Majorizing Stress Formula Two

Jan de Leeuw - University of California Los Angeles

Started April 19 2024, Version of April 21, 2024

Abstract

Modifications of the smacof algorithm for multidimensional scaling are proposed that provide a convergent majorization algorithm for Kruskal's stress formula two.

Contents

Note: This is a working paper which will be expanded/updated frequently. All suggestions for improvement are welcome.

1 Introduction

The loss function minimized in the current R implementations of the smacof programs for MDS (De Leeuw and Mair (2009), Mair, Groenen, and De Leeuw (2022)) is Kruskal's original *stress* (Kruskal (1964a), Kruskal (1964b)). It is defined as

$$\sigma_1(X) := \frac{\sum\sum w_{ij}(\delta_{ij} - d_{ij}(X)^2)}{\sum\sum w_{ij}d_{ij}^2(X)}. \tag{1}$$

We assume throughout, without loss of generality, that dissimilarities δ_{ij} and the weights w_{ij} are non-negative, and that the weights add up to one. The double summation is over all pairs of indices (i, j) with $i > j$, i.e, over the elements below the diagonal of the matrices Δ, W, and $D(X)$.

In Kruskal (1965) a different loss function was used (in the context of fitting a linear model). In MDS this loss function is

$$\sigma_2(X) := \frac{\sum\sum w_{ij}(\delta_{ij} - d_{ij}(X)^2)}{\sum\sum w_{ij}(d_{ij}(X) - \overline{d}(X))^2}, \tag{2}$$

where

$$\overline{d}(X) = \sum\sum w_{ij}d_{ij}(X). \tag{3}$$

In Kruskal and Carroll (1969), in the section written by Kruskal (p. 652), we see

> In several of my scaling programs, I refer to these expressions as "stress formula one" and "stress formula two", respectively. Historically, stress formula one was the only badness-of-fit function used for some time. Stress formula two has been used more recently and I now tend to recommend it.

Another early adopter (Roskam (1968), p. 34) says

> While the original formula is adequate for completely ordered B-data, we found it is not adequate with completely ordered A-data.

The distinction between A-data and B-data comes from Coombs (1964). For B-data the δ_{ij} are dissimilarties between pairs of elements of a single set, while for A-data they are dissimilarities between two different sets, a row-set and a column-set. Moreover both Kruskal and Roskam found that having the variance of the distances in the denominator of stress has major advantages for conditional A-data, in which only comparisons of dissimilarities with in the same row are meaningful.

In this paper we will extend the theory and algorithm of smacof to stress formula two.

2 Problem

We want to minimize Kruskal's *stress formula two* from (2) over the $n \times p$ configuration matrices X.

It is convenient to have some notation for the numerator and denominator of the two stress formulas.

$$\sigma_R(X) := \sum\sum w_{ij}(\delta_{ij} - d_{ij}(X))^2, \tag{4a}$$
$$\eta_1^2(X) := \sum\sum w_{ij}d_{ij}^2(X), \tag{4b}$$
$$\eta_2^2(X) := \sum\sum w_{ij}(d_{ij}(X) - \overline{d}(X))^2, \tag{4c}$$

Kruskal terms σ_R from definition (4a) the *raw stress*.

There have not been any systematic comparisons of the two stress formulas that I am aware of. Kruskal (in Kruskal and Carroll (1969), p. 652) says

> For any given configuration, of course, stress formula two yields a substantially larger value than stress formula one, perhaps twice as large in many cases. However, in typical multidimensional scaling applications, minimizing stress formula two typically yields very similar configurations to minimizing stress formula one.

We can get some idea about the difference in scale from the results

$$\frac{\sigma_1(X)}{\sigma_2(X)} = \frac{\eta_2^2(X)}{\eta_1^2(X)} \geq \min_X \frac{\eta_2^2(X)}{\eta_1^2(X)} \tag{5}$$

De Leeuw and Stoop (1984) show that in the one-dimensional case with $p = 1$ and with all w_{ij} equal, this implies

$$\sigma_1(X) \geq \frac{1}{3}\frac{n-2}{n}\sigma_2(X). \tag{6}$$

Thus in this special case σ_1 is three to nine times as large as σ_2. In general the bound in equation (6) depends on the weights, on the dimensionality p, and on the order n of the problem.

As a qualitative statement, supported by the computations of De Leeuw and Stoop (1984), we can perhaps say that minimizing σ_1 tends to give optimal configurations in which distances are more equal than those in configurations that minimize σ_2. One thing is for sure, however. If X is a regular simplex in $n-1$ dimensions then σ_2 is not even defined. Or, to put it differently, if all δ_{ij} are equal the minimum of σ_2 in $n-1$ dimensions does not exist.

