Fitting Mixed-Effects Models Using Efficient EM-Type Algorithms

David A. VAN DYK

In recent years numerous advances in EM methodology have led to algorithms which can be very efficient when compared with both their EM predecessors and other numerical methods (e.g., algorithms based on Newton–Raphson). This article combines several of these new methods to develop a set of mode-finding algorithms for the popular mixed-effects model which are both fast and more reliable than such standard algorithms as `proc mixed` in SAS. We present efficient algorithms for maximum likelihood (ML), restricted maximum likelihood (REML), and computing posterior modes with conjugate proper and improper priors. These algorithms are not only useful in their own right, but also illustrate how parameter expansion, conditional data augmentation, and the ECME algorithm can be used in conjunction to form efficient algorithms. In particular, we illustrate a difficulty in using the typically very efficient PXEM (parameter-expanded EM) for posterior calculations, but show how algorithms based on conditional data augmentation can be used. Finally, we present a result that extends Hobert and Casella's result on the propriety of the posterior for the mixed-effects model under an improper prior, an important concern in Bayesian analysis involving these models that when not properly understood has lead to difficulties in several applications.

Key Words: EM algorithm; ECME algorithm; Gaussian hierarchical models; Posterior inference; PXEM algorithm; Random-effects models; REML; Variance-component models; Working parameters.

1. INTRODUCTION

The EM algorithm (Dempster, Laird, and Rubin 1977) has long been a popular tool for statistical analysis in the presence of missing data or in problems that can be formulated as such. Fitting mixed-effects models is among the most important uses of the EM algorithm as illustrated by the great variety and number of applications (see, e.g., Meng and Pedlow 1992; Meng and van Dyk 1997) and its development in the statistical literature (e.g., Laird 1982; Laird and Ware 1982; Dempster, Selwyn, Patel, and Roth 1984; Laird, Lange, and Stram 1987; Liu and Rubin 1994; and Meng and van Dyk 1998). There is no doubt that the reason for EM's popularity compared with other numerical methods (e.g., Newton–Raphson, as developed by Thompson and Meyer

David A. van Dyk is an Assistant Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: vandyk@stat.harvard.edu).

Journal of Computational and Graphical Statistics, Volume 9, Number 1, Pages 78–98

1986; Lindstrom and Bates 1988; and Callahan and Harville 1991; see also Harville 1977) which can be much faster than the early EM implementations is EM's superior stability properties (e.g., monotone convergence in log-likelihood or log posterior). For example, without extra computational effort, such numerical methods can converge to negative variance estimates (e.g., Thompson and Meyer 1986; Callahan and Harville 1991). Even implementations released in standard software which incorporate special monitoring can converge to a point outside the parameter space (e.g., SAS; see Section 4) or to the wrong point within the parameter space (e.g., S-Plus; see Meng and van Dyk 1998 and Section 4). The primary goal of this article is to build algorithms that are very fast but maintain the stability properties of EM-type algorithms.

This goal maintains the spirit of several recent advances in EM methodology. For example, Meng and van Dyk (1998) (see also Foulley and Quaas 1995) developed an alternative EM-type implementation (i.e., an ECME algorithm) for the mixed-effects model that substantially reduced the computational effort for obtaining maximum likelihood (ML) and restricted maximum likelihood (REML) estimates compared with earlier implementations (e.g., Laird, Lange, and Stram 1987). This adaptation was further improved by Liu, Rubin, and Wu (1998) in the special case of ML estimation with univariate response (i.e., a type of regression with heterogeneous residual variance) using the PXEM algorithm. In this article, we show how PXEM can be used for ML and REML model fitting of much more general mixed-effects models. We also show how ECME methodology can be used to eliminate data augmentation for the residual variance parameter in addition to the fixed-effects parameters while maintaining an algorithm which is completely in closed form. This extends Liu and Rubin's (1995) ECME algorithm which starts with the less efficient algorithm of Laird, Lange, and Stram (1987) and requires nested iterations for the ECME update of the residual variance.

