A Tutorial on MM Algorithms

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]
David Geffen School of Medicine at UCLA
Los Angeles, CA 90095-1766

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

December 10, 2003

Abstract

Most problems in frequentist statistics involve optimization of a function such as a likelihood or a sum of squares. EM algorithms are among the most effective algorithms for maximum likelihood estimation because they consistently drive the likelihood uphill by maximizing a simple surrogate function for the loglikelihood. Iterative optimization of a surrogate function as exemplified by an EM algorithm does not necessarily require missing data. Indeed, every EM algorithm is a special case of the more general class of MM optimization algorithms, which typically exploit convexity rather than missing data in majorizing or minorizing an objective function. In our opinion, MM algorithms deserve to part of the standard toolkit of professional statisticians. The current article explains the principle behind MM algorithms, suggests some methods for constructing them, and discusses some of their attractive features. We include numerous examples throughout the article to illustrate the concepts described. In addition to surveying previous work on MM algorithms, this article introduces some new material on constrained optimization and standard error estimation.

Key words and phrases: constrained optimization, EM algorithm, majorization, minorization, Newton-Raphson

1 Introduction

Maximum likelihood and least squares are the dominant forms of estimation in frequentist statistics. Toy optimization problems designed for classroom presentation can be solved analytically, but most practical maximum likelihood and least squares estimation problems must be solved numerically. In the current article, we discuss an optimization method that typically relies on convexity arguments and is a generalization of the well-known EM algorithm method (Dempster et al., 1977; McLachlan and Krishnan, 1997). We call any algorithm based on this iterative method an MM algorithm.

To our knowledge, the general principle behind MM algorithms was first enunciated by the numerical analysts Ortega and Rheinboldt (1970) in the context of line search methods. de Leeuw and Heiser (1977) present an MM algorithm for multidimensional scaling contemporary with the classic Dempster et al. (1977) paper on EM algorithms. Although the work of de Leeuw and Heiser did not spark the same explosion of interest from the statistical community set off by the Dempster et al. (1977) paper, steady development of MM algorithms has continued. The MM principle reappears, among other places, in robust regression (Huber, 1981), in correspondence analysis (Heiser, 1987), in the quadratic lower bound principle of Böhning and Lindsay (1988), in the psychometrics literature on least squares (Bijleveld and de Leeuw, 1991; Kiers and Ten Berge, 1992), and in medical imaging (De Pierro, 1995; Lange and Fessler, 1995). The recent survey articles of de Leeuw (1994), Heiser (1995), Becker et al. (1997), and Lange et al. (2000) deal with the general principle, but it is not until the rejoinder of Hunter and Lange (2000a) that the acronym MM first appears. This acronym pays homage to the earlier names "majorization" and "iterative majorization" of the MM principle, emphasizes its crucial link to the better-known EM principle, and diminishes the possibility of confusion with the distinct subject in mathematics known as majorization (Marshall and Olkin, 1979). Recent work has demonstrated the utility of MM algorithms in a broad range of statistical contexts, including quantile regression (Hunter and Lange, 2000b), survival analysis (Hunter and Lange, 2002), paired and multiple comparisons (Hunter, 2004), variable selection (Hunter and Li, 2002), and DNA sequence analysis (Sabatti and Lange, 2002).

One of the virtues of the MM acronym is that it does double duty. In minimization problems, the first M of MM stands for majorize and the second M for minimize. In maximization problems, the first M stands for minorize and the second M for maximize. (We define the terms "majorize" and "minorize" in Section 2.) A successful MM algorithm substitutes a simple optimization problem for a difficult optimization problem. Simplicity can be attained by (a) avoiding large matrix inversions, (b) linearizing an optimization problem, (c) separating the parameters of an optimization problem, (d) dealing with equality and inequality constraints gracefully, or (e) turning a nondifferentiable problem into a smooth problem. Iteration is the price we pay for simplifying the original problem.

In our view, MM algorithms are easier to understand and sometimes easier to apply than EM algorithms. Although we have no intention of detracting from EM algorithms, their dominance over MM algorithms is a historical accident. An EM algorithm operates by identifying a theoretical complete data space. In the E step of the algorithm, the conditional expectation of the complete data loglikelihood is calculated with respect to the observed data. The surrogate function created by the E step is, up to a constant, a minorizing function. In the M step, this minorizing function is maximized with respect to the parameters of the underlying model; thus, every EM algorithm is an example of an MM algorithm. Construction of an EM algorithm sometimes demands creativity in identifying the complete data and technical skill in calculating an often complicated conditional expectation and then maximizing it analytically.

In contrast, typical applications of MM revolve around careful inspection of a loglikelihood or other objective function to be optimized, with particular attention paid to convexity and inequalities. Thus, success with MM algorithms and success with EM algorithms hinge on somewhat different mathematical maneuvers. However, the skills required by most MM algorithms are no harder to master than the skills required by most EM algorithms. The purpose of this article is to present some strategies for constructing MM algorithms and to illustrate various aspects of these algorithms through the study of specific examples.

We conclude this section with a note on nomenclature. Just as EM is more a prescription for creating algorithms than an actual algorithm, MM refers not to a single algorithm but to a class of algorithms. Thus, this article refers to specific EM and MM algorithms but never to "*the* MM algorithm" or "*the* EM algorithm".

2 The MM Philosophy

Let $\theta^{(m)}$ represent a fixed value of the parameter θ, and let $g(\theta \mid \theta^{(m)})$ denote a real-valued function of θ whose form depends on $\theta^{(m)}$. The function $g(\theta \mid \theta^{(m)})$ is said to majorize a real-valued function $f(\theta)$ at the point $\theta^{(m)}$ provided

$$\begin{array}{rcl} g(\theta \mid \theta^{(m)}) & \geq & f(\theta) \text{ for all } \theta, \\ g(\theta^{(m)} \mid \theta^{(m)}) & = & f(\theta^{(m)}). \end{array} \tag{1}$$

In other words, the surface $\theta \mapsto g(\theta \mid \theta^{(m)})$ lies above the surface $f(\theta)$ and is tangent to it at the point $\theta = \theta^{(m)}$. The function $g(\theta \mid \theta^{(m)})$ is said to minorize $f(\theta)$ at $\theta^{(m)}$ if $-g(\theta \mid \theta^{(m)})$ majorizes $-f(\theta)$ at $\theta^{(m)}$.

