# Computing Estimates in the Proportional Odds Model

David R. Hunter[1]
Kenneth Lange[2]

Department of Statistics[1]
Penn State University
University Park, PA 16802-2111

Departments of Biomathematics and Human Genetics[2]
UCLA School of Medicine
Los Angeles, CA 90024

email: dhunter@stat.psu.edu[1]
phone: (814) 863-0979 [1]
fax: (814) 863-7114 [1]

Research supported in part by USPHS
grants GM53275[2] and MH59490[2].

Proposed running head: Computing MLE for proportional odds

Submitted to Annals of the Institute of Statistical Mathematics
September 7, 1999
Resubmitted June 1, 2000

Resubmitted October 23, 2000

**Abstract**

The semiparametric proportional odds model for survival data is useful when mortality rates of different groups converge over time. However, fitting the model by maximum likelihood proves computationally cumbersome for large datasets because the number of parameters exceeds the number of uncensored

observations. We present here an alternative to the standard Newton-Raphson method of maximum likelihood estimation. Our algorithm, an example of a minorization-maximization (MM) algorithm, is guaranteed to converge to the maximum likelihood estimate whenever it exists. For large problems, both the algorithm and its quasi-Newton accelerated counterpart outperform Newton-Raphson by more than two orders of magnitude.

**Key Words:** majorization, proportional odds, Newton-Raphson, quasi-Newton, survival analysis.

# 1 Introduction

In a survival analysis setting with right-censored data, we observe $n$ iid copies of the paired random variables $(\min\{T, C\}, \delta)$, where $T$ is the time until the occurrence of some event of interest, $C$ is a censoring time, and $\delta$ is the censoring indicator $1_{\{T<C\}}$. The proportional odds model (Bennett, 1983) may be used to model survival data in which mortality rates for separate groups of patients converge over time. For example, if we have two groups of patients, say a treatment group ($i = 1$) and a control group ($i = 0$), then we may postulate that the odds ratio

$$r = \frac{F_1(t)[1 - F_0(t)]}{F_0(t)[1 - F_1(t)]} \tag{1.1}$$

remains constant over time. In expression (1.1), $F_i(t)$ denotes the cumulative distribution function of $T$ for group $i$. Following the development of Bennett (1983), we may extend the model to the case in which we measure several continuous covariates on each patient. Letting $z_i$ denote a $p$-vector of measured covariates on patient $i$ and $\beta$ a $p$-vector of parameters, we define $r_i = \exp(z_i^t \beta)$, and the model becomes

$$\frac{F(t; r_i)}{1 - F(t; r_i)} = \frac{F_0(t)}{1 - F_0(t)} r_i, \tag{1.2}$$

where $F_0(t)$ is some baseline distribution function and $F(t, r)$ is a family of distribution functions indexed by the proportionality constant $r$.

Letting $H(t) = F_0(t)/[1 - F_0(t)]$ denote the baseline odds of failure by time $t$, we may rearrange model (1.2) to obtain the survivor function

$$1 - F(t; r_i) = \frac{\exp[-z_i^t \beta]}{H(t) + \exp[-z_i^t \beta]} \tag{1.3}$$

and ultimately the hazard function

$$-\frac{d}{dt} \ln [1 - F(t; r_i)] .$$

Under the proportional odds model, the ratio of hazard functions for individuals $i$ and $j$ converges monotonically to 1 as $t$ increases. As Bennett (1983) suggests, this model may be particularly useful in modeling an effective cure, where the mortality of a control population approaches that of a treatment population.

In contrast to the Cox proportional hazards model (Cox, 1972), the estimation of $\beta$ via maximum likelihood is complicated by the fact that the parameter $F_0(t)$ must be estimated along with $\beta$. As summarized by Murphy *et al.* (1997), several authors have proposed estimators for $\beta$ in the proportional odds model. For example, Cheng *et al.* (1995) noted the equivalence of the model to a linear transformation model and introduced a procedure based on estimating equations. The current article treats a method shown by Murphy *et al.* (1997) to produce a consistent and asymptotically normal estimator. When $\beta$ lies in a compact set $\mathcal{B} \subset R^p$, they show that the maximum likelihood estimator of $F_0(t)$ is a discrete distribution supported on the set of uncensored survival times. We describe the likelihood function and comment on the necessity of assuming that $\beta \in \mathcal{B}$ in Section 3.

The main purpose of this article is to present an algorithm for computing the maximum likelihood estimate $[\hat{\beta}, \hat{H}(t)]$ under the natural constraints that $H(0) = 0$ and $H(t)$ is constant except at uncensored survival times $t$, where it has positive jumps. This algorithm is a special case of the MM (minorization-maximization) algorithm explained in Section 2. Section 3 records the likelihood function to be

maximized and describes a reparameterization which renders it strictly concave. Section 4 derives the MM algorithm that maximizes this likelihood. Section 5 suggests a quasi-Newton acceleration of the MM algorithm. Section 6 summarizes numerical tests on large problems that show the computational superiority of MM and accelerated versions over the Newton-Raphson method. Proofs of all propositions appear in the appendix.

## 2 MM Algorithms

Described as early as 1970 by Ortega and Rheinboldt (1970, p. 253), the MM algorithm has surfaced from time to time in the statistical literature under names such as iterative majorization (Heiser 1995), EM algorithms without missing data (Becker *et al.* 1997), and optimization transfer (Lange *et al.* 2000). The initials MM are suggested in a rejoinder by Hunter and Lange (2000) to emphasize the relationship of the technique to the ubiquitous EM algorithm, which is a specific example of MM. The letter pair MM is intentionally ambiguous; it can also stand for majorization-minimization in problems for which minimization, rather than maximization, of an objective function is the goal. Theory and reviews of past work on these algorithms may be found in de Leeuw (1994), Heiser (1995), Becker *et al.* (1997), and Lange *et al.* (2000).

In essence, an MM algorithm replaces a difficult optimization problem by a sequence of easier optimization problems. In most cases, the solutions of the substitute problems converge to a solution of the original problem. Suppose we want to maximize the continuous function $L(\theta) : R^p \to R$. If $\theta^k$ denotes the current iterate in finding the maximum point, then an MM algorithm proceeds in two steps. First, we concoct surrogate functions $Q(\theta \mid \theta^k)$ that minorize $L(\theta)$ in the sense that

$$Q(\theta^k \mid \theta^k) = L(\theta^k) \tag{2.1}$$

$$Q(\theta \mid \theta^k) \quad \leq \quad L(\theta) \tag{2.2}$$

for all $\theta$ and $\theta^k$. This slightly odd functional notation is adopted from the EM algorithm literature (Dempster *et al.* 1977) and emphasizes the role of $\theta^k$ as indexing a set of real-valued functions on $R^p$, each bounded above by $L(\theta)$. The index, $\theta^k$, is the point of tangency between $Q(\theta \mid \theta^k)$ and $L(\theta)$. It is a challenge to construct useful surrogate functions. This is exactly what is accomplished by the E step of a well-conceived EM algorithm. More generally, MM algorithms tend to rely on surrogate functions constructed from convexity arguments.

