OPTIMIZING FUNCTIONS OF SQUARED DISTANCES

JAN DE LEEUW, PATRICK J.F. GROENEN, AND RAOUL PIETERSZ

Abstract. A general algorithm is developed for minimizing any function ϕ of squared Euclidean distances between n points that has a bounded second Hessian. The algorithm solves a sequence of eigenvalue problems. We apply the algorithms to squared distance scaling, distance completion, unfolding, and fitting distance-based choice models.

1. Introduction

In multidimensional scaling (MDS) and related techniques we typically minimize loss functions of the form

$$\sigma(X) = \phi(d_{12}(X), \cdots, d_{n-1,n}(X))$$

over all $n \times p$ *configurations* X. Here $d_{ij}(X)$ is the *squared* Euclidean distance between the points (rows) i and j in the configuration, and ϕ is any real valued function. We use functions of *squared* distances for mathematical convenience, but of course any function of the distances can also be written as a function of the squared distances, and vice versa.

We introduce some convenient notation to work with squared distances. First,

$$d_{ij}(X) = (e_i - e_j)'XX'(e_i - e_j) = \mathbf{tr}\, CA_{ij},$$

where $C = XX'$ and $A_{ij} = (e_i - e_j)(e_i - e_j)'$. Thus A_{ij} has $+1$ for elements (i, i) and (j, j) and -1 for elements (i, j) and (j, i). All other elements are zero. Also define $d_{ij}(C) = c_{ii} + c_{jj} - 2c_{ij}$.

Date: February 7, 2006.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Multidimensional Scaling, Majorization Algorithms.

It follows that instead of minimizing $\sigma(X)$ over X we can also minimize

$$\sigma(C) = \phi(d_{12}(C), \cdots, d_{n-1,n}(C))$$

over all $n \times n$ positive semi-definite matrices C of rank less than or equal to p.

2. Derivatives

By the chain rule

$$\frac{\partial \phi}{\partial C} = \sum\sum_{1 \leq i < j \leq n} \frac{\partial \phi}{\partial d_{ij}} A_{ij}. \tag{1a}$$

and

$$\frac{\partial^2 \phi}{\partial c_{\alpha\beta} \partial c_{\gamma\delta}} = \sum\sum_{1 \leq i < j \leq n} \sum\sum_{1 \leq k < \ell \leq n} \frac{\partial^2 \phi}{\partial d_{ij} \partial d_{k\ell}} [A_{ij}]_{\alpha\beta} [A_{k\ell}]_{\gamma\delta}. \tag{1b}$$

Here the notation $[A_{ij}]_{\alpha\beta}$ is used for element (α, β) of matrix A_{ij}.

We write $M(C)$ for the $n^2 \times n^2$ matrix of second partials defined by (1b), and $H(C)$ for the $\binom{n}{x} \times \binom{n}{2}$ matrix with elements

$$h_{\{ij\}\{k\ell\}}(C) = \frac{\partial^2 \phi}{\partial d_{ij} \partial d_{k\ell}}.$$

We also use the Loewner ordering of symmetric matrices. If A and B are two symmetric matrices, then $A \lesssim B$ means that $B - A$ is positive semi-definite. Moreover $\mathbf{vec}(A)$ strings out a matrix to a vector.

Theorem 2.1. *If there is a diagonal B with $H(C) \lesssim B$, and $\lambda(V)$ is the largest eigenvalue of*

$$V = \sum\sum_{1 \leq i < j \leq n} b_{ij} \mathbf{vec}(A_{ij}) \mathbf{vec}(A_{ij})'.$$

then $M(C) \lesssim \lambda(V) I$.

Proof. If U is an arbitrary symmetric matrix, then

$$\sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \sum_{\gamma=1}^{n} \sum_{\delta=1}^{n} u_{\alpha\beta} u_{\gamma\delta} \frac{\partial^2 \phi}{\partial c_{\alpha\beta} \partial c_{\gamma\delta}} = \sum\sum_{1 \leq i < j \leq n} \sum\sum_{1 \leq k < \ell \leq n} \frac{\partial^2 \phi}{\partial d_{ij} \partial d_{k\ell}} \mathbf{tr}\ U A_{ij} \mathbf{tr}\ U A_{k\ell}.$$

Thus

$$\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n}\sum_{\gamma=1}^{n}\sum_{\delta=1}^{n} u_{\alpha\beta}u_{\gamma\delta}\frac{\partial^2\phi}{\partial c_{\alpha\beta}\partial c_{\gamma\delta}} \leq \sum\sum_{1\leq i<j\leq n} b_{ij}\{\mathbf{tr}\ UA_{ij}\}^2 \leq \lambda(V)\mathbf{tr}\ U^2.$$

□

Corollary 2.2. *If $H(C) \lesssim \beta I$ then $M(C) \lesssim 2n\beta I$.*

Proof. We merely have to show that $\lambda(V) = 2n$ if

$$V = \sum\sum_{1\leq i<j\leq n} \mathbf{vec}(A_{ij})\mathbf{vec}(A_{ij})'.$$

Instead of looking at the $n^2 \times n^2$ matrix V, we look at the $\binom{n}{2} \times \binom{n}{2}$ matrix S with elements $s_{ij,k\ell} = \mathbf{tr}\ A_{ij}A_{k\ell}$. S has the same eigenvalues as V. All elements of S are non-negative, the diagonal elements are $+4$, and each row has $2n-4$ elements equal to $+1$. Thus the largest eigenvalue, corresponding to an eigenvector with all elements equal to $+1$, is $2n$. □

Corollary 2.3. *If $H(C) \lesssim B$, where B is symmetric, then $M(C) \lesssim \tau(V)I$, with*

$$\tau(V) = 4\ \mathbf{tr}\ B + \sum b_{\{i,j\},\{k,\ell\}},$$

where summation is over the $n(n-1)(n-2)$ pairs (i, j) and (k, ℓ) that have exactly one index in common.

Proof. The largest eigenvalue $\lambda(V)$ is always less than or equal to the trace

$$\begin{aligned}\mathbf{tr}\ V &= \sum\sum_{1\leq i<j\leq n}\sum\sum_{1\leq k<\ell\leq n} b_{\{i,j\},\{k,\ell\}}\mathbf{tr}\ A_{ij}A_{k\ell} = \\ &= \sum\sum_{1\leq i<j\leq n}\sum\sum_{1\leq k<\ell\leq n} b_{\{i,j\},\{k,\ell\}}(\delta^{ik} - \delta^{i\ell} - \delta^{jk} + \delta^{j\ell})^2.\end{aligned}$$

Collecting terms, as in the proof of Corollary 2.2, gives the required result.

□

3. Majorization

To state our main result, which gives the basic property of our algorithm, we define

$$B(C) = \sum\sum_{1\leq i<j\leq n} \left.\frac{\partial\phi}{\partial d_{ij}}\right|_{C} A_{ij},$$

$$\overline{C} = C - \frac{1}{\mu}B(C),$$

Theorem 3.1. *If $\mu \geq \lambda(V)$ and*

$$C^{+} = \mathbf{argmin}\{\mathbf{tr}\ (C-\overline{C})^{2} \mid C \gtrsim 0, \mathbf{rank}(C) \leq p\}.$$

then $\sigma(C^{+}) \leq \sigma(C)$. Moreover if $C^{+} \neq C$ then $\sigma(C^{+}) < \sigma(C)$.

Proof. The results so far imply that for all pairs C and $\tilde{C}$ we have $\sigma(\tilde{C}) \leq \eta(\tilde{C}, C)$, where the majorization function η is defined by

$$\eta(\tilde{C}, C) = \sigma(\tilde{C}) + \mathbf{tr}\ B(\tilde{C})(C-\tilde{C}) + \frac{1}{2}\mu\ \mathbf{tr}\ (C-\tilde{C})^{2}.$$

By completing the square we can rewrite this as

$$\eta(\tilde{C}, C) = \sigma(\tilde{C}) + \frac{1}{2}\mu\ \mathbf{tr}\ (C-\overline{C})^{2} - \frac{1}{2}\mu\ \mathbf{tr}\ B(\tilde{C})^{2}.$$

Thus we derive the sandwich inequality from majorization theory [De Leeuw, 1994],

$$\sigma(C^{+}) \leq \eta(C^{+}, C) \leq \eta(C, C) = \sigma(C).$$

□

The minimization in Theorem 3.1 is easily accomplished by

4. Examples

4.1. **Squared Distance Scaling.** Let us look at what is probably the most straghtforward example of our theory. Suppose the function we are minimizing is *sstress* [Takane et al., 1977], i.e.

$$\sigma(C) = \frac{1}{2}\sum\sum_{1\leq i<j\leq n} w_{ij}(\delta_{ij} - d_{ij}(C))^{2} \tag{2}$$

Clearly we can take $B = W$ in this case, and apply Theorem 2.1, or use $\beta = \max w_{ij}$ and apply Corollary 2.2. By the general reasoning in the previous sections, these provide convergent algorithms.

Many different algorithms have been proposed to minimize the loss function (2). Foremost of these is perhaps the `ALSCAL` method [Takane et al., 1977], which is of the cyclic coordinate descent type. One `ALSCAL` iteration consists of a cycle over all np coordinates of X, minimizing loss over one coordinate at a time, while keeping the other coordinates fixed at their current values. Since the loss function is a multivariate quartic in X, the coordinate subproblems can be solved by finding the roots of a univariate cubic (and choosing the one corresponding to the minimum).

Even before `ALSCAL`, De Leeuw [1975] proposed an augmentation algorithm to minimize (2), in the case in which there are no weights. At the time, the algorithm was called `ELEGANT`. It corresponds to our squared distance algorithm using Corollary 2.3, and it consequently uses the bound

$$\tau(V) = 4 \sum_{1\leq i<j\leq n}\sum w_{ij}.$$

The 1975 paper was never published, but the algorithm has been discussed by Takane [1977] and Browne [1987]. They did not include the original derivation nor a convergence proof. This paper provides one version of that proof, quite different from the original one, and much more generally applicable. Our majorization proof also applies to the case of unequal weights, and to a much more general class of functions.

Euclidean Completion.

Suppose, more generally, that

$$\sigma(C) = \frac{1}{2}\sum_{1\leq i<j\leq n}\sum w_{ij}(\psi(\delta_{ij}) - \psi(d_{ij}(C)))^2. \tag{3}$$

Our approach is of somewhat limited value for this class of loss functions, because the most popular choices of ψ lead to unbounded second derivatives.

The Hessian H is diagonal in the case of loss function (3), with elements

$$h_{\{ij\},\{ij\}} = -w_{ij}\psi(\delta_{ij})\psi''(d_{ij}) + w_{ij}[\{\psi'(d_{ij})\}^2 + \psi(d_{ij})\psi''(d_{ij})].$$

If ψ is the square root, for instance, we see that

$$h_{\{ij\},\{ij\}} = \frac{1}{4}w_{ij}\psi(\delta_{ij})d_{ij}^{-3/2}.$$

This cannot be bounded, because it will become arbitrary large if i and j get arbitrary close to each other. Similar problems occur if ψ is the logarithm. On the other hand, if $\psi(d) = \exp(-d)$, then we can derive the useful bound

$$h_{\{ij\},\{ij\}} \leq 2w_{ij}\exp(-2d_{ij}) \leq 2w_{ij}.$$

4.2. **Unfolding.** There are some special cases of (2), which deserve some attention. In (unweighted, metric, squared distance) unfolding, for example, in which we have unit weights for off-diagonal distances with $1 \leq i \leq n_1$ and $n_1 + 1 \leq j \leq n_1 + n_2$ and zero weights elsewhere. Computations like the ones in Corollary 2.2 show that $\lambda(V) = n_1 + n_2 + 2$.

4.3. **Choice model.**

$$\sigma(C) = -N\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\log\pi_{ij}(C) + (1 - y_{ij})\log(1 - \pi_{ij})$$

$$\pi_{ij}(C) = \frac{1}{1 + \alpha_i\beta_j\exp(-d_{ij}(C))}.$$

$$\sigma(C) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\{y_{ij}d_{ij}(C) - \log(1 + \exp(d_{ij}(C))\}$$

4.4. **Gaussian Ordination.**

4.5. **Shepard-Luce Model.**

$$\sigma(C) = -N\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}\log\pi_{ij}(C)$$

$$\pi_{ij}(C) = \frac{\alpha_i\beta_j\exp(-d_{ij}(C))}{\sum_{k=1}^{n}\sum_{\ell=1}^{n}\alpha_k\beta_\ell\exp(-d_{k\ell}(C))}$$

References

M.W. Browne. The Young-Householder Algorithm and the Least Squares Multdimensional Scaling of Squared Distances. *Journal of Classification*, 4:175–190, 1987.

J. De Leeuw. Block Relaxation Methods in Statistics. In H.H. Bock, W. Lenski, and M.M. Richter, editors, *Information Systems and Data Analysis*, Berlin, 1994. Springer Verlag.

J. De Leeuw. An Alternating Least Squares Approach to Squared Distance Scaling. 1975.

Y. Takane. On the Relations among Four Methods of Multidimensional Scaling. *Behaviormetrika*, 4:29–42, 1977.

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

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

E-mail address: deleeuw@stat.ucla.edu

URL: http://gifi.stat.ucla.edu

Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

E-mail address: groenen@few.eur.nl

Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

E-mail address: pietersz@few.eur.nl