Minimizing fStress and rStress by Majorizing Gauss-Newton

Jan de Leeuw

January 29, 2026

TBD

Table of contents

Note: This is a working manuscript which will be expanded/updated frequently. All suggestions for improvement are welcome. All Rmd, tex, html, pdf, R, and C files are in the public domain. Attribution will be appreciated, but is not required. The files can be found at https://github.com/deleeuw/rStress

1 Loss Functions

The Multidimensional Scaling (MDS) loss function fStress (Groenen, De Leeuw, and Mathar (1995), De Leeuw (2017)) is defined as

$$\sigma_f(x) := \frac{1}{2}\sum_{k=1}^{K} w_k(f(\delta_k) - f(d_k(x)))^2, \tag{1}$$

with f increasing and differentiable in the open interval $(0, +\infty)$. In (1) the w_k are positive *weights*, the δ_k are known *dissimilarities*. The vector x has the coordinates of an number of points in $\mathbb{R}^p$, and the d_k are Euclidean *distances* between pairs of these points. Metric least squares MDS minimizes fStress over x.

fStress was introduced and studied in Groenen, De Leeuw, and Mathar (1995). No explicit algorithm to minimize it was given, but the paper has formulas for the first and second derivatives. De Leeuw (2017) uses the multivariate Faà di Bruno formula to derive derivatives up to order four. These derivatives can be used in general purpose minimization methods.

An important special case of fStress is rStress (also known as powerStress), which is

$$\sigma_r(x) := \frac{1}{2}\sum_{k=1}^{K} w_k(\delta_k^r - d_k^r(x))^2 \tag{2}$$

Special cases of rStress are Kruskal's stress (Kruskal (1964a), Kruskal (1964b)) with $r = 1$, sstress by Takane, Young, and De Leeuw (1977) with $r = 2$, and logarithmic stress with $r \rightarrow 0$ by Ramsay (1977). There have been various attempts to extend the majorization (or MM) method for MDS (De Leeuw (1977)) to rStress. References and links to various unpublished reports are in De Leeuw (2017). The recent smacofx package (Rusch et al. (In Press)) has R code for the rstressMin() function that implements one of these techniques.

Minimizing of either fStress or rStress over x is a metric MDS problem. The δ_k, and consequently the $f(\delta_k)$, are K known numbers. In non-metric MDS the loss functions are minimized over both x and δ, where δ is constrained to be in a set $\Delta \subseteq \mathbb{R}^K$. For the ordinal version of non-metric MDS, for example, we require $\delta_1 \leq \cdots \leq \delta_K$. Since f is increasing, we can write fStress simply as

$$\sigma_f(x, \delta) := \frac{1}{2}\sum_{k=1}^{K} w_k(\delta_k - f(d_k(x)))^2, \tag{3}$$

where δ is no longer a vector of known dissimilarities, but any vector monotone with the dissimilarities. These transformed or scaled dissimilarities are often called *disparities*. Non-metric rStress is simply (3) with $f(d_k(x)) = d_k^r(x)$.

In order to exclude the trivial solution with $x = 0$ and $\delta = 0$ in addition we impose the normalization constraint

$$\eta^2(\delta) := \frac{1}{2}\sum_{k=1}^K w_k\delta_k^2 = 1. \tag{4}$$

This formulation of non-metric MDS can be generalized to $\delta \in \Delta$, where Δ is a convex cone in $\mathbb{R}^K$. This means we require the disparities to be in the intersection of the cone Δ and the sphere Σ defined by (4).

2 Alternating Least Squares

The technique proposed in this paper to minimize non-metric fStress or rStress is in the Alternating Least Squares (ALS) class. We start with an initial estimate of x. We then minimize σ_f over δ in $\Delta \cap \Sigma$ for the current $d(x)$. The minimizing δ is then used to minimize fStress over x for the current disparities. These two steps are alternated until x and δ do not change any more. Starting in iteration $\nu = 0$ we compute

$$\delta^{(\nu)} = \mathop{\text{argmin}}_{\delta\in\Delta\cap\Sigma} \sigma_f(x^{(\nu)}, \delta), \tag{5a}$$

$$x^{(\nu+1)} = \mathop{\text{argmin}}_{x} \sigma_f(x, \delta^{(\nu)}). \tag{5b}$$

We then increase ν by one and go into the next iteration. And so on, until convergence.

2.1 Normalized Cone Regression

The step (5a) is comparatively easy. We compute the least squares projection of $f(d(x))$ on the cone Δ and then normalize this projection to give it length one. If Δ is the cone of monotone vectors, as in ordinal MDS, the cone projection is *monotone regression*. The fact that projection on the intersection of Δ and Σ is equivalent to normalizing the projection of Δ is due to De Leeuw (1975) and more generally (and more rigorously) to Bauschke, Bui, and Wang (2018). Since this step is the same as in the standard MDS algorithms such as smacof (De Leeuw and Mair (2009), Mair, Groenen, and De Leeuw (2022)) we do not go into details here, and just refer to the literature.

