LEAST SQUARES METRIC MULTIDIMENSIONAL SCALING

JAN DE LEEUW

Abstract. We study the properties of Kruskal's *stress* loss function used in multidimensional scaling. In particular, we discuss and extend what is known about the local minima of the function.

1. Introduction

In this paper we study the properties of the function

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

Here the w_{ij} and the δ_{ij} are known non-negative numbers, called, respectively, *weights* and *dissimilarities*. X is an unknown $n \times p$ matrix called the *configuration*, and the $d_{ij}(X)$ are the Euclidean distances between the rows of X. Thus

$$d_{ij}(X) = \sqrt{\sum_{s=1}^{p}(x_{is} - x_{js})^2}.$$

Minimizing (1) over all configurations is one form of metric multidimensional scaling (MDS). Nonmetric MDS methods often have to solve a sequence of metric MDS problems The loss function (1) is usually called *stress*, and it was first used in MDS by Kruskal [1964a,b].

1.1. Simplification.

We use some notation first introduced by De Leeuw and Heiser [1980]. If the e_i are unit vectors (columns of the identity matrix) then

$$d_{ij}^2(X) = (e_i - e_j)'XX'(e_i - e_j) = \mathbf{tr}\ X'A_{ij}X$$

Date: March 17, 2005.

2000 *Mathematics Subject Classification.* 62H25,62H30.

Key words and phrases. Multivariate Analysis, Multidimensional Scaling.

with $A_{ij} = (e_i - e_j)(e_i - e_j)'$. Without loss of generality we can assume that

$$\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n w_{ij}\delta_{ij}^2 = 1,$$

and thus

$$\sigma(X) = 1 - \sum_{i=1}^n\sum_{j=1}^n w_{ij}\delta_{ij}\sqrt{\mathbf{tr}\ X'A_{ij}X} + \frac{1}{2}\mathbf{tr}\ X'VX, \tag{2}$$

where

$$V = \sum_{i=1}^n\sum_{j=1}^n w_{ij}A_{ij}.$$

By introducing the matrix-valued function

$$B(X) = \sum_{i=1}^n\sum_{j=1}^n w_{ij}\frac{\delta_{ij}}{d_{ij}(X)}A_{ij},$$

and the shorthand

$$\begin{aligned}\eta^2(X) &= \mathbf{tr}\ X'VX,\\ \rho(X) &= \mathbf{tr}\ X'B(X)X,\end{aligned}$$

we can write

$$\sigma(X) = 1 - \rho(X) + \frac{1}{2}\eta^2(X). \tag{3}$$

Representation (3) shows that *stress* is the difference of a convex quadratic $\eta^2(X)$ and a positively homogeneous non-negative convex function $\rho(X)$. In other words, it is a dc-function [Tao and Souad, 1986]. This is the basis of the majorization algorithm

$$X \leftarrow V^{-1}B(X)X, \tag{4}$$

which was first discussed by Guttman [1968], then shown to be globally convergent by De Leeuw [1977], and to have a linear convergence rate by De Leeuw [1988].

2. Full-dimensional Scaling

If we reformulate the MDS problem in terms of $C = XX'$ we obtain from (2)

$$\sigma(C) = 1 - \sum_{i=1}^n\sum_{j=1}^n w_{ij}\delta_{ij}\sqrt{\mathbf{tr}\ A_{ij}C} + \frac{1}{2}\mathbf{tr}\ VC, \tag{5}$$

which must be minimized over all positive semi-definite C with $\mathbf{rank}(C) \leq p$. The requirement that C is positive semi-definite is also written as $C \gtrsim 0$. It defines a closed convex cone in the space of all $n \times n$ symmetric matrices.

Theorem 2.1. *$\sigma(C)$ is strictly convex on the cone $C \gtrsim 0$.*

Proof. The term $\mathbf{tr}\ VC$ is linear in C. The term $\sqrt{\mathbf{tr}\ A_{ij}C}$ is the square root of a non-negative linear function, and thus it is strictly concave. □

If we ignore the rank constraint, and only require $C \gtrsim 0$, then the corresponding MDS problem is called *full-dimensional scaling* or FDS [De Leeuw, 1993; Leeuw and Groenen, 1993]. We have seen that the FDS problem has no local minima and a unique solution C_F. We call $\mathbf{rank}(C_F)$ the *FDS rank* of the dissimilarities Δ, and write it as $r_F(\Delta)$.

It is clear that the global minimum in an MDS problem with $p \geq r_F(\Delta)$ can be found simply by computing C_F and choosing X such that $C_F = XX'$. This suggests and alternative way to doing metric MDS, which very close to the classical Torgerson method. If we want to compute the $p-$dimensional MDS solution, we first compute C_F and then use the dominant p eigenvalues and corresponding eigenvectors as our solution. This method has no local minimum problems and provides nested solutions. We can compute the FDS solution by applying algorithm (4) with $p \geq r_F(\Delta)$, for instance with $p = n$.

