Least squares multidimensional scaling with transformed distances

Patrick J.F. Groenen[1], Jan de Leeuw[2] and Rudolf Mathar[3]

[1]Department of Data Theory, University of Leiden
P.O. Box 9555, 2300 RB Leiden, The Netherlands
[2]Interdivisional Program in Statistics, UCLA
[3]Institute of Statistics, Aachen University of Technology
Wüllnerstraße 3, D-5100 Aachen, Germany

Summary: We consider a general least squares loss function for multidimensional scaling. Special cases of this loss function are STRESS, S-STRESS, and MULTISCALE. Several analytic results are presented. In particular, we present the gradient and Hessian, and look at the differentiability at a local minimum. We also consider fulldimensional scaling and indicate when a global minimum can be obtained. Furthermore, we treat the problem of inverse multidimensional scaling, where the aim is to find those dissimilarity matrices for which a fixed configuration is a stationary point.

1. Introduction

Various loss functions exist for performing multidimensional scaling (MDS) that all aim at representing n objects in a p dimensional space such that the distances correspond in some optimal sense to fixed nonnegative dissimilarity measures δ_{ij} for every pair of objects i, j. Here, we consider the general least squares loss function

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

It is convenient to express the squared distance between row i and row j of the $n \times p$ coordinate matrix $\mathbf{X}$ as $d_{ij}(\mathbf{X}) = \text{tr}\,(\mathbf{X}'\mathbf{A}_{ij}\mathbf{X})$, where $\mathbf{A}_{ij} = (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)'$ with $\mathbf{e}_i$ equal to column i of the identity matrix. $w_{ij} = w_{ji}$ are fixed nonnegative weights with $w_{ii} = 0$. The function $f(z)$ could be any function from $\Re^1$ to $\Re^1$, although we shall assume that $f(z)$ is twice continuously differentiable over the domain $(0, \infty)$ and that the inverse function $f^{-1}(z)$ exists such that $f^{-1}(f(z)) = f(f^{-1}(z)) = z$. We focus on three particular examples of $f(z)$, i.e., $f(z) = z^{1/2}$ gives Kruskal's (1964) raw STRESS function, $f(z) = z$ gives S-STRESS (Takane, Young, and de Leeuw (1977)), and $f(z) = \log(z)$ gives Ramsay's (1977) MULTISCALE loss function. For these cases several algorithms for minimizing (1) over $\mathbf{X}$ exist, notably for S-STRESS the ALSCAL algorithm (Takane et al. (1977)), an algorithm of Glunt, Hayden, and Liu (1991), and the Newton–Raphson algorithm of Browne (1987). For the STRESS case the KYST algorithm (Kruskal, Young,

and Seery (1977)) and SMACOF of de Leeuw and Heiser (1980) can be used. An algorithm based on a probabilistic version of (1) with replications was presented by Stoop, Heiser, and de Leeuw (1981).

In the next section we present the gradient and Hessian of the general least squares MDS loss function (1) and investigate several useful properties. Then we consider a special case, fulldimensional scaling, and indicate in what situations a global minimum can be obtained. One of the problems of the algorithms above is that they usually stop at a local minimum, which need not be the global minimum. In order to get a better understanding of the local minimum problem we also study its inverse problem; what dissimilarity matrices Δ have some given $\mathbf{X}$ as local minimum. This problem of *inverse scaling* has been discussed first in de Leeuw and Groenen (1993) using STRESS.

2. The gradient and Hessian

For a local minimum $\mathbf{X}^*$ we need that, if it exists, the gradient equals zero and the Hessian is nonnegative definite. Explicit formulae for gradient and Hessian are given below.

A necessary condition for a stationary point $\mathbf{X}$ is that the gradient of $\sigma(\mathbf{X}, \Delta)$ is equal to zero. This gradient, if it exists, can be written as

$$\frac{\partial \sigma(\mathbf{X}, \Delta)}{\partial \mathbf{x}_s} = 4 \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(f(d_{ij}(\mathbf{X})) - f(\delta_{ij}) \right) f'(d_{ij}(\mathbf{X})) \mathbf{A}_{ij} \mathbf{x}_s, \tag{2}$$

