DETERMINISTIC METHODS FOR OPTIMIZING INTEGRALS OCCURRING IN ITEM RESPONSE THEORY

JAN DE LEEUW

ABSTRACT.

Gott würfelt nicht
A. Einstein

1. INTRODUCTION

Principal Component Analysis (PCA) and Factor Analysis (FA) methods for matrices of zeroes and ones were discussed in De Leeuw [2003a,b]. Methods based on correspondence analysis, on least squares multidimensional scaling, and on Bernoulli likelihood were distinguished. The likelihood methods minimize the deviance, which is a function of the form

$$\mathcal{D}(A, B) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\log F(q_{ij}\langle a_i, b_j\rangle), \tag{1}$$

where $\langle a, b\rangle$ is the inner product, F is a known one-dimensional cdf such as the standard normal or the standard logistic, and q_{ij} is the binary data, linearly transformed to ± 1. Geometrically, minimizing this loss function means we represent the row as points, and the columns as hyperplanes optimally separating points corresponding to the zeroes in the column from those corresponding to the ones.

If the number of columns is small compared to the number of rows then the geometry tells us the location of the row points will be not very well determined, and in particular row points can be moved to infinity along directions of recession. The minimum will not be attained. In statistical terminology this happens because there are too many incidental parameters (the a_i in Equation (1)).

Date: October 7, 2003.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis.

On common solution to this problem is to assume the a_i are realizations of a random vector, with distribution G, and to minimize the marginal deviance

$$\mathcal{D}(B,G)=-\sum_{i=1}^n\log\int\exp\{\sum_{j=1}^m\log F(q_{ij}\langle a_i,b_j\rangle)\}dG(a). \tag{2}$$

We can distinguish two special cases here. In the first case G is completely known, for instance we set it to the multivariate standard normal, and in the second case G is completely unknown, except for the fact that it must be a cdf. In this second case we minimize the marginal deviance over both B and G. Obviously there could be intermediate cases, in which G is partially known, but we do not discuss these here.

The problem we study in this article has both (1) and (2) as special cases. We minimize marginal deviances of the form

$$\mathcal{D}(B,G)=-\sum_{i=1}^n\log\int\exp\{\sum_{j=1}^m\log F(q_{ij}\langle a,b_j\rangle)\}dG_i(a). \tag{3}$$

Thus each row has its own cdf G_i, which can again be either completely known or completely unknown.

This integral is studied in component or factor analysis of binary matrices, for instance in the item response theory of educational statistics [McDonald, 1997; Reckase, 1997] and in the roll call analysis of political science [Clinton et al., 2003]. In recent publications, the minimization problem is often solved by using some form of Monte Carlo method, often by Markov Chain Monte Carlo [Meng and Schilling, 1996; Beguin and Glas, 2001; Jackman, 2001]. In this paper we study some alternative minimization methods based on *majorization* (also known as *variational bounding*).

These methods are not necessarily superior to MCMC. They do guarantee convergence from any starting point and a monotone decreasing sequence of loss function values $\mathcal{D}$. Thus one would at least expect the convergence is more regular and somewhat easier to monitor than MCMC convergence. Moreover each step of the algorithm computes the partial singular value decomposition (SVD) of a matrix, which can be done efficiently in interpreted languages with fast linear algebra routines, such as R or Matlab. And finally initial convergence of our majorization algorithms is extremely fast, which means that, at the very least, they provide excellent initializers for other algorithms.

2. SINGLE STEP G_i

Suppose $G_i(a)$ steps from zero to one at an unknown a_i. Then loss function (3) becomes loss function (1), and we are in the situation discussed in detail in De Leeuw [2003b]. We repeat the key result here.

Suppose we have a variational bound of the form

$$-\log F(x) \leq -\log F(y) + h(y)(x-y) + \frac{1}{2}w(y)(x-y)^2. \tag{4}$$

for all x and y, with equality if and only if $x = y$. Observe that by completing the square we can also write (4) as

$$-\log F(x) \leq -\log F(y) + \frac{1}{2}w(y)(x-z(y))^2 - \frac{1}{2}\frac{h^2(y)}{w(y)}, \tag{5a}$$

where

$$z(y) = y - \frac{h(y)}{w(y)}. \tag{5b}$$

Specific instances of such bounds for a logistic and normal cdf F are given in De Leeuw [2003b].

