Statistica Sinica **5**(1995), 1-18

A QUASI-NEWTON ACCELERATION OF THE EM ALGORITHM

Kenneth Lange

University of Michigan

Abstract. The EM algorithm is one of the most commonly used methods of maximum likelihood estimation. In many practical applications, it converges at a frustratingly slow linear rate. The current paper considers an acceleration of the EM algorithm based on classical quasi-Newton optimization techniques. This acceleration seeks to steer the EM algorithm gradually toward the Newton-Raphson algorithm, which has a quadratic rate of convergence. The fundamental difference between the current algorithm and a naive quasi-Newton algorithm is that the early stages of the current algorithm resemble the EM algorithm rather than steepest ascent. Numerical examples involving the Dirichlet distribution, a mixture of Poisson distributions, and a repeated measures model illustrate the potential of the current algorithm.

Key words and phrases: Dirichlet distribution, maximum likelihood, repeated measures model, secant approximation.

1. Introduction

The EM algorithm (Dempster et al. (1977)) is one of the pillars of modern computational statistics. Numerical stability and computational simplicity alike recommend it. Unfortunately, the EM algorithm can suffer from extremely slow convergence in problems with sizable amounts of missing data. This defect has prompted a number of suggestions for accelerating the algorithm. For instance, Louis (1982) advocates classical Aitken acceleration. This tactic is useful for problems with a moderate number of parameters, but the required matrix inversions render it less effective for problems with a large number of parameters. Furthermore, Aitken acceleration can only improve the linear rate of convergence of the underlying EM algorithm. One should aim for superlinear convergence. Accordingly, Jamshidian and Jennrich (1993) have recently recommended a conjugate gradient version of the EM algorithm. This hybrid algorithm shows great promise. It operates by using the EM algorithm steps as generalized gradients in a conjugate gradient search.

The present paper takes a different tack. In modern optimization theory, quasi-Newton methods are the principal competitors with conjugate gradient methods. Quasi-Newton methods are predicated on the philosophy that all fast algorithms should attempt to approximate the Newton-Raphson algorithm. At the same time they should avoid the faults of Newton-Raphson. These faults include explicit evaluation of the Hessian of the objective function and the tendency to head toward saddle points and local minima as often as toward local maxima. Quasi-Newton algorithms circumvent the first fault by gradually constructing an approximate Hessian from the gradient of the objective function evaluated at the successive points encountered by the algorithm. They circumvent the second fault by forcing the approximate Hessian to always be negative definite.

Even with these safeguards, quasi-Newton algorithms are less than ideal in many statistical applications. One of their least desirable features is that they typically start by approximating the Hessian by the identity matrix. This initial approximation may be poorly scaled to the problem at hand. Hence, the algorithms can wildly overshoot or undershoot the maximum of the objective function along the direction of the current step. It can take many iterations before a decently scaled approximation is built up.

For this reason other algorithms such as Fisher scoring and the Gauss-Newton algorithm of nonlinear least squares remain popular in statistics. The experience of numerical analysts in addressing the defects of the Gauss-Newton algorithm are particularly instructive. In nonlinear least squares one minimizes

$$f(\theta) = \frac{1}{2}\sum_{i=1}^{n}[y_i - \mu_i(\theta)]^2$$

over the parameter vector θ. The Gauss-Newton algorithm approximates the Hessian

$$d^2 f(\theta) = \sum_{i=1}^{n} d\mu_i(\theta) d\mu_i(\theta)^t - \sum_{i=1}^{n}[y_i - \mu_i(\theta)] d^2\mu_i(\theta) \tag{1}$$

by retaining the first sum; namely,

$$d^2 f(\theta) \approx \sum_{i=1}^{n} d\mu_i(\theta) d\mu_i(\theta)^t. \tag{2}$$

On problems with small residuals $y_i - \mu_i(\theta)$, Gauss-Newton is very close to Newton-Raphson. On large residual problems, the divergence between the two algorithms is greater. In any case the approximate Hessian (2) is positive definite, and Gauss-Newton leads to a well-behaved descent algorithm. In seeking to improve the Gauss-Newton algorithm, one can retain the first sum exactly as given in (1) and attempt to approximate the second sum. This successful strategy is described in the survey article by Nazareth (1980).

The current paper carries out an analogous strategy for the EM algorithm by decomposing the observed information matrix into an EM part and a remainder. The remainder is approximated in typical quasi-Newton fashion using a sequence of secant conditions. The early stages of the new algorithm resemble a close approximation to the EM algorithm known as the EM gradient algorithm (Lange (1994)). Later stages approximate Newton-Raphson, and intermediate stages make a graceful transition between these two extremes. Thus, the new algorithm appears to combine the stability and early rapid progress of EM with the later superlinear convergence of Newton-Raphson. In the next section we motivate and derive the algorithm. This general overview is followed by a concise summary of the algorithm and then its application to several numerical examples. The paper concludes with a discussion of the merits and extensions of the algorithm.

2. Derivation of the Algorithm

Let us begin by briefly reviewing the EM algorithm (Dempster et al. (1977), Little and Rubin (1987)). Underlying the observed data Y is the complete data X; some function $t(X) = Y$ relates the two. Although the statistician has control over how the complete data are defined, the sensible procedure is to chose X so that it is trivial to estimate by maximum likelihood the parameters θ of a model explaining X and hence Y. The complete data X is postulated to have probability density $f(X \mid \theta)$ with respect to some fixed measure. In the E step of the EM algorithm, the conditional expectation

$$Q(\theta \mid \theta^n) = E(\ln[f(X \mid \theta)] \mid Y, \theta^n)$$

is computed. Here θ^n is the current estimated value of θ. (The superscript n will always refer to iteration number in the sequel.) In the M step, the θ maximizing $Q(\theta \mid \theta^n)$ is found. This yields the new parameter estimate θ^{n+1}, and this two step procedure is repeated until convergence occurs. The essence of the EM algorithm is a maximization transfer principle; maximizing $Q(\theta \mid \theta^n)$ with respect to its left entry θ forces an increase in the loglikelihood $L(\theta)$ of the observed data Y. This property is a consequence of a well-known information theory inequality (Rao (1973)).