It seems difficult to derive more precise information about the FDS rank. The *Gower Conjecture* is that $r_F(\Delta)$ is less than or equal to the *Torgerson rank* $r_T(\Delta)$, which is the number of positive eigenvalues of $-\frac{1}{2}J_n\Delta^{(2)}J_n$, with J_n the centering operator, and $\Delta^{(2)}$ the matrix of squared dissimilarities.

More insight can be gained by looking at the stationary equations of the FDS problem. Using the theory in Rockafellar [1970, Theorem 31.4], we see that C_F is the unique solution of

$$V - B(C) \gtrsim 0, \tag{6a}$$

$$C \gtrsim 0, \tag{6b}$$

$$C(V - B(C)) = 0. \tag{6c}$$

Thus

$$r_F(\Delta) \leq n - \mathbf{rank}(V - B(C_F)).$$

Theorem 2.2. *If $X = V^+B(X)X$ and $B(X) \lesssim I$ then X is the global minimizer of σ.*

Proof. If X satisfies the conditions in the theorem, then XX' satisfies (6), and thus $XX' = C_F$. □

3. Using a basis of configurations

The X, θ, A parametrizations. Invariance under choice of basis.

It is explained in De Leeuw [1993] that the metric multidimensional scaling problem can be reformulated advantageously by using a basis of configuration matrices. We repeat this argument here. Because of inderterminacy due to rotation and translation the space of configuration matrices has dimension $m = np - \frac{1}{2}p(p+1)$. Suppose $Y_1, \cdots, Y_m$ is a basis for this space, then we can write any X as a linear combination $X = \theta_1 Y_1 + \cdots + \theta_m Y_m$ and we can think of *stress* as a function of θ. In this formulation there are no indeterminacies any more due to rotation and translation, because these have been eliminated by choosing the basis. Moreover we have gained some generality, because we can use the same notation to restrict the configuration to any subspace of the space of configuration matrices by choosing a suitable basis. Thus we can incorporate some of the restrictions discussed, for example, by De Leeuw and Heiser [1980]. In the future m will refer to the number of elements in any basis we have chosen.

We make an additional notational simplification. Use

$$d_{ij}^2(X) = (e_i - e_j)'XX'(e_i - e_j) = \sum_{s=1}^m \sum_{t=1}^m \theta_s\theta_t(e_i - e_j)'Y_sY_t'(e_i - e_j)$$

to define the matrices C_{ij}, of order r, by

$$\{C_{ij}\}_{st} = (e_i - e_j)'Y_sY_t'(e_i - e_j). \tag{7}$$

The matrices C_{ij} are symmetric and positive semi-definite. Moreover $d_{ij}(X) = \sqrt{\theta'C_{ij}\theta}$.

Some additional generality can be gained by not necessarily fitting all dissimilarities, but a selected subset of them. If we do this, we can also get rid of the double indexing, which is just a nuisance. Combining our results so far, we define the (metric, Euclidean) multidimensional scaling problem as minimization of *stress*,

given by

$$\sigma(\theta) = \frac{1}{2}\sum_{i=1}^{n} w_i(\delta_i - \sqrt{\theta' C_i \theta})^2. \tag{8}$$

Minimizing *stress* defined in this way applies to any set of positive semi-definite matrices C_i, although abviously we will be most interested of course in matrices defined by (7).

We continue to simplify the problem somewhat more. Clearly we can suppose without loss of generality that the w_i are positive. In addition, we can use the fact that if we replace the C_i by $\tilde{C}_i = S'C_iS$ we are still solving the same problem. We have, with obvious notation, $\sigma(\theta) = \tilde{\sigma}(S^{-1}\theta)$. This shows we can assume that there is no non-zero vector in the intersection of the null spaces of the C_i. If there was such a vector, we could find an S such that all $\tilde{C}_i$ has zeroes in their last row and column, and we could formulate the problem using a θ with fewer elements.

In fact, we can assume without loss of generality, by choosing S appropriately, that the matrices C_i satisfy

$$\sum_{i=1}^{n} w_i C_i = I.$$

And finally, again without loss of generality, we can assume that the dissimilarities are scaled in such a way that

$$\frac{1}{2}\sum_{i=1}^{n} w_i \delta_i^2 = 1,$$

With these simplification

$$\sigma(\theta) = 1 - \rho(\theta) + \frac{1}{2}\theta'\theta, \tag{9a}$$

where

$$\rho(\theta) = \sum_{i=1}^{n} w_i \delta_i \sqrt{\theta' C_i \theta}. \tag{9b}$$

This is the final form of the MDS loss function and minimization problem we will study in this paper.

4. Convexification

5. Stress on a line

6. Derivatives

One important property of *stress* is that it is not smooth. It is continuous everywhere, but it is not differentiable at points where $\theta'C_i\theta = 0$ for some i. Until further notice we will suppose that θ is such that $\theta'C_i\theta > 0$ for all i, and also that $\delta_i > 0$ for all i. We will deal with the consequences of not making these assumptions in a later section.

For the derivatives of *stress* at θ we find

$$\mathcal{D}\sigma(\theta) = \theta - B(\theta)\theta, \tag{10a}$$

where

$$B(\theta) = \sum_{i=1}^{n} w_i r_i(\theta) C_i, \tag{10b}$$

and

$$r_i(\theta) = \frac{\delta_i}{d_i(\theta)}. \tag{10c}$$

are the *residuals*. Also define $r_+(\theta)$ and $r_-(\theta)$ as the maximum and minimum residual.

