Power-Stress for Multidimensional Scaling

P.J.F. Groenen* J. de Leeuw†

January 1, 2010

Abstract

Several loss functions exist for doing multidimensional scaling. Two important ones are based on the sum of squared differences of distances and dissimilarities (Stress) and on differences of squared distances and squared dissimilarities (S-Stress). The Power-Stress loss function incorporates these loss functions as it takes the sum of squared differences of distances and dissimilarities to some power larger than one. In this paper, we propose a majorization algorithm to minimize the Power-Stress loss function. In the case of the power one (Stress), the new algorithm simplifies to the well-known SMACOF algorithm for MDS. An important advantage of this new algorithm is that as with any majorizing algorithm, a monotonically nonincreasing series of Power-Stress values is obtained that in almost all practical situations ends up in a local minimum.

Keywords: Stress, S-Stress, Multidimensional scaling, iterative majorization.

1 Introduction

In this paper, we study the least-squares multidimensional scaling of power differences. This problem is formalized by minimizing the Power-Stress loss function, that is,

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

*Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (e-mail: groenen@few.eur.nl)

†Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA (e-mail: deleeuw@stat.ucla.edu)

over $\mathbf{X}$. Here $\mathbf{X}$ is an $n \times p$ *configuration*, the w_{ij}'s are known nonnegative *weights*, the δ_{ij}'s are known *dissimilarities*, and $d_{ij}(\mathbf{X})$ is the *Euclidean distance* between rows i and j of $\mathbf{X}$. Thus, we fit distances raised to some power $\lambda \geq 1$ to the dissimilarities raised to the same power. Larger choices of λ leads to emphasizing the fit of larger dissimilarities and conversely the smaller λ to less emphasis on fitting the larger dissimilarities. Note that the summation is done only over the upper triangular elements of the dissimilarity matrix. The weights w_{ij} are assumed to be irreducible, that is, there does not exist two or more subsets of objects such that all weights between objects belonging to different subsets is zero. This assumption avoids the situation where the MDS problem can be split into two or more independent MDS problems.

We need some convenient matrix expressions for the squared Euclidean distance, that is,

$$d_{ij}^2(\mathbf{X}) = (\mathbf{e}_i - \mathbf{e}_j)'\mathbf{X}\mathbf{X}'(\mathbf{e}_i - \mathbf{e}_j) = \text{tr}\,\mathbf{X}\mathbf{A}_{ij}\mathbf{X}, \tag{2}$$

with $\mathbf{e}_i$ and $\mathbf{e}_j$ columns i and j of the $n \times n$ identity matrix and $\mathbf{A}_{ij}$ the matrix

$$\mathbf{A}_{ij} = (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)'. \tag{3}$$

2 The Power-Stress Majorization Algorithm

Let us analyze (1) more closely and first expand this function as

$$\begin{aligned} \sigma(\mathbf{X}) &= \sum_{i<j} w_{ij}\delta_{ij}^{2\lambda} + \sum_{i<j} w_{ij} d_{ij}^{2\lambda}(\mathbf{X}) - 2\sum_{i<j} w_{ij}\delta_{ij} d_{ij}^{\lambda}(\mathbf{X}) \\ &= \eta_\delta + \eta^2(\mathbf{X}) - 2\rho(\mathbf{X}). \end{aligned} \tag{4}$$

Using majorization, we develop an algorithm in steps. First, we majorize $-\rho(\mathbf{X})$, then $\eta^2(\mathbf{X})$, followed by combining the two results int a single algorithm.