Ordinarily, $\theta^{(m)}$ represents the current iterate in a search of the surface $f(\theta)$. In a majorize-minimize MM algorithm, we minimize the majorizing function $g(\theta \mid \theta^{(m)})$ rather than the actual function $f(\theta)$. If $\theta^{(m+1)}$ denotes the minimizer of $g(\theta \mid \theta^{(m)})$, then we can show that the MM procedure forces $f(\theta)$ downhill. Indeed, the inequality

$$\begin{aligned} f(\theta^{(m+1)}) &= g(\theta^{(m+1)} \mid \theta^{(m)}) + f(\theta^{(m+1)}) - g(\theta^{(m+1)} \mid \theta^{(m)}) \\ &\leq g(\theta^{(m)} \mid \theta^{(m)}) + f(\theta^{(m)}) - g(\theta^{(m)} \mid \theta^{(m)}) \qquad (2) \\ &= f(\theta^{(m)}) \end{aligned}$$

follows directly from the fact $g(\theta^{(m+1)} \mid \theta^{(m)}) \leq g(\theta^{(m)} \mid \theta^{(m)})$ and definition (1). The descent property (2) lends an MM algorithm remarkable numerical stability. With straightforward changes, the MM recipe also applies to maximization rather than minimization: To maximize a function $f(\theta)$, we minorize it by a surrogate function $g(\theta \mid \theta^{(m)})$ and maximize $g(\theta \mid \theta^{(m)})$ to produce the next iterate $\theta^{(m+1)}$.

2.1 Calculation of Sample Quantiles

As a one-dimensional example, consider the problem of computing a sample quantile from a sample $x_1, \ldots, x_n$ of n real numbers. One can readily prove (Hunter and Lange, 2000b) that for $q \in (0,1)$, a qth sample quantile of $x_1, \ldots, x_n$ minimizes the function

$$f(\theta) = \sum_{i=1}^{n} \rho_q(x_i - \theta), \qquad (3)$$

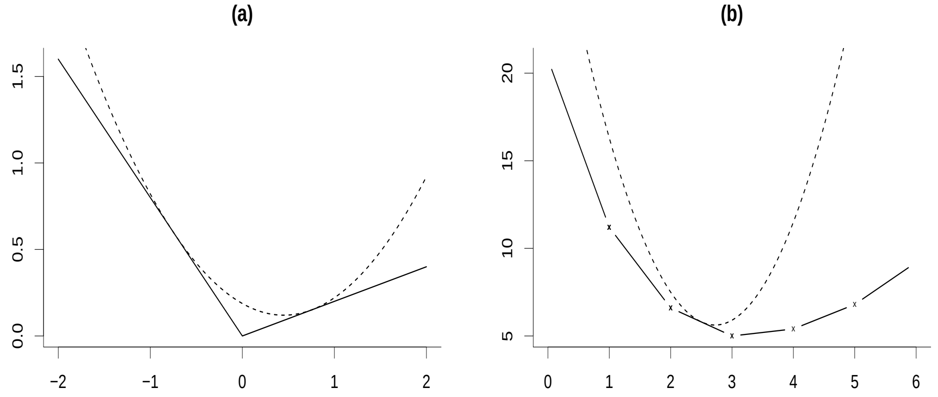

Figure 1: For $q = 0.8$, (a) depicts the "vee" function $\rho_q(\theta)$ and its quadratic majorizing function for $\theta^{(m)} = -0.75$; (b) shows the objective function $f(\theta)$ that is minimized by the 0.8 quantile of the sample $1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 5$, along with its quadratic majorizer, for $\theta^{(m)} = 2.5$.

where $\rho_q(\theta)$ is the "vee" function

$$\rho_q(\theta) \quad = \quad \begin{cases} q\theta & \theta \geq 0 \\ -(1-q)\theta & \theta < 0. \end{cases}$$

When $q = 1/2$, this function is proportional to the absolute value function; for $q \neq 1/2$, the "vee" is tilted to one side or the other. As seen in Figure 1(a), it is possible to majorize the "vee" function at any nonzero point by a simple quadratic function. Specifically, for a given $\theta^{(m)} \neq 0$, $\rho_q(\theta)$ is majorized at $\pm\theta^{(m)}$ by

$$\zeta_q(\theta \mid \theta^{(m)}) \quad = \quad \frac{1}{4}\left\{\frac{\theta^2}{|\theta^{(m)}|} + (4q-2)\theta + |\theta^{(m)}|\right\}.$$

Fortunately, the majorization relation between functions is closed under the formation of sums, nonnegative products, limits, and composition with an increasing function. These rules permit us to work piecemeal in simplifying complicated objective functions. Thus, the function $f(\theta)$ of equation (3) is majorized at the point

$\theta^{(m)}$ by

$$g(\theta \mid \theta^{(m)}) \quad = \quad \sum_{i=1}^{n} \zeta_q(x_i - \theta \mid x_i - \theta^{(m)}). \tag{4}$$

The function $f(\theta)$ and its majorizer $g(\theta \mid \theta^{(m)})$ are shown in Figure 1(b) for a particular sample of size $n = 12$.

Setting the first derivative of $g(\theta \mid \theta^{(m)})$ equal to zero gives the minimum point

$$\theta^{(m+1)} \quad = \quad \frac{n(2q-1) + \sum_{i=1}^{n} w_i^{(m)} x_i}{\sum_{i=1}^{n} w_i^{(m)}}, \tag{5}$$

where the weight $w_i^{(m)} = |x_i - \theta^{(m)}|^{-1}$ depends on $\theta^{(m)}$. A flaw of algorithm (5) is that the weight $w_i^{(m)}$ is undefined whenever $\theta^{(m)} = x_i$. In mending this flaw, Hunter and Lange (2000b) also discuss the broader technique of quantile regression introduced by Koenker and Bassett (1978). From a computational perspective, the most fascinating thing about the quantile-finding algorithm is that it avoids sorting and relies entirely on arithmetic and iteration. For the case of the sample median ($q = 1/2$), algorithm (5) is found in Schlossmacher (1973) and is shown to be an MM algorithm by Lange and Sinsheimer (1993) and Heiser (1995).

