MAJORIZATION IN MULTINORMAL MAXIMUM LIKELIHOOD ESTIMATION

JAN DE LEEUW

1. Introduction

The *deviance*[1] of n observations y_i from a p-dimensional centered multivariate normal distribution $\mathcal{N}(0,\Sigma)$ is, except for constants which do not depend on the parameters,

$$\mathbf{f}(\Sigma)=n\log\mathbf{det}(\Sigma)+\mathbf{tr}\ \Sigma^{-1}S, \tag{1}$$

where

$$S\triangleq\frac{1}{n}\sum_{i=1}^n y_iy_i' \tag{2}$$

is the *sample dispersion matrix*. We assume, throughout, that S is positive definite[2].

The problem we study is minimization of $\mathcal{D}(\Sigma)$ over subsets of the cone of positive definite matrices. Alternatively, we also look at minimization of

$$\mathbf{g}(\Omega)\triangleq-\log\mathbf{det}(\Omega)+\mathbf{tr}\ \Omega S, \tag{3}$$

over subsets of the cone of positive definite matrices. This is basically the same problem, except that in (1) the deviance is a function of the *covariance matrix* Σ, while in (3) it is a function of the *concentration matrix* $\Omega=\Sigma^{-1}$.

Remember that in both (1) and (3) the matrix S is a known positive definite matrix.

Date: February 3, 2003.

[1]Two times the negative log-likelihood.

[2]If it isn't, just add $n^{-1}I$.

Multivariate analysis problems are often formulated slightly differently, by writing the covariance matrix in the form $\sigma^2\Sigma$, where both σ^2 and $\Sigma \in \mathcal{S}$ must be estimated. Now

$$\mathcal{f}(\sigma^2\Sigma) = p\log\sigma^2 + \log\mathbf{det}(\Sigma) + \frac{1}{\sigma^2}\mathbf{tr}\ \Sigma^{-1}S.$$

The minimum over σ^2 is attained at $\hat\sigma^2 = p^{-1}\mathbf{tr}\ \Sigma^{-1}S$, and this minimum is equal to

$$\mathcal{f}(\hat\sigma^2\Sigma) = p\log\mathbf{tr}\ \Sigma^{-1}S + \log\mathbf{det}(\Sigma) + (p - p\log p).$$

Thus, ignoring irrelevant constants, we can also study the problem of minimizing

$$\mathcal{h}(\Sigma) \triangleq \log\mathbf{det}(\Sigma) + p\log\mathbf{tr}\ \Sigma^{-1}S \tag{4}$$

over Σ in some set of positive definite matrices.

1.1. **Majorization.** This paper is about majorization algorithms to minimize loss functions such as f, g, and h. Majorization algorithms are iterative algorithms, where in each iteration we find a majorization function and we minimize it. The majorization function has to be chosen in such a way that it is always above the loss function we are trying to minimize, except that it touches the loss function at the current point. For details we refer to the Appendix.

1.2. **Unrestricted Optimization.** If there are no restrictions on Σ, then the maximum likelihood estimate is just S. This is a classical result, we just give a proof to illustrate our basic techniques.

Lemma 1. *If A is a positive definite matrix and B is a real symmetric matrix, then there a non-singular matrix T and a diagonal matrix Λ such that $A = TT'$ and $B = T\Lambda T'$.*

Proof. Use the eigen-decomposition $A = K\Phi^2K'$. Define $\tilde B \triangleq \Phi^{-1}K'BK\Phi^{-1}$, and suppose $L\Lambda L'$ is the eigen-decomposition of $\tilde B$. Set $T \triangleq K\Phi L$. Clearly $TT' = A$. Moreover $T\Lambda T' = K\Phi L\Lambda L'\Phi K' = K\Phi\tilde B\Phi K' = B$. □

Theorem 2. *$f(\Sigma) \geq \log\mathbf{det}(S) + p$, with equality if and only if $\Sigma = S$.*

Proof. Suppose $\hat{\Sigma}$ minimizes f. Then by the Lemma there is a non-singular T and a positive definite diagonal Λ such that $S = TT'$ and $\hat{\Sigma} = T\Lambda T'$. Thus

$$\begin{aligned} f(\hat{\Sigma}) &= \log \mathbf{det}(S) + \log \mathbf{det}(\Lambda) + \mathbf{tr}\ \Lambda^{-1} \\ &= \log \mathbf{det}(S) + \sum_{s=1}^{p} \left(\log \lambda_s + \lambda_s^{-1}\right). \end{aligned}$$

