A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms

GREG C. G. WEI and MARTIN A. TANNER*

The first part of this article presents the Monte Carlo implementation of the E step of the EM algorithm. Given the current guess to the maximizer of the posterior distribution, latent data patterns are generated from the conditional predictive distribution. The expected value of the augmented log-posterior is then updated as a mixture of augmented log-posteriors, mixed over the generated latent data patterns (multiple imputations). In the M step of the algorithm, this mixture is maximized to obtain the update to the maximizer of the observed posterior. The gradient and Hessian of the observed log posterior are also expressed as mixtures, mixed over the multiple imputations. The relation between the Monte Carlo EM (MCEM) algorithm and the data augmentation algorithm is noted. Two modifications to the MCEM algorithm (the poor man's data augmentation algorithms), which allow for the calculation of the *entire* posterior, are then presented. These approximations serve as diagnostics for the validity of the normal approximation to the posterior, as well as starting points for the full data augmentation analysis. The methodology is illustrated with two examples.

KEY WORDS: Bayesian inference; Multiple imputation; Simulation

1. INTRODUCTION

The EM algorithm (Dempster, Laird, and Rubin 1977) is a powerful computational technique for locating a maximizer of a posterior distribution. Rather than attempt to maximize a complicated posterior, the EM algorithm requires a series of maximizations of functions. The EM algorithm is best suited for situations where the construction of each function and each maximization is straightforward. In Section 3 of this article, we present the Monte Carlo implementation of the E step of the EM algorithm, thereby expanding the scope of application of the algorithm. To compute the expectation of the log-posterior, latent data patterns are generated from the conditional predictive distribution, given the current guess to the maximizer of the posterior. The expected value of the log-posterior is then updated as a mixture of augmented log-posteriors, mixed over the generated latent data patterns (multiple imputations). The gradient and Hessian of the observed log posterior are similarly expressed as mixtures, mixed over the generated latent data patterns. The relationship between the Monte Carlo EM (MCEM) algorithm and the data augmentation algorithm (Tanner and Wong 1987) is then noted. The multiple imputations in the data augmentation algorithm are drawn from the current approximation to the predictive distribution, and the corresponding values in the MCEM algorithm are drawn from the conditional predictive distribution, conditional on the current approximation to the maximizer of the observed posterior. This observation suggests two modifications of the MCEM algorithm, the poor man's data augmentation algorithms, which allow for the estimation of the *entire* posterior. In Section 4, the methodology is illustrated in the context of regression analysis with censored data.

2. THE EM ALGORITHM

The EM algorithm (Dempster et al. 1977) is an iterative method for the computation of the maximizer of the posterior density. Before proceeding to describe the algorithm, we review basic terminology related to the algorithm. The basic idea behind the EM algorithm is to augment the *observed data* y by a quantity z, which will be referred to as *latent data*. It is assumed that, given both y and z, it is straightforward to calculate and maximize the expectation of the augmented log-posterior $\log(p(\theta \mid y, z))$. To obtain the maximizer of the observed posterior $p(\theta \mid y)$, one first computes the expectation of $\log(p(\theta \mid y, z))$ with respect to the conditional predictive distribution $p(z \mid y, \theta^{(i)})$, where $\theta^{(i)}$ is the current approximation to the mode of the observed posterior. This is known as the E step. In the M step, one obtains the maximizer of this conditional expectation. The conditional predictive distribution is then updated using the new maximizer and the algorithm is iterated. [Regarding issues of convergence, see Dempster et al. (1977), Wu (1983), and Boyles (1983).]

More formally, define the Q function as

$$Q(\theta, \theta_o) = \int_Z \log(p(\theta \mid z, y)) p(z \mid \theta_o, y)\, dz, \quad (2.1)$$

where Z denotes the sample space for the latent data z. By Jensen's inequality, it follows that if θ is chosen such that $Q(\theta, \theta_o) \geq Q(\theta_o, \theta_o)$, then $\log(p(\theta \mid y))$ will be greater

* Greg C. G. Wei is Senior Biostatistician, Department of Statistics and Data Analysis, Marion Merrell Dow Inc., Kansas City, MO 64137. Martin A. Tanner is Professor, Departments of Biostatistics and Statistics, University of Rochester, Rochester, NY 14642. Tanner was supported by National Institutes of Health Grant R01-CA35464. The authors wish to thank R. E. Kass for calling the first- and second-order approximations discussed in Tierney, Kass, and Kadane (1986) to their attention. The authors also thank Cliff Clogg, Tom Louis, Don Rubin, the associate editor, and the referees for their comments and suggestions. A portion of this work was completed while both authors were at the University of Wisconsin–Madison.

