ON AITKEN'S METHOD AND OTHER APPROACHES FOR ACCELERATING CONVERGENCE OF THE EM ALGORITHM

G.J. McLachlan

Department of Mathematics

The University of Queensland

Brisbane 4072 Australia

Summary

The EM algorithm of Dempster, Laird, and Rubin (1977) is a broadly applicable approach that has been widely applied to the iterative computation of maximum likelihood estimates in a variety of incomplete-data problems. A criticism that has been levelled at the EM algorithm is that its convergence can be quite slow. Unfortunately, methods to accelerate the EM algorithm do tend to sacrifice the simplicity it usually enjoys. In this paper, we review some available methods for accelerating convergence of the EM algorithm. In particular, we consider the use of the multivariate version of Aitken's method for EM acceleration. Other methods to be considered include the conjugate gradient approach of Jamshidian and Jennrich (1993) and the quasi-Newton approach of Lange (1995b). We also consider the recently proposed ECME algorithm of Liu and Rubin (1994).

1. Introduction

A criticism that has been levelled at the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977) is that its convergence can be quite slow. We shall therefore consider some methods that have been proposed for accelerating the EM algorithm. However, as remarked by Lange (1995b), it is likely that no acceleration method can match the stability and simplicity of the unadorned EM algorithm. The reader is referred to McLachlan and Krishnan (1995) for a fuller account of this topic.

2. Formulation of the EM Algorithm

We let $\boldsymbol{Y}$ be a p-dimensional random vector with probability density function (p.d.f.) $g(\boldsymbol{y};\, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_d)^T$ is the vector containing the unknown parameters in the postulated form for the p.d.f. of $\boldsymbol{Y}$. Here the superscript T denotes the transpose of a vector or a matrix. The parameter space is denoted by $\boldsymbol{\Omega}$. The vector $\boldsymbol{\Psi}$ is to be estimated by maximum likelihood. The likelihood function for $\boldsymbol{\Psi}$ formed from the observed data $\boldsymbol{y}$ is given by $L(\boldsymbol{\Psi}) = g(\boldsymbol{y};\, \boldsymbol{\Psi})$. In the EM framework where $\boldsymbol{y}$ is viewed as being incomplete, we let $g_c(\boldsymbol{x};\, \boldsymbol{\Psi})$ denote the p.d.f. of the random vector $\boldsymbol{X}$ corresponding to the complete-data vector $\boldsymbol{x}$. Then the complete-data log likelihood function that could be formed for $\boldsymbol{\Psi}$ if $\boldsymbol{x}$ were fully observable is given by $\log L_c(\boldsymbol{\Psi}) = \log g_c(\boldsymbol{x};\, \boldsymbol{\Psi})$.

The EM algorithm approaches the problem of solving the incomplete-data likelihood equation indirectly by proceeding iteratively in terms of the complete-data log likelihood function $\log\, L_c(\boldsymbol{\Psi})$. Since the complete-data vector $\boldsymbol{x}$ is not fully observable, the complete-data log likelihood function is replaced by its conditional expectation given $\boldsymbol{y}$, using the current fit for $\boldsymbol{\Psi}$. More specifically, let $\boldsymbol{\Psi}^{(k)}$ be the value of $\boldsymbol{\Psi}$ after the kth iteration. Then on the $(k+1)$th iteration, the E- and M-steps are defined as follows:

E-step. Calculate $Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})$, where

$$Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\{\log L_c(\boldsymbol{\Psi}) \mid \boldsymbol{y}\}.$$

M-step. Choose $\boldsymbol{\Psi}^{(k+1)}$ to be any value of $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$ that globally maximizes $Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})$ with respect to $\boldsymbol{\Psi}$.

The E- and M-steps are alternated repeatedly until the likelihood changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\boldsymbol{\Psi}^{(k)})\}$. Dempster et al. (1977) showed that the (incomplete-data) likelihood function $L(\boldsymbol{\Psi})$ is not decreased after an EM iteration. Hence convergence must be obtained for a sequence of likelihood values that are bounded above.

