MAJORIZATION METHODS FOR LOGISTIC REGRESSION

JAN DE LEEUW

1. Introduction

The majorization method [De Leeuw, 1994; Heiser, 1995; Lange et al., 2000; Ahn et al., 2006] for minimization of real valued loss functions has become very popular in statistics and computer science (under a wide variety of names). We give a brief introduction. Suppose the problem is to minimize a real valued function $\phi(\bullet)$ over $\Theta \subseteq \mathbb{R}^p$,

We say that a real valued function $\psi(\bullet)$ *majorizes* $\phi(\bullet)$ *over* Θ *in* $\xi \in \Theta$ if

$$\phi(\theta) \leq \psi(\theta) \qquad \forall \theta \in \Theta, \tag{1a}$$

$$\phi(\xi) = \psi(\xi). \tag{1b}$$

In words, $\psi(\bullet)$ must be above $\phi(\bullet)$ in all of Θ, and touches $\phi(\bullet)$ in ξ. We say that $\psi(\bullet)$ *strictly majorizes* $\phi(\bullet)$ *over* Θ *in* $\xi \in \Theta$ if we have (1a) and

$$\phi(\theta) = \psi(\theta) \text{ if and only if } \theta = \xi. \tag{2}$$

In words, $\psi(\bullet)$ must be above $\phi(\bullet)$ in all of Θ, and touches $\phi(\bullet)$ *only* in ξ.

Now suppose that we have a function $\psi(\bullet, \bullet)$ on $\Theta \otimes \Theta$ such that

$$\phi(\theta) \leq \psi(\theta.\xi) \qquad \forall \theta, \xi \in \Theta, \tag{3a}$$

$$\phi(\xi) = \psi(\xi, \xi) \qquad \forall \xi \in \Theta. \tag{3b}$$

Date: June 4, 2007.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis.

Thus for all $\xi \in \Theta$ the function $\psi(\bullet, \xi)$ majorizes $\phi(\bullet)$ over Θ in ξ. In this case we simply say that $\psi(\bullet, \bullet)$ *majorizes* $\psi(\bullet)$ *over* Θ. Also, $\psi(\bullet, \bullet)$ *strictly majorizes* $\psi(\bullet)$ *over* Θ if for all $\xi \in \Theta$ the function $\psi(\bullet, \xi)$ strictly majorizes $\phi(\bullet)$ over Θ in ξ. Jacobson and Fessler [2004] call the function $\psi(\bullet, \bullet)$ a *majorant generator*. From the computational point of view the trick is to find a majorant generator which is relatively simple to minimize over θ for each ξ.

Each majorization function can be used to define an algorithm. In each step of such a *majorization algorithm* we find the update $\theta^{(k+1)}$ by minimizing $\psi(\bullet, \theta^{(k)})$ over Θ, i.e. we choose

$$\theta^{(k+1)} \in \underset{\theta \in \Theta}{\mathbf{Argmin}}\, \psi(\theta, \theta^{(k)}).$$

The minimum of $\psi(\bullet, \theta^{(k)})$ over Θ may not be unique, and consequently $\mathbf{Argmin}(\bullet)$ is a set-valued map. If the minimum is unique, we use the single-valued version and set

$$\theta^{(k+1)} = \underset{\theta \in \Theta}{\mathbf{argmin}}\, \psi(\theta, \theta^{(k)}).$$

The algorithm includes a simple stopping rule. If

$$\theta^{(k)} \in \underset{\theta \in \Theta}{\mathbf{Argmin}}\, \psi(\theta, \theta^{(k)})$$

then we stop. If we never stop, then we obviously generate an infinite sequence.

Now suppose the algorithm generates an infinite sequence. For each step of the algorithm the *sandwich inequality*

$$\phi(\theta^{(k+1)}) \leq \psi(\theta^{(k+1)}, \theta^{(k)}) < \psi(\theta^{(k)}, \theta^{(k)}) = \phi(\theta^{(k)}) \tag{4}$$

shows that an iteration decreases the value of the loss function [De Leeuw, 1994]. The strict inequality $\psi(\theta^{(k+1)}, \theta^{(k)}) < \psi(\theta^{(k)}, \theta^{(k)})$ follows from the fact that we do not stop, which implies that $\theta^{(k)}$ is not a minimizer of $\psi(\bullet, \theta^{(k)})$. This is used to prove convergence of the algorithm, using general results such as those of Zangwill [1969].