Journal of the American Statistical Association
September 1990, Vol. 85, No. 411, Theory and Methods

than or equal to $\log(p(\theta_o \mid y))$. In this way, given the current approximation to the maximizer of the observed posterior ($\theta^{(i)}$), the E step of the EM algorithm is defined by computing $Q(\theta, \theta^{(i)}) = \int_Z \log(p(\theta \mid z, y))p(z \mid \theta^{(i)}, y)\, dz$. The M step then consists of maximizing the Q function with respect to θ to obtain the update $\theta^{(i+1)}$.

3. THE MONTE CARLO IMPLEMENTATION AND THE RELATIONSHIP TO DATA AUGMENTATION

3.1 The MCEM Algorithm

To perform the integration in (2.1), we propose to use the method of Monte Carlo to obtain the MCEM algorithm. In particular, Equation (2.1) motivates the following scheme: Given the current approximation to the maximizer $\theta^{(i)}$, (a) generate a sample $z^{(1)}, \ldots, z^{(m)}$ from the current approximation to the conditional predictive distribution $p(z \mid \theta^{(i)}, y)$ and (b) update the current approximation to $Q_{i+1}(\theta, \theta^{(i)})$ to be the mixture of augmented log-posteriors of θ, mixed over the latent data patterns from (a).

$$Q_{i+1}(\theta, \theta^{(i)}) = \frac{1}{m}\sum_{j=1}^{m} \log(p(\theta \mid z^{(j)}, y)). \qquad (3.1)$$

The M step then consists of maximizing the right side of (3.1). The conditional predictive distribution is then updated using the new maximizer, and the algorithm is iterated.

Remark 1. The often-referred-to "EM-type" algorithm is obtained when m is equal to 1 and $z^{(1)} = \tilde{z}$ is some "good" summary of $p(z \mid \theta^{(i)}, y)$, such as a mode or expected value. The iterative EM-type algorithm consists of forming $\log(p(\theta \mid \tilde{z}, y))$, maximizing this function over θ, and then computing the updated $\tilde{z}$ using $p(z \mid \theta^{(i+1)}, y)$. When $\log(p(\theta \mid z, y))$ is linear in z, the EM and EM-type algorithms [with $\mathrm{E}(p(z \mid \theta^{(i)}, y)) = \tilde{z}$] both yield the maximizer of the observed posterior.

Remark 2. Rubin (1987) referred to the quantities $z^{(1)}, \ldots, z^{(m)}$ as *multiple imputations*.

Remark 3. A referee has pointed out that Monte Carlo may also be used to locate the maximizer of the augmented posterior (see Diggle and Gratton 1984). Optimization of the augmented posterior in high-dimensional problems may best be handled by using conjugate gradient or even quasi-Newton methods (see Fletcher 1980).

Remark 4. This algorithm is modified for the calculation of maximum likelihood estimates, rather than posterior modes, by adopting a flat prior. Note that a particular specification of the prior may complicate or simplify the M step.

Two important considerations regarding the implementation of this Monte Carlo algorithm are the monitoring convergence of the algorithm and the specification of m. Regarding the specification of m, it is noted that it is inefficient to start with a large value of m when the current approximation to the maximizer may be far from the true value. Rather, it is recommended that one increase m as the current approximation moves closer to the true maximizer. One may monitor the convergence of the algorithm by plotting (or simply examining a table of) $\theta^{(i)}$ versus the iteration (i). After a certain number of iterations, the plot will reveal that the process has stabilized; that is, there will be random fluctuation about the $\theta = \hat{\theta}$ line. At such a point, one may terminate the algorithm or continue with a larger value of m that will further decrease the system variability.