3. EM Mapping

Any instance of the EM algorithm as described above implicitly defines a mapping $\boldsymbol{\Psi} \rightarrow \boldsymbol{M}(\boldsymbol{\Psi})$, from the parameter space of $\boldsymbol{\Psi}$, $\boldsymbol{\Omega}$, to itself such that

$$\boldsymbol{\Psi}^{(k+1)} = \boldsymbol{M}(\boldsymbol{\Psi}^{(k)}) \qquad (k = 0, 1, 2, \ldots).$$

If $\boldsymbol{\Psi}^{(k)}$ converges to some point $\boldsymbol{\Psi}^*$ and $\boldsymbol{M}(\boldsymbol{\Psi})$ is continuous, then $\boldsymbol{\Psi}^*$ must satisfy $\boldsymbol{\Psi}^* = \boldsymbol{M}(\boldsymbol{\Psi}^*)$. Thus $\boldsymbol{\Psi}^*$ is a fixed point of the map $\boldsymbol{M}$. By a Taylor series expansion of $\boldsymbol{\Psi}^{(k+1)} = \boldsymbol{M}(\boldsymbol{\Psi}^{(k)})$, we have in a neighbourhood of $\boldsymbol{\Psi}^*$ that

$$\boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^* \approx \boldsymbol{J}(\boldsymbol{\Psi}^*)(\boldsymbol{\Psi}^{(k)} - \boldsymbol{\Psi}^*), \tag{1}$$

where $\boldsymbol{J}(\boldsymbol{\Psi})$ is the $d \times d$ Jacobian matrix for $\boldsymbol{M}(\boldsymbol{\Psi}) = (M_1(\boldsymbol{\Psi}), \ldots, M_d(\boldsymbol{\Psi}))^T$, having (i,j)th element $J_{ij}(\boldsymbol{\Psi})$ equal to $J_{ij}(\boldsymbol{\Psi}) = \partial M_i(\boldsymbol{\Psi})/\partial \Psi_j$. Thus, in a neighbourhood of $\boldsymbol{\Psi}^*$, the EM algorithm is essentially a linear iteration with rate matrix $\boldsymbol{J}(\boldsymbol{\Psi}^*)$, since $J(\boldsymbol{\Psi}^*)$ is typically nonzero.

It is assumed henceforth that $\{\boldsymbol{\Psi}^{(k)}\}$ is an EM sequence for which

$$\partial Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{\Psi} = \mathbf{0}$$

is satisfied by $\boldsymbol{\Psi} = \boldsymbol{\Psi}^{(k+1)}$, which will be the case with standard complete-data estimation. Dempster et al. (1977) showed that if $\boldsymbol{\Psi}^{(k)}$ converges to a point $\boldsymbol{\Psi}^*$, then

$$\boldsymbol{J}(\boldsymbol{\Psi}^*) = \boldsymbol{I} - \boldsymbol{\mathcal{I}}_c^{-1}(\boldsymbol{\Psi}^*;\, \boldsymbol{y})\boldsymbol{I}(\boldsymbol{\Psi}^*;\, \boldsymbol{y}), \tag{2}$$

where

$$\boldsymbol{I}(\boldsymbol{\Psi};\, \boldsymbol{y}) = -\partial^2 \log L(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T$$

and

$$\boldsymbol{\mathcal{I}}_c(\boldsymbol{\Psi};\, \boldsymbol{y}) = E_{\boldsymbol{\Psi}}\{\boldsymbol{I}_c(\boldsymbol{\Psi};\, \boldsymbol{X}) \mid \boldsymbol{y}\},$$

and $\boldsymbol{I}_c(\boldsymbol{\Psi};\, \boldsymbol{x}) = -\partial^2 \log L_c(\boldsymbol{\Psi})/\boldsymbol{\Psi}\boldsymbol{\Psi}^T$. Also, $\boldsymbol{I}$ is the $d \times d$ identity matrix.

Exceptions to the convergence of the EM algorithm to a local maximum of $L(\boldsymbol{\Psi})$ occur if $\boldsymbol{J}(\boldsymbol{\Psi}^*)$ has eigenvalues exceeding unity. An eigenvalue of $\boldsymbol{J}(\boldsymbol{\Psi}^*)$ which is unity in a neighbourhood of $\boldsymbol{\Psi}^*$ implies a ridge in $L(\boldsymbol{\Psi}^*)$ through $\boldsymbol{\Psi}^*$. Generally, we expect $\partial^2 \log L(\boldsymbol{\Psi}^*)/\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T$ to be negative semidefinite, if not negative definite, in which case the eigenvalues of $\boldsymbol{J}(\boldsymbol{\Psi}^*)$ all lie in $[0,\, 1]$ or $[0,\, 1)$, respectively.