2.1 Majorizing $-\rho(\mathbf{X})$

The term $-\rho(\mathbf{X})$ consists of a weighted sum of $-d_{ij}^{\lambda}(\mathbf{X})$ with the weights $w_{ij}\delta_{ij}^{\lambda}$ being nonnegative. The function $d_{ij}^{\lambda}(\mathbf{X})$ is a convex function in $\mathbf{X}$ and raising it to the power $\lambda \geq 1$ keeps in convex. Therefore, $-d_{ij}^{\lambda}(\mathbf{X})$ is a concave function in $\mathbf{X}$. Standard majorizing theory tells that any concave function can be majorized by a linear function in $\mathbf{X}$. The requirement is that at a support point $\mathbf{Y}$, the majorizing function should touch the original (concave) function, so that, the functions values and first derivatives are equal provided the first derivative exists. Let

$$\nabla\left(-d_{ij}^{\lambda}(\mathbf{X})\right) = \frac{\partial\left(-d_{ij}^{\lambda}(\mathbf{X})\right)}{\partial\mathbf{X}} = -\lambda d_{ij}^{\lambda-2}(\mathbf{X})\mathbf{A}_{ij}\mathbf{X}$$

be the first derivative of $\left(-d_{ij}^{\lambda}(\mathbf{X})\right)$. Then, the following majorization inequality holds,

$$-d_{ij}^{\lambda}(\mathbf{X}) \leq -\lambda d_{ij}^{\lambda-2}(\mathbf{Y})\text{tr } \mathbf{X}'\mathbf{A}_{ij}\mathbf{Y} + (\lambda-1)d_{ij}^{\lambda}(\mathbf{Y}) \tag{5}$$

Consequently, multiplying the left and right hand side of (5) by $w_{ij}\delta_{ij}$ and summing over all combinations of i, j gives

$$\begin{aligned} -\rho(\mathbf{X}) &= -\sum_{i<j} w_{ij}\delta_{ij}^{\lambda} d_{ij}^{\lambda}(\mathbf{X}) \\ &\leq -\lambda \sum_{i<j} w_{ij}\delta_{ij}^{\lambda} d_{ij}^{\lambda-2}(\mathbf{Y})\text{tr } \mathbf{X}'\mathbf{A}_{ij}\mathbf{Y} + (\lambda-1)\sum_{i<j} w_{ij}\delta_{ij}^{\lambda} d_{ij}^{\lambda}(\mathbf{Y}) \\ &= -\lambda \sum_{i<j} w_{ij}\delta_{ij}^{\lambda} d_{ij}^{\lambda-2}(\mathbf{Y})\text{tr } \mathbf{X}'\mathbf{A}_{ij}\mathbf{Y} + (\lambda-1)\rho(\mathbf{Y}) \end{aligned} \tag{6}$$

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

The term $\eta^2(\mathbf{X})$ equals the weighted sum of elements $w_{ij}d_{ij}^{2\lambda}(\mathbf{X})$. For the moment, we focus on $d_{ij}^{2\lambda}(\mathbf{X})$ and assume that λ is a positive integer. To derive a majorizing inequality, consider (2). There, the matrix $\mathbf{A}_{ij}$ has largest eigenvalue 2, so that the matrix $\mathbf{A}_{ij} - 2\mathbf{I}$ is negative semidefinite, implying that

$$\begin{aligned} \text{tr } \mathbf{X}(\mathbf{A}_{ij} - 2\mathbf{I})\mathbf{X} &\leq 0 \\ d_{ij}^2(\mathbf{X}) = \text{tr } \mathbf{X}\mathbf{A}_{ij}\mathbf{X} &\leq 2\text{tr } \mathbf{X}'\mathbf{X}. \end{aligned}$$

By rewriting $d_{ij}^{2\lambda}(\mathbf{X})$ and using the previous inequality $\lambda - 1$ times, we obtain the inequality

$$\begin{aligned} d_{ij}^{2\lambda}(\mathbf{X}) &= d_{ij}^2(\mathbf{X})d_{ij}^{2(\lambda-1)}(\mathbf{X}) \\ &= (\mathbf{e}_i - \mathbf{e}_j)'\mathbf{X}\mathbf{X}'(\mathbf{e}_i - \mathbf{e}_j)\left[(\mathbf{e}_i - \mathbf{e}_j)'\mathbf{X}\mathbf{X}'(\mathbf{e}_i - \mathbf{e}_j)\right]^{\lambda-1} \\ &= \text{tr } \mathbf{X}'\mathbf{A}_{ij}\mathbf{X}\left[\mathbf{X}'\mathbf{A}_{ij}\mathbf{X}\right]^{\lambda-1} \\ &\leq 2^{\lambda-1}\text{tr } \mathbf{X}'\mathbf{A}_{ij}\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{\lambda-1} = g_{ij}(\mathbf{X}). \end{aligned} \tag{7}$$