2.2 Majorization

The second step, minimizing over x for fixed current δ, is much more complicated. There is no analytic solution, as in the first step, and minimization requires an iterative process of its own. Thus the second step requires an infinite number of "inner" iterations. As a compromise, we deviate from the strict ALS framework by not minimizing fStress over x for fixed δ, but by merely taking a single one of the "inner" iterations and merely decrease fStress. If this is done judiciously we still obtain a convergent sequence of updates. This is also what is done in other MDS algorithms such as smacof (De Leeuw (1977)) and alscal (Takane, Young, and De Leeuw (1977)).

In smacof the inner iteration step is a majorization step, by now more commonly known as an MM step. Briefly, we find a function $\kappa_f(x;y)$ such that

- $\kappa_f(x,y) \leq \sigma_f(x)$ for all x and y in Ξ, and
- $\kappa_f(y;y) = \sigma_f(y)$ for all y in Ξ.

An inner iteration is of the form

$$x^{(\mu+1)} = \underset{x\in\Xi}{\text{argmin}}\,\kappa(x, x^{(\mu)}) \tag{6}$$

Remember that this inner iterative process goes on within step 2 of each ALS update. Now, from (6),

$$\sigma_f(x^{(\mu+1)}) \leq \kappa_f(x^{(\mu+1)}, x^{(\mu)}) \leq \kappa_f(x^{(\mu)}, x^{(\mu)}) = \sigma_f(x^{(\mu)}).\textbf{(??)} \tag{7}$$

$$

We can approximate $f(d(x))$ near $d(y)$ with

$$f(d_k(x)) \approx f(d_k(y)) + \mathcal{D}f(d_k(y))(d_k(x) - d_k(y)). \tag{8}$$

Define

$$\eta_f(x;y) := \sum_{k=1}^{K} w_k(\delta_k - f(d_k(y)) - \mathcal{D}f(d_k(y))(d_k(x) - d_k(y)))^2. \tag{9}$$

Note that $\eta_f(x;x) = \sigma_f(x)$ for all x and if f is the identity then $\eta_f(x;y) = \sigma_f(x)$ for all x and y. Define

$$\tilde{w}_k(y) := w_k\{\mathcal{D}f(d_k(y))\}^2, \tag{10a}$$

$$\tilde{\delta}_k(y) := \frac{\delta_k - f(d_k(y))}{\mathcal{D}f(d_k(y))} + d_k(y). \tag{10b}$$

Then

$$\eta_f(x;y) = \sum_{k=1}^{K} \tilde{w}_k(y)(\tilde{\delta}_k(y) - d_k(x))^2$$

Now majorize.

$$\eta_f(x;y)=C+\sum_{k=1}^K\tilde w_k(y)d_k^2(x)-2\sum_{k=1}^K\tilde w_k(y)\tilde\delta_k(y)d_k(x)$$

Now if $\delta_k(y)>0$ we use

$$d_k(x)\geq\frac{1}{d_k(y)}\mathrm{tr}\ X'A_kY$$

and if $\delta_k(y)<0$ we use

$$d_k(x)\leq\frac{1}{2}\frac{1}{d_k(y)}\{d_k^2(y)+d_k^2(x)\}$$

Thus

$$\sum_{k=1}^K\tilde w_k(y)\tilde\delta_k(y)d_k(x)\geq\sum_{\tilde\delta_k(y)>0}w_k\frac{\delta_k(y)}{d_k(y)}x'A_ky+\frac{1}{2}\sum_{\tilde\delta_k(y)<0}w_k\frac{\delta_k(y)}{d_k(y)}x'A_kx+C$$

$$V:=\sum_{k=1}^K\tilde w_kA_k, B(y):=\sum_{\tilde\delta_k(y)>0}w_k\frac{\delta_k(y)}{d_k(y)}A_k, H(y):=\frac{1}{2}\sum_{\tilde\delta_k(y)<0}w_k\frac{\delta_k(y)}{d_k(y)}A_k.$$

$$\sigma_f(x)\leq C-2x'B(y)y+x'(V+H(y))x$$

$$\sigma_f(x^+)\approx\eta_f(x^+;x)\leq\eta_f(x,x)=\sigma_f(x).$$

3 Gauss-Newton approximation to rStress

$$\sigma_r(x)\approx\sum_{k=1}^K w_k(\delta_k^r-d_k^r(y)-rd_k^{r-1}(y)(d_k(x)-d_k(y)))^2= \tag{11}$$

$$\sum_{k=1}^K w_k\{rd_k^{r-1}(y)\}^2(\frac{\delta_k^r-d_k^r(y)}{rd_k^{r-1}(y)}-(d_k(x)-d_k(y)))^2 \tag{12}$$