3.1.1. Derivatives in the Context of the MCEM Algorithm. The gradient (score function) and Hessian (observed Fisher information) of the observed log posterior are both of use in accelerating the EM algorithm (Hartley 1958; Louis 1982; Meilijson 1989), and the Hessian is of use in specifying the variance–covariance matrix of the normal approximation to the observed posterior. The gradient of the observed log posterior is given by

$$\int_Z \frac{D\log(p(\theta \mid y, z))}{D\theta} p(z \mid y, \theta)\, dz$$

(see Louis 1982; Meilijson 1989). Given a sample $z^{(1)}, \ldots, z^{(m)}$ from the current approximation to the conditional predictive distribution $p(z \mid \theta^{(i)}, y)$, the current approximation to the gradient is given by

$$\frac{1}{m}\sum_{j=1}^{m} \frac{D\log(p(\theta \mid z^{(j)}, y))}{D\theta}. \qquad (3.2)$$

The Hessian of the observed log posterior is given by

$$\int_Z \frac{D^2\log(p(\theta \mid y, z))}{D^2\theta} p(z \mid y, \theta)\, dz + \int_Z \left(\frac{D\log(p(\theta \mid y, z))}{D\theta}\right)^2 p(z \mid y, \theta)\, dz - \left[\int_Z \frac{D\log(p(\theta \mid y, z))}{D\theta} p(z \mid y, \theta)\, dz\right]^2$$

(see Louis 1982; Meilijson 1989). Given the sample of latent data patterns, the current approximation to the Hessian is given by

$$\frac{1}{m}\sum_{j=1}^{m} \frac{D^2\log(p(\theta \mid y, z^{(j)}))}{D^2\theta} \times \frac{1}{m}\sum_{j=1}^{m}\left(\frac{D\log(p(\theta \mid y, z^{(j)}))}{D\theta}\right)^2 - \left[\frac{1}{m}\sum_{j=1}^{m}\left(\frac{D\log(p(\theta \mid y, z^{(j)}))}{D\theta}\right)\right]^2. \qquad (3.3)$$

Wei (1989) presented a least squares estimate of the first term in (3.3).

3.1.2. The Genetic Linkage Model. We consider the genetic linkage model examined in Dempster et al. (1977), Louis (1982), Rao (1973), and Tanner and Wong (1987). The model is a multinomial with four categories, having observed counts of $y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$ and cell probabilities specified by $(1/2 + \theta/4, (1 -$

Table 1. MCEM History—Linkage Data

Iteration	*Theta*
1	.5833
2	.6222
3	.6192
4	.6321
5	.6153
6	.6259
7	.6238
8	.6245
9	.6270
10	.6265
11	.6264
12	.6270

$\theta)/4, (1 - \theta)/4, \theta/4)$. The observed data y are augmented by splitting the first cell of the multinomial into two cells, one having probability 1/2, the other having probability $\theta/4$. In this way, the augmented data set is given by $x = (x_1, x_2, x_3, x_4, x_5)$, where $x_1 + x_2 = 125$, $x_3 = y_2$, $x_4 = y_3$, and $x_5 = y_4$. Note that the augmented log-posterior [for a uniform prior on (0, 1)] is equal to $(x_2 + x_5)\log(\theta) + (x_3 + x_4)\log(1 - \theta) + C$. Given the current approximation to the maximizer of the observed posterior ($\theta^{(i)}$), the Monte Carlo E step is given by (a) draw $z^{(1)}, z^{(2)}, \ldots, z^{(m)}$ from the distribution Bi(125, $\theta^{(i)}/(\theta^{(i)} + 2)$) and (b) form $Q_{i+1}(\theta, \theta^{(i)}) = 1/m \sum_{j=1}^{m} \log(p(\theta \mid z^{(j)}, y))$. In the M step maximize Q_{i+1} over θ to obtain $\theta^{(i+1)}$. Note that $Q_{i+1}(\theta, \theta^{(i)})$ can be written as $(\text{ave} + x_5)\log(\theta) + (x_3 + x_4)\log(1 - \theta)$, where $\text{ave} = 1/m \sum_{j=1}^{m} z^{(j)}$.