4. Multivariate Version of Aitken's Acceleration Method

The most commonly used method for EM acceleration is the multivariate version of Aitken's acceleration method. Suppose that $\boldsymbol{\Psi}^{(k)} \to \boldsymbol{\Psi}^*$, as $k \to \infty$. Then we can express $\boldsymbol{\Psi}^*$ as

$$\boldsymbol{\Psi}^* = \boldsymbol{\Psi}^{(k)} + \sum_{h=1}^{\infty}(\boldsymbol{\Psi}^{(h+k)} - \boldsymbol{\Psi}^{(h+k-1)}). \tag{3}$$

Now

$$\begin{aligned}
\boldsymbol{\Psi}^{(h+k)} - \boldsymbol{\Psi}^{(h+k-1)} &= \boldsymbol{M}(\boldsymbol{\Psi}^{(h+k-1)}) - \boldsymbol{M}(\boldsymbol{\Psi}^{(h+k-2)}) &(4)\\
&\approx \boldsymbol{J}(\boldsymbol{\Psi}^{(h+k-2)})(\boldsymbol{\Psi}^{(h+k-1)} - \boldsymbol{\Psi}^{(h+k-2)}) &(5)\\
&\approx \boldsymbol{J}(\boldsymbol{\Psi}^*)(\boldsymbol{\Psi}^{(h+k-1)} - \boldsymbol{\Psi}^{(h+k-2)}), &(6)
\end{aligned}$$

since

$$\boldsymbol{J}(\boldsymbol{\Psi}^{(h+k)}) = \boldsymbol{J}(\boldsymbol{\Psi}^*)$$

for k sufficiently large. The approximation (5) to (4) is obtained by a linear Taylor series expansion of $\boldsymbol{M}(\boldsymbol{\Psi}^{(h+k-1)})$ about the point $\boldsymbol{\Psi}^{(h+k-2)}$, as in (1). Repeated application of (6) in (3) gives

$$\begin{aligned} \boldsymbol{\Psi}^* &\approx \boldsymbol{\Psi}^{(k)} + \sum_{h=0}^{\infty} \{\boldsymbol{J}(\boldsymbol{\Psi}^*)\}^h (\boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^{(k)}) \\ &= \boldsymbol{\Psi}^{(k)} + \{\boldsymbol{I} - \boldsymbol{J}(\boldsymbol{\Psi}^*)\}^{-1} (\boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^{(k)}), \end{aligned} \tag{7}$$

as the power series

$$\sum_{h=0}^{\infty} \{\boldsymbol{J}(\boldsymbol{\Psi}^*)\}^h$$

converges to $\{\boldsymbol{I} - \boldsymbol{J}(\boldsymbol{\Psi}^*)\}^{-1}$ if $\boldsymbol{J}(\boldsymbol{\Psi}^*)$ has all its eigenvalues between 0 and 1.

The multivariate version (7) of Aitken's acceleration method suggests trying the sequence of iterates $\{\boldsymbol{\Psi}_A^{(k)}\}$, where $\boldsymbol{\Psi}_A^{(k+1)}$ is defined by

$$\boldsymbol{\Psi}_A^{(k+1)} = \boldsymbol{\Psi}_A^{(k)} + \{\boldsymbol{I} - \boldsymbol{J}(\boldsymbol{\Psi}_A^{(k)})\}^{-1} (\boldsymbol{\Psi}_{EMA}^{(k+1)} - \boldsymbol{\Psi}_A^{(k)}), \tag{8}$$

where $\boldsymbol{\Psi}_{EMA}^{(k+1)}$ is the EM iterate produced using $\boldsymbol{\Psi}_A^{(k)}$ as the current fit for $\boldsymbol{\Psi}$.

Hence this method proceeds on the $(k+1)$th iteration by first producing $\boldsymbol{\Psi}_{EMA}^{(k+1)}$ using an EM iteration with $\boldsymbol{\Psi}_A^{(k)}$ as the current fit for $\boldsymbol{\Psi}$. One then uses the EM iterate $\boldsymbol{\Psi}_{EMA}^{(k+1)}$ in Aitken's acceleration procedure (7) to yield the final iterate $\boldsymbol{\Psi}_A^{(k+1)}$ on the $(k+1)$th iteration. This is the method proposed by Louis (1982) for speeding up the convergence of the EM algorithm.