Corresponding to the function $Q(\theta \mid \theta^n)$ is the slightly mysterious decomposition

$$L(\theta) = Q(\theta \mid \theta^n) - H(\theta \mid \theta^n) \tag{3}$$

of the loglikelihood $L(\theta)$ (Dempster et al. (1977)). Equation (3) leads immediately to the further decomposition

$$-d^2L(\theta) = -d^{20}Q(\theta \mid \theta^n) + d^{20}H(\theta \mid \theta^n) \tag{4}$$

of the observed information matrix. In Equation (4) the operator d^{20} takes second partials with respect to the θ variables of $Q(\theta \mid \theta^n)$ and $H(\theta \mid \theta^n)$. By analogy with the Gauss-Newton approximation (2), the crude approximation

$$d^2L(\theta) \approx d^{20}Q(\theta \mid \theta^n) \tag{5}$$

gives rise to the EM gradient algorithm (Lange (1993))

$$\begin{aligned} \theta^{n+1} &= \theta^n - d^{20}Q(\theta^n \mid \theta^n)^{-1} dL(\theta^n) \\ &= \theta^n - d^{20}Q(\theta^n \mid \theta^n)^{-1} d^{10}Q(\theta^n \mid \theta^n). \end{aligned} \tag{6}$$

This algorithm substitutes one step of Newton-Raphson on $Q(\theta \mid \theta^n)$ for the M step of the EM algorithm. Because $L(\theta) - Q(\theta \mid \theta^n)$ has it minimum at $\theta = \theta^n$, the score equality

$$dL(\theta) = d^{10}Q(\theta \mid \theta) \tag{7}$$

holds at $\theta = \theta^n$ whenever θ^n is an interior point of the parameter feasible region. This accounts for the second line of (6).

The EM gradient algorithm (6) is interesting in its own right. It avoids explicit solution of the M step of the EM algorithm while preserving the local convergence properties of the EM algorithm (Lange (1993)). Because the EM gradient algorithm so closely resembles the EM algorithm, it also tends to share the desirable stability properties of the EM algorithm. Our goal, however, is to improve the convergence rate of the EM algorithm. Thus, we need to amend the EM gradient algorithm so that the missing piece $d^{20}H(\theta^n \mid \theta^n)$ of the observed information matrix is approximated. In other words we need to perform approximate Newton-Raphson for maximizing $L(\theta)$ instead of exact Newton-Raphson for maximizing $Q(\theta \mid \theta^n)$.

At one point our analogy with amendment of the Gauss-Newton algorithm potentially breaks down. The matrix approximation (5) may not be negative definite, and consequently the EM gradient algorithm may not be an ascent algorithm. In practice, $d^{20}Q(\theta^n \mid \theta^n)$ is almost always negative definite or can be rendered so by a change of variables. If the complete data X belong to a regular exponential family, then $-d^{20}Q(\theta^n \mid \theta^n)$ can be identified with the expected information of the complete data. In this instance, the EM gradient algorithm coincides with the earlier ascent algorithm of Titterington (1984). When the complete data do not belong to a regular exponential family, often $d^{20}Q(\theta^n \mid \theta^n)$ is diagonal, and negative definiteness is simple to check. Our examples illustrate these points, as well as the ease with which $d^{20}Q(\theta^n \mid \theta^n)$ can be evaluated. Certainly, $d^{20}Q(\theta^n \mid \theta^n)$ is typically much easier to evaluate than $d^2L(\theta^n)$.

Returning to our main concern, let B^n be the current approximation to the Hessian $d^{20}H(\theta^n \mid \theta^n)$. The natural replacement for the EM gradient algorithm (6) is

$$\theta^{n+1} = \theta^n - [d^{20}Q(\theta^n \mid \theta^n) - B^n]^{-1} d^{10}Q(\theta^n \mid \theta^n). \tag{8}$$

Three important questions now arise. First, what initial value should B^1 assume? A good choice is the zero matrix $B^1 = \mathbf{0}$. Indeed with this choice, the first step of the algorithm (8) is an EM gradient step.

Second, how can we update B^n to produce a better approximation to $d^{20}H(\theta^n \mid \theta^n)$? Experience with standard quasi-Newton methods suggests the importance of the secant condition

$$d^{10}H(\theta^{n-1} \mid \theta^n) - d^{10}H(\theta^n \mid \theta^n) \approx d^{20}H(\theta^n \mid \theta^n)(\theta^{n-1} - \theta^n).$$

The secant condition (9) is a first order Taylor's expansion that can be incorporated into the overall approximation process by requiring that B^n be a small rank perturbation of B^{n-1} satisfying

$$\begin{aligned} B^n s^n &= g^n, \\ s^n &= \theta^{n-1} - \theta^n, \\ g^n &= d^{10}H(\theta^{n-1} \mid \theta^n) - d^{10}H(\theta^n \mid \theta^n). \end{aligned} \tag{9}$$

Davidon's (1959) symmetric, rank-one update is defined by

$$B^n = B^{n-1} + c^n v^n (v^n)^t, \tag{10}$$

with constant c^n and vector v^n specified as

$$\begin{aligned} c^n &= \frac{1}{(g^n - B^{n-1}s^n)^t s^n}, \\ v^n &= g^n - B^{n-1}s^n. \end{aligned} \tag{11}$$

It is straightforward to check that Davidon's update is the unique symmetric, rank-one update satisfying Condition (9).