Now $\log \lambda_s + \lambda_s^{-1}$ has a minimum equal to 1 at λ equal to 1. This is easily verified using derivatives. Thus $f(\Sigma) \geq \log \mathbf{det}(S) + p$. □

2. Smoothness and Convexity

2.1. **Convexity.** The following result is due to Ky Fan. See Beckenbach and Bellman [1965, Chapter 2, Paragraph 9] or Magnus and Neudecker [1998, Chapter 11, Section 22]. We give a proof using our basic lemma.

Theorem 3. $\log \mathbf{det}(\Sigma)$ *is concave in* Σ *on* $\mathcal{P}$.

Proof.

$$\begin{aligned} \log \mathbf{det}(\alpha\Sigma_1 + (1-\alpha)\Sigma_2) &= \log \mathbf{det}(\alpha T\Lambda T' + (1-\alpha)TT') = \\ &= \log \mathbf{det}(TT') + \log \mathbf{det}(\alpha\Lambda + (1-\alpha)I) = \\ &= \log \mathbf{det}(TT') + \sum_{i=1}^{n} \log(\alpha\lambda_i + (1-\alpha)). \end{aligned}$$

Because $\log(x)$ is concave on the positive half line

$$\log(\alpha\lambda_i + (1-\alpha)) \geq \alpha \log \lambda_i$$

and thus

$$\log \mathbf{det}(\alpha\Sigma_1 + (1-\alpha)\Sigma_2) \geq \alpha \log \mathbf{det}(\Sigma_1) + (1-\alpha) \log \mathbf{det}(\Sigma_2).$$

□

Theorem 4. $\mathbf{tr}\ \Sigma^{-1}C$ *is convex in* Σ *on* $\mathcal{P}$.

Proof.

$$\begin{aligned}\mathbf{tr}(\alpha\Sigma_1+(1-\alpha)\Sigma_2)^{-1}S&=\mathbf{tr}(\alpha T\Lambda T'+(1-\alpha)TT')^{-1}S=\\&=\mathbf{tr}(\alpha\Lambda+(1-\alpha)I)^{-1}T^{-1}ST^{-T}=\\&=\sum_{i=1}^n\frac{1}{\alpha\lambda_i+(1-\alpha)}\tilde{c}_{ii}\end{aligned}$$

with $\tilde{S}=T^{-1}CT^{-T}$.

Because x^{-1} is convex in x on the positive half-line, we see that

$$\frac{1}{\alpha\lambda_i+(1-\alpha)}\leq\alpha\frac{1}{\lambda_i}+(1-\alpha),$$

and thus

$$\mathbf{tr}(\alpha\Sigma_1+(1-\alpha)\Sigma_2)^{-1}S\leq\alpha\mathbf{tr}\ \Sigma_1^{-1}S+(1-\alpha)\mathbf{tr}\ \Sigma_2^{-1}S.$$

□

Corollary 5. *$g(\Omega)$ is convex in Ω on $\mathcal{P}$.*

Proof. $-\log\mathbf{det}(\Omega)$ is convex, and $\mathbf{tr}\ \Omega S$ is linear, and thus convex. □

2.2. **Differentiation.** A convenient text on matrix differentiation is Magnus and Neudecker [1998]. It contains the results in this section. We state the theorems and give the proofs in a slightly different form, using directional or Gateaux derivatives.

Theorem 6.

$$\log\mathbf{det}(\Sigma+\epsilon\Delta)=\log\mathbf{det}(\Sigma)+\epsilon\mathbf{tr}\ \Sigma^{-1}\Delta-\frac{1}{2}\epsilon^2\mathbf{tr}\ \Sigma^{-1}\Delta\Sigma^{-1}\Delta+o(\epsilon^2).$$

Proof.

$$\log\mathbf{det}(\Sigma+\epsilon\Delta)=\log\mathbf{det}(TT'+\epsilon T\Lambda T')=\log\mathbf{det}(TT')+\log\mathbf{det}(I+\epsilon\Lambda).$$