The bounds can be used to implement a simple majorization algorithm. Start with some $A^{(0)}$ and $B^{(0)}$. Suppose $A^{(\nu)}$ and $B^{(\nu)}$ are the current best solution. We update them to find a better solution in two steps, similar to the E-step and the M-step in the EM-algorithm.

Algorithm 2.1 (Majorization)**.**

Step $\nu(1)$**:** *Compute the matrices* $W^{(\nu)}$, $H^{(\nu)}$ *and* $Z^{(\nu)}$ *with elements*

$$\begin{aligned} w_{ij}^{(\nu)} &= w(q_{ij}\langle a_i^{(\nu)}, b_j^{(\nu)}\rangle), \\ h_{ij}^{(\nu)} &= h(q_{ij}\langle a_i^{(\nu)}, b_j^{(\nu)}\rangle), \\ z_{ij}^{(\nu)} &= \langle a_i^{(\nu)}, b_j^{(\nu)}\rangle - q_{ij}\frac{h_{ij}^{(\nu)}}{w_{ij}^{(\nu)}} \end{aligned}$$

Step $\nu(2)$**:** *Solve the least squares matrix approximation problem*

$$\min_{A,B} \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^{(\nu)}(\langle a_i, b_j\rangle - z_{ij}^{(\nu)})^2$$

by using the (weighted) SVD. This gives $A^{(\nu+1)}$ *and* $B^{(\nu+1)}$.

Theorem 2.1. (1) *The Majorization Algorithm 2.1 produces a decreasing sequence* $\mathcal{D}(A^{(\nu)}, B^{(\nu)})$ *of loss function values.*

(2) *A necessary condition for a minimum of the loss function* $\mathcal{D}$ *at* (A, B) *is that the algorithm is stationary at* (A, B).

(3) *All accumulation points of the sequence* $(A^{(\nu)}, B^{(\nu)})$ *of iterates are stationary points of the algorithm.*

Proof. Substitute $q_{ij}\langle a_i, b_j\rangle$ for x and $q_{ij}\langle a_i^{(\nu)}, b_j^{(\nu)}\rangle$ for y in Equation (5). Then

$$-\log F(q_{ij}\langle a_i, b_j\rangle) \leq -\log F(q_{ij}\langle a_i^{(\nu)}, b_j^{(\nu)}\rangle) - \frac{1}{2}\frac{(h_{ij}^{(\nu)})^2}{w_{ij}^{(\nu)}} + \\ + \frac{1}{2}w_{ij}^{(\nu)}(\langle a_i, b_j\rangle - z_{ij}^{(\nu)})^2$$

Now sum over i and j to obtain

$$\mathcal{D}(A, B) \leq \mathcal{D}(A^{(\nu)}, B^{(\nu)}) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(h_{ij}^{(\nu)})^2}{w_{ij}^{(\nu)}} + \\ + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}^{(\nu)}(\langle a_i, b_j\rangle - z_{ij}^{(\nu)})^2$$

The right hand side of this inequality, which we can write as $\mathcal{E}(A, B; A^{(\nu)}, B^{(\nu)})$, is the majorization function. It is minimized over A and B, producing $A^{(\nu+1)}$ and $B^{(\nu+1)}$. Now

$$\mathcal{D}(A^{(\nu+1)}, B^{(\nu+1)}) \leq \mathcal{E}(A^{(\nu+1)}, B^{(\nu+1)}; A^{(\nu)}, B^{(\nu)}) \leq \\ \leq \mathcal{E}(A^{(\nu)}, B^{(\nu)}; A^{(\nu)}, B^{(\nu)}) = \mathcal{D}(A^{(\nu)}, B^{(\nu)})$$

This proves the first part of the theorem.

