IMPROVING THE EM ALGORITHM

David Lansky and George Casella, Biometrics Unit
337 Warren Hall, Cornell University
Ithaca, N.Y. 14853

BU-1081-M

November 10, 1990

ABSTRACT

The EM algorithm is often a practical method for obtaining maximum likelihood estimates. For the vector parameter case, we provide a faster method than Meng and Rubin (1989) for obtaining the derivative of the EM mapping, which can be used to obtain the observed variance-covariance matrix. Our method exhibits good behavior for a simple example. Aitken's acceleration is commonly used to speed convergence of EM near a solution. Because Aitken's acceleration often fails to converge we propose a mixture of EM and Aitken accelerated EM which satisfies the generalized EM (GEM) criteria, assuring convergence. We show that such a mixture sequence exists and demonstrate good convergence behavior for a heuristic approximation to this mixture.

1. INTRODUCTION

The EM algorithm is often a good iterative approach for difficult maximum likelihood estimation problems, providing answers that are as good as the likelihood surface allows. The major shortcoming of EM is that it is often slow to converge, particularly near the end of the search.

We represent EM as a mapping from a parameter estimate, $\theta^{(k)}$, to a new estimate $\theta^{(k+1)}$,

$$\theta^{(k+1)} = \mathrm{M}(\theta^{(k)}).$$

Associated with this mapping, we define the derivative of the mapping, $\boldsymbol{DM}(\theta^{(k)})$, as the Jacobian of the component derivatives of $\mathrm{M}(\theta^{(k)})$ with respect to $\theta^{(k)}$,

$$\boldsymbol{DM}(\theta^{(k)}) = \{r_{ij}\},$$

where $r_{ij} = \dfrac{\partial \mathrm{M}(\theta^{(k)})_i}{\partial \theta_j^{(k)}}$.

The derivative of the EM mapping has interesting and useful properties from both algorithmic and statistical viewpoints. For likelihoods with continuous derivatives in θ, the derivative of the mapping, $\boldsymbol{DM}(\theta^{(k)})$, converges to $\boldsymbol{DM}(\theta^*)$ as $\theta^{(k)} \to \theta^*$, where θ^* is a local maximum of the likelihood surface.

Aitken's acceleration, applied to EM, does not converge reliably unless it is started close to a local maximum (Louis, 1982). Louis (1982) suggests using EM for early iterations, where it is stable to poor starting values, and Aitken accelerated EM for late iterations, where EM is slow to converge. We propose criteria for deciding when to begin accelerating EM.

We seek an accelerated version of EM that will preserve both the stability of EM to a wide range of starting values and its easy implementation for statisticians. We show that mixing the EM and Aitken steps can yield a generalized EM (GEM) step, which will preserve the good convergence properties of EM (Wu, 1983). We present a heuristic mixture that performs well.

2. DEFINITION OF EM AND NOTATION

EM is formally defined following Dempster, Laird and Rubin, (1977; subsequently denoted as DLR, 1977) as follows. Postulate two sample spaces $\mathfrak{X}$ and $\mathfrak{Y}$ and a many-to-one mapping from $\mathfrak{X}$ to $\mathfrak{Y}$, denoted $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$. The observed data $\boldsymbol{y}$ are a realization from $\mathfrak{Y}$. We observe $\boldsymbol{x}$ in $\mathfrak{X}$ only indirectly through $\boldsymbol{y}$. We call $\boldsymbol{x}$ the "complete data" and $\boldsymbol{y}$ the "incomplete data". The subset of $\mathfrak{X}$ in which the complete data lie is denoted $\mathfrak{X}(\boldsymbol{y}) = \{\boldsymbol{x} : \boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})\}$. The family of densities $\mathrm{f}(\boldsymbol{x} \mid \theta)$ induces a related family of sampling densities $\mathrm{g}(\boldsymbol{y} \mid \theta)$,

$$\mathrm{g}(\boldsymbol{y} \mid \theta) = \int_{\mathfrak{X}(\boldsymbol{y})} \mathrm{f}(\boldsymbol{x} \mid \theta)\, \mathrm{d}\boldsymbol{x}.$$

The common parameterization, θ, for f and g is an essential feature of the EM setup.