Now

$$\begin{aligned}\mathbf{det}(I+\epsilon\Lambda) &= 1+\epsilon\sum_s \lambda_s + \epsilon^2 \sum_{s<t}\sum \lambda_s\lambda_t + o(\epsilon^2) = \\ &= 1+\epsilon\sum_s \lambda_s + \frac{1}{2}\epsilon^2\{(\sum_s \lambda_s)^2 - \sum_s \lambda_s^2\} + o(\epsilon^2),\end{aligned}$$

and thus

$$\begin{aligned}\log \mathbf{det}(I+\epsilon\Lambda) &= \epsilon\sum_s \lambda_s - \frac{1}{2}\epsilon^2\sum_s \lambda_s^2 + o(\epsilon^2) = \\ &= \epsilon\mathbf{tr}\ \Lambda - \frac{1}{2}\epsilon^2\mathbf{tr}\ \Lambda^2 + o(\epsilon^2).\end{aligned}$$

Because $\Sigma^{-1}\Delta = T^{-T}\Lambda T'$ we see that $\mathbf{tr}\ \Sigma^{-1}\Delta = \mathbf{tr}\ \Lambda$ and moreover $\mathbf{tr}\ \Sigma^{-1}\Delta\Sigma^{-1}\Delta = \mathbf{tr}\ \Lambda^2$, giving the required result. □

Theorem 7.

$$\begin{aligned}&\mathbf{tr}\ (\Sigma+\epsilon\Delta)^{-1}S = \\ &\quad \mathbf{tr}\ \Sigma^{-1}S - \epsilon\mathbf{tr}\ \Sigma^{-1}\Delta\Sigma^{-1}S + \epsilon^2\mathbf{tr}\ \Sigma^{-1}\Delta\Sigma^{-1}\Delta\Sigma^{-1}S + o(\epsilon)^2.\end{aligned}$$

Proof. Let $\tilde{S} \triangleq T^{-1}ST^{-T}$. Then

$$\begin{aligned}&\mathbf{tr}\ (\Sigma+\epsilon\Delta)^{-1}S = \mathbf{tr}\ (I+\epsilon\Lambda)^{-1}\tilde{S} = \mathbf{tr}\ (I-\epsilon\Lambda+\epsilon^2\Lambda^2-\cdots)\tilde{S} = \\ &\quad \mathbf{tr}\ (\tilde{S}-\epsilon\Lambda\tilde{S}+\epsilon^2\Lambda^2\tilde{S}-\cdots) = \mathbf{tr}\ \tilde{S} - \epsilon\mathbf{tr}\ \Lambda\tilde{S} + \epsilon^2\mathbf{tr}\ \Lambda^2\tilde{S} - \cdots\end{aligned}$$

Now $\mathbf{tr}\ \tilde{S} = \mathbf{tr}\ \Sigma^{-1}S$, and using $\Sigma^{-1}\Delta = T^{-T}\Lambda T'$ we see immediately that $\mathbf{tr}\ \Lambda\tilde{S} = \mathbf{tr}\ \Sigma^{-1}\Delta\Sigma^{-1}S$, and $\mathbf{tr}\ \Lambda^2\tilde{S} = \mathbf{tr}\ \Sigma^{-1}\Delta\Sigma^{-1}\Delta\Sigma^{-1}S$. □

3. Majorization

Theorem 8.

$$\begin{aligned}\log \mathbf{det}(\Sigma_1) &\le \log \mathbf{det}(\Sigma_2) + \mathbf{tr}\ \Sigma_2^{-1}(\Sigma_1-\Sigma_2), \\ \log \mathbf{det}(\Sigma_1) &\ge \log \mathbf{det}(\Sigma_2) + \mathbf{tr}\ \Sigma_1^{-1}(\Sigma_1-\Sigma_2).\end{aligned}$$

Proof. To prove the first inequality we use the mean value theorem in the form

$$\log\mathbf{det}(\Sigma_1)\leq\log\mathbf{det}(\Sigma_2)+\max_{0\leq\alpha\leq 1}\mathbf{tr}(\alpha\Sigma_1+(1-\alpha)\Sigma_2)^{-1}(\Sigma_1-\Sigma_2)$$

Now, using the simultaneous diagonalization lemma,

$$\mathbf{tr}(\alpha\Sigma_1+(1-\alpha)\Sigma_2)^{-1}(\Sigma_1-\Sigma_2)=\sum_{s=1}^p\frac{\lambda_s-1}{\alpha\lambda_s+(1-\alpha)},$$

which is an decreasing function of α. Thus it attains its maximum at $\alpha=0$. It attains its minimum at $\alpha=1$, which proves the second inequality[3]. □

It should be emphasized that for majorization the first inequality is the useful one. It restates the well-known fact that a differentiable concave function can be majorized by a linear function at any point (or, to put it differently, a concave function is below any of its tangents). The second inequality is just given for completeness.

