40

FINDING A POSITIVE SEMIDEFINITE MATRIX OF PRESCRIBED RANK r IN A NONLINEAR DIFFERENTIABLE MANIFOLD

by

Jan de Leeuw*
Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

In this paper we study some technical aspects of the problem of choosing the parameters of a matrix function in such a way that the matrix is approximately positive semidefinite (psd) of rank r. Applications to multidimensional scaling are discussed, but not analyzed in detail. They will be analyzed in detail in some companion papers, which use the technical results listed in this paper.

* On leave of absence from the University of Leiden, Leiden, The Netherlands.

FINDING A POSITIVE SEMIDEFINITE MATRIX OF PRESCRIBED RANK r IN A NONLINEAR DIFFERENTIABLE MANIFOLD

by

Jan de Leeuw*
Bell Laboratories
Murray Hill, New Jersey 07974

1. Problem

Consider a square symmetric matrix $A(\theta)$ whose elements are functions of real parameters $\theta_1,\ldots,\theta_u$ assuming values in an open set Θ. We assume that in Θ the matrix functions $G_s(\theta)$ and $H_{st}(\theta)$ $(s,t = 1,\ldots,u)$ defined by

$$g_{ij}^{s}(\hat{\theta}) \triangleq \left.\frac{\partial a_{ij}}{\partial\theta_s}\right|_{\theta=\hat{\theta}}$$

$$h_{ij}^{st}(\hat{\theta}) \triangleq \left.\frac{\partial^2 a_{ij}}{\partial\theta_s\partial\theta_t}\right|_{\theta=\hat{\theta}}$$

exist and are continuous functions of their arguments. The eigenvalues of $A(\theta)$ are written as $\lambda_1(\theta) \geq \lambda_2(\theta) \geq \cdots \geq \lambda_n(\theta)$ and the corresponding orthonormal eigenvectors are $x_1(\theta),\ldots,x_n(\theta)$. They are usually collected in the diagonal matrix function $\Lambda(\theta)$ or the orthonormal matrix function $X(\theta)$. Several important data analytic problems in the multidimensional

---

* On leave of absence from the University of Leiden, Leiden, The Netherlands.

scaling and factor analysis area can be formulated in terms of finding the value of θ which makes $A(\theta)$ as close as possible to the set of matrices of rank r, or to the set of positive semidefinite (psd) matrices of rank r. In this paper we try to solve this problem by maximizing a particular function of the eigenvalues of $A(\theta)$. The precise reasons for choosing this particular parameterization of scaling problems, and the precise reason for choosing this particular function of the eigenvalues are outlined in two companion papers. The function we use to measure the degree of "positive semi-definiteness" and "rank r-ness" simultaneously is

$$\frac{\sum_{p=1}^{r} \lambda_p^2(\theta)}{\sum_{p=1}^{n} \lambda_p^2(\theta)} \triangleq \frac{2S_r(\theta)}{2T(\theta)} \triangleq u_r(\theta)$$

Clearly $0 \leq u_r(\theta) \leq 1$ with $u_r(\theta) = 1$ iff $\lambda_p(\theta) = 0$ for all $p = r+1,\ldots,n$ and $\lambda_p(\theta) > 0$ for at least one $1 \leq p \leq r$ iff $A(\theta)$ is psd of rank $1 \leq \rho \leq r$. Moreover $u_r(\theta) = 0$ iff $\lambda_p(\theta) = 0$ for all $p = 1,\ldots,r$ and $\lambda_p(\theta) < 0$ for at least one $r + 1 \leq \rho \leq n$ iff $A(\theta)$ is nsd of rank $1 \leq \rho \leq n - r$. An alternative, and perhaps slightly more satisfactory, way of measuring the same property of $A(\theta)$ can be obtained by defining

$$\lambda_p^+(\theta) \triangleq \tfrac{1}{2}\{|\lambda_p(\theta)| + \lambda_p(\theta)\}$$

$$\lambda_p^-(\theta) \triangleq \tfrac{1}{2}\{|\lambda_p(\theta)| - \lambda_p(\theta)\}$$

