IMS Collections
Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton
Vol. 10 (2013) 43–62

DOI: 10.1214/12-IMSCOLL1003

On Convergence Properties of the Monte Carlo EM Algorithm

Ronald C. Neath

Hunter College, City University of New York

Abstract: The Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) is a popular method for computing maximum likelihood estimates (MLEs) in problems with missing data. Each iteration of the algorithm formally consists of an E-step: evaluate the expected complete-data log-likelihood given the observed data, with expectation taken at current parameter estimate; and an M-step: maximize the resulting expression to find the updated estimate. Conditions that guarantee convergence of the EM sequence to a unique MLE were found by Boyles (1983) and Wu (1983). In complicated models for high-dimensional data, it is common to encounter an intractable integral in the E-step. The Monte Carlo EM algorithm of Wei and Tanner (1990) works around this difficulty by maximizing instead a Monte Carlo approximation to the appropriate conditional expectation. Convergence properties of Monte Carlo EM have been studied, most notably, by Chan and Ledolter (1995) and Fort and Moulines (2003).

The goal of this review paper is to provide an accessible but rigorous introduction to the convergence properties of EM and Monte Carlo EM. No previous knowledge of the EM algorithm is assumed. We demonstrate the implementation of EM and Monte Carlo EM in two simple but realistic examples. We show that if the EM algorithm converges it converges to a stationary point of the likelihood, and that the rate of convergence is linear at best. For Monte Carlo EM we present a readable proof of the main result of Chan and Ledolter (1995), and state without proof the conclusions of Fort and Moulines (2003). An important practical implication of Fort and Moulines's (2003) result relates to the determination of Monte Carlo sample sizes in MCEM; we provide a brief review of the literature (Booth and Hobert, 1999; Caffo, Jank and Jones, 2005) on that problem.

1. Introduction: The Monte Carlo EM algorithm

The expectation-maximization, or EM algorithm, is an algorithm for maximizing likelihood functions, especially in the presence of missing data. When EM works, the algorithm's output is a sequence of parameter values that converges to the maximum likelihood estimate (MLE). The seminal paper on EM, and that which gave the algorithm its name, is the article by Dempster, Laird and Rubin (1977). A book length treatment is given by McLachlan and Krishnan (1997).

Consider a statistical model in which the random vector (Y, U), $Y \in \mathbb{R}^N$ and $U \in \mathbb{R}^q$, has distribution given by $f(y, u; \theta)$, a density with respect to the measure $\lambda \times \mu$, where λ and μ are measures on $\mathbb{R}^N$ and $\mathbb{R}^q$ respectively, and indexed by the unknown parameter $\theta \in \Theta$. We refer to (Y, U) as the "complete data" but only

Department of Mathematics and Statistics Hunter College, City University of New York, e-mail: rneath@hunter.cuny.edu

AMS 2000 subject classifications: Primary 62-02

Keywords and phrases: convergence, EM algorithm, maximum likelihood, mixed model, Monte Carlo

$Y = y$ is observed; U represents the unobserved or "missing" data. The MLE of θ is the value $\hat{\theta}$ which maximizes the likelihood function

$$L(\theta; y) = \int f(y, u; \theta)\mu(du) \tag{1}$$

or, equivalently, the log likelihood $l(\theta; y) = \log L(\theta; y)$. The EM algorithm can be used to find $\hat{\theta}$ even if the integral in (1) is intractable. Define the Q-function, a mapping on $\Theta \times \Theta$, by

$$Q(\theta|\tilde{\theta}; y) = \mathrm{E}\left\{\log f(y, U; \theta) \mid y; \tilde{\theta}\right\} , \tag{2}$$

that is, the expected value of the "complete data" log-likelihood at θ, given the observed data, this conditional expectation evaluated under $\tilde{\theta}$. Each EM iteration formally consists of an E-step, to evaluate the conditional expectation in (2), and an M-step, to maximize it. More precisely, if $\theta^{(t)}$ is the parameter value as of the tth iteration, the update $\theta^{(t+1)}$ is chosen such that $Q(\theta^{(t+1)}|\theta^{(t)}; y) \geq Q(\theta|\theta^{(t)}; y)$ for all $\theta \in \Theta$. Under regularity conditions (Boyles, 1983; Wu, 1983, and see Section 3 below), and given a suitable starting value $\theta^{(0)}$, the resulting sequence $\left\{\theta^{(t)} : t = 0, 1, \ldots\right\}$ will converge to a local maximizer of L.

If the integral in (2) admits a closed form solution, the implementation of EM is straightforward (though the M-step may still require a numerical optimization scheme such as Newton-Raphson). Suppose it does not. As noted, the evaluation of (2) requires taking an expectation with respect to the conditional distribution of the missing data U, given observed data $Y = y$. If one has the means to simulate random draws from this target distribution, the Q-function can be approximated by Monte Carlo integration. Let $u^{(1)}, \ldots, u^{(m)}$ denote a random sample from $h(u|y; \tilde{\theta}) = f(y, u; \tilde{\theta})/L(\tilde{\theta}; y)$. Then a Monte Carlo approximation to (2) is given by

$$Q_m(\theta|\tilde{\theta}; y) = \frac{1}{m}\sum_{k=1}^{m} \log f(y, u^{(k)}; \theta) .$$

In the Monte Carlo EM algorithm (MCEM), first introduced by Wei and Tanner (1990), the update $\theta^{(t+1)}$ is the value of θ that maximizes $Q_m(\theta|\theta^{(t)}; y)$.

Applications of EM and MCEM have been numerous; in this work we focus on one in particular, the two-stage hierarchical model, introduced in Section 2. We give two simple but realistic examples from this class of models, and demonstrate the implementation of EM and MCEM in those two problems. In Section 3 we discuss convergence properties of the EM algorithm. Of course, the question of convergence for MCEM is far more complicated, and an accessible discussion of the major results in this area is the main objective of this review paper. In Section 4 we provide a rigorous but accessible review of the two seminal papers on MCEM convergence, those of Chan and Ledolter (1995) and Fort and Moulines (2003). We make some concluding remarks in Section 5.

2. Application: The two-stage hierarchical model

Let $Y = (Y_1, \ldots, Y_N)^T$, where each Y_i is a random variable in $\mathbb{R}^1$, denote the observable data. In a *two-stage hierarchical model*, the distribution of Y is specified conditionally on some unobservable random quantity $U = (U_1, \ldots, U_q)^T$. Specifically, we assume that conditional on $U = u$, the Y_i are independent with conditional

densities denoted by $f_i(y_i|u_i;\theta_1)$, where $\theta_1 \in \Theta_1$ is an unknown parameter and each f_i is a density with respect to Lebesgue or counting measure. The f_i may also depend on an observable covariate x_i though this dependence is suppressed in our notation. Define $f(y|u;\theta_1) = \prod_{i=1}^N f_i(y_i|u;\theta_1)$, a density on $\mathbb{R}^N$, and this completes specification of the first level, or *stage*, of the hierarchy. At the second stage we specify a marginal distribution for U, defined by $h(u;\theta_2)$, a density on $\mathbb{R}^q$ that depends on the unknown parameter $\theta_2 \in \Theta_2$. Assume the parameter spaces Θ_1 and Θ_2 are open subsets of $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Let $d = d_1 + d_2$. The unknown parameter $\theta = (\theta_1, \theta_2)$ lies in the parameter space $\Theta = \Theta_1 \times \Theta_2$, an open subset of $\mathbb{R}^d$.

Suppose we wish to compute maximum likelihood estimates (MLEs) of θ_1 and θ_2. Were the random effects U observable the likelihood function would be given by what we will call the *complete data likelihood* $L_c(\theta; y, u) = f(y|u;\theta_1)h(u;\theta_2)$. But since only the data Y are observed, the random effects must be integrated out of L_C yielding the likelihood function

$$L(\theta; y) = \int L_c(\theta; y, u)du = \int f(y|u;\theta_1)h(u;\theta_2)du \,. \tag{3}$$

We wish to find the value of θ that maximizes L, that is, the MLE $\hat{\theta}$.

It will most often be the case that the integral in (3) is intractable. Booth, Hobert and Jank (2001) provide a very nice summary of numerical and Monte Carlo methods available for maximum likelihood in this problem, arriving at the conclusion that "Monte Carlo EM is generally the simplest and most efficient Monte Carlo fitting algorithm for two-stage hierarchical models." As noted above, the EM algorithm is a general method for maximum likelihood in the presence of missing data; hierarchical models are cast in this light by viewing the unobserved random effects as "missing".

Let $l_c = \log L_c$ denote the complete data log likelihood, so

$$l_c(\theta; y, u) = \log f(y|u;\theta_1) + \log h(u;\theta_2) \,.$$

Thus in the setting of hierarchical models, the EM update rule introduced in Section 1 can be written

$$\begin{aligned} \theta_1^{(t+1)} &= \arg\max \mathrm{E}\left\{\log f(y|U;\theta_1)\middle|\; y;\theta^{(t)}\right\} \,, \\ \theta_2^{(t+1)} &= \arg\max \mathrm{E}\left\{\log h(U;\theta_2)\middle|\; y;\theta^{(t)}\right\} \,, \end{aligned} \tag{4}$$

that is, the update of θ_1 and that of θ_2 can be considered separately.