Because $g(\theta \mid \theta^{(m)})$ in equation (4) is a quadratic function of θ, expression (5) coincides with the more general Newton-Raphson update

$$\theta^{(m+1)} \quad = \quad \theta^{(m)} - \left[\nabla^2 g(\theta^{(m)} \mid \theta^{(m)})\right]^{-1} \nabla g(\theta^{(m)} \mid \theta^{(m)}), \tag{6}$$

where $\nabla g(\theta^{(m)} \mid \theta^{(m)})$ and $\nabla^2 g(\theta^{(m)} \mid \theta^{(m)})$ denote the gradient vector and the Hessian matrix of $g(\theta \mid \theta^{(m)})$ evaluated at $\theta^{(m)}$. Since the descent property (2) depends only on decreasing $g(\theta \mid \theta^{(m)})$ and not on minimizing it, the update (6) can serve in cases where $g(\theta \mid \theta^{(m)})$ lacks a closed-form minimizer, provided this update decreases the value of $g(\theta \mid \theta^{(m)})$. In the context of EM algorithms, Dempster et al. (1977) call an algorithm that reduces $g(\theta \mid \theta^{(m)})$ without actually minimizing it a generalized EM (GEM) algorithm. The specific case of equation (6), which we call a gradient MM algorithm, is studied in the EM context by Lange (1995a), who points out that update (6) saves us from performing iterations within iterations and yet still displays the same local rate of convergence as a full MM algorithm that minimizes $g(\theta \mid \theta^{(m)})$ at each iteration.

3 Tricks of the Trade

In the quantile example of Section 2.1, the convex "vee" function admits a quadratic majorizer as depicted in Figure 1(a). In general, many majorizing or minorizing relationships may be derived from various inequalities stemming from convexity or concavity. This section outlines some common inequalities used to construct majorizing or minorizing functions for various types of objective functions.

3.1 Jensen's Inequality

Jensen's inequality states for a convex function $\kappa(x)$ and any random variable X that $\kappa[\mathrm{E}\,(X)] \leq \mathrm{E}\,[\kappa(X)]$. Since $-\ln(x)$ is a convex function, we conclude for probability densities $a(x)$ and $b(x)$ that

$$-\ln\left\{\mathrm{E}\left[\frac{a(X)}{b(X)}\right]\right\} \quad \leq \quad -\mathrm{E}\left[\ln\frac{a(X)}{b(X)}\right].$$

If X has the density $b(x)$, then $\mathrm{E}\,[a(X)/b(X)] = 1$, so the left hand side above vanishes and we obtain

$$\mathrm{E}\,[\ln a(X)] \quad \leq \quad \mathrm{E}\,[\ln b(X)]\,,$$

which is sometimes known as the information inequality. It is this inequality that guarantees that a minorizing function is constructed in the E-step of any EM algorithm (de Leeuw, 1994; Heiser, 1995), making every EM algorithm an MM algorithm.

3.2 Minorization via Supporting Hyperplanes

Jensen's inequality is easily derived from the supporting hyperplane property of a convex function: Any linear function tangent to the graph of a convex function is a minorizer at the point of tangency. Thus, if $\kappa(\theta)$ is convex and differentiable, then

$$\kappa(\theta) \quad \geq \quad \kappa(\theta^{(m)}) + \nabla\kappa(\theta^{(m)})^t(\theta - \theta^{(m)}), \tag{7}$$

with equality when $\theta = \theta^{(m)}$. This inequality is illustrated by the example of Section 7 involving constrained optimization.

3.3 Majorization via the Definition of Convexity

If we wish to majorize a convex function instead of minorizing it, then we can use the standard definition of convexity; namely, $\kappa(t)$ is convex if and only if

$$\kappa\Big(\sum_i \alpha_i t_i\Big) \quad \leq \quad \sum_i \alpha_i \kappa(t_i) \tag{8}$$

for any finite collection of points t_i and corresponding multipliers α_i with $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. Application of definition (8) is particularly effective when $\kappa(t)$ is composed with a linear function $x^t\theta$. For instance, suppose for vectors x, θ, and $\theta^{(m)}$ that we make the substitution $t_i = x_i(\theta_i - \theta_i^{(m)})/\alpha_i + x^t\theta^{(m)}$. Inequality (8) then becomes

$$\kappa(x^t\theta) \quad \leq \quad \sum_i \alpha_i \kappa\left[\frac{x_i}{\alpha_i}(\theta_i - \theta_i^{(m)}) + x^t\theta^{(m)}\right]. \tag{9}$$

Alternatively, if all components of x, θ, and $\theta^{(m)}$ are positive, then we may take $t_i = x^t\theta^{(m)}\theta_i/\theta_i^{(m)}$ and $\alpha_i = x_i\theta_i^{(m)}/x^t\theta^{(m)}$. Now inequality (8) becomes

$$\kappa(x^t\theta) \quad \leq \quad \sum_i \frac{x_i\theta_i^{(m)}}{x^t\theta^{(m)}}\kappa\left[\frac{x^t\theta^{(m)}\theta_i}{\theta_i^{(m)}}\right]. \tag{10}$$

Inequalities (9) and (10) have been used to construct MM algorithms in the contexts of medical imaging (De Pierro, 1995; Lange and Fessler, 1995) and least-squares estimation without matrix inversion (Becker et al., 1997).

3.4 Majorization via a Quadratic Upper Bound

If a convex function $\kappa(\theta)$ is twice differentiable and has bounded curvature, then we can majorize $\kappa(\theta)$ by a quadratic function with sufficiently high curvature and tangent to $\kappa(\theta)$ at $\theta^{(m)}$ (Böhning and Lindsay, 1988). In algebraic terms, if we can find a positive definite matrix M such that $M - \nabla^2\kappa(\theta)$ is nonnegative definite for all θ, then

$$\kappa(\theta) \;\leq\; \kappa(\theta^{(m)}) + \nabla\kappa(\theta^{(m)})^t(\theta - \theta^{(m)}) + \frac{1}{2}(\theta - \theta^{(m)})^t M(\theta - \theta^{(m)})$$

provides a quadratic upper bound. For example, Heiser (1995) notes in the unidimensional case that

$$\frac{1}{\theta} \;\leq\; \frac{1}{\theta^{(m)}} - \frac{\theta - \theta^{(m)}}{(\theta^{(m)})^2} + \frac{(\theta - \theta^{(m)})^2}{c^3}$$

for $0 < c \leq \min\{\theta, \theta^{(m)}\}$. The corresponding quadratic lower bound principle for minorization is the basis for the logistic regression example of Section 6.