The usual identities and inequalities associated with this decomposition are

$$\lambda_p^+(\theta) + \lambda_p^-(\theta) = |\lambda_p(\theta)| .$$

$$\lambda_p^+(\theta) - \lambda_p^-(\theta) = \lambda_p(\theta).$$

$$\lambda_p^+(\theta)\lambda_p^-(\theta) = 0.$$

$$\lambda_p^+(\theta) \geq 0 \quad \text{and} \quad \lambda_p^+(\theta) = 0 \quad \text{iff} \quad \lambda_p(\theta) \leq 0.$$

$$\lambda_p^-(\theta) \geq 0 \quad \text{and} \quad \lambda_p^-(\theta) = 0 \quad \text{iff} \quad \lambda_p(\theta) \geq 0.$$

$$\left[\lambda_p^+(\theta) \pm \lambda_p^-(\theta)\right]^2 = \left[\lambda_p^+(\theta)\right]^2 + \left[\lambda_p^-(\theta)\right]^2 .$$

The modified gain function is simply obtained by defining

$$\tilde{S}_r(\theta) \triangleq \tfrac{1}{2} \sum_{p=1}^{r} \left[\lambda_p^+(\theta)\right]^2$$

$$\tilde{\alpha}_r(\theta) \triangleq \frac{2\tilde{S}_r(\theta)}{2T(\theta)}$$

Now $\tilde{\alpha}_r(\theta) = 1$ iff $\alpha_r(\theta) = 1$, but $\tilde{\alpha}_r(\theta) = 0$ iff $\lambda_p(\theta) \leq 0$ for all $p = 1,\ldots,r$ and $\lambda_p(\theta) < 0$ for at least one $r + 1 \leq p \leq n$ iff $A(\theta)$ is nsd of rank $1 \leq \rho \leq n$.

3. Derivatives

We intend to use gradient and modified gradient methods to maximize our gain functions in the particular cases we are interested in. For this purpose it is useful to have general expressions for the derivatives of these functions. In the appendix we have collected some more or less well known results from the perturbation theory of eigenvectors and eigenvalues of symmetric matrices. Since no confusion is possible we surpress the dependence of all expressions on the value of $\theta \in \Theta$. For the denominator we find, trivially,

$$\frac{\partial T}{\partial \theta_s} = \mathrm{tr}[AG_s],$$

$$\frac{\partial^2 T}{\partial \theta_s \partial \theta_t} = \mathrm{tr}[AH_{st}] + \mathrm{tr}[G_s G_t].$$

From the results in the appendix

$$\frac{\partial S_r}{\partial \theta_s} = \mathrm{tr}[A_r G_s],$$

with the matrix $A_r = \{a_{ij}^r\}$ defined by

$$a_{ij}^r = \sum_{p=1}^{r} \lambda_p x_{ip} x_{jp}.$$

For the second partials of S_r the results are more complicated.

If we define

$$v_{pq} = \begin{cases} \dfrac{\lambda_p}{\lambda_q - \lambda_p} & p \neq q \\ 1 & p = q \end{cases}$$

$$w_{ijk\ell} = \sum_{p=1}^{r} \sum_{q=1}^{n} v_{pq} x_{ip} x_{jq} x_{kp} x_{\ell q}$$

then

$$\frac{\partial^2 S_r}{\partial \theta_s \partial \theta_t} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{\ell=1}^{n} g_{ij}^{s} g_{k\ell}^{t} w_{ijk\ell} + \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}^{st} a_{ij}^{r}$$

For $\tilde{S}_r$ we find the slightly modified expressions

$$\frac{\partial \tilde{S}_r}{\partial \theta_s} = tr[\tilde{A}_r G_s]$$

with

$$\tilde{a}_{ij}^{r} = \sum_{p=1}^{r} \lambda_p^{+} x_{ip} x_{jp}$$

and

$$\frac{\partial^2 \tilde{S}_r}{\partial \theta_s \partial \theta_t} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{\ell=1}^{n} g_{ij}^{s} g_{ij}^{t} \tilde{w}_{ijk\ell} + \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij}^{st} \tilde{a}_{ij}^{r}$$

with

$$w_{ijk\ell} = \sum_{p=1}^{r} \sum_{q=1}^{n} \tilde{u}_{pq} x_{ip} x_{jq} x_{kp} x_{\ell q}$$

and

$$\tilde{v}_{pq} = \begin{cases} \dfrac{\lambda_p^+}{\lambda_q - \lambda_p} & p \neq q \\ \tfrac{1}{2}\left(\operatorname{sign}(\lambda_p) + 1\right) & p = q \end{cases}$$

Although it is clear that the expressions in this section are powerful enough to use any second order method for solving our optimization problem, we will discuss the actual computational procedures only in connection with the various examples in later papers.