Table 1 presents the history of an implementation of the MCEM algorithm for this problem. The algorithm was initiated with $\theta = .4$ and m was taken to be 10 (1,000) for iterations 1–8 (9–12). The true maximizer of the observed posterior is equal (to four places) to .6268. As can be seen from the table, for $m = 10$, the process seems to stabilize by the eighth iteration to yield a value of the maximizer correct to two decimal places. From the results of iterations 9–11, the maximizer is determined (to three decimal places) to be .627.

3.2 The Relation to Data Augmentation

At this point, it is instructive to note the relationship between the Monte Carlo implementation of the EM algorithm and the data augmentation algorithm (Tanner and Wong 1987). The data augmentation algorithm is an iterative method for the computation of the *entire* posterior density, rather than just a maximizer. Given the current approximation $g_i(\theta)$ to the observed posterior $p(\theta \mid y)$, the data augmentation algorithm specifies that one (a) generate a sample $z^{(1)}, \ldots, z^{(m)}$ from the current approximation to the predictive distribution $p(z \mid y)$ and (b) update the current approximation to $p(\theta \mid y)$ to be the mixture of augmented posteriors of θ, given the augmented data from (a), that is, $g_{i+1}(\theta) = 1/m \sum_{j=1}^{m} p(\theta \mid z^{(j)}, y)$. Steps (a) and (b) are then iterated. Tanner and Wong (1987) presented regularity conditions under which the algorithm converges. The $z^{(j)}$'s from the final iteration facilitate the computation of the expectation of any functional of the parameters. The case when $m = 1$ is of special interest. See Gelfand and Smith (1990).

To generate a sample of latent data, given the current guess to the posterior, Tanner and Wong (1987) suggested the following: (a1) generate θ from $g_i(\theta)$ and (a2) generate z from $p(z \mid \phi, y)$, where ϕ is the value of the parameter generated in (a1). Note that the multiple imputations in the data augmentation algorithm are drawn from the current approximation to $p(z \mid y)$. In the MCEM algorithm the multiple imputations are drawn from $p(z \mid \theta^{(i)}, y)$, where $\theta^{(i)}$ is the current approximation to the maximizer of the observed posterior.

3.3 The Poor Man's Data Augmentation (PMDA) Algorithms

3.3.1. PMDA 1. The final observation in the preceding section suggests a simple modification to the MCEM algorithm that yields an estimate of the *entire* posterior, rather than just a maximizer and the curvature at this point to specify a normal approximation to the observed posterior. Equation (2.4) of Tiemey, Kass, and Kadane (1986) implies that $p(z \mid y) = p(z \mid y, \hat{\theta})(1 + O(n^{-1}))$, where $\hat{\theta}$ is the mode of the observed posterior. This equation suggests that having obtained the mode of the observed posterior ($\hat{\theta}$), the following *noniterative* algorithm (the PMDA algorithm) will yield an approximation to the observed posterior. The algorithm is called a "poor man's" version of the data augmentation algorithm because it is intended for those who cannot afford to sample from $p(z \mid y)$. (a) Generate a sample $z^{(1)}, \ldots, z^{(m)}$ from the conditional predictive distribution $p(z \mid \hat{\theta}, y)$. (b) Approximate the observed posterior by the mixture of augmented posteriors of θ, mixed over the latent data patterns from (a).

$$\frac{1}{m}\sum_{j=1}^{m} p(\theta \mid z^{(j)}, y). \tag{3.4}$$

In this way, in large samples, Equation (3.4) will provide a refinement to the normal approximation of the observed posterior. In small samples, Equation (3.4) may be used as a diagnostic to the normal approximation. In particular, if evidence of skewness or multimodality is detected in (3.4), then the normal approximation to the observed posterior may be misleading. In such a case, one may wish to proceed to the full data augmentation algorithm using (3.4) as a starting point (g_o) for the algorithm. In general, PMDA can provide a good starting point for the data augmentation algorithm.