If one or both of the expectations in (4) is intractable, one might employ the Monte Carlo EM algorithm. The MCEM update rule for the two-stage hierarchical model is given here. Let $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})$ denote the current parameter value; then $\theta^{(t+1)}$ is found by

1. Simulate $u^{(t,1)}, \ldots, u^{(t,m)}$, a random sample from the conditional density $h(u|y;\theta^{(t)})$;
2. Compute updates

$$\begin{aligned} \theta_1^{(t+1)} &= \arg\max \left\{ \frac{1}{m} \sum_{k=1}^m \log f(y|u^{(t,k)};\theta_1) \right\} \\ \theta_2^{(t+1)} &= \arg\max \left\{ \frac{1}{m} \sum_{k=1}^m \log h(u^{(t,k)};\theta_2) \right\} \,. \end{aligned}$$

The "target density" for the Monte Carlo E-step (step 1) is the conditional density of the random effects given the data,

$$h(u|y;\theta) \propto f(y|u;\theta_1)h(u;\theta_2)\ . \tag{5}$$

If direct simulation from (5) is impossible, one might resort to a Markov chain Monte Carlo (MCMC) method such as the Metropolis-Hastings algorithm. In this case the sample $\left\{u^{(t,k)} : k = 1,\ldots,m\right\}$ is an ergodic Markov chain having $h(u|y;\theta^{(t)})$ as its unique stationary density (see, for example, Robert and Casella, 2004). An alternative approach is to compute a quasi-Monte Carlo or *randomized quasi-Monte Carlo* (L'Ecuyer and Lemieux, 2002) approximation to the Q-function with the goal of reducing Monte Carlo error and hence increasing the efficiency of the algorithm. We will not consider quasi-Monte Carlo methods any further in this report; the interested reader is referred to Jank (2004).

2.1. Example 1: A linear mixed model

Table 1 contains a data set for an experiment described by Snedecor and Cochran (1989). The experiment involved six bulls and very many cows. From each bull, some number of semen samples was taken, and each of these samples was used in an attempt to artificially inseminate a large number of cows. Some attempts were successful and some were not; let Y_{ij} denote the success rate (percentage of conceptions) for sample j from bull i, for $j = 1,\ldots,n_i$ and $i = 1,\ldots,q = 6$; here $N = \sum_{i=1}^{q} n_i$. Consider the one-way random effects model

$$y_{ij} = \mu + u_i + e_{ij}$$

where μ is the overall mean, u_i is the ith bull effect, and e_{ij} is a residual error term. As the six bulls were a random sample from a larger population of bulls, the u_i are modeled as independent and identically distributed (i.i.d.) random effects. Model specification is completed by a distribution assumption on the bull effect and error term; we take

$$u_i \sim \text{ iid Normal}\left(0,\sigma_u^2\right)\ ; \quad \text{independent of} \quad e_{ij} \sim \text{ iid Normal}\left(0,\sigma_e^2\right)\ .$$

When there exists a conjugate relationship between f and h, as in the normal linear mixed model, the integral in (3) can be solved explicitly. The resulting log-likelihood can be maximized numerically (or analytically in the case of balanced data $n_i \equiv n$); for the bulls data we obtain $\hat{\mu} = 53.318$, $\hat{\sigma}_u^2 = 54.821$, and $\hat{\sigma}_e^2 = 249.23$.

TABLE 1
Bovine artificial insemination data of Example 1 (Snedecor and Cochran, 1989)

Bull (i)	n_i	Percentage of conception
1	5	46, 31, 37, 62, 30
2	2	70, 59
3	7	52, 44, 57, 40, 67, 64, 70
4	5	47, 21, 70, 46, 14
5	7	42, 64, 50, 69, 77, 81, 87
6	9	35, 68, 59, 38, 57, 76, 57, 29, 60
Total	35	

Consider the EM algorithm. We find it more convenient to work with an equivalent version of the model in which $y_{ij} = u_i + e_{ij}$ and the u_i are i.i.d. Normal(μ, σ_u^2). Under this reparameterization the complete data log-likelihood of $\theta = (\mu, \sigma_u^2, \sigma_e^2)$ is

$$l_c(\theta; y, u) = -\frac{N}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{q}\sum_{j=1}^{n_i}(y_{ij} - u_i)^2 - \frac{q}{2}\log(\sigma_u^2) - \frac{1}{2\sigma_u^2}\sum_{i=1}^{q}(u_i - \mu)^2 \ .$$

Owing to the conjugacy it is straightforward to show that

$$(6)\quad U_i|(Y = y; \theta)\ \ i = 1, \ldots, q\ \text{ are indep Normal}\left(\frac{\sigma_e^2\mu + n_i\sigma_u^2\bar{y}_i}{\sigma_e^2 + n_i\sigma_u^2}, \frac{\sigma_e^2\sigma_u^2}{\sigma_e^2 + n_i\sigma_u^2}\right) \ .$$

Denote the conditional mean and variance of U_i given $Y = y$ by $\hat{u}_i$ and $\hat{V}_i$, respectively. Then the EM update rule is given by

$$\begin{aligned}
\mu^{(t+1)} &= \frac{1}{q}\sum_{i=1}^{q}\hat{u}_i^{(t)} \\
\sigma_u^{2^{(t+1)}} &= \frac{1}{q}\sum_{i=1}^{q}\left(\hat{V}_i^{(t)} + \left[\hat{u}_i^{(t)}\right]^2\right) - \left[\mu^{(t+1)}\right]^2 \\
\sigma_e^{2^{(t+1)}} &= \frac{1}{N}\sum_{i=1}^{q}\left[\sum_{j=1}^{n_i} y_{ij}^2 - 2n_i\bar{y}_i\hat{u}_i^{(t)} + n_i\left(\hat{V}_i^{(t)} + \left[\hat{u}_i^{(t)}\right]^2\right)\right] \ .
\end{aligned}$$

Given the existence of a closed form EM update, there is no practical reason to resort to Monte Carlo EM (indeed there was no practical need for EM, as we found a closed form expression for the likelihood as well), but we will consider MCEM for illustration. Let $u^{(t,1)}, \ldots, u^{(t,m)}$ denote a sequence of simulated draws from $h(u|y; \theta^{(t)})$, given at (6). The MCEM update rule for $\theta^{(t+1)}$ is

$$\begin{aligned}
\mu^{(t+1)} &= \frac{1}{mq}\sum_{k=1}^{m}\sum_{i=1}^{q} u_i^{(t,k)} \\
\sigma_u^{2^{(t+1)}} &= \frac{1}{mq}\sum_{k=1}^{m}\sum_{i=1}^{q}\left(u_i^{(t,k)} - \mu^{(t+1)}\right)^2 \\
\sigma_e^{2^{(t+1)}} &= \frac{1}{mN}\sum_{k=1}^{m}\sum_{i=1}^{q}\sum_{j=1}^{n_i}\left(y_{ij} - u_i^{(t,k)}\right)^2 \ .
\end{aligned}$$

We ran three independent MCEM runs of 20 iterations each, starting at the point $(\mu^{(0)}, \sigma_u^{2^{(0)}}, \sigma_e^{2^{(0)}}) = (55, 45, 260)$. For each update we used Monte Carlo sample size $m = 10^4$; results are shown in Figure 1. The three dashed lines indicate the paths of the three MCEM runs, and the solid line shows that of ordinary (deterministic) EM. We did three more runs with starting values closer to the MLE and using $m = 10^5$; those results are summarized in Figure 2.

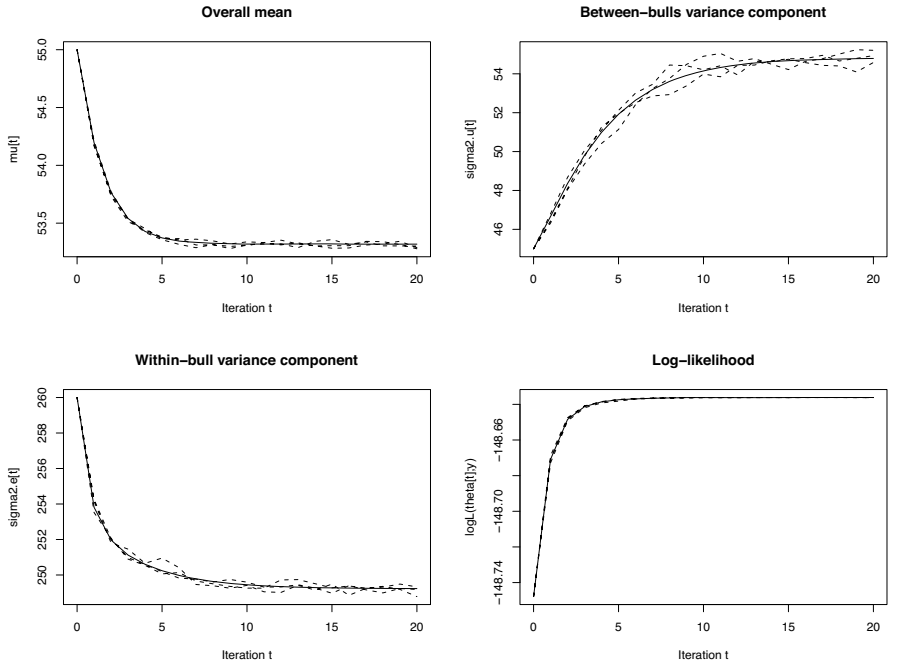

FIG 1. *Trace plots for Monte Carlo EM in Example 1, based on Monte Carlo sample size* $m = 10^4$ *at each iteration. Top left plot is overall mean* μ*, top right and bottom left are variance components* σ_u^2 *and* σ_e^2*, respectively. Bottom right plot shows log-likelihood evaluated at current parameter value. The solid line is deterministic EM and the three dashed lines correspond to three independent runs of Monte Carlo EM.*