The second step in an MM iteration is to maximize $Q(\theta \mid \theta^k)$ with respect to $\theta$. We take the next iterate $\theta^{k+1}$ to maximize $Q(\theta \mid \theta^k)$ or in difficult problems to simply increase $Q(\theta \mid \theta^k)$. The inequality

$$Q(\theta^{k+1} \mid \theta^k) \quad \geq \quad Q(\theta^k \mid \theta^k) \tag{2.3}$$

together with conditions (2.1) and (2.2) gives the ascent property

$$L(\theta^{k+1}) \quad \geq \quad L(\theta^k). \tag{2.4}$$

The ascent property (2.4) makes MM algorithms very attractive computationally because unlike many other numerical methods, each iteration of an MM algorithm drives the value of the objective function in the desired direction. However, it is clearly possible in many problems to construct a sequence $\{\theta^k\}$ satisfying inequality (2.4) that does not converge at all. Thus, it is natural to ask whether an MM algorithm is guaranteed to converge and, if so, whether it converges to a local maximum. Without some additional hypotheses, the answers are no, as standard counterexamples for the EM algorithm demonstrate (McLachlan and Krishnan, 1997). In the case of proportional odds, however, we show in Section 3 that it is possible to reparameterize the problem so that the loglikelihood is strictly concave. The following proposition, whose proof is contained in the appendix, is relevant to the convergence of our proportional odds algorithm.

**Proposition 1** *Suppose the surrogate function $Q : R^p \times R^p \to R$ satisfies conditions (2.1) and (2.2), and the MM map function $T(\theta) : \theta^k \mapsto \theta^{k+1}$ satisfies inequality (2.3). If (a) $L(\theta)$ is upper compact, in the sense that $\{\theta \in \Theta : L(\theta) \geq c\}$ is compact for any constant $c$; (b) $L(\theta)$ is strictly concave; (c) $T(\theta)$ is continuous; (d) fixed points of $T(\theta)$ coincide with stationary points of $L(\theta)$; and (e) $L[T(\theta)] = L(\theta)$ only if $T(\theta) = \theta$, then $\lim_k \theta^k = \theta^*$, the unique maximizer of $L(\theta)$.*

We discuss the implications of Proposition 1 in Section 4.

# 3 Reparameterizing the Proportional Odds Model

Murphy *et al.* (1997) write the likelihood for the proportional odds model as

$$\prod_{i=1}^{n} \left( \frac{e^{-z_i^t \beta}}{H(Y_i) + e^{-z_i^t \beta}} \right) \left( \frac{\Delta H(Y_i)}{H(Y_i-) + e^{-z_i^t \beta}} \right)^{\delta_i}, \tag{3.1}$$

where the data on the $i$th individual consist of a survival time $Y_i$, a censoring indicator $\delta_i$, and a $p$-vector $z_i$ of covariates for some $p < n$. For the sake of convenience, we assume without loss of generality that the observations are ordered so that $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ and, in the case of ties, uncensored observations come first—that is, $Y_i = Y_j$ and $i < j$ together imply $\delta_i \geq \delta_j$. The baseline odds function $H(y)$ is a right-continuous step function with jumps only at uncensored survival times. For an uncensored observation $Y_i$, $\Delta H(Y_i)$ denotes the jump in the value of $H(y)$ at $Y_i$ and $H(Y_i-)$ denotes the left-hand limit of $H(y)$ at $Y_i$; thus, $H(Y_i) = H(Y_i-) + \Delta H(Y_i)$ for uncensored $Y_i$.

If $Y_n$ is uncensored, then the ratio $\Delta H(Y_n)/(H(Y_n) + e^{-z_n^t \beta})$ corresponding to this largest observation can be made arbitrarily close to 1 by letting $\Delta H(Y_n)$ grow without bound. In the limit as $\Delta H(Y_n)$ tends to $\infty$, the last term in the likelihood (3.1) becomes

$$\frac{e^{-z_n^t \beta}}{H(Y_n-) + e^{-z_n^t \beta}}.$$

In other words, if $\delta_n = 1$ we treat $Y_n$ and any observations tied with $Y_n$ as if they are censored data points in all that follows, thus avoiding a situation in which the estimation algorithm diverges. Unless there are ties, this simply amounts to setting $\delta_n = 0$. We note also that if $Y_1$ is censored, then $H(Y_1) = 0$ and the first term in the likelihood is simply 1. Thus, any censored observations occurring prior to the first uncensored observation do not contribute to the likelihood at all, so they are dropped. As a result, we assume in all that follows that $\delta_1 = 1$.

The parameters in this model are the $p$-vector $\beta$ and the jumps $\Delta H(Y_i)$ of the nondecreasing, right-continuous step function $H$. Let $m$ be the number of jumps—i.e., the number of unique, uncensored survival times. Letting $U_1 < U_2 < \cdots < U_m$ denote these unique uncensored survival times, we define new parameters $\gamma_j = \ln[\Delta H(U_j)]$ for $1 \leq j \leq m$. For a given observation $Y_i$, the value of the baseline odds function at that point, $H(Y_i)$, may be obtained by summing the jumps $\Delta H(U_j)$ for all $U_j \leq Y_i$. To aid notation, we let $w_i$ be the index of the largest jump point less than or equal to $Y_i$; that is, $w_i = \max\{j : U_j \leq Y_i\}$. With this notation, we may write $H(Y_i) = \sum_{j=1}^{w_i} \exp(\gamma_j)$ for each $i$; and if $\delta_i = 1$, then $\Delta H(Y_i) = \exp(\gamma_{w_i})$ and $H(Y_i-) = \sum_{j=1}^{w_i-1} \exp(\gamma_j)$ (or $H(Y_i-) = 0$ if $w_i = 1$). Thus, if $\theta \in R^{p+m}$ is defined to be the parameter vector $(\beta^t, \gamma^t)^t$, the likelihood (3.1) may be expressed as

$$\prod_{i=1}^{n} \left( \frac{e^{-z_i^t \beta}}{D_i(\theta)} \right) \left( \frac{e^{\gamma_{w_i}}}{D_i(\theta) - e^{\gamma_{w_i}}} \right)^{\delta_i}, \tag{3.2}$$

where

$$D_i(\theta) \;=\; e^{-z_i^t \beta} + \sum_{j=1}^{w_i} e^{\gamma_j}. \tag{3.3}$$