where $f'(z)$ denotes the first derivative of f at z and $\mathbf{x}_s$ is column s of $\mathbf{X}$. Since $\mathbf{A}_{ij}$ is double centered (has row and column sums equal to zero) we may assume that $\mathbf{X}$ also has column sum zero. It is not difficult to see that if $\mathbf{X}^*$ has zero gradient, $\mathbf{X}^*\mathbf{T}$ with $\mathbf{T}$ a rotation matrix ($\mathbf{T}\mathbf{T}' = \mathbf{I}$) is also a stationary point since distances do not change under rotation of $\mathbf{X}$. Note that in general (2) may not be defined everywhere. Especially when zero distances occur $f(z)$ or $f'(z)$ may not be defined. For S-STRESS this problem does not occur which was an important reason for proposing this MDS loss function.

Of course, the set of configurations with zero gradient includes local minima, local maxima and saddle points. If the gradient of $\sigma(\mathbf{X}, \Delta)$ at $\mathbf{X}$ is zero and the Hessian $\mathbf{H}$ is positive definite, i.e., $\mathbf{y}'\mathbf{H}\mathbf{y} > 0$ for all $\mathbf{y} \neq 0$, then we have a strict local minimum at $\mathbf{X}$. Moreover, a necessary condition for a local minimum of σ is that the gradient vanishes and the Hessian $\mathbf{H}$ is positive semidefinite, i.e., $\mathbf{y}'\mathbf{H}\mathbf{y} \geq 0$ for all $\mathbf{y}$. The Hessian is a $p \times p$ partitioned block matrix with blocks

$$\begin{aligned} \mathbf{H}_{st} \;=\; & 4\beta^{st} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (f(d_{ij}(\mathbf{X})) - f(\delta_{ij})) f'(d_{ij}(\mathbf{X})) \mathbf{A}_{ij} + \\ & 8 \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (f(d_{ij}(\mathbf{X})) - f(\delta_{ij})) f''(d_{ij}(\mathbf{X})) \mathbf{A}_{ij} \mathbf{x}_s \mathbf{x}_t' \mathbf{A}_{ij} + \end{aligned}$$

$$8\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(f'(d_{ij}(\mathbf{X})))^2\mathbf{A}_{ij}\mathbf{x}_s\mathbf{x}_t'\mathbf{A}_{ij} \tag{3}$$

of size $n \times n$, where $\beta^{st} = 1$ if $s = t$ and $\beta^{st} = 0$ otherwise. In Table 1 we give the particular gradients and in Table 2 the Hessians of STRESS, S-STRESS and MULTISCALE.

Some properties of the Hessian can be derived. If $\mathbf{H}$ is defined, then $\mathbf{H}$ has only real eigenvalues, since $\mathbf{H}$ is symmetric. Furthermore, $\mathbf{H}$ is rank deficient, which implies that $\mathbf{H}$ has zero eigenvalues. If the $n \times p$ vector $\mathbf{y}$ is an eigenvector corresponding to a zero eigenvalue, then $\sum_{t=1}^{p} \mathbf{H}_{st}\mathbf{y}_t = \mathbf{0}$, where $\mathbf{y}$ is partitioned in p $n \times 1$ vectors $\mathbf{y}_t$. The Hessian $\mathbf{H}$ has p eigenvalues equal to zero corresponding to the p orthogonal eigenvectors for which $\mathbf{y}_s = \mathbf{1}$ if $s = t$ and $\mathbf{y}_s = \mathbf{0}$ if $s \neq t$. In addition, if $\mathbf{X}$ is a stationary point, then $\mathbf{H}$ has at least $p(p-1)/2$ additional eigenvalues equal to zero. Let $\mathbf{Y} = (\mathbf{y}_1|\ldots|\mathbf{y}_p) = \mathbf{XS}$ with $\mathbf{S}$ skewsymmetric, i.e., $\mathbf{S} = -\mathbf{S}'$. Without loss of generality we may assume that $\mathbf{X}$ is centered and of rank p. For any s consider $\sum_{t=1}^{p} \mathbf{H}_{st}\mathbf{y}_t$, which is equal to zero if $\mathbf{y}$ is the eigenvector corresponding to a zero eigenvalue. The first term of (3) becomes zero, because it is multiplied with linear combinations of columns of a stationary point $\mathbf{X}$ and the gradient is zero at stationary points. Furthermore, multiplying the last two terms of (3) by $\mathbf{y}_t$ gives