Note that both sides of the inequality above are twice differentiable for $\lambda \geq 1$. Let $\nabla^2 d_{ij}^{2\lambda}(\mathbf{X}) = \mathbf{H}_d$ be the Hessian of $d_{ij}^{2\lambda}(\mathbf{X})$ and $\nabla^2 g_{ij}(\mathbf{X}) = \mathbf{H}_g$ be the Hessian of $g_{ij}(\mathbf{X})$. Then, the inequality $d_{ij}^{2\lambda}(\mathbf{X}) \leq g_{ij}(\mathbf{X})$ implies that $\mathbf{H}_g - \mathbf{H}_d$ is positive semidefinite. Let

$$\begin{aligned} h_{ij}(\mathbf{X}, \mathbf{Y}) &= g_{ij}(\mathbf{X}) - \text{tr } \mathbf{X}'[\nabla g_{ij}(\mathbf{Y}) - \nabla d_{ij}^{2\lambda}(\mathbf{Y})] \\ &\quad + d_{ij}^{2\lambda}(\mathbf{Y}) - g_{ij}(\mathbf{Y}) + \text{tr } \mathbf{Y}'[\nabla g_{ij}(\mathbf{Y}) - \nabla d_{ij}^{2\lambda}(\mathbf{Y})], \end{aligned} \tag{8}$$

where

$$\nabla g_{ij}(\mathbf{X}) = \frac{\partial \left(2^{\lambda-1}\text{tr } \mathbf{X}'\mathbf{A}_{ij}\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{\lambda-1}\right)}{\partial \mathbf{X}} \quad = \quad \lambda 2^{\lambda}\mathbf{A}_{ij}\mathbf{X}(\mathbf{X}'\mathbf{X})^{\lambda-1}$$

and

$$\nabla \left(d_{ij}^{2\lambda}(\mathbf{X})\right) = \frac{\partial \left(d_{ij}^{2\lambda}(\mathbf{X})\right)}{\partial \mathbf{X}} \quad = \quad 2\lambda d_{ij}^{2\lambda-2}(\mathbf{X})\mathbf{A}_{ij}\mathbf{X}.$$

Then, $h_{ij}(\mathbf{X}, \mathbf{Y})$ is a majorizing function of $d_{ij}^{2\lambda}(\mathbf{Y})$. To prove so, we have to establish that (a) $d_{ij}^{2\lambda}(\mathbf{X}) \leq h_{ij}(\mathbf{X}, \mathbf{Y})$ for all $\mathbf{X}$ and (b) $d_{ij}^{2\lambda}(\mathbf{Y}) \leq h_{ij}(\mathbf{Y}, \mathbf{Y})$ at the supporting point $\mathbf{Y}$. Requirement (b) is easily derived by substituting $\mathbf{Y}$ for $\mathbf{X}$ in (8). To prove (a), we use the fact that the Hessian of $h_{ij}(\mathbf{X}, \mathbf{Y})$ is equal to $\mathbf{H}_g$ as $h_{ij}(\mathbf{X}, \mathbf{Y})$ is the sum of $g_{ij}(\mathbf{X})$ and linear or constant terms in $\mathbf{X}$. Consider the difference function $h_{ij}(\mathbf{X}, \mathbf{Y}) - d_{ij}^{2\lambda}(\mathbf{X})$. The Hessian of this difference function equals $\mathbf{H}_g - \mathbf{H}_d$ and is positive semidefinite (proven above) so that the difference function is convex, which only has global minima. A minimum of the difference function is found at $\mathbf{Y}$ because (1) $h_{ij}(\mathbf{Y}, \mathbf{Y}) = d_{ij}^{2\lambda}(\mathbf{Y})$ so that the difference $h_{ij}(\mathbf{Y}, \mathbf{Y}) - d_{ij}^{2\lambda}(\mathbf{Y}) = 0$ and (2) if the gradient of the difference function vanishes at $\mathbf{Y}$. The gradient of $h_{ij}(\mathbf{X}, \mathbf{Y})$ equals