Louis (1982) suggests making use of the relationship (2) to estimate $\boldsymbol{J}(\boldsymbol{\Psi}_A^{(k)})$ in (8). Using (2) in (8) gives

$$\boldsymbol{\Psi}_A^{(k+1)} = \boldsymbol{\Psi}_A^{(k)} + \boldsymbol{I}^{-1}(\boldsymbol{\Psi}_A^{(k)}; \boldsymbol{y}) \boldsymbol{\mathcal{I}}_c(\boldsymbol{\Psi}_A^{(k)}; \boldsymbol{y}) (\boldsymbol{\Psi}_{EMA}^{(k+1)} - \boldsymbol{\Psi}_A^{(k)}). \tag{9}$$

The relationship (2) for values of $\boldsymbol{\Psi}$ other than $\boldsymbol{\Psi}^*$ is an approximation useful only local to the MLE $\boldsymbol{\Psi}^*$, and so should not be used until some EM iterations have been performed. As noted by Meilijson (1989), the use of (9) is approximately equivalent to using the Newton-Raphson algorithm to find a zero of the (incomplete-data) score statistic $\boldsymbol{S}(\boldsymbol{y};\boldsymbol{\Psi})=\partial\log L(\boldsymbol{\Psi})/\partial\boldsymbol{\Psi}$. It can be shown that

$$\boldsymbol{S}(\boldsymbol{y};\,\boldsymbol{\Psi}_A^{(k)})\approx\boldsymbol{\mathcal{I}}_c(\boldsymbol{\Psi}_A^{(k)};\,\boldsymbol{y})(\boldsymbol{\Psi}_{EMA}^{(k+1)}-\boldsymbol{\Psi}_A^{(k)}). \tag{10}$$

On substituting this result in (9), we obtain

$$\boldsymbol{\Psi}_A^{(k+1)}\approx\boldsymbol{\Psi}_A^{(k)}+\boldsymbol{I}^{-1}(\boldsymbol{\Psi}_A^{(k)};\,\boldsymbol{y})\boldsymbol{S}(\boldsymbol{y};\,\boldsymbol{\Psi}_A^{(k)}). \tag{11}$$

The right-hand side of (11) is the iterate produced on the $(k+1)$th iteration of the Newton-Raphson procedure applied to finding a zero of $\boldsymbol{S}(\boldsymbol{y};\,\boldsymbol{\Psi})$. Hence the use of Aitken's procedure as applied by Louis (1982) is essentially equivalent to the Newton-Raphson procedure applied to $\boldsymbol{S}(\boldsymbol{y};\,\boldsymbol{\Psi})$.

In the case of independent and identically distributed observations, $\boldsymbol{w}_1,\,\ldots,\,\boldsymbol{w}_n$, Meilijson (1989) considered the Newton-Raphson procedure (11) with $\boldsymbol{I}(\boldsymbol{\Psi};\,\boldsymbol{y})$ approximated by the so-called empirical information matrix

$$\boldsymbol{I}_e(\boldsymbol{\Psi};\,\boldsymbol{y})=\sum_{j=1}^{n}\boldsymbol{s}(\boldsymbol{w}_j;\,\boldsymbol{\Psi})\boldsymbol{s}^T(\boldsymbol{w}_j;\,\boldsymbol{\Psi})-n^{-1}\boldsymbol{S}(\boldsymbol{y};\,\boldsymbol{\Psi})\boldsymbol{S}^T(\boldsymbol{y};\,\boldsymbol{\Psi}),$$

where $\boldsymbol{s}(\boldsymbol{w}_j;\,\boldsymbol{\Psi})$ is the score statistic based on just the single observation $\boldsymbol{w}_j$.

