BU-1170-M

CONVERGENCE AND INVARIANCE PROPERTIES OF THE EM ALGORITHM

David Lansky, Loyola University Medical Center
G. Casella and C. McCulloch, Cornell University, Biometrics Unit.
D. Lansky, Preclinical Statistics, Searle, 4901 Searle Pkwy, Skokie, IL 60077

KEY WORDS: Maximum likelihood, Posterior mode, Aitken's acceleration

ABSTRACT

The EM algorithm is often used to advantage on difficult maximum likelihood problems where there is a similar, easier problem which requires more data. Unfortunately, EM is often slow to converge, particularly near the end of the iterative sequence of estimates. We show that EM is strongly invariant to the parameterization chosen. The Aitken acceleration, which uses the Jacobian of the EM mapping ($\boldsymbol{J}$), can improve the convergence rate, particularly late in the EM sequence. However, Aitken's accelerated EM generally has a smaller radius of convergence than EM. Significant obstacles to regular use of Aitken's acceleration include: 1) no reliable predictors of convergence for Aitken's accelerated EM, and 2) approximating $\boldsymbol{J}$ can be computationally expensive. We present results on the convergence rate of $\boldsymbol{J}$ in exponential family problems which support variations on Aitken's acceleration methods. Our results provide a theoretical basis for improved acceleration methods for EM.

INTRODUCTION

Aside from problems caused by difficult likelihood surfaces, the major disadvantages of EM are its slow convergence and the lack of a built-in variance estimator. The latter problem can be addressed by any of several procedures that correct the complete data observed information after EM has converged (Louis, 1982; Meng and Rubin, 1991). Variance correction, slow convergence and Aitken's acceleration of EM are all related through the derivative of the EM mapping $\boldsymbol{DM}$. Application of Aitken's acceleration to EM is largely based on methods derived from experience (for example see Laird et al, 1987). Our results on the invariance of EM, the convergence rate of EM and the convergence rate of $\boldsymbol{DM}$ should allow a more organized approach to acceleration of EM. Technically, EM is a q-linear algorithm (Dennis and Schnabel, 1983) with a rate constant that can be near 1 (if the rate constant is equal to 1 the sequence does not move at all).

DEFINITION OF EM & NOTATION

EM is formally defined, following Dempster, Laird and Rubin (1977) (subsequently denoted as DLR), by postulating two sample spaces, $\mathfrak{X}$ and $\mathfrak{Y}$, and a many-to-one mapping from $\mathfrak{X}$ to $\mathfrak{Y}$; the observed (or "incomplete") data $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ are a realization from $\mathfrak{Y}$. The subset of $\mathfrak{X}$ in which $\boldsymbol{x}$, the "complete" data, lies is denoted $\mathfrak{X}(\boldsymbol{y}) = \{\boldsymbol{x}:\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})\}$. The family $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ induces a related family $g(\boldsymbol{y} \mid \boldsymbol{\theta})$,

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

The common parameterization, $\boldsymbol{\theta}$, for f and g is an essential feature of the EM setup. We represent EM as a mapping from a parameter estimate, $\boldsymbol{\theta}^{(k)}$, to a new estimate $\boldsymbol{\theta}^{(k+1)}$,

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

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

$$\boldsymbol{J}(\boldsymbol{\theta}) = \{d_{ij}\}, \quad \text{where } d_{ij} = \frac{\partial M(\boldsymbol{\theta})_i}{\partial \boldsymbol{\theta}_j},$$

$$\text{for } i,\ j = 1, \ldots, p.$$

Note that $\boldsymbol{J}(\boldsymbol{\theta})$, a p by p matrix, is the transpose of the matrix $\boldsymbol{DM}(\boldsymbol{\theta})$ that is used by DLR (1977) and Meng and Rubin (1989 and 1991). The formal definition of EM (DLR, 1977) does not refer to Expectation and Maximization steps; it is based on maximizing the expected log of the complete data likelihood at each step, so that

$$\boldsymbol{\theta}^{(k+1)} = M\left(\boldsymbol{\theta}^{(k)}\right) =$$

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

where

$$Q(\theta' \mid \theta) = E\left(\log f(x \mid \theta') \mid y, \theta\right).$$

DLR (1977) also define generalized EM (GEM) by relaxing the maximization to

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

we suggest that this should be called E generalized M because the maximization (M step) is being generalized. For exponential family problems, DM is the ratio of the information for the missing data conditional on the observed data to the unconditional observed information for the complete data (DLR, 1977; Louis, 1982; Meilijson, 1989; Meng and Rubin, 1989).