$$\nabla h_{ij}(\mathbf{X}, \mathbf{Y}) \quad = \quad \nabla g_{ij}(\mathbf{X}) - [\nabla g_{ij}(\mathbf{Y}) + \nabla d_{ij}^{2\lambda}(\mathbf{Y})]$$

which is equal to $\nabla d_{ij}^{2\lambda}(\mathbf{Y})$ at $\mathbf{X} = \mathbf{Y}$, hence the gradient

$$\nabla[h_{ij}(\mathbf{Y}, \mathbf{Y}) - \nabla d_{ij}^{2\lambda}(\mathbf{Y})] \quad = \quad \nabla d_{ij}^{2\lambda}(\mathbf{Y}) - \nabla d_{ij}^{2\lambda}(\mathbf{Y}) = \mathbf{0},$$

so it vanishes indeed. These steps prove that

$$d_{ij}^{2\lambda}(\mathbf{X}) \leq h_{ij}(\mathbf{X}, \mathbf{Y}) \tag{9}$$

for all $\mathbf{X}$ and $\mathbf{Y}$, with equality if $\mathbf{X} = \mathbf{Y}$.

Multiplying $d_{ij}^{2\lambda}(\mathbf{X})$ by w_{ij} and summing over i, j gives

$$\begin{aligned}
\eta^2(\mathbf{X}) &= -\sum_{i<j} w_{ij} d_{ij}^{2\lambda}(\mathbf{X}) \\
&\leq \sum_{i<j} w_{ij} h_{ij}(\mathbf{X}) \\
&= \sum_{i<j} w_{ij}(g_{ij}(\mathbf{X}) - \text{tr } \mathbf{X}'[\nabla g_{ij}(\mathbf{Y}) - \nabla d_{ij}^{2\lambda}(\mathbf{Y})] \\
&\quad + d_{ij}^{2\lambda}(\mathbf{Y}) - g_{ij}(\mathbf{Y}) + \text{tr } \mathbf{Y}'[\nabla g_{ij}(\mathbf{Y}) - \nabla d_{ij}^{2\lambda}(\mathbf{Y})]) \\
&= \sum_{i<j} w_{ij}(g_{ij}(\mathbf{X}) - \text{tr } \mathbf{X}'[\nabla g_{ij}(\mathbf{Y}) - \nabla d_{ij}^{2\lambda}(\mathbf{Y})]) + c_h,
\end{aligned} \tag{10}$$

where c_h contains the constant terms in $\mathbf{X}$. The term $\sum_{i<j} w_{ij} g_{ij}(\mathbf{X})$ can be conveniently expressed as

$$\eta^2(\mathbf{X}) \sum_{i<j} w_{ij} g_{ij}(\mathbf{X}) \;=\; 2^{\lambda-1} \text{tr}\, \mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{\lambda-1} \tag{11}$$

with $\mathbf{V} = \sum_{i<j} w_{ij} \mathbf{A}_{ij}$.

2.3 Getting the Update

The update can be derived from the combined majorizing functions of $\eta^2(\mathbf{X})$ and $\rho(\mathbf{X})$.