This simple rank-one update is more parsimonious than standard rank-two updates. Empirical evidence also suggests that it provides an approximate Hessian superior to the Davidon-Fletcher-Powell (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS) symmetric rank-two updates (Conn et al. (1991)). At first glance it appears desirable that B^n be kept negative definite since $H(\theta \mid \theta^n)$ attains its maximum at $\theta = \theta^n$. This requirement cannot be maintained by Davidon's update but is possible with the DFP and BFGS updates. However, for the algorithm defined by (8) to be an ascent algorithm, it is far more important that the difference $d^{20}Q(\theta^n \mid \theta^n) - B^n$ be kept negative definite. Another concern is that the constant c^n is undefined when the inner product $(g^n - B^{n-1}s^n)^t s^n = 0$. In such situations or when $(g^n - B^{n-1}s^n)^t s^n$ is small compared to $\|g^n - B^{n-1}s^n\| \cdot \|s^n\|$, we simply take $B^n = B^{n-1}$.

As noted above, one can anticipate problems with the algorithm (8) when the matrix $d^{20}Q(\theta^n \mid \theta^n) - B^n$ fails to be negative definite. If this is the case, we replace (8) by

$$\theta^{n+1} = \theta^n - [d^{20}Q(\theta^n \mid \theta^n) - (\frac{1}{2})^m B^n]^{-1} d^{10}Q(\theta^n \mid \theta^n), \tag{12}$$

where m is the smallest positive integer making $d^{20}Q(\theta^n \mid \theta^n) - (1/2)^m B^n$ negative definite. This tactic tends to preserve as much of the approximate information contained in B^n as is consistent with the algorithm (12) being an ascent algorithm. The particular form of the factor $(1/2)^m$ is merely a convenient one. Observe that it is straightforward to check the negative definiteness of the matrix $d^{20}Q(\theta^n \mid \theta^n) - (1/2)^m B^n$ in the process of inverting it. If this matrix is inverted by sweeping, then each diagonal pivot encountered must be negative just prior to being swept (Little and Rubin (1987), Thisted (1988)).

It is noteworthy that almost all relevant quantities for the above algorithm can be computed in terms of $Q(\theta \mid \theta^n)$ and its derivatives. In particular, the difference of gradients in (9) can be rewritten as

$$\begin{aligned} g^n &= d^{10}H(\theta^{n-1} \mid \theta^n) - d^{10}H(\theta^n \mid \theta^n) \\ &= d^{10}Q(\theta^{n-1} \mid \theta^n) - d^{10}Q(\theta^{n-1} \mid \theta^{n-1}) \end{aligned} \tag{13}$$

using the fact $d^{10}H(\theta^n \mid \theta^n) = 0$ and the identity (7) at $\theta = \theta^{n-1}$.

The only relevant quantity not expressible in terms of $Q(\theta \mid \theta^n)$ or its derivatives is the loglikelihood $L(\theta)$, which is useful in monitoring the progress of the algorithm. If the algorithm defined by (8) and (12) overshoots at any given iteration, then some form of step-decrementing will guarantee an increase in $L(\theta)$. In practice, we follow Powell's (1978) suggestion and fit a quadratic to the function $r \to L(\theta^n + r[\theta^{n+1} - \theta^n])$ through the values $L(\theta^{n+1})$ and $L(\theta^n)$ with slope $dL(\theta^n)^t(\theta^{n+1} - \theta^n)$ at $r = 0$. If the maximum of the quadratic occurs at r_{max}, then we step back to $r = \max\{r_{max}, 0.1\}$. If this procedure still does not yield an increase in $L(\theta)$, then it can be repeated. In practice, one or two step decrements invariably give the desired increase in $L(\theta)$.

3. Recapitulation of the Algorithm

For the sake of clarity, let us now summarize how the algorithm is implemented. The basic idea is to employ Equation (8) with appropriate safeguards designed to force an increase in the loglikelihood $L(\theta)$. In the update (8), θ^n is the current iterate and θ^{n+1} is the next iterate. The function $Q(\theta \mid \theta^n)$ is the standard function associated with the E step of the EM algorithm. The matrix B^n approximates the second differential $d^{20}H(\theta^n \mid \theta^n)$ of the function $H(\theta^n \mid \theta^n)$ defined in Equation (3). If this approximation is good, then the update (8) is close to the Newton-Raphson update for maximizing $L(\theta)$. From the initial value $B^1 = \mathbf{0}$, the matrix B^n is updated via Equation (10) using the quantities defined by equations (9), (11) and (13). If the inner product $(g^n - B^{n-1}s^n)^t s^n$ is small compared to $\|g^n - B^{n-1}s^n\| \cdot \|s^n\|$, then the update (10) is omitted and $B^n = B^{n-1}$ is used instead.

The safeguards for the algorithm occur at two levels. To insure that the next iterate extends in an uphill direction from the current iterate, the usual update (8) is replaced by the modified update (12) when the matrix difference $d^{20}Q(\theta^n \mid \theta^n) - B^n$ fails to be negative definite. If $d^{20}Q(\theta^n \mid \theta^n)$ is negative definite and the power m occurring in (12) is sufficiently large, then the modified matrix $d^{20}Q(\theta^n \mid \theta^n) - (1/2)^m B^n$ is certain to be negative definite as well. We choose m to be the smallest nonnegative integer such that the modified matrix passes the negative definiteness test. Determination of whether the matrix $d^{20}Q(\theta^n \mid \theta^n) - (1/2)^m B^n$ is negative definite can be made in the process of inverting it by sweeping.