The derivative of the EM mapping has interesting and useful properties from both algorithmic and statistical viewpoints. For likelihoods, $f(\theta)$, with continuous derivatives in θ on the interior of the support, the mapping $M(\theta)$ and its derivative $DM(\theta)$ are also continuous (for f continuous and differentiable in θ, then $\log f$, the integral of $\log f$, and the maximum of $Q(\theta' \mid \theta)$) share these properties. The derivative of the mapping, $DM(\theta^{(k)})$, converges to $DM(\theta^*)$ as $\theta^{(k)} \rightarrow \theta^*$, where θ^* is a fixed point of the EM mapping M, and, we hope, a local maximum of the likelihood surface. The largest eigenvalue of DM gives the asymptotic rate of convergence of EM (DLR, 1977).

Meng and Rubin (1991) suggest that a transformation of the parameters can improve the accuracy of the asymptotic variance estimates for the final MLEs. Additionally, they argue that transformations can improve the stability of their Supplemented EM (SEM) and the numerical stability of EM. They state that the rate of convergence of EM is invariant to one-to-one differentiable transformations of the parameters. We agree with their statements about EM; we will present and discuss a stronger invariance property of EM.

Aitken's acceleration is often applied to EM, following Louis (1982). The Aitken accelerated EM estimate, $\theta_{\mathrm{A}}^{(k+1)}$, is given by

$$\theta_{\mathrm{A}}^{(k+1)} = M(\theta^{(k)}) + (I - DM(\theta^*))^{-1}(M(\theta^{(k)}) - \theta^{(k)}),$$

where I is the identity matrix. Generally $DM(\theta^*)$ is not available because we do not know θ^*; a common practice is to use $DM(\theta^{(k)})$ as an expensive (in computing time) approximation to $DM(\theta^*)$. When Aitken's acceleration is used, the rules for starting Aitken's and how often to use an accelerated step are ad hoc (Laird, et al, 1987).

Aitken's acceleration of EM can yield an algorithm with a q-quadratic convergence rate when it converges. Meilijson (1989) writes the Aitken step in a way that suggests a quasi-Newton step which provides an appealing motivation for construction of an updating scheme for $DM(\theta^{(k)})$. Unfortunately, this updating scheme assumes that $I - DM(\theta^*)$ is symmetric and close to the identity matrix. The fact that these conditions are generally not satisfied may explain why Aitken's acceleration has not been implemented with an update step. In any case, most quasi-Newton methods would only assure convergence of $\theta^{(k)}$, not convergence of $DM(\theta^{(k)})$. Because we are specifically interested in $DM(\theta^{(k)})$, we would like our approximation to converge to $DM(\theta^*)$.

INVARIANCE PROPERTIES

THEOREM 1.

Let $\{\theta^{(k)}\}$ represent a sequence of EM steps and let $\phi^{(k)} = c(\theta^{(k)})$, with $\phi^* = \theta^*$, where $c(\cdot)$ is one-to-one, differentiable, and has full rank matrix derivative $DM_{\frac{\partial \theta}{\partial \phi}}$. Further suppose that both $f(x \mid \theta)$ and $g(y \mid \theta)$ are members of the exponential family. Then, the eigenvalues of $DM(\theta^*)$ are the same as the eigenvalues of $DM(\phi^*)$, where $\theta^* \equiv \lim_{k \rightarrow \infty} \{\theta^{(k)}\}$.

PROOF

For the exponential family, the matrix DM, at the maximum likelihood estimator, θ^* or ϕ^*, is a function of the Fisher information for the complete data, I_x, and the incomplete data, I_y,

$$DM = \left(I_x - I_y\right) I_x^{-1}, \tag{1}$$

(DLR, 1977 and Meng and Rubin, 1991). The Fisher information, I_ϕ, is related to I_θ via

$$I_\phi = J_{\frac{\partial \theta}{\partial \phi}} I_\theta J'_{\frac{\partial \theta}{\partial \phi}}, \tag{2}$$

where $J_{\frac{\partial\theta}{\partial\phi}}$ is the Jacobian of the transformation from θ to ϕ. Using (1) and (2) write $DM(\phi^*)$ as a function of $DM(\theta^*)$ and J;

$$DM(\phi^*) = J\, DM(\theta^*)\, J^{-1}. \tag{3}$$

This shows that $DM(\phi^*)$ and $DM(\theta^*)$, are similar matrices (Searle, 1982); and therefore have identical eigenvalues. □