This article is organized into four additional sections. In Section 2, after a brief review of EM, ECME (Liu and Rubin 1995), and working parameters (Meng and van Dyk 1997), we extend parameter-expanded EM (PXEM; Liu, Rubin, and Wu 1998) to compute posterior modes and illustrate why this efficient algorithm can be difficult to use in this setting. Section 3 uses the methods developed and reviewed in Section 2 to construct several new algorithms for fitting variations of the mixed-effects model. Section 4 illustrates briefly the computational speed and stability of the algorithms relative to commercially available software. Finally, Section 5 contains concluding remarks and an Appendix proves a result on the propriety of the posterior distributions when using certain improper priors.

2. EFFICIENT DATA AUGMENTATION

2.1 Background on EM-Type Algorithms

The EM algorithm is designed to compute a (local) mode, $\theta^\star$, of $\ell(\theta|Y_{\text{obs}}) = \log p(Y_{\text{obs}}|\theta) + \log p(\theta)$, where the parameter θ is allowed to vary over some space Θ and Y_{obs} is the observed data. For likelihood calculations, $\log p(\theta) = 0$ for all $\theta \in \Theta$ and $\ell(\theta|Y_{\text{obs}})$ is the log-likelihood; for Bayesian calculations $\ell(\theta|Y_{\text{obs}})$ refers to the log posterior. Throughout this article, the notation ℓ is used for a log posterior or a log-likelihood.

A data-augmentation scheme, $p(Y_{\text{aug}}|\theta)$ is a model defined so that

$$\int_{\{Y_{\text{aug}}:\mathcal{M}(Y_{\text{aug}})=Y_{\text{obs}}\}} p(Y_{\text{aug}}|\theta)dY_{\text{aug}} = p(Y_{\text{obs}}|\theta), \tag{2.1}$$

where $\mathcal{M}$ is some many-to-one mapping. EM iteratively computes θ by setting $\theta^{(t+1)} = \text{argmax}_{\theta\in\Theta}Q(\theta|\theta^{(t)})$, where $Q(\theta|\theta^{(t)}) = \int \ell(\theta|Y_{\text{aug}})p(Y_{\text{aug}}|\theta^{(t)}, Y_{\text{obs}})dY_{\text{aug}}$, with $\ell(\theta|Y_{\text{aug}}) = \log p(Y_{\text{aug}}|\theta) + \log p(\theta)$. (Here and henceforth integration over Y_{aug} is over the set $\{Y_{\text{aug}} : \mathcal{M}(Y_{\text{aug}}) = Y_{\text{obs}}\}$.) Computing the expectation, $Q(\theta|\theta^{(t)})$, is known as the E-step, while the maximization operation is known as the M-step. It can be shown that this procedure assures that $\ell(\theta^{(t+1)}|Y_{\text{obs}}) \geq \ell(\theta^{(t)}|Y_{\text{obs}})$ and typically converges to a (local) maximum of $\ell(\theta|Y_{\text{obs}})$ (Dempster, Laird, and Rubin 1977; Wu 1983).

The choice of the data-augmentation scheme, $p(Y_{\text{aug}}|\theta)$, in (2.1) is not unique. In fact, this choice can greatly affect the rate of convergence of EM-type algorithms (Meng and van Dyk 1997; Liu, Rubin, and Wu 1998) and their stochastic counterparts such as the data augmentation algorithm (Tanner and Wong 1987; Meng and van Dyk 1999; Liu and Wu 1999; van Dyk and Meng in press). In order to choose $p(Y_{\text{aug}}|\theta)$ to result in efficient algorithms, the working parameter approach (e.g., Meng and van Dyk 1999) parameterizes the data-augmentation scheme so that

$$\int_{\{Y_{\text{aug}}:\mathcal{M}(Y_{\text{aug}})=Y_{\text{obs}}\}} p(Y_{\text{aug}}|\theta,\alpha)dY_{\text{aug}} = p(Y_{\text{obs}}|\theta), \tag{2.2}$$

for each α in some class, A. (Likewise, we sometimes index $Q(\theta|\theta')$ by the working parameter for clarity—that is, $Q_\alpha(\theta|\theta')$.) The method of conditional augmentation suggests choosing α to minimize (i.e., optimize) the global rate of convergence (e.g., Meng and Rubin 1994) of EM—that is, the largest eigenvalue of the matrix fraction of missing information,