The second level of safeguards involves checking whether the proposed step (12) actually leads to an increase in $L(\theta)$. If this is not the case, some form of step decrementing must be instituted. Because the algorithm moves locally uphill, step decrementing is bound to succeed. We have suggested fitting a quadratic to the loglikelihood curve between θ^n and the proposed θ^{n+1}. Probably, a simple step-halving strategy would be equally effective.

4. Examples

4.1. Dirichlet distribution

The Dirichlet distribution is useful for modeling data on proportions (Kingman (1993)). Let $X_1, \ldots, X_k$ be independent random variables such that X_i has gamma density $x_i^{\theta_i - 1} e^{-x_i} \Gamma(\theta_i)^{-1}$, $x_i > 0$. A Dirichlet random vector $Y = (Y_1, \ldots, Y_k)^t$ is defined by setting its ith component equal to the proportion $Y_i = X_i (\sum_{j=1}^k X_j)^{-1}$. It can be shown that Y has regular exponential density

$$\frac{\Gamma(\sum_{i=1}^k \theta_i)}{\prod_{i=1}^k \Gamma(\theta_i)} \prod_{i=1}^k y_i^{\theta_i - 1}$$

on the simplex $\{y = (y_1, \ldots, y_k)^t : y_1 > 0, \ldots, y_k > 0, \sum_{i=1}^k y_i = 1\}$ endowed with the uniform measure. The random vector Y constitutes the observed data, and the underlying random vector $X = (X_1, \dots, X_k)^t$ constitutes the complete data.

For an i.i.d. sample $Y^1 = y^1, \dots, Y^m = y^m$ from the Dirichlet distribution, it is tempting to estimate the parameter vector $\theta = (\theta_1, \dots, \theta_k)^t$ by the EM algorithm. Let $X^1, \dots, X^m$ be the corresponding complete data. It is immediately evident that

$$\begin{aligned} Q(\theta \mid \theta^n) = &- m\sum_{i=1}^{k} \ln\Gamma(\theta_i) + \sum_{i=1}^{k}(\theta_i - 1)\sum_{j=1}^{m} E(\ln[X_i^j] \mid Y^j = y^j, \theta^n) \\ &- \sum_{i=1}^{k}\sum_{j=1}^{m} E(X_i^j \mid Y^j = y^j, \theta^n). \end{aligned} \tag{14}$$

Owing to the presence of the terms $\ln\Gamma(\theta_i)$ in (14), one cannot solve the M step analytically. However, the EM gradient algorithm is trivial to implement since the score vector is easily computed and the Hessian matrix $d^{20}Q(\theta \mid \theta^n)$ is diagonal with ith diagonal entry $-m\frac{d^2}{d\theta_i^2}\ln\Gamma(\theta_i)$, which is negative because $\ln\Gamma(r)$ is strictly convex. The gradient $d^{10}Q(\theta^{n-1} \mid \theta^n)$ is a little trickier to evaluate for accelerated EM since the conditional expectations $E(\ln[X_i^j] \mid Y^j = y^j, \theta^n)$ can no longer be ignored. However, because the identity (7) holds at $\theta = \theta^n$, Equation (14) implies

$$\begin{aligned} \sum_{j=1}^{m} E(\ln[X_i^j] \mid Y^j = y^j, \theta^n) &= \frac{\partial}{\partial\theta_i}L(\theta^n) + m\frac{\partial}{\partial\theta_i}\ln\Gamma(\theta_i^n) \\ &= m\frac{\partial}{\partial\theta_i}\ln\Gamma(\sum_{l=1}^{k}\theta_l^n) + \sum_{j=1}^{m}\ln y_i^j. \end{aligned}$$

Scoring is an attractive alternative to the EM gradient algorithm in this particular example (Narayanan (1991)). The data of Mosimann (1962) on the relative frequencies of $k = 3$ serum proteins in $m = 23$ young, white Pekin ducklings furnish an interesting test case for comparing the EM gradient algorithm, our accelerated EM algorithm, and scoring. Starting from $\theta^1 = (1., 1., 1.)^t$, all three algorithms converge smoothly to the maximum point $(3.22, 20.38, 21.69)^t$, with the loglikelihood showing a steady increase to its maximum value of 73.1250 along the way. The EM gradient algorithm takes 287 iterations for the loglikelihood to achieve this final value, while the accelerated EM algorithm and scoring take 8 and 9 iterations, respectively.

Table 1. Performance of accelerated EM on the Dirichlet distribution data of Mosimann (1962).

Iteration	Extra Steps	Exponent	$L(\theta)$	θ_1	θ_2	θ_3
1	0	0	15.9424	1.000	1.000	1.000
2	0	0	24.7300	.2113	1.418	1.457
3	0	0	41.3402	.3897	2.650	2.760
4	0	1	49.1425	.6143	3.271	3.445
5	0	1	53.3627	.8222	3.827	4.045
6	0	0	73.0122	3.368	22.19	23.59
7	0	2	73.0524	3.445	22.05	23.47
8	0	0	73.1250	3.217	20.40	21.70
9	0	0	73.1250	3.217	20.39	21.69
10	0	0	73.1250	3.215	20.38	21.69

Table 1 records the behavior of the accelerated EM algorithm on this problem. The 'exponent' column of Table 1 refers to the minimum nonnegative integer m required to make $d^{20}Q(\theta^n \mid \theta^n) - (1/2)^m B^n$ negative definite. The 'extra steps' column refers to the number of step decrements taken in order to produce an increase in $L(\theta)$ at a given iteration. In this problem step decrementing was never necessary. All computations were done using a special version of the author's Fortran optimization program SEARCH (Lange (1994)).