2. Linear Logistic Regression

2.1. **Loss Function.** In *logistic regression* we minimize a negative log-likelihood function of the form

$$\phi(\theta) = -\sum_{i=1}^{n} \{y_i \log F(x_i'\theta) + (1 - y_i) \log(1 - F(x_i'\theta))\}, \tag{5}$$

where

$$F(x) = \frac{1}{1 + \exp(-x)} \tag{6}$$

is the cumulative logistic distribution function. The y_i are our binary data, and the x_i are vectors with the values of p predictors.

In many practical applications, such as bio-assay or psychophysics, the number of different values of the predictors x_i may be much smaller than n. In other words, we have a design in which the values of the regressors are replicated. In this case we usually write

$$\phi(\theta) = -\sum_{j=1}^{m} \left\{ n_j \log F(x_j'\theta) + (N_j - n_j) \log(1 - F(x_j'\theta)) \right\}, \tag{7}$$

Here n_j is the number of positive outcomes for treatment x_j, and N_j is the total number of times treatment x_j is offered.

We will continue to work with (5), because it is somewhat more convenient. If we define $z_i = (1 - 2y_i)x_i$, for example, then the log-likelihood can be written in the more compact form

$$\phi(\theta) = -\sum_{i=1}^{n} \log F(z_i'\theta) = \sum_{i=1}^{n} \log(1 + \exp(-z_i'\theta)),$$

2.2. **Boundedness and Convexity.** Because $0 < F(x) < 1$ for all x, it follows directly that $\phi(\theta) > 0$ for all θ. Thus the negative log-likelihood is bounded below by zero. Since it is continuous (in fact, infinitely many times differentiable) this implies that $\hat{\phi} = \inf_\theta \phi(\theta)$ is finite and non-negative. Observe that $\hat{\phi} \leq \phi(0) = n \log 2$, with

equality if and only if Z is centered, i.e. the columns of Z add up to zero.

We see, from Appendix C,

$$\mathcal{D}\phi(\theta) = -Z'g(\theta),$$

where $g(\theta)$ has elements $1 - F(z_i'\theta)$, and

$$\mathcal{D}^2\phi(\theta) = Z'H(\theta)Z,$$

where $H(\theta)$ is diagonal, with elements $F(z_i'\theta)(1 - F(z_i'\theta))$.

The diagonal elements of $H(\theta)$ satisfy $0 < h_{ii}(\theta) \leq \frac{1}{4}$. Thus $H(\theta)$, and consequently $\mathcal{D}^2\phi(\theta)$ is positive semi-definite for all θ, and $\phi(\theta)$ is convex. Each local minimum is a global minimum, and the set of θ where the minimum is attained is a compact convex set (which may be empty). Moreover $\textbf{rank}(\mathcal{D}^2\phi(\theta)) = \textbf{rank}(Z)$, so if we parametrize our problem such that Z is of full column-rank p we see that $\mathcal{D}^2\phi(\theta)$ is positive definite and the negative log-likelihood is strictly convex. If the minimum is attained, it is attained at a unique point.

2.3. **Existence.** We still have to investigate if the minimum is indeed attained, i.e. if maximum likelihood estimates exist or if there is a $\hat{\theta}$ such that $\hat{\phi} = \phi(\hat{\theta})$. For logistic regression the existence problem was first studied systematically by Albert and Anderson [1984], with improvements by Sandler and Duffy [1986]. For general discrete exponential families results were given by Jacobsen [1989]. We give a slightly different treatment, using general tools from convex analysis. Also compare Kaufmann [1988]. If maximum likelihood estimates do not exist, we can still compute *extended maximum likelihood estimates* [Haberman, 1974, Appendix B], in which some or all of the components of θ are equal to $\pm\infty$. Computational aspects of extended maximum likelihood estimation are discussed in Clarkson and Jennrich [1991].