We wll first look at stationary points, that is, points where the derivatives are zero. [1]

Theorem 6.1. *If $r_+(\theta) < 1$ or if $r_-(\theta) > 1$ then θ is not a stationary point.*

Proof. θ is a stationary point of stress if $B(\theta)\theta = \theta$, i.e. if θ is an eigenvector with unit eigenvalue of $B(\theta)$. But if $r_+(\theta) < 1$ then $B(\theta) \preceq r_+(\theta)I \prec I$ and thus $B(\theta)$ cannot have an eigenvalue equal to one. In the same way, if $r_-(\theta) > 1$ then $B(\theta) \succeq r_-(\theta)I \succ I$. □

The theorem says that if the distances are all less than or equal to the dissimilarities, with at least one inequality strict, we cannot be at a stationary point. The

[1] We use symbols like $\preceq$ for the usual ordering of matrices. Thus $A \succ B$ means that $A - B$ is positive definite, $A \preceq B$ means that $B - A$ is positive semidefinite, and so on.

same thing is true if the distances are all larger than or equal to the dissimilarities. Observe that the set of θ for which $d_i(\theta) \leq \delta_i$ for all i is a convex intersection of ellipsoids, containing the origin, while the set of θ such that $d_i(\theta) \geq \delta_i$ for all i is the intersection of the complements of these ellipsoids. Thus there is a convex set containing the origin which has no stationary points, and a sphere with center at the origin outside of which there are no stationary points. We will get more information by looking at the second derivatives.

For the second derivatives at θ we find

$$\mathcal{D}^2\sigma(\theta) = I - H(\theta), \tag{11a}$$

where

$$H(\theta) = \sum_{i=1}^n w_i r_i(\theta)\left\{C_i - \frac{C_i\theta\theta' C_i}{\theta' C_i\theta}\right\}. \tag{11b}$$

Here are some simple, and mostly obvious, facts about the second derivatives.

- By Cauchy-Schwartz we see that $0 \lesssim H(\theta) \leq B(\theta)$, for all θ. Thus $I-B(\theta) \leq \mathcal{D}^2\sigma(\theta) \leq I$.
- $H(\theta)\xi = 0$ if and only if $\xi = \theta$, i.e. $H(\theta)$ has only a single zero eigenvalue, and is of rank $p-1$.
- If σ has a local minimum at θ, then $\mathcal{D}^2\sigma(\theta) \geq 0$ and thus $H(\theta) \leq I$. If the local minimum is strict, then $\mathcal{D}^2\sigma(\theta) > 0$ and $H(\theta) < I$.
- For any positive number λ we have $H(\lambda\theta) = \lambda^{-1}H(\theta)$. Thus $\lim_{\lambda\to\infty} \mathcal{D}^2\sigma(\lambda\theta) = I$ and for all sufficiently large λ the Hessian will be positive definite.Moreover if λ is sufficiently small, $\mathcal{D}^2\sigma(\lambda\theta)$ will have one eigenvalue equal to one (corresponding with eigenvector θ), while the other eigenvalues will be negative.

$$(1 - r_+(\theta))I + r_+(\theta)P(\theta) \leq \mathcal{D}^2\sigma(\theta) \leq (1 - r_-(\theta))I + r_-(\theta)P(\theta)$$

where

$$P(\theta) = \sum_{i=1}^n w_i \frac{C_i\theta\theta' C_i}{\theta' C_i\theta}$$

Observe that if $\delta_i = d_i(\theta)$ for all i, that is, if we have *perfect fit*, then $\mathcal{D}^2\sigma(\theta) = P(\theta)$. By Cauchy-Schwartz again, $P(\theta) \leq I$ for all θ.

The new parametrization also makes it possible to calculate third-order partial derivatives.The resulting expression is quite compact, while it obviously will be very unattractive in the original configuration parametrization. Define the vectors

where $\eta^i = C_i\theta$. The elements of C_i are written as $c^i_{\alpha\beta}$. We also use $e_i(\theta)$ for $\theta'C_i\theta$, that is, for $d_i^2(\theta)$.

$$\frac{\partial^3\sigma}{\partial\theta_\alpha\partial\theta_\beta\partial\theta_\gamma} = -\frac{\partial h_{\alpha\beta}(\theta)}{\partial\theta_\gamma} = \\ -\sum_{i=1}^n w_i\frac{r_i(\theta)}{e_i(\theta)}\left[\eta_{i\gamma}(\theta)c_{i\alpha\beta} - \eta_{i\alpha}(\theta)c_{i\beta\gamma} - \eta_{i\beta}(\theta)c_{i\alpha\gamma} + \frac{\eta_{i\alpha}(\theta)\eta_{i\beta}(\theta)\eta_{i\gamma}(\theta)}{e_i(\theta)}\right],$$

and thus

$$\sum_{\alpha=1}^n\sum_{\beta=1}^n\sum_{\gamma=1}^n\xi_\alpha\xi_\beta\xi_\gamma\frac{\partial^3\sigma}{\partial\theta_\alpha\partial\theta_\beta\partial\theta_\gamma} = \sum_{i=1}^n w_i r_i(\theta)\left(\frac{\xi'C_i\theta}{\theta'C_i\theta}\right)\left[\xi'C_i\xi - \frac{(\xi'C_i\theta)^2}{\theta'C_i\theta}\right].$$