4.2. Mixture of Poissons

Hasselblad (1969) and Titterington et al. (1985) consider mortality data from the *London Times* newspaper during the three years 1910-1912. The observed data consist of the number of days $Y_i = y_i$ on which there were i death notices for women aged 80 years and over, $i = 0, \ldots, 9$. A single Poisson distribution fits these data poorly, but a mixture of two Poisson distributions gives an acceptable fit. In the mixture model the complete data can be viewed as (Y_i, Z_i), where Z_i is the number of days out of the Y_i days belonging to population 1. Evidently, the complete data likelihood is proportional to

$$\prod_i \left[\pi(\mu_1)^i e^{-\mu_1}\right]^{Z_i} \left[(1-\pi)(\mu_2)^i e^{-\mu_2}\right]^{Y_i - Z_i}.$$

Here μ_1 and μ_2 are the means for populations 1 and 2, respectively, and π is the admixture proportion for population 1.

The usual application of Bayes' theorem implies

$$Q(\theta \mid \theta^n) = \sum_i w_i(\ln \pi + i \ln \mu_1 - \mu_1) + \sum_i (y_i - w_i)[\ln(1-\pi) + i \ln \mu_2 - \mu_2],$$

where $\theta = (\mu_1, \mu_2, \pi)^t$ and where

$$w_i = E(Z_i \mid Y_i = y_i, \theta^n) = y_i \frac{\pi^n (\mu_1^n)^i e^{-\mu_1^n}}{\pi^n (\mu_1^n)^i e^{-\mu_1^n} + (1-\pi^n)(\mu_2^n)^i e^{-\mu_2^n}}.$$

Clearly in this example the Hessian $d^{20}Q(\theta \mid \theta^n)$ is diagonal and negative definite.

Starting from the moment estimates $(\mu_1^1, \mu_2^1, \pi^1) = (1.101, 2.582, .2870)$ (Hasselblad (1969)), the EM gradient algorithm takes an excruciating 535 iterations for the loglikelihood to attain its maximum of -1989.946. Even worse it takes 1749 iterations for the parameters to reach the maximum likelihood estimates $(\hat{\mu}_1, \hat{\mu}_2, \hat{\pi}) = (1.256, 2.663, .3599)$. The sizable difference in convergence rates to the maximum loglikelihood and the maximum likelihood estimates indicates that the likelihood surface is quite flat. In contrast, the accelerated EM algorithm converges to the maximum loglikelihood in 11 iterations and to the maximum likelihood estimates in 16 iterations. Table 2 charts the progress of the accelerated EM algorithm on this problem. Titterington et al. (1985) report that Newton-Raphson typically takes 8 to 11 iterations to converge when it converges for these data. For about a third of their initial points, Newton-Raphson fails.

Table 2. Performance of accelerated EM on the Poisson mixture data of Hasselblad (1969).

Iteration	Extra Steps	Exponent	$L(\theta)$	μ_1	μ_2	π
1	0	0	-1990.038	1.101	2.582	.2870
2	0	0	-1990.033	1.105	2.580	.2870
3	0	0	-1990.024	1.119	2.576	.2876
4	0	0	-1990.018	1.127	2.579	.2905
5	0	1	-1990.016	1.127	2.580	.2913
6	0	0	-1989.971	1.297	2.677	.3723
7	0	1	-1989.951	1.286	2.676	.3734
8	0	1	-1989.949	1.283	2.678	.3735
9	0	1	-1989.949	1.282	2.679	.3734
10	0	1	-1989.948	1.280	2.679	.3734
12	0	0	-1989.946	1.250	2.659	.3562
14	1	0	-1989.946	1.256	2.664	.3600
16	0	0	-1989.946	1.256	2.663	.3599

4.3. Repeated measures models

The classical random effects, repeated measures model postulates r-variate observations Y of the form

$$Y = A\beta + BU + V, \tag{15}$$

where A and B are known constant matrices, β is a parameter vector of fixed effects, U is an s-variate vector of random effects, and V is an r-variate vector of random errors. The random vectors U and V are assumed to be independent and normally distributed with $\mathbf{0}$ means and covariance matrices Δ and $\sigma^2 I$, respectively. Although many applications dictate additional structure on Δ (Jennrich and Schluchter (1986)), for our purposes it is more interesting to suppose that Δ is unstructured. Collectively, β, σ^2, and the lower triangular entries of Δ constitute the parameters θ of the model. These parameters are also appropriate for a random sample $Y^1 = y^1, \ldots, Y^m = y^m$, in which the number of components r^i and the constant matrices A^i and B^i vary from case to case.