For mathematical convenience, we turn our attention to maximizing the loglikelihood

$$L(\theta) \;=\; \sum_{i=1}^{n} -z_i^t \beta - \ln D_i(\theta) + \delta_i \left\{ \gamma_{w_i} - \ln[D_i(\theta) - e^{\gamma_{w_i}}] \right\}. \tag{3.4}$$

The nonlinear terms above, $\ln D_i(\theta)$ and $\ln[D_i(\theta) - e^{\gamma w_i}]$, are convex because a sum of logconvex functions is logconvex (Seneta 1973). Thus, the loglikelihood is concave after reparameterization. In view of Proposition 1, the real question of interest is whether this concavity is strict. Proposition 2 gives a necessary and sufficient condition for strict concavity to fail. The $n \times p$ parameter matrix $Z$, whose rows consist of the row vectors $z_i^t$, is assumed to be of full rank.

**Proposition 2** *The concave loglikelihood $L(\theta)$ fails to be strictly concave if and only if the vectors $z_1, z_2, \ldots, z_n$ all lie on a $p$-dimensional hyperplane; that is, strict concavity fails if and only if there exists some $p$-vector $v$ such that $Zv = \mathbf{1}$, where $Z$ is the $n \times p$ full-rank covariate matrix and $\mathbf{1}$ denotes the $n$-vector whose entries are all equal to one.*

Note that randomly perturbing a single element of $Z$ destroys the equality $Zv = \mathbf{1}$ in the unusual circumstance that it holds. In light of Proposition 2, we therefore assume that the reparameterized loglikelihood (3.4) is strictly concave in what follows.

Strict concavity rules out multiple maxima, but it does not guarantee the existence of a maximum. In fact, it is possible that no MLE exists, as the following proposition demonstrates.

**Proposition 3** $\sup_{\theta \in R^{m+p}} L(\theta) = 0$ *if and only if there exists $\beta^* \in R^p$ such that $z_i^t \beta^* > z_k^t \beta^*$ whenever $\delta_i = \delta_k = 1$ and $Y_i < Y_k$ or whenever $\delta_i = 1 - \delta_k = 1$ and $Y_i \leq Y_k$. In particular, no maximizer of the likelihood exists if there is a $\beta^* \in R^p$ such that $z_1^t \beta^* > z_2^t \beta^* > \cdots > z_n^t \beta^*$.*

By inspection, the likelihood (3.2) must be strictly less than one; thus, the statement that $\sup_\theta L(\theta) = 0$ implies that no maximizer of the likelihood exists in $R^{m+p}$. Proposition 3 is most useful as a demonstration that there are arbitrarily large datasets for which no maximizer of the likelihood exists. The second half of Proposition 3 makes it clear that constructing such artificial datasets is typically

a simple matter of rearranging the rows of the covariate matrix $Z$. As a practical matter, however, such datasets are extremely rare. For instance, in the case $p = 1$, a maximizer typically exists unless the covariates $z_1, z_2, \ldots, z_n$ form a strictly increasing or decreasing sequence. In real examples, convergence of the MM algorithm is a reliable test of the existence of a maximizer.

# 4 MM for Proportional Odds

The most difficult task in producing any MM algorithm is the creation of a good surrogate function. To this end, we exploit the inequality

$$\ln b \quad \leq \quad \ln a + b/a - 1 \tag{4.1}$$

for positive $a$ and $b$. Inequality (4.1) simply says that the tangent line at the point $a$ majorizes the concave function $\ln b$. Applying this inequality to equation (3.4) and defining a term $C_i(\theta^k) = 1 + \delta_i - \ln D_i(\theta^k) - \delta_i \ln[D_i(\theta^k) - e^{\gamma_{w_i}}]$ that does not depend on $\theta$ gives the surrogate function

$$Q(\theta \mid \theta^k) \quad = \quad \sum_{i=1}^{n} \left[ -z_i^t \beta + \delta_i \gamma_{w_i} - \frac{D_i(\theta)}{D_i(\theta^k)} - \frac{\delta_i [D_i(\theta) - e^{\gamma_{w_i}}]}{[D_i(\theta^k) - e^{\gamma_{w_i}^k}]} + C_i(\theta^k) \right] \tag{4.2}$$

minorizing $L(\theta)$ at $\theta^k$. Ignoring the irrelevant constant $C_i(\theta^k)$, note that $Q(\theta \mid \theta^k)$ partially separates the parameters in the sense that

$$Q(\theta \mid \theta^k) \quad = \quad \sum_{i=1}^{n} f_i(\beta \mid \theta^k) + \sum_{j=1}^{m} g_j(\gamma_j \mid \theta^k) \tag{4.3}$$

for appropriately defined functions

$$f_i(\beta \mid \theta^k) \quad = \quad -z_i^t \beta - e^{-z_i^t \beta} \left[ \frac{1}{D_i(\theta^k)} + \frac{\delta_i}{D_i(\theta^k) - e^{\gamma_{w_i}^k}} \right] \tag{4.4}$$

and

$$g_j(\gamma_j \mid \theta^k) \quad = \quad u_j \gamma_j - e^{\gamma_j} \left[ \sum_{i:w_i \geq j} \frac{1}{D_i(\theta^k)} + \sum_{i:w_i > j} \frac{\delta_i}{D_i(\theta^k) - e^{\gamma_{w_i}^k}} \right]. \tag{4.5}$$

In equation (4.5), $u_j = \#\{i : \delta_i = 1, w_i = j\}$ is the number of uncensored observations at the $j$th jump point. Thus, maximization of $Q(\theta \mid \theta^k)$ with respect to $\theta$ may be accomplished one parameter entry at a time except for the $\beta$ vector, which must be tackled all at once. Since the dimension $p$ of $\beta$ is generally much smaller than $p + m$, this is a considerable simplification.