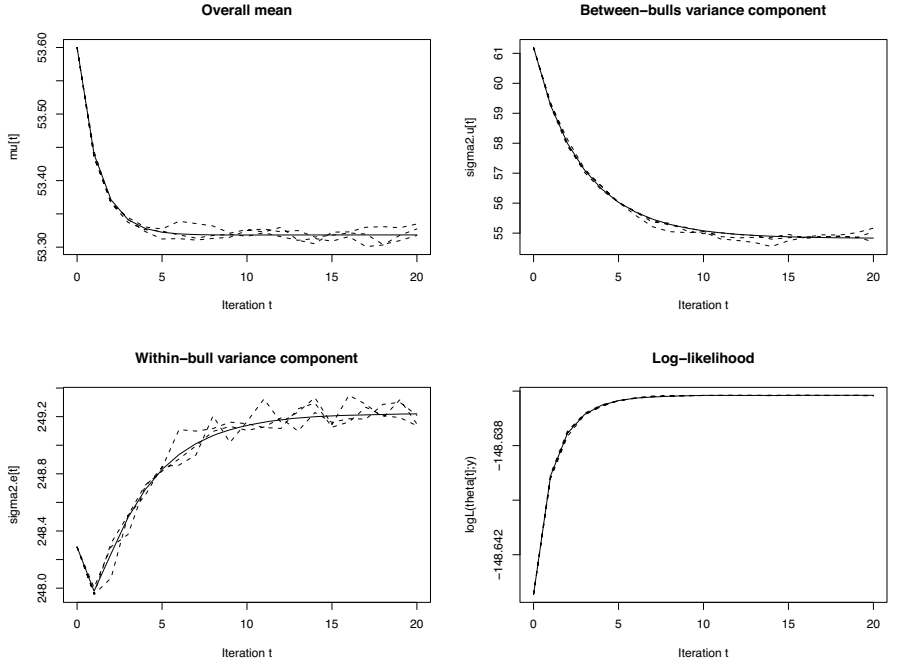

FIG 2. *Analogous to Figure 1, but with* $m = 10^5$ *and starting values chosen closer to the true MLE.*

2.2. Example 2: A logit-normal generalized linear mixed model

Let $Y = \{Y_{ij} : j = 1, \ldots, n_i;\ i = 1, \ldots, q\}$ denote a set of binary response variables; here again one can think of Y_{ij} as the jth response for the ith subject. Let x_{ij} be a covariate (or vector of covariates) associated with the i, j observation. Conditional on the random effects $U = u \in \mathbb{R}^q$, the responses are independent Bernoulli(π_{ij}) where

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta x_{ij} + u_i \ .$$

Let $U_1, \ldots, U_q$ be independent and identically distributed as Normal$(0, \sigma^2)$. The likelihood is given by

$$L(\beta, \sigma^2; y) = \left(\sigma^2\right)^{-q/2} \times \int_{\mathbb{R}^q} \exp\left\{\sum_{i=1}^{q}\sum_{j=1}^{n_i}\left[y_{ij}\left(\beta x_{ij} + u_i\right) - \log\left(1 + e^{\beta x_{ij}+u_i}\right)\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{q} u_i^2\right\} du.$$

The above model has been used by several authors (Booth and Hobert, 1999; Caffo, Jank and Jones, 2005; McCulloch, 1997) as a benchmark for comparing Monte Carlo methods of maximum likelihood. We consider here a data set generated by Booth and Hobert (1999, Table 2) with $n_i = 15$, $q = 10$, and $x_{ij} = j/15$ for each i, j. For these data the MLEs are known to be $(\hat{\beta}, \hat{\sigma}^2) = (6.132, 1.766)$.

A version of the complete data log-likelihood is given by

$$l_c(\beta, \sigma^2; y, u) = -\frac{q}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{q} u_i^2 + \sum_{i=1}^{q}\sum_{j=1}^{n_i}\left[\beta x_{ij} y_{ij} - \log\left(1 + e^{\beta x_{ij}+u_i}\right)\right] \ .$$

To apply the EM algorithm in this problem we would need to compute the (conditional) expectation of l_c with respect to the density

$$h(u|y;\theta) \propto \exp\left\{\sum_{i=1}^{q}\sum_{j=1}^{n_i}\left[y_{ij}u_i - \log\left(1 + e^{\beta x_{ij}+u_i}\right)\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{q} u_i^2\right\} \ . \tag{7}$$

Clearly this integral will be intractable. Thus we consider a Monte Carlo EM algorithm, which requires the means to simulate random draws from the distribution given by (7). McCulloch (1997) employed a variable-at-a-time Metropolis-Hastings independence sampler with Normal$(0, \sigma^2)$ proposals, which Johnson, Jones and Neath (2011) have shown is uniformly ergodic.

Trace plots for three independent runs of MCEM are shown in the left hand panels of Figure 3. The starting values for these runs were $(\beta^{(0)}, \sigma^{2(0)}) = (2, 1)$, and we ran 35 updates with Monte Carlo sample size $m = 10^4$ at each iteration. We conducted three more runs of 25 iterations, starting at $(\beta^{(0)}, \sigma^{2(0)}) = (6, 2)$, with $m = 10^5$; results are shown in the right hand panels of Figure 3.

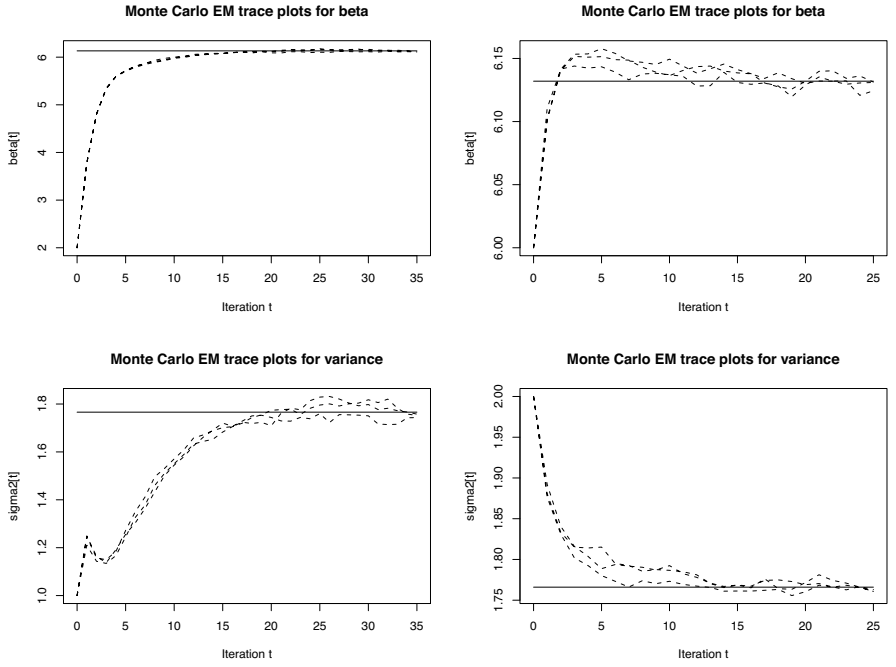

FIG 3. *Monte Carlo EM trace plots for logit-normal model of Example 2. Top panels show β, bottom panels show σ^2. Three dashed lines correspond to three independent runs of MCEM, with solid horizontal line drawn at true MLE. Runs in left hand panels used Monte Carlo sample size $m = 10^4$ at each iteration; in right hand panels we used $m = 10^5$ with starting values closer to the true MLE.*

3. Convergence properties of ordinary EM

The basic convergence properties of the EM algorithm were established by Boyles (1983) and Wu (1983). The presentation given here draws heavily from Geyer (1998). We will show that if an EM sequence converges, its limit must be a stationary point of the log-likelihood. We then present conditions that guarantee the convergence of EM, with additional conditions that guarantee convergence to the MLE. We conclude this section with a proof that the EM algorithm cannot produce a superlinearly convergent sequence.

We begin by proving the *ascent property* of the EM algorithm, which guarantees that an EM update will never decrease the value of the likelihood function, that is, if $\{\theta^{(t)}\}$ is an EM sequence, then $l(\theta^{(t+1)}; y) \geq l(\theta^{(t)}; y)$ for each t.