More recently, Jamshidian and Jennrich (1993) noted that the accelerated sequence (8) as proposed by Louis (1982) is precisely the same as that obtained by applying the Newton-Raphson method to find a zero of the difference

$$\boldsymbol{\delta}(\boldsymbol{\Psi})=\boldsymbol{M}(\boldsymbol{\Psi})-\boldsymbol{\Psi},$$

where $\boldsymbol{M}$ is the map defined by the EM sequence. To see this, we can write (8) as

$$\boldsymbol{\Psi}_A^{(k+1)}=\boldsymbol{\Psi}_A^{(k)}+\{\boldsymbol{I}-\boldsymbol{J}(\boldsymbol{\Psi}_A^{(k)})\}^{-1}\{\boldsymbol{M}(\boldsymbol{\Psi}_A^{(k)})-\boldsymbol{\Psi}_A^{(k)}\}, \tag{12}$$

since $\boldsymbol{\Psi}_{EMA}^{(k+1)}=\boldsymbol{M}(\boldsymbol{\Psi}_A^{(k)})$. As the gradient of $\boldsymbol{\delta}(\boldsymbol{\Psi})$ is $-\{\boldsymbol{I}-\boldsymbol{J}(\boldsymbol{\Psi})\}$, (12) corresponds exactly to applying the Newton-Raphson procedure to find a zero of $\boldsymbol{\delta}(\boldsymbol{\Psi})$.

5. Conjugate Gradient Algorithm

Jamshidian and Jennrich (1993) proposed an acceleration method based on conjugate gradients. The key is that the change in $\boldsymbol{\Psi}$ after an EM iteration can be viewed (approximately at least) as a generalized gradient, making it natural to apply generalized conjugate gradient methods in an attempt to accelerate the EM algorithm.

Using (10),

$$\begin{aligned}\boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^{(k)} &= \boldsymbol{M}(\boldsymbol{\Psi}^{(k)}) - \boldsymbol{\Psi}^{(k)} \\ &\approx \boldsymbol{\mathcal{I}}_c^{-1}(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{y})\boldsymbol{S}(\boldsymbol{y};\, \boldsymbol{\Psi}^{(k)}).\end{aligned}$$

Thus the change in $\boldsymbol{\Psi}$ after an EM iteration, $\boldsymbol{\delta}(\boldsymbol{\Psi})$, is approximately equal to

$$\boldsymbol{\delta}(\boldsymbol{\Psi}) \approx \boldsymbol{\mathcal{I}}_c^{-1}(\boldsymbol{\Psi};\, \boldsymbol{y})\boldsymbol{S}(\boldsymbol{y};\, \boldsymbol{\Psi}). \tag{13}$$

Typically $\boldsymbol{\mathcal{I}}_c(\boldsymbol{\Psi};\, \boldsymbol{y})$ is positive definite. Thus (13) shows that when $\boldsymbol{M}(\boldsymbol{\Psi})$ is near to $\boldsymbol{\Psi}$, $\boldsymbol{\delta}(\boldsymbol{\Psi})$ is, to a good approximation, a generalized gradient of $\log L(\boldsymbol{\Psi})$. Jamshidian and Jennrich (1993) therefore proposed accelerating the EM algorithm by applying the generalized conjugate gradient algorithm. They call the resulting algorithm the accelerated EM (AEM) algorithm. Each step of this algorithm begins with an EM iteration. First, its direction is modified, and then its length is optimized.

Although the AEM algorithm is fairly simple, it is more complex than the EM algorithm itself. In addition to the EM iterations, one must compute the score statistic $\boldsymbol{S}(\boldsymbol{y};\, \boldsymbol{\Psi})$. Frequently, however, the latter is either available from the EM code or is obtainable with minor modification. The biggest complication is the line search required in the first step. An algorithm for this search is given in the Appendix of Jamshidian of Jennrich (1993). Their experience suggests that a simple line search is sufficient. To demonstrate the effectiveness of the AEM algorithm, Jamshidian and Jennrich (1993) applied it to several problems in areas that included the estimation of a covariance matrix from incomplete multivariate normal data, confirmatory factor analysis, and repeated measures analysis. In terms of operation counts, for all of the comparative examples considered the AEM algorithm increases the speed of the EM algorithm, in some cases by a factor of 10 or more.