Because the convergence rate of EM is given by the largest eigenvalue of DM (DLR, 1977), the invariance of eigenvalues of DM to one-to-one differentiable transformations do yield invariant convergence rates as Meng and Rubin (1991) state, yet the complete result above is stronger than their statement. The next theorem will establish a stronger invariance property of EM. The invariance property of maximum likelihood estimators (see, for example Casella and Berger, 1990) applies directly to each conditional maximization in the EM sequence; hence, the entire sequence is MLE-invariant to transformations. This next theorem is restricted to one-to-one and onto transformations to ensure that the conditioning is unique under the transformation and its inverse.

THEOREM 2.

Assume a one-to-one and onto transformation c, such that $\phi = c(\theta)$ and $\theta = c^{-1}(\phi)$. Let $M^*(\phi)$ and $L^*(\phi \mid x)$ denote the EM mapping and likelihood in the transformed parameter space. Further assume that, for any y and $\theta^{(k)}$

$$Q(\theta \mid \theta^{(k)}) = \int_{\mathfrak{B}(y)} \log L(\theta \mid x)\; L(\theta^{(k)} \mid x)\, dx$$

has a unique maximum. Then the EM mapping and $c(\cdot)$ are commutative in the sense that

$$\phi^{(k+1)} \equiv M^*(c(\theta^{(k)})) = c(M(\theta^{(k)})).$$

PROOF:

The set of values of θ which maximize Q then correspond to a set of values of ϕ which maximize Q^*. The maximum of the likelihood in the parameter space, and the unimodality of the likelihood yield a unique maximizer of Q. With a one-to-one transformation, this maximum corresponds to the unique maximizer of Q^*

$$\begin{aligned}
\phi^{(k+1)} &= \operatorname*{Argmax}_{\phi} Q^*(\phi \mid \phi^{(k)}) \\
&= \{\phi : Q^*(\phi \mid \phi^{(k)}) \geq Q^*(\tilde{\phi} \mid \phi^{(k)})\ \forall\ \tilde{\phi}\,\} \\
&\equiv \operatorname*{Argmax}_{\phi} \int_{\mathfrak{B}(y)} \log L^*(\phi \mid x)\; L^*(\phi^{(k)} \mid x)\, dx \\
&= c\Big(\operatorname*{Argmax}_{\theta} \int_{\mathfrak{B}(y)} \log L(\theta \mid x)\; L(\theta^{(k)} \mid x)\, dx\Big) \\
&= c(M(\theta^{(k)})).
\end{aligned}$$

An induction argument then completes the proof:

1) $\phi^{(0)} = c(\theta^{(0)})$ and $\theta^{(0)} = c^{-1}(\phi^{(0)})$

2) apply the above argument to any other value of k, then by induction to all k. □

In general, to get from the maximum of a likelihood surface to a maximum likelihood estimate requires the standard, difficult to check, EM assumptions on f that Wu (1983) imposed for GEM convergence. Transformations of θ which are not one-to-one will not necessarily yield invariant EM sequences. For example, with bivariate normal data $y \sim N(\mu, \sigma^2)$, where we transform μ to μ^2; under the μ parameterization, $E(\sum x_1) = (n_{x_1} - n_{y_1})\mu_1^{(k)} + \sum y_1$, but for the μ^2 parameterization the expectation, $E(\sum x_1) = \sum y_1$, where x is the complete data. The invariance result may not be fully applicable in practice, as a simple transformation such as $c(\theta) = (1 \times 10^7) \times \theta$, could cause overflow problems.

Even though these conditions are quite restrictive, this is a surprisingly strong property for a sequence. For transformations that are not one-to-one, but are locally one-to-one, this result will apply locally. These strong invariance properties indicate that we can, under certain restrictions, choose the parameterization that is most convenient at each stage in the EM sequence, perhaps using different parameterizations as the sequence approaches convergence or when we start using an acceleration scheme.

CONVERGENCE OF $DM(\theta^{(k)})$

For complete data densities that have two or more continuous derivatives (i.e. exponential families), the derivative of the EM mapping at $\theta^{(k)}$ converges to the derivative of the EM

mapping at θ^*. Motivated by the observation that EM tends to take many small steps with little change in direction at each step (in our simulations and those described by Lindstrom and Bates, 1988) we will compare the convergence rate of $\{DM(\theta^{(k)})\}$ to the convergence rate of $\{\theta^{(k)}\}$. For exponential family problems we show that the asymptotic convergence rate of DM is the same as the asymptotic convergence rate of $\theta^{(k)}$.