Theorem 9.

$$\begin{aligned}\mathbf{tr}\ \Sigma_1^{-1}S&\leq\mathbf{tr}\ \Sigma_2^{-1}S-\mathbf{tr}\ \Sigma_1^{-1}(\Sigma_1-\Sigma_2)\Sigma_1^{-1}S,\\ \mathbf{tr}\ \Sigma_1^{-1}S&\geq\mathbf{tr}\ \Sigma_2^{-1}S-\mathbf{tr}\ \Sigma_2^{-1}(\Sigma_1-\Sigma_2)\Sigma_2^{-1}S.\end{aligned}$$

Proof. The proof is very similar to that of Theorem 8, except that we now have

$$\mathbf{tr}\ (\alpha\Sigma_1+(1-\alpha)\Sigma_2)^{-1}(\Sigma_1-\Sigma_2)(\alpha\Sigma_1+(1-\alpha)\Sigma_2)^{-1}S=\sum_{s=1}^p\frac{(\lambda_s-1)\tilde{s}_{ss}}{(\alpha\lambda_s+(1-\alpha))^2},$$

where $\tilde{S}=T^{-1}ST^{-T}$. Again, this is decreasing in α, which gives the required result. □

[3]The second inequality can also be proved by interchanging Σ_1 and Σ_2 in the first inequality.

In Theorem 9 it is the other way around, the second inequality is the useful one. It says that a differentiable convex function can be minorized by a linear function, and that a convex function majorizes all its tangents.

3.1. **Quadratic Approximation.** We will now try to extend the techniques of the previous section to quadratic, instead of linear, majorization. We start with

$$\log \mathbf{det}(\Sigma_1) \leq \log \mathbf{det}(\Sigma_2) + \mathbf{tr}\ \Sigma_2^{-1}(\Sigma_1 - \Sigma_2) -$$
$$-\frac{1}{2}\min_{0\leq\alpha\leq 1} \mathbf{tr}\,(\alpha\Sigma_1+(1-\alpha)\Sigma_2)^{-1}(\Sigma_1-\Sigma_2)(\alpha\Sigma_1+(1-\alpha)\Sigma_2)^{-1}(\Sigma_1-\Sigma_2).$$

Now clearly

$$\mathbf{tr}\ (\alpha\Sigma_1 + (1-\alpha)\Sigma_2)^{-1}(\Sigma_1 - \Sigma_2)(\alpha\Sigma_1 + (1-\alpha)\Sigma_2)^{-1}(\Sigma_1 - \Sigma_2) = \sum_{s=1}^{p} \frac{(\lambda_s - 1)^2}{[\alpha\lambda_s + (1-\alpha)]^2}.$$

Unfortunately the terms of this sum are decreasing if $\lambda_s > 1$ and increasing if $\lambda_s < 1$. Thus the situtation is not as simple as it is with linear approximation.

$$\lambda^2_{\max}((\alpha\Sigma_1 + (1-\alpha)\Sigma_2)^{-1})\mathbf{tr}(\Sigma_1 - \Sigma_2)^2 = \frac{1}{\lambda^2_{\min}(\alpha\Sigma_1 + (1-\alpha)\Sigma_2)}\mathbf{tr}(\Sigma_1 - \Sigma_2)^2.$$

Now $\lambda_{\min}$ is concave. Thus

$$\frac{1}{\lambda^2_{\min}(\alpha\Sigma_1 + (1-\alpha)\Sigma_2)} \leq \frac{1}{(\alpha\lambda_{\min}(\Sigma_1) + (1-\alpha)\lambda_{\min}(\Sigma_2))^2} \leq \min(\lambda_{\min}(\Sigma_1), \lambda_{\min}(\Sigma_2))^{-2}.$$

Appendix A. Majorization Methods

A.1. **General Principles.** The algorithms proposed in this paper are all of the majorization type. Majorization is discussed in general terms in de Leeuw [1994], Heiser [1995], Lange et al. [2000].