Define

$$\begin{aligned} R(\theta|\tilde{\theta}; y) &= \mathrm{E}\left\{\log h(U|y;\theta)\middle|\ y;\tilde{\theta}\right\} \\ &= \mathrm{E}\left\{\log f(y,U;\theta)\middle|\ y;\tilde{\theta}\right\} - \mathrm{E}\left\{\log f(y;\theta)\middle|\ y;\tilde{\theta}\right\} \\ &= Q(\theta|\tilde{\theta}; y) - l(\theta; y)\ . \end{aligned} \tag{8}$$

We now show that, for fixed $\tilde{\theta}$, $R(\theta|\tilde{\theta}; y)$ attains its maximum at $\theta = \tilde{\theta}$.

Lemma 1. *For any $\tilde{\theta} \in \Theta$, $R(\tilde{\theta}|\tilde{\theta}; y) \geq R(\theta|\tilde{\theta}; y)$ for all θ.*

Proof.

$$R(\theta|\tilde{\theta};y) - R(\tilde{\theta}|\tilde{\theta};y) = \mathrm{E}\left\{\log\left(\frac{h(U|y;\theta)}{h(U|y;\tilde{\theta})}\right)\Big|\ y;\ \tilde{\theta}\right\} \leq \log\left(\mathrm{E}\left\{\frac{h(U|y;\theta)}{h(U|y;\tilde{\theta})}\Big|\ y;\ \tilde{\theta}\right\}\right)$$

by the conditional Jensen inequality (see Billingsley, 1995, page 449); now

$$\mathrm{E}\left\{\frac{h(U|y;\theta)}{h(U|y;\tilde{\theta})}\Big|\ y;\ \tilde{\theta}\right\} = \int \frac{h(u|y;\theta)}{h(u|y;\tilde{\theta})} h(u|y;\tilde{\theta})du = \int h(u|y;\theta)du = 1$$

and thus $R(\theta|\tilde{\theta};y) - R(\tilde{\theta}|\tilde{\theta};y) \leq \log(1) = 0$. □

Theorem 1. *If* $Q(\theta|\tilde{\theta};y) \geq Q(\tilde{\theta}|\tilde{\theta};y)$*, then* $l(\theta;y) \geq l(\tilde{\theta};y)$*. If* $Q(\theta|\tilde{\theta};y) > Q(\tilde{\theta}|\tilde{\theta};y)$*, then* $l(\theta;y) > l(\tilde{\theta};y)$*.*

Proof. By (8) and Lemma 1,

$$\begin{aligned} l(\theta;y) - l(\tilde{\theta};y) &= Q(\theta|\tilde{\theta};y) - Q(\tilde{\theta}|\tilde{\theta};y) - \left[R(\theta|\tilde{\theta};y) - R(\tilde{\theta}|\tilde{\theta};y)\right] \\ &\geq Q(\theta|\tilde{\theta};y) - Q(\tilde{\theta}|\tilde{\theta};y) \end{aligned}$$

□

The ascent property of EM follows immediately from Theorem 1: since $\theta^{(t+1)}$ is chosen to maximize $Q(\theta|\theta^{(t)};y)$, it must be that $Q(\theta^{(t+1)}|\theta^{(t)};y) \geq Q(\theta^{(t)}|\theta^{(t)};y)$ and thus $l(\theta^{(t+1)};y) \geq l(\theta^{(t)};y)$. This is an appealing property, as it guarantees that an EM update will never take a step in the wrong direction. Of course, this result tells us absolutely nothing about the convergence of an EM sequence.

We now show that if an EM sequence converges, it converges to a stationary point of the log-likelihood. Unless otherwise noted, ∇ will denote differentiation with respect to the first argument.

Theorem 2. *Suppose the mapping* $(\theta, \tilde{\theta}) \mapsto \nabla Q(\theta|\tilde{\theta};y)$ *is jointly continuous. If* θ^* *is the limit of an EM sequence* $\{\theta^{(t)}\}$*, then* $\nabla l(\theta^*;y) = 0$*.*

Proof. Since $\theta^{(t+1)}$ maximizes $Q(\theta|\theta^{(t)};y)$ at each t we have $\nabla Q(\theta^{(t+1)}|\theta^{(t)};y) = 0$ at each t. By the continuity assumption $\nabla Q(\theta^{(t+1)}|\theta^{(t)};y) \to \nabla Q(\theta^*|\theta^*;y)$ as $t \to \infty$ and thus $\nabla Q(\theta^*|\theta^*;y) = 0$. Let R be as defined at (8), and note

$$\begin{aligned} \nabla R(\theta|\theta;y) &= \int \left[\frac{\partial}{\partial\theta}\log h(u|y;\theta)\right] h(u|y;\theta)du \\ &= \int \frac{\frac{\partial}{\partial\theta}h(u|y;\theta)}{h(u|y;\theta)} h(u|y;\theta)du \\ &= \frac{\partial}{\partial\theta}\int h(u|y;\theta)du = \frac{\partial}{\partial\theta}(1) = 0\ . \end{aligned}$$

It then follows from (8) that

$$\nabla l(\theta^*;y) = \nabla Q(\theta^*|\theta^*;y) = 0\ .$$

□

From Theorem 2 we have that *if* the EM algorithm converges, it converges to a stationary point of l; we as yet have no guarantee that EM converges. By the ascent property, the limit of an EM sequence (if it exists) cannot be a local minimum. It

can, however, be a local but not global maximum (Wu, 1983, cites several examples) or a saddlepoint (Murray, 1977, gives an example).

We will now specify conditions that do guarantee the convergence of the EM algorithm. We define a generalized EM sequence as one in which each update increases the Q-function, but does not necessarily maximize it.

Definition 1. A *generalized EM (GEM) sequence* is a sequence of parameter values $\{\theta^{(t)}\}$ satisfying $Q(\theta^{(t+1)}|\theta^{(t)}; y) \geq Q(\theta^{(t)}|\theta^{(t)}; y)$ for each t.

It is immediately clear from Theorem 1 that a GEM sequence enjoys the ascent property $l(\theta^{(t+1)}; y) \geq l(\theta^{(t)}; y)$. The conclusion of Theorem 2, that the limit of an EM sequence (if it exists) must be a stationary point of l, does not hold for GEM without additional assumptions.

Consider a sequence of parameter values $\{\theta^{(t)}\}$ satisfying $\theta^{(t+1)} \in M(\theta^{(t)})$ for some point-to-set mapping M. For example, a GEM sequence can be formulated in this manner by taking $M(\tilde{\theta}) = \left\{\theta : Q(\theta|\tilde{\theta}; y) \geq Q(\tilde{\theta}|\tilde{\theta}; y)\right\}$. We will indicate a point-to-set mapping M in Θ by the notation $M : \Theta \rightrightarrows \Theta$.

Definition 2. The point-to-set mapping $M : \Theta \rightrightarrows \Theta$ is *outer semicontinuous* if the graph of M,

$$\left\{(\theta, \tilde{\theta}) \in \Theta \times \Theta : \theta \in M(\tilde{\theta})\right\}$$

is a closed set; that is, if for any convergent sequence $\left\{(\theta^{(t)}, \tilde{\theta}^{(t)})\right\}$ satisfying $\theta^{(t)} \in M(\tilde{\theta}^{(t)})$ for each t, the limit $(\theta^*, \tilde{\theta}^*)$ satisfies $\theta^* \in M(\tilde{\theta}^*)$.

The following theorem gives a set of conditions under which every cluster point of a GEM sequence lies in a particular set $\Gamma \subset \Theta$.

Theorem 3. *Let $\Gamma \subset \Theta$ and $M : \Theta \rightrightarrows \Theta$ be such that the following conditions hold.*

1. $M(\tilde{\theta}) \subset \left\{\theta : Q(\theta|\tilde{\theta}; y) \geq Q(\tilde{\theta}|\tilde{\theta}; y)\right\}$ *when* $\tilde{\theta} \in \Gamma$.
2. $M(\tilde{\theta}) \subset \left\{\theta : Q(\theta|\tilde{\theta}; y) > Q(\tilde{\theta}|\tilde{\theta}; y)\right\}$ *when* $\tilde{\theta} \in \Theta \setminus \Gamma$.
3. *The restriction of M to $\Theta \setminus \Gamma$ is outer semicontinuous.*

Further suppose that the log-likelihood l is continuous, that the level set $\{\theta : l(\theta; y) \geq l(\theta^{(0)}; y)\}$ is compact, and let the sequence $\{\theta^{(t)} : t = 0, 1, 2, \ldots\}$ be such that $\theta^{(t+1)} \in M(\theta^{(t)})$ for each t. Then $l(\theta^{(t)}; y)$ converges to a limit, and every cluster point of $\{\theta^{(t)}\}$ is contained in Γ.

Proof. By assumption the log-likelihood is bounded above. Also, $\{\theta^{(t)}\}$ is a GEM sequence, hence $l(\theta^{(t)}; y)$ is nondecreasing, so it converges to a limit λ.

Suppose to get a contradiction there exists a subsequence $\theta^{(t_k)} \to \theta^* \notin \Gamma$. Consider the subsequence $\{\theta^{(t_k+1)}\}$. By the ascent property $l(\theta^{(t_k+1)}; y) \geq l(\theta^{(0)}; y)$ for each k, so the compactness assumption guarantees that $\{\theta^{(t_k+1)}\}$ has a convergent subsequence with limit θ^{**}. Further, $\theta^{**} \in M(\theta^*)$ by the outer semicontinuity of M, and thus $Q(\theta^{**}|\theta^*; y) > Q(\theta^*|\theta^*; y)$ and thus $l(\theta^{**}; y) > l(\theta^*; y)$ by assumption 2 and Theorem 1, respectively. But $l(\theta^{**}; y) = \lambda = l(\theta^*; y)$ by continuity of l, a contradiction.