3. Applications

In this section we indicate some possible applications of the basic formula's of this paper to problems in scaling. The advantages over alternative formulations are only briefly hinted at, they will be described more extensively in subsequent papers dealing with each of these special cases. The purpose of this section is merely to indicate how standard problems can be reformulated in this new parametrization, and to give the reader an idea what this reformulation looks like.

.1 Additive Constants

In metric Euclidean scaling a common problem is to find a constant γ (sometimes required to be not smaller than a given γ_0) such that $\delta_{ij} + \gamma$ is approximately a set of distances between n-points in r-space. We can write this (using fuzzy equality symbols) as

$$d_{ij} \approx \delta_{ij} + \gamma$$

or

$$d_{ij}^2 \approx \delta_{ij}^2 + 2\gamma\delta_{ij} + \gamma^2$$

By using doubly centering we find that this is equivalent to

$$C_1 + 2\gamma C_2 + \gamma^2 C_3 \underset{\sim}{\in} A_r$$

with A_r the psd matrices of rank r, and C_1, C_2, C_3 the doubly centered versions of $\{\delta_{ij}^2\}$, $\{\delta_{ij}\}$, and the identity. Existing methods for finding additive constants are given by Messick and Abelson (1956), and Cooper (1972). The method proposed by Messick and Abelson starts with some $\gamma^{(0)}$, and then proceeds as follows

<u>step 1</u>: compute $A(\gamma^{(k)}) = C_1 + 2\gamma^{(k)}C_2 + (\gamma^{(k)})^2 C_3$

<u>step 2</u>: compute $\lambda_1(\gamma^{(k)}), \ldots, \lambda_r(\gamma^{(k)})$

<u>step 3</u>: find $\gamma^{(k+1)}$ as the appropriate solution of

$$\mathrm{tr}(C_1) + 2\gamma\,\mathrm{tr}(C_2) + \gamma^2\mathrm{tr}(C_3) = \sum_{p=1}^{r} \lambda_p(\gamma^{(k)})$$

This method is not exactly rigorous, but it seems to work quite well in practice. It can, moreover, be used to derive a starting point for more sophisticated procedures. The method proposed by Cooper (1972) is more satisfactory. It minimizes

$$S = \sum_{i=1}^{n} \sum_{j=1}^{n} (\delta_{ij} + \gamma - d_{ij})^2$$

over coordinate values and additive constant. Although this is probably the most straightforward approach there really is no need to use this particular parametrization of the problem. The one in terms of the eigenvalues of the matrix

$$A(\gamma) = C_1 + 2\gamma C_2 + \gamma^2 C_3$$

is much more compact and computationally much more efficient. For these one-parameter families of matrices very efficient algorithms are available. The papers of Lancaster (1963, 1964 a,b) give a nice summary of available results. They also generalize some of the results in our appendix for this particular case. It seems reasonable to conjecture that an algorithm that starts with a couple of Messick-Abelson iterations and then ends with a number of Newton-Raphson iterations will be quite efficient. This makes it also quite trivial to deal with the possible constraint $\gamma \geq \gamma_0$

3.2 Missing Data in Multidimensional Scaling

The missing data problem in multidimensional scaling can be formulated as

$$C_0 + \theta_1 T_1 + \ldots + \theta T_p \underset{\sim}{\in} A_r$$

Here C_0 is the doubly centered version of the matrix of squared distances (with missing data first replaced by zeroes) and $T_1, \ldots, T_p$ correspond with the missing data (each empty cell i,j has a matrix, which is the doubly centered version of the matrix E_{ij} with only element i,j equal to unity and all other elements equal to zero). For this linear problem the matrices H_{st} are, of course, all equal to zero and the matrices G_s are equal to T_s, and independent of the value of θ. The numerator $T(\theta)$ is a quadratic function of θ. A suggested computational procedure is to use the linear approximation to $S_r(\theta)$ and the quadratic "approximation" to $T(\theta)$. We can then reduce the fractional objective function to a linear one with quadratic constraint, and this transforms by using other known tricks to a quadratic program. The advantage is, of course, that linear equality and inequality constraints on the θ_s can be incorporated very easily.

3.3 Nonmetric Scaling

One of the more interesting applications is an alternative parametrization of nonmetric multidimensional scaling. This can be written as

$$\theta_1 T_1 + \ldots + \theta_u T_u \underset{\sim}{\in} A_r$$

Now the T_s are the doubly centered versions of the edges of the polyhedral cone in matrix space containing the admissible solutions. Observe that this problem is both linear and homogeneous and consequently we use the numerator $T(\theta)$ here in a different sense as in our previous, nonhomogeneous, examples. In this case we cannot do without it, in previous cases we could minimize $S_r(\theta)$ without involving $T(\theta)$. Observe also that we have to require $\theta \geq 0$ here. Other homogeneous models which do not require $\theta \geq 0$ are also possible (DeLeeuw 1973,1974).