3.5 The Arithmetic-Geometric Mean Inequality

The arithmetic-geometric mean inequality is a special case of inequality (8). Taking $\kappa(t) = e^t$ and $\alpha_i = 1/m$ yields

$$\exp\left(\frac{1}{m}\sum_{i=1}^m t_i\right) \;\leq\; \frac{1}{m}\sum_{i=1}^m e^{t_i}.$$

If we let $x_i = e^{t_i}$, then we obtain the standard form

$$\sqrt[m]{\prod_{i=1}^m x_i} \;\leq\; \frac{1}{m}\sum_{i=1}^m x_i \qquad (11)$$

of the arithmetic-geometric mean inequality. Because the exponential function is strictly convex, equality holds if and only if all of the x_i are equal. Inequality (11) is helpful in constructing the majorizer

$$x_1 x_2 \quad \leq \quad x_1^2 \frac{x_2^{(m)}}{2x_1^{(m)}} + x_2^2 \frac{x_1^{(m)}}{2x_2^{(m)}} \tag{12}$$

of the product of two positive numbers. This inequality is used in the sports contest model of Section 4.

3.6 The Cauchy-Schwartz Inequality

The Cauchy-Schwartz inequality for the Euclidean norm is a special case of inequality (7). The function $\kappa(\theta) = \|\theta\|$ is convex because it satisfies the triangle inequality and the homogeneity condition $\|\alpha\theta\| = |\alpha| \cdot \|\theta\|$. Since $\kappa(\theta) = \sqrt{\sum_i \theta_i^2}$, we see that $\nabla\kappa(\theta) = \theta/\|\theta\|$, and therefore inequality (7) gives

$$\|\theta\| \quad \geq \quad \|\theta^{(m)}\| + \frac{(\theta - \theta^{(m)})^t \theta^{(m)}}{\|\theta^{(m)}\|} \quad = \quad \frac{\theta^t \theta^{(m)}}{\|\theta^{(m)}\|}, \tag{13}$$

which is the Cauchy-Schwartz inequality. de Leeuw and Heiser (1977) and Groenen (1993) use inequality (13) to derive MM algorithms for multidimensional scaling.

4 Separation of Parameters and Cyclic MM

One of the key criteria in judging minorizing or majorizing functions is their ease of optimization. Successful MM algorithms in high-dimensional parameter spaces often rely on surrogate functions in which the individual parameter components are separated. In other words, the surrogate function mapping $\theta \in U \subset R^d \to R$ reduces to the sum of d real-valued functions taking the real-valued arguments θ_1 through θ_d. Since the d univariate functions may be optimized one by one, this makes the surrogate function easier to optimize at each iteration.

4.1 Poisson Sports Model

Consider a simplified version of a model proposed by Maher (1982) for a sports contest between two individuals or teams in which the number of points scored by team i against team j follows a Poisson process with intensity $e^{o_i - d_j}$, where o_i is an "offensive strength" parameter for team i and d_j is a "defensive strength" parameter for team j. If t_{ij} is the length of time that i plays j and p_{ij} is the number of points that i scores against j, then the corresponding Poisson loglikelihood function is

$$\ell_{ij}(\theta) \quad = \quad p_{ij}(o_i - d_j) - t_{ij}e^{o_i - d_j} + p_{ij}\ln t_{ij} - \ln p_{ij}!, \tag{14}$$

where $\theta = (o, d)$ is the parameter vector. Note that the parameters should satisfy a linear constraint, such as $\sum_i o_i + \sum_j d_j = 0$, in order for the model be identifiable; otherwise, it is clearly possible to add the same constant to each o_i and d_j without altering the likelihood. We make two simplifying assumptions. First, different games are independent of each other. Second, each team's point total within a single game is independent of its opponent's point total. The second assumption is more suspect than the first since it implies that a team's offensive and defensive performances are somehow unrelated to one another; nonetheless the model gives an interesting first approximation to reality. Under these assumptions, the full data loglikelihood is obtained by summing $\ell_{ij}(\theta)$ over all pairs (i, j). Setting the partial derivatives of the loglikelihood equal to zero leads to the equations

$$e^{-\hat{d}_j} \quad = \quad \frac{\sum_i p_{ij}}{\sum_i t_{ij}e^{\hat{o}_i}} \qquad \text{and} \qquad e^{\hat{o}_i} \quad = \quad \frac{\sum_j p_{ij}}{\sum_j t_{ij}e^{-\hat{d}_j}}$$

satisfied by the maximum likelihood estimate $(\hat{o}, \hat{d})$. These equations do not admit a closed-form solution, so we turn to an MM algorithm.

Because the task is to maximize the loglikelihood (14), we need a minorizing function. Focusing on the $-t_{ij}e^{o_i-d_j}$ term, we may use inequality (12) to show that

$$-t_{ij}e^{o_i-d_j} \quad \geq \quad -\frac{t_{ij}}{2}\frac{e^{2o_i}}{e^{o_i^{(m)}+d_j^{(m)}}} - \frac{t_{ij}}{2}e^{-2d_j}e^{o_i^{(m)}+d_j^{(m)}}. \tag{15}$$

Although the right side of the above inequality may appear more complicated than the left side, it is actually simpler in one important respect — the parameter components o_i and d_j are separated on the right side but not on the left. Summing the loglikelihood (14) over all pairs (i,j) and invoking inequality (15) yields the function

$$g(\theta \mid \theta^{(m)}) \quad = \quad \sum_i \sum_j \left[p_{ij}(o_i - d_j) - \frac{t_{ij}}{2}\frac{e^{2o_i}}{e^{o_i^{(m)}+d_j^{(m)}}} - \frac{t_{ij}}{2}e^{-2d_j}e^{o_i^{(m)}+d_j^{(m)}} \right]$$

minorizing the full loglikelihood up to an additive constant independent of θ. The fact that the components of θ are separated by $g(\theta \mid \theta^{(m)})$ permits us to update parameters one by one and substantially reduces computational costs. Setting the partial derivatives of $g(\theta \mid \theta^{(m)})$ equal to zero yields the updates

$$o_i^{(m+1)} = \frac{1}{2}\ln\left\{\frac{\sum_j p_{ij}}{\sum_j t_{ij}e^{-o_i^{(m)}-d_j^{(m)}}}\right\}, \; d_j^{(m+1)} = -\frac{1}{2}\ln\left\{\frac{\sum_i p_{ij}}{\sum_i t_{ij}e^{o_i^{(m)}+d_j^{(m)}}}\right\}. \tag{16}$$

The question now arises as to whether one should modify algorithm (16) so that updated subsets of the parameters are used as soon as they become available. For instance, if we update the o vector before the d vector in each iteration of algorithm (16), we could replace the formula for $d_j^{(m+1)}$ above by

$$d_j^{(m+1)} = -\frac{1}{2}\ln\left\{\frac{\sum_i p_{ij}}{\sum_i t_{ij}e^{o_i^{(m+1)}+d_j^{(m)}}}\right\}. \tag{17}$$

In practice, an MM algorithm often takes fewer iterations when we cycle through the parameters updating one at a time than when we update the whole vector at once as in algorithm (16). We call such versions of MM algorithms cyclic MM algorithms; they generalize the ECM algorithms of Meng and Rubin (1993). A cyclic MM algorithm always drives the objective function in the right direction; indeed, every iteration of a cyclic MM algorithm is simply an MM iteration on a reduced parameter set.