The second part also follows directly. If (A, B) is a minimizer, and the algorithm is not stationary at (A, B), then we can update (A, B) to a solution with a lower function value, which contradicts the assumption that (A, B) is the minimum.

For the third part we apply general results on majorization given, for instance, in De Leeuw [1994]; Heiser [1995]; Lange et al. [2000]. □

For implementation details we refer to De Leeuw [2003b]. The most important choice we have to make is how to compute the weighted SVD. We can apply majorization once again, as in Groenen et al. [2003], to arrive at an unweighted

SVD. We can also use alternating least squares (a.k.a. criss-cross regression or NILES/NIPALS) iterations, as in Wold [1966a,b]; Daugavet [1968]. These alternate updating A for fixed B and B for fixed A, which are both linear regression problems, any number of times within on e major iteration. Alternatively, we can also choose to use a uniform majorization, in which $w(y)$ is independent of y. This leads directly to an unweighted SVD.

3. MULTIPLE STEP G_i

Let us assume that the G_i are step functions, stepping at a_{ik}. We must distinguish two cases. In the first the location and the size of the steps are unknown. This occurs if we try to approximate an unknown G_i by a step function. It also occurs if we optimize over all cdf's G_i, because the optimum G_i in that case will be a step function [Rustagi, 1976, Chapter IV]. The second case has known location and size of steps. This occurs if we approximate an integral by a linear quadrature formula, such as a product form of the Gauss-Hermite rule or a quasi-Monte Carlo rule.

We will outline the algorithm for the case in which the steps are unknown. Modifications for known steps are obvious. To simplify notation we assume that we have the same number of steps for each i. Using the step function representation we can write

$$\mathcal{D}(A, B, \pi) = -\sum_{i=1}^{n}\log\sum_{k=1}^{K}\pi_{ik}\exp\{\sum_{j=1}^{m}\log F(q_{ij}\langle a_{ik}, b_j\rangle)\}. \tag{6}$$

This loss function must be minimized over A and B and π, for various values of both K and p. We also need some convenient abbreviations. Define

$$\theta_{ik}(A, B, \pi) = \pi_{ik}\exp\{\sum_{j=1}^{m}\log F(q_{ij}\langle a_{ik}, b_j\rangle)\},$$

and

$$\xi_{ik}(A, B, \pi) = \frac{\theta_{ik}(A, B, \pi)}{\theta_{i\star}(A, B, \pi)},$$

where we use the convention of replacing an index over which we have summed by a star. Thus, for instance,

$$\mathcal{D}(A, B, \pi) = -\sum_{i=1}^{n}\log\theta_{i\star}(A, B, \pi).$$

The algorithm in this case starts with some $A^{(0)}$, $B^{(0)}$ and $\pi^{(0)}$. Suppose $A^{(\nu)}$, ($B^{(\nu)}$ and $\pi^{(\nu)}$ are the current best solution. We update them to find a better solution in

three steps, using majorization twice. The first majorization is based on Jensen's inequality, it is the same as the majorization used in the EM algorithm. The second majorization is based on the quadratic bound in Equation (4).

Algorithm 3.1 (Majorization)**.**

Step $\nu(1)$**:** *Compute the arrays* $W^{(\nu)}$, $H^{(\nu)}$*and* $Z^{(\nu)}$ *with elements*

$$\begin{aligned} w_{ijk}^{(\nu)} &= w(q_{ij}\langle a_{ik}^{(\nu)}, b_j^{(\nu)}\rangle),\\ h_{ijk}^{(\nu)} &= h(q_{ij}\langle a_{ik}^{(\nu)}, b_j^{(\nu)}\rangle),\\ z_{ijk}^{(\nu)} &= \langle a_{ik}^{(\nu)}, b_j^{(\nu)}\rangle - q_{ij}\frac{h_{ijk}^{(\nu)}}{w_{ijk}^{(\nu)}}. \end{aligned}$$

Also compute the matrices

$$\begin{aligned} \theta_{ik}^{(\nu)} &= \theta_{ik}(A^{(\nu)}, B^{(\nu)}, \pi^{(\nu)}),\\ \xi_{ik}^{(\nu)} &= \xi_{ik}(A^{(\nu)}, B^{(\nu)}, \pi^{(\nu)}), \end{aligned}$$

and the composite weights $\eta_{ijk}^{(\nu)} = \xi_{ik}^{(\nu)} w_{ijk}^{(\nu)}$.