To increase the value of $f(\beta \mid \theta^k) = \sum_{i=1}^n f_i(\beta \mid \theta^k)$ with respect to $\beta$, we use a Newton-Raphson approach. This requires the first differential

$$df(\beta \mid \theta^k) = -\sum_{i=1}^n \left[1 - \frac{e^{-z_i^t\beta}}{D_i(\theta^k)} - \frac{\delta_i e^{-z_i^t\beta}}{D_i(\theta^k) - e^{\gamma_{w_i}^k}}\right] z_i^t \tag{4.6}$$

and second differential

$$d^2 f(\beta \mid \theta^k) = -\sum_{i=1}^n z_i z_i^t e^{-z_i^t\beta} \left[\frac{1}{D_i(\theta^k)} + \frac{\delta_i}{D_i(\theta^k) - e^{\gamma_{w_i}^k}}\right] \tag{4.7}$$

of $f(\beta \mid \theta^k)$ with respect to its first argument, where we abuse notation slightly by identifying differentials of a function with the appropriate matrices of partial derivatives; i.e., the first differential is the transpose of the gradient and the second differential is the Hessian matrix. Fortunately, $d^2 f(\beta \mid \theta^k)$ is negative definite provided $Z$ has full rank. Therefore, we can guarantee an increase in the value of $f(\beta \mid \theta^k)$ by taking a sufficiently small positive step in the Newton direction

$$\Delta\beta^k = -d^2 f(\beta^k \mid \theta^k)^{-1} df(\beta^k \mid \theta^k)^t. \tag{4.8}$$

To this end, we define the constant

$$\alpha^k = \arg\max_{\alpha \in (0,1]} f(\beta^k + \alpha\Delta\beta^k \mid \theta^k) \tag{4.9}$$

and the next iterate

$$\beta^{k+1} = \beta^k + \alpha^k \Delta\beta^k. \tag{4.10}$$

Maximizing the sum $\sum_{j=1}^m g_j(\gamma_j \mid \theta^k)$ with respect to $\gamma$ simply involves setting

$g_j'(\gamma_j \mid \theta^k) = 0$ and solving for $\gamma_j$ for each $j$. This gives the closed-form solution

$$\gamma_j^{k+1} = \ln u_j - \ln \left[ \sum_{i:w_i \geq j} \frac{1}{D_i(\theta^k)} + \sum_{i:w_i > j} \frac{\delta_i}{D_i(\theta^k) - e^{\gamma_{w_i}^k}} \right] \tag{4.11}$$

for each $1 \leq j \leq m$. Note that for computing the profile likelihood

$$\text{Prlik}(\hat{\beta}) \quad = \quad \sup_{\gamma} L\left[\left(\hat{\beta}^t, \gamma^t\right)^t\right],$$

only equation (4.11) is needed; the matrix inversion and line search of equations (4.8) and (4.9) become unnecessary. The profile likelihood may be used, for example, to estimate standard errors consistently (Murphy *et al.* 1997).

Proposition 4, whose proof consists of checking the hypotheses of Proposition 1, shows that under mild assumptions, the MM map $T : \theta^k \mapsto \theta^{k+1}$ given by equations (4.8) through (4.11) yields a sequence which converges to the unique maximum likelihood estimate.

**Proposition 4** *Assuming that $Z$ has full rank and that $L(\theta)$ is strictly concave and upper compact, the sequence $\{\theta^k\}$ of iterates generated by the MM algorithm in equations (4.8) through (4.11) converges to the unique maximizer of $L(\theta)$.*

One may avoid the line search implied by equation (4.9) by employing a much simpler version of the search such as step-halving—that is, defining for $\Delta\beta^k \neq 0$

$$\alpha^k \quad = \quad \max\left\{\alpha = 2^{-\nu} : f(\beta^k + \alpha\Delta\beta^k \mid \theta^k) > f(\beta^k \mid \theta^k), \nu = 0, 1, 2, \ldots\right\}. \tag{4.12}$$

Such an approach tends to save a great deal of computational time; however, it destroys the continuity of the MM map, which is a vital element of the proof of Proposition 4. We show below that despite this problem, the algorithm must converge to the correct vector if it converges at all.

**Proposition 5** *If the hypotheses of Proposition 4 hold except that the sequence $\{\theta^k\}$ is generated by the algorithm employing definition (4.12) in place of definition (4.9), then $\lim_{k\to\infty} \theta^k = \theta^*$ implies that $\theta^*$ is the maximizer of $L(\theta)$.*

In practice, we have never seen an example in which the MM algorithm based on equation (4.12) fails to converge when a unique MLE exists (in fact, we are not certain whether it is even possible to construct such an example). We therefore recommend the step-halving version of the MM algorithm despite the small gap in its convergence proof. In the worst case, failure to converge in a reasonable amount of time should prompt the user to examine the sequence of iterates; if they appear to be heading toward $\infty$, probably no MLE exists, whereas if they appear to oscillate, the line search of equation (4.9) should be used. Our numerical tests of Section 6 use equation (4.12), though the MATLAB code we make available has the capability of using either step-halving or a line search (using the MATLAB function `fminbnd`) to compute $\alpha^k$.

## 5 Accelerated MM

The strengths of the Newton-Raphson method and MM are complementary. On one hand, Newton-Raphson converges at a fast quadratic rate of convergence near an optimum point. On the other hand, MM avoids large matrix inversions and consequently uses many fewer computations per iteration. MM is also guaranteed to increase the likelihood at each iteration, particularly far from an optimum, where Newton-Raphson can be erratic. One can accelerate MM by building over successive iterations better and better approximations to the inverse Hessian and using them to approximate Newton-Raphson. The resulting quasi-Newton algorithm starts as MM and ends as an approximate Newton-Raphson method without ever inverting a large matrix (Jamshidian and Jennrich 1997; Lange *et al.* 2000).

To implement quasi-Newton acceleration, view the MM map $\theta^{k+1} = T(\theta^k)$ as a gradient method of the form $T(\theta^k) = \theta^k - A_k dL(\theta^k)^t$. Next, approximate the difference $M_k \approx d^2L(\theta^k)^{-1} - A_k$ and write the quasi-Newton iteration as

$$\theta^{k+1} = \theta^k - (A_k + M_k)dL(\theta^k)^t. \tag{5.1}$$

The Taylor approximation

$$d^2L(\theta^k)^{-1}s_k \;\approx\; \theta^k - \theta^{k-1}$$

with $s_k^t = dL(\theta^k) - dL(\theta^{k-1})$ yields the inverse secant condition

$$M_k s_k \;=\; \theta^k - \theta^{k-1} - A_k s_k \;\equiv\; r_k \tag{5.2}$$

for updating $M_k$. Davidon's (1959) parsimonious symmetric, rank-one update of $M_k$ satisfying condition (5.2) is

$$M_k \;=\; M_{k-1} + q_k q_k^t / c_k, \tag{5.3}$$

where $q_k = r_k - M_{k-1}s_k$ and $c_k = q_k^t s_k$. More complicated updates are possible, but recent experience suggests that these are no better than Davidon's rank one update (Conn *et al.* 1991; Khalfan *et al.* 1993). In order to start the algorithm, some choice of $M_0$ is required, and we set $M_0 = \mathbf{0}$.