$$\begin{aligned}
&8\sum_{t=1}^{p}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left((f(d_{ij}(\mathbf{X})) - f(\delta_{ij}))f''(d_{ij}(\mathbf{X})) + (f'(d_{ij}(\mathbf{X})))^2\right)\\
&\mathbf{A}_{ij}\mathbf{x}_s\mathbf{x}_t'\mathbf{A}_{ij}\mathbf{y}_t =\\
&8\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left((f(d_{ij}(\mathbf{X})) - f(\delta_{ij}))\, f''(d_{ij}(\mathbf{X})) + (f'(d_{ij}(\mathbf{X})))^2\right)\\
&\mathbf{A}_{ij}\mathbf{x}_s\sum_{t=1}^{p}\mathbf{x}_t'\mathbf{A}_{ij}\mathbf{y}_t.
\end{aligned} \tag{4}$$

The factor $\sum_{t=1}^{p} \mathbf{x}_t'\mathbf{A}_{ij}\mathbf{y}_t$ can be simplified into tr $(\mathbf{X}'\mathbf{A}_{ij}\mathbf{XS})$, which is zero, since it is the trace of the product of a symmetric and a skewsymmetric matrix. Thus all the terms that constitute $\sum_{t=1}^{p} \mathbf{H}_{st}\mathbf{y}_t$ are equal to zero, which proves that $\mathbf{y}$ is an eigenvector with zero eigenvalue. There are $p(p-1)/2$ linearly independent skewsymmetric matrices $\mathbf{S}$, which lead to linearly independent eigenvectors of the above type. This shows the assertion.

If at a stationary point $\mathbf{X}$ the Hessian $\mathbf{H}$ has exactly $p(p+1)/2$ zero eigenvalues and all other eigenvalues are positive, then we call $\mathbf{X}$ a *strict* local minimum.

3. Differentiability at a local minimum

To investigate differentiability of $\sigma(\mathbf{X}, \Delta)$ we set $f(z) = g(z^{1/2})$, where we assume the function $g : [0, \infty) \to \Re^1$ to be differentiable with right sided

Table 1: The gradients of STRESS, S-STRESS and MULTISCALE.

$f(z)$	$f'(z)$	Name	Gradient
$z^{1/2}$	$\frac{1}{2}z^{-1/2}$	STRESS	$2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(1-\sqrt{\delta_{ij}d_{ij}^{-1}(\mathbf{X})})\mathbf{A}_{ij}\mathbf{X}$
z	1	S-STRESS	$4\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(d_{ij}(\mathbf{X})-\delta_{ij})\mathbf{A}_{ij}\mathbf{X}$
$\log(z)$	z^{-1}	MULTISCALE	$4\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(d_{ij}(\mathbf{X})\delta_{ij}^{-1})d_{ij}^{-1}(\mathbf{X})\mathbf{A}_{ij}\mathbf{X}$

Table 2: The Hessians of STRESS, S-STRESS and MULTISCALE.

$f''(z)$	Name	$\mathbf{H}_{st}$
$-\frac{1}{4}z^{-3/2}$	STRESS	$2\beta^{st}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(1-\delta_{ij}^{1/2}d_{ij}^{-1/2}(\mathbf{X}))\mathbf{A}_{ij}+$ $2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij}^{1/2}d_{ij}^{-3/2}(\mathbf{X})\mathbf{A}_{ij}\mathbf{x}_s\mathbf{x}_t'\mathbf{A}_{ij}$
0	S-STRESS	$4\beta^{st}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(d_{ij}(\mathbf{X})-\delta_{ij})\mathbf{A}_{ij}+$ $8\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\mathbf{A}_{ij}\mathbf{x}_s\mathbf{x}_t'\mathbf{A}_{ij}$
$-z^{-2}$	MULTI-SCALE	$4\beta^{st}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(d_{ij}(\mathbf{X})\delta_{ij}^{-1}))d_{ij}^{-1}(\mathbf{X})\mathbf{A}_{ij}+$ $8\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(1-\log(d_{ij}(\mathbf{X})\delta_{ij}^{-1}))d_{ij}^{-2}(\mathbf{X})\mathbf{A}_{ij}\mathbf{x}_s\mathbf{x}_t'\mathbf{A}_{ij}$