3 Notation

Now for some notation. As in standard MDS theory (De Leeuw (1977), De Leeuw and Heiser (1977), De Leeuw (1988)) we use the matrices

$$A_{ij} := (e_i - e_j)(e_i - e_j)', \tag{7}$$

where e_i are unit vectors with element i equal to one and the other $n-1$ elements equal to zero. Thus A_{ij} has elements (i,i) and (j,j) equal to $+1$, elements (i,j) and (j,i) equal to -1, and all other elements equal to zero. The usefulness of the A_{ij} in MDS derives mainly from the formula

$$d_{ij}^2(X) = \text{tr } X' A_{ij} X. \tag{8}$$

Using the A_{ij} we now define other matrices, also standard in MDS,

$$V := \sum\sum w_{ij} A_{ij}, \tag{9a}$$

$$B(X) := \sum\sum w_{ij} \frac{\delta_{ij}}{d_{ij}(X)} A_{ij}. \tag{9b}$$

Note that B is a matrix-valued function, not a single matrix. For completeness also define

$$\eta^2(\Delta) := \sum\sum w_{ij} \delta_{ij}^2. \tag{10}$$

Because we are dealing with stress formula two we also need the non-standard definition

$$M(X) := \overline{d}(X) \sum\sum \frac{w_{ij}}{d_{ij}(X)} A_{ij}. \tag{11}$$

In both definitions (9b) and (11) the summation is over pairs (i,j) with $d_{ij}(X) > 0$. Of course we can also omit all pairs from the summation for which $w_{ij} = 0$.

4 Majorization

In this section we construct a convergent majorization algorithm (De Leeuw (1994)) (also known as an MM algorithm, Lange (2016)) to minimize stress formula two.

The first step is to turn the minimization of a ratio of two functions into the iterative minimization of a difference of the two functions. This is a classical trick in fractional programming, usually attributed to Dinkelbach (1967). Define

$$\omega(X,Y) := \sum\sum w_{ij}(\delta_{ij} - d_{ij}(X))^2 - \sigma(Y)\{\sum\sum w_{ij}(d_{ij}(X) - \overline{d}(X))^2\} \tag{12}$$

Lemma 4.1. *If* $\omega(X,Y) < \omega(Y,Y) = 0$ *then* $\sigma(X) < \sigma(Y)$.

Proof. This is embarassingly simple. Direct substitution shows $\omega(X,X) = 0$ for all X. Also $\omega(X,Y) < 0$ if and only if

$$\sum\sum w_{ij}(\delta_{ij} - d_{ij}(X))^2 < \sigma(Y)\{\sum\sum w_{ij}(d_{ij}(X) - \overline{d}(X))^2\} \tag{13}$$

Dividing both sides by $\{\sum\sum w_{ij}(d_{ij}(X) - \overline{d}(X))^2\}$ shows that $\sigma(X) < \sigma(Y)$. □

It follows from lemma 4.1 that if we are in iteration k, with tentative solution $X^{(k)}$, then finding any $X^{(k+1)}$ such that $\omega(X^{(k+1)}, X^{(k)}) < 0$ will decrease stress. We accomplish this by performing one or more majorization steps decreasing $\omega(X, X^{(k)})$.

From definitions (9a), (9b), (10), and (11)

$$\omega(X,Y) = \eta^2(\Delta) + (1 - \sigma(Y))\text{tr } X'VX - 2 \text{ tr } X'B(X)X + \text{tr } X'M(X)X \tag{14}$$

Lemma 4.2. *For all* X *and* Y

$$\textit{tr } X'B(X)X \geq \textit{tr } X'B(Y)Y, \tag{15}$$

with equality if $X = Y$.

Proof. By Cauchy-Schwartz

$$d_{ij}(X) \geq \frac{1}{d_{ij}(Y)}\text{tr } X'A_{ij}Y \tag{16}$$

Multiplying both sides by $w_{ij}\delta_{ij}$ and summing proves the lemma. □

Lemma 4.3. *For all* X *and* Y

$$\textit{tr } X'M(X)X \leq \textit{tr } X'M(Y)X, \tag{17}$$

with equality if $X = Y$.