Jamshidian and Jennrich (1993) specify the EM algorithm for this repeated measures model and compare it to the modified EM algorithm of Laird and Ware (1982). In performing the E step of the algorithm, it is convenient to identify the complete data for case i as the pair U^i and $W^i = A^i\beta + V^i$. Under this identification,

$$\begin{aligned}
Q(\theta \mid \theta^n) &= \sum_{i=1}^m Q_i(\theta \mid \theta^n) \\
&= -\frac{m}{2}\ln|\Delta| - \frac{\ln\sigma^2}{2}\sum_{i=1}^m r^i \\
&\quad -\frac{1}{2}\sum_{i=1}^m E([U^i]^t\Delta^{-1}U^i \mid Y^i = y^i, \theta^n) \\
&\quad -\frac{1}{2\sigma^2}\sum_{i=1}^m E([W^i - A^i\beta]^t[W^i - A^i\beta] \mid Y^i = y^i, \theta^n). \qquad (16)
\end{aligned}$$

To evaluate the conditional expectations in (16), recall that for a random vector Z with mean μ and covariance Σ, the quadratic form Z^tFZ has expectation

$$E(Z^tFZ) = \mathrm{tr}(F\Sigma) + \mu^tF\mu = \mathrm{tr}(F[\Sigma + \mu\mu^t]).$$

It follows that

$$\begin{aligned}
Q(\theta \mid \theta^n) = &-\frac{m}{2}\ln|\Delta| - \frac{\ln\sigma^2}{2}\sum_{i=1}^m r^i - \frac{1}{2}\mathrm{tr}(\Delta^{-1}C) - \frac{1}{2\sigma^2}\mathrm{tr}(D) \\
&-\frac{1}{2\sigma^2}\sum_{i=1}^m (e^i - A^i\beta)^t(e^i - A^i\beta),
\end{aligned}$$

where

$$\begin{aligned} C &= \sum_{i=1}^{m} \operatorname{Var}(U^i \mid Y^i = y^i, \theta^n) + \sum_{i=1}^{m} E(U^i \mid Y^i = y^i, \theta^n) E(U^i \mid Y^i = y^i, \theta^n)^t, \\ D &= \sum_{i=1}^{m} \operatorname{Var}(W^i \mid Y^i = y^i, \theta^n), \\ e_i &= E(W^i \mid Y^i = y^i, \theta^n). \end{aligned}$$

The above conditional expectations and covariances are well known (Rao (1973)).

One can make $Q(\theta \mid \theta^n)$ concave by reparameterizing. Following Burridge (1981) and Pratt (1981), define $\alpha = \beta/\sigma$ and $\omega = \sigma^{-1}$. The obvious completion of this partial reparameterization is to let Γ be the lower triangular Cholesky decomposition of Δ^{-1} (Horn and Johnson (1985)). Since $\Gamma\Gamma^t = \Delta^{-1}$, these substitutions yield

$$\begin{aligned} Q(\theta \mid \theta^n) = {} & m \sum_{k=1}^{s} \ln \Gamma_{kk} + \ln \omega \sum_{i=1}^{m} r^i - \frac{1}{2} \operatorname{tr}(\Gamma^t C \Gamma) - \frac{\omega^2}{2} \operatorname{tr}(D) \\ & - \frac{1}{2} \sum_{i=1}^{m} (\omega e^i - A^i \alpha)^t (\omega e^i - A^i \alpha), \end{aligned} \tag{17}$$

where $\Gamma_{kk} > 0$ is the kth diagonal element of Γ.

It is easy to check that all terms of $Q(\theta \mid \theta^n)$ in (13) are concave except possibly for $-\frac{1}{2}\operatorname{tr}(\Gamma^t C \Gamma)$. To verify concavity for this term, note that $\operatorname{tr}(G^t C H)$ defines an inner product on $s \times s$ matrices G and H. Here it is crucial that C be positive definite. Because the function x^2 is convex, the corresponding matrix norm $\| \cdot \|_C$ satisfies

$$\begin{aligned} \|\lambda \Gamma_1 + (1-\lambda)\Gamma_2\|_C^2 &\le (\lambda \|\Gamma_1\|_C + (1-\lambda)\|\Gamma_2\|_C)^2 \\ &\le \lambda \|\Gamma_1\|_C^2 + (1-\lambda)\|\Gamma_2\|_C^2 \end{aligned} \tag{18}$$

for all $0 \le \lambda \le 1$. Equality holds in the first inequality of (18) if and only if Γ_1 is a constant multiple of Γ_2. Equality then holds in the second inequality of (18) only if the constant multiplier is 1. Hence, strict inequality holds unless $\Gamma_1 = \Gamma_2$.

Calculation of the gradient and Hessian of $Q(\theta \mid \theta^n)$ is straightforward. Partial derivatives with respect to the entries of Γ can be simplified by employing the identities

$$\begin{aligned} \operatorname{tr}(E_{jk}^t G) &= G_{jk}, \\ \operatorname{tr}(G E_{jk}) &= G_{kj}, \\ \operatorname{tr}(E_{jk}^t G E_{mn}) &= G_{jm} \delta_{kn}, \end{aligned}$$

where the matrix E_{jk} has all entries 0 except for a 1 in row j and column k, and where δ_{kn} is Kronecker's delta.

The classical growth curve data of Potthoff and Roy (1964) furnish an interesting test case for the accelerated EM algorithm. These data record the distances measured from the center of the pituitary to the pterygomaxillary fissure on 11 girls and 16 boys at ages 8, 10, 12, and 14. A simple model ignoring the obvious sex differences in the data can be constructed by taking the same design matrices A^i and B^i in (15)

$$A^i = B^i \ = \begin{pmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{pmatrix}$$

for all 27 cases.

Because of the symmetry of this model, the initial values suggested by Cooke (Laird et al. (1987)) coincide with the maximum likelihood estimates of 12.79, .5039, and .7633 for α_1, α_2, and ω, respectively. Starting with these values and the identity matrix for Γ, the accelerated EM algorithm converges to the maximum loglikelihood of -120.3604 in 17 iterations. The algorithm takes 19 iterations to converge to the maximum likelihood estimate of Γ given in Table 3. In contrast, the EM gradient algorithm takes 144 iterations to attain the maximum loglikelihood and 213 iterations to reach the maximum likelihood estimate of Γ.