6. Hybrid Methods for Finding MLE

Various authors, including Redner and Walker (1984), propose a hybrid approach to the computation of the MLE that would switch from the EM algorithm after a few iterations to the Newton-Raphson or some quasi-Newton method. The idea is to use the EM algorithm initially to take advantage of its good global convergence properties and to then exploit the rapid local convergence of Newton-type methods by switching to such a method. It can be seen that the method of Louis (1982) is a hybrid algorithm of this nature, for in effect after a few initial EM iterations, it uses the Newton-Raphson method to accelerate convergence after performing each subsequent EM iteration. Of course there is no guarantee that these hybrid algorithms increase the likelihood $L(\boldsymbol{\Psi})$ monotonically. Hybrid algorithms have been considered also by Jones and McLachlan (1992), and Aitkin and Aitkin (1995), among others.

7. EM Gradient Algorithm

Consider the sequence of iterates $\{\boldsymbol{\Psi}^{(k)}\}$, where $\boldsymbol{\Psi}^{(k+1)}$ is defined to be

$$\boldsymbol{\Psi}^{(k+1)} = \boldsymbol{\Psi}^{(k)} - s^{(k)}\boldsymbol{\delta}^{(k)}, \tag{14}$$

where

$$\boldsymbol{\delta}^{(k)} = [\partial^2 Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})/\partial\boldsymbol{\Psi}\partial\boldsymbol{\Psi}^T]^{-1}_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k)}}[\partial Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})/\partial\boldsymbol{\Psi}]_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k)}},$$

and where $0 < s^{(k)} \leq 1$. Rai and Matthews (1993) considered the choice of the constant $s^{(k)}$ so that (14) defines a GEM algorithm. That is, so that

$$Q(\boldsymbol{\Psi}^{(k+1)};\, \boldsymbol{\Psi}^{(k)}) \geq Q(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{\Psi}^{(k)}),$$

which ensures the monotonicity of the sequence $\{L(\boldsymbol{\Psi}^{(k)})\}$.

Lange (1995a) considered the sequence of iterates defined by (14) with $s^{(k)} = 1$. He called this algorithm the EM gradient algorithm. Lange (1995b) subsequently used the EM gradient algorithm to form the basis of a quasi-Newton approach to accelerate convergence of the EM algorithm. But as pointed out by Lange (1995a), the EM gradient algorithm is an interesting algorithm in its own right. Since the Newton-Raphson method converges quickly, the local properties of the EM gradient algorithm are almost identical with those of the EM algorithm.

The matrix

$$[\partial^2 Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})/\partial\boldsymbol{\Psi}\boldsymbol{\Psi}^T]_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k)}} = -\boldsymbol{\mathcal{I}}_c(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{y}) \tag{15}$$

may not be negative definite, and consequently the EM gradient algorithm is not necessarily an ascent algorithm. However, in practice, the matrix (15) is typically negative definite. In the case of the complete-data density being a member of the regular exponential family with natural parameter $\boldsymbol{\Psi}$, we have that this matrix is equal to the complete-data information matrix, and so negative definiteness is automatic. In this case, the EM gradient algorithm coincides with the ascent algorithm of Titterington (1984).

8. A Quasi-Newton Acceleration of the EM Algorithm

In the case of $s^{(k)} = 1$, (14) defines the first iterate obtained when using the Newton-Raphson procedure on the M-step where the intent is to maximize $Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})$. Under regularity conditions, we have that

$$[\partial Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{\Psi}]_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k)}} = \boldsymbol{S}(\boldsymbol{y};\, \boldsymbol{\Psi}^{(k)})$$

and

$$[\partial^2 Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T]_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k)}} = -\boldsymbol{\mathcal{I}}_c(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{y}).$$

Then we can express (14) with $s^{(k)} = 1$ as

$$\boldsymbol{\Psi}^{(k+1)} = \boldsymbol{\Psi}^{(k)} + \boldsymbol{\mathcal{I}}_c^{(-1}(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{y}) \boldsymbol{S}(\boldsymbol{y};\, \boldsymbol{\Psi}^{(k)}). \tag{16}$$

On contrasting (16) with (11), the use of the EM gradient algorithm can be viewed as using the Newton-Raphson method to find a zero of the likelihood equation, but with the approximation

$$\boldsymbol{I}(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{y}) \approx \boldsymbol{\mathcal{I}}_c(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{y}).$$