Compare this with

$$\sum_{\alpha=1}^n\sum_{\beta=1}^n\xi_\alpha\xi_\beta h_{\alpha\beta}(\theta) = \sum_{i=1}^n w_i r_i(\theta)\left[\xi'C_i\xi - \frac{(\xi'C_i\theta)^2}{\theta'C_i\theta}\right].$$

7. Directional Derivatives

Although distance $d_i(\theta) = \sqrt{\theta'C_i\theta}$ is not differentiable at zero, the one-sided directional derivatives

$$\nabla d_i(\theta,\xi) = \lim_{\lambda\downarrow 0}\frac{d_i(\theta+\lambda\xi) - d_i(\theta)}{\lambda}$$

exist everywhere. In fact

$$\nabla d_i(\theta,\xi) = \begin{cases} d_i(\xi) & \text{if } d_i(\theta) = 0,\\ \xi'C_i\theta/d_i(\theta) & \text{otherwise.}\end{cases}$$

Suppose $I_0(\theta)$ is the set of indices for which $\theta'C_i\theta = 0$, and $I_+(\theta)$ is the rest. Then extend the definition of B, given before in (10b).

$$B(\theta) = \sum_{i=1}^n w_i s_i(\theta)C_i,$$

where

$$s_i(\theta) = \begin{cases} r_i(\theta) & \text{if } i\in I_+(\theta),\\ \text{arbitrary} & \text{if } i\in I_0(\theta).\end{cases}$$

Using this B, the Guttman transform is generalized as $\Gamma(\theta) = B(\theta)\theta$. Observe that the Guttman transform is the same, no matter what we choose for the $s_i(\theta)$ with $i\in I_0(\theta)$. With these definitions

$$\nabla\sigma(\theta,\xi) = \xi'(\theta-\Gamma(\theta)) - \sum_{i\in I_0(\theta)} w_i\delta_i d_i(\xi) \tag{12}$$

Using the directional derivative we can prove two interesting results, orginally due to De Leeuw [1993] and De Leeuw [1984]. The proofs are more elegant and shorter, because of the alternative parametrization of the MDS problem.

Theorem 7.1. *σ has a local maximum at zero, and no other local maxima.*

Proof. At zero we have $\nabla\sigma(0,\xi) = -\rho(\xi) \leq 0$, so *stress* decreases in every direction. Suppose $\theta \neq 0$ is a local maximum. Then $\sigma(\theta + \lambda\theta)$ should have a local maximum at $\lambda = 0$. But $\sigma(\theta+\lambda\theta)$ is a convex quadratic in λ, which means it cannot have a local maximum. □

Theorem 7.2. *If $\delta_i > 0$ for all i, then $d_i(\theta) > 0$ at a local minima.*

Proof. If σ has a local minimum at θ the one-sided directional derivatives in all directions are non-negative. Thus, from (12), a necessary condition for a local minimum is that θ is a fixed point of the Guttman transform, as before, and that $\sum_{i\in I_0(\theta)} w_i\delta_i d_i(\xi) = 0$. Suppose $d_i(\theta) = 0$ and $\delta_i > 0$ at a local minimum. Choose ξ such that $d_i(\xi) > 0$. Then $\nabla\sigma(\theta,\xi) \leq -w_i\delta_i d_i(\xi) < 0$, which means ξ is a descent direction, and θ canot be a local minimum. □

8. Subgradients

9. Homogeneity

We have seen that the stationary equations $\mathcal{D}(\theta) = 0$ define the non-linear eigenvalue problem $B(\theta)\theta = \theta$. The relationships with eigenvalue problems can be made even more explicit.

Theorem 9.1. *Suppose $\hat\theta$ maximizes $\rho(\theta)$ over the sphere $\theta'\theta = 1$, or, equivalently, over the ball $\theta'\theta \leq 1$. Then*

$$\ddot\theta = \left[\frac{\rho(\hat\theta)}{\hat\theta'\hat\theta}\right]\hat\theta$$

minimizes $\sigma(\theta)$.

Proof. We see that for $\lambda \geq 0$

$$\sigma(\lambda\theta) = 1 - \lambda\rho(\theta) + \frac{1}{2}\lambda^2\theta'\theta,$$

and consequently

$$\min_{\lambda\geq 0}\sigma(\lambda\theta) = 1 - \frac{\rho^2(\theta)}{\theta'\theta}.$$

Thus

$$\min_{\theta}\sigma(\theta)=1-\left[\max_{\theta}\frac{\rho(\theta)}{\sqrt{\theta'\theta}}\right]^2.$$

Because ρ is positive homogeneous, maximizing the ratio is equivalent to maximizing ρ over the unit sphere. □

10. Nonpositive dissimilarities and/or weights

11. Pictures of Stress

We can make our results more specific if we look at the case in which θ only has two elements. In an MDS context, this means we look at configurations in the two-dimensional subspace $X = \theta_1 X_i + \theta_2 X_2$. We first look at contourplots of σ as a function of θ_1 and θ_2. On left in Figure 1 we see a global picture of the function, very much like a convex quadratic valley with a single little hill in the center. The valley is not equally deep everywhere, and on the right in Figure 1 we zoom in on the deepest spot.