Proof. Start with the trivial result

$$\sum\sum w_{ij}d_{ij}(X)=\sum\sum \frac{w_{ij}}{d_{ij}(Y)}d_{ij}(X)d_{ij}(Y). \tag{18}$$

By Cauchy-Schwartz

$$\overline{d}(X)\leq\sqrt{\sum\sum \frac{w_{ij}}{d_{ij}(Y)}d_{ij}^2(X)}\sqrt{\sum\sum \frac{w_{ij}}{d_{ij}(Y)}d_{ij}^2(Y)} \tag{19}$$

Squaring both sides proves the lemma. □

We are now ready for the main result.

Theorem 4.1. *Suppose* $\sigma_2(X^{(0)})\leq 1$. *The update*

$$X^{(k+1)}=\{(1-\sigma_2(X^{(k)}))V+\sigma_2(X^{(k)})M(X^{(k)})\}^+B(X^{(k)})X^{(k)} \tag{20}$$

defines a convergent majorization algorithm.

Proof. Using the definitions in equations (9a), (9b), (10), and (11) define

$$\xi(X,Y):=\eta^2(\Delta)+(1-\sigma(Y))\text{tr }X'VX-2\text{tr }X'B(Y)Y+\sigma(Y)\text{tr }X'M(Y)X. \tag{21}$$

From lemmas 4.2 and 4.3 $\omega(X,Y)\leq\xi(X,Y)$ with equality if $X=Y$. In particular

$$\omega(X^{(k+1)},X^{(k)})\leq\xi(X^{(k+1)},X^{(k)}). \tag{22a}$$

The update $X^{(k+1)}$ minimizes $\xi(X,X^{(k)})$ and thus

$$\xi(X^{(k+1)},X^{(k)})\leq\xi(X^{(k)},X^{(k)})=\omega(X^{(k)},X^{(k)}). \tag{22b}$$

Combining equations (22a) and (22b), and using lemma 4.1, shows that also $\sigma_2(X^{(k+1)})\leq\sigma_2(X^{(k)})$. □

In order to guarantee that $X^{(k+1)}$ minimizes $\xi(X,X^{(k)})$ it is sufficient that $\sigma_2(X^{(k)})\leq 1$, because then $(1-\sigma_2(X^{(k)}))V+\sigma_2(X^{(k)})M(X^{(k)})$ is positive semi-definite. But if $\sigma_2(X^{(0)})\leq 1$ then $\sigma_2(X^{(k)})\leq 1$ for all k. Since σ_2 is mainly used in non-metric scaling in which the δ_{ij} are optimal transformations or quantifications the condition $\sigma_2(X^{(k)})\leq 1$ will be automatically satisfied for all k.

5 Derivatives

The derivatives of stress formula two are

$$\mathcal{D}\sigma_2(X) = \frac{\mathcal{D}\sigma_R(X) - \sigma_2(X)\mathcal{D}\eta_2^2(X)}{\eta_2^2(X)} \tag{23}$$

Now

$$\mathcal{D}\sigma_R(X) = -2\sum\sum w_{ij}(\delta_{ij} - d_{ij}(X))\mathcal{D}d_{ij}(X), \tag{24a}$$
$$\mathcal{D}\eta_2^2(X) = 2\sum\sum w_{ij}\mathcal{D}d_{ij}^2(X) - 2\overline{d}(X)\sum\sum w_{ij}\mathcal{D}d_{ij}(X), \tag{24b}$$

and

$$\mathcal{D}d_{ij}(X) = \frac{1}{d_{ij}(X)}A_{ij}X. \tag{25}$$

And thus, using definitions (9a), (9b), and (11)

$$\mathcal{D}\sigma_R(X) = 2(V - B(X))X, \tag{26a}$$
$$\mathcal{D}\eta_2^2(X) = 2(V - M(X))X. \tag{26b}$$

It follows that $\mathcal{D}\sigma_2(X) = 0$ if and only if

$$X = \{(1 - \sigma_2(X))V + \sigma_2(X)M(X)\}^{+}B(X)X. \tag{27}$$

We can summarize the results of our computations in a theorem.