3.3.2. PMDA 2. Equation (3.4) is an approximation because the multiple imputations are sampled from $p(z \mid \hat{\theta}, y)$ rather than from $p(z \mid y)$. If $p(z \mid y)$ is straightforward to evaluate as a function of z, the observed posterior is easily calculated by using the technique of importance sampling (Ripley 1987). To calculate the observed posterior, given a sample $z^{(1)}, z^{(2)}, \ldots, z^{(m)}$ from $p(z \mid \hat{\theta},$

y), assign the weights

$$w_j = \frac{p(z^{(j)} \mid y)}{p(z^{(j)} \mid \hat{\theta}, y)} \tag{3.5}$$

to the imputations. That is, replace Equation (3.4) with

$$\frac{\sum_{j=1}^{m} w_j p(\theta \mid z^{(j)}, y)}{\sum_{j=1}^{m} w_j}. \tag{3.6}$$

In practice, $p(z^{(j)} \mid y)$ may be difficult to compute and a second-order approximation to $p(z \mid y)$ is available. Note that $p(z \mid y) = \int_\theta p(z \mid \theta, y)p(\theta \mid y)\, d\theta = \mathrm{E}(p(z \mid \theta, y))$. This observation suggests that one may use equation (2.5a) of Tierney et al. (1986) to obtain a second-order approximation to $p(z \mid y)$ to be used in place of $p(z \mid y)$ in Equation (3.5) above. To motivate this approximation, note that the $-nh^*$ function of Tierney et al. [1986; they typically take $-nh^*$ to be the sum of the log-posterior distribution ($-nh$) and the log of the function that is integrated against the posterior] in the present context is given by $-nh^*(\theta, z) = \log(p(\theta \mid y, z)) + \log(p(z \mid y))$. In this way, the maximizer of function $-nh^*(\theta, z^{(j)})$ with respect to θ is equal to the corresponding maximizer of $\log(p(\theta \mid y, z^{(j)}))$. Similarly, the $-nh$ function of Tierney et al. (1986) in the present context is given by $-nh(\theta, z) = \log(p(\theta \mid y, z)) - \log(p(z \mid \theta, y)) + \log(p(z \mid y))$. Let the maximizer of $p(\theta \mid y, z^{(j)})$ and $p(\theta \mid y)$ be denoted as $\hat{\theta}_j^*$ and $\hat{\theta}$, respectively. (Note that the latter maximizer is provided as output from the MCEM algorithm and the former maximizer is typically easy to obtain.) In this way, equation (2.5a) of Tierney et al. (1986) yields the second-order approximation

$$p(z^{(j)} \mid y) \propto (\det \Sigma^*)^{1/2} \frac{p(\hat{\theta}_j^* \mid y, z^{(j)})}{p(\hat{\theta} \mid y, z^{(j)})} p(z^{(j)} \mid y, \hat{\theta}),$$

where Σ^* is minus the inverse Hessian of $\log(p(\theta \mid y, z^{(j)}))$ evaluated at $\hat{\theta}_j^*$. Plugging this approximation into Equation (3.5), we have

$$w_j = [\det \Sigma^*]^{1/2} \frac{p(\hat{\theta}_j^* \mid y, z^{(j)})}{p(\hat{\theta} \mid y, z^{(j)})}.$$

3.3.3. The Genetic Linkage Example (continued). To illustrate the PMDA algorithms, consider the following small-sample data set for the genetic linkage model: (14, 0, 1, 5). [See Tanner and Wong (1987) for the full data augmentation analysis of these data.] The MCEM algorithm was run with $m = 5{,}000$ for 15 iterations, yielding $\hat{\theta} = .9034$. Using this value of $\hat{\theta}$, 5,000 samples were drawn from the conditional predictive distribution $p(z \mid y, \hat{\theta})$. Figure 1 presents the mixture of augmented posteriors, mixed over these PMDA 1 imputations (dotted line), along with the exact observed posterior (solid line) and the PMDA 2 mixture (dashed line). As can be seen from the figure, in this case PMDA 1 successfully recovers the highly skewed shape of the observed posterior. PMDA 2 is even more congruent with the true observed posterior.