It is useful to first discuss the case $\hat{\phi} = 0$. This happens if and only if the system of linear inequalities $Z\theta > 0$ has a solution, i.e. we can choose θ such that $x_i'\theta < 0$ for all i such that $y_i = 1$ and $x_i'\theta > 0$ for all i such that $y_i = 0$. If θ is such that $Z\theta > 0$ we can simply use $\lim_{\lambda\to\infty}\phi(\lambda\theta) = 0$. Geometrically $Z\theta > 0$ means there is a hyperplane strictly separating the two sets of points X_1 and X_0 in $\mathbb{R}^p$, with X_1 corresponding with the x_i for which $y_i = 1$ and X_0 with the x_i for which $y_i = 0$. In the terminology of Albert and Anderson [1984] this means we have *complete separation*.

Now consider the case where the minimum is attained at some θ. To study this we use the *asymptotic function* (also known as the *recession function*), defined by

$$\phi'_\infty(d) = \sup_{\tau>0}\frac{\phi(x+\tau d)-\phi(x)}{\tau} = \lim_{\tau\to\infty}\frac{\phi(x+\tau d)-\phi(x)}{\tau}.$$

Observe that, as the notation suggest, the value of the recession function only depends on the direction d, and not on x [Hiriart-Urruty and Lemaréchal, 1993, p. 178–183]. The general result we need is that $\phi'_\infty(d) > 0$ for all $d \neq 0$ is necessary and sufficient for $\phi(\bullet)$ to have a nonempty (and thus compact and convex) set of minimum points.

Define $\phi_i(\theta) = -\log F(z_i'\theta) = \log(1+\exp(-z_i'\theta))$. By simple calculation

$$(\phi_i)'_\infty(d) = \begin{cases} 1 & \text{if } z_i'd < 0, \\ 0 & \text{if } z_i'd \geq 0. \end{cases}$$

It follows that $\phi'_\infty(d) = \sum_{i=1}^n(\phi_i)'_\infty(d)$ is the number of i for which $z_i'd < 0$. We conclude that the minimum exists if and only if for all $d \neq 0$ there is at least one i such that $z_i'd < 0$. Or, in other words, if and only if the homogeneous system of linear inequalities $Z\theta \geq 0$ only has the trivial solution $\theta = 0$.

Maximum likelihood estimates do *not* exist if and only if $Z\theta \geq 0$ has a non-trivial solution θ. In this last case we have *complete separation* if there is a θ such that $Z\theta > 0$. We have *partial separation* if

there is no complete separation, but there is a solution θ in which some of the components of $Z\theta$ are positive and some are zero.

2.4. **Systems of Linear Inequalities.** It follows from these considerations that optimizing the logistic regression loss function can be used to compute solutions to systems of linear inequalities. This is closely related to a proposal by Motzkin [1952], taken up later by Stewart [1987], to solve $Z\theta \geq 0$ by minimizing

$$\phi_{MS}(\theta) = \sum_{i=1}^{n} \exp(-z_i'\theta). \tag{8}$$

See also Borwein and Lewis [2000, p. 23–27]. Similar to the negative log-likelihood $\phi(\bullet)$ the function $\phi_{MS}(\bullet)$ is convex and has a minimum if and only if the system $Z\theta \geq 0$ does *not* have a non-trivial solution. Stewart suggests using Newton's method with line search to minimize (8), and studies its properties.

In a related development Chen and Mangasarian [1996] solve the system $Z\theta \leq 0$ by minimizing

$$\phi_{CM}(\theta, \tau) = \sum_{i=1}^{n} p(z_i'\theta, \tau), \tag{9}$$

where

$$p(x, \tau) = x + \frac{1}{\tau}\log(1 + \exp(-\tau x)).$$

The function $p(x, \tau)$ is a smoothed (infinitely differentiable, strictly convex, and strictly increasing) version of $(x)_+ = \max(x, 0)$. In fact $p(x, \tau) > (x)_+$ for all x, and $\lim_{\tau\to\infty} p(x, \tau) = (x)_+$. The relationship with the negative logistic log-likelihood is

$$\phi(\theta) = \phi_{CM}(\theta, 1) - r'\theta,$$

where

$$r = \sum_{i=1}^{n} y_i z_i.$$

2.5. **Newton's Method.** The standard algorithm for linear logistic maximum likelihood is Newton's method. We discuss it briefly for reference purposes. Newton's method takes the simple form

$$\theta^{(k+1)} = \theta^{(k)} - (Z'H(\theta^{(k)})Z)^{-1}Z'g(\theta^{(k)}).$$

Applying the Newton algorithm without safeguards usually leads to problems. If the quadratic approximation is poor, the algorithm may make steps which are too large and this can lead to non-convergence. We also have to take into account that the minimum may not exist, and thus the iterates cannot possibly converge.