Step $\nu(2)$**:** *Solve the least squares matrix approximation problem*

$$\min_{A,B} \sum_{k=1}^{K}\sum_{j=1}^{m} \eta_{ijk}^{(\nu)}(\overline{z}_{ijk}^{(\nu)} - \langle a_{ik}, b_j\rangle)^2$$

by using the (weighted) SVD. This gives $A^{(\nu+1)}$ *and* $B^{(\nu+1)}$.

Step $\nu(3)$**:** *Update* π *by*

$$\pi_{ik}^{(\nu+1)} = \xi_{ik}^{(\nu)}.$$

Theorem 3.1. (1) *The Majorization Algorithm 3.1 produces a decreasing sequence* $\mathcal{D}(A^{(\nu)}, B^{(\nu)}, \pi^{(\nu)})$ *of loss function values.*

(2) *A necessary condition for a minimum of the loss function* $\mathcal{D}$ *at* (A, B, π) *is that the algorithm is stationary at* (A, B, π).

(3) *All accumulation points of the sequence* $(A^{(\nu)}, B^{(\nu)}, \pi^{(\nu)})$ *of iterates are stationary points of the algorithm.*

Proof. By Jensen's inequality

$$\log \frac{\theta_{i\star}(A, B, \pi)}{\theta_{i\star}(A^{(\nu)}, B^{(\nu)}, \pi^{(\nu)})} = \log \sum_{k=1}^{K} \xi_{ik}^{(\nu)} \frac{\theta_{ik}(A, B, \pi)}{\tilde{\theta}_{ik}^{(\nu)}} \geq \sum_{k=1}^{K} \xi_{ik}^{(\nu)} \log \frac{\theta_{ik}(A, B, \pi)}{\theta_{ik}^{(\nu)}},$$

and thus

$$\mathcal{D}(A, B, \pi) \leq \mathcal{D}(A^{(\nu)}, B^{(\nu)}, \pi^{(\nu)}) + \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{ik}^{(\nu)} \log \theta_{ik}^{(\nu)} - \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{ik}^{(\nu)} \log \pi_{ik} - \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{ik}^{(\nu)} \sum_{j=1}^{m} \log F(q_{ij} \langle a_{ik}, b_j \rangle). \tag{7}$$

Equation (7) is the first majorization, which shows us how to optimize over π to decrease the loss. We now use the variational bound to majorize and simplify the last term of Equation (7). By Equation (4)

$$-\sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{ik}^{(\nu)} \sum_{j=1}^{m} \log F(q_{ij} \langle a_{ik}, b_j \rangle) \leq - \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{ik}^{(\nu)} \sum_{j=1}^{m} \log F(q_{ij} \langle a_{ik}^{(\nu)}, b_j^{(\nu)} \rangle) - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \xi_{ik}^{(\nu)} \sum_{j=1}^{m} \frac{(h_{ijk}^{(\nu)})^2}{w_{ijk}^{(\nu)}} + \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{m} \eta_{ijk}^{(\nu)} (\langle a_{ik}, b_j \rangle - z_{ijk}^{(\nu)})^2. \tag{8}$$

Combining Equations (7) and (8) gives us the final majorization, and shows us how to update A and B.

The other statements in the Theorem follow in the same way as in the proof of Theorem 2.1. □

In an important special case we require all G_i to be the same. Write a_k and π_k for the steps. Define

$$\overline{z}_{jk}^{(\nu)} = \frac{\sum_{i=1}^{n} \eta_{ijk}^{(\nu)} z_{ijk}^{(\nu)}}{\eta_{\star jk}^{(\nu)}}$$