In a majorization algorithm the goal is to optimize a function $\phi(\theta)$ over $\theta \in \Theta$, with $\Theta \subseteq \mathbb{R}^p$. Suppose that a function $\psi(\theta, \xi)$ defined on $\Theta \times \Theta$ satisfies

$$\phi(\theta) \geq \psi(\theta, \xi) \text{ for all } \theta, \xi \in \Theta, \tag{5a}$$

$$\phi(\theta) = \psi(\theta, \theta) \text{ for all } \theta \in \Theta. \tag{5b}$$

Thus, for a fixed ξ, $\psi(\bullet, \xi)$ is below ϕ, and it touches ϕ at the point $(\xi, \phi(\xi))$. We then say that $\phi(\theta)$ *majorizes* $\psi(\theta, \xi)$ or that $\psi(\theta, \xi)$ *minorizes* $\phi(\theta)$.

There are two key theorems associated with these definitions.

Theorem 10. *If ϕ attains its maximum on Θ at $\hat{\theta}$, then $\psi(\bullet, \hat{\theta})$ also attains its maximum on Θ at $\hat{\theta}$.*

Proof. Suppose $\psi(\tilde{\theta}, \hat{\theta}) > \psi(\hat{\theta}, \hat{\theta})$ for some $\tilde{\theta} \in \Theta$. Then, by (5a) and (5b), $\phi(\tilde{\theta}) \geq \psi(\tilde{\theta}, \hat{\theta}) > \psi(\hat{\theta}, \hat{\theta}) = \phi(\hat{\theta})$, which contradicts the definition of $\hat{\theta}$ as the maximizer of ϕ on Θ. □

Theorem 11. *If $\tilde{\theta} \in \Theta$ and $\hat{\theta}$ maximizes $\psi(\bullet, \tilde{\theta})$ over Θ, then $\phi(\hat{\theta}) \geq \phi(\tilde{\theta})$.*

Proof. By (5a) we have $\phi(\hat{\theta}) \geq \psi(\hat{\theta}, \tilde{\theta})$. By the definition of $\hat{\theta}$ we have $\psi(\hat{\theta}, \tilde{\theta}) \geq \psi(\tilde{\theta}, \tilde{\theta})$. And by (5b) we have $\psi(\tilde{\theta}, \tilde{\theta}) = \phi(\tilde{\theta})$. Combining these three results we get the result. □

These two results suggest the following algorithm for maximizing $\phi(\theta)$.

Step 1:: Given a value $\theta^{(k)}$ construct a minorizing function $\psi(\theta^{(k)}, \xi)$.
Step 2:: Maximize $\psi(\theta^{(k)}, \xi)$ with respect to ξ. Set $\theta^{(k+1)} = \xi^{\max}$.

Step 3:: If $|\phi(\theta^{(k+1)} - \phi(\theta^{(k)}| < \epsilon$ for some predetermined $\epsilon > 0$ stop; else go to Step 1.

In order for this algorithm to be of practical use, the minorizing function ψ needs to be easy to maximize, otherwise nothing substantial is gained by following this route. Notice, that in case we are interested to minimize ϕ, we have to find a majorizing function ψ that needs to be minimized in Step 2.

We demonstrate next how the idea behind majorization works with a simple example.