3. Quadratic Majorization

3.1. **Uniform Quadratic Majorization.** By the mean value theorem

$$\begin{aligned}\phi(\theta) \leq \phi(\theta^{(k)}) - (\theta - \theta^{(k)})'Z'g(\theta^{(k)}) + \\ + \frac{1}{2}\sup_{0\leq\lambda\leq 1}(\theta - \theta^{(k)})'Z'H(\lambda\theta + (1-\lambda)\theta^{(k)})Z(\theta - \theta^{(k)}),\end{aligned}$$

and since

$$H(\lambda\theta + (1-\lambda)\theta^{(k)}) \leq \frac{1}{4}I,$$

we have

$$\begin{aligned}\phi(\theta) \leq \phi(\theta^{(k)}) - (\theta - \theta^{(k)})'Z'g(\theta^{(k)}) + \\ + \frac{1}{8}(\theta - \theta^{(k)})'Z'Z(\theta - \theta^{(k)}).\end{aligned}$$

The corresponding majorization algorithm is

$$\theta^{(k+1)} = \theta^{(k)} + 4(Z'Z)^{-1}Z'g(\theta^{(k)}). \tag{10}$$

Alternatively, we can use any matrix norm $\|Z'Z\|$ to derive the simpler algorithm

$$\theta^{(k+1)} = \theta^{(k)} + \frac{4}{\|Z'Z\|}Z'g(\theta^{(k)}). \tag{11}$$

The obvious choice in (11) is to use the spectral norm, i.e. the largest eigenvalue of $Z'Z$.

3.2. **Relaxation.** For the sandwich inequality to apply it is not necessary to actually minimize the majorization function. It suffices to decrease it. In fact, let $F(\bullet)$ be a mapping of Θ into Θ such that

$$\psi(F(\theta^{(k)}),\theta^{(k)})\leq\psi(\theta^{(k)},\theta^{(k)})).$$

Then the sandwich inequality still applies, and under strict majorization we still have $\phi(\theta^{(k+1)})<\phi(\theta^{(k)})$ if we set $\theta^{(k+1)}=F(\theta^{(k)})$.

In our quadratic majorization example we can consider the algorithm

$$\theta^{(k+1)}=\theta^{(k)}-K(\pi(2+\theta^{(k)})-\pi(1-\theta^{(k)})).$$

Now

$$\psi(\theta^{(k+1)},\theta^{(k)})=\phi(\theta^{(k)})+(\frac{1}{4}K^2-K)(\pi(2+\theta^{(k)})-\pi(1-\theta^{(k)}))^2,$$

and thus $\psi(\theta^{(k+1)},\theta^{(k)})\leq\phi(\theta^{(k)})$ for $0\leq K\leq 4$.

The case $K=0$ is uninteresting, because any point is a fixed point and nothing changes. The case $K=4$ is of some interest, however. We move to a point equally far from the minimum as the current solution, or majorization point, but on the other side of the minimum. This is sometimes known as *over-relaxation*. At this over-relaxed point we have $\psi(\theta^{(k+1)},\theta^{(k)})=\phi(\theta^{(k)})$, but in the case of strict relaxation this still gives $\phi(\theta^{(k+1)})<\phi(\theta^{(k)})$.

The linear convergence rate of the algorithm with step-size K is

$$|1-K\phi''(-.5))|\approx|1-0.2982929K|$$

Thus for $K=4$ we obtain a rate of 0.1931716 and convergence which is about twice as fast as before (see the "overrel" column in Table **??**). The first two iterations of the over-relaxed algorithm are in Figure **??**.

[Figure 1 about here.]

For the very special choice

$$K = \frac{1}{2 + \sigma(1.5)} \approx 3.352410$$

we have superlinear convergence (but of course we can only use this step-size if we already know the solution). See the "optrel" column in Table **??**.

3.3. **Sharp Quadratic Majorization.**

4. Cubic Majorization

So far our majorization methods have linear convergence (unless we are very lucky). It is quite straightforward, however, to construct majorization methods with superlinear convergence.