Thus all cluster points of $\{\theta^{(t)}\}$ are in Γ. □

In the obvious application of Theorem 3 the solution set Γ is taken to be the set of stationary points of the log-likelihood. We now have a set of conditions under which the EM algorithm is guaranteed to converge to the unique MLE $\hat{\theta}$.

Corollary 1. *If the conditions of Theorem 3 hold and the set* Γ *consists of a single point* $\hat{\theta}$*, then the sequence* $\left\{\theta^{(t)}\right\}$ *converges to* $\hat{\theta}$*.*

Unfortunately, these conditions can be difficult or impossible to verify in many practical applications. Further, the rate of convergence of the EM algorithm cannot be superlinear, as we show here.

Definition 3. The sequence $\left\{\theta^{(t)}\right\}$ converging to $\hat{\theta}$ is said to converge *superlinearly* if

$$\theta^{(t+1)} - \hat{\theta} = o\left(||\theta^{(t)} - \hat{\theta}||\right)$$

as $t \to \infty$, where $||\cdot||$ denotes the standard Euclidean norm.

Lemma 2. *Suppose the log-likelihood is twice continuously differentiable with a local maximum at* $\hat{\theta}$ *and suppose that* $\nabla^2 l(\hat{\theta}; y)$ *is nonsingular and negative definite. Further suppose that* $\nabla^2 Q(\hat{\theta}|\hat{\theta}; y)$ *is nonsingular and negative definite and* $\nabla^2 Q(\hat{\theta}|\hat{\theta}; y) - \nabla^2 l(\hat{\theta}; y)$ *is nonsingular. Define the sequence* $\left\{\theta^{(t)}\right\}$ *by*

$$\theta^{(t+1)} = \theta^{(t)} - \left[\nabla^2 Q(\theta^{(t)}|\theta^{(t)}; y)\right]^{-1} \nabla Q(\theta^{(t)}|\theta^{(t)}; y) \tag{9}$$

and suppose that $\theta^{(t)} \to \hat{\theta}$*. Then the convergence is not superlinear.*

Proof. Let δ_{NR} denote the Newton-Raphson update increment for the optimization of l, that is, if $\left\{\theta'^{(t)}\right\}$ is a Newton-Raphson sequence then $\theta'^{(t+1)} = \theta'^{(t)} + \delta_{NR}(\theta'^{(t)})$ for each t, or

$$\delta_{NR}(\theta) = -\left[\nabla^2 l(\theta; y)\right]^{-1} \nabla l(\theta; y) \ .$$

Since $\nabla^2 l(\theta; y)$ is continuous and $\nabla^2 l(\hat{\theta}; y)$ is nonsingular, it must be that $\nabla^2 l(\theta; y)$ is invertible in a neighborhood of $\hat{\theta}$, and thus $\delta_{NR}(\theta^{(t)})$ is well-defined for sufficiently large t.

By convergence of $\left\{\theta^{(t)}\right\}$ and the continuity of ∇l, $\nabla l(\theta^{(t)}; y) \to \nabla l(\hat{\theta}; y) = 0$. Together with the continuity of $\nabla^2 l(\theta; y)$, this guarantees that

$$\delta_{NR}(\theta^{(t)}) = -\left[\nabla^2 l(\theta^{(t)}; y)\right]^{-1} \nabla l(\theta^{(t)}; y) \to \left[\nabla^2 l(\hat{\theta}; y)\right]^{-1} \cdot 0 = 0$$

as $t \to \infty$. Now, consider the sequence $\left\{\nabla l(\theta^{(t)}; y)/||\nabla l(\theta^{(t)}; y)||\right\}$. This sequence lives on the unit sphere, a compact set, and hence has a convergent subsequence. Let $\{t_k\}$ denote the indices of a convergent subsequence and b its limit. Then

$$\frac{\theta^{(t_k+1)} - \theta^{(t_k)}}{||\nabla l(\theta^{(t_k)}; y)||} = \frac{-\left[\nabla^2 Q(\theta^{(t_k)}|\theta^{(t_k)}; y)\right]^{-1} \nabla l(\theta^{(t_k)}; y)}{||\nabla l(\theta^{(t_k)}; y)||} \to -\left[\nabla^2 Q(\hat{\theta}|\hat{\theta}; y)\right]^{-1} b \tag{10}$$

and

$$\frac{\delta_{NR}(\theta^{(t_k)})}{||\nabla l(\theta^{(t_k)}; y)||} = \frac{-\left[\nabla^2 l(\theta^{(t_k)}; y)\right]^{-1} \nabla l(\theta^{(t_k)}; y)}{||\nabla l(\theta^{(t_k)}; y)||} \to -\left[\nabla^2 l(\hat{\theta} y)\right]^{-1} b \tag{11}$$

as $k \to \infty$. The equality in (10) follows from the fact that $\nabla Q(\theta|\theta; y) = \nabla l(\theta; y)$ for any θ.

Suppose the sequence $\left\{\theta^{(t)}\right\}$ does converge superlinearly. Then it is asymptotically equivalent to Newton-Raphson by the Dennis-Moré characterization theorem

(see, for example, Fletcher, 1987), and thus the (sub)sequences defined in (10) and (11) must have the same limit. Then $\left[\nabla^2 Q(\hat{\theta}|\hat{\theta};y)\right]^{-1} b = \left[\nabla^2 l(\hat{\theta};y)\right]^{-1} b = c$. So

$$\left[\nabla^2 Q(\hat{\theta}|\hat{\theta};y) - \nabla^2 l(\hat{\theta};y)\right] c = 0$$

and thus $c = 0$ since $\nabla^2 Q(\hat{\theta}|\hat{\theta};y) - \nabla^2 l(\hat{\theta};y)$ is full rank. But b must be on the unit sphere, a contradiction.

Thus the convergence of $\left\{\theta^{(t)}\right\}$ to $\hat{\theta}$ is not superlinear. □

The algorithm defined at (9), with update rule given by a single Newton-Raphson iteration toward the maximum of the Q-function, was first introduced by Lange (1995) and is known as the *EM gradient algorithm*. Details are beyond the scope of this report, but roughly speaking, the convergence properties of the EM algorithm are equally enjoyed by Lange's (1995) EM gradient algorithm. Thus while Lemma 2 takes the convergence of the EM gradient sequence as a given, there is no sacrifice in the applicability of the result, as the EM gradient converges to a local maximum under essentially the same conditions as does the EM algorithm.

Theorem 4. *Suppose the EM sequence* $\left\{\theta^{(t)}\right\}$ *converges to a point* $\theta^* \in \Theta$, *a stationary point of the log-likelihood. Further suppose that* $l(\theta;y)$, $Q(\theta|\tilde{\theta};y)$, *and* $R(\theta|\tilde{\theta};y)$ *are twice continuously differentiable in* θ *and that* $\nabla^2 l(\theta^*;y)$, $\nabla^2 Q(\theta^*|\theta^*;y)$, *and* $\nabla^2 R(\theta^*|\theta^*;y)$ *have full rank. Then the convergence cannot be superlinear.*

Proof. Let δ_{EG} denote the EM gradient update increment, that is, if $\left\{\theta'^{(t)}\right\}$ is an EM gradient sequence then $\theta'^{(t+1)} = \theta'^{(t)} + \delta_{EG}(\theta'^{(t)})$ for each t:

$$\delta_{EG}(\theta) = -\left[\nabla^2 Q(\theta|\theta;y)\right]^{-1} \nabla Q(\theta|\theta;y) \; .$$

Define δ_{EM} analogously, so $\theta + \delta_{EG}(\theta)$ represents the first iteration in a Newton-Raphson routine starting at θ and converging to $\theta + \delta_{EM}(\theta)$. Since Newton-Raphson converges superlinearly in this subproblem (see, for example, Fletcher, 1987, Theorem 3.1.1), we have

$$\theta + \delta_{EG}(\theta) - [\theta + \delta_{EM}(\theta)] = o\left(||\delta_{EM}(\theta)||\right)$$

or

$$\delta_{EG}(\theta) = \delta_{EM}(\theta) + o\left(||\delta_{EM}(\theta)||\right) \; ,$$

and thus the EM gradient algorithm (9) is asymptotically equivalent to the EM algorithm. But EM gradient is not superlinearly convergent by Lemma 2, and thus neither is the EM algorithm. □

4. Some convergence results for Monte Carlo EM

It seems a statement of the obvious (and an understatement at that) to point out that the study of convergence properties of Monte Carlo EM is more complicated than that of ordinary EM. Even before coming to face the complexity of the mathematical arguments, one must determine which notion of "convergence" one wishes to consider – what exactly is going to infinity? We mention here three distinct approaches to the problem.

