Multidimensional Scaling of Power-Transformed Dissimilarities

Jan de Leeuw

January 19, 2025

We discuss a non-linear multidimensiopnal scaling method in which the transformations of the dissimilarities are power functions. In the optimal scaling phase of the alternating least squares algorithm the loss function is optimized over the power in the transformation.

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 various files can be found at https://github.com/deleeuw in the smacofPO directory of the repositories smacofCode, smacofManual, and smacofExamples.

1 Introduction

In least squares MDS we minimize *stress*, defined as

$$\sigma(X, r) := \frac{1}{2} \sum_{1 \le i < j \le n} \sum w_{ij} (\hat{d}_{ij} - d_{ij}(X))^2. \tag{1}$$

over the *configurations* $X \in \mathfrak{X} \subseteq \mathbb{R}^{n \times p}$ and over the *disparities* $\hat{D} \in \Delta \subseteq \mathbb{R}^{n \times n}$. (The symbol $:=$ is used for definitions). Assume, without loss of generality, that the *weights* w_{ij} add up to one. The double summations in the definition of stress is over the elements above the diagonal of the symmetric matrices $\hat{D}$ and D.

In metric MDS the set of disparities Δ is the singleton $\{\Delta_0\}$, with Δ_0 the observed *dissimilarities*. In non-metric MDS Δ is the set of all non-decreasing monotone transformations of Δ_0, and in non-linear MDS it is the set of all polynomial or splinical transformations, possibly also required to be montonic. In *additive constant* MDS Δ is the set of all $\hat{D}$ of the form $\Delta_0 + \alpha(E - I)$, where I is the identity and E has all elements equal to one. In *interval* MDS we require $\hat{D}_- \le \hat{D} \le \hat{D}_+$ elementwise, where $\hat{D}_-$ and $\hat{D}_+$ are two given matrices of *disparity bounds*.

In this chapter we study and implement another set Δ, the set of all Δ_0^r, the elementwise powers of the dissimilarities. This definition has some advantages and some disadvantages. Polynomials are often criticized as unsuitable for approximation purposes because of their rigitidy. The values of a polynomial in an interval, however small, determine the shape of the polynomial on the whole real line. This is one of the reasons for the popularity of splines, which are piecewise polynomials joined with a certain degree of smoothness at the knots. Splines are also popular because of their generality: polynomials on an interval are splines without interior knots, while step functions are splines of degree zero.

The set of all monotone functions for Δ, as in the original non-metric proposals of Kruskal (1964) and Guttman (1968), provides a great deal of flexibility. As the case of non-metric unfolding shows, there can actually be too much flexibility, leading to perfect but trivial solutions of the MDS problem.

In terms of flexibility the power MDS studied in this paper performs badly. There is only one single parameter that completely determines the shape of the function on the non-negative real line. But this rigidity can also be seen as an advantage. If the power function fits the data well then it will presumably be quite stable under small perturbations of the data. There are other advantages. Power functions x^r have some nice properties: they always start at the origin and they are monotone, either increasing or decreasing depending on the values of x and r. Moreover for positive powers they are convex, for negative powers they are concave. In psychophysics power functions are prominent because of the work of Stevens (1957) and Luce

(1959). And, perhaps most importantly, in many cases non-metric and non-linear MDS compute optimal transformations that look a lot like power functions, with some irregularities that may be mostly due to measurement errors. Verbally describing what these optimal transformations look like often amounts to "they look like a power function with positive exponent of about two".

2 Loss Function

So let us now redefine stress as

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

and consider the problem of minimizing thus stress over both configurations X and powers r. Throughout the chapter we follow the convention that $0^0 = 1$.

The algorithm we will use is *alternating least squares (ALS)*, i.e. we alternate minimization over X for the current bext value of r and minimization over r for the current best value of X. In this chapter we will only consider the second *optimal scaling* phase of the ALS process, computing the optimal r for given X, because minimizing over X for fixed r is a standard metric MDS problem.

Minimizing stress from Equation 2 differs from the more familiar forms of non-linear and non-metric scaling because the set of admissible optimal scaling is not positively homogeneous. The set of matrices Δ_0^r does not define a cone, let alone a convex cone. It is also worth noting that the matrix $E - I$, with all off-diagonal disparities equal to one, is always in Δ (it corresponds with $r = 0$).

Minimizing Equation 2 over r for given X is similar to two other MDS problems. Historically the first problem is to find the Minkowski power metric that best fits a set of dissimilarities or disparities. We minimize

$$\sigma(X,r) := \frac{1}{2}\sum\sum w_{ij}(\delta_{ij} - d_{ij}^{\{r\}}(X))^2, \tag{3}$$

with

$$d_{ij}^{\{r\}}(X) = \{\sum |x_{is} - x_{js}|^r\}^{1/r}. \tag{4}$$

This particular problem has mainly been used in comparing minimum stress for the city block metric ($r = 1$) and the Euclidean metric ($r = 2$).

A second similar problem is minimization of a form of power stress defined by

$$\sigma(X,r) := \frac{1}{2}\sum\sum w_{ij}(\delta_{ij} - d_{ij}^r(X))^2. \tag{5}$$

Minimizing the loss function in Equation 5 for various values of r has been studied by Groenen and De Leeuw (2010), De Leeuw (2014), De Leeuw, Groenen, and Mair (2016b), De Leeuw, Groenen, and Mair (2016a). For both power stress and Minkovski stress mostly the minimization over X for fixed values of the power r have been considered. Minimization over r is addressed, if at all, by comparing the minimum values of stress over X for different values of r and then choosing or guessing the r corresponding with the smallest value of minimum stress. See, for example, figure 18 in Kruskal (1964).

We can formalize this search strategy using the *marginal function*

$$\sigma_\star(r) := \frac{1}{2}\min_X \sum\sum w_{ij}(\delta_{ij}^r - d_{ij}(X))^2. \tag{6}$$

Also define, for later use,

$$X(r) := \underset{X}{\text{argmin}}\, \sigma(X,r) = \{X \mid \sigma(X,r) = \sigma_\star(r)\}. \tag{7}$$

The idea of the search strategy is to compute the value of the marginal function at a number of values of r, and then interpolate to approximate the minimum over r. There is nothing wrong with this, but it is somewhat ad-hoc and potentially rather expensive. It also supposes, of course, that in computing the marginal function the global minimum over X for given r has been found.

Zero and infinity, the extreme values of r, are of special interest. For $r = 0$ the situation is clear.

$$\sigma(X,0) := \frac{1}{2}\sum\sum w_{ij}(\hat\delta_{ij} - d_{ij}(X))^2. \tag{8}$$

with $\hat\delta_{ij} = 1$. Computing $\sigma_\star(0)$, i.e. minimizing $\sigma(X,0)$ over X, means fitting p-dimensional distances to the distance matrix of an $(n-1)$-dimensional regular simplex. This problem has been studied, in a different context, by De Leeuw and Stoop (1984). They compute $\sigma_\star(0)$ and the corresponding configurations $X(0)$ for various values of the number of objects n and the number of dimensions p. For $n \leq 8$ the optimal configuration has its points equally spaced on a circle, for $n > 8$ points are equally spaced on two or more concentric circles. Of course the minimum is far from unique, because we can permute the points on the circles however we want without changing stress.

If $r \rightarrow +\infty$ limit behavior depends on Δ.

3 Theory

3.1 Derivatives of stress

If $f(r) = x^r$ then

$$\mathcal{D}f(r) = x^r \log x, \tag{9a}$$
$$\mathcal{D}^2 f(r) = x^r (\log x)^2. \tag{9b}$$

It follows that

- if $x < 1$ then f is decreasing,
- if $x > 1$ then f in increasing,
- if $x = 1$ then f is constant,
- f is convex.

Now define

$$\eta^2(r) := \sum\sum w_{ij}\{\delta_{ij}^r\}^2, \tag{10a}$$
$$\rho(r) := \sum\sum w_{ij} d_{ij}(X)\delta_{ij}^r, \tag{10b}$$
$$\omega^2 := \sum\sum w_{ij} d_{ij}^2(X), \tag{10c}$$

so that

$$\sigma(r) = \frac{1}{2}\{\eta^2(r) - 2\rho(r) + \omega^2\}. \tag{11}$$

Now

- both η^2 and ρ are convex,
- if $\delta_{ij} \leq 1$ for all (i, j) then both η^2 and ρ are non-increasing,
- if $\delta_{ij} \geq 1$ for all (i, j) then both η^2 and ρ are non-decreasing.

Using equation (9a) the first derivative of stress is

$$\mathcal{D}\sigma(r) = \sum\sum w_{ij}\delta_{ij}^r \log \delta_{ij}(\delta_{ij}^r - d_{ij}(X)), \tag{12}$$

and using (9b) the second derivative is

$$\mathcal{D}^2\sigma(r) = \sum\sum w_{ij}\delta_{ij}^r (\log \delta_{ij})^2 (2\delta_{ij}^r - d_{ij}(X)) \tag{13}$$

If either $\delta_{ij}\leq 1$ for all (i,j) or $\delta_{ij}\geq 1$ for all (i,j) then all quantities $w_{ij}\delta_{ij}^r\log\delta_{ij}$ have the same sign, and we see that $\mathcal{D}\sigma(r)\geq 0$ if

$$\frac{\sum\sum w_{ij}\delta_{ij}^r|\log\delta_{ij}|\delta_{ij}^r}{\sum\sum w_{ij}\delta_{ij}^r|\log\delta_{ij}|d_{ij}^2(X)}\geq 1.$$

Without any further conditions we have $\mathcal{D}\sigma(r)\geq 0$ if

$$\frac{\sum\sum w_{ij}\delta_{ij}^r(\log\delta_{ij})^2\delta_{ij}^r}{\sum\sum w_{ij}\delta_{ij}^r(\log\delta_{ij})^2d_{ij}^2(X)}\geq\frac{1}{2}.$$

In a decent fit we will have for all or most (i,j)

$$\frac{d_{ij}(X)}{\delta_{ij}^r}\leq 2,(\#eq:decent) \tag{14}$$

and thus $\mathcal{D}^2\sigma(r)\geq 0$.

In an excellent fit $\delta_{ij}^r\approx d_{ij}(X)$ and

$$\mathcal{D}^2\sigma(r)\approx\sum\sum w_{ij}(\delta_{ij}^r\log\delta_{ij})^2,(\#eq:excellent) \tag{15}$$

which is obviously non-negative, and can be used in a Gauss-Newton approximation of stress.

Because of some examples we will discuss later on in this paper the derivatives at $r=0$ are of special interest. First

$$\mathcal{D}\sigma(0)=\sum\sum w_{ij}\log\delta_{ij}(1-d_{ij}(X)),$$

and thus $\mathcal{D}\sigma(0)=0$ if

$$\frac{\sum\sum w_{ij}\log\delta_{ij}d_{ij}(X)}{\sum\sum w_{ij}\log\delta_{ij}}=1.$$

Also

$$\mathcal{D}^2\sigma(0)=\sum\sum w_{ij}(\log\delta_{ij})^2(2-d_{ij}(X)),$$

and thus $\mathcal{D}^2\sigma(0)\geq 0$ if

$$\frac{\sum\sum w_{ij}(\log\delta_{ij})^2d_{ij}(X)}{\sum\sum w_{ij}(\log\delta_{ij})^2}\leq 2.$$

3.2 Marginal Function

Continuous

Directional derivative:

$$\mathfrak{D}\sigma_\star(r) := \lim_{\epsilon\downarrow 0}\frac{\sigma_\star(r+\epsilon)-\sigma_\star(r)}{\epsilon} =$$
$$\mathfrak{D}\sigma_\star(r) = \sum\sum w_{ij}\delta_{ij}^r\log\delta_{ij}(\delta_{ij}^r - d_{ij}(X(r)))$$

Again, the directional derivative at zero is

$$\mathfrak{D}\sigma_\star(0) = \sum\sum w_{ij}\log\delta_{ij}(1-d_{ij}(X(0)))$$

where $X(0)$ is now the metric MDS solution if all dissimilarities are equal to one. This configuration has been studied in detail by De Leeuw and Stoop (1984), where it is shown that for small n we find n points equally spaced on a circle, while for larger n it becomes points equally spaced on several concentric circles.

4 Algorithm

We'll use the R function optimize() to find the optimal power r for fixed X. Using optimize() is safe, but somewhat brute force and probably not efficient. We don't use information from previous iterations, so every iteration has a "cold start". Given the convexity properties of the loss function we could probably use a lightly safeguarded Newton method for efficiency. Also, our algorithm uses only a single Guttman transform per major iteration. Performing more Guttman iterations between upgrades of r may also improve performance.

5 Examples

5.1 Artificial

We start with an aritificial example in which perfect fit is possible. Configuration X consists of 10 points equally spaced on a circle. Define dissimilarities as $\delta_{ij}=d_{ij}^2(X)$. Note that for antipodal points δ_{ij} is as large as four.

```
s <- seq(0, 2 * pi, length = 11)
x <- cbind(sin(s), cos(s))[1:10, ]
delta <- dist(x) ^ 2
harti <- smacofPO(as.matrix(delta), itmax = 1000, verbose = FALSE)
```

Convergence in 70 iterations to stress 1.2290209×10^{-8} and power 0.5000151. smacofPO finds the square root, the inverse of the square.

Next define $\delta_{ij}=\sqrt{d_{ij}(X)}$.

Convergence in 2 iterations to stress 2.2837148×10^{-9} and power 2.0000198. smacofPO finds the squares, the inverse of the square root.

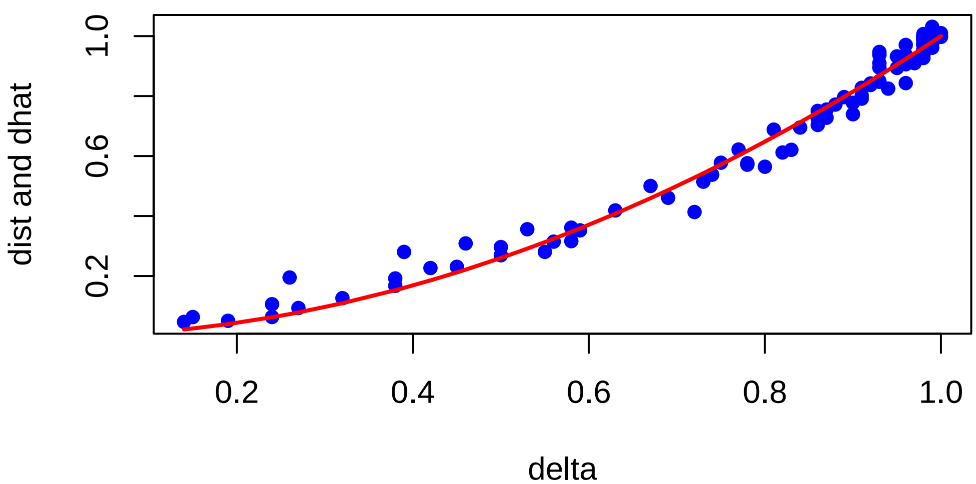
5.2 Ekman (1954)

```
hzero <- smacofPO(as.matrix(ekman), interval = c(0, 0), itmax = 1000, eps = 1e-15, ver
```

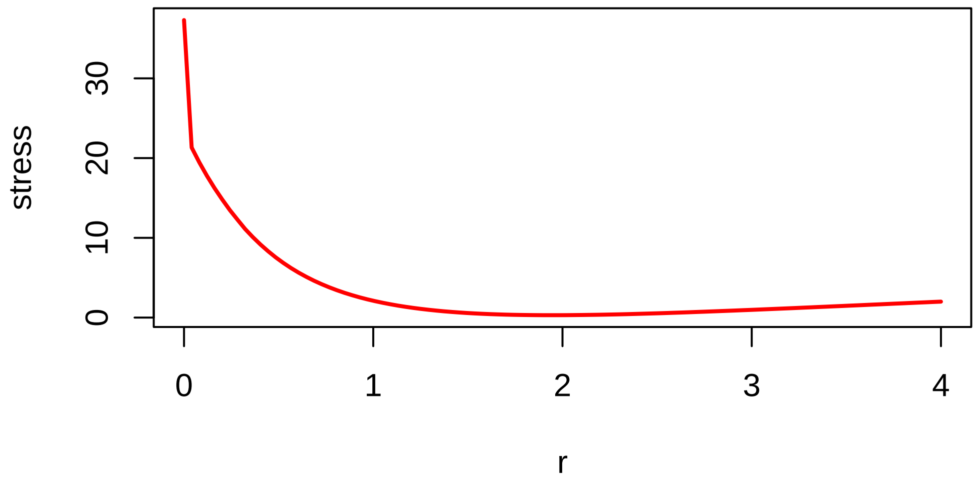
Stress at $r=0$ is 33.382 and the right derivative of the marginal function at zero is -25.5134287. The largest δ_{ij} is 1 and the smallest 0.14.

```
itel   1 sold  0.826422 smid  0.403104 snew  0.402945 pow   2.009280
itel   2 sold  0.402945 smid  0.332216 snew  0.331764 pow   1.993584
itel   3 sold  0.331764 smid  0.310839 snew  0.310361 pow   1.977552
itel   4 sold  0.310361 smid  0.303911 snew  0.303642 pow   1.965610
itel   5 sold  0.303642 smid  0.301562 snew  0.301436 pow   1.957466
itel   6 sold  0.301436 smid  0.300732 snew  0.300677 pow   1.952087
itel   7 sold  0.300677 smid  0.300428 snew  0.300405 pow   1.948579
itel   8 sold  0.300405 smid  0.300313 snew  0.300303 pow   1.946302
itel   9 sold  0.300303 smid  0.300269 snew  0.300264 pow   1.944833
```

```
itel  10 sold  0.300264 smid  0.300251 snew  0.300249 pow   1.943882
itel  11 sold  0.300249 smid  0.300244 snew  0.300243 pow   1.943267
itel  12 sold  0.300243 smid  0.300241 snew  0.300241 pow   1.942870
itel  13 sold  0.300241 smid  0.300240 snew  0.300240 pow   1.942614
itel  14 sold  0.300240 smid  0.300239 snew  0.300239 pow   1.942449
itel  15 sold  0.300239 smid  0.300239 snew  0.300239 pow   1.942343
itel  16 sold  0.300239 smid  0.300239 snew  0.300239 pow   1.942274
itel  17 sold  0.300239 smid  0.300239 snew  0.300239 pow   1.942230
itel  18 sold  0.300239 smid  0.300239 snew  0.300239 pow   1.942201
itel  19 sold  0.300239 smid  0.300239 snew  0.300239 pow   1.942183
itel  20 sold  0.300239 smid  0.300239 snew  0.300239 pow   1.942171
itel  21 sold  0.300239 smid  0.300239 snew  0.300239 pow   1.942163
itel  22 sold  0.300239 smid  0.300239 snew  0.300239 pow   1.942158
itel  23 sold  0.300239 smid  0.300239 snew  0.300239 pow   1.942155
itel  24 sold  0.300239 smid  0.300239 snew  0.300239 pow   1.942153
```



Convergence in 24 iterations to stress 0.3002389 and power 1.9421533.



5.3 De Gruijter (1967)

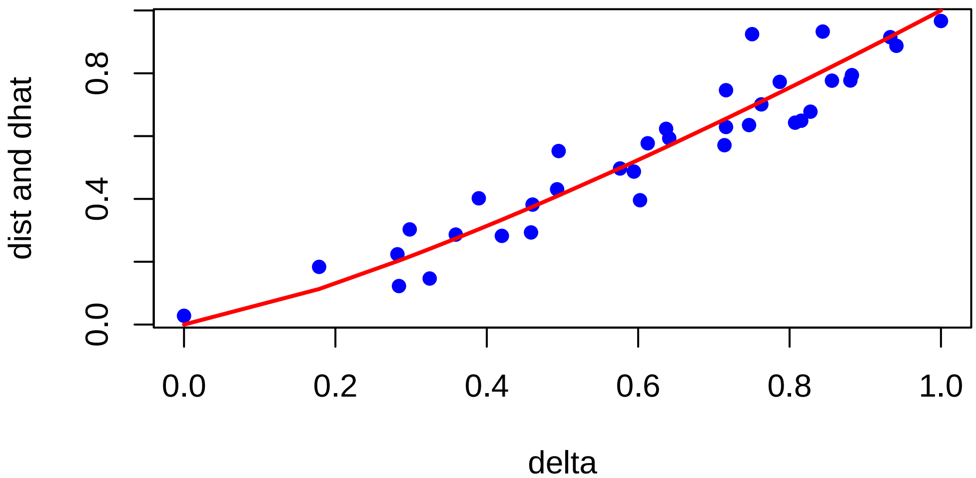
```
hzero <- smacofPO(1 - diag(9), interval = c(0, 0), eps = 1e-15, itmax = 10000, verbose
```

Stress at $r = 0$ is 2074.22 and the right derivative of the marginal function at zero is 13.1124395. The largest δ_{ij} is 8.13 and the smallest 3.2.