Because the MM algorithm is an ascent algorithm,

$$L(\theta^k) \;\leq\; L[\theta^k - A_k dL(\theta^k)^t] \;=\; L[\theta_{\mathrm{MM}}^{k+1}],$$

where $\theta_{\mathrm{MM}}^{k+1}$ is the MM version of the updated $\theta^k$. However, there is no guarantee that defining $\theta^{k+1}$ via equation (5.1) will lead to an increase in the likelihood. If not, we satisfy ourselves with the MM update. Thus, the accelerated version of MM is

$$\theta^{k+1} \;=\; \arg\max \left\{ L\left[\theta_{\mathrm{MM}}^{k+1} - M_k dL(\theta^k)^t\right], L\left[\theta_{\mathrm{MM}}^{k+1}\right] \right\}. \tag{5.4}$$

It is possible to accelerate the MM algorithm without explicitly computing the $A_k$ matrix. Using the approximation $A_k \approx A_{k-1}$, we may express the inverse secant condition (5.2) in terms of MM increments $\Delta\theta^k = -A_k dL(\theta^k)^t$ as suggested by Jamshidian and Jennrich (1997). Thus, condition (5.2) is replaced by

$$M_k s_k \;=\; \theta^k - \theta^{k-1} - \Delta\theta^k + \Delta\theta^{k-1} \;\equiv\; r_k. \tag{5.5}$$

We then carry out the updates (5.3) and (5.4) with this change. In preparing the numerical tests of the next section, we tested both of these versions of quasi-Newton acceleration. We find that secant condition (5.5) generally yields a slightly faster algorithm than secant condition (5.2); thus, Section 6 discusses only the former, which enjoys the additional advantage that computation of the matrix $A_k$ is unnecessary.

In implementing the quasi-Newton algorithm (5.4), it is wise to avoid computing $M_k dL(\theta^k)^t$ and $M_{k-1}s_k$ by left-multiplying column vectors by the large square matrix $M_k$. Instead, we write for example

$$M_k dL(\theta^k)^t \quad = \quad M_0 dL(\theta^k)^t + \sum_{j=1}^{k} c_j [dL(\theta^k) q_j]\, q_j.$$

For the typical choice $M_0 = \mathbf{0}$, these maneuvers do not involve the storage or multiplication of any $(p+m) \times (p+m)$ matrices. Although the program must store the constants $c_k$ and the vectors $q_k$ for all iterations, this tactic yields savings in storage and computation whenever the number of iterations is less than $p+m$, which, as Table 1 suggests, is true for every problem we tested.

Because each iteration of the accelerated algorithm (5.4) is at least as good as the unaccelerated MM algorithm by design, it is easy to show that it enjoys the same convergence properties as the MM algorithm. In other words, if algorithm (5.4) converges, then it converges to the MLE; and if the MM map is continuous, as it is when a line search is used to determine $\alpha^k$, algorithm (5.4) is guaranteed to converge when an MLE exists. As we do for the unaccelerated version of MM, we test an accelerated MM algorithm in Section 6 that uses the step-halving of equation (4.12) instead of the line search of equation (4.9) to compute $\alpha^k$.

# 6 Numerical Results

We present in Figure 1 a graphical summary of many numerical tests of the MM algorithm and its accelerated version as coded in MATLAB. The floating point

operations, or FLOPs, reported by MATLAB provide a fair basis for comparing algorithms. FLOPs are better than elapsed time, which varies widely depending on the compiler and machine chosen, and total iterations, which fail to take into account the fact that each MM iteration is very fast. We report iteration counts in Table 1 in case the reader is interested. All of the MATLAB code used for these tests is available online at `http://www.stat.psu.edu/~dhunter/matlab`.

As a basis for comparison, we also test a Newton-Raphson method. Hessian-based methods such as Newton-Raphson are currently the most common approach to estimating the proportional odds parameters; Murphy *et al.* (1997), for example, use the IMSL minimization routine UMIAH, which employs a modification of the Newton-Raphson algorithm. Although it is not possible to determine exactly how a commercially sold routine like UMIAH optimizes an objective function, the algorithm of Gay (1981) cited in the UMIAH documentation does in fact invert the Hessian matrix (or a perturbation thereof, if the Hessian is not negative definite) at each iteration. Because our reparameterization (3.4) makes the Hessian negative definite, we therefore believe the simple Newton-Raphson method we test operates on the same principle as the algorithm used by Murphy *et al.* Regardless of the specifics of the Hessian-based algorithm that might be employed, even the task of computing the Hessian matrix in this problem is enormously complicated and therefore the MM algorithm we propose is much simpler to code than any Hessian-based algorithm.

For each test problem, we simulate data using $p = 4$ and $\beta = (1, 1, 1, 1)^t$. The function $H(t)$ is the identity. The $n \times p$ covariate matrix $Z$ has independent uniform (0,1) entries. The survival times $Y_i$ are generated by the inversion method (Ripley 1987) employing equation (1.3). We censor at a rate of 10% using two different methods: independent censoring and dependent censoring. Independent censoring ignores the value of $Z$ for a given individual and censors an observation whenever

it is larger than the 90th percentile of the empirical distribution of survival times. Dependent censoring censors any observation larger than the 90th percentile given its value of $Z$.

| $n$ | Newton-Raphson | Unaccelerated MM | Accelerated MM |
|---|---|---|---|
| 50 | 8.0 | 1398.5 | 20.0 |
| 100 | 8.0 | 1429.5 | 21.0 |
| 200 | 8.0 | 1433.0 | 21.5 |
| 300 | 9.5 | 1505.5 | 21.5 |
| 400 | 9.0 | 1440.5 | 21.0 |
| 500 | 9.5 | 1517.0 | 22.5 |
| 600 | 9.0 | 1469.0 | 22.0 |
| 700 | 10.0 | 1501.0 | 22.0 |
| 800 | 10.0 | 1526.5 | 23.0 |
| 900 | 10.0 | 1476.0 | 23.0 |
| 1000 | 10.0 | 1529.5 | 23.0 |
| 1200 | 9.5 | 1548.0 | 22.5 |

Table 1: Median number of iterations until convergence for some representative values of $n$ depicted in Figure 1.