The first serious effort in establishing convergence properties of MCEM is that of Chan and Ledolter (1995), who treat the data as fixed, and hold the Monte Carlo

sample size m constant across MCEM iterations. They then let m go to infinity, and study the asymptotic properties of the MCEM *sequence* as a Monte Carlo approximation to the ordinary EM sequence with the same starting value (whose convergence properties are well understood). We will discuss Chan and Ledolter's (1995) results in considerable detail in subsection 4.1. On the other hand, unless the Monte Carlo sample size is allowed to increase with the iteration count, there is no chance for convergence in the usual sense (convergence to the MLE) because of persistent Monte Carlo error.

In the version of MCEM considered by Sherman, Ho and Dalal (1997), the Monte Carlo E-step is carried out by running multiple (independent) Markov chains generated by a Gibbs sampler. Their theoretical results are built on allowing the number of chains, the length of each chain, and the number of EM iterations T to all tend to infinity, as does the data sample size N. They then prove $\sqrt{N}$-consistency and asymptotic normality of the estimator $\theta^{(T)}$. In other words, Sherman, Ho and Dalal (1997) found conditions under which the MCEM approximation to the MLE enjoys the same asymptotic properties as the MLE itself. This represents yet another possible notion of "convergence" of MCEM, though not one that we will pursue any further in the present paper.

Fort and Moulines (2003) treat the data as fixed, the Monte Carlo sample size as increasing (deterministically) across MCEM iterations, and establish a.s. convergence of the sequence as the iteration count goes to infinity. We consider this the strongest known result on the asymptotic properties of MCEM, as this notion of convergence seems the most consistent with that of ordinary (deterministic) EM. We summarize Fort and Moulines (2003) main conclusions in subsection 4.2.

4.1. A result of Chan and Ledolter (1995)

Chan and Ledolter (1995) showed that, given a suitable starting value, a sequence of parameter values generated by the Monte Carlo EM algorithm will get arbitrarily close to a maximizer of the observed likelihood with high probability. Their main result is given as Theorem 5 below. We first establish one more convergence property of deterministic EM, also attributable to Chan and Ledolter (1995).

Let $M_{EM} : \Theta \to \Theta$ denote the mapping given by the deterministic EM update rule, that is, $M_{EM}(\tilde{\theta}) = \arg\max Q(\theta|\tilde{\theta}; y)$.

Lemma 3. *(Lemma 1 of Chan and Ledolter, 1995) Suppose θ^* is a local maximizer of the log-likelihood $l(\theta; y)$, a continuous function of θ, and that there exists a neighborhood in which θ^* is the only stationary point. Then for any neighborhood $\mathcal{N}$ of θ^*, there exists a neighborhood $\mathcal{N}^*$ such that an EM sequence $\{\theta^{(t)} : t = 0, 1, 2, \ldots\}$ started at any $\theta^{(0)} \in \mathcal{N}^*$, satisfies (i) $\theta^{(t)} \in \mathcal{N}$ for all $t = 1, 2, \ldots$; and (ii) $\theta^{(t)} \to \theta^*$ as $t \to \infty$.*

Proof. Let $\mathcal{N}$ be a neighborhood of θ^*. There exists a compact, connected subneighborhood $\mathcal{N}^* \subset \mathcal{N}$ such that (i) $l(\theta; y)$ attains its maximum over $\mathcal{N}^*$ at θ^*, (ii) $\mathcal{N}^*$ contains no other stationary points of l, and (ii) there exists $\varepsilon > 0$ such that $l(\theta; y) \geq l(\theta^*; y) - \varepsilon$ for all $\theta \in \mathcal{N}^*$. It follows from these conditions that $M_{EM}(\theta) \in \mathcal{N}^*$ for any $\theta \in \mathcal{N}^*$; thus an EM sequence $\{\theta^{(t)}\}$ with $\theta^{(0)} \in \mathcal{N}^*$ satisfies $\theta^{(t)} \in \mathcal{N}^*$, and thus $\theta^{(t)} \in \mathcal{N}$, for all $t = 1, 2, \ldots$.

Continue to assume that $\theta^{(0)} \in \mathcal{N}^*$ and consider the EM sequence $\{\theta^{(t)}\}$. Now the sequence $\{l(\theta^{(t)}; y)\}$ is nondecreasing and bounded above by $l(\theta^*; y)$, and thus

converges to a finite limit, call it λ. The sequence $\{\theta^{(t)}\}$ lives in $\mathcal{N}^*$, a compact set; let $\{\theta^{(t_k)}\}$ be a convergent subsequence and denote its limit by $\theta^{**} \in \mathcal{N}^*$.

Suppose $\theta^{**} \neq \theta^*$. Then $l(\theta^{(t_k+1)}; y) \to l(M_{EM}(\theta^{**}); y) > l(\theta^{**}; y) = \lambda$. That is, the subsequence $\{l(\theta^{(t_k+1)}; y)\}$ converges to a limit greater than λ, a contradiction.

Thus any convergent subsequence of $\{\theta^{(t)}\}$ must converge to θ^*; thus $\{\theta^{(t)}\}$ converges to θ^*. □

In the terminology of the stability theory of dynamical systems (see, for example, Arrowsmith and Place, 1992, section 3.5), the lemma asserts that an isolated local maximizer θ^* of $l(\theta; y)$ is an *asymptotically stable fixed point* for the EM algorithm. Practically, Lemma 3 tells us that an EM sequence with a sufficiently close starting value will remain arbitrarily close to θ^* (by stability) as well as converge to θ^*.

Theorem 5. *(Theorem 1 of Chan and Ledolter, 1995). Let $\{\theta^{(t)}\}$ denote a Monte Carlo EM sequence based on Monte Carlo sample sizes $m_t \equiv m$, and suppose that the MCEM update $\mathcal{M}_m(\tilde{\theta}) := \arg\max Q_m(\theta|\tilde{\theta}; y)$ converges in probability to $M_{EM}(\tilde{\theta})$ as $m \to \infty$. Further suppose that this convergence is uniform on compact subsets of Θ. Let θ^* be an isolated local maximizer of $l(\theta; y)$, a continuous function of θ. Then there exists a neighborhood of θ^* such that for any starting value $\theta^{(0)}$ in that neighborhood and for any $\varepsilon > 0$, there exists T_0 such that*

$$\Pr\left\{||\theta^{(t)} - \theta^*|| < \varepsilon \text{ for some } t \leq T_0\right\} \to 1 \tag{12}$$

as the Monte Carlo sample size $m \to \infty$.

Proof. Let $\mathcal{N}$ be the set defined as $\mathcal{N}^*$ in the proof of Lemma 3, so that $\mathcal{N}$ is compact and connected, contains θ^*, and $M_{EM}(\theta) \in \mathcal{N}$ for any $\theta \in \mathcal{N}$. For any $\varepsilon > 0$, we will find T_0 such that (12) holds for any $\theta^{(0)} \in \mathcal{N}$.

Let $\varepsilon > 0$ be given. First, there exists a positive number $\varepsilon_1 \leq \varepsilon$ such that $\mathcal{N}_1 := \{\theta \in \mathcal{N} : ||\theta - \theta^*|| \geq \varepsilon_1\}$ is nonempty; if $\theta \in \mathcal{N}_1$, then $M_{EM}(\theta) \neq \theta$. By the ascent property and by continuity of l in θ there exist $\delta,\ \delta_1 > 0$ such that for any $\theta \in \mathcal{N}_1$, if $||\theta' - M_{EM}(\theta)|| < \delta$, then $l(\theta'; y) - l(\theta; y) > \delta_1$.

By construction of $\mathcal{N}$, there exists $\delta_2 > 0$ such that for any $\theta \in \mathcal{N}$, any θ' with $||\theta' - M_{EM}(\theta)|| < \delta_2$ is also in $\mathcal{N}$. Without loss of generality we can take $\delta_2 < \delta$. Thus we have that for any $\theta \in \mathcal{N}_1$, any θ' with $||\theta' - M_{EM}(\theta)|| < \delta_2$ is also in $\mathcal{N}$ (though not necessarily in $\mathcal{N}_1$) and $l(\theta'; y) - l(\theta; y) > \delta_1$. Let

$$R = \sup_{\theta, \theta' \in \mathcal{N}} \{l(\theta; y) - l(\theta'; y)\} < \infty \tag{13}$$

and let $T_0 = \lfloor R/\delta_1 \rfloor + 1$, where $\lfloor \cdot \rfloor$ denotes the greatest integer function.

Now, suppose an element of the MCEM sequence $\theta^{(t)} = \tilde{\theta} \in \mathcal{N}$. Then the probability that its MCEM update $\theta^{(t+1)} = \mathcal{M}_m(\theta^{(t)})$ is also in $\mathcal{N}$ is

$$\Pr\left\{\theta^{(t+1)} \in \mathcal{N} \mid \theta^{(t)} = \tilde{\theta}\right\} \geq \Pr\left\{||\theta^{(t+1)} - M_{EM}(\theta^{(t)})|| < \delta_2 \mid \theta^{(t)} = \tilde{\theta}\right\} \tag{14}$$

by the definition of δ_2. Denote a lower bound on the right hand side of (14) by $p = p(\delta_2, m) > 0$ and note that (i) p can be chosen not to depend on the value of $\tilde{\theta} \in \mathcal{N}$ by the compactness of $\mathcal{N}$ and the uniformity of convergence $\mathcal{M}_m(\theta) \to M_{EM}(\theta)$ over compact subsets of Θ; and (ii) for fixed δ_2, $p(\delta_2, m) \to 1$ as $m \to \infty$.