Table 3. Performance of accelerated EM on the growth curve data of Potthoff and Roy (1964).

Iteration	Extra Steps	Exponent	$L(\theta)$	Γ_{11}	Γ_{21}	Γ_{22}
1	0	0	-152.5679	1.000	0.000	1.000
2	0	0	-136.1852	.9893	.1384	1.921
3	0	0	-125.2459	.9748	.2489	3.339
4	0	0	-121.4986	.9548	.3303	4.696
5	0	0	-121.0150	.9302	.3326	5.322
6	0	0	-120.8511	.8853	.2573	5.570
7	1	0	-120.7068	.8269	.1598	5.741
8	1	0	-120.6306	.7626	.0817	6.030
9	0	1	-120.5940	.7429	.1235	6.248
11	0	0	-120.5446	.6533	.1233	6.497
13	0	0	-120.3783	.4857	.3082	5.894
15	0	0	-120.3615	.4653	.3237	5.722
17	0	0	-120.3604	.4559	.3255	5.719
19	0	0	-120.3604	.4558	.3258	5.719

5. Discussion

Two problems beset the EM algorithm. First, either the E step or the M step may not be explicitly computable. For instance, with a structured covariance matrix Δ in the repeated measures model, it is often impossible to solve the M step exactly. The EM gradient algorithm circumvents difficulties in carrying out the M step. Second, the common rate of convergence of the EM and EM gradient algorithms can be painfully low. Our three examples illustrate the striking improvements in computational speed possible with quasi-Newton acceleration of the EM gradient algorithm. However, it should be borne in mind that on many problems the EM algorithm converges quickly, and no acceleration technique will produce orders of magnitude improvement. It is also likely that no accelerated version of the EM algorithm can match the stability and simplicity of the unadorned EM algorithm. Despite these reservations, acceleration is an attractive adjunct to current implementations of the EM algorithm.

For the sake of brevity, the present paper does not attempt an empirical comparison of the quasi-Newton acceleration of the EM algorithm with the conjugate gradient acceleration (Jamshidian and Jennrich (1993)). Such a comparison would be valuable. In the absence of empirical evidence, some advantages and disadvantages of the two algorithms are nonetheless obvious. Since it entails neither storage nor inversion of an approximate observed information matrix, the conjugate gradient acceleration will be particularly useful in problems with large numbers of parameters. Thus, EM image reconstruction algorithms are prime candidates for acceleration by conjugate gradients (Shepp and Vardi (1982), Lange and Carson (1984)). In this regard it is noteworthy that vector extrapolation techniques involving no large matrix inversions (Smith et al. (1987)) have already shown impressive gains in computational speed with no loss in image clarity (Rajeevan et al. (1992)).

In some instances, the quasi-Newton acceleration also does not require explicit storage or inversion of the approximate observed information matrix. Suppose $A = d^{20}Q(\theta^1 \mid \theta^1)$ is diagonal. Then A^{-1} is diagonal as well, and the inverse of a small-rank perturbation

$$\begin{aligned} A + \sum_{i=1}^{n} u^i (v^i)^t &= A + UV^t, \\ U &= (u^1, \ldots, u^n), \\ V &= (v^1, \ldots, v^n) \end{aligned}$$

can be computed easily via Woodbury's version of the Sherman-Morrison formula (Press et al. (1987)). Even more useful is the slight adaptation

$$(A + UV^t)^{-1} w = A^{-1} w - A^{-1} U (I + V^t A^{-1} U)^{-1} V^t A^{-1} w \tag{19}$$

of the Woodbury formula applied to a vector w. The matrix $I + V^t A^{-1} U$ appearing in (19) is just $n \times n$ and therefore easy to invert for n small.

The conjugate gradient acceleration does not adapt easily to problems with parameter bounds and constraints. In contrast, our scheme for approximating the observed information matrix combines gracefully with the method of recursive quadratic programming (Luenberger (1984), which does take into account parameter bounds and constraints. The quasi-Newton acceleration relies on the identity (7) for interior points θ. When θ is a boundary point or constraints eliminate the interior of the parameter domain, the identity may still hold. In such cases one should inquire whether the domain can be enlarged in violation of the bounds or in violation of the constraints but in such an manner that the given θ becomes an interior point. This enlargement must be compatible with $L(\theta)$ remaining a loglikelihood. If enlargement of the parameter domain is possible, then the identity (7) is still valid.

Negative definiteness of $d^{20}Q(\theta \mid \theta)$ is an essential feature of the quasi-Newton algorithm, but obviously not of the EM algorithm itself. When the complete data X come from a regular exponential family, then negative definiteness is automatic. In this instance, the complete data loglikelihood can be written as

$$\ln[f(X \mid \theta)] = \sum_j a_j(\theta) S_j(X) + b(\theta)$$

with the $a_j(\theta)$ linear and the $S_j(X)$ sufficient statistics. It turns out that $-d^{20}Q(\theta \mid \theta) = -d^2 b(\theta)$ can be identified with the expected information matrix of the complete data (Jennrich and Moore (1975)). Let us stress that regular exponential families are not the only distributional families having the negative definiteness property. The repeated measures model is a case in point.

A Bayesian version of the quasi-Newton acceleration is certainly possible. Any log prior $R(\theta)$ is left untouched by the E step of the algorithm and added to $Q(\theta \mid \theta^n)$. Since $R(\theta)$ is typically chosen negative definite, the sum $Q(\theta \mid \theta^n) + R(\theta)$ is even more negative definite than $Q(\theta \mid \theta^n)$ itself. Thus, finding posterior modes fits well into the present scheme.