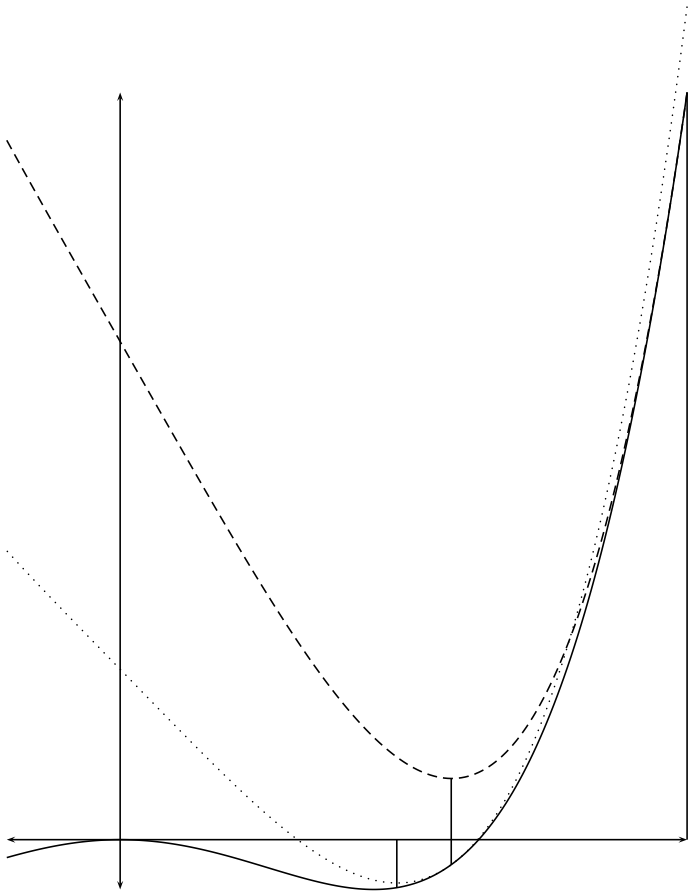

Figure 1: Majorization

Example 1. *This is an artificial example, chosen for its simplicity. Consider* $\phi(\theta) = \theta^4 - 10\theta^2,\ \theta \in \mathbb{R}$. *Because* $\theta^2 \geq \xi^2 + 2\xi(\theta - \xi) = 2\xi\theta - \xi^2$ *we see that* $\psi(\theta, \xi) = \theta^4 - 20\xi\theta + 10\xi^2$ *is a suitable majorization function. The majorization algorithm is* $\theta^+ = \sqrt[3]{5\xi}$.

The algorithm is illustrated in Figure A.1. We start with $\theta(0) = 5$. *Then* $\psi(\theta, 5)$ *is the dashed function. It is minimized at* $\theta^{(1)} \approx 2.924$, *where* $\psi(\theta^{(1)}, 5) \approx 30.70$, *and* $\phi(\theta^{(1)}) \approx -12.56$. *We then majorize by using the dotted function* $\psi(\theta, \theta^{(1)})$, *which has its minimum at about* 2.44, *equal to*

about -21.79. *The corresponding value of* ϕ *at this point is about* -24.1. *Thus we are rapidly getting close to the local minimum at* $\sqrt{5}$, *with value* 25. *The linear convergence rate at this point is* $\frac{1}{3}$.

We briefly address next some convergence issues (for a general discussion see the book by Zangwill [1969] and also Meyer [1976]). If ϕ is bounded above (below) on Θ, then the algorithm generates a bounded increasing sequence of function values $\phi(\theta^{(k)}$, thus it converges to $\phi(\theta^{\infty})$. For example, continuity of ϕ and compactness of Θ would suffice for establishing the result. Moreover with some addiitonal mild continuity considerations we get that $||\theta^{(k)} - \theta^{(k+1)}|| \rightarrow 0$ [de Leeuw, 1990], which in turn implies, because of a result by Ostrowski [1966], that θ converges either to a stable point or to a continuum of limit points. Hence, majorization algorithms for all practical purposes find local optima, and by starting the algorithm at different initial values global optima can be located.

References

E. F. Beckenbach and R. Bellman. *Inequalities*. Springer-Verlag, Berlin, Germany, 1965.

J. de Leeuw. Generalized eigenvalue problems with positive semidefinite matrices. *Psychometrika*, 47:87–94, 1982.

J. de Leeuw. Multivariate analysis with optimal scaling. In S. Das Gupta and J. Sethuraman, editors, *Progress in Multivariate Analysis*, Calcutta, India, 1990. Indian Statistical Institute.

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.

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. T. Jensen, S. Johansen, and S. L. Lauritzen. Globally convergent algorithms for maximizing a likelihood function. *Biometrika*, 78:867–877, 1991.

K. Lange, D.R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9:1–20, 2000.

J.R. Magnus and H. Neudecker. *Matrix Differential Calculus, with Applications in Statistics and Econometrix*. Wiley, New York, 1998.

R. R. Meyer. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *Journal of Computer and System Sciences*, 12:108–121, 1976.

W. Oberhofer and J. Kmenta. A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica*, 42: 579–590, 1974.

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

W. I. Zangwill. *Nonlinear Programming: a Unified Approach*. Prentice-Hall, Englewood-Cliffs, N.J., 1969.