$$\begin{aligned}
\sigma(\mathbf{X}) \;\leq\; & 2^{\lambda-1} \text{tr}\, \mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{\lambda-1} - \text{tr}\, \mathbf{X}' \left(\sum_{i<j} w_{ij} [\nabla g_{ij}(\mathbf{Y}) - \nabla d_{ij}^{2\lambda}(\mathbf{Y})] \right) \\
& -2\text{tr}\, \mathbf{X}' \left(\lambda \sum_{i<j} w_{ij} \delta_{ij}^{\lambda} d_{ij}^{\lambda-2}(\mathbf{Y}) \mathbf{A}_{ij} \mathbf{Y} \right) + c_h + 2(\lambda-1)\rho(\mathbf{Y}) + \eta_\delta \\
= \; & 2^{\lambda-1} \text{tr}\, \mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{\lambda-1} - \lambda 2^{\lambda} \text{tr}\, \mathbf{X}'\mathbf{V}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{\lambda-1} \\
& -2\lambda \text{tr}\, \mathbf{X}' \left(\sum_{i<j} w_{ij} \left[\delta_{ij}^{\lambda} d_{ij}^{\lambda-2}(\mathbf{Y}) - d_{ij}^{2\lambda-2}(\mathbf{Y}) \right] \mathbf{A}_{ij} \right) \mathbf{Y} + c \\
= \; & 2^{\lambda-1} \text{tr}\, \mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{\lambda-1} - \lambda 2^{\lambda} \text{tr}\, \mathbf{X}'\mathbf{V}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{\lambda-1} \\
& -2\lambda \text{tr}\, \mathbf{X}'\mathbf{B}(\mathbf{Y})\mathbf{Y} + c,
\end{aligned} \tag{12}$$

where $\mathbf{B}(\mathbf{Y}) = \sum_{i<j} w_{ij} \left[\delta_{ij}^{\lambda} d_{ij}^{\lambda-2}(\mathbf{Y}) - d_{ij}^{2\lambda-2}(\mathbf{Y}) \right] \mathbf{A}_{ij}$ and $c = c_h + 2(\lambda - 1)\rho(\mathbf{Y}) + \eta_\delta$. The majorizing function in (12) can also be expressed as

$$2^{\lambda-1} \text{tr}\, \mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{\lambda-1} - \lambda 2^{\lambda} \text{tr}\, \mathbf{X}'\mathbf{V}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{\lambda-1} - 2\lambda \text{tr}\, \mathbf{X}'\mathbf{B}(\mathbf{Y})\mathbf{Y} + c, \tag{13}$$

To minimize the majorizing function, its gradient must be equal to zero yielding the following linear system:

$$\begin{aligned}
& \lambda 2^{\lambda} \mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{\lambda-1} - \lambda 2^{\lambda} \mathbf{V}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{\lambda-1} - 2\lambda \mathbf{B}(\mathbf{Y})\mathbf{Y} = \mathbf{0} \\
& \mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{\lambda-1} \;=\; \mathbf{V}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{\lambda-1} + 2^{1-\lambda} \mathbf{B}(\mathbf{Y})\mathbf{Y} \\
& \mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{\lambda-1} \;=\; \mathbf{Z}
\end{aligned}$$

with $\mathbf{Z} = \mathbf{V}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{\lambda-1} + 2^{1-\lambda}\mathbf{B}(\mathbf{Y})\mathbf{Y}$. To solve (14) for $\mathbf{X}$, consider the singular value decomposition of $\mathbf{V}^{-}\mathbf{Z}$ (with $\mathbf{V}^{-}$ a generalized inverse of $\mathbf{V}$, that is, $\mathbf{V}^{-}\mathbf{Z} = \mathbf{P}\mathbf{\Phi}\mathbf{Q}'$ with $\mathbf{P}'\mathbf{P} = \mathbf{Q}'\mathbf{Q} = \mathbf{I}$ and $\mathbf{\Phi}$ diagonal with nonnegative values. Choosing the update

$$\mathbf{X}^{+} = \mathbf{P}\mathbf{\Phi}^{1/(2\lambda-1)}\mathbf{Q}' \tag{14}$$

satisfies the zero gradient condition (14) so that the majorizing function at the right of (12) is minimized.

2.4 Special Case of $\lambda = 1$: Stress

2.5 Special Case of $\lambda = 2$: S-Stress

Leeuw (1977)

References

Leeuw, J. D. (1977). Applications of convex analysis to multidimensional scaling. In B. V. C. E. Al. (Ed.), *Recent advantages in statistics.* Amsterdam, Netherlands: North Holland Publishing Company.