Define

$$\mu(\theta) = \sigma'(\theta) = \pi''(\theta) = \pi(\theta)(1 - \pi(\theta))(1 - 2\pi(\theta)).$$

Some simple computation gives

$$-\frac{1}{18}\sqrt{3} \leq \mu(\theta) \leq +\frac{1}{18}\sqrt{3}.$$

This means that

$$\frac{1}{9}\sqrt{3} \leq \phi'''(\theta) = \mu(2 + \theta) - \mu(1 - \theta) \leq \frac{1}{9}\sqrt{3},$$

and thus

$$\begin{aligned}\psi(\theta, \xi) = \phi(\xi) + (\pi(2 + \xi) - \pi(1 - \xi))(\theta - \xi) + \\ + \frac{1}{2}(\sigma(1 - \xi) + \sigma(2 + \xi))(\theta - \xi)^2 + \frac{\sqrt{3}}{54}|\theta - \xi|^3\end{aligned}$$

is a majorization of $\phi(\bullet)$. The majorization function seems somewhat non-standard, because it involves the absolute value of the

cubic term. Nevertheless it is two times continuously differentiable. In fact, it is also strictly convex, because the second derivative is

$$\mathcal{D}_{11}\psi(\theta,\xi)=\begin{cases}(\sigma(1-\xi)+\sigma(2+\xi))+\frac{\sqrt{3}}{9}(\theta-\xi) & \text{for } \theta\geq\xi,\\(\sigma(1-\xi)+\sigma(2+\xi))-\frac{\sqrt{3}}{9}(\theta-\xi) & \text{for } \theta\leq\xi.\end{cases}$$

which is clearly positive.

To find the minimum we set the first derivative equal to zero. The first derivative at ξ is equal to $\pi(2+\xi)-\pi(1-\xi)$. If this is positive then the minimum is attained at a value smaller than ξ. In this case the quadratic

$$\pi(2+\xi)-\pi(1-\xi)+(\sigma(1-\xi)+\sigma(2+\xi))\zeta-\frac{\sqrt{3}}{18}\zeta^2=0$$

has two real roots $\zeta_1<0<\zeta_2$ and the minimum we look for is attained at $\xi+\zeta_1$. If the derivative at zero $\pi(2+\xi)-\pi(1-\xi)$ is negative, then

$$(\pi(2+\xi)-\pi(1-\xi))+(\sigma(1-\xi)+\sigma(2+\xi))\zeta+\frac{\sqrt{3}}{18}\zeta^2=0$$

again has two real roots $\zeta_1<0<\zeta_2$ and the minimum of the majorization function is attained at $\xi+\zeta_2$.

For the derivative of the algorithmic map $\xi+\zeta(\xi)$ we find

$$1-\frac{\phi''(\xi)+\zeta(\xi)\phi'''(\xi)}{\phi''(\xi)+\zeta(\xi)\frac{1}{18}\sqrt{3}}.$$

At a fixed point $\zeta(\xi)=0$ and thus the derivative is zero, which implies superlinear convergence.

5. Quartic Majorization

Define

$$\lambda(\theta)=\pi'''(\theta)=\pi(\theta)(1-\pi(\theta))(1-6\pi(\theta)+6\pi^2(\theta)).$$

We find that

$$-\frac{1}{24}\leq\lambda(\theta)\leq\frac{1}{8}.$$

Thus

$$\phi''''(\theta) = \lambda(2+\theta) + \lambda(1-\theta) \leq \frac{1}{4}.$$

The majorization function is

$$\begin{aligned}\psi(\theta,\xi) &= \phi(\xi) + (\pi(2+\xi) - \pi(1-\xi))(\theta-\xi) + \\ &+ \frac{1}{2}(\sigma(1-\xi) + \sigma(2+\xi))(\theta-\xi)^2 + \frac{1}{6}(\mu(2+\xi) - \mu(1-\xi))(\theta-\xi)^3 + \\ &+ \frac{1}{96}(\theta-\xi)^4.\end{aligned}$$

The second partials are

$$\begin{aligned}\mathcal{D}_{11}\psi(\theta,\xi) &= (\sigma(1-\xi) + \sigma(2+\xi)) + \\ &(\mu(2+\xi) - \mu(1-\xi))(\theta-\xi) + \frac{1}{8}(\theta-\xi)^2.\end{aligned}$$

This quadratic has no real roots (conjecture so far), and since it is positive for $\theta = \xi$ the quartic majorization function is strictly convex.