Consider running a Monte Carlo EM algorithm for T_0 updates. For any starting value $\theta^{(0)} \in \mathcal{N}$,

(15)
$$\Pr\left\{\theta^{(t)} \in \mathcal{N} \text{ for } t = 0, 1, \ldots, T_0\right\} \geq \Pr\left\{||\theta^{(t+1)} - M_{EM}(\theta^{(t)})|| < \delta_2 \text{ for } t = 0, 1, \ldots, T_0 - 1\right\},$$

and since each Monte Carlo EM update is calculated independently, the right hand side of (15) is bounded below by $p(\delta_2, m)^{T_0}$.

Now, suppose that $\theta^{(0)} \in \mathcal{N}$, and that $||\theta^{(t+1)} - M_{EM}(\theta^{(t)})|| < \delta_2$ for each t, and thus $\theta^{(t)} \in \mathcal{N}$ for each t. Suppose to get a contradiction that $||\theta^{(t)} - \theta^*|| \geq \varepsilon_1$, that is, that $\theta^{(t)} \in \mathcal{N}_1$ for each $t = 0, 1, \ldots, T_0$. Then $l(\theta^{(t+1)}; y) - l(\theta^{(t)}; y) > \delta_1$ for each $t = 0, 1, \ldots, T_0 - 1$, and thus $l(\theta^{(T_0)}; y) - l(\theta^{(0)}; y) > \delta_1 T_0 > R$. But that contradicts (13), the definition of R, since $\theta^{(0)}$ and $\theta^{(T_0)}$ are both in $\mathcal{N}$.

Thus it must be that if $\theta^{(0)} \in \mathcal{N}$ and $||\theta^{(t+1)} - M_{EM}(\theta^{(t)})|| < \delta_2$ for each t, then $||\theta^{(t)} - \theta^*|| < \varepsilon_1 \leq \varepsilon$ for some t, which occurs with probability not less than $p(\delta_2, m)^{T_0}$, which converges to 1 as $m \to \infty$. □

A couple of remarks are in order. First, we note that the assumptions of Theorem 5 are slightly different than those made by Chan and Ledolter (1995) in that where we assumed uniform convergence of the Monte Carlo EM update, Chan and Ledolter (1995) assumed conditions on the form of the log-likelihood sufficient to guarantee it. Secondly, the conclusion of Theorem 5, while interesting, is unsatisfying in at least one respect: It does not guarantee the convergence of an MCEM sequence in any meaningful sense. Practically, what this theorem tells us is that if you run the algorithm long enough (at least T_0 iterations), the resulting sequence will, with high probability, *at some point* get arbitrarily close to the MLE. But to an analyst examining the output of an MCEM run, even a very long one, there is no way to know when that has happened, if at all. A more powerful result would be one that specifies conditions under which the algorithm gets close to the MLE and stays there.

4.2. A result of Fort and Moulines (2003)

Fort and Moulines (2003) used the ergodic theory of Markov chains to prove the almost sure (a.s.) convergence of a variation of the Monte Carlo EM algorithm. We will state their assumptions and main conclusion; the proof is highly technical and beyond the scope of this report.

We will state Fort and Moulines (2003) convergence result assuming that the Monte Carlo E-step is accomplished by i.i.d. sampling. In fact the result holds more generally under Markov chain Monte Carlo methods, assuming the underlying Markov transition kernel is *uniformly ergodic* (see, for example, Jones and Hobert, 2001).

Fort and Moulines (2003) consider a variation of Monte Carlo EM they call *stable MCEM*, which we define here. Let $\{\mathcal{K}_t : t = 0, 1, 2, \ldots\}$ be a sequence of compact subsets of Θ satisfying

(16)
$$\mathcal{K}_t \subset \mathcal{K}_{t+1} \text{ for each } t, \quad \text{and} \quad \bigcup_{t=0}^{\infty} \mathcal{K}_t = \Theta\,.$$

Set $p_0 = 0$ and choose $\theta^{(0)} \in \mathcal{K}_0$. Given $\theta^{(t)}$ and p_t, the stable MCEM update rule for $\theta^{(t+1)}$ and p_{t+1} is given by

1. Let θ' be the ordinary MCEM update as defined in Section 1.
2. If $\theta' \in \mathcal{K}_{p_t}$, then $\theta^{(t+1)} = \theta'$ and $p_{t+1} = p_t$.
   If $\theta' \notin \mathcal{K}_{p_t}$, then $\theta^{(t+1)} = \theta^{(0)}$ and $p_{t+1} = p_t + 1$.

Thus in stable MCEM, any time the ordinary MCEM update falls outside a specific set, the algorithm is reinitialized at the point $\theta^{(0)}$; p_t counts the cumulative number of reinitializations as of update t. Fort and Moulines (2003) showed that under appropriate assumptions (see Theorem 6 below), $\{p_t\}$ is a.s. finite.

We will assume that the complete data model $f(y, u; \theta)$ is from the class of *curved exponential families*: Let $\mathcal{Y} \subset \mathbb{R}^N$ denote the range of Y and $\mathcal{U} \subset \mathbb{R}^q$ the range of U. We assume that for some integer k there exist functions $\phi : \Theta \to \mathbb{R}^1$, $\psi : \Theta \to \mathbb{R}^k$, and $S : \mathcal{Y} \times \mathcal{U} \to \mathcal{S} \subset \mathbb{R}^k$ such that

$$l_c(\theta; y, u) = \log f(y, u; \theta) = \psi(\theta)^T S(y, u) + \phi(\theta) \; .$$

Since l_c depends on (y, u) only through $s = S(y, u)$ we can write $l_c(\theta; s) = \psi(\theta)^T s + \phi(\theta)$. Note that the curved exponential families include the linear mixed model of Example 1 in Section 2, but not the logit-normal GLMM of Example 2.

We will further assume that

1. ϕ and ψ are continuous on Θ, S is continuous on $\mathcal{Y} \times \mathcal{U}$;
2. for all $\theta \in \Theta$, $\bar{S}(\theta; y) := \mathrm{E}\{S(y, U) \mid y;\ \theta\}$ is finite and continuous on Θ;
3. there exists a continuous function $\hat{\theta} : \mathcal{S} \to \Theta$ such that for all $s \in \mathcal{S}$, $l_c(\hat{\theta}(s); s) = \sup_{\theta \in \Theta} l_c(\theta; s)$;
4. the observed data log-likelihood $l(\theta; y)$ is continuous on Θ, and for any λ, the level set $\{\theta \in \Theta : l(\theta; y) \geq \lambda\}$ is compact;
5. the set of fixed points of the EM algorithm is compact.

Let Γ denote the set of fixed points of the EM algorithm; in a curved exponential family, and using the notation introduced above, $\Gamma = \left\{\theta \in \Theta : \hat{\theta}(\bar{S}(\theta; y)) = \theta\right\}$. As shown by Wu (1983, Theorem 2), under the above assumptions, if Θ is open and ϕ and ψ are differentiable on Θ, then $l(\theta; y)$ is differentiable on Θ and $\Gamma = \{\theta \in \Theta : \nabla l(\theta; y) = 0\}$. In other words, the set of fixed points of the EM algorithm coincides with the set of stationary points of the log-likelihood $l(\theta; y)$; see also our Theorem 2.

Finally, note that assumptions 4. and 5. guarantee that the set $\{l(\theta; y) : \theta \in \Gamma\}$ is compact as well. We can now state Fort and Moulines's (2003) main result. We will denote the *closure* of a sequence by $\mathrm{Cl}(\cdot)$, so that $\mathrm{Cl}(\{\theta^{(t)}\})$ represents the union of the sequence $\{\theta^{(t)}\}$ itself with its limit points.

Theorem 6. *(Theorem 3 of Fort and Moulines, 2003) Assume the complete data model is from the class of curved exponential families, and the model satisfies assumptions* 1. *through* 6. *above. Consider an implementation of the stable MCEM algorithm using a sequence of sets $\{\mathcal{K}_t\}$ satisfying (16). Let $\theta^{(0)} \in \mathcal{K}_0$ and suppose the Monte Carlo sample sizes $\{m_t\}$ satisfy $\sum_{t=0}^{\infty} m_t^{-1} < \infty$. Then*

1. *(a) $\lim_{t\to\infty} p_t < \infty$ with probability 1 (w.p. 1) and $\limsup_{t\to\infty} ||\theta^{(t)}|| < \infty$ w.p. 1;*
   (b) $\{l(\theta^{(t)}; y)\}$ converges w.p. 1 to a connected component of $l(\Gamma; y)$ where Γ denotes the set of stationary points of $l(\theta; y)$ (and fixed points of the EM algorithm).

2. *If* $\{l(\theta;y):\theta\in\Gamma\cap \mathrm{Cl}(\{\theta^{(t)}\})\}$ *has an empty interior, then* $\{l(\theta^{(t)};y)\}$ *converges w.p. 1 to a point* λ^* *and* $\{\theta^{(t)}\}$ *converges to the set* $\{\theta : l(\theta;y)=\lambda^*\}$.