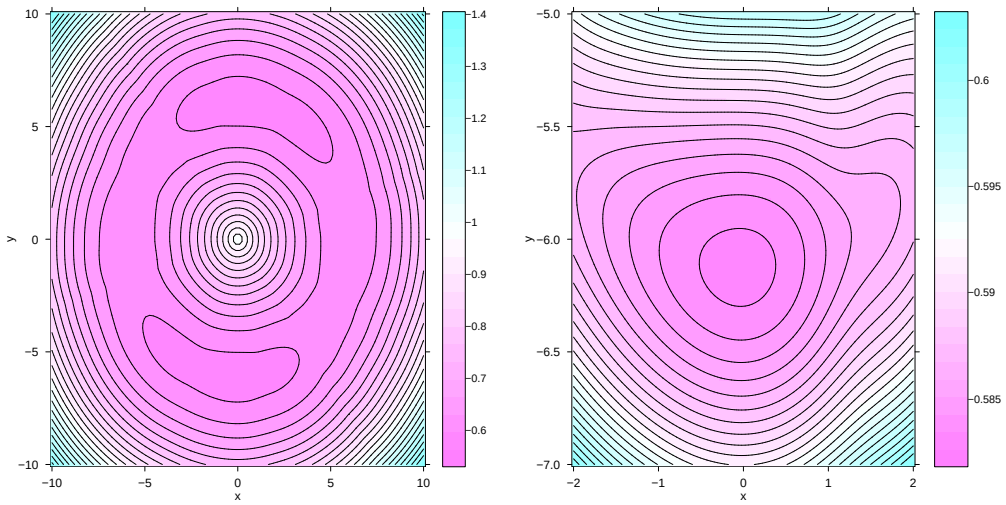

Figure 1. Contourplots

The same two regions are plotted in Figure 2, this time as wireframe plots. We basically see the same qualitative features, in a slightly different form.

This suggest another way to plot the function. We draw the contour lines of ρ and see where they intersect the unit circle. The contour lines will define concentric

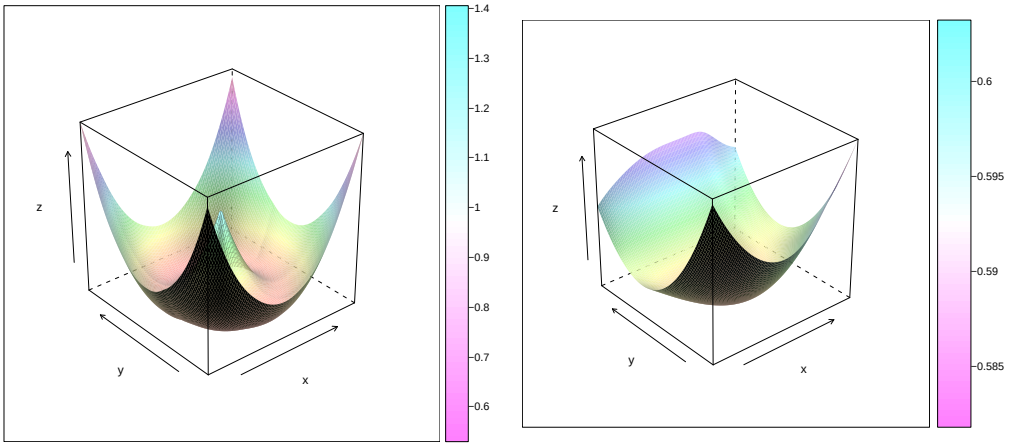

FIGURE 2. Wireframes

convex sets, and we can find the largest contour still intersecting the circle to define the maximum of ρ, and thus the minimum of σ. This is illustrated in Figure 3. On the left we see the global picture again, with the contour lines of ρ drawn at 0.1 intervals. The picture shows that the maximum of ρ on the unit circle is a little over 0.90, and we zoom in on the right by drawing contour lines at 0.01 intervals, showing that the maximum ρ is about 0.92, corresponding wih a σ equal to 0.15.