Following Dennis and Schnabel (1983), the quotient or q-rate of convergence, R, of the sequence $DM(\theta^{(k)})$, is given by

$$R = \lim_{k\to\infty} \frac{\left\| DM(\theta^{(k+1)}) - DM(\theta^*) \right\|}{\left\| DM(\theta^{(k)}) - DM(\theta^*) \right\|}, \qquad (4)$$

for a given norm; we will use a Frobenius norm

$$\| A \|_F \equiv \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij}^2}.$$

Before establishing the conditions such that the convergence rate of $\{DM^{(k)}\}$ is the same as the convergence rate of $\{\theta^{(k)}\}$, we will construct a general expression for the convergence rate of $\{DM^{(k)}\}$. A Taylor series approximation of the matrix DM is given by

$$DM(\theta^{(k)}) = DM(\theta^*)$$

$$+ \left\{ \left(\frac{\partial DM(\theta^*)_{ij}}{\partial \theta'} \right) (\theta^{(k)} - \theta^*) \right\}_{i,j=1..p} + \varepsilon^{(k)}, \qquad (5)$$

where $\varepsilon^{(k)} \le b\left((\theta^{(k)} - \theta^*)^2\right)$ or $O\left((\theta^{(k)} - \theta^*)^2\right)$.

THEOREM 3.

For an instance of the EM algorithm (with parameter θ), the Frobenius norm asymptotic convergence rate of $\{DM(\theta^{(k)})\}$

$$R_{DM} = \lim_{k\to\infty} \frac{\left\| DM(\theta^{(k+1)}) - DM(\theta^*) \right\|_F}{\left\| DM(\theta^{(k)}) - DM(\theta^*) \right\|_F}$$

is

$$= \lim_{k\to\infty} \sqrt{\frac{(\theta^{(k+1)} - \theta^*) B (\theta^{(k+1)} - \theta^*)}{(\theta^{(k)} - \theta^*) B (\theta^{(k)} - \theta^*)}},$$

where $B = \sum_{i=1}^{p} \sum_{j=1}^{p} B_{ij}$, with

$$B_{ij} = \left(\frac{\partial DM(\theta^*)_{ij}}{\partial \theta} \right) \left(\frac{\partial DM(\theta^*)_{ij}}{\partial \theta'} \right).$$

PROOF

The asymptotic convergence rate of $\{DM(\theta^{(k)})\}$, from (4), is

$$R_{DM} = \lim_{k\to\infty} \frac{\left\| DM(\theta^{(k+1)}) - DM(\theta^*) \right\|_F}{\left\| DM(\theta^{(k)}) - DM(\theta^*) \right\|_F}.$$

Substituting (5) into our expression for R_{DM} we obtain

$$R_{DM} =$$

$$\lim_{k\to\infty} \frac{\left\| \left\{ \left(\frac{\partial DM(\theta^*)_{ij}}{\partial \theta'} \right) (\theta^{(k+1)} - \theta^*) \right\}_{i,j=1..p} + \varepsilon^{(k+1)} \right\|_F}{\left\| \left\{ \left(\frac{\partial DM(\theta^*)_{ij}}{\partial \theta'} \right) (\theta^{(k)} - \theta^*) \right\}_{i,j=1..p} + \varepsilon^{(k)} \right\|_F}.$$

Because $DM(\theta^*)$ is fixed but not zero so is $\partial DM(\theta^*)_{ij}/\partial\theta'$; meanwhile

$$(\theta^{(k+1)} - \theta^*) \underset{k\to\infty}{\longrightarrow} 0,$$

hence $\varepsilon^{(k)}$ and $\varepsilon^{(k+1)}$ terms will be small compared to the first term, $O\left((\theta^{(k)} - \theta^*)^2\right)$ and $O\left((\theta^{(k+1)} - \theta^*)^2\right)$ respectively. Thus, we can ignore the $\varepsilon^{(k)}$ and $\varepsilon^{(k+1)}$ for $\theta^{(k)}$ close enough to θ^*. The squared Frobenius norm of the denominator then becomes,

$$\sum_{i=1}^{p} \sum_{j=1}^{p} d^{(k)\prime} \left(\frac{\partial DM(\theta^*)_{ij}}{\partial \theta} \right) \left(\frac{\partial DM(\theta^*)_{ij}}{\partial \theta'} \right) d^{(k)}$$

$$= d^{(k)\prime} B d^{(k)}, \text{ for } d = (\theta^{(k)} - \theta^*) \text{ and } B \text{ as}$$

defined in the hypotheses. Substitution norms in the numerator and denominator will finish the proof. □