The EM mapping usually consists of two steps an Expectation step and a Maximization step. In practice these steps are combined. The general definition of EM does not refer to these E and M steps; instead we maximize the expected log of the complete data likelihood at each step, so that

$$\theta^{(k+1)} = \mathrm{M}\left(\theta^{(k)}\right)$$
$$= \left\{\theta^{(k+1)}: \mathrm{Q}\left(\theta^{(k+1)} \mid \theta^{(k)}\right) = \max_{\theta \in \Omega} \mathrm{Q}\left(\theta \mid \theta^{(k)}\right)\right\},$$

where

$$\mathrm{Q}(\theta' \mid \theta) = \mathrm{E}\left(\log \mathrm{f}(\boldsymbol{x} \mid \theta') \mid \boldsymbol{y}, \theta\right)$$
$$= \int_{\mathfrak{B}(\boldsymbol{y})} \log\left\{\mathrm{f}(\boldsymbol{x} \mid \theta')\right\} \mathrm{f}(\boldsymbol{x} \mid \theta)\, \mathrm{d}\boldsymbol{x}.$$

For GEM (Generalized EM) we need only increase Q rather than maximize Q at each step, satisfying

$$\mathrm{Q}\left(\theta^{(k+1)} \mid \theta^{(k)}\right) \geq \mathrm{Q}\left(\theta^{(k)} \mid \theta^{(k)}\right).$$

Practical use of EM demands a density $\mathrm{f}(\boldsymbol{x} \mid \theta)$ where at least one of the E and M steps is easy to calculate. For the E-step we calculate,

$$t^{(k)} = \mathrm{E}_{\mathfrak{B}(\boldsymbol{y})}\left\{t(\boldsymbol{x}) \mid \boldsymbol{y}, \theta^{(k)}\right\}$$

where $t(\boldsymbol{x})$ is the set of sufficient statistics for θ. The M-step is then,

$$\theta^{(k+1)} = \mathrm{MLE}_{\mathrm{f}}(\theta \mid t^{(k)}).$$

EM then maximizes $\mathcal{L}(\theta \mid \boldsymbol{y})$ without evaluating $\mathcal{L}(\theta \mid \boldsymbol{y})$, $\mathcal{L}(\theta \mid \boldsymbol{x})$ or even $\mathrm{Q}(\theta' \mid \theta)$.

3. PROPERTIES OF $\boldsymbol{DM}(\theta^{(k)})$

We desire an estimate of $\boldsymbol{I}_{\boldsymbol{y}}^{-1}$ ($\boldsymbol{I}_{\boldsymbol{y}}$ and $\boldsymbol{I}_{\boldsymbol{x}}$ are the Fisher Information for the observed and complete data respectively) to construct variance estimators based on the observed (incomplete data) Fisher Information (Efron and Hinkley, 1978). We use EM when it is difficult to evaluate the observed data likelihood. Thus, it is usually easier to evaluate $\boldsymbol{I}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{DM}(\theta^*)$ and exploit the exponential family relationship reported in DLR (1977)

$$\boldsymbol{DM}(\theta^*) = \mathrm{V}(\boldsymbol{t}(\boldsymbol{x}) \mid \theta^*, \boldsymbol{y})\, \mathrm{V}^{-1}(\boldsymbol{t}(\boldsymbol{x}) \mid \theta^*)$$

where $\mathrm{V}(\boldsymbol{t}(\boldsymbol{x}) \mid \theta^*, \boldsymbol{y}) = \boldsymbol{I}_{\boldsymbol{x} \mid \boldsymbol{y}}^{-1}$ and $\boldsymbol{I}_{\boldsymbol{x} \mid \boldsymbol{y}} = \boldsymbol{I}_{\boldsymbol{x}} - \boldsymbol{I}_{\boldsymbol{y}}$.