Team	$\hat{o}_i + \hat{d}_i$	Wins	Team	$\hat{o}_i + \hat{d}_i$	Wins
Cleveland	-0.0994	17	Phoenix	0.0166	44
Denver	-0.0845	17	New Orleans	0.0169	47
Toronto	-0.0647	24	Philadelphia	0.0187	48
Miami	-0.0581	25	Houston	0.0205	43
Chicago	-0.0544	30	Minnesota	0.0259	51
Atlanta	-0.0402	35	LA Lakers	0.0277	50
LA Clippers	-0.0355	27	Indiana	0.0296	48
Memphis	-0.0255	28	Utah	0.0299	47
New York	-0.0164	37	Portland	0.0320	50
Washington	-0.0153	37	Detroit	0.0336	50
Boston	-0.0077	44	New Jersey	0.0481	49
Golden State	-0.0051	38	San Antonio	0.0611	60
Orlando	-0.0039	42	Sacramento	0.0686	59
Milwaukee	-0.0027	42	Dallas	0.0804	60
Seattle	0.0039	40			

Table 1: Ranking of all 29 NBA teams on the basis of the 2002-2003 regular season according to their estimated offensive strength plus defensive strength. Each team played 82 games.

4.2 Application to National Basketball Association Results

Table 1 summarizes our application of the Poisson sports model to the results of the 2002–2003 regular season of the National Basketball Association. In these data, t_{ij} is measured in minutes. A regular game lasts 48 minutes, and each overtime period, if necessary, adds five minutes. Thus, team i is expected to score $48e^{\hat{o}_i - \hat{d}_j}$ points against team j when the two teams meet and do not tie. Team i is ranked higher than team j if $\hat{o}_i - \hat{d}_j > \hat{o}_j - \hat{d}_i$, which is equivalent to $\hat{o}_i + \hat{d}_i > \hat{o}_j + \hat{d}_j$.

It is worth emphasizing some of the virtues of the model. First, the ranking of the 29 NBA teams on the basis of the estimated sums $\hat{o}_i + \hat{d}_i$ for the 2002-2003 regular season is not perfectly consistent with their cumulative wins; strength of schedule and margins of victory are reflected in the model. Second, the model gives the point-spread function for a particular game as the difference of two independent Poisson random variables. Third, one can easily amend the model to rank individual players rather than teams by assigning to each player an offensive and defensive intensity parameter. If each game is divided into time segments punctuated by substitutions, then the MM algorithm can be adapted to estimate the assigned player intensities. This might provide a rational basis for salary negotiations that takes into account subtle differences between players not reflected in traditional sports statistics.

Finally, the NBA data set sheds light on the comparative speeds of the original MM algorithm (16) and its cyclic modification (17). The cyclic MM algorithm converged in fewer iterations (25 instead of 28). However, because of the additional work required to recompute the denominators in equation (17), the cyclic version required slightly more floating-point operations as counted by MATLAB (301,157 instead of 289,998).

5 Speed of Convergence

MM algorithms and Newton-Raphson algorithms have complementary strengths. On one hand, Newton-Raphson algorithms boast a quadratic rate of convergence as they near a local optimum point θ^*. In other words, under certain general conditions,

$$\lim_{m\to\infty} \frac{\|\theta^{(m+1)} - \theta^*\|}{\|\theta^{(m)} - \theta^*\|^2} \quad = \quad c$$

for some constant c. This quadratic rate of convergence is much faster than the linear rate of convergence

$$\lim_{m\to\infty} \frac{\|\theta^{(m+1)} - \theta^*\|}{\|\theta^{(m)} - \theta^*\|} \;=\; c \;<\; 1 \tag{18}$$

displayed by typical MM algorithms. Hence, Newton-Raphson algorithms tend to require fewer iterations than MM algorithms. On the other hand, an iteration of a Newton-Raphson algorithm can be far more computationally onerous than an iteration of an MM algorithm. Examination of the form

$$\theta^{(m+1)} \;=\; \theta^{(m)} - \nabla^2 f(\theta^{(m)})^{-1} \nabla f(\theta^{(m)})$$

of a Newton-Raphson iteration reveals that it requires evaluation and inversion of the Hessian matrix $\nabla^2 f(\theta^{(m)})$. If θ has p components, then the number of calculations needed to invert the $p \times p$ matrix $\nabla^2 f(\theta)$ is roughly proportional to p^3. By contrast, an MM algorithm that separates parameters usually takes on the order of p or p^2 arithmetic operations per iteration. Thus, well-designed MM algorithms tend to require more iterations but simpler iterations than Newton-Raphson. For this reason MM algorithms sometimes enjoy an advantage in computational speed.

For example, the Poisson process scoring model for the NBA data set of Section 4 has 57 parameters (two for each of 29 teams minus one for the linear constraint). A single matrix inversion of a 57×57 matrix requires roughly 387,000 floating point operations according to MATLAB. Thus, even a single Newton-Raphson iteration requires more computation in this example than the 300,000 floating point operations required for the MM algorithm to converge completely in 28 iterations. Numerical stability also enters the balance sheet. A Newton-Raphson algorithm can behave poorly if started too far from an optimum point. By contrast, MM algorithms are guaranteed to appropriately increase or decrease the value of the objective function at every iteration.

Other types of deterministic optimization algorithms, such as Fisher scoring, quasi-Newton methods, or gradient-free methods like Nelder-Mead, occupy a kind of middle ground. Although none of them can match Newton-Raphson in required iterations until convergence, each has its own merits. The expected information matrix used in Fisher scoring is sometimes easier to evaluate than the observed information matrix of Newton-Raphson. Scoring does not automatically lead to an increase in the loglikelihood, but at least (unlike Newton-Raphson) it can always be made to do so if some form of backtracking is incorporated. Quasi-Newton methods mitigate or even eliminate the need for matrix inversion. The Nelder-Mead approach is applicable in situations where the objective function is nondifferentiable. Because of the complexities of practical problems, it is impossible to declare any optimization algorithm best overall. In our experience, however, MM algorithms are often difficult to beat in terms of stability and computational simplicity.