derivative $g'(0)$ at $z = 0$. Then $f(d_{ij}(\mathbf{X})) = g(\tilde{d}_{ij}(\mathbf{X}))$ with

$$\tilde{d}_{ij}(\mathbf{X}) = \left(\sum_{s=1}^{p}(x_{is} - x_{js})^2\right)^{1/2},$$

the distance between points i and j. It is quite natural to assume that $g(0) = 0$, i.e., zero distances and dissimilarities are not transformed to positive values, and $g'(z) \geq 0$ for all $z \geq 0$, which means that the transformation g of distances is monotone. In the limit, $g'(0) = \infty$ is also allowed.

Obviously, if $g'(0) = 0$ then $\sigma(\mathbf{X}, \Delta)$ is differentiable for all $\mathbf{X}$, no matter if zero distances occur. Thus we investigate the case that $g'(0) > 0$, such that nondifferentiable points may be encountered. Examples of corresponding transformations are the utility functions $g(z) = \ln(z+1)$ and $g(z) = 1-e^{-\lambda z}$, $\lambda > 0$, and furthermore the class of functions $g_\lambda(z) = z^\lambda$, $0 < \lambda \leq 1$.

$g_1(z) = z$, e.g., yields STRESS via $f(z) = z^{1/2}$. In this case de Leeuw (1984) has shown that STRESS is differentiable at a local minimum, provided $w_{ij}\delta_{ij} > 0$ for all $i \neq j$. He calls such data *usable*. This result has been extended to arbitrary Minkowski ℓ_p–distances by Groenen, Mathar and Heiser (1992). We follow the basic idea to evaluate directional derivatives of $\sigma(\mathbf{X}, \Delta)$. The directional derivative of σ at $\mathbf{X}$ in direction $\mathbf{Y}$ is defined by

$$\nabla\sigma(\mathbf{X}; \mathbf{Y}) = \lim_{\varepsilon \downarrow 0} \frac{\sigma(\mathbf{X} + \varepsilon\mathbf{Y}, \Delta) - \sigma(\mathbf{X}, \Delta)}{\varepsilon},$$

and always exists if f is differentiable. The directional derivatives of the compositions $f \circ d_{ij} = g \circ \tilde{d}_{ij}$ and $f^2 \circ d_{ij} = g^2 \circ \tilde{d}_{ij}$ are given by

$$\nabla g \circ \tilde{d}_{ij}(\mathbf{X}; \mathbf{Y}) = \begin{cases} g'(\tilde{d}_{ij}(\mathbf{X}))\tilde{d}_{ij}(\mathbf{Y}), & \text{if } \tilde{d}_{ij}(\mathbf{X}) = 0 \\ \frac{g'(\tilde{d}_{ij}(\mathrm{X}))}{\tilde{d}_{ij}(\mathrm{X})} \sum_{s=1}^{p}(x_{is} - x_{js})(y_{is} - y_{js}), & \text{if } \tilde{d}_{ij}(\mathbf{X}) \neq 0, \end{cases}$$

and

$$\nabla g^2 \circ \tilde{d}_{ij}(\mathbf{X}; \mathbf{Y}) = \begin{cases} 2g(\tilde{d}_{ij}(\mathbf{X}))g'(\tilde{d}_{ij}(\mathbf{X}))\tilde{d}_{ij}(\mathbf{Y}), & \text{if } \tilde{d}_{ij}(\mathbf{X}) = 0 \\ \frac{2g(\tilde{d}_{ij}(\mathrm{X}))g'(\tilde{d}_{ij}(\mathrm{X}))\sum_{s=1}^{p}(x_{is}-x_{js})(y_{is}-y_{js})}{\tilde{d}_{ij}(\mathrm{X})}, & \text{if } \tilde{d}_{ij}(\mathbf{X}) \neq 0. \end{cases}$$