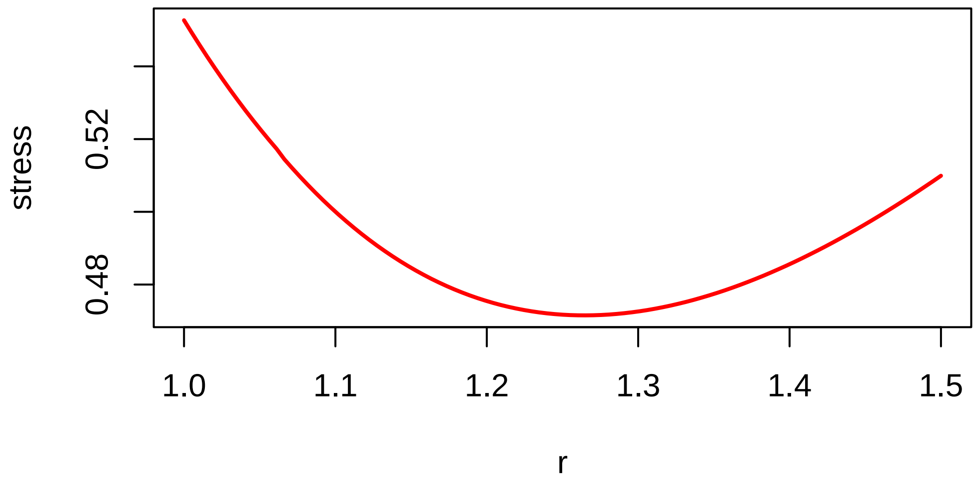
0.5125615

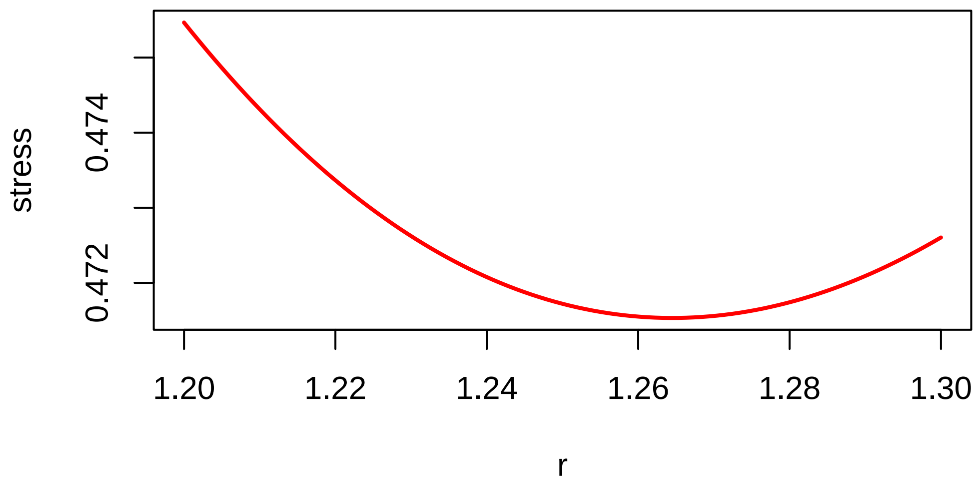
5.3.1 One

```
[1] -Inf
```



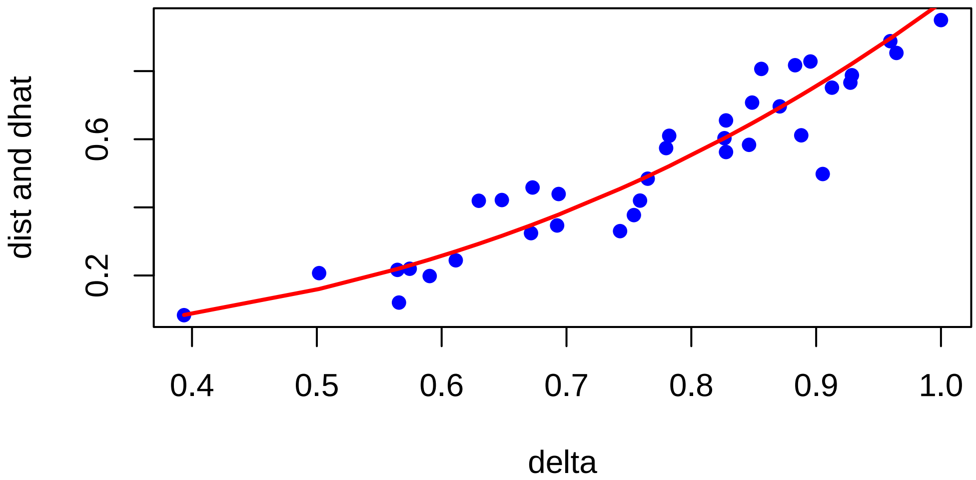
Convergence in 247 iterations to stress 0.4715325 and power 1.2644972.



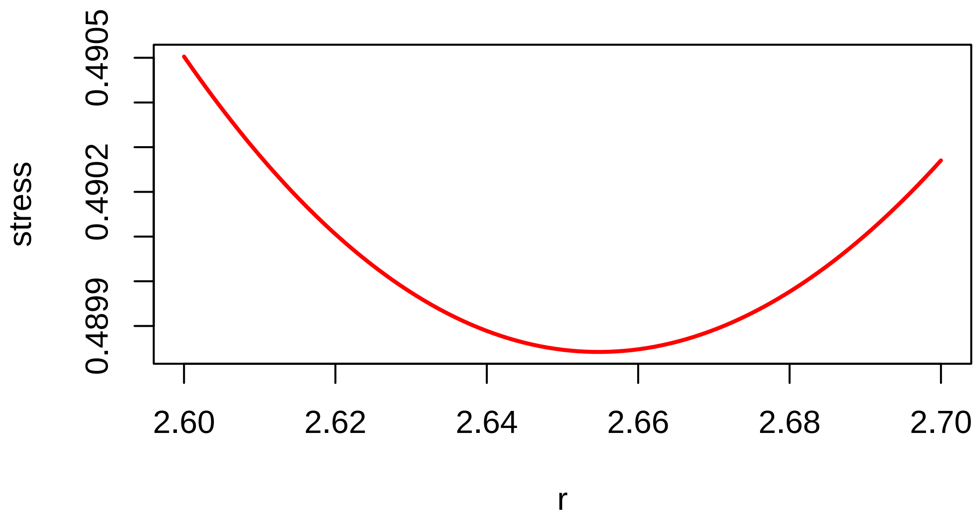


5.3.2 Two

```
[1] -2.427043
```

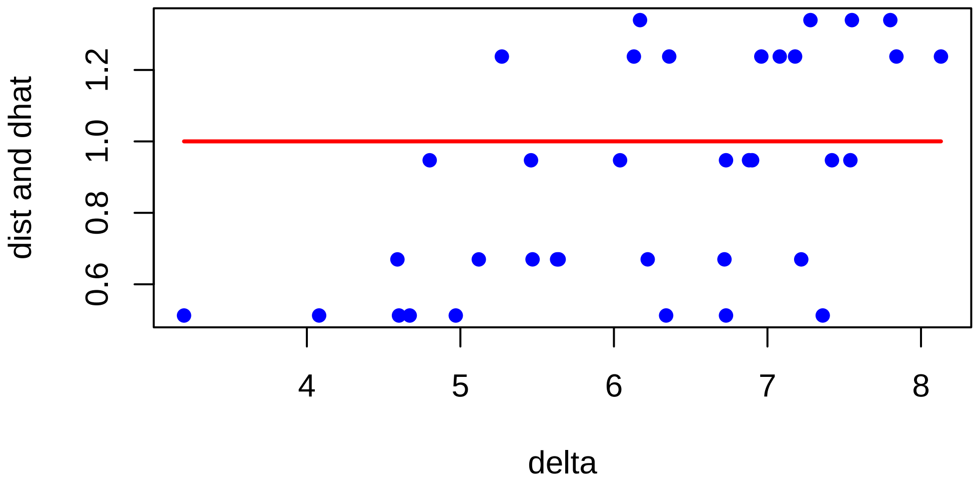


Convergence in 86 iterations to stress 0.4898419 and power 2.6547464.