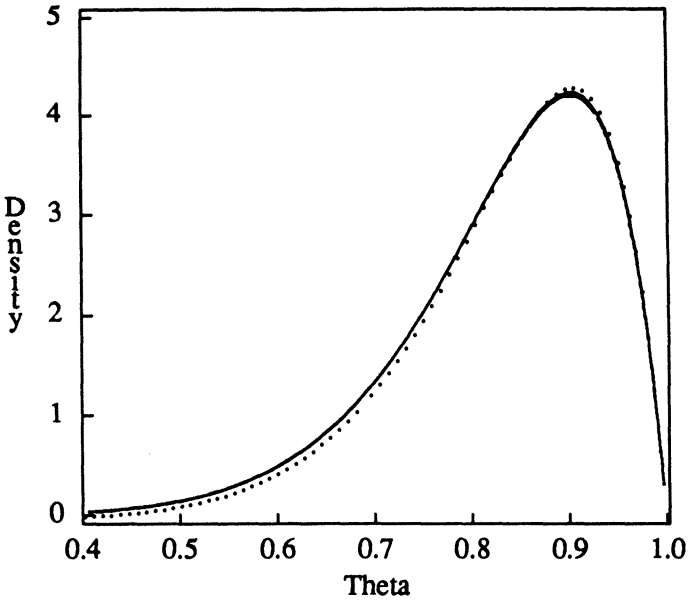

Figure 1. True Posterior and the PMDA Estimates.

In practice, the magnitude of the error of the first-order approximation will not be known. Having obtained PMDA 1, one may wish to "check" it against PMDA 2. If they are in accord, there may be little reason to doubt the validity of either approximation. If they are in discord, it is not clear whether the error in PMDA 2 can be neglected. Alternatively, one may wish to proceed to the implementation of the full data augmentation algorithm using PMDA 2 [see Eq. (3.6)] as a starting point (i.e., $g_0(\theta)$] for the data agumentation algorithm. Tanner and Wong (1987) gave conditions under which the data augmentation algorithm converges to be observed posterior. The error in the data augmentation algorithm (due to the Monte Carlo variation) is easily quantified.

4. REGRESSION ANALYSIS WITH CENSORED DATA

As a second example, we consider the motorette data set reported in Crawford (1970) and analyzed in Schmee and Hahn (1979). The data represent the results of temperature-accelerated life tests on electrical insulation in 40 motorettes, when 10 motorettes were tested at each of four temperatures in degrees Centigrade (150°, 170°, 190°, and 220°). Motorettes were on study for different periods depending on temperature, resulting in a total of 17 failed and 23 unfailed units. Following Schmee and Hahn (1979), the model used to analyze these data is $\mu_i = \rho_0 + \rho_1 v_i + \sigma\varepsilon_i$ ($i = 1, \ldots, 40$), where $\mu_i = \log(i\text{th failure time})$, $v_i = 1{,}000/(T_i + 273.2)$, T_i is the ith level of temperature, and the errors are assumed to follow a standard normal distribution. V will be used to denote the design matrix for this data set.

Before proceeding with the MCEM and PMDA algorithms, two questions must be addressed. For this problem, what is the functional form of the augmented

posterior $p(\theta \mid z, y)$? What is the conditional predictive distribution $p(z \mid \theta, y)$?

Regarding the form of the augmented posterior, it is well known that the marginal posterior distribution can be factored exactly into the product of the marginal density of σ^2 and the conditional density of $\boldsymbol{\rho}$ given σ^2. Moreover, for the prior $p(\sigma^2, \boldsymbol{\rho}) \propto \sigma^{-2}$, the marginal of σ^2 is that of the distribution of the random variable $((n - p)s^2)/(\chi^2_{(n-p)})$, where χ^2_v is a chi-squared random variable with degree of freedom, v, whereas the conditional marginal posterior distribution of $\boldsymbol{\rho}$ is a multivariate normal distribution (given σ^2) centered at the least squares estimate of $\boldsymbol{\rho}$ for the augmented data set (Box and Tiao 1973).

In answer to the second question, the conditional predictive distribution for a right-censored observation (μ_{t_0}) is the conditional normal distribution $\phi(s)/(1 - \Phi(z_{t_0}))$, where $\phi(s)$ and $\Phi(s)$ are the density and cdf of the standard normal distribution, respectively, and z_{t_0} is the value $z_{t_0} = (u_{t_0} - \mathbf{v}_{t_0}^T\boldsymbol{\rho})/\sigma$. Regarding the implementation of the M step, it is noted that

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{m} (\mathbf{u}_i - V\hat{\boldsymbol{\rho}})'(\mathbf{u}_i - V\hat{\boldsymbol{\rho}})}{m(n + 2)}$$

and $\hat{\boldsymbol{\rho}} = (V'V)^{-1}V'(\sum_{i=1}^{m} \mathbf{u}_i)/m$, where $\mathbf{u}_i$ is the ith augmented data set, maximize the Q function given in Equation (3.1).