Setting the derivative equal to zero means solving a cubic with only one real root. This root gives the minimum of the majorization function.

Appendix A. Numerical Examples

Consider the data from Maxwell [1961, page 64] in Table C. They indicate the number of boys in a clinic classified as inveterate liars by the resident psychiatrist. We do a simple logistic regression on age, which means Z has a column of ones and a column with the numbers one to five. The maximum likelihood solution for the intercept is $\theta_0 \approx -1.1971$, while that for the slope is $\theta_1 \approx 0.2737$.

[Table 1 about here.]

[Table 2 about here.]

For completeness, we also study the quite different Cancer Remission data of Lee [1974], given in Table C. These data have a binary

outcome (given in the last column), indicating in which of 27 patients remission occurred. There are six variables (plus the intercept), indicating the results of several medical tests.

Appendix B. Additive Regression Functions

Theorem B.1. *For a function of the form*

$$\phi(\theta) = \sum_{i=1}^{n} f(z_i'\theta)$$

we have

$$\phi^{(r)}(\theta) = \sum_{i=1}^{n} f^{(r)}(z_i'\theta)\overbrace{z_i \otimes \cdots \otimes z_i}^{r \text{ times}}.$$

Appendix C. The log-logistic

Theorem C.1. *The function $f(x) = -\log F(x) = \log(1 + \exp(-x))$ is strictly convex.*

Proof. Elementary computation gives

$$\begin{aligned} f'(x) &= 1 - F(x), \\ f''(x) &= F(x)(1 - F(x)). \end{aligned}$$

Thus the first derivative is strictly decreasing, the second derivative is positive. This proves strict convexity. □

Theorem C.2. *The r-th derivative $f^{(r)}(x)$ is a polynomial in $F(x)$ of degree r. Consequently for all r there are two finite real numbers $m_r < M_r$ such that $m_r \leq f^{(r)}(x) \leq M_r$ for all x.*

Proof. From the previous theorem we see the result is true for $r = 1$ and $r = 2$. Now proceed by induction. If $f^{(r)}(x) = P_r(F(x))$ for some polynomial P_r of degree r, then $f^{(r+1)}(x) = P_r'(F(x))F(x)(1-$

$F(x)$), which is indeed a polynomial in $F(x)$ of degree $r+1$. In addition

$$\sup_x f^{(r)}(x) = \max_{0\leq s\leq 1} P_r(s),$$
$$\inf_x f^{(r)}(x) = \min_{0\leq s\leq 1} P_r(s),$$

and the quantities on the right-hand side are clearly finite. □

We illustrate Theorem C.2 by computiong some higher derivatives

$$f^{(3)}(x) = -F(x)(1-F(x))(1-2F(x)),$$

$$f^{(4)}(x) = -F(x)(1-F(x))(1-6F(x)+6F^2(x)),$$

$$f^{(5)}(x) = -F(x)(1-F(x))(1-2F(x))(1-12F(x)+12F^2(x))$$

which implies

$$-\frac{1}{18}\sqrt{3} \leq f^{(3)}(x) \leq +\frac{1}{18}\sqrt{3},$$
$$-\frac{1}{24} \leq f^{(4)}(x) \leq \frac{1}{8}$$

[Figure 2 about here.]

We now look more generally at the polynomials P_r. From the proof of Theorem C.2 we see that for $r>1$ we have $P_r(0) = P_r(1) = 0$. Because $P_2(s) = P_2(1-s)$ we see that actually $P_r(s) = P_r(1-s)$ for all even r and $P_r(s) = -P_r(1-s)$ for all odd $r>1$. This implies that $P_r(\frac{1}{2}) = 0$ for all odd $r>1$.

In fact we can get further than this and derive an explicit form for the polynomials. The difference/differential equation we have to solve is

$$P_{r+1}(x) = x(1-x)P_r'(x),$$

where $P_1(x) = 1-x$. The general solution (Tom Ferguson, personal communication, 03/06/05) is

$$P_r(x) = \sum_{j=1}^{r} (j-1)!S(j,r)(1-x)^j$$

where the $S(j,r)$ are the Stirling numbers of the second kind (the number of ways of partitioning r elements into j non-empty subsets).