Parameter asymptotic standard errors are in principle available by inverting the final approximate observed information matrix $-d^{20}Q(\theta^n \mid \theta^n) + B^n$. Experience with classical quasi-Newton methods suggests caution however (Thisted (1988)). Convergence may well occur before a sufficiently good approximation to $d^2 L(\theta^n)$ is constructed. Numerical differentiation of the identity (7) at the optimum point θ^∞ is probably the safer course (Meilijson (1989)). Alternatively, one can harness the EM algorithm directly to perform the necessary numerical differentiations (Meng and Rubin (1991)).

By now it should be evident that quasi-Newton acceleration of the EM algorithm raises more questions than a single paper can hope to answer. Our examples hardly exhaust the applications of the algorithm. Further theoretical development of the algorithm is also bound to be challenging, if for no other reason than that important foundational issues for classical quasi-Newton algorithms are still unresolved after more than two decades of intensive work (Nocedal (1992)). These barriers to understanding should not be allowed to detract from the potential of the quasi-Newton acceleration. The missing data paradigm is ubiquitous in statistical applications, and the EM algorithm enjoys ever wider use. Any measure that improves the EM algorithm can only benefit consumers of statistics.

Acknowledgments

I thank Stella Grosser for her help in programming the repeated measures example. Stella and Janet Sinsheimer read a preliminary draft of the paper and suggested improvements. This research was supported in part by the University of California, Los Angeles and USPHS Grant CA 16042.

References

Burridge, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. J. Roy. Statist. Soc. Ser.B **43**, 41-45.

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

Davidon, W. C. (1959). Variable metric methods for minimization. AEC Research and Development Report ANL-5990, Argonne National Laboratory.

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

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. J. Amer. Statist. Assoc. **64**, 1459-1471.

Horn, R. A. and Johnson, C. R. (1985). Matrix Analysis. Cambridge University Press, Cambridge, pp. 406-407.

Jamshidian, M. and Jennrich, R. I. (1993). Conjugate gradient acceleration of the EM algorithm. J. Amer. Statist. Assoc. **88**, 221-228.

Jennrich, R. I. and Moore, R. H. (1975). Maximum likelihood estimation by means of nonlinear least squares. Proceedings of the Statistical Computing Section: American Statistical Association, 57-65.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. Biometrics **42**, 805-820.

Kingman, J. F. C. (1993). Poisson Processes. Oxford University Press, Oxford, pp. 90-91.

Laird, N. M., Lange, N. and Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. J. Amer. Statist. Assoc. **82**, 97-105.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics **38**, 963-974.

Lange, K. (1994). SEARCH: A Fortran Program for Optimization. Department of Biostatistics, University of Michigan.

Lange, K. (1994). A gradient algorithm locally equivalent to the EM algorithm. J. Roy. Statist. Soc. Ser.B (in press).

Lange, K. and Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography. J. Comput. Assist. Tomography **8**, 306-316.

Little, R. J. A. and Rubin, D. B. (1987). Statistical Analysis with Missing Data. John Wiley, New York, pp. 112-119, 127-141.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. J. Roy. Statist. Soc. Ser.B **44**, 226-233.

Luenberger, D. G. (1984). Linear and Nonlinear Programming, 2nd edition. Addison-Wesley, Reading, Massachusetts, pp. 446-449.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. J. Roy. Statist. Soc. Ser.B **51**, 127-138.

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

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika **49**, 65-82.

Narayanan, A. (1991). Algorithm AS 266: Maximum likelihood estimation of the parameters of the Dirichlet distribution. Appl. Statist. **40**, 365-374.

Nazareth, L. (1980). Some recent approaches to solving large residual nonlinear least squares problems. SIAM Rev. **22**, 1-11.

Nocedal, J. (1992). Theory of algorithms for unconstrained optimization. Acta Numerica 1992, 199-242.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika **51**, 313-326.

Powell, M. J. D. (1978). A fast algorithm for nonlinearly constrained optimization calculations. Numerical Analysis: Proceedings of the Biennial Dundee Conference, 1977 (Edited by Watson GA), Springer, New York.

Pratt, J. W. (1981). Concavity of the log likelihood. J. Amer. Statist. Assoc. **76**, 103-106.

Press W. H., Flannery B. P., Teuklosky S. A. and Vetterling W. T. (1989). Numerical Recipes. The Art of Scientific Computing. Cambridge Univ. Press, Cambridge.

Rajeevan, N., Rajgopal, K. and Krishna, G. (1992). Vector extrapolated fast maximum likelihood estimation for emission tomography. IEEE Trans. Med. Imaging **11**, 9-20.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications, 2nd edition. John Wiley, New York, pp. 58-59, 522-523.

Shepp, L. A. and Vardi, Y. (1982). Maximum likelihood estimation for emission tomography. IEEE Trans. Med. Imaging MI-1, 113-121.

Smith, D. A., Ford, W. F. and Sidi, A. (1987). Extrapolation methods for vector sequences. SIAM Rev. **29**, 199-233.

Thisted, R. A. (1988). Elements of Statistical Computing: Numerical Computation. Chapman and Hall, New York, pp. 84-92, 209.

Titterington, D. M. (1984). Recursive parameter estimation using incomplete data. J. Roy. Statist. Soc. Ser.B **46**, 257-267.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). Statistical Analysis of Finite Mixture Distributions. John Wiley, New York, pp. 89-90.

Department of Biostatistics, School Public Health, University of Michigan, Ann Arbor, MI 48109, U.S.A.

(Received March 1994; accepted April 1994)