Table 2 presents the history of an implementation of the MCEM algorithm for these data. The algorithm was initiated with $\rho_0 = -4.931$, $\rho_1 = 3.747$, and $\sigma^2 = .0247$. The value of m was equal to 50 (5,000) for iterations 1–14 (15–18). From the final three iterations, it can be seen that the maximum posterior estimates of ρ_0, ρ_1, and σ^2 are -5.96, 4.28, and .0589, respectively. [The corresponding estimates under a flat prior are -6.02, 4.31, and .067, respectively. These values agree with the corresponding figures given in Aitkin (1981).]

To illustrate the PMDA algorithms, we will examine the

Table 2. MCEM History—Motorette Data

Iteration	ρ_0	ρ_1	σ^2 ($\times 100$)
1	−5.27	3.93	3.36
2	−5.61	4.10	4.07
3	−5.64	4.12	4.65
4	−5.77	4.18	5.08
5	−5.81	4.20	5.31
6	−5.84	4.22	5.41
7	−5.85	4.23	5.62
8	−5.86	4.23	5.70
9	−5.94	4.27	5.72
10	−5.88	4.24	5.82
11	−5.97	4.29	5.93
12	−5.88	4.24	5.80
13	−5.78	4.19	5.77
14	−5.94	4.27	5.67
15	−5.97	4.28	5.84
16	−5.96	4.28	5.89
17	−5.96	4.28	5.89
18	−5.96	4.28	5.89

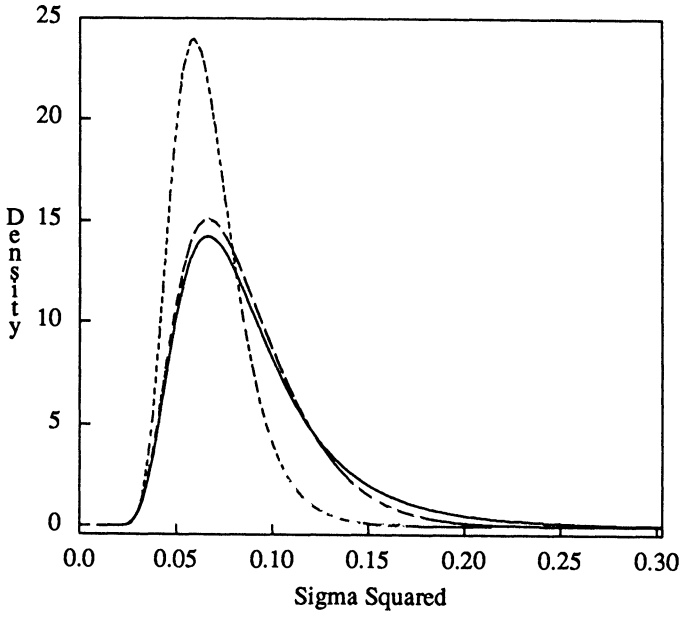

Figure 2. Data Augmentation and PMDA Estimates.

σ^2 marginal. Note that an estimate of $p(\sigma^2 \mid y)$ based on Equation (3.4) is readily available since

$$p(\sigma^2 \mid y) = \int p(\sigma^2, \boldsymbol{\rho} \mid y)\, d\boldsymbol{\rho} = \frac{1}{m}\sum_{i=1}^{m} \int p(\sigma^2, \boldsymbol{\rho} \mid \mathbf{u}_i)\, d\boldsymbol{\rho}$$

$$\propto \frac{1}{m}\sum_{i=1}^{m} \frac{s_i^{38}}{(\sigma^2)^{20}} \exp\left(-\frac{19s_i^2}{\sigma^2}\right),$$

where s_i^2 is the least squares estimate of σ^2 for the ith augmented data set, due to the inverse chi-squared conditional normal factorization.