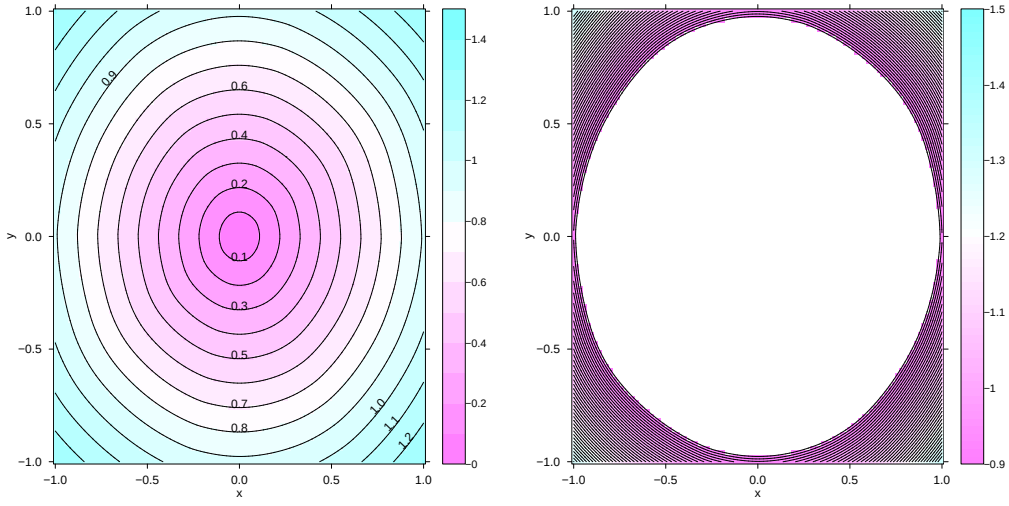

FIGURE 3. Contours of Rho

This representation is still somewhat wasteful, since we are really only interested in the θ on the unit circle. Thus we can take $\theta_1 = \sin(\zeta)$ and $\theta_2 = \cos(\zeta)$ and plot ρ as a function of $0 \leq \zeta \leq \pi$. This is shown in Figure 4, where again we see that the maximum of ρ is about 0.92. Observe there is a local minimum of ρ where it is not differentiable.

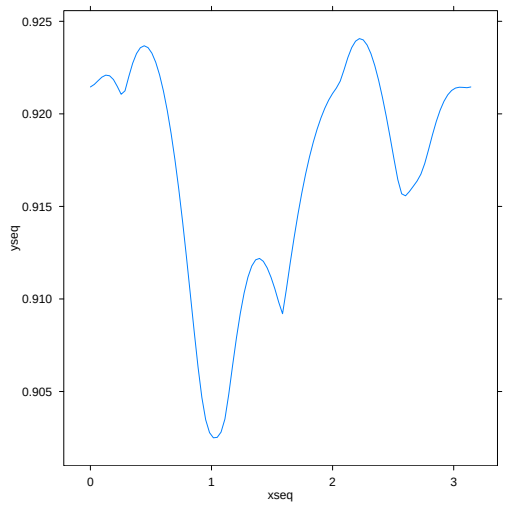

FIGURE 4. Rho on the Circle

12. INVERSE MDS

In the inverse MDS problem we have a given θ and we are looking for the δ_i for which this θ is a local minimum. Thus instead of solving $B(\theta)\theta = \theta$ for given δ we solve for δ for given θ. Define $u_i = C_i\theta$, and collect the u_i as columns in an $m \times n$ matrix U. Now find v such that $Uv = 0$.

Theorem 12.1. *The general solution for δ is $\delta_i = d_i(\theta)(1 + \frac{1}{w_i}v_i)$.*

13. MAJORIZATION

If we introduce the *Guttman transform* [Guttman, 1968] by $\Gamma(\theta) = B(\theta)\theta$, then θ is stationary if it is equal to its Guttman transform (that is, if it is a *fixed point* of the Guttman transform).

The stationary equations $\theta = \Gamma(\theta)$ suggest the iterative algorithm

$$\theta^{(k+1)} = \Gamma(\theta^{(k)}). \tag{13}$$

This was first proposed in Guttman [1968], but a convergence proof was not given until De Leeuw [1977]. The proof was based on the represenation in the previous section, which reduces the problem to maximizing a ratio of norms. For maximizing such a ratio, the algorithm of Robert [1967] is available and immediately shows convergence.

A simpler proof for the Euclidean case was developed by De Leeuw and Heiser [1977] and further streamlined by De Leeuw and Heiser [1980].

We can minorize ρ by a family of linear functions and majorize it by a family of quadratic functions.

Theorem 13.1. *For all θ, ξ we have*

$$\xi' B(\xi)\theta \leq \rho(\theta) \leq \frac{1}{2}\{\theta' B(\xi)\theta + \rho(\xi)\}$$

with equality on both sides if and only if $\theta = \xi$.

Proof. This first inequality follows from applying Cauchy-Schwartz to $\sqrt{\theta' C\theta}$, the second from applying the arithmetic-geometric mean inequality. □

$$\begin{aligned}\sigma(\theta) &\leq \sigma(\xi) + (\theta - \xi)'\mathcal{D}\sigma(\xi) + \frac{1}{2}\max_{0\leq\lambda\leq 1}(\theta - \xi)'\mathcal{D}^2(\lambda\xi + (1-\lambda)\theta)(\theta - \xi)\\ &\leq \sigma(\xi) + (\theta - \xi)'\mathcal{D}\sigma(\xi) + \frac{1}{2}(\theta - \xi)'(\theta - \xi).\end{aligned}$$

and thus

$$\begin{aligned}\sigma(\theta) \leq \sigma(\xi) &- \frac{1}{4}(\xi - \mathcal{D}\sigma(\xi))'(\xi - \mathcal{D}\sigma(\xi))\\ &+ \frac{1}{2}(\theta - (\xi - \mathcal{D}\sigma(\xi))'(\theta - (\xi - \mathcal{D}\sigma(\xi))\end{aligned}$$