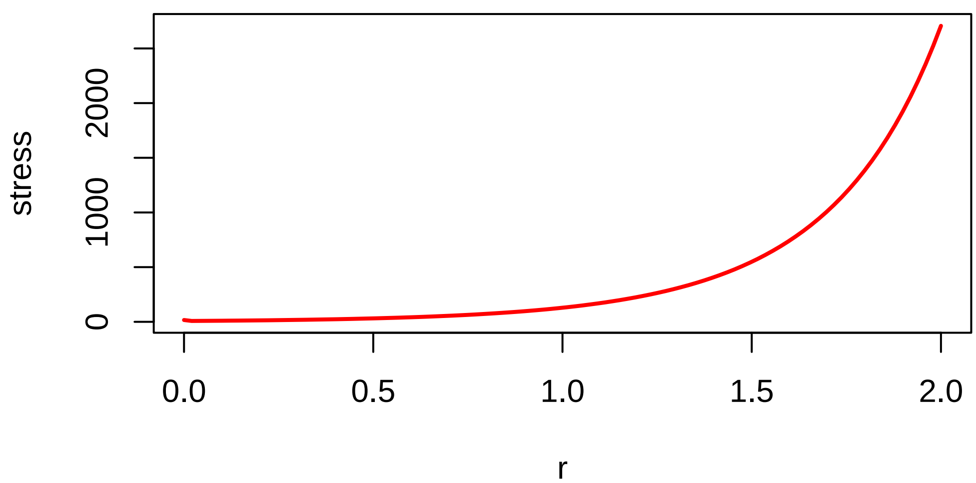
5.3.3 Three

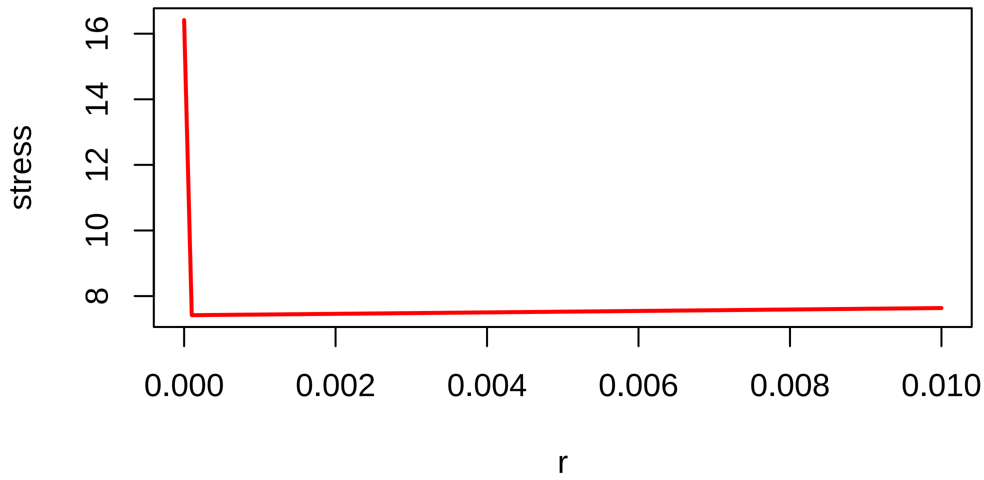
```
[1] 13.11244
```



Convergence in 287 iterations to stress 7.4170928 and power 7.6608779×10^{-5}.

5.4 Deeper



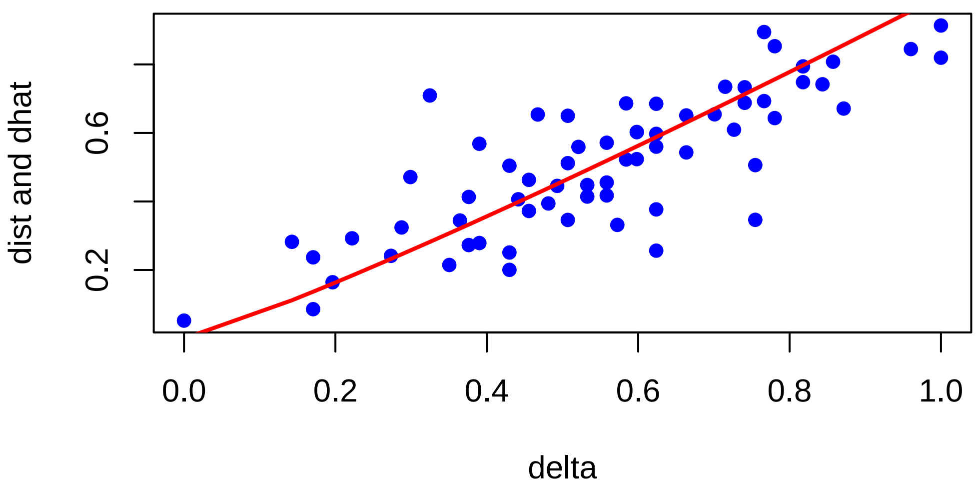


5.5 Wish (1971)

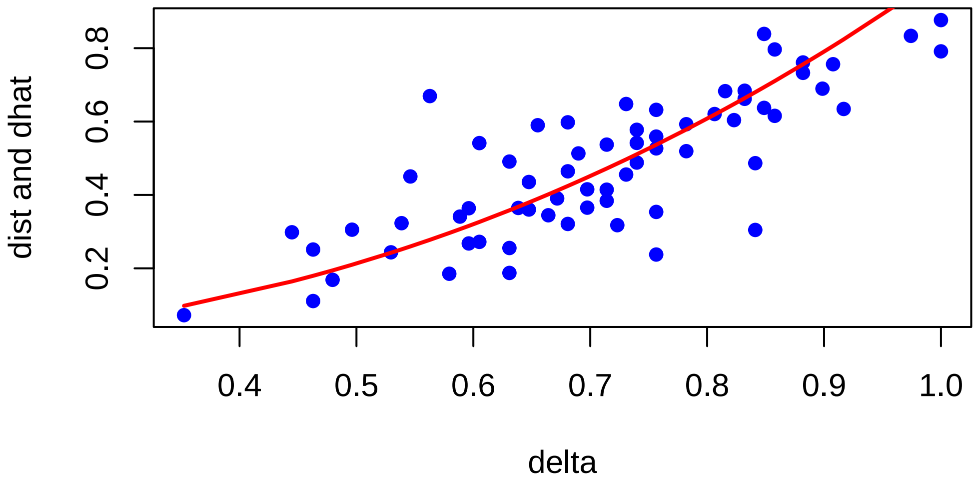
```
hzero <- smacofPO(1 - diag(12), interval = c(0, 0), verbose = FALSE)
```

Stress at $r = 0$ is 1931.9714 and the right derivative of the marginal function at zero is 25.2469345. The largest δ_{ij} is 6.61 and the smallest 2.33.