6 Standard Error Estimates

In most cases, a maximum likelihood estimator has asymptotic covariance matrix equal to the inverse of the expected information matrix. In practice, the expected information matrix is often well-approximated by the observed information matrix $-\nabla^2\ell(\theta)$ computed by differentiating the loglikelihood $\ell(\theta)$ twice. Thus, after the MLE $\hat{\theta}$ has been found, a standard error of $\hat{\theta}$ can be obtained by taking square roots of the diagonal entries of the inverse of $-\nabla^2\ell(\hat{\theta})$. In some problems, however, direct calculation of $\nabla^2\ell(\hat{\theta})$ is difficult. Here we propose two numerical approximations to this matrix that exploit quantities readily obtained by running an MM algorithm.

Let $g(\theta \mid \theta^{(m)})$ denote a minorizing function of the loglikelihood $\ell(\theta)$ at the point $\theta^{(m)}$, and define

$$M(\vartheta) = \arg\max_{\theta} g(\theta \mid \vartheta)$$

to be the MM algorithm map taking $\theta^{(m)}$ to $\theta^{(m+1)}$.

6.1 Numerical Differentiation via MM

The two numerical approximations to $-\nabla^2\ell(\hat{\theta})$ are based on the formulas

$$\begin{aligned} \nabla^2\ell(\hat{\theta}) &= \nabla^2 g(\hat{\theta} \mid \hat{\theta})\left[I - \nabla M(\hat{\theta})\right] && (19) \\ &= \nabla^2 g(\hat{\theta} \mid \hat{\theta}) + \left[\frac{\partial}{\partial\vartheta}\nabla g(\hat{\theta} \mid \vartheta)\right]_{\vartheta=\hat{\theta}}, && (20) \end{aligned}$$

where I denotes the identity matrix. These formulas are derived in Lange (1999) using two simple facts: First, the tangency of $\ell(\theta)$ and its minorizer imply that their gradient vectors are equal at the point of minorization; and second, the gradient of $g(\theta \mid \theta^{(m)})$ at its maximizer $M(\theta^{(m)})$ is zero. Alternative derivations of formulas (19) and (20) are given by Meng and Rubin (1991) and Oakes (1999), respectively. Although these formulas have been applied to standard error estimation in the EM algorithm literature — Meng and Rubin (1991) base their SEM idea on formula (19) — to our knowledge, neither has been applied in the broader context of MM algorithms.

Approximation of $\nabla^2\ell(\hat{\theta})$ using equation (19) requires a numerical approximation of the Jacobian matrix $\nabla M(\theta)$, whose i, j entry equals

$$\frac{\partial}{\partial\theta_j} M_i(\theta) = \lim_{\delta\to 0} \frac{M_i(\theta + \delta e_j) - M_i(\theta)}{\delta}, \qquad (21)$$

where the vector e_j is the jth standard basis vector having a one in its jth component and zeros elsewhere. Since $M(\hat{\theta}) = \hat{\theta}$, the jth column of $\nabla M(\hat{\theta})$ may be approximated using only output from the corresponding MM algorithm by (a) iterating until $\hat{\theta}$ is found, (b) altering the jth component of $\hat{\theta}$ by a small amount δ_j, (c) applying the MM algorithm to this altered θ, (d) subtracting $\hat{\theta}$ from the result, and (e) dividing by δ_j. Approximation of $\nabla^2\ell(\hat{\theta})$ using equation (20) is analogous except it involves numerically approximating the Jacobian of $h(\vartheta) = \nabla g(\hat{\theta} \mid \vartheta)$. In this case one may exploit the fact that $h(\hat{\theta})$ is zero.

6.2 An MM Algorithm for Logistic Regression

To illustrate these ideas and facilitate comparison of the various numerical methods, we consider an example in which the Hessian of the loglikelihood is easy to compute. Böhning and Lindsay (1988) apply the quadratic bound principle of Section 3.4 to the case of logistic regression, in which we have an $n \times 1$ vector Y of binary responses and an $n \times p$ matrix X of predictors. The model stipulates that the probability $\pi_i(\theta)$ that $Y_i = 1$ equals $\exp\{\theta^t x_i\}/\left(1 + \exp\{\theta^t x_i\}\right)$, Straightforward differentiation of the resulting loglikelihood function shows that

$$\nabla^2\ell(\theta) \quad = \quad -\sum_{i=1}^{n} \pi_i(\theta)[1 - \pi_i(\theta)]x_i x_i^t.$$

Since $\pi_i(\theta)[1 - \pi_i(\theta)]$ is bounded above by $\frac{1}{4}$, we may define the negative definite matrix $B = -\frac{1}{4}X^t X$ and conclude that $\nabla^2\ell(\theta) - B$ is nonnegative definite as desired. Therefore, the quadratic function

$$g(\theta \mid \theta^{(m)}) \quad = \quad \ell(\theta^{(m)}) + \nabla\ell(\theta^{(m)})^t(\theta - \theta^{(m)}) + \frac{1}{2}(\theta - \theta^{(m)})^t B(\theta - \theta^{(m)})$$

minorizes $\ell(\theta)$ at $\theta^{(m)}$. The MM algorithm proceeds by maximizing this quadratic, giving

$$\begin{aligned} \theta^{(m+1)} &= \theta^{(m)} - B^{-1}\nabla\ell(\theta^{(m)}) \\ &= \theta^{(m)} - 4(X^t X)^{-1}X^t[Y - \pi(\theta^{(m)})]. \end{aligned} \qquad (22)$$

Since the MM algorithm of equation (22) needs to invert X^tX only once, it enjoys an increasing computational advantage over Newton-Raphson as the number of predictors p increases (Böhning and Lindsay, 1988).

		Standard errors based on:		
Variable	$\hat{\theta}$	Exact $\nabla^2\ell(\hat{\theta})$	Eqn (19)	Eqn (20)
Constant	0.48062	1.1969	1.1984	1.1984
AGE	−0.029549	0.037031	0.037081	0.037081
LWT	−0.015424	0.0069194	0.0069336	0.0069336
RACE2	1.2723	0.52736	0.52753	0.52753
RACE3	0.8805	0.44079	0.44076	0.44076
SMOKE	0.93885	0.40215	0.40219	0.40219
PTL	0.54334	0.34541	0.34545	0.34545
HT	1.8633	0.69754	0.69811	0.69811
UI	0.76765	0.45932	0.45933	0.45933
FTV	0.065302	0.17240	0.17251	0.17251

Table 2: Estimated coefficients and standard errors for the low birth weight logistic regression example.