Now use $\xi - \mathcal{D}\sigma(\xi) = \Gamma(\xi)$ to obtain

$$\sigma(\theta) \leq \sigma(\xi) - \frac{1}{4}\Gamma(\xi)'\Gamma(\xi) + \frac{1}{2}(\theta - \Gamma(\xi))'(\theta - \Gamma(\xi))$$

Our alternative parametrization also makes it easier to apply the basic theorem by Ostrowski [1966], and simplifies the results of De Leeuw [1988]. We write$\|A\|_\infty$ for the sup-norm of a square matrix, i.e. for the modulus of the largest eigenvalue. Remember that a point of attraction of

Theorem 13.2. *Suppose $\|H(\theta)\|_\infty < 1$. Then θ is a point of attraction of the iteration* (13)*. Moreover, convergence is linear, with rate $\|H(\theta)\|_\infty$.*

Proof. Ortega and Rheinboldt [1970, section] □

14. Newton's Method

The update iterations for Newton's method take a simple form in our current parametrization.

$$\theta^{(k+1)} = \theta^{(k)} - [I - H(\theta^{(k)})]^{-1}(\theta^{(k)} - B(\theta^{(k)})\theta^{(k)}) = [I - H(\theta^{(k)})]^{-1}\Gamma(\theta^{(k)})$$

Nonmonotone line search. Safeguarding. If H is small, then Newton is majorization.

15. Example

For our examples we will analyze mapping four points in two dimensions. As a basis for the configuration matrices we use five matrices, code in R to compute the basis is in the appendix.

The first example takes all dissimilarities equal. Our algorithms converge to three types of stationary points. The global minimum are four points in the corners of a square. Stress is .0286 and the smallest eigenvalue of the Hessian is 0.1595. There is also a non-isolated local minimum formed by three points in the corners of an equilateral triangle, and the fourth point in the centroid of the triangle. This has stress 0.0670, and the smallest eigenvalue of the Hessian is zero (which is why this is a non-isolated minimum). Finally there is a saddle point, with the four points equally spaced on a line. The stress is 0.1667 and the smallest eigenvalue is −0.7977. Let's call these types of stationary points A,B, and C.

Our algorithms are Newton, Majorization, Relaxed majorization with factor 1.5, Relaxed majorization with factor 1.9, Hybrid with 10 majorization steps, Hybrid with 25 majorization steps. We did one hundred runs of each, starting with different random θ. We stoped iterating when the largest component of the gradient had absolute value less than 1e-6.

The results are interesting. Newton converges very fast, but has serious local minimum (and even saddle point) problems. Saddle points are points of repulsion for the majorization algorithm, and even non-isolated stationary points such as the triangle seems to repulse the majorization iterations. Majorization is pretty fast in this case, because we have linear convergence to the square with rate $1 - 0.1595 =$

Algorithm	Mean It	StDev It	Square	Triangle	Line
Newton	9.4	3.4	36	51	13
Majorization	62.3	10.4	100	0	0
Relax 1.5	152.5	264.7	87	13	0
Relax 1.9	223.2	226.5	80	20	0
Hybrid 10	14.3	2.4	80	18	2
Hybrid 25	27.4	1.6	93	7	0

0.8404. The relaxed iterations are disappointing, but further analysis explains why this happens. If iterations converge to the triangle with centroid, they do so sublinearly. If they converge to the square, then relaxation with factor 1.9 gives rate 0.6968 and relaxation with rate 1.5 gives rate 0.7606. Thus convergence rate does become better, but unfortunately the frequency of the triangle increases. The two hybrid methods work pretty well. On the average they need only about 3 additional Newton iterations, and with enough majorization iterations the undesirable solutions become rare.

We get the same result if we iterate to higher precision (1e-10). Newton uses on the average 15.4 iterations, but it only converges to the square in 30% of the cases. Majorization uses 113.5 iterations, and always finds the square.

In our second example we take dissimilarities equal to the distances between the four points at the corners of the unit square.

Appendix A. Code

To create the basis and the C-matrices we use

```
makeX<-function (n,p) {
r<-(p*n)-(p*(p+1)/2); l<-1
x<-array(0,c(n,p,r))
for (i in 1:p)
        {
        qrq<-qr.Q(qr(outer(1:(n-i+1),0:(n-i),"^")))[,2:(n-i
            +1)]
        for (k in 1:(n-i))
                {
                x[1:(n-i+1),i,l]<-qrq[,k]
                l<-l+1
                }
        }
return(x)
}

makeCfromX<-function(x) {
n<-dim(x)[3]; m<-dim(x)[1]; mm<-m*(m-1)/2
c<-array(0,c(n,n,mm))
for (s in 1:n) for (t in 1:n)
        {
        prd<-x[,,s]%*%t(x[,,t]); k<-1
        for (i in 1:(m-1)) for (j in (i+1):m)
                {
                c[s,t,k]<-prd[i,i]+prd[j,j]-(prd[i,j]+prd[j
                    ,i])
                k<-k+1
                }
        }
return(c/m)
}
```