$$DM^{\text{EM}}(\alpha) = I - I_{\text{obs}}I_{\text{aug}}^{-1}(\alpha), \tag{2.3}$$

where I_{obs} is the observed Fisher information and

$$I_{\text{aug}}(\alpha) = \text{E}\left[-\frac{\partial^2\ell(\theta|Y_{\text{aug}},\alpha)}{\partial\theta\cdot\partial\theta^\top}\middle|Y_{\text{obs}},\theta\right]\Bigg|_{\theta=\theta^\star},$$

is the expected augmented Fisher information. Note that we adopt the traditional terms (e.g., Fisher information) of the EM literature, which focuses on likelihood calculations, even though we also deal with more general posterior computations. If we choose α to minimize $I_{\text{aug}}(\alpha)$ in a positive semidefinite ordering sense we optimize the global rate of convergence. This idea has led to a number of very efficient algorithms for fitting multivariate t models, probit regression models, mixed-effects models, Poisson models for image reconstruction, and factor analysis models either directly (Fessler and Hero 1994, 1995; Meng and van Dyk 1997, 1998; van Dyk in press a) or indirectly through the PXEM algorithm (which is discussed in detail in Section 2.2; see also Liu, Rubin, and Wu 1998).

Liu and Rubin (1994) also realized that reducing $I_{\text{aug}}(\alpha)$ is the key to speeding up EM. In their ECME algorithm, the augmented information is reduced to the observed information for some of the parameters. For example, a simple ECME algorithm dichotomizes the model parameter, $\theta = (\theta_1, \theta_2)$. The M-step of EM is then broken into two steps. In the first step we set $\theta^{(t+1/2)}$ to the maximizer of $Q(\theta|\theta^{(t)})$ as a function of θ, subject to the constraint $\theta_2 = \theta_2^{(t)}$. In the second step θ is set to the maximizer of $\ell(\theta|Y_{\text{obs}})$ subject to the constraint $\theta_1 = \theta_1^{(t+1/2)}$. Since there is no data augmentation in the second step, we expect the algorithm to converge more quickly than EM, as was verified by Liu and Rubin (1994, 1995) in several examples.

We use both ECME and conditional augmentation to build efficient algorithms for fitting mixed-effects models in Section 3. First, however, we turn our attention to an important extension of conditional augmentation, namely the PXEM algorithm.

2.2 PXEM for Bayesian Calculations

Liu, Rubin, and Wu (1998) presented the PXEM algorithm as a fast adaptation of conditional augmentation in the special case when $p(\theta) \propto 1$; for example, in maximum likelihood estimation. Simply speaking, instead of conditioning on an optimal value of the working parameter, α, PXEM fits α in the iteration. Here we outline a generalization of this algorithm by using this same novel idea to compute posterior modes. We start by defining $Q_{\text{px}}(\theta, \alpha|\theta', \alpha_0) = \int \log[p(Y_{\text{aug}}|\theta, \alpha)p(\theta)]p(Y_{\text{aug}}|Y_{\text{obs}}, \theta', \alpha_0)dY_{\text{aug}}$. We then define $(\theta^{(t+1)}, \alpha^{(t+1)})$ as the maximizer of $Q_{\text{px}}(\theta, \alpha|\theta^{(t)}, \alpha_0)$, where α_0 is some fixed value. (Note $\alpha^{(t+1)}$ is not used subsequently.) Since $p(Y_{\text{obs}}|\theta) = p(Y_{\text{obs}}|\theta, \alpha) = p(Y_{\text{aug}}|\theta, \alpha)/p(Y_{\text{aug}}|Y_{\text{obs}}, \theta, \alpha)$ for any $\alpha \in A$ and any Y_{aug} such that $\mathcal{M}(Y_{\text{aug}}) = Y_{\text{obs}}$, we have

$$\ell(\theta|Y_{\text{obs}}) = Q_{\text{px}}(\theta, \alpha|\theta^{(t)}, \alpha_0) - \int \log[p(Y_{\text{aug}}|Y_{\text{obs}}, \theta, \alpha)]p(Y_{\text{aug}}|Y_{\text{obs}}, \theta^{(t)}, \alpha_0)dY_{\text{aug}}.$$