6.3 Application to Low Birth Weight Data

We now test the standard error approximations based on equations (19) and (20) on the low birth weight dataset of Hosmer and Lemeshow (1989). This dataset involves 189 observations and eight maternal predictors. The response is 0 or 1 according to whether an infant is born underweight, defined as less than 2.5 kilograms. The predictors include mother's age in years (AGE), weight at last menstrual period (LWT), race (RACE2 and RACE3), smoking status during pregnancy (SMOKE), number of previous premature labors (PTL), presence of hypertension history (HT), presence of uterine irritability (UI), and number of physician visits during the first trimester (FTV). Each of these predictors is quantitative except for race, which is a 3-level factor with level 1 for whites, level 2 for blacks, and level 3 for other races. Table 2 shows the maximum likelihood estimates and asymptotic standard errors for the 10 parameters. The differentiation increment δ_j was $\hat{\theta}_j/1000$ for each parameter θ_j. The standard error approximations in the two rightmost columns turn out to be the same in this example, but in other models they will differ. The close agreement of the approximations with the "gold standard" based on the exact value $\nabla^2\ell(\hat{\theta})$ is clearly good enough for practical purposes.

7 Handling Constraints

Many optimization problems impose constraints on parameters. For example, parameters are often required to be nonnegative. Here we discuss a majorization technique that in a sense eliminates inequality constraints. For this adaptive barrier method (Censor and Zenios, 1992; Lange, 1994) to work, an initial point $\theta^{(0)}$ must be selected with all inequality constraints strictly satisfied. The barrier method confines subsequent iterates to the interior of the parameter space but allows strict inequalities to become equalities in the limit.

Consider the problem of minimizing $f(\theta)$ subject to the constraints $v_j(\theta) \geq 0$ for $1 \leq j \leq q$, where each $v_j(\theta)$ is a concave, differentiable function. Since $-v_j(\theta)$ is convex, we know from inequality (7) that

$$v_j(\theta^{(m)}) - v_j(\theta) \quad \geq \quad \nabla v_j(\theta^{(m)})^t(\theta^{(m)} - \theta).$$

Application of the similar inequality $\ln s - \ln t \geq s^{-1}(t - s)$ implies that

$$v_j(\theta^{(m)}) \left[-\ln v_j(\theta) + \ln v_j(\theta^{(m)}) \right] \quad \geq \quad v_j(\theta^{(m)}) - v_j(\theta).$$

Adding the last two inequalities, we see that

$$v_j(\theta^{(m)})\left[-\ln v_j(\theta)+\ln v_j(\theta^{(m)})\right]+\nabla v_j(\theta^{(m)})^t(\theta-\theta^{(m)}) \quad \geq \quad 0,$$

with equality when $\theta = \theta^{(m)}$. Summing over j and multiplying by a positive tuning parameter ω, we construct the function

$$g(\theta \mid \theta^{(m)}) \quad = \quad f(\theta)+\omega\sum_{j=1}^{q}\left[v_j(\theta^{(m)})\ln\frac{v_j(\theta^{(m)})}{v_j(\theta)}+(\theta-\theta^{(m)})^t\nabla v_j(\theta^{(m)})\right] \tag{23}$$

majorizing $f(\theta)$ at $\theta^{(m)}$. The presence of the term $\ln v_j(\theta)$ in equation (23) prevents $v_j(\theta^{(m+1)}) \leq 0$ from occurring. The multiplier $v_j(\theta^{(m)})$ of $\ln v_j(\theta)$ gradually adapts and allows $v_j(\theta^{(m+1)})$ to tend to 0 if it is inclined to do so. When there are equality constraints $A\theta = b$ in addition to the inequality constraints $v_j(\theta) \geq 0$, these should be enforced during the minimization of $g(\theta \mid \theta^{(m)})$.

7.1 Multinomial Sampling

To gain a feel for how these ideas work in practice, consider the problem of maximum likelihood estimation given a random sample of size n from a multinomial distribution. If there are q categories and n_i observations fall in category i, then the loglikelihood reduces to $\sum_i n_i \ln \theta_i$ plus a constant. The components of the parameter vector θ satisfy $\theta_i \geq 0$ and $\sum_i \theta_i = 1$. Although it is well known that the maximum likelihood estimates are given by $\hat{\theta}_i = n_i/n$, this example is instructive because it is explicitly solvable and demonstrates the linear rate of convergence of the proposed MM algorithm.

To minimize the negative loglikelihood $f(\theta) = -\sum_i n_i \ln \theta_i$ subject to the q inequality constraints $v_i(\theta) = \theta_i \geq 0$ and the equality constraint $\sum_i \theta_i = 1$, we construct the majorizing function

$$g(\theta \mid \theta^{(m)}) \quad = \quad f(\theta) - \omega \sum_{i=1}^{q} \theta_i^{(m)} \ln \theta_i + \omega \sum_{i=1}^{q} \theta_i$$

suggested in equation (23), omitting irrelevant constants. We minimize $g(\theta \mid \theta^{(m)})$ while enforcing $\sum_i \theta_i = 1$ by introducing a Lagrange multiplier and looking for a stationary point of the Lagrangian

$$h(\theta) \quad = \quad g(\theta \mid \theta^{(m)}) + \lambda\Big(\sum_i \theta - 1\Big).$$

Setting $\partial h(\theta)/\partial \theta_i$ equal to zero and multiplying by θ_i gives

$$-n_i - \omega\theta_i^{(m)} + \omega\theta_i + \lambda\theta_i \quad = \quad 0.$$

Summing on i reveals that $\lambda = n$ and yields the update

$$\theta_i^{(m+1)} \quad = \quad \frac{n_i + \omega\theta^{(m)}}{n + \omega}.$$

Hence, all iterates have positive components if they start with positive components. The final rearrangement

$$\theta_i^{(m+1)} - \frac{n_i}{n} \quad = \quad \frac{\omega}{n+\omega}\left(\theta_i^{(m)} - \frac{n_i}{n}\right).$$

demonstrates that $\theta^{(m)}$ approaches the estimate $\hat{\theta}$ at the linear rate $\omega/(n+\omega)$, regardless of whether $\hat{\theta}$ occurs on the boundary of the parameter space where one or more of its components $\hat{\theta}_i$ equal zero.