The program to perform the various algorithms is

```
smacof<-function(delta,p=2,theta="c",eps=1e-6,type="s",
    relax=1.9,verbose=FALSE,switch=0) {
delta<-delta/sqrt(sum(delta^2)); itel<-1; nn<-length(delta)
    ; n<-(1+sqrt(1+8*nn))/2; m<-(p*n)-(p*(p+1)/2)
xxx<-makeX(n,p)
ccc<-makeCfromX(xxx)
if (theta=="c") theta<-rep(1,m)
        else if (theta=="r") theta<-rchisq(m,1)
repeat{
h<-b<-matrix(0,m,m)
for (i in 1:nn)
        {
        cc<-ccc[,,i]
        v<-as.vector(cc%*%theta)
        d<-sum(theta*v)
        r<-delta[i]/sqrt(d)
        b<-b+r*cc
        h<-h+r*(cc-(outer(v,v)/d))
        }
gut<-b%*%theta
if (type=="s") ups<-gut
if (type=="a") ups<-(relax*gut)-(relax-1.0)*theta
if (type=="n") ups<-solve(diag(m)-h,gut)
if (type=="b") if (itel < switch) ups<-gut else ups<-solve(
    diag(m)-h,gut)
ops<-max(abs(theta-ups))
grd<-max(abs(theta-gut))
str<-1+sum(theta^2)-2*sum(theta*gut)
if (verbose)
        cat("Iteration: ",formatC(itel,digits=5,width=3),
        "Change: ",formatC(ops,digits=10,width=15,format="f
            "),
        "Maxgrad: ",formatC(grd,digits=10,width=15,format="
            f"),
        "Stress: ",formatC(str,digits=10,width=15,format="f
            "),"\n")
```

```
    if (grd<eps) break
32  theta<-ups
    itel<-itel+1
34  }
    eb<-eigen(b)$values; eh<-eigen(diag(m)-h)$values
36  cat("Iteration:␣",formatC(itel,digits=5,width=3),
    "Change:␣",formatC(ops,digits=10,width=15,format="f"),
38  "Maxgrad:␣",formatC(grd,digits=10,width=15,format="f"),
    "Stress:␣",formatC(str,digits=10,width=15,format="f"),"\n")
40  cat("Eigenvalues␣B:␣", formatC(eb, digits=10,width=15,
        format="f"),"\n")
    cat("Eigenvalues␣D:␣", formatC(eh, digits=10,width=15,
        format="f"),"\n")
42  x<-matrix(0,n,p)
    for (i in 1:m) x<-x+theta[i]*xxx[,,i]
44  list(itel=itel,stress=str,eval=eh[length(eh)],sol=x)
    }
```

References

J. De Leeuw. Applications of Convex Analysis to Multidimensional Scaling. In J.R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, editors, *Recent developments in statistics*, pages 133–145, Amsterdam, The Netherlands, 1977. North Holland Publishing Company.

J. De Leeuw. Differentiability of Kruskal's Stress at a Local Minimum. *Psychometrika*, 49:111–113, 1984.

J. De Leeuw. Convergence of the Majorization Method for Multidimensional Scaling. *Journal of Classification*, 5:163–180, 1988.

J. De Leeuw. Fitting Distances by Least Squares. Technical Report UCLA Statistics Series 130, Interdivisonal Program in Statistics, UCLA, Los Angeles, California, 1993.

J. De Leeuw and W. J. Heiser. Convergence of Correction Matrix Algorithms for Multidimensional Scaling. In J.C. Lingoes, editor, *Geometric representations of relational data*, pages 735–752. Mathesis Press, Ann Arbor, Michigan, 1977.

J. De Leeuw and W. J. Heiser. Multidimensional Scaling with Restrictions on the Configuration. In P.R. Krishnaiah, editor, *Multivariate Analysis, volume V*,

pages 501–522, Amsterdam, The Netherlands, 1980. North Holland Publishing Company.

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

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

J.B. Kruskal. Nonmetric Multidimensional Scaling: a Numerical Method. *Psychometrika*, 29:115–129, 1964b.

J. De Leeuw and P.J.F. Groenen. Inverse scaling. Technical Report UCLA Statistics Series, Interdivisonal Program in Statistics, UCLA, Los Angeles, California, 1993.

J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, N.Y., 1970.

A. M. Ostrowski. *Solution of Equations and Systems of Equations*. Academic Press, New York, N.Y., 1966.

F. Robert. Calcul du Rapport Maximal de Deux Normes sur $\mathbb{R}^n$. *Revue Francaise d'Automatique, d'Informatique Et De Recherche Operationelle*, 1:97–118, 1967.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970.

Pham Dinh Tao and El Bernoussie Souad. Algorithms for Solving a Class of Nonconvex Optimization Problems. Methods of Subgradients. In J.-B. Hiriart-Urruty, editor, *Fermat days 1985: Mathematics for Optimization*, pages 249–272, Amsterdam, The Netherlands, 1986. North Holland Publishing Company.

Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu