus)$ and in addition of $w_{ij}\, \varepsilon$ over all $(i,j) \in S(N) \cap S(O)$; the relevant off-diagonal elements of the matrices $\underline{\mathbf{W}}(\mathbf{Y})$ and $\underline{\mathbf{C}}(\mathbf{Y})$ have to be specified as

$$\underline{w}_{ij}(\mathbf{Y}) = w_{ij} \qquad \text{for } (i,j) \in S(P) \cap S(O), \tag{36a}$$

$$\underline{w}_{ij}(\mathbf{Y}) = \{w_{ij}\, (\varepsilon + |\delta_{ij}|^2)\}\, / \varepsilon \qquad \text{for } (i,j) \in S(N) \cap S(O), \tag{36b}$$

$$\underline{c}_{ij}(\mathbf{Y}) = 0 \qquad \text{for } (i,j) \in S(P) \cap S(O), \tag{37a}$$

$$\underline{c}_{ij}(\mathbf{Y}) = 0 \qquad \text{for } (i,j) \in S(N) \cap S(O). \tag{37b}$$

It now follows from (35) that $\underline{\mu}(\Delta,\mathbf{X},\mathbf{Y}) \geq \sigma(\Delta,\mathbf{X})$, but we merely have $\underline{\mu}(\Delta,\mathbf{Y},\mathbf{Y}) \approx \sigma(\Delta,\mathbf{Y})$ in the generalized algorithm, where the order of the approximation is ε. More precisely, from (35) it can be seen that for $(i,j) \in S(N) \cap S(O)$ the contribution h_{ij} to the (positive) difference between STRESS and the majorizing function at $\mathbf{X} = \mathbf{Y}$ is

$$\begin{aligned} h_{ij} \quad &= w_{ij}\, \varepsilon + \{w_{ij}\, |\delta_{ij}|^2 / \varepsilon\}\, d^2_{ij}(\mathbf{Y}) - 2\, w_{ij}\, |\delta_{ij}|\, d_{ij}(\mathbf{Y}) \\ &= w_{ij}\, \varepsilon\, [1 - |\delta_{ij}| d_{ij}(\mathbf{Y}) / \varepsilon]^2\,. \end{aligned} \tag{38}$$

So if $d_{ij}(\mathbf{Y}) = 0$, then $h_{ij} = w_{ij}\, \varepsilon$, where the weights w_{ij} need not bother us as we can assume without loss of generality that they are between zero and one. This property implies that we can come as closely to the STRESS function as is needed to keep the successive function values

monotonically decreasing, by choosing ε smaller than the stopping criterion used to terminate the process.

In practice, we never test $d_{ij}(\mathbf{Y})$ for exact equality to zero, but use some small positive bounding value beyond which the mechanism of this section is to be invoked. Such a bound β can be derived from (38) by requiring that $h_{ij} \leq w_{ij}\,\varepsilon$ whenever $d_{ij}(\mathbf{Y}) \leq \beta$; thus the squared term in (38) should be smaller than one. This requirement yields $\beta = 2\,\varepsilon / |\delta_{ij}|$, which can be sharpened to $\beta = 2\,\varepsilon$ when $|\delta_{ij}| \leq 1$, since in the latter case $1 \leq 1 / |\delta_{ij}|$ implies $2\,\varepsilon \leq 2\,\varepsilon / |\delta_{ij}|$. In general, when δ_{ij} is negative $d_{ij}(\mathbf{Y})$ will tend to be(come) small; then (31b) shows that the new weight $\underline{w}_{ij}(\mathbf{Y})$ will become relatively large, thus enforcing $d_{ij}(\mathbf{X})$ to remain small, and when $d_{ij}(\mathbf{Y})$ approaches zero then (36b) yields a very large $\underline{w}_{ij}(\mathbf{Y})$, except when δ_{ij} approaches zero as well. The latter effect is reassuring, since it marks the continuity from (31b) to (31a) and from (36b) to (36a) when $\delta_{ij} \rightarrow 0$ from below. If δ_{ij} is only slightly smaller than zero the weights $\underline{w}_{ij}(\mathbf{Y})$ will never become exceedingly large (e.g., when $|\delta_{ij}| = \sqrt{\varepsilon}$ the weight is merely doubled); only for very large violations of positivity there might be reason to worry about $\underline{\mathbf{V}}(\mathbf{Y})$ becoming ill-conditioned due to extreme variation in its elements.

The slight modification with respect to Heiser (1987a) concerns the following. In that paper the loss function studied is of the same form as $\rho(\Delta,\mathbf{X})$, but with $f_{ij} = w_{ij}\,|\delta_{ij}|$ as general nonnegative weights. Since f_{ij}, like w_{ij}, can have zero elements it cannot be used for division of ε, as is done in (34) with $|\delta_{ij}|$ (which is always positive). The correction with $|\delta_{ij}|$ leads to a weight $\underline{w}_{ij}(\mathbf{Y})$ in (36b) that more strongly depends on the amount of nonpositivity, which is desirable in the present context.

7. Circumstances in which negative dissimilarities arise

Three cases in which negative dissimilarities are known to arise will be discussed in this section. The first is when $\sigma(\Delta,\mathbf{X})$ is optimized over Δ as well, yielding the so-called *pseudo-distances* to which the distances have to be fitted, and when it is *normalized on the variance* of the pseudo-distances. The second case, also involving pseudo-distances, is independent of the

normalization, but appears to be a possibility whenever the class of regression functions that is set up to regress $\mathbf{D}(\mathbf{X})$ on Δ includes an *intercept*. Finally, we may have to deal with negative quantities in the role of dissimilarities when fitting the *City Block model* dimension after dimension (even if the original dissimilarities are positive and remain fixed).

7.1. *Normalization on the variance*. Optimization of $\sigma(\Delta,\mathbf{X})$ over Δ can be done in various ways, e.g. over the entire set of dissimilarities, or over subsets of them separately. It is therefore convenient to drop the double subscripting and to collect all relevant dissimilarities in a vector $\boldsymbol{\delta}$, and all corresponding distances in a vector $\mathbf{d}$, both of length m. In most cases some normalization is needed in order to exclude collapse of all quantities to zero, or to avoid other undesirable effects (Kruskal and Carroll, 1969). Let us now consider the following problem involving one of these normalized loss functions:

$$\textit{Problem A:} \quad \min_{\boldsymbol{\delta}\in\Gamma} \quad \| \boldsymbol{\delta} - \mathbf{d} \|^2 / \boldsymbol{\delta}'\mathbf{J}\boldsymbol{\delta} \ . \tag{39}$$

The class of regression functions is denoted by Γ, and we will restrict attention to the common "ordinal" case where Γ is the convex cone of monotonic transformations of some initial set of dissimilarities. The symbol $\mathbf{J}$ is used for the centering operator, i.e. $\mathbf{J} = \mathbf{I} - \mathbf{ee}'/\mathbf{e}'\mathbf{e}$, with $\mathbf{e}$ a vector of ones. So the normalized loss function in problem A is the squared Euclidean distance between $\boldsymbol{\delta}$ and $\mathbf{d}$ divided by the variance of $\boldsymbol{\delta}$.