Every maximization begins iterating from the zero vector $(\beta^0, \gamma^0)^t = (\mathbf{0}, \mathbf{0})^t$. We safeguard the Newton-Raphson method tested here by step halving. Thus, if the step $a_k[d^2L(\theta^k)]^{-1}dL(\theta^k)^t$ with $a_k = -1$ produces a decrease in the likelihood, then we try $a_k = -1/2$, $a_k = -1/4$, and so forth until the likelihood increases. As noted earlier, the MM algorithms (accelerated and unaccelerated) also use step halving to determine $\alpha^k$, as in equation (4.12), instead of conducting the line search of equation (4.9).

Declaring convergence is difficult because MM algorithms tend to take very small steps as they approach the maximizer of the likelihood. Ideally, one might stop the algorithm when the relative change in the value of the loglikelihood is less than some small constant, but this approach would eventually declare convergence even in a case when no MLE exists and the algorithms diverge. We therefore declare

convergence for all algorithms when

$$\max\left\{\left|\frac{L(\theta^k)-L(\theta^{k-1})}{L(\theta^k)}\right|, \|\theta^k-\theta^{k-1}\|_2\right\} < 10^{-8}.$$

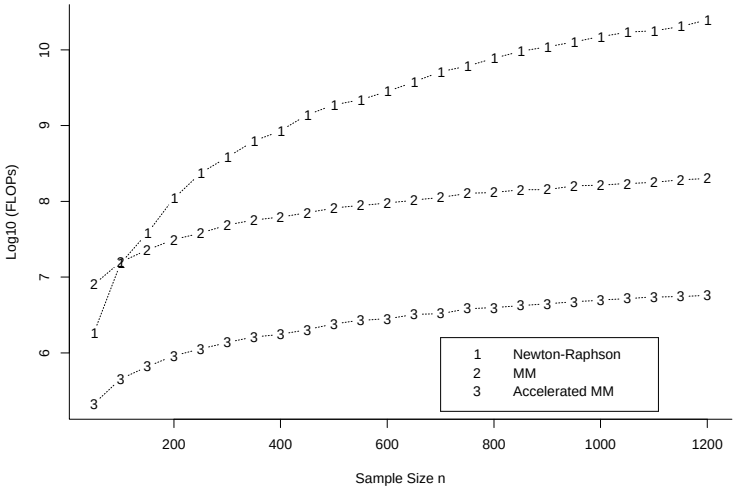


Figure 1: Floating point operations (FLOPs) required for examples with dependent censoring. Each point is the median of ten runs.

Figure 1 shows the clear advantage in speed of the MM algorithm and its accelerated counterpart over Newton-Raphson. It is noteworthy that MATLAB takes roughly $1.5 \times 10^9$ FLOPs to invert a single $900 \times 900$ matrix, the approximate size of the Hessian matrix when $n = 1000$ under 10% censoring. Exploiting the symmetry of the Hessian, we might reduce this number by half. For problems this large, the graph shows that MM converges using fewer operations than inverting the Hessian even once. To its detriment, Newton-Raphson must invert the Hessian repeatedly. Thus, although the particular implementation of the Newton-Raphson idea we use may not be the same as implementations used by others, it is clear that any Newton-Raphson method requires an enormous amount of computation just to perform the necessary large matrix inversion at each iteration.

Results of tests of the routines on independent censoring problems are nearly identical to those displayed in Figure 1. Thus, we include only the dependent censoring results here. We have observed results as striking as those in Figure 1 with other datasets we tested. Finally, we remark that the simplicity of the MM algorithm makes it relatively easy to code, particularly when compared with the complicated Hessian matrix calculations necessary for coding the Newton-Raphson method.

# 7 Appendix

**Proof of Proposition 1**. Upper compactness and the ascent property imply that there is some subsequence $\{\theta^{k_n}\}_{n\geq 1}$ of $\{\theta^k\}_{k\geq 1}$ with limit $\hat{\theta}$. Since $L(\theta^{k_{n+1}}) \geq L[T(\theta^{k_n})] \geq L(\theta^{k_n})$, the continuity of $L(\theta)$ and $T(\theta)$ yields in the limit $L(\hat{\theta}) \geq L[T(\hat{\theta})] \geq L(\hat{\theta})$, so that $L[T(\hat{\theta})] = L(\hat{\theta})$. In other words, $\hat{\theta}$ is a stationary point of $L(\theta)$. Strict concavity of $L(\theta)$ implies that there is at most one stationary point, namely the maximizer of $L(\theta)$. ■

**Proof of Proposition 2**. The proof uses a form of Hölder's inequality (Magnus and Neudecker 1988). For positive numbers $c_1, c_2, \ldots, c_N$ and $d_1, d_2, \ldots, d_N$ and $\alpha \in (0,1)$,

$$\ln \sum_{k=1}^{N} c_k^{\alpha} d_k^{1-\alpha} \leq \alpha \ln \sum_{k=1}^{N} c_k + (1-\alpha) \ln \sum_{k=1}^{N} d_k,$$

with equality if and only if there exists some $\lambda > 0$ such that $c_k = \lambda d_k$ for all $k$. Because of definition (3.3), Hölder's inequality establishes the concavity of $-\ln D_i(\theta)$ and $-\ln[D_i(\theta) - e^{\gamma w_i}]$ for all $i$. Strict concavity of just one of the $-\ln D_i(\theta)$ (or of $-\ln[D_i(\theta) - e^{\gamma w_i}]$ if $\delta_i = 1$) implies strict concavity of $L(\theta)$. Strict concavity of $-\ln D_i(\theta)$ fails if and only if there exist $\theta^1 \neq \theta^2$ and $\alpha \in (0,1)$ such that

$$D_i(\alpha\theta^1 + (1-\alpha)\theta^2) \quad = \quad \left[D_i(\theta^1)\right]^{\alpha} \left[D_i(\theta^2)\right]^{1-\alpha}.$$