Let

$$\tilde w_k(y):=r^2w_kd_k^{2(r-1)}(y)$$

and

$$\tilde\delta_k(y):=\frac{\delta_k^r-d_k^r(y)}{rd_k^{r-1}(y)}+d_k(y)=\frac{\delta_k^r+(r-1)d_k^r(y)}{rd_k^{r-1}(y)}$$

then

$$\sigma_r(x;y)\approx\sum_{k=1}^K\tilde w_k(y)(\tilde\delta_k(y)-d_k(x))^2$$

which can be minimized by majorization. Also if $x=y$ then $\sigma_r(x;x)=\sigma_r(x)$.

4 rStress

For $r \geq 1$ we have $\delta_k(y) \geq 0$.

For $r = 1$ we have $\tilde{w}_k(y) = w_k$ and $\tilde{\delta}_k(y) = \delta_k$. This is regular smacof.

For $r = 2$ we have $\tilde{w}_k(y) = 4w_k d_k^2(y)$ and $\tilde{\delta}_k(y) = \frac{\delta_k^2 + d_k^2(y)}{2d_k(y)}$.

5 Negative $\tilde{\delta}_k$

If some of the $\tilde{\delta}_k$ are negative we may use the AM/GM inequality for majorization as in Heiser (1991). For now the program gives an error.

6 Nonmetric

In the metric case we have to decide if we want to fit d^r to δ or to δ^r. This can be handled in the R driver. In the non-metric case it does not make a difference which of the two we choose.

$$\sigma(x, \hat{d}) =$$

References

Bauschke, Heinz H., Minh N. Bui, and Xianfu Wang. 2018. "Projecting onto the Intersection of a Cone and a Sphere." *SIAM Journal on Optimization* 28: 2158–88.

De Leeuw, Jan. 1975. "A Normalized Cone Regression Approach to Alternating Least Squares Algorithms." Department of Data Theory FSW/RUL. https://jansweb.netlify.app/publication/deleeuw-u-75-a/deleeuw-u-75-a.pdf.

———. 1977. "Applications of Convex Analysis to Multidimensional Scaling." In *Recent Developments in Statistics*, edited by J. R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, 133–45. Amsterdam, The Netherlands: North Holland Publishing Company.

———. 2017. "Higher Partials of fStress. Who Needs Them ?" 2017. https://jansweb.netlify.app/publication/deleeuw-e-17-r/deleeuw-e-17-r.pdf.

De Leeuw, Jan, and Patrick Mair. 2009. "Multidimensional Scaling Using Majorization: SMACOF in R." *Journal of Statistical Software* 31 (3): 1–30. https://www.jstatsoft.org/article/view/v031i03.

Groenen, Patrick J. F., Jan De Leeuw, and Rudolf Mathar. 1995. "Least Squares Multidimensional Scaling with Transformed Distances." In *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis and Knowledge Organization*, edited by W. Gaul and D. Pfeifer. Berlin, Germany: Springer Verlag. https://jansweb.netlify.app/publication/groenen-deleeuw-mathar-c-95/groenen-deleeuw-mathar-c-95.pdf.

Heiser, W. J. 1991. "A Generalized Majorization Method for Least Squares Multidimensional Scaling of Pseudodistances that May Be Negative." *Psychometrika* 56 (1): 7–27.

Kruskal, Joseph B. 1964a. "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis." *Psychometrika* 29: 1–27.

———. 1964b. "Nonmetric Multidimensional Scaling: a Numerical Method." *Psychometrika* 29: 115–29.

Mair, Patrick, Patrick J. F. Groenen, and Jan De Leeuw. 2022. "More on Multidimensional Scaling in R: smacof Version 2." *Journal of Statistical Software* 102 (10): 1–47. https://www.jstatsoft.org/article/view/v102i10.

Ramsay, James O. 1977. "Maximum Likelihood Estimation in Multidimensional Scaling." *Psychometrika* 42: 241–66.

Rusch, Thomas, Jan De Leeuw, Patrick Mair, and Kurt Hornik. In Press. "Flexible Multidimensional Scaling with the r Package Smacofx." *Journal of Statistical Software*, In Press. https://jansweb.netlify.app/publication/rusch-deleeuw-mair-hornik-a-25/rusch-deleeuw-mair-hornik-a-25.pdf.

Takane, Yoshio, Forrest W. Young, and Jan De Leeuw. 1977. "Nonmetric Individual Differences in Multidimensional Scaling: An Alternating Least Squares Method with Optimal Scaling Features." *Psychometrika* 42: 7–67.