3.4 Additional

Clearly additional examples can be generated by the two observations that our development also applies to the scalar product model, and, with the slight modifications mentioned in the appendix, to rectangular and asymmetric matrices. More complicated function $a_{ij}(\theta)$ can be used to define metric and nonmetric versions of spherical, hyperbolic, and elliptic multidimensional scaling. Slightly more complicated functions of the $\lambda_i(\theta)$ can be used to define nonlinear versions of multiple factor analysis. All this will be considered in later papers.

Acknowledgment

Comments by Joe Kruskal were very helpful in preparing this revision of the original paper.

APPENDIX

It is well known that each real symmetric matrix A can be written in the form $A = X\Lambda X'$ with $X'X = XX' = I$ and Λ a diagonal matrix. We are interested in the changes in this decomposition if we make a small symmetric perturbation of the form $\tilde{A} = A + \Delta$. If all elements of Λ are unequal it can be proved rigorously that the values in $\tilde{\Lambda}$ are analytic functions of the δ_{ij}. Define $T = X'\Delta X$. The first order approximation is

$$\tilde{\Lambda} = \Lambda + \text{diag}(T).$$

Define $H = \{h_{pq}\}$ by

$$h_{pq} = \begin{cases} -\dfrac{t_{pq}}{\lambda_p - \lambda_q} & p \neq q \\ 0 & p = q \end{cases}$$

The matrix G is defined simply as $H + I$. The second order approximation is

$$\tilde{\Lambda} = \Lambda + \text{diag}(T) + \text{diag}(TH) = \Lambda + \text{diag}(TG)$$

In element wise notation

$$\tilde{\lambda}_p = \lambda_p + \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} x_{ip} x_{jp}$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \delta_{ij} \delta_{k\ell} \sum_{\substack{q \neq p \\ q=1}}^{n} \frac{1}{\lambda_q - \lambda_p} x_{ip} x_{jq} x_{kp} x_{\ell q}$$

If the elements of the matrix A are functions of parameters $\theta_1, \ldots, \theta_r$, then we find the formula's

$$\frac{\partial \lambda_p}{\partial \theta_s} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial \lambda_p}{\partial a_{ij}} \frac{\partial a_{ij}}{\partial \theta_s} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial a_{ij}}{\partial \theta_s} x_{ip} x_{jp}$$

$$\frac{\partial^2 \lambda_p}{\partial \theta_s \partial \theta_t} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \frac{\partial^2 \lambda_p}{\partial a_{ij} \partial a_{k\ell}} \frac{\partial a_{ij}}{\partial \theta_s} \frac{\partial a_{k\ell}}{\partial \theta_t} + \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial \lambda_p}{\partial a_{ij}} \frac{\partial^2 a_{ij}}{\partial \theta_s \partial \theta_t}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \frac{\partial a_{ij}}{\partial \theta_s} \frac{\partial a_{k\ell}}{\partial \theta_t} \sum_{\substack{q \neq p \\ q=1}}^{n} \frac{1}{\lambda_q - \lambda_p} x_{ip} x_{jq} x_{kp} x_{\ell q}$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 a_{ij}}{\partial \theta_s \partial \theta_t} x_{ip} x_{jp}$$

Although we do not need the results in this paper we give, for reasons of completeness, a similar expression for the first derivatives of the eigenvectors. The first order expression is simply

$$\tilde{X} = X + XH = XG.$$

In elementwise notation

$$\tilde{x}_{kp} = x_{kp} + \sum_{i=1}^{n}\sum_{j=1}^{n} e_{ij} \sum_{\substack{q\neq p\\ q=1}}^{n} \frac{1}{\lambda_q-\lambda_p}\, x_{kq}x_{ip}x_{jq}$$

It follows that

$$\frac{\partial x_{kp}}{\partial \sigma_s} = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial e_{ij}}{\partial \sigma_s} \sum_{\substack{q\neq p\\ q=1}}^{n} \frac{1}{\lambda_q-\lambda_p}\, x_{kq}x_{ip}x_{jq}$$

Differentiating this expression once again would give us second derivatives but they do not look very interesting and are not needed in most applications. In a separate paper these expressions will be generalized to the nonsymmetric and to the rectangular case. We shall also comment there on the problems connected with equal eigenvalues.