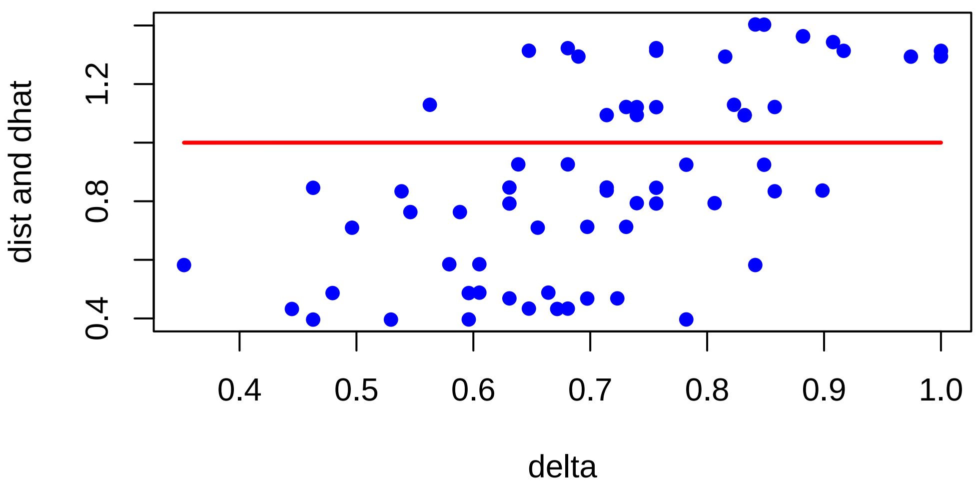
0.4304788



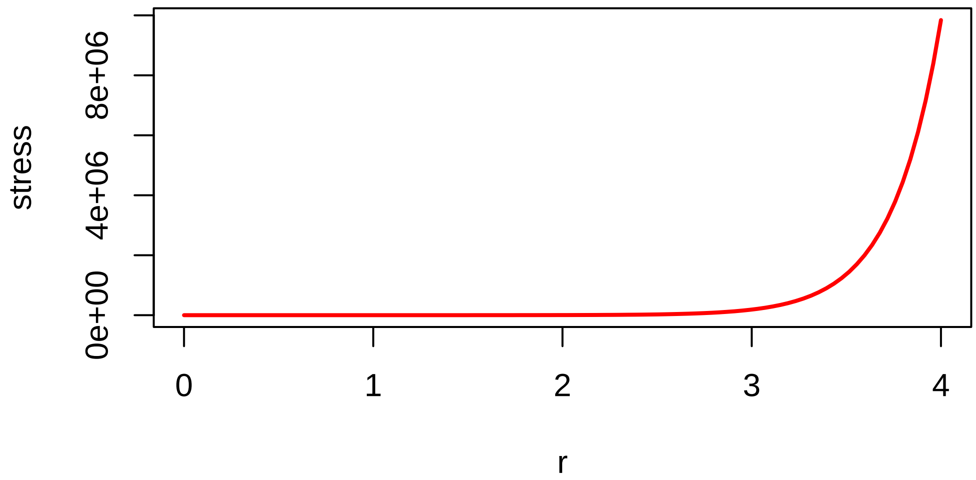
Convergence in 148 iterations to stress 2.3250238 and power 1.1255951.



Convergence in 166 iterations to stress 2.0296426 and power 2.2292659.



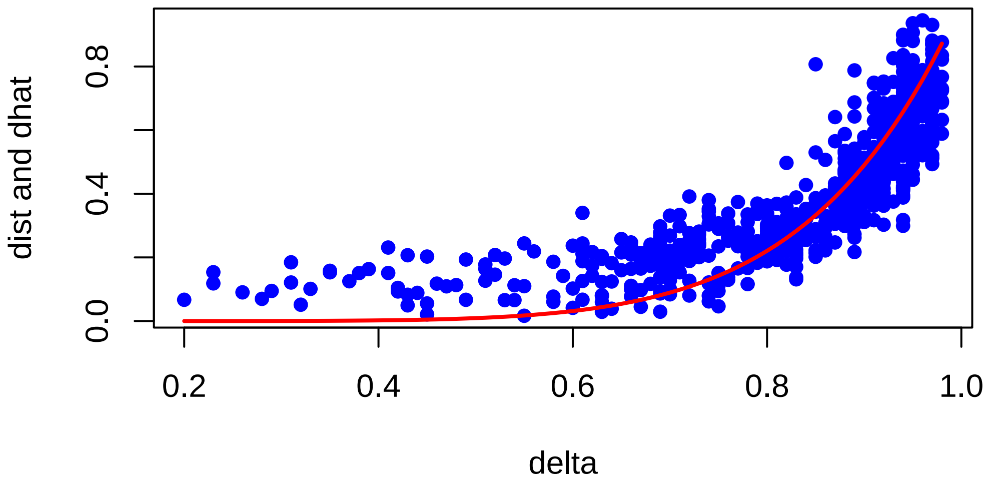
Convergence in 2656 iterations to stress 15.9244051 and power 6.4120229×10^{-5}.



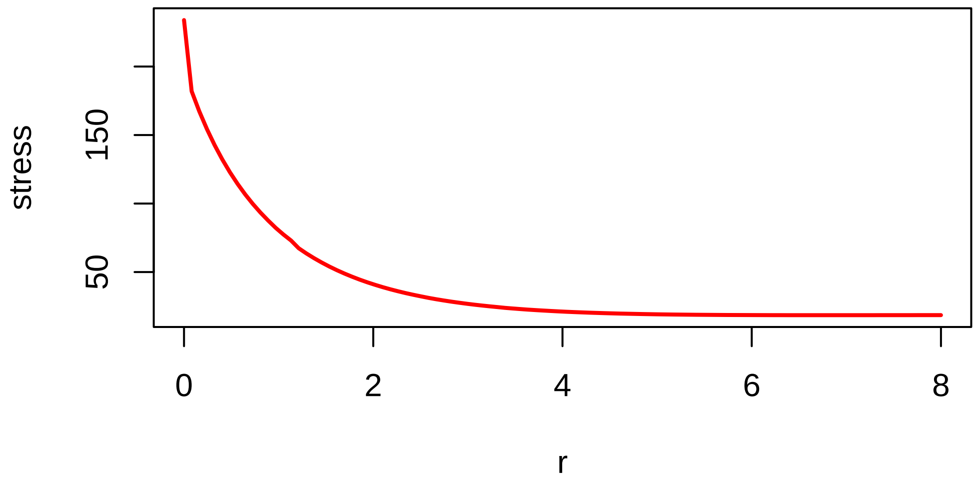
5.6 Rothkopf (1957)

```
hzero <- smacofPO(1 - diag(36), xe = matrix(rnorm(72), 36, 2), interval = c(0, 0), ver
hone <- smacofPO(as.matrix(morse), xe = NULL, interval = c(1, 1), verbose = FALSE, itm
```

Stress at $r = 0$ is 104.2792 and the right derivative of the marginal function at zero is -45.7472567. The largest δ_{ij} is 0.98 and the smallest 0.2. Stress at $r = 1$ is 80.3674371.



Convergence in 253 iterations to stress 18.5193005 and power 6.7735801.


stress
50
150
0
2
4
6
8
r

6 Code

```
smacofPO <-
  function(delta,
           interval = c(0, 4),
           xold = NULL,
           itmax = 1000,
           eps = 1e-10,
           verbose = TRUE) {
    nobj <- nrow(delta)
    dd <- delta ^ 2
    rd <- rowSums(dd) / nobj
    sd <- mean(delta)
    ce <- -.5 * (dd - outer(rd, rd) + sd)
    ee <- eigen(ce)
    xe <- ee$vectors[, 1:2] %*% diag(sqrt(ee$values[1:2]))
    de <- as.matrix(dist(xe))
    if (interval[1] == interval[2]) {
      r <- interval[1]
      fixed <- TRUE
    } else {
      r <- (interval[1] + interval[2]) / 2
    }
    g <- function(r, delta, de) {
      return(sum(((delta ^ r) - de) ^ 2))
    }
    ep <- delta ^ r
    sold <- sum((ep - de) ^ 2)
    itel <- 1
    repeat {
      b <- -ep / (de + diag(nobj))
      diag(b) <- -rowSums(b)
      xe <- (b %*% xe) / nobj
      de <- as.matrix(dist(xe))
      smid <- sum((ep - de) ^ 2)
      if (!fixed) {
        r <- optimize(g, interval = interval, delta = delta, de = de)$minimum
      }
      ep <- delta ^ r
```

```r
    snew <- sum((ep - de) ^ 2)
    if (verbose) {
      cat(
        "itel ",
        formatC(itel, format = "d"),
        "sold ",
        formatC(sold, digits = 6, format = "f"),
        "smid ",
        formatC(smid, digits = 6, format = "f"),
        "snew ",
        formatC(snew, digits = 6, format = "f"),
        "pow  ",
        formatC(r, digits = 6, format = "f"),
        "\n"
      )
    }
    if (((sold - snew) < 1e-10) || (itel == itmax)) {
      break
    }
    itel <- itel + 1
    sold <- snew
  }
  return(list(
    x = xe,
    d = de,
    e = ep,
    r = r,
    itel = itel,
    stress = snew
  ))
}
```

References

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. 2014. "Minimizing rStress Using Nested Majorization." UCLA Department of Statistics. https://jansweb.netlify.app/publication/deleeuw-u-14-c/deleeuw-u-14-c.pdf.

De Leeuw, J., P. Groenen, and P. Mair. 2016a. "Minimizing qStress for Small q." 2016. https://jansweb.netlify.app/publication/deleeuw-groenen-mair-e-16-h/deleeuw-groenen-mair-e-16-h.pdf.

———. 2016b. "Minimizing rStress Using Majorization." 2016. https://jansweb.netlify.app/publication/deleeuw-groenen-mair-e-16-a/deleeuw-groenen-mair-e-16-a.pdf.

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

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

Groenen, P. J. F., and J. De Leeuw. 2010. "Power-Stress for Multidimensional Scaling." https://jansweb.netlify.app/publication/groenen-deleeuw-u-10/groenen-deleeuw-u-10.pdf.

Guttman, L. 1968. "A General Nonmetric Technique for Fitting the Smallest Coordinate Space for a Configuration of Points." *Psychometrika* 33: 469–506.

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

Luce, R. D. 1959. "On the Possible Psychophysical Laws." *Psychologicsal Review* 66 (2): 81–95.

Rothkopf, E. Z. 1957. "A Measure of Stimulus Similarity and Errors in some Paired-associate Learning." *Journal of Experimental Psychology* 53: 94–101.

Stevens, S. S. 1957. "On the Psychophysical Law." *Psychological Review* 64 (3): 153–81.

Wish, M. 1971. "Individual Differences in Perceptions and Preferences Among Nations." In *Attitude Research Reaches New Heights*, edited by C. W. King and D. Tigert, 312–28. American Marketing Association.