It is often the case that the set Γ is made up of isolated points; the above theorem then guarantees pointwise convergence of $\{l(\theta^{(t)};y)\}$ to a stationary point of $l(\theta;y)$. If Γ consists of a single point $\hat{\theta}$, the theorem guarantees that $l(\theta^{(t)};y)\to l(\hat{\theta};y)$ w.p. 1 and $\theta^{(t)}\to\hat{\theta}$ w.p. 1, analogous to Corollary 1.

Finally, we note that the assumption that $\sum m_t^{-1}<\infty$ can be weakened in many instances, but is necessarily of the form $\sum m_t^{-p}<\infty$ for some $p\geq 1$.

5. Remarks: Lessons for the (MC)EM practitioner

We conclude with a brief discussion of the practical implications of the convergence results of Sections 3 and 4. First, as we noted in our discussion following Theorem 2, even when EM converges, there is no guarantee in general that it has converged to a global maximum. In more complex settings such as mixture models, or model-based clustering, the likelihood function may have multiple optima, most of which will be local optima. While the EM algorithm may converge, its limit point is sub-optimal. Solutions to overcome local optima can include merging the ideas of the EM algorithm with those of global optimization. One example is described in the paper by Heath, Fu and Jank (2009) who combine EM with the cross-entropy method and model reference adaptive search, two global optimization heuristics. Another example can be found in Tu, Ball and Jank (2008), who combine the EM algorithm with the genetic algorithm to model flight delay distributions.

With respect to Monte Carlo EM, as we have previously noted, the Monte Carlo sample size must be increased with the iteration count; otherwise there is no chance for convergence in the usual sense, due to the persistence of Monte Carlo error. The convergence results of section 4.2 (Fort and Moulines, 2003) require $\sum m_t^{-1}<\infty$. Intuitively it makes sense to start the algorithm with modest simulation sizes: when the parameter value is relatively far from the MLE, the (deterministic) EM update makes a substantial jump, and less precision is required for the Monte Carlo approximation to that jump. When the parameter value is close to the MLE, as will be the case after a number of iterations, the EM update is a small step, and greater precision is required for the Monte Carlo approximation.

Thus it is clear that m_t must be an increasing function of t, though it is not at all clear what might be an appropriate form. In fact there exists a literature, beginning with Booth and Hobert (1999), on *automated* Monte Carlo EM algorithms, in which the simulation size for each Monte Carlo E-step is determined internally to the algorithm, based on some rule for assessing the level of precision required for the Monte Carlo approximation at hand. Other authors who have contributed to this literature include Levine and Casella (2001) and Caffo, Jank and Jones (2005).

One can view the Monte Carlo EM update to the parameter value $\theta^{(t)}$ as an estimate of the deterministic EM update $M_{EM}(\theta^{(t)})$. In Booth and Hobert's (1999) algorithm, each MCEM update requires the computation of an asymptotic confidence region for $M_{EM}(\theta^{(t)})$ in addition to the point estimate $\mathcal{M}_{m_t}(\theta^{(t)})$. If $\theta^{(t)}$ falls within this confidence region, we must accept that the current parameter value $\theta^{(t)}$ is statistically indistinguishable from its EM update $M_{EM}(\theta^{(t)})$. This suggests that the MCEM update was "swamped by Monte Carlo error," and thus the simulation size must be increased at the next iteration. The reader is referred to Booth and Hobert (1999) for details and examples. Levine and Casella (2001)

use a regeneration-based approach to Monte Carlo standard errors in computing their confidence region.

The Ascent-based Monte Carlo EM algorithm of Caffo, Jank and Jones (2005) seeks to prevent the MCEM update from being swamped by Monte Carlo error by successively appending the Monte Carlo sample until one has a pre-specified level of confidence that the proposed update increases the log-likelihood over the current parameter value, that is, until we are confident that indeed $l(\theta^{(t+1)}; y) \geq l(\theta^{(t)}; y)$. Recall that this ascent property is guaranteed for ordinary EM (Theorem 1). Since the MCEM update maximizes an estimate of the Q-function rather than the Q-function itself, there is no ascent property for MCEM in general. But a parameter update computed according to the Ascent-based MCEM rule will increase the log-likelihood with high probability. Again the reader is referred to the source (Caffo, Jank and Jones, 2005) for details. Empirical comparisons between Ascent-based MCEM and Booth and Hobert's (1999) algorithm can be found in Caffo, Jank and Jones (2005) and Neath (2006).

A second practical implication of the convergence properties of Monte Carlo EM relates to convergence criteria, or *stopping rules* for the algorithm. At what point should the MCEM iterations be terminated and the current parameter value accepted as the MLE? The usual stopping rules employed in a deterministic iterative algorithm like ordinary EM terminate when it is apparent that further iterations (i) will not substantively change the approximation to the MLE, or (ii) will not substantively change the value of the objective (likelihood) function. For example, one might terminate at the first iteration t to satisfy

$$\max_i \left\{ \frac{|\theta_i^{(t)} - \theta_i^{(t-1)}|}{|\theta_i^{(t)}| + \delta} \right\} < \epsilon \tag{17}$$

for user-specified δ and ϵ, where the maximum is taken over components of the parameter vector. In Monte Carlo EM, such criteria run the risk of terminating too early, as (17) may be attained only because of Monte Carlo error in the update. An obvious but inelegant solution is to terminate only after (17) is met for, say, three consecutive iterations. This is the stopping rule recommended by Booth and Hobert (1999). Other MCEM stopping rules considered in the literature include Chan and Ledolter's (1995) suggestion to terminate at the first iteration where $l(\theta^{(t)}; y) - l(\theta^{(t-1)}; y)$ is stochastically small; in a similar vein Caffo, Jank and Jones (2005) terminate when an asymptotic upper bound on $Q(\theta^{(t)}|\theta^{(t-1)}; y) - Q(\theta^{(t-1)}|\theta^{(t-1)}; y)$ falls below a pre-specified tolerance.

Finally, we note that while our focus throughout has been on finding a good approximation to the MLE, meaningful statistical inference requires at minimum a reliable estimate of the standard error as well. A formula in Louis (1982) expresses the observed Fisher Information as an expectation taken with respect to the conditional distribution of the unobserved data given the observed data. Thus a Monte Carlo approximation to the inverse covariance matrix of the MLE is readily available from the simulation already conducted to compute the final MCEM update.

References

Arrowsmith, D. K. and Place, C. M. (1992). *Dynamical Systems: Differential Equations, Maps and Chaotic Behavior*. Chapman & Hall, London.

Billingsley, P. (1995). *Probability and Measure*, third ed. Wiley, New York.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society,* Series B **61** 265–285.

Booth, J. G., Hobert, J. P. and Jank, W. (2001). A survey of Monte Carlo algorithms for maximizing the likelihood of a two-stage hierarchical model. *Statistical Modelling: An International Journal* **1** 333–349.

Boyles, R. A. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society,* Series B **45** 47–50.

Caffo, B. S., Jank, W. and Jones, G. L. (2005). Ascent-based Monte Carlo EM. *Journal of the Royal Statistical Society,* Series B **67** 235–251.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series involving counts. *Journal of the American Statistical Association* **90** 242–252.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* Series B **39** 1–22.

Fletcher, R. (1987). *Practical Methods of Optimization*, Second ed. Wiley, New York.

Fort, G. and Moulines, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics* **31** 1220–1259.

Geyer, C. J. (1998). Course notes: Inequality-constrained statistical inference, School of Statistics, University of Minnesota.

Heath, J. W., Fu, M. C. and Jank, W. (2009). New global optimization algorithms for model-based clustering. *Computational Statistics & Data Analysis* **53** 3999–4017.

Jank, W. (2004). Quasi-Monte Carlo sampling to improve the efficiency of Monte Carlo EM. *Computational Statistics & Data Analysis* **48** 685–701.

Johnson, A. A., Jones, G. L. and Neath, R. C. (2011). Component-wise Markov chain Monte Carlo. *ArXiv e-prints.*

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science* **16** 312–334.

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

L'Ecuyer, P. and Lemieux, C. (2002). Recent advances in randomized quasi-Monte Carlo methods. In *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications* (M. Dror, P. L'Ecuyer and F. Szidarovski, eds.) 419–474. Kluwer Academic Publishers, Norwell, Massachusetts.

Levine, R. A. and Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics* **10** 422–439.

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.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92** 162–170.

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

Murray, G. D. (1977). Discussion of the paper by Professor Dempster et al. *Journal of the Royal Statistical Society,* Series B **39** 27–28.

Neath, R. C. (2006). Monte Carlo methods for likelihood-based inference in hierarchical models PhD thesis, University of Minnesota, School of Statistics.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, second ed. Springer-Verlag, New York.

Sherman, R. P., Ho, Y.-Y. K. and Dalal, S. R. (1997). Conditions for convergence of Monte Carlo EM sequences with an application to product diffusion modeling. *Econometrics Journal* **2** 248–267.

Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods*, eighth ed. Iowa State University Press, Ames.

Tu, Y., Ball, M. and Jank, W. (2008). Estimating flight departure delay distributions–a statistical approach with long-term trend and short-term pattern. *Journal of the American Statistical Association* **103** 112–125.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85** 699–704.

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