By completing the square in the last term of Equation (8) we see that

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m}\eta_{ijk}^{(\nu)}(\langle a_k,b_j\rangle-z_{ijk}^{(\nu)})^2= \tag{9}$$
$$=\frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{m}\eta_{\star jk}^{(\nu)}(\langle a_k,b_j\rangle-\overline{z}_{jk}^{(\nu)})^2+\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\eta_{ijk}^{(\nu)}(z_{ijk}^{(\nu)}-\overline{z}_{jk}^{(\nu)})^2$$

Thus we update A and B by minimizing

$$\frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{m}\eta_{\star jk}^{(\nu)}(\langle a_k,b_j\rangle-\overline{z}_{jk}^{(\nu)})^2,$$

and we update π by

$$\pi_k^{(\nu+1)}=\frac{1}{K}\xi_{\star k}^{(\nu)}.$$

4. QUADRATURE

If we use a linear quadrature rule to approximate the integrals, then again

$$\mathcal{D}(B)=-\sum_{i=1}^{n}\log\sum_{k=1}^{K}\pi_k\exp\{\sum_{j=1}^{m}\log F(q_{ij}\langle a_k,b_j\rangle)\}. \tag{10}$$

but now the π_k and the a_k are known. The algorithm from the previous section simplifies accordingly, and updating B in each iteration is just a linear regression problem. Alternatively, we can stop after the first majorization of Equation (7) and update B by solving m logit or probit regression problems (or whatever other type of regression problem is specified by F).

5. DISCUSSION

REFERENCES

A.A. Beguin and C.W. Glas. MCMC Estimation and some Model-fit Analysis of Multidimensional IRT Models. *Psychometrika*, 66:541–562, 2001.

J. Clinton, S. Jackman, and D. Rivers. The Statistical Analysis of Roll Call Data. URL `http://jackman.stanford.edu/papers/masterideal.pdf`. 2003.

V.A. Daugavet. Variant of the Stepped Exponential Method of Finding Some of the First Characteristics Values of a Symmetric Matrix. *USSR Computation and Mathematical Physics*, 8(1):212–223, 1968.

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. Homogeneity Analysis of Pavings. URL `http://jackman.stanford.edu/ideal/MeasurementConference/abstracts/homPeig.pdf`. August 2003a.

J. De Leeuw. Principal Component Analysis of Binary Data. Applications to Roll-Call-Analysis. UCLA Statistics Preprints 364, UCLA Department of Statistics, 2003b. URL `http://preprints.stat.ucla.edu/364/`.

P.J.F. Groenen, P. Giaquinto, and H.L Kiers. Weighted Majorization Algorithms for Weighted Least Squares Decomposition Models. Technical Report EI 2003-09, Econometric Institute, Erasmus University, Rotterdam, Netherlands, 2003.

W.J. Heiser. Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis. In W.J. Krzanowski, editor, *Recent Advanmtages in Descriptive Multivariate Analysis*. Oxford: Clarendon Press, 1995.

S. Jackman. Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking. *Political Analysis*, 9 (3):227–241, 2001.

K. Lange, D.R. Hunter, and I. Yang. Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational and Graphical Statistics*, 9:1–20, 2000.

R.P. McDonald. Normal Ogive Multidimensional Model. In W.J. van der Linden and R.K. Hambleton, editors, *Handbook of Item Response Theory*. Springer, 1997.

X.-L. Meng and S. Schilling. Fitting Full-Information Item Factor Analysis Models and an Empirical Investigation of Bridge Sampling. *Journal of the American Statistical Association*, 91:1254–1267, 1996.

M.D. Reckase. A Linear Logistic Multidimensional Model. In W.J. van der Linden and R.K. Hambleton, editors, *Handbook of Item Response Theory*. Springer, 1997.

J.S. Rustagi. *Variational Methods in Statistics*. Number 121 in Mathematics in Science and Engineering. Academic Press, 1976.

H. Wold. Estimation of Principal Components and Related Models by Iterative Least Squares. In P.R. Krishnaiah, editor, *Multivariate Analysis*. Academic Press, 1966a.

H. Wold. Nonlinear Estimation by Iterative Least Squares Procedures. In F.N. David, editor, *Research Papers in Statistics. Festschrift for J. Neyman.* Wiley, 1966b.

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