As shown by Kruskal and Carroll (1969) problem A can be solved by relating it to a number of somewhat simpler problems. First, let us write $\boldsymbol{\delta} = \boldsymbol{\delta}_1 + \mu\mathbf{e}$, where $\boldsymbol{\delta}_1$ is in deviation from its mean, registered in the new parameter μ. The fact that Γ is a convex cone implies that if $\boldsymbol{\delta}\in\Gamma$ then $\boldsymbol{\delta}_1\in\Gamma$ as well, so that it is sufficient to find optimal $\boldsymbol{\delta}_1$ and μ. Thus problem A is equivalent to

$$\textit{Problem B:} \quad \min_{\mu,\ \boldsymbol{\delta}_1\in\Gamma} \quad \| \boldsymbol{\delta}_1 + \mu\mathbf{e} - \mathbf{d} \|^2 / \boldsymbol{\delta}_1'\boldsymbol{\delta}_1. \tag{40}$$

Solving problem B for μ yields $\mu^* = (\mathbf{e}'\mathbf{d})/m$, and when we define $\mathbf{d}_1 = \mathbf{d} - \mu^*\mathbf{e}$ we get

$$\textit{Problem C:} \quad \min_{\boldsymbol{\delta}_1\in\Gamma} \quad \| \mathbf{d}_1 - \boldsymbol{\delta}_1 \|^2 / \boldsymbol{\delta}_1'\boldsymbol{\delta}_1. \tag{41}$$

Using Kruskal and Carroll's arguments it can be shown that if δ_2^* solves the unnormalized

$$\textit{Problem D:} \qquad \min_{\delta_2 \in \Gamma} \quad \| \mathbf{d}_1 - \delta_2 \|^2, \tag{42}$$

then the rescaled vector $\delta_1^* = (\mathbf{d}_1'\mathbf{d}_1)/(\delta_2^{*\prime}\delta_2^*)\,\delta_2^*$ solves problem C. Once we know δ_1^* it is easy to obtain the pseudo-distances $\delta^* = \delta_1^* + \mu^*\mathbf{e}$ as the solution to problem A. Problem D is a monotonic regression problem, the efficient solution of which is standard (Kruskal, 1964).

The fact that problem A may lead to negative pseudo-distances was noticed by Heiser (1981, ch.7), based on the observation that the rescaling factor to go from δ_2^* to δ_1^* is larger than one, by the general variance diminishing property of regression: $(\mathbf{d}_1'\mathbf{d}_1) \geq (\delta_2^{*\prime}\delta_2^*)$. As a consequence, when some of the smallest elements of δ_2^* remain equal to those of $\mathbf{d}_1$ (when there are no order reversals in the small distances), but some of the larger elements of δ_2^* do get tied, the rescaling factor causes the small elements of δ_1^* to get smaller than those of $\mathbf{d}_1$ (both vectors are in deviation from their mean). Then adding the mean distance μ^* may not be enough to obtain positivity for the smallest values of δ^*. This effect is illustrated in Table 1, where the distances in

Table 1. Example of monotonic regression with normalization on the variance, yielding a negative pseudo-distance.

$\mathbf{d}$	$\mathbf{d}_1$	δ_2^*	δ_1^*	δ^*
1	-4	-4	-6.133	-1.133
3	-2	-2	-3.067	1.933
8	3	1	1.533	6.533
4	-1	1	1.533	6.533
9	4	2	3.067	8.067
5	0	2	3.067	8.067
30	0	0	0.000	30.000

the leftmost column are those among four points along a line, with a spacing of 1, 3,and 5 units. The mean of the distances is 5; the second column gives the centered distances $\mathbf{d}_1$, and the third column the standard monotonic regression assuming the order of the dissimilarities coincides with

the order of the rows. The sum of squares of $\mathbf{d}_1$ is 46, the sum of squares of $\boldsymbol{\delta}_2^*$ is 30, so the rescaling factor is 1.533. This factor results in a value of -6.133 for the smallest element of the rescaled $\boldsymbol{\delta}_1^*$, and therefore in a smallest pseudo-distance of -1.133.

The remedy against negative pseudo-distances proposed by Heiser (1981) was to restrict the normalized monotonic regression to remain nonnegative, using a specialized quadratic programming approach. The possibility of obtaining negative transformed dissimilarities was also considered in Stoop and De Leeuw (1982), who proposed to add a step-size procedure to the standard SMACOF approach (with the aim to diminishing the step-size whenever STRESS starts to increase). Both proposals could be regarded as somewhat *ad hoc*, but are obvious candidates to compare the present procedure with.

7.2. *Regression including an intercept.* In many cases of metric multidimensional scaling the data are recorded on a dissimilarity scale with arbitrary origin (an "interval" scale). The fact that shifts along the dissimilarity scale may greatly influence the solution obtained, and even the estimated dimensionality of the configuration of points, is a recurring worry in the theory of MDS (called the *additive constant problem*). In the STRESS framework a simple solution (discussed at length in, e.g., Cooper, 1972, and Roskam, 1972) would seem to be to define Γ as the subspace of linear transformations of some initially given set of dissimilarities $\boldsymbol{\delta}_0$. Since our argument does not depend on the particular normalization chosen, we consider the unnormalized

$$\textit{Problem E:} \qquad \min_{\alpha,\mu} \quad \| (\alpha\boldsymbol{\delta}_0 + \mu\mathbf{e}) - \mathbf{d} \|^2 . \tag{43}$$

The parameter estimates of this linear regression problem are well-known to be $\mu^* = \mu_d - \alpha^*\mu_\delta$, where μ_d and μ_δ are the mean of the distances and the mean of the dissimilarities, respectively, and $\alpha^* = \mathbf{d}'\mathbf{J}\boldsymbol{\delta}_0 / \boldsymbol{\delta}_0'\mathbf{J}\boldsymbol{\delta}_0$.

The circumstances under which $\boldsymbol{\delta}^* = \alpha^*\boldsymbol{\delta}_0 + \mu^*\mathbf{e}$ will obtain negative elements are perhaps similar to those of the previous case: when the plot of $\mathbf{d}$ against $\boldsymbol{\delta}_0$ exhibits acceleration, i.e. when piece-wise linear functions would get steeper slopes from left to right, then a global linear function might pass through the $\boldsymbol{\delta}_0$ axis beyond the few smallest dissimilarities. An example is given in

Table 2, with the same distances as the ones used in Table 1, and with dissimilarities chosen in the

Table 2. Example of linear regression with intercept, yielding a negative pseudo-distance.

δ_0	**d**	$\mathbf{J}\delta_0$	**Jd**	δ^*	$\delta_\mu{}^*$
1	1	-6	-4	-0.04	-1.0
7	3	0	-2	5.00	5.0
7	4	0	-1	5.00	5.0
8	5	1	0	5.84	6.0
9	8	2	3	6.68	7.0
10	9	3	4	7.52	8.0
42	30	0	0	30.00	30.0

way just indicated. The means are $\mu_\delta = 7$ and $\mu_d = 5$; the cross product of **Jd** and $\mathbf{J}\delta_0$ is 42, the sum of squares of $\mathbf{J}\delta_0$ is 50, and therefore the slope coefficient $\alpha^* = .84$ and the additive constant $\mu^* = -.88$. From this line we get a smallest pseudo-distance of -.04 (for sure, not *very* negative, but the example was kept as realistic as possible). The right-most column of Table 2 gives the result for the case of an additive constant only, denoted by $\delta_\mu{}^*$, where the slope is fixed at 1. The effect becomes more pronounced, because "regression to the mean" is precluded.

Thus we expect Cooper's (1972) and Roskam's (1972) algorithms for solving the additive constant problem to exhibit irregular, or less efficient, convergence behavior, since these authors do not seem to have been aware of the problem. The phenomenon of negative predicted values might not be immediately evident from the point of view of normal regression theory, as under the usual assumptions regression plots are cigar shaped, or - under less usual assumptions - they can at least be inspected for nonlinearity. In the multidimensional scaling context the shape of the regression changes every iteration, and nonpositivity due to (perhaps unexpected) nonlinearity has to be dealt with explicitly.

7.3. *Dimension-wise fitting of the City Block model.* The third case is quite different from the previous two, because we will now consider a different *distance* model, while keeping the

dissimilarities fixed, for convenience. In the *City Block-*, *Manhattan-*, or L_1-model the distance between two points in p-dimensional space is given by

$$ {}_1d_{ij}(\mathbf{X}) \equiv \sum_a | x_{ia} - x_{ja} | . \tag{44} $$

An important characteristic of the City Block distance is that it is *additive* across dimensions, i.e. we may express it as the sum of "one-dimensional Euclidean" distances:

$$ {}_1d_{ij}(\mathbf{X}) = \sum_a d_{ij}(\mathbf{x}_a) , \tag{45} $$

where $\mathbf{x}_a$ is used to denote the coordinates on the ath dimension, or the ath column of $\mathbf{X}$. Additivity allows us to define the residual quantities

$$ {}_a\delta_{ij} = \delta_{ij} - \sum_{b \neq a} d_{ij}(\mathbf{x}_b) , \tag{46} $$

the "dissimilarities corrected for the contribution of the other dimensions", so that STRESS for the ath dimension of City Block scaling, holding the other dimensions fixed at their current values, can be written as:

$$ \sigma_1(\mathbf{x}_a) = \sum_i \sum_j w_{ij} [{}_a\delta_{ij} - d_{ij}(\mathbf{x}_a)]^2 . \tag{47} $$

Stated this way it would seem that all we have to do for City Block scaling is repeatedly solving a one-dimensional MDS problem with updated residual dissimilarities. Hubert and Arabie (1988) have reported their disappointing experiences with this idea (e.g., in a number of cases they were unable to recover the generating configuration of errorless data; also, they observed non-monotone convergence behavior). One major reason for these problems would seem to be that (46) allows negative ${}_a\delta_{ij}$ to occur (Hubert, 1987, has confirmed that negativity is indeed present in his testruns). So the generalized SMACOF algorithm is called for when (47) is to be minimized reliably.

A second possibility suggests itself if we have a closer look at the structure of STRESS under the City Block model. For the sake of simplicity, we will do this for the two-dimensional case; no new complications arise for higher dimensionalities. Using (45), two-dimensional City Block scaling (3) can be re-expressed as the minimization of

$$\begin{aligned}\sigma(X) = \quad & \Sigma_i\Sigma_j\, w_{ij}\, \delta^2_{ij} \\ & + \Sigma_i\Sigma_j\, w_{ij}\, d^2_{ij}(\mathbf{x}_1) + \Sigma_i\Sigma_j\, w_{ij}\, d^2_{ij}(\mathbf{x}_2) \\ & + 2\,\Sigma_i\Sigma_j\, w_{ij}\, d_{ij}(\mathbf{x}_1)\, d_{ij}(\mathbf{x}_2) \\ & - 2\,\Sigma_i\Sigma_j\, w_{ij}\, \delta_{ij}\, d_{ij}(\mathbf{x}_1) - 2\,\Sigma_i\Sigma_j\, w_{ij}\, \delta_{ij}\, d_{ij}(\mathbf{x}_2)\,. \end{aligned} \tag{48}$$

So the only term that prevents straightforward dimension-wise fitting is the cross product of distances, the term on the third line of (48). Now either we can "redistribute" this term over the last two cross product terms, yielding the approach of the previous paragraph with quantities like (46), or we could majorize it directly with the basic inequality (29b), using $d_{ij}(\mathbf{x}_2)$ or $d_{ij}(\mathbf{x}_1)$ in the role of $|\delta_{ij}|$. A detailed comparison of these two possibilities would bring us outside the scope of this report.

Finally, two technical remarks are in order. First, the term dimension-wise fitting is used here in the sense of *alternating least squares* , i.e. one dimension is fitted at a time, while keeping the other coordinates fixed at their current values. This ensures decrease of the loss function across subproblems. The term is not used in the sense of *successive fitting* of dimensions, i.e. fitting one dimension first, followed by subtracting its contribution from the data, fitting a second dimension next, and so on, without ever going back to previous dimensions. Unlike what is true for classical MVA techniques such as principal components analysis, these two approaches do not amount to the same thing here. Secondly, in one way or the other we end up with one-dimensional MDS problems, and it is well-known (De Leeuw and Heiser, 1977; Defays, 1978; Heiser, 1981; Hubert and Arabie, 1986, 1988) that globally optimal solutions are especially hard to reach, due to the combinatorial nature of the one-dimensional case. So special attention is required for solving the subproblems of the alternating least squares scheme.

8. Discussion

In this report the SMACOF algorithm model for multidimensional scaling was generalized in order to keep the iterative fitting process convergent when negative dissimilarities are present. It

was shown that there are at least three different situations for which this generalization is needed. For all three situations, further work is required to check whether or not additional complications arise, e.g. whether the local minimum problem becomes more serious than usual.

A further line of future research would be to study alternative ways of proceeding when the previous distance is not usable, i.e. when $d_{ij}(\mathbf{Y}) = 0$; two possibilities are:

(a) restrict $\mathbf{x}_i = \mathbf{x}_j$ for the next iteration(s);

(b) work out how to fix one of the points while trying to improve the other .

Ideally, one would like to be able to prove that whenever $d_{ij}(\mathbf{Y}) = 0$ occurs, optimal $d_{ij}(\mathbf{X}^*) = 0$. When this conjecture would turn out to be true, one would also have to adjust De Leeuw's (1984b) proof that - for ordinary MDS - at the minimum all $d_{ij}(\mathbf{X}^*)$ are usable.

9. References

Cooper, L.G. (1972), A new solution to the additive constant problem in metric multidimensional scaling. *Psychometrika, 37*, 311-322.

Defays, D. (1978), A short note on a method of seriation. *British J. Math. Stat. Psych., 31*, 49-53.

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

De Leeuw, J. (1984a), *Convergence of the majorization algorithm for multidimensional scaling*. Internal Report RR-84-07, Dept. of Data Theory, University of Leiden.

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

De Leeuw, J. and Heiser,W.J. (1980), Convergence of correction matrix algorithms for multidimensional scaling. In J.C. Lingoes et al. (Eds.), *Geometric Representations of Relational Data*, Ann Arbor: Mathesis Press.

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

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

Heiser, W.J. (1981), *Unfolding Analysis of Proximity Data*. Unpublished Doctoral Dissertation, University of Leiden.

Heiser, W.J. (1986), *A majorization algorithm for the reciprocal location problem* . Internal report RR-86-12, Dept. of Data Theory, University of Leiden.

Heiser, W.J. (1987a), *Notes on the L ARAMP algorithm.* Internal report RR-87-04, Dept. of Data Theory, University of Leiden.

Heiser, W.J. (1987b), Correspondence analysis with least absolute residuals, *Computational Statistics & Data Analysis, 5*, 337-356.

Hubert, L. (1987), *Personal communication.*

Hubert, L. , and Arabie, P. (1986), Unidimensional scaling and combinatorial optimization. In J. De Leeuw et al. (Eds.), *Multidimensional Data Analysis*, Leiden: DSWO Press.

Hubert, L. , and Arabie, P. (1988), Relying on necessary conditions for optimization: unidimensional scaling and some extensions. In H.H. Bock (Ed.), *Classification and Related Methods of Data Analysis*, Amsterdam: North-Holland.

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

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

Kruskal, J.B. and Carroll, J.D. (1969), Geometrical models and badness-of-fit funtions. In P.R. Krishnaiah (Ed.), *Multivariate Analysis, Vol 2*, Amsterdam: North-Holland.

Roskam, E.E. (1972), Multidimensional scaling by metric transformation of data. *Nederlands Tijdschrift voor de Psychologie, 27*, 486-508.

Stoop, I. and De Leeuw, J. (1982), *How to use SMACOF-1B*. Internal Report, Dept. of Data Theory, University of Leiden.