For $\mathbf{X} \in \Re^{n\times p}$ define $\mathcal{P} = \{(i,j) \mid i \neq j,\ \tilde{d}_{ij}(\mathbf{X}) \neq 0\}$ and correspondingly $\mathcal{Q} = \{(i,j) \mid i \neq j,\ \tilde{d}_{ij}(\mathbf{X}) = 0\}$. From the above representations we obtain the directional derivative of σ as

$$\begin{aligned} &\nabla\sigma(\mathbf{X}; \mathbf{Y}) \\ &= \sum_{i\neq j} w_{ij}\nabla g^2 \circ \tilde{d}_{ij}(\mathbf{X}; \mathbf{Y}) - 2\sum_{i\neq j} w_{ij} g(\delta_{ij}^{1/2})\nabla g \circ \tilde{d}_{ij}(\mathbf{X}; \mathbf{Y}) \\ &= \sum_{(i,j)\in\mathcal{P}} \frac{2w_{ij}g'(\tilde{d}_{ij}(\mathbf{X}))\sum_{s=1}^{p}(x_{is} - x_{js})(y_{is} - y_{js})}{\tilde{d}_{ij}(\mathbf{X})}[g(\tilde{d}_{ij}(\mathbf{X})) - g(\delta_{ij}^{1/2})] \\ &\quad + \sum_{(i,j)\in\mathcal{Q}} 2w_{ij}g'(\tilde{d}_{ij}(\mathbf{X}))\tilde{d}_{ij}(\mathbf{Y})[g(\tilde{d}_{ij}(\mathbf{X})) - g(\delta_{ij}^{1/2})]. \end{aligned} \tag{5}$$

¿From this it easily follows that for all $\mathbf{X}, \mathbf{Y}$

$$\nabla\sigma(\mathbf{X};\mathbf{Y}) + \nabla\sigma(\mathbf{X};-\mathbf{Y}) = 4\sum_{(i,j)\in\mathcal{Q}} w_{ij}g'(\tilde{d}_{ij}(\mathbf{X}))\tilde{d}_{ij}(\mathbf{Y})[g(\tilde{d}_{ij}(\mathbf{X})) - g(\delta_{ij}^{1/2})].$$

If $\mathbf{X}$ is a local minimum, the directional derivative in all directions is nonnegative. This yields

$$\sum_{(i,j)\in\mathcal{Q}} w_{ij}g'(0)\tilde{d}_{ij}(\mathbf{Y})[g(0) - g(\delta_{ij}^{1/2})] \geq 0 \tag{6}$$

for all $\mathbf{Y}$. Now choose $\mathbf{Y}$ such that $\tilde{d}_{ij}(\mathbf{Y}) > 0$ for all $i \neq j$. Because of the assumptions $g'(0) > 0$ and $g(0) = 0$, (6) can happen for usable data only if $\mathcal{Q} = \emptyset$. Thus, at a local minimum $\mathbf{X}$ it holds that $\tilde{d}_{ij}(\mathbf{X}) > 0$ for all $i \neq j$.

In summary, we have shown the following result: for usable data ($w_{ij}\delta_{ij} > 0$ for all $i \neq j$), for any differentiable transformation g with $g(0) = 0$, $g'(0) > 0$, and $f(z) = g(z^{1/2})$, the general least squares loss function $\sigma(\mathbf{X}, \Delta)$ is differentiable at any local minimum $\mathbf{X}$. If $g'(0) = 0$ then $\sigma(\mathbf{X}, \Delta)$ is differentiable for all $\mathbf{X}$. Thus for usable data STRESS and S-STRESS are differentiable at a local minimum, but for MULTISCALE this need not be so.

4. Fulldimensional scaling

For fulldimensional scaling, where $p = n-1$, every local minimum is a global minimum for some choices of f. This can be seen by using

$$d_{ij}(\mathbf{X}) = \text{tr } \mathbf{A}_{ij}\mathbf{X}\mathbf{X}' = \text{tr } \mathbf{A}_{ij}\mathbf{C} = c_{ii} + c_{jj} - 2c_{ij} \tag{7}$$

with the only requirement that $\mathbf{C}$ is in the cone of positive semi definite double centered symmetric (DCS) matrices. and rewriting (1) as

$$\sigma(\mathbf{C}, \Delta) = \sum_{i=1}^{n}\sum_{j=1}^{n}(f^2(\delta_{ij}) + f^2(\text{tr } \mathbf{A}_{ij}\mathbf{C}) - 2f(\text{tr } \mathbf{A}_{ij}\mathbf{C})f(\delta_{ij})). \tag{8}$$