We are now ready to show that the convergence rate of $DM(\theta^{(k)})$ matches that of $\theta^{(k)}$. For members of the exponential family, apply the result from Theorem 1 to the result of Theorem 3. The result is that the asymptotic convergence rate of DM is the same as the asymptotic convergence rate of the parameter for exponential families EM problems. Further, for members of a regular exponential family, the convergence rates of the EM sequence $\{\theta^{(k)}\}$ and the associated sequence $\{DM(\theta^{(k)})\}$ are both are invariant to one-to-one transformations of the parameter θ.

THEOREM 4

For any EM sequence from an exponential family the asymptotic convergence rates of $\{DM(\theta^{(k)})\}$ and $\{\theta^{(k)}\}$ are the same. That is,

$$R_{DM} = \lim_{k\to\infty} \frac{\left\| DM(\theta^{(k+1)}) - DM(\theta^*) \right\|_F}{\left\| DM(\theta^{(k)}) - DM(\theta^*) \right\|_F}$$

$$= \lim_{k\to\infty} \frac{\left\| \theta^{(k+1)} - \theta^* \right\|_F}{\left\| \theta^{(k)} - \theta^* \right\|_F} = R_{\theta}.$$

PROOF

Apply Theorem 3 and let $K = UD^{1/2}V'$, for $B = UDV'$ so that $K'K = B$. We now have, $\phi = K\theta$ and $\theta = K^{-1}\phi$. We can now write

$$R_{DM} = \lim_{k\to\infty} \sqrt{\frac{(\theta^{(k+1)} - \theta^*)K'K(\theta^{(k+1)} - \theta^*)}{(\theta^{(k)} - \theta^*)K'K(\theta^{(k)} - \theta^*)}}$$

$$= \lim_{k\to\infty} \frac{\left\| \phi^{(k+1)} - \phi^* \right\|_F}{\left\| \phi^{(k)} - \phi^* \right\|_F} = R_{\phi} = R_{\theta},$$

where the last equality is assured by Theorem 1 and that the convergence rate of EM is given by the largest eigenvalue of DM (DLR, 1977).□

SIMULATION ON SMALL SAMPLES

To demonstrate that the asymptotic (in k) results of Theorem 4 are useful for values of k small enough that $\left\| \theta^{(k)} - \theta^* \right\|$ is not trivial we studied a simple example. We compared the convergence behavior of $\{\theta^{(k)}\}$ with the convergence behavior of $\{DM(\theta^{(k)})\}$ in a bivariate normal problem.

For bivariate normal data with some observations missing one member of the pair (all pairs with missing data have the same element missing), we examined four parameterizations. We chose this example because there is an exact solution to the incomplete data likelihood, hence we can directly evaluate θ^* and $DM(\theta^*)$. The four parameterizations can be named the correlation (θ), covariance (β), regression ($\overset{\circ}{\beta}$) and natural parameterization (ϕ), with parameters that are related by

$$\theta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\};$$
$$\beta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12}\},$$

where $\sigma_{12} = \rho\sqrt{\sigma_1^2\sigma_2^2}$;

$$\overset{\circ}{\beta} = \{\mu_1, \sigma_1^2, \beta_{2.0}, \beta_{2.01}, \sigma_{2.21}^2\},$$

where $\mu_2 = \beta_{2.0} + \beta_{2.01}\mu_1$,

$$\sigma_2^2 = \sigma_{2.21}^2 + \beta_{2.01}^2\sigma_1^2,$$

$$\rho = \beta_{2.01}\sqrt{\sigma_1^2 / \sigma_2^2};$$

and finally,

$$\phi = \{(V^{-1}\mu)_1, (V^{-1}\mu)_2, -(V^{-1})_{1,1},$$
$$-(V^{-1})_{2,2}, -(V^{-1})_{1,2}\},$$

where $\mu = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$, and $V = \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}$.