[Table 3 about here.]

REFERENCES

S. Ahn, J.A. Fessler, Doron Blatt, and A.O. Hero. Convergent Incremental Optimization Transfer Algorithms: Application to Tomography. *IEEE Transactions on Medical Imaging*, 25:283–296, 2006.

A. Albert and J.A. Anderson. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 71: 1–10, 1984.

J.M. Borwein and A.S. Lewis. *Convex Analysis and Nonlinear Optimization*. Number 3 in CMS Books in Mathematics. Springer-Verlag, 2000.

C. Chen and O. L. Mangasarian. Smoothing Methods for Convex Inequalities and Linear Complementary Problems. *Mathematical Programming*, 71:51–69, 1996.

D.B. Clarkson and R.I. Jennrich. Computing Extended Maximum Likelihood Estimates for Linear Parameter Models. *Journal of the Royal Statistical Society B*, 53:417–426, 1991.

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.

S.J. Haberman. *The Analysis of Frequency Data*. University of Chicago Press, 1974.

W.J. Heiser. Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis. In W.J. Krzanowski, editor, *Recent Advances in Descriptive Multivariate Analysis*, pages 157–189. Oxford: Clarendon Press, 1995.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Number 305 in Grundlehren der mathematischen Wissenschaften. Springer Verlag, New York, 1993.

M. Jacobsen. Existence and Unicity of MLEs in Discrete Exponential Family Distributions. *Scandinavian Journal of Statistics*, 16:335–349, 1989.

M.W. Jacobson and J.A. Fessler. Properties of MM Algorithms on Convex Feasible Sets: Extended Version. Technical Report 353, Communication and Signal Porcessing Laboratory, Depratment of EECS, University of Michigan, November 2004. URL `http://www.eecs.umich.edu/~fessler/papers/files/tr/04,jacobson.pdf`.

H. Kaufmann. On Existence and Uniqueness of a Vector Minimizing a Convex Function. *Methods and Models of Operations Research*, 32:357–373, 1988.

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

E.T. Lee. A Computer Program for Linear Logistic Regression Analysis. *Computer Programs in Biomedicine*, 4:80–92, 1974.

A.E. Maxwell. *Analysing Qualitative Data*. Chapman & Hall, London, GB, 1961.

T.S. Motzkin. New Techniques for Linear Inequalities and Optimization. In *Project SCOOP Symposium on Linear Inequalities and Programming*, Washington, D.C., 1952. Planning Research Division, U.S. Air Force.

T.J. Sandler and D.E. Duffy. A Note on A. Albert and J.A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 73:755–758, 1986.

G.W. Stewart. An Iterative Method for Solving Linear Inequalities. Technical Report TR-1833, Department of Computer Science, University of Maryland, 1987.

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

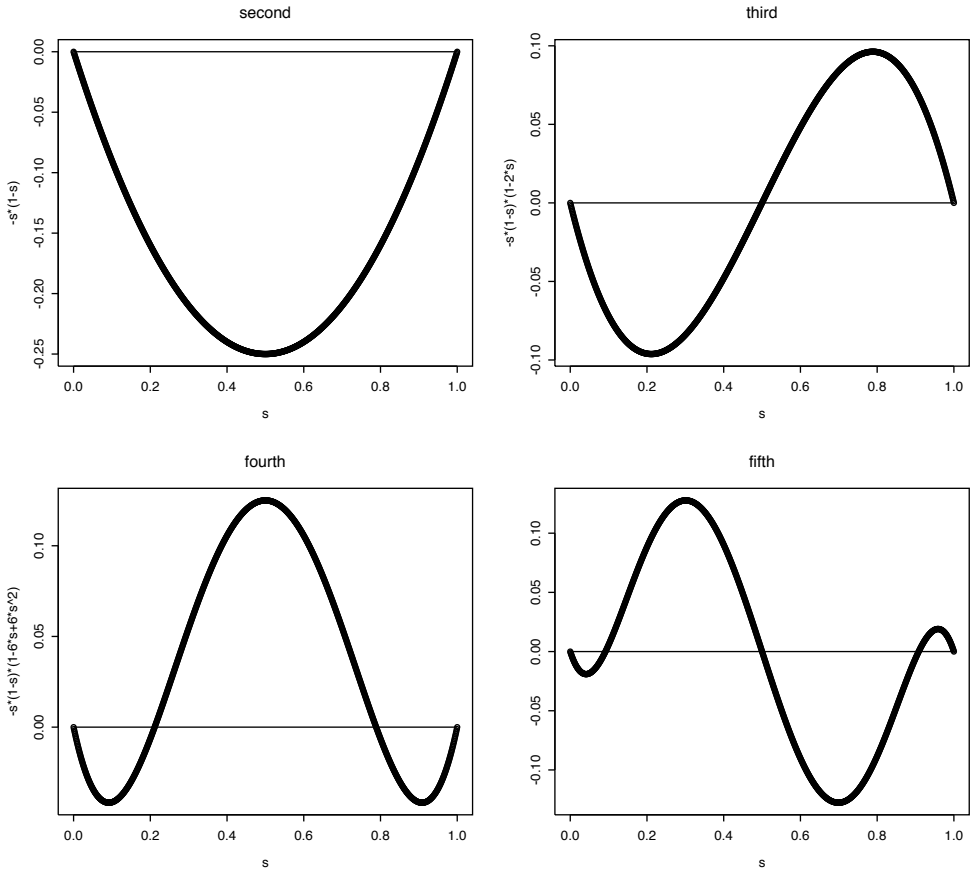
Department of Statistics, University of California, Los Angeles, CA 90095-1554