Since the first term on the right is maximized by $(\theta, \alpha) = (\theta^{(t+1)}, \alpha^{(t+1)})$, and the second is minimized by $(\theta, \alpha) = (\theta^{(t)}, \alpha_0)$, we have $\ell(\theta^{(t+1)}|Y_{\text{obs}}) \geq \ell(\theta^{(t)}|Y_{\text{obs}})$. That is, this generalization of the PXEM algorithm converges monotonically in log posterior. Following Wu (1983) we can further obtain that this algorithm converges to a stationary point or local maximum of the posterior. These results hold for any value of α_0 such that all quantities exist. In fact, the value of α_0 is generally irrelevant for a PXEM iteration and is simply set to some convenient value (e.g., $\alpha_0 = 1$ for scale working parameters and $\alpha_0 = 0$ for location working parameters).

We expect this algorithm to perform at least as well as an algorithm that fixes α (i.e., conditional data augmentation) in terms of the global rate of convergence because it essentially removes the conditioning on α in the data-augmentation scheme. Removing this conditioning reduces I_{aug} (in a positive semidefinite ordering sense) and thus improves the rate of convergence of EM (see Meng and van Dyk 1997 and Liu, Rubin, and Wu 1998 for details). Liu, Rubin, and Wu (1998) gave an alternative explanation—that by fitting α, we are performing a covariance adjustment to capitalize on information in the data-augmentation scheme. They also illustrated the substantial computational advantage PXEM can offer over other EM-type algorithms for ML estimation.

Unfortunately, this algorithm can be difficult to use for some Bayesian computations since the maximizer of $Q_{\text{px}}(\theta, \alpha|\theta^{(t)}, \alpha_0)$ may not exist in closed form, even when the corresponding maximum likelihood PXEM algorithm is in closed form and a conjugate prior is used. A simple random-effects model illustrates both this difficulty and the potential computation advantage. Suppose

$$Y_i = \alpha Z_i b_i + e_i \quad \text{with} \quad b_i \overset{\text{iid}}{\sim} N(0, \tau^2/\alpha^2) \quad \text{and} \quad e_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{2.4}$$

for $i = 1, \ldots, m$, where b_i and e_i are independent, $\{(Y_i, Z_i), i = 1, \ldots, m\}$ are the observed data, α is a working parameter, and all quantities are scalars. (It can easily be verified that $p(Y_{\text{obs}}|\theta, \alpha)$, with $\theta = (\sigma^2, \tau^2)$, does not depend on α.) We set $Y_{\text{aug}} = \{(Y_i, Z_i, b_i), i = 1, \ldots, m\}$ and $\alpha_0 = 1$ to define the data-augmentation scheme via (2.4). We consider the independent priors $\sigma^2 \sim \nu\sigma_0^2/\chi_\nu^2$ and $\tau^2 \sim \eta\tau_0^2/\chi_\eta^2$, which are conjugate for $p(Y_{\text{aug}}|\theta, \alpha_0)$, in which case

$$Q_{\text{px}}(\theta, \alpha|\theta', \alpha_0) = \mathrm{E}\Bigg[-\left(\frac{m+\nu}{2} + 1\right)\log(\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{m}(Y_i - \alpha Z_i b_i)^2 + \nu\sigma_0^2\right] \\ -\frac{m}{2}\log\left(\frac{\tau^2}{\alpha^2}\right) - \frac{\alpha^2}{2\tau^2}\sum_{i=1}^{m} b_i^2 - \left(\frac{\eta}{2} + 1\right)\log\tau^2 - \frac{\eta\tau_0^2}{2\tau^2}\Bigg| Y_{\text{obs}}, \theta', \alpha_0\Bigg]. \tag{2.5}$$

In the absence of prior information, the usual strategy is to reparameterize (e.g., set $\xi = \tau^2/\alpha^2$) in order to simplify the optimization of $Q_{\text{px}}(\theta, \alpha|\theta', \alpha_0)$. Although this works nicely with maximum likelihood it clearly does not work with (2.5). It can be shown that introducing a proper prior for α destroys the computational advantage of parameter expansion, while introducing a dependent prior (e.g., $\tau^2|\alpha \sim \eta\tau_0^2\alpha^2/\chi_\eta^2$) alters the model and $\theta^\star$. Thus, it seems difficult to optimize (2.5) without resorting to iterative numerical methods, the cost of which is likely to outweigh the benefit of PXEM. (This is certainly true with multiple random effects, in which case, the scalar τ^2 is replaced by a variance-covariance matrix.) Thus, we typically use conditional augmentation for Bayesian mode finding, at least when we use fully proper priors.

In cases where $Q_{\text{px}}(\theta, \alpha|\theta', \alpha_0)$ is easy to maximize, however, the algorithm can be very fast. Suppose for example we use the improper prior $p(\theta) \propto (\sigma^2)^{-1}(\tau^2)^{-1/2}$ (i.e., $\tau_0^2 = 0, \eta = -1$). Hobert and Casella (1996) verified that this prior results in a proper posterior as long as $m, n \geq 2$; see also the Appendix. In this case (2.5) is unbounded for τ^2 near zero. Thus, the PXEM algorithm converges in one step to the global mode $\tau^2 = 0$. Figure 1, for example, shows a contour plot of the posterior surface for an artificial dataset of size 100. For this dataset, a standard EM algorithm which fixes $\alpha = 1$ converges with global rate equal to one (empirical result) and thus takes many iterates to coverage. The second plot in Figure 1 shows a cross section of the posterior surface with $\sigma^2 = 1$. The conditional posterior is bimodal and a standard EM algorithm which computes $(\tau^2)^\star$ given $\sigma^2 = 1$ with α fixed at one converges to a (local) mode for $(\tau^2)^{(0)} \geq .026$, the local minimum (again this is an empirical result). The PXEM algorithm, on the other hand, again converges to the global mode in one step for any $(\tau^2)^{(0)}$.

We emphasize that Figure 1 is included to compare the behavior of PXEM and the standard EM algorithm. It is not necessarily better to converge to the global mode.

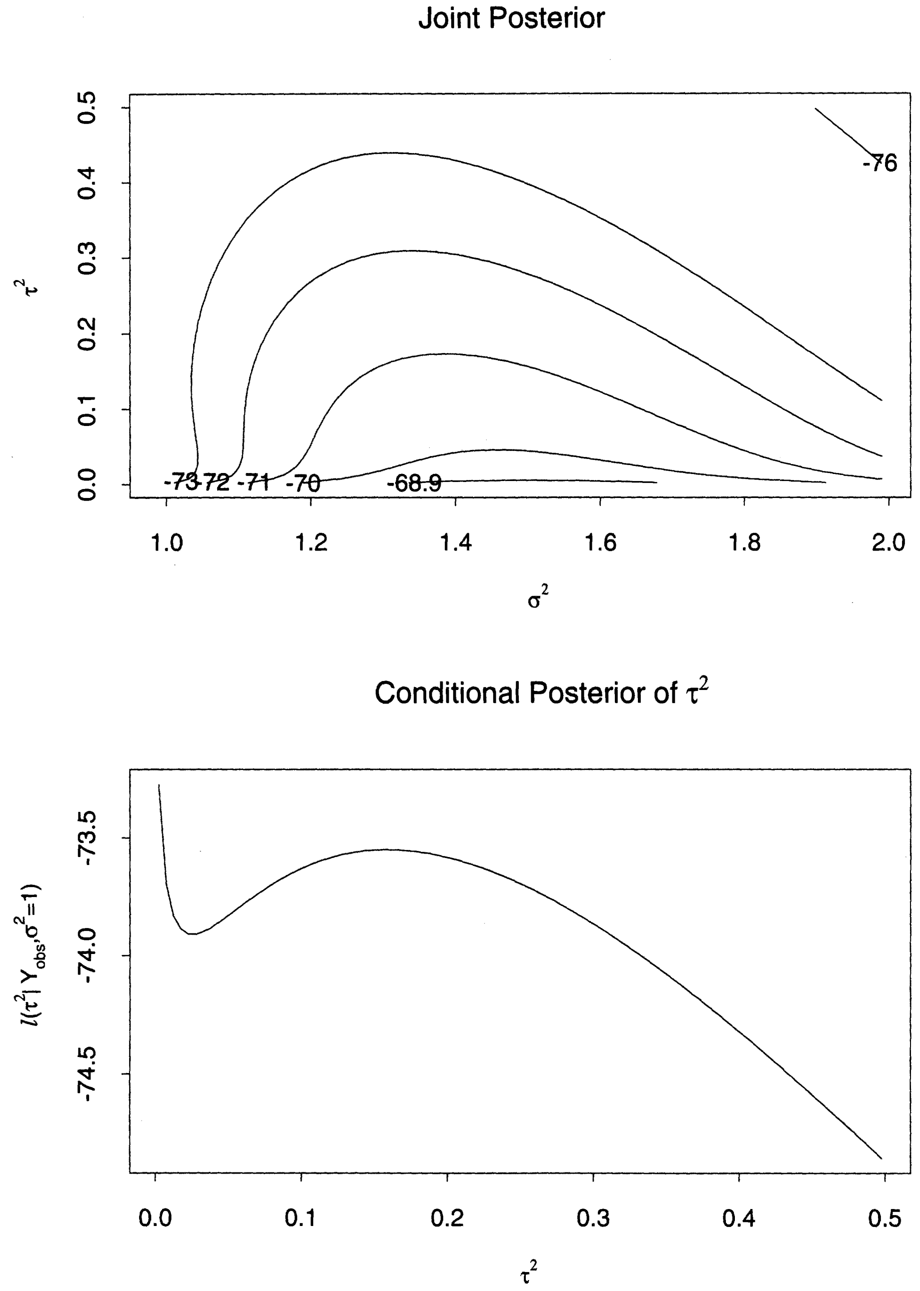

Figure 1. The posterior surface of the parameters in model (2.4) using an artificial dataset. The first plot is a contour plot of the joint posterior of (σ^2, τ^2); the second plot shows the conditional posterior of τ^2 given $\sigma^2 = 1$. In both cases PXEM converges to the global model in one step while the standard EM algorithm can be very slow to converge to a (local) mode.

In particular, the local (conditional) mode contains more of the posterior probability and is driven by the data rather than the prior. Our conclusion is that this improper prior should be used only with great care—the mode is not a sufficient summary of the posterior, at least on the original scale. It is noteworthy that using PXEM brings attention to this difficulty with the posterior.

3. ALGORITHMS FOR FITTING MIXED-EFFECTS MODELS

3.1 EFFICIENT ALGORITHMS WITH PROPER PRIORS

We begin in the fully Bayesian setting because the model notation is the most general and the algorithms that rely on conditional augmentation are simpler and serve as building blocks for the algorithms for computing REML estimates, posterior modes with improper priors (both in Section 3.2), and ML estimates (Section 3.3).

We consider the mixed-effects model of the general form

$$Y_i = X_i\beta + Z_ib_i + e_i \quad \text{with} \quad b_i \overset{\text{iid}}{\sim} N(0, \sigma^2 D) \quad \text{and} \quad e_i \overset{\text{ind}}{\sim} N(0, \sigma^2 R_i), \qquad (3.1)$$

where Y_i is $n_i \times 1$ for $i = 1, \dots, m$, X_i and Z_i are known covariates of dimension $n_i \times p$ and $n_i \times q$, respectively, β is a $p \times 1$ vector of fixed effects, $b_i = (b_{i1}, \dots, b_{iq})^\top$ are $q \times 1$ vectors of random effects, R_i are known $n_i \times n_i$ positive definite matrixes, and b_i and e_i are independent for each i. We parameterize the variance of the random effects in terms of the residual variance σ^2 to facilitate computation but have occasion to use both the parameterizations $T = \sigma^2 D$ and $LL^\top = D$, where L is the Cholesky factor of D (i.e., L