Suppose that $f(z) \geq 0$ for $z \geq 0$ and that $f(z)$ is concave. This makes the third term of (8) a convex function in $\mathbf{C}$. If additionally $f^2(z)$ is convex, then $\sigma(\mathbf{C}, \Delta)$ is a convex function in $\mathbf{C}$. Thus, minimizing $\sigma(\mathbf{C}, \Delta)$ over $\mathbf{C}$ is minimizing a convex function over a convex set. Then any local minimum is a global minimum. It is easy to see that $f(z) = z$ and $f(z) = z^{1/2}$ satisfies these requirements, so that fulldimensional scaling for STRESS and S-STRESS results in a global minimum. In fact, $f(z) = z^\lambda$ with $\frac{1}{2} \leq \lambda \leq 1$ satisfies the requirement for a global minimum of fulldimensional scaling. Gaffke and Mathar (1989) proposed a special algorithm for S-STRESS with $p = n - 1$ based on cyclic projection.

Critchley (1986) and Bailey and Gower (1990) prove that the rank of the fulldimensional scaling solution of S-STRESS can never be larger than the number of positive eigenvalues of $-\frac{1}{2}\mathbf{J}\Delta\mathbf{J}$, where $\mathbf{J}$ is the centering operator $\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'$. Numerical experiments with fulldimensional scaling of STRESS suggest that the same rank conditions also holds for STRESS, although no proof for this conjecture exists yet. It may even be the case that this assertion holds for all f for which $\sigma(\mathbf{C}, \Delta)$ is a convex function.

5. Inverse scaling

Instead of finding the configurations which are optimal for given dissimilarities, we now look for dissimilarities for which a given configuration is optimal.

Let $f(\delta_{ij}) = f(d_{ij}(\mathbf{X})) - e_{ij}$. Inserting this in (2) gives

$$-\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}e_{ij}f'(d_{ij}(\mathbf{X}))\mathbf{A}_{ij}\mathbf{X}. \tag{9}$$

By substituting $e_{ij} = t_{ij}/(w_{ij}f'(d_{ij}(\mathbf{X})))$ for $i \neq j$ and $e_{ii} = 0$ into (9) we have that $\mathbf{X}$ is a stationary point if the gradient (2) equals zero, or, equivalently, if

$$-\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} t_{ij}\mathbf{A}_{ij}\mathbf{X} = \mathbf{0}. \tag{10}$$

Of course we assume that $f'(d_{ij}(\mathbf{X}))$ exists for all pairs ij with $i \neq j$. Furthermore, we assume that $\mathbf{X}$ is centered, i.e., $\mathbf{X}'\mathbf{1} = \mathbf{0}$.

We should realize that the $\mathbf{A}_{ij}$ form a basis of the space of double centered symmetric (DCS) matrices. Condition (10) simply translates into $\mathbf{T}\mathbf{X} = \mathbf{0}$ such that $\mathbf{T}$ is DCS. But any DCS matrix $\mathbf{T}$ satisfying $\mathbf{T}\mathbf{X} = \mathbf{0}$ and $\mathbf{T}\mathbf{1} = \mathbf{0}$ can be expressed as $\mathbf{K}\mathbf{M}\mathbf{K}'$ with $\mathbf{M}$ symmetric and $(\mathbf{K}|\frac{1}{\sqrt{n}}\mathbf{1})$ an orthonormal basis of the nullspace of $\mathbf{X}'$, i.e., $\mathbf{K}'\mathbf{X} = \mathbf{0}$ and $\mathbf{K}'\mathbf{1} = \mathbf{0}$. If r is the rank of $\mathbf{X}$ then the rank of $\mathbf{K}$ equals $n - r - 1$. Since $\mathbf{M}$ is symmetric there are $(n-r)(n-r-1)/2$ independent solutions. Note that the diagonal elements t_{ii} are left free, so that they can be chosen such that $\mathbf{T}$ becomes DCS.