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

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

List of Figures

FIGURE 2. Derivatives of the log-logistic



List of Tables

TABLE 1. Boys' Ratings on a Lie Scale

age	n	N-n	N
5-7	6	15	21
8-9	18	31	49
10-11	19	31	50
12-13	27	32	59
14-15	25	19	44

TABLE 2. Cancer Remission Data

	A	B	C	D	E	F	
1	0.80	0.83	0.66	1.9	1.100	0.996	1
1	0.90	0.36	0.32	1.4	0.740	0.992	1
1	0.80	0.88	0.70	0.8	0.176	0.982	0
1	1.00	0.87	0.87	0.7	1.053	0.986	0
1	0.90	0.75	0.68	1.3	0.519	0.980	1
1	1.00	0.65	0.65	0.6	0.519	0.982	0
1	0.95	0.97	0.92	1.0	1.230	0.992	1
1	0.95	0.87	0.83	1.9	1.354	1.020	0
1	1.00	0.45	0.45	0.8	0.322	0.999	0
1	0.95	0.36	0.34	0.5	0.000	1.038	0
1	0.85	0.39	0.33	0.7	0.279	0.988	0
1	0.70	0.76	0.53	1.2	0.146	0.982	0
1	0.80	0.46	0.37	0.4	0.380	1.006	0
1	0.20	0.39	0.08	0.8	0.114	0.990	0
1	1.00	0.90	0.90	1.1	1.037	0.990	0
1	1.00	0.84	0.84	1.9	2.064	1.020	1
1	0.65	0.42	0.27	0.5	0.114	1.014	0
1	1.00	0.75	0.75	1.0	1.322	1.004	0
1	0.50	0.44	0.22	0.6	0.114	0.990	0
1	1.00	0.63	0.63	1.1	1.072	0.986	1
1	1.00	0.33	0.33	0.4	0.176	1.010	0
1	0.90	0.93	0.84	0.6	1.591	1.020	0
1	1.00	0.58	0.58	1.0	0.531	1.002	1
1	0.95	0.32	0.30	1.6	0.886	0.988	0
1	1.00	0.60	0.60	1.7	0.964	0.990	1
1	1.00	0.69	0.69	0.9	0.398	0.986	1
1	1.00	0.73	0.73	0.7	0.398	0.986	0

	9	8	7	6	5	4	3	2	1
1	0	0	0	0	0	0	0	0	-1
2	0	0	0	0	0	0	0	1	-1
3	0	0	0	0	0	0	-2	3	-1
4	0	0	0	0	0	6	-12	7	-1
5	0	0	0	0	-24	60	-50	15	-1
6	0	0	0	120	-360	390	-180	31	-1
7	0	0	-720	2520	-3360	2100	-602	63	-1
8	0	5040	-20160	31920	-25200	10206	-1932	127	-1
9	-40320	181440	-332640	317520	-166824	46620	-6050	255	-1