8 Discussion

This article is meant to whet readers' appetites, not satiate them. We have omitted much. For instance, there is a great deal known about the convergence properties of MM algorithms that is too mathematically demanding to present here. Fortunately, almost all results from the EM algorithm literature (Wu, 1983; Lange, 1995a; McLachlan and Krishnan, 1997; Lange, 1999) carry over without change to MM algorithms. Furthermore, there are several methods for accelerating EM algorithms that are also applicable to accelerating MM algorithms (Heiser, 1995; Lange, 1995b; Jamshidian and Jennrich, 1997; Lange et al., 2000).

Although this survey article necessarily reports much that is already known, there are some new results here. Our MM treatment of constrained optimization in Section 7 is more general than previous versions in the literature (Censor and Zenios, 1992; Lange, 1994). The application of equation (20) to the estimation of standard errors in MM algorithms is new, as is the extension of the SEM idea of Meng and Rubin (1991) to the MM case.

There are so many examples of MM algorithms in the literature that we are unable to cite them all. Readers should be on the lookout for these and for known EM algorithms that can be explained more simply as MM algorithms. Even more importantly, we hope this article will stimulate readers to discover new MM algorithms.

References

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

C. C. J. H. Bijleveld and J. de Leeuw (1991), Fitting longitudinal reduced-rank regression models by alternating least squares, *Psychometrika*, **56**, 433–447.

D. Böhning and B. G. Lindsay (1988), Monotonicity of quadratic approximation algorithms, *Ann. Instit. Stat. Math.*, **40**, 641–663.

Y. Censor and S. A. Zenios (1992), Proximal minimization with D-functions, *J. Optimization Theory Appl.* **73**, 451–464.

J. de Leeuw (1994), Block relaxation algorithms in statistics, in *Information Systems and Data Analysis* (ed. H. H. Bock, W. Lenski, and M. M. Richter), pp. 308–325. Berlin: Springer-Verlag.

J. de Leeuw and W. J. Heiser (1977), Convergence of correction matrix algorithms for multidimensional scaling, in *Geometric Representations of Relational Data* (ed. J. C. Lingoes, E. Roskam, and I. Borg), pp. 735–752. Ann Arbor: Mathesis Press.

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

A. R. De Pierro (1995), A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography, *IEEE Trans. Med. Imaging*, **14**, 132–137.

P. J. F. Groenen (1993), *The Majorization Approach to Multidimensional Scaling: Some Problems and Extensions*, DSWO Press, Leiden, the Netherlands.

W. J. Heiser (1987), Correspondence analysis with least absolute residuals, *Comput. Stat. Data Analysis*, **5**, 337–356.

W. J. Heiser (1995), Convergent computing by iterative majorization: theory and applications in multidimensional data analysis, in *Recent Advances in Descrip-*

tive Multivariate Analysis (ed. W. J. Krzanowski), pp. 157–189. Oxford: Clarendon Press.

D. W. Hosmer and S. Lemeshow (1989), *Applied Logistic Regression*, Wiley, New York.

P. J. Huber (1981), *Robust Statistics*, Wiley, New York.

D. R. Hunter (2004), MM algorithms for generalized Bradley-Terry models, *Annals Stat.*, to appear.

D. R. Hunter and K. Lange (2000a), Rejoinder to discussion of "Optimization transfer using surrogate objective functions", *J. Comput. Graphical Stat.* **9**, 52–59.

D. R. Hunter and K. Lange (2000b), Quantile regression via an MM algorithm, *J. Comput. Graphical Stat.* **9**, 60–77.

D. R. Hunter and K. Lange (2002), Computing estimates in the proportional odds model, *Ann. Inst. Stat. Math.* **54**, 155–168.

D. R. Hunter and R. Li (2002), A connection between variable selection and EM-type algorithms, Pennsylvania State University statistics department technical report 0201.

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

H. A. L. Kiers and J. M. F. Ten Berge (1992), Minimization of a class of matrix trace functions by means of refined majorization, *Psychometrika*, **57**, 371–382.

R. Koenker and G. Bassett (1978), Regression quantiles, *Econometrica*, **46**, 33–50.

K. Lange (1994), An adaptive barrier method for convex programming, *Methods Applications Analysis*, **1**, 392–402.

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

K. Lange (1995b), A quasi-Newton acceleration of the EM algorithm, *Statistica Sinica*, **5**, 1–18.

K. Lange (1999), *Numerical Analysis for Statisticians*, Springer-Verlag, New York.

K. Lange and J. A. Fessler (1995), Globally convergent algorithms for maximum a posteriori transmission tomography, *IEEE Trans. Image Processing*, **4**, 1430–1438.

K. Lange, D. R. Hunter, and I. Yang (2000), Optimization transfer using surrogate objective functions (with discussion), *J. Comput. Graphical Stat.* **9**, 1–20.

K. Lange and J. S. Sinsheimer (1993) Normal/independent distributions and their applications in robust regression. *J. Comput. Graphical Stat.* **2**, 175-198

D. G. Luenberger (1984), *Linear and Nonlinear Programming, 2nd ed.*, Addison-Wesley, Reading, MA.

M. J. Maher (1982), Modelling association football scores, *Statistica Neerlandica*, **36**: 109–118.

A. W. Marshall and I. Olkin (1979), *Inequalities: Theory of Majorization and its Applications*, Academic, San Diego.

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

X-L Meng and D. B. Rubin (1991), Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm, *J. Amer. Stat. Assoc.*, **86**, 899–909.

X-L Meng and D. B. Rubin (1993), Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika*, **80**, 267–278.

D. Oakes (1999), Direct calculation of the information matrix via the EM algorithm, *J. Roy. Stat. Soc. B*, **61**, Part 2, 479–482.

J. M. Ortega and W. C. Rheinboldt (1970), *Iterative Solutions of Nonlinear Equations in Several Variables*, Academic, New York, pp. 253–255.

C. Sabatti and K. Lange (2002), Genomewide motif identification using a dictionary model, *Proceedings IEEE* **90**, 1803–1810.

E. J. Schlossmacher (1973), An iterative technique for absolute deviations curve fitting, *J. Amer. Stat. Assoc.*, **68**, 857–859.

C. F. J. Wu (1983). On the convergence properties of the EM algorithm, *Annals Stat.*, **11**, 95–103.