Since the dissimilarities are required to be nonnegative and $f(\delta_{ij}) \in \Omega$ with $\Omega = \text{range}(f)$ must hold, certain restrictions on t_{ij} are necessary. In particular, it must be ensured that

$$f^{-1}\left(f(d_{ij}(\mathbf{X})) - \frac{t_{ij}}{w_{ij}f'(d_{ij}(\mathbf{X}))}\right) \geq 0 \tag{11}$$

and that

$$f(d_{ij}(\mathbf{X})) - \frac{t_{ij}}{w_{ij}f'(d_{ij}(\mathbf{X}))} \in \Omega. \tag{12}$$

For S-STRESS both requirements lead to $t_{ij} \leq w_{ij} d_{ij}(\mathbf{X})$. The second requirement imposes restrictions on $\mathbf{T}$ for STRESS. Since $\delta_{ij}^{1/2} \geq 0$ we have that $d_{ij}^{1/2}(\mathbf{X}) - t_{ij}/(2w_{ij}d_{ij}^{1/2}(\mathbf{X})) \geq 0$, or, equivalently, $t_{ij} \leq w_{ij}/2$. For MULTISCALE no restriction is needed on t_{ij}, because $\log(\delta_{ij})$ has $\Omega = \Re^1$ and

$$\log(\delta_{ij}) = \log\left(d_{ij}(\mathbf{X})e^{-t_{ij}w_{ij}^{-1}d_{ij}^{-1}(\mathrm{X})}\right) \tag{13}$$

so that the domain of right logarithm is positive for every t_{ij}.

De Leeuw and Groenen (1993) proved for STRESS by making use of the inequality constraints that inverse scaling defines a closed, convex polyhedron that contains the matrix of distances of $\mathbf{X}$. For S-STRESS a similar result can be proved, but not for MULTISCALE.

Thus we can find a dissimilarity matrix for which the gradient is zero, given a configuration $\mathbf{X}$. That only means that $\mathbf{X}$ is a stationary point for any of those dissimilarity matrices obtained by inverse scaling. But $\mathbf{X}$ can be a local minimum, a local maximum or a saddle point. If we wish to find only those dissimilarity matrices for which $\mathbf{X}$ is a strict local minimum, then we have to impose the additional constraint that the Hessian is positive semidefinite, where the only zero eigenvalues are those indicated in section 3. For more details of the STRESS case, we refer to de Leeuw and Groenen (1993).

References:

BAILEY, R., and GOWER, J. C. (1990): Approximating a symmetric matrix. *Psychometrika*, *55*, 665–675.

BROWNE, M. W. (1987): The Young-Housholder algorithm and the least squares multdimensional scaling of squared distances. *Journal of Classification*, *4*, 175–190.

CRITCHLEY, F. (1986): Dimensionality theorems in multidimensional scaling and hierarchical cluster analysis. In: E. Diday, Y. Escoufier, L. Lebart, J. Lepage, Y. Schektman, and R. Tomassone (eds.), Informatics, IV, North-Holland, Amsterdam, 45–70.

DE LEEUW, J. (1984): Differentiability of Kruskal's Stress at a local minimum. *Psychometrika*, *49*, 111–113.

DE LEEUW, J., and GROENEN, P.J.F. (1993): Inverse scaling. Tech. rep. 144, UCLA Statistics Series, Interdivisonal Program in Statistics, UCLA, Los Angeles, California.

DE LEEUW, J., and HEISER, W. J. (1980): Multidimensional scaling with restrictions on the configuration. In: Krishnaiah, P. (ed), *Multivariate Analysis, volume V*. North Holland, Amsterdam, 501–522.

GAFFKE, N., and MATHAR, R. (1989): A cyclic projection algorithm via duality. *Metrika*, *36*, 29–54.

GLUNT, W., HAYDEN, T., and LIU, W.-M. (1991): The embedding problem for predistance matrices. *Bulletin of Mathematical Biology*, *53*, 769–796.

GROENEN, P. J.F., MATHAR, R., and HEISER, W. J. (1992): The majorization approach to multidimensional scaling for Minkowski distances. Tech. rep. RR-92-11, Department of Data Theory, Leiden.

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

KRUSKAL, J. B., YOUNG, F. W., and SEERY, J. (1977): How to use KYST-2, a very flexible program to do multidimensional scaling. Tech. rep. AT&T Bell Laboratories, Murray Hill, New Jersey.

RAMSAY, J. O. (1977): Maximum likelihood estimation in MDS. *Psychometrika*, *42*, 241–266.

STOOP, I., HEISER, W.J., and DE LEEUW, J. (1981): How to use SMACOF-I A. Tech. rep. Department of Data Theory, Leiden.

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