Theorem 5.1. *X is a fixed point of the majorization iterations if and only if $\mathcal{D}\sigma_2(X) = 0$.*

6 Examples

6.1 Ekman

Our first example are the obligatory color data from Ekman (1954). The stress2 program produces the following sequence of stress formula two values and converges in 28 iterations.

```
## itel  1 sold  0.1577255150 snew  0.1321216983
## itel  2 sold  0.1321216983 snew  0.1207395499
## itel  3 sold  0.1207395499 snew  0.1156260670
## itel  4 sold  0.1156260670 snew  0.1135043532
## itel  5 sold  0.1135043532 snew  0.1126543441
## itel  6 sold  0.1126543441 snew  0.1123159800
## itel  7 sold  0.1123159800 snew  0.1121798388
## itel  8 sold  0.1121798388 snew  0.1121239038
## itel  9 sold  0.1121239038 snew  0.1121002964
## itel  10 sold  0.1121002964 snew  0.1120900307
## itel  11 sold  0.1120900307 snew  0.1120854276
## itel  12 sold  0.1120854276 snew  0.1120833009
## itel  13 sold  0.1120833009 snew  0.1120822904
## itel  14 sold  0.1120822904 snew  0.1120817979
## itel  15 sold  0.1120817979 snew  0.1120815523
## itel  16 sold  0.1120815523 snew  0.1120814273
## itel  17 sold  0.1120814273 snew  0.1120813627
## itel  18 sold  0.1120813627 snew  0.1120813287
## itel  19 sold  0.1120813287 snew  0.1120813107
## itel  20 sold  0.1120813107 snew  0.1120813010
## itel  21 sold  0.1120813010 snew  0.1120812957
## itel  22 sold  0.1120812957 snew  0.1120812929
## itel  23 sold  0.1120812929 snew  0.1120812913
## itel  24 sold  0.1120812913 snew  0.1120812904
## itel  25 sold  0.1120812904 snew  0.1120812899
## itel  26 sold  0.1120812899 snew  0.1120812897
## itel  27 sold  0.1120812897 snew  0.1120812895
## itel  28 sold  0.1120812895 snew  0.1120812894
```

The optimum configuration is in figure 1, which can be compared with the solution minimizing raw stress (which is identical up to a scale factor with the solution minimizing stress formula one) in figure 2. The raw stress solution reaches stress formula one equal to 0.5278528 in 32 iterations. The two optimal configurations are virtually identical.

6.2 De Gruijter

The Ekman data have an excellent fit in two dimensions and the optimum configuration is extremely stable over variations in the MDS method. The data from De Gruijter (1967) on the similarities between nine Dutch political parties in 1966 have a worse fit, and much less stability.

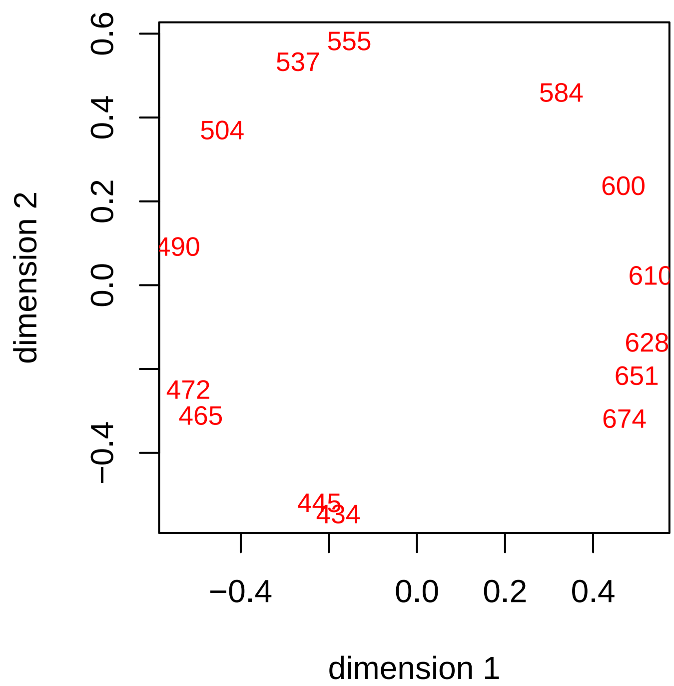

Figure 1: Ekman Metric Stress 2 Solution

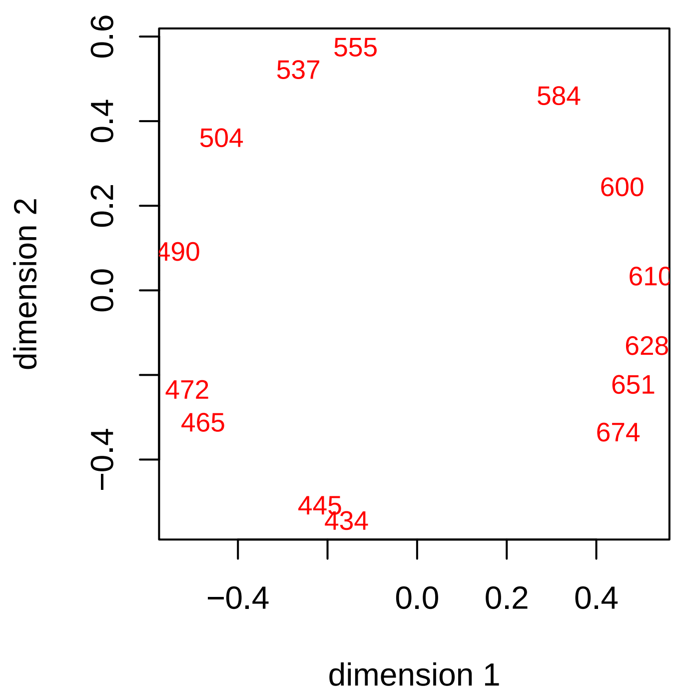

Figure 2: Ekman Metric Raw Stress Solution

The solution minimizing stress formula two has a loss of 0.3482919 and uses 230 iterations. Minimizing raw stress finds stress 9.4408856 and uses 244 iterations. The optimal configurations in figures 3 and 4 are similar, but definitely not the same. Specifically the position of D66 (a "pragmatic" party, ideologically neither left nor right, established in 1966, i.e. in the year of the study) differs a lot between solutions.