Therefore, Hölder's inequality implies that $L(\theta)$ fails to be strictly concave if and only if there exist two distinct parameter vectors $\theta^1 = [(\beta^1)^t, (\gamma^1)^t]^t$ and $\theta^2 = [(\beta^2)^t, (\gamma^2)^t]^t$ and constants $\lambda_1, \lambda_2, \ldots, \lambda_n$ such that for all $i$, $\exp(-z_i^t\beta^1) = \lambda_i \exp(-z_i^t\beta^2)$ and $\exp(\gamma_j^1) = \lambda_i \exp(\gamma_j^2)$ for $j \leq w_i$. In particular, $\lambda_i = \exp(\gamma_1^1 - \gamma_1^2)$ for every $i$, so we may drop the subscripts on the $\lambda_i$ and refer to a single positive constant $\lambda$. Furthermore, $\lambda$ cannot equal 1; otherwise, we would have $\gamma^1 = \gamma^2$ and $z_i(\beta^1 - \beta^2) = 0$ for each $i$, which is a contradiction since $\theta^1 \neq \theta^2$ and $Z$ was assumed to have full rank. We conclude that strict concavity of $L(\theta)$ fails if and only if there exist $\theta^1$, $\theta^2$, and positive $\lambda \neq 1$ such that $z_i^t(\beta^2 - \beta^1)/\ln\lambda = 1$ for every $i$. Setting $v = (\beta^2 - \beta^1)/\ln\lambda$ completes the proof. ■

**Proof of Proposition 3**. Letting $\theta = (\beta^t, \gamma^t)^t$, we first set

$$\varphi_i(\theta) = \begin{cases} \left[1 + \sum_{j \leq w_i} e^{\gamma_j + z_i^t\beta}\right]^{-1} & \text{if } \delta_i = 0 \\ \left[1 + \sum_{j < w_i} e^{\gamma_j + z_i^t\beta}\right]^{-1} \left[1 + e^{-z_i^t\beta - \gamma_{w_i}} + \sum_{j < w_i} e^{\gamma_j - \gamma_{w_i}}\right]^{-1} & \text{if } \delta_i = 1. \end{cases}$$

Then the likelihood function (3.2) may be written $\prod_{i=1}^n \varphi_i(\theta)$. The supremum of the loglikelihood is zero if and only if all $\varphi_i(\theta)$ may be made simultaneously arbitrarily close to 1. In other words, $\sup_\theta L(\theta) = 0$ if and only if for any $\epsilon > 0$, there exists $\theta = (\beta^t, \gamma^t)^t$ such that

1. Whenever $\delta_i = 0$, $\exp(\gamma_j + z_i^t\beta) < \epsilon$ for all $j \leq w_i$.
2. Whenever $\delta_i = 1$,
   (a) $\exp(\gamma_j + z_i^t\beta) < \epsilon$ for all $j < w_i$,
   (b) $\exp(\gamma_j - \gamma_{w_i}) < \epsilon$ for all $j < w_i$, and
   (c) $\exp(-z_i^t\beta - \gamma_{w_i}) < \epsilon$.

Under these conditions, $\varphi_i(\theta)$ may be made arbitrarily close to 1 for each $i$.

Assume that $\sup_\theta L(\theta) = 0$. Fix positive $\epsilon < 1$. There exists $\theta^*$ satisfying 1 and 2 above. If $\delta_i = \delta_k = 1$ and $Y_i < Y_k$, then $w_i < w_k$; whence $-z_i^t\beta^* < \gamma_{w_i}^*$ by 2c

and $\gamma^*_{w_i} < -z_k^t\beta^*$ by 2a. We conclude that $z_i^t\beta^* > z_k^t\beta^*$ as required. Furthermore, if $\delta_i = 1$, $\delta_k = 0$ and $Y_i \leq Y_k$, then $w_i \leq w_k$; whence $-z_i^t\beta^* < \gamma^*_{w_i}$ by 2c and $\gamma^*_{w_i} < -z_k^t\beta^*$ by 1. We conclude that $z_i^t\beta^* > z_k^t\beta^*$ as required.

Conversely, assume there exists $\beta^*$ such that $z_i^t\beta^* > z_k^t\beta^*$ whenever $\delta_i = \delta_k = 1$ and $Y_i < Y_k$ or whenever $\delta_i = 1 - \delta_k = 1$ and $Y_i \leq Y_k$. Then it is possible to choose $\gamma^*$ such that whenever $\delta_i = \delta_k = 1$ and $Y_i < Y_k$, which implies $w_i < w_k$, we have $-z_i^t\beta^* < \gamma^*_{w_i} < -z_k^t\beta^* < \gamma^*_{w_k}$. This choice of $\gamma^*$ may be made so that in addition, whenever $\delta_i = 1$, $\delta_k = 0$, and $Y_i \leq Y_k$, which implies $w_i \leq w_k$, we have $-z_i^t\beta^* < \gamma^*_{w_i} \leq \gamma^*_{w_k} < -z_k^t\beta^*$, with equality only if $w_i = w_k$. Letting $\epsilon = 1$ for the moment, it is trivial to verify that for this choice of $\beta^*$ and $\gamma^*$, conditions 1, 2a, 2b, and 2c are satisfied. Therefore, for an arbirary $\epsilon > 0$, these four conditions may be satisfied by replacing $\beta^*$ and $\gamma^*$ by $K\beta^*$ and $K\gamma^*$ for some $K > 1$. This proves that $\sup_\theta L(\theta) = 0$. ■

**Proof of Proposition 4**. It is only necessary to show that conditions (c), (d), and (e) of Proposition 1 hold. Let $T : \theta^k \mapsto \theta^{k+1}$ denote the MM map. At a fixed point of the algorithm, equation (4.8) and the fact that $g_j'(\gamma_j^{k+1} \mid \theta^k) = 0$ imply that $dQ(\theta^k \mid \theta^k) = dL(\theta^k) = \mathbf{0}$, so fixed points of $T(\theta)$ coincide with stationary points of $L(\theta)$. Observe that the first differentials of $Q(\theta \mid \theta^k)$ and $L(\theta)$ are equal at $\theta = \theta^k$ because these functions are tangent at $\theta = \theta^k$. If $L(\theta^{k+1}) = L(\theta^k)$, then $Q(\theta^{k+1} \mid \theta^k) = Q(\theta^k \mid \theta^k)$. Given that $Q(\theta \mid \theta^k)$ has a unique maximizer, $\theta^{k+1} = \theta^k$.