All numerical approximations of $DM(\theta^{(k)})$ used a fixed step size, forward difference method. The forward difference numerical approximation to $DM(\theta^*)$ is not identical to the closed-form value of $DM(\theta^*)$ (the norm of this difference is approximately 10^{-5}). When the difference between DM^* and $DM^{(k)}$ becomes much larger than the difference between $DM^{(k)}$ and $DM^{(k+1)}$ the convergence rate for DM goes to one (Figure 1). If one were using a lack-of-progress convergence criterion (Bates and Watts, 1981) with a tolerance of 10^{-5}, the EM and DM sequences would stop before this point. Hence, for the bivariate normal example, in each parameterization, the convergence rate of $DM(\theta^{(k)})$ is very close to the convergence rate of $\theta^{(k)}$ at values of k where we can make practical use of this convergence property (Figure 1). Figures for all four parameterizations are similar. For more simulation results and additional details see Lansky et al. (1992).

Our results on the convergence rate of DM support the practice of using recent relative changes in the slow-changing components of θ to approximate λ, the largest eigenvalue of $DM(\theta^*)$, as suggested by Laird et al. (1987). Further, our results suggest that the Laird et al. method could be extended when different components of θ are changing at different rates, to estimate other eigenvalues of $DM(\theta^*)$, as long as the relative change for each component is stable over several EM steps.

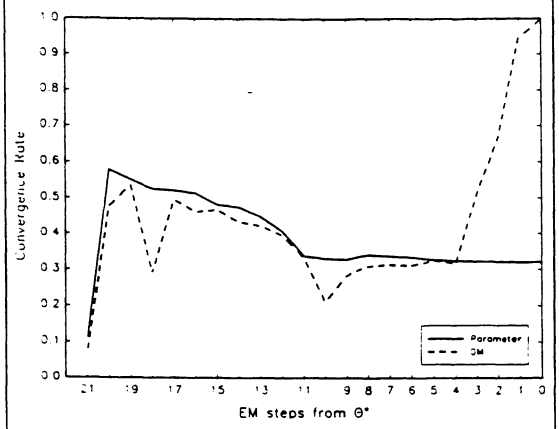

Figure 1. The median convergence rate of the forward difference numerical approximation to $DM(\theta^{(k)})$ compared to the median convergence rate for $\theta^{(k)}$ for 50 samples from the bivariate normal problem.

$$R_{\boldsymbol{x}}^{k} = \left\| \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{*} \right\| \Big/ \left\| \boldsymbol{x}^{(k)} - \boldsymbol{x}^{*} \right\|,$$

where $\boldsymbol{x}$ is either $\boldsymbol{\theta}$ or $DM(\boldsymbol{\theta})$. The numerical approximation to $DM(\boldsymbol{\theta}^{*})$ does not equal $DM(\boldsymbol{\theta}^{*})$, the norm of the difference is about 1×10^{-5}, this explains the rise in R_{DM}^{k} on the right of the figure.

ACKNOWLEDGEMENTS

Carlos Castillo-Chavez has helped both with his encouragement and with financial support from NSF grant DMS-8906580. Ziding Feng (Fred Hutchinson Cancer Research center) and Art Roth (Searle) both made suggestions which improved this manuscript.

REFERENCES

Bates, D. M. and Watts, D. G. (1981) A Relative Offset Orthogonality Convergence Criterion for Nonlinear Least Squares, Technometrics 23(2):179-183.

Casella, G. and Berger, R. L. (1990) *Statistical Inference,* Wadsworth and Brooks/Cole, Pacific Grove, CA.

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.

Lansky, D. M., Casella, G. and McCulloch C. (1992) Convergence and Invariance of the EM Algorithm, Cornell University Biometrics Unit BU-1170-M.

Laird, N. M., Lange, N. and Stram, D. (1987) Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm. *JASA* 82(397):97-105.

Lindstrom, M. J. and Bates, D. M. (1988) Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data, *JASA* 83(404):1014-1022.

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.

Meng, X. and Rubin, D. B. (1991) Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm, *JASA* 86(416):899-909.

Peters, B. C. and Walker, H. F. (1978) An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions. *SIAM J. Appl. Math* 35(2):362-378.

Redner, R. A. and Walker, H. F. (1984) Mixture Densities, Maximum Likelihood and The EM Algorithm, *SIAM Review* 26(2):195-239.

Searle, S. R. (1982) *Matrix Algebra Useful for Statistics*, Wiley, New York, NY.