Figure 2 presents PMDA 1 (short dashed line), PMDA 2 (long dashed line), and the data augmentation estimate (solid line) of the marginal based on $m = 5{,}000$. Clearly, the normal approximation to the marginal does not seem appropriate in this case. [In fact, the normal approximation is not appropriate even on the $\log(\sigma)$ scale.] PMDA 1 gives a hint of the skew in the marginal posterior. PMDA 2 does, however, represent an improvement.

In this example, there is a noticeable discrepancy between PMDA 1 and PMDA 2. In practice, having noticed such a discrepancy, one may want to run several iterations of the data augmentation algorithm, using PMDA 2 [see Eq. (3.6)] as a starting point for the data augmentation algorithm.

5. WHAT'S OUT THERE?

Several algorithms are available that make use of the data augmentation principle: augment the data to facilitate the analysis. The EM and MCEM algorithms provide a minimal amount of information to the data analyst. These algorithms yield the location of the normal approximation to the posterior distribution. Modifications by Louis (1982) and Meilijson (1989) allow for the specification of the scale

of the normal approximation in the EM context. The PMDA algorithms, in contrast, approximate the *entire* posterior distribution, thus allowing for nonquadratic log-posteriors. The sampling importance resampling (SIR) algorithm (Rubin 1987) is a noniterative algorithm for the estimation of the entire posterior. It is most efficient when a good approximation to the posterior (importance function) is available. At the upper end are the data augmentation algorithm and the Gibbs sampler. [Regarding the Gibbs sampler, see Geman and Geman (1984), Li (1988), and Gelfand and Smith (1990).] The latter two algorithms are iterative and can be shown to converge to the true posterior distribution under mild regularity conditions. They allow for the calculation of the posterior distribution of any functional of the parameters, for example, the content and boundary of the highest posterior density region (Wei 1989). In this regard, it is noted that the PMDA algorithms can provide good starting points to the SIR, data augmentation, and Gibbs sampler algorithms.

[*Received June 1989. Revised January 1990.*]

REFERENCES

Aitkin, M. (1981), "A Note on the Regresssion Analysis of Censored Data," *Technometrics,* 23, 161–163.

Box, G. E. P. and Tiao, G. (1973), *Bayesian Inference in Statistical Analysis,* Reading, MA: Addison-Wesley.

Boyles, R. A. (1983), "On the Convergence of the EM Algorithm," *Journal of the Royal Statistical Society,* Ser. B, 45, 47–50.

Crawford, D. E. (1970). "Analysis of Incomplete Life Test Data on Motorettes," *Insulation/Circuits,* 16, 43–48.

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,* Ser. B, 39, 1–38.

Diggle, P. J., and Gratton, R. J. (1984), "Monte Carlo Methods of Inference for Implicit Statistical Models," *Journal of the Royal Statistical Society,* Ser. B, 46, 193–227.

Fletcher, R. (1980), *Practical Methods of Optimization,* New York: John Wiley.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398–409.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721–741.

Hartley, H. O. (1958), "Maximum Likelihood Estimation from Incomplete Data," *Biometrics,* 27, 783–823.

Li, K. H. (1988), "Imputation Using Markov Chains," *Journal of Statistical Computation and Simulation,* 30, 57–79.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society,* Ser. B, 44, 226–233.

Meilijson, I. (1989), "A Fast Improvement to the EM Algorithm on Its Own Terms," *Journal of the Royal Statistical Society,* Ser. B, 51, 127–138.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications,* New York: John Wiley.

Ripley, B. D. (1987), *Stochastic Simulation,* New York: John Wiley.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys,* New York: John Wiley.

Schmee, J., and Hahn, G. J. (1979), "A Simple Method for Regression Analysis With Censored Data," *Technometrics,* 21, 417–432.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association,* 82, 528–550.

Tierney, L., Kass, R. E., and Kadane, J. B. (1986), "Approximation of Posterior Expectations and Variances Using Laplace's Method," Technical Report 385, Department of Statistics, Carnegie-Mellon University.

Wei, G. C. G. (1989), "Posterior Distribution Computations With Applications to Censored Regression Data," unpublished Ph.D. thesis, Department of Statistics, University of Wisconsin–Madison.

Wu, C. F. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics,* 11, 95–103.