Hence, we can derive

$$\boldsymbol{I}_{\boldsymbol{y}}^{-1} = (\boldsymbol{I} - \boldsymbol{DM}(\theta^*)^{-1})^{-1} \boldsymbol{I}_{\boldsymbol{x}}^{-1},$$

which, as Meilijson (1989) has pointed out, is incorrectly reported in Louis (1982). The asymptotic rate of convergence of EM is a one minus the largest eigenvalue of $\boldsymbol{DM}(\theta^*)$ when this eigenvalue is less than one (DLR, 1977). Under mild regularity conditions, $\boldsymbol{DM}(\theta^{(k)})$ converges to $\boldsymbol{DM}(\theta^*)$. We propose using a near convergence criterion (technically lack of change) on $\boldsymbol{DM}(\theta^{(k)})$ to indicate that EM is "stable." We call EM "stable" when it is taking many small steps in essentially the same direction.

4. ESTIMATING $\hat{\boldsymbol{DM}}(\theta^{(k)})$ and $\hat{\boldsymbol{DM}}(\theta^*)$

Meng and Rubin (1989) propose estimating $\boldsymbol{DM}$ from forced EM steps. For the vector parameter case they take differences from θ^* at each step using

$$\hat{\boldsymbol{DM}}(\theta^*) = \lim_{k \to \infty} \hat{\boldsymbol{DM}}_{\mathrm{m}}(\theta^{(k)}),$$

and

$$\hat{\boldsymbol{DM}}_{\mathrm{m}}(\theta^{(k)}) = \left\{\mathrm{r}_{ij}^{(k)}\right\},$$

with

$$\mathrm{r}_{ij}^{(k)} = \frac{\tilde{\theta}_j^{(k+1)}(i) - \theta_j^*}{\theta_i^{(k)} - \theta_i^*}$$

and

$$\tilde{\theta}^{(k+1)}(i) = \mathrm{M}\{\theta_1^*, \theta_2^*, \ldots, \theta_{i-1}^*, \theta_i^{(k)}, \theta_{i-1}^*, \ldots, \theta_d^*\}.$$

Note that their method yields the transpose of the usual Jacobian. This approach requires knowledge of θ^*; hence, we must first find $\hat{\theta}^*$, the MLE for θ. This requires running EM twice, first to estimate $\hat{\theta}^*$, and second until all elements of $\hat{\boldsymbol{DM}}_{\mathrm{m}}(\theta^{(k)})$ have satisfied appropriate convergence criteria. This is a substantial duplication of EM steps.

We extend Meng and Rubin's (1989) idea from the scalar parameter case to the multiparameter

case, using the step sizes of the mapping to estimate the derivative of the mapping. Dennis and Schnabel (1983) recommend using these step sizes for derivative estimation. We use

$$\hat{DM}_1(\theta^{(k)}) = \left\{q_{ij}^{(k)}\right\}$$

where

$$q_{ij}^{(k)} = \frac{\overset{\circ}{\theta}{}_i^{(k+1)}(j) - \theta_i^{(k+1)}}{\theta_j^{(k+1)} - \theta_j^{(k)}}$$

and

$$\overset{\circ}{\theta}{}^{(k+1)}(j) = M\{\theta_1^{(k)}, \ldots, \theta_{j-1}^{(k)}, \theta_j^{(k+1)}, \theta_{j-2}^{(k)}, \ldots, \theta_d^{(k)}\}.$$

For both methods, we can stop taking the jth forced EM step when row j of $\boldsymbol{DM}$ has converged. In our example both methods exhibit convergence before the denominator becomes too small. Calculating $\hat{DM}(\theta^*) = \lim_{k\to\infty} \hat{DM}_1(\theta^{(k)})$ on the first (and only) pass through the EM sequence, has worked well in simulation studies on a simple example. This approach appears promising for more complex problems.

5. EM, AITKEN'S AND A MIXTURE OF THEM

The Aitken acceleration estimate, $\theta_A^{(k+1)}$, can be written in terms of the current estimate of θ, $\theta^{(k)}$, the next EM estimate of θ, $\theta^{(k+1)}$, and the inverse of $(I - \boldsymbol{DM}(\theta^*))$,

$$\theta_A^{(k+1)} \overset{\&}{=} \theta^{(k)} + \left(I - \boldsymbol{DM}(\theta^*)\right)^{-1}\left(\theta^{(k+1)} - \theta^{(k)}\right),$$

assuming that all the eigenvalues of $\boldsymbol{DM}(\theta^*)$ are less than one in absolute value. $\boldsymbol{DM}(\theta^{(k)})$ is often used in place of $\boldsymbol{DM}(\theta^*)$ in Aitken's acceleration (Louis, 1982).