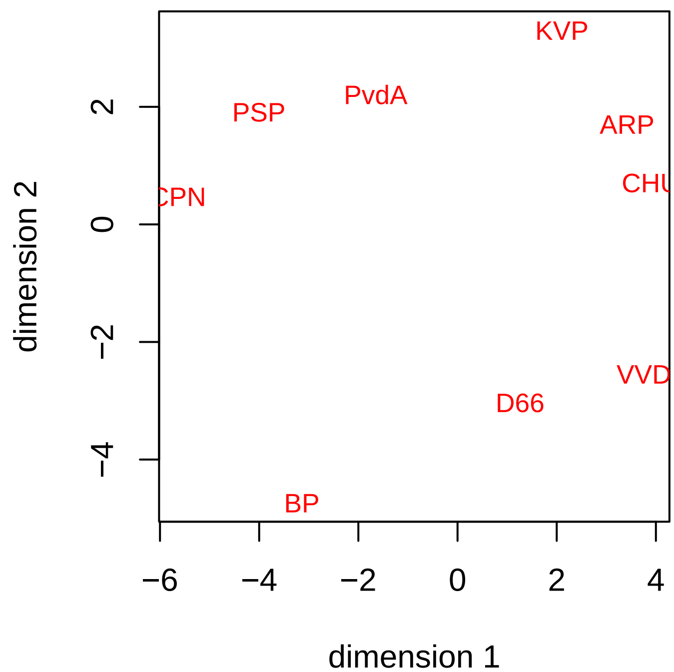

Figure 3: Gruijter Metric Stress 2 Solution

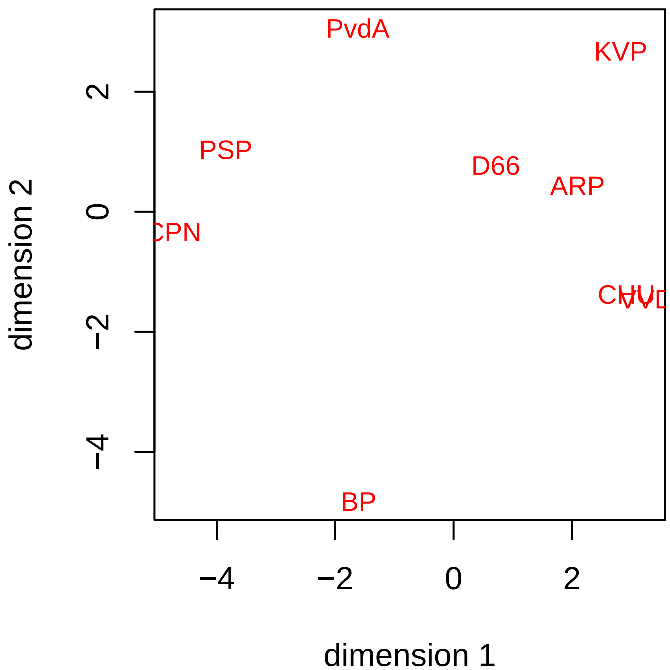

Figure 4: Gruijter Metric Raw Stress Solution

7 Appendix: Code

7.1 stress2.R

```
stress2 <-
  function(delta,
           wmat = 1 - diag(nrow(delta)),
           ndim = 2,
           itmax = 1000,
           eps = 1e-10,
           verbose = TRUE) {
    itel <- 1
    n <- nrow(delta)
    wmat <- wmat / sum(wmat)
    vmat <- -wmat
    diag(vmat) <- -rowSums(vmat)
    xold <- torgerson(delta, ndim)
    dold <- as.matrix(dist(xold))
    enum <- sum(wmat * delta * dold)
    eden <- sum(wmat * dold ^ 2)
    lbda <- enum / eden
    dold <- lbda * dold
    xold <- lbda * xold
    aold <- sum(wmat * dold)
    sold <- sum(wmat * (delta - dold) ^ 2) / sum(wmat * (dold - aold) ^ 2)
    repeat {
      mmat <- -aold * wmat / (dold + diag(n))
      diag(mmat) <- -rowSums(mmat)
      bmat <- -wmat * delta / (dold + diag(n))
      diag(bmat) <- -rowSums(bmat)
      umat <- ((1 - sold) * vmat) + (sold * mmat)
      uinv <- solve(umat + 1/n) - 1/n
      xnew <- uinv %*% bmat %*% xold
      dnew <- as.matrix(dist(xnew))
      anew <- sum(wmat * dnew)
      snew <- sum(wmat * (delta - dnew) ^ 2) / sum(wmat * (dnew - anew) ^ 2)
      if (verbose) {
        cat(
          "itel ",
          formatC(itel, format = "d"),
          "sold ",
          formatC(sold, digits = 10, format = "f"),
          "snew ",
          formatC(snew, digits = 10, format = "f"),
```

```
          "\n"
        )
      }
      if ((itel == itmax) || ((sold - snew) < eps)) {
        break
      }
      sold <- snew
      dold <- dnew
      xold <- xnew
      aold <- anew
      itel <- itel + 1
    }
    return(list(
      x = xnew,
      s = snew,
      d = dnew,
      b = bmat,
      m = mmat,
      w = wmat,
      a = anew,
      u = umat,
      itel = itel
    ))
  }

torgerson <- function(delta, ndim) {
  dd <- delta ^ 2
  rd <- apply(dd, 1, mean)
  rr <- mean(dd)
  cc <- -.5 * (dd - outer(rd, rd, "+") + rr)
  ec <- eigen(cc)
  xx <- ec$vectors[, 1:ndim] %*% diag(sqrt(ec$values[1:ndim]))
  return(xx)
}
```

References

Coombs, C. H. 1964. *A Theory of Data*. Wiley.

De Gruijter, D. N. M. 1967. "The Cognitive Structure of Dutch Political Parties in 1966." Report E019-67. Psychological Institute, University of Leiden.

De Leeuw, J. 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.

———. 1988. "Convergence of the Majorization Method for Multidimensional Scaling." *Journal of Classification* 5: 163–80.

———. 1994. "Block Relaxation Algorithms in Statistics." In *Information Systems and Data Analysis*, edited by H. H. Bock, W. Lenski, and M. M. Richter, 308–24. Berlin: Springer Verlag. https://jansweb.netlify.app/publication/deleeuw-c-94-c/deleeuw-c-94-c.pdf.

De Leeuw, J., and W. J. Heiser. 1977. "Convergence of Correction Matrix Algorithms for Multidimensional Scaling." In *Geometric Representations of Relational Data*, edited by J. C. Lingoes, 735–53. Ann Arbor, Michigan: Mathesis Press.

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

De Leeuw, J., and I. Stoop. 1984. "Upper Bounds for Kruskal's Stress." *Psychometrika* 49: 391–402.

Dinkelbach, W. 1967. "On Nonlinear Fractional Programming." *Management Science* 13: 492–98.

Ekman, G. 1954. "Dimensions of Color Vision." *Journal of Psychology* 38: 467–74.

Kruskal, J. 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.

———. 1965. "Analysis of Factorial Experiments by Estimating Monotone Transformations of the Data." *Journal of the Royal Statistical Society* B27: 251–63.

Kruskal, J. B., and J. D. Carroll. 1969. "Geometrical Models and Badness of Fit Functions." In *Multivariate Analysis, Volume II*, edited by P. R. Krishnaiah, 639–71. North Holland Publishing Company.

Lange, K. 2016. *MM Optimization Algorithms*. SIAM.

Mair, P., P. J. F. Groenen, and J. 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.

Roskam, E. E. 1968. "Metric Analysis of Ordinal Data in Psychology." PhD thesis, University of Leiden.