It only remains to show that $T(\theta)$ is continuous. For this purpose, we follow closely an argument contained in Lange (1995a). Let $\{\phi^k\}_{k\geq 1}$ be any sequence of vectors in $R^{m+p}$ which converges to a limit $\phi$. The upper compactness of $L(\theta)$ and the ascent property (2.4) imply that all $T(\phi^k)$ belong to the same compact set. Let $\{T(\phi^{k_n})\}_{n\geq 1}$ be a subsequence with limit $\psi$. We must show that $\psi = T(\phi)$. Since the $\gamma$-update (4.11) is continuous by inspection, we need only show that the $\beta$ step (4.10) is continuous. Denote by a subscripted $\beta$ that portion of a parameter vector

corresponding to the $\beta$ components; thus, we wish to prove that $\psi_\beta = T_\beta(\phi)$. To accomplish this, we rewrite equation (4.10) as

$$\theta_\beta^{k+1} \;=\; \theta_\beta^k + \alpha(\theta^k)\delta(\theta^k),$$

where $\delta(\theta)$ denotes the $\beta$-search direction and $\alpha(\theta)$ is the appropriate step size maximizing the value of the function $a \mapsto f(\theta_\beta + a\delta(\theta) \mid \theta)$.

If $\delta(\phi) = \mathbf{0}$, then the strict concavity of $f(\beta \mid \phi)$ in its first argument implies that $\phi_\beta$ is the unique maximizer of $f(\beta \mid \phi)$. Because $f(\beta \mid \theta)$ is jointly continuous in its two arguments,

$$\begin{aligned} f(\psi_\beta \mid \phi) &= \lim_n f[T_\beta(\phi^{k_n}) \mid \phi^{k_n}] \\ &\geq \lim_n f(\phi_\beta^{k_n} \mid \phi^{k_n}) \\ &= f(\phi_\beta \mid \phi), \end{aligned}$$

which entails $\psi_\beta = \phi_\beta = T_\beta(\phi)$.

On the other hand, if $\delta(\phi) \neq 0$ and $\psi_\beta \neq \phi_\beta$, then

$$\frac{\psi_\beta - \phi_\beta}{\|\psi_\beta - \phi_\beta\|} \;=\; \lim_{n\to\infty} \frac{T_\beta(\phi^{k_n}) - \phi_\beta^{k_n}}{\|T_\beta(\phi^{k_n}) - \phi_\beta^{k_n}\|} \;=\; \lim_{n\to\infty} \frac{\delta(\phi^{k_n})}{\|\delta(\phi^{k_n})\|} \;=\; \frac{\delta(\phi)}{\|\delta(\phi)\|}$$

since $\delta(\theta)$ is continuous in $\theta$. Therefore, $\psi_\beta = \phi_\beta + c\delta(\phi)$ for $c = \|\psi_\beta - \phi_\beta\|/\|\delta(\phi)\|$, which means that $\phi_\beta = T_\beta(\phi)$ as long as $c \in (0, 1]$ and $f(\psi_\beta \mid \phi) \geq f[\phi_\beta + s\delta(\phi) \mid \phi]$ for all $s \in (0, 1]$. The first fact is verified by noting that

$$c \;=\; \lim_{n\to\infty} \frac{\|T_\beta(\phi^{k_n}) - \phi_\beta^{k_n}\|}{\|\delta(\phi^{k_n})\|} \;\leq\; 1$$

and the second by

$$\begin{aligned} f(\psi_\beta \mid \phi) &= \lim_{n\to\infty} f[T_\beta(\phi^{k_n}) \mid \phi^{k_n}] \\ &\geq \lim_{n\to\infty} f[\phi_\beta^{k_n} + s\delta(\phi^{k_n}) \mid \phi^{k_n}] \\ &= f[\phi_\beta + s\delta(\phi) \mid \phi]. \end{aligned}$$

■

**Proof of Proposition 5**. In the proofs of Propositions 1 and 4, continuity of the MM map $T : \theta^k \mapsto \theta^{k+1}$ was used only to show that $\lim_{n\to\infty} T(\theta^{k_n}) = \theta^*$ as $n \to \infty$, where $\{\theta^{k_n}\}$ is some subsequence with limit $\theta^*$. In this case, continuity fails but we are given that $\lim_{k\to\infty} T(\theta^k) = \lim_{k\to\infty} \theta^{k+1} = \theta^*$. The rest of the proof of Proposition 4 now goes through. ■

## References


Becker, M. P., Yang, I., and Lange, K. (1997). EM algorithms without missing data, *Statist. Meth. Med. Res.*, **6**, 38–54.

Bennett, S. (1983). Analysis of survival data by the proportional odds model, *Statist. in Medicine*, **2**, 273–277.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995). Analysis of transformation models with censored data, *Biometrika*, **82**, 835–845.

Conn, A. R., Gould, I. M., and Toint, P. L. (1991). Convergence of quasi-Newton matrices generated by the symmetric rank one update, *Math. Prog.*, **50**, 177–195.

Cox, D. R. (1972). Regression models and life tables (with discussion), *J. Roy. Statist. Soc. Ser. B*, **34**, 187–220.

Davidon, W. C. (1959). Variable metric methods for minimization, *AEC R&D Report ANL-5990*, Argonne National Laboratory.

de Leeuw, J. (1994). Block-relaxation algorithms in statistics, *Information Systems and Data Analysis* (eds. H. H. Bock *et al.*), 308–325, Springer-Verlag, Berlin.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. Ser. B*, **39**, 1–38.

Gay, D. M. (1981). Computing Optimal Locally Constrained Steps, *SIAM J. Sci. Stat. Comput.*, **2**, 186–197.

Heiser, W. J. (1995). Convergent computation by iterative majorization, *Recent Advances in Descriptive Multivariate Analysis* (ed. W. J. Krzanowski), 157–189, Oxford University Press, New York.

Hunter, D. R. and Lange, K. (2000). Rejoinder to discussion of optimization transfer using surrogate objective functions, *J. Comp. Graph. Statist.*, **9**, 52–59.

Jamshidian, M. and Jennrich, R. I. (1997). Quasi-Newton acceleration of the EM algorithm, *J. Roy. Statist. Soc. Ser. B*, **59**, 569–587.

Khalfan, H. F., Byrd, R. H., and Schnabel, R. B. (1993). A theoretical and experimental study of the symmetric rank one update, *SIAM J. Opt.*, **3**, 1–24.

Lange, K. (1995a). A gradient algorithm locally equivalent to the EM algorithm, *J. Roy. Statist. Soc. Ser. B*, **57**. 425–437.

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer algorithms using surrogate objective functions (with discussion), *J. Comp. Graph. Statist.*, **9**, 1–59.

Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, New York.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, Wiley, New York.

Murphy, S. A., Rossini, A. J., and Van der Vaart, A. W. (1997). MLE in the proportional odds model, *J. Am. Statist. Assoc.*, **92**, 968–976.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, Orlando.

Ripley, B. D. (1987). *Stochastic Simulation*, Wiley, New York.

Seneta, E. (1973). *Non-Negative Matrices*, Wiley, New York.