One of the most useful properties of EM and GEM is their reliable convergence. Combining the EM and Aitken acceleration steps so that each mixed step is a GEM step would assure convergence.

Theorem 1. For $\theta^{(k)}$ an estimate of θ, there exists a constant $\Lambda^{(k)}$, $0<\Lambda^{(k)}<1$, such that

$$\theta^{(k+1)} = \Lambda^{(k)}\theta_A^{(k+1)} + (1-\Lambda^{(k)})\theta_{EM}^{(k+1)}$$

will satisfy the GEM criteria, with $\theta_{EM}^{(k+1)}$ and $\theta_A^{(k+1)}$ the usual EM and Aitken estimates.

Proof: $\theta_{EM}^{(k+1)}$ will maximize $Q\left(\theta \mid \theta^{(k)}\right)$ and $\theta_A^{(k+1)}$ may increase or decrease $Q\left(\theta \mid \theta^{(k)}\right)$. Yet, for a GEM step we need only satisfy $Q\left(\theta \mid \theta^{(k)}\right) \geq Q\left(\theta^{(k)} \mid \theta^{(k)}\right)$; there is clearly a constant $\Lambda^{(k)}$, $0 < \Lambda^{(k)} < 1$, which will produce the desired result. □

By induction we can establish that a sequence of Λs, which produce a GEM sequence, exists. We expect some of these mixtures to converge more rapidly than EM. For this paper we suggest some functions for $\Lambda^{(k)}$. A simple choice for $\Lambda^{(k)}$, $\frac{1}{2}$, performs surprisingly well for our three parameter EM problem. Other functions we examine include powers of the cosine of the angle between the EM and Aitken step, and the ratio of the lengths of the EM and Aitken steps. When EM is "stable" it takes many small steps in the same direction. This motivates our use of the cosine to weight the Aitken step more heavily when it goes in approximately the same direction as EM. Higher powers of the cosine weight EM more heavily. The relative lengths of the EM and Aitken steps are the only other current step information available. Because EM takes small steps, relative step length seems less likely than cosine to perform well.

6. EXAMPLE

To evaluate candidates for $\Lambda^{(k)}$ we constructed a simple example. We sampled from a bivariate normal population $(x,y) \sim \text{BVN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ with some pairs incomplete (all missing observations were from y). Most of our simulations used $n_x = 15$ and $n_y = 5$, with $\theta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\} = \{5,10,3,3,.9\}$. Other sample sizes and parameter values gave similar results. For each sample we construct MLEs for μ_1 and σ_1^2 using all n_x observations. Using only the n_y observations with both x and y observed we construct starting estimates of μ_2, σ_2^2 and ρ. Three sequences were constructed for each sample: 1) EM, 2) EM followed by the Aitken accelerated EM, and 3) EM followed by the mixture described in section 5. The latter

two sequences changed from EM to accelerated versions when $DM(\theta^{(k)})$ is "stable" $|\max_{i,j}(DM(\theta^{(k+1)}) - DM(\theta^{(k)}))| < 0.4$. We used a fixed step size numerical estimate of $DM(\theta^*)$ because we cannot guarantee the convergence of $\theta^{(k)}$ which is required by both our and Meng and Rubin's (1989) DM estimation methods.

Table 1 indicates that the 4th power of the cosine performed about as well as Aitken's method. Aitken's failed to converge for approximately 25% of our samples while the cosine based mixtures converged slightly more often, but requiring slightly more steps. The methods based on relative step lengths did not perform well, as expected. Figure 1 illustrates a situation where Aitken's acceleration and the cosine mixture both converge, yet the cosine requires fewer steps. The mixture path is closer to the EM sequence and smoother than the Aitken's acceleration path. Figure 2 illustrates a case where the Aitken acceleration fails to converge.

Table 1. Comparison of EM, Aitken acceleration and mixtures of the two, in terms of number of times that each failed to converge (EM never failed) and the average number of steps to convergence in 100 trials.