The quasi-Newton acceleration procedure proposed by Lange (1995b) defines $\boldsymbol{\Psi}^{(k+1)}$ to be

$$\boldsymbol{\Psi}^{(k+1)} = \boldsymbol{\Psi}^{(k)} + \{\boldsymbol{\mathcal{I}}_c(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{y}) + \boldsymbol{B}^{(k)}\}^{-1} \boldsymbol{S}(\boldsymbol{y};\, \boldsymbol{\Psi}^{(k)}), \tag{17}$$

where

$$\boldsymbol{B}^{(k)} = \boldsymbol{B}^{(k-1)} + c^{(k)} \boldsymbol{v}^{(k)} \boldsymbol{v}^{(k)^T},$$

and where

$$c^{(k)} = -1/\{\boldsymbol{v}^{(k)^T}(\boldsymbol{\Psi}^{(k)} - \boldsymbol{\Psi}^{(k-1)})\},$$

$$\boldsymbol{v}^{(k)} = \boldsymbol{h}^{(k)} + \boldsymbol{B}^{(k-1)}(\boldsymbol{\Psi}^{(k)} - \boldsymbol{\Psi}^{(k-1)}),$$

and

$$\boldsymbol{h}^{(k)} = [\partial Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)})/\partial \boldsymbol{\Psi}]_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k-1)}} - [\partial Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k-1)})/\partial \boldsymbol{\Psi}]_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(k-1)}}.$$

When the matrix

$$\boldsymbol{\mathcal{I}}_c(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{y}) + \boldsymbol{B}^{(k)}$$

fails to be positive definite, Lange (1995b) suggests replacing it by

$$\boldsymbol{\mathcal{I}}_c(\boldsymbol{\Psi}^{(k)};\, \boldsymbol{y}) + (\tfrac{1}{2})^m \boldsymbol{B}^{(k)}, \tag{18}$$

where m is the smallest positive integer such that (18) is positive definite. If (17) does not lead to an increase in $L(\boldsymbol{\Psi})$, then some form of step decrementing must be implemented.

As Lange (1995b) points out, the early stages of the algorithm resemble a close approximation to the EM algorithm, later stages approximate Newton-Raphson, and intermediate stages make a graceful transition between the two extremes.

9. ECME Algorithm

Liu and Rubin (1994) presented an extension of then ECM algorithm called the ECME (expectation - conditional maximization either) algorithm. Here the "either" refers to the fact that with this extension, some or all of the CM-steps of the ECM algorithm are replaced by steps that conditionally maximize the incomplete-data log likelihood function and not the Q-function. Hence with the ECME algorithm, each CM-step either maximizes $Q(\boldsymbol{\Psi}\,;\, \boldsymbol{\Psi}^{(k)})$ or the actual (incomplete-data) log likelihood function, $\log L(\boldsymbol{\Psi})$, subject to the same constraints on $\boldsymbol{\Psi}$. Typically, the ECME algorithm is more tedious to code than the ECM algorithm of Meng and Rubin (1993), but the reward of faster convergence is often worthwhile especially because it allows convergence to be more easily assessed. It is is nearly always faster than the EM and ECM algorithms in terms of number of iterations and moreover can be faster in total computer time by orders of magnitude. The ECME algorithm shares with both the EM and ECM algorithms their stable monotone convergence; that is, $L(\boldsymbol{\Psi}^{(k+1)}) \geq L(\boldsymbol{\Psi}^{(k)})$. Analogous convergence results hold for the ECME algorithm as for the ECM and EM algorithms.

References

Aitkin, M. and Aitkin, I (1995). A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing*. To appear.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data *via* the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B* **39**, 1–38.

Jamshidian, M. and Jennrich, R.I. (1993). Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association* **88**, 221–228.

Jones, P.N. and McLachlan, G.J. (1992). Improving the convergence rate of the EM algorithm for a mixture model fitted to grouped truncated data. *Journal of Statistical Computation and Simulation* **43**, 31–44.

Lange, K. (1995a). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society Series B* **57**, 425–437.

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

Liu, C. and Rubin, D.B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 633–648.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B* **44**, 226–233.

McLachlan, G.J. and Krishnan, T. (1995). *The EM Algorithm and Extensions.* New York: Wiley. To be published.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society Series B* **51**, 127–138.

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

Rai, S.N. and Matthews, D.E. (1993). Improving the EM algorithm. *Biometrics* **49**, 587–591.

Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the E.M. algorithm. *SIAM Review* **26**, 195–239.

Titterington, D.M. (1984). Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society Series B*, **46**, 257–267.