Method	percent convergence	mean number of steps
EM	100	62.7
Aitken	75	5.4
$\cos^4$	76	6.8

7. SUMMARY

We modify Meng and Rubin's (1989) method of estimating the derivative of the EM mapping to require only one EM sequence. Their and our method relies on the convergence of the derivative of the EM mapping which depends upon convergence of $\theta^{(k)} \to \theta^*$. We propose near convergence of the derivative of the EM mapping to indicate when to begin using Aitken-like acceleration methods.

We show that a there exists a sequence of steps, composed of convex mixtures of EM and Aitken steps, which is a GEM sequence. For several simple heuristic approximations to this mixture we demonstrate convergence for a simple EM problem.

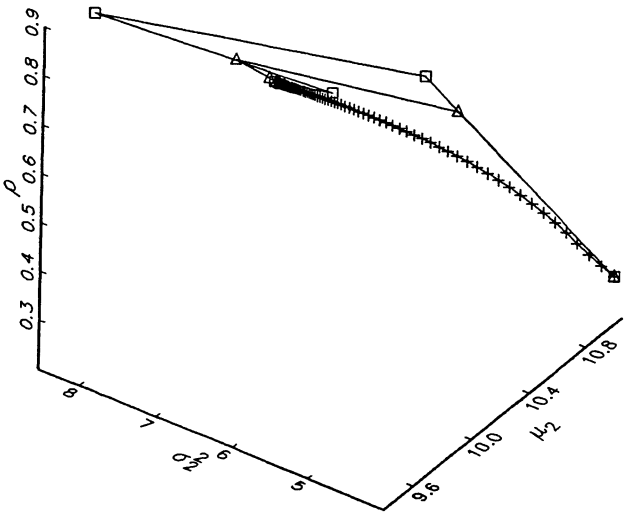

Figure 1. Sequences of EM (+) , Aitken (□) and mixed (Δ) estimates where the mixture does very well. Note the extremely slow progress of the EM algorithm, particularly in the last part of the sequence.

ACKNOWLEDGEMENTS

This research has been partially supported by NSF grant DMS-8906580 to C. Castillo-Chavez. The authors thank Charles McCulloch for his suggestions and critical comments.

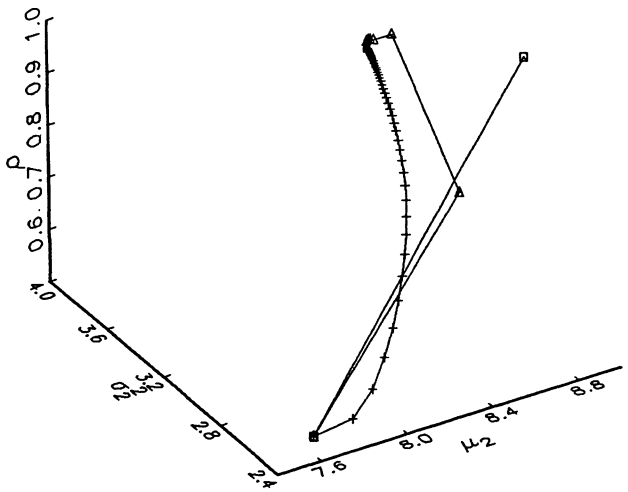

Figure 2. Sequences of EM (+), Aitken (□) and mixed (Δ) estimates where the Aitken sequence fails to converge. At the last point in the Aitken sequence shown, at the top right of the figure, the next step for the Aitken method would be out of the parameter space.

REFERENCES

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion), JRSS B 39:1-38

Dennis, J.E. and Schnabel, R.B. (1983) Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Prentice-Hall Inc., Englewood Cliffs, NJ

Efron, B. and Hinkley, D. V. (1978) Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information, Biometrika 65(3):457-487

Little, R. J. A. and Rubin, D. B. (1987) Statistical Analysis With Missing Data, Wiley

Louis, T. A. (1982) Finding the Observed Information Matrix when Using the EM Algorithm, JRSS B 44(2):226-233

Meilijson, I. (1989) A Fast Improvement to the EM Algorithm on its Own Terms, JRSS B 51:127-138

Meng, X. and Rubin, D. B. (1989) Obtaining Asymptotic Variance-Covariance Matrices By The EM Algorithm, Dept. of Statistics, Harvard University, Cambridge, MA 02138

Wu, J. (1983) On the Convergence Properties of the EM Algorithm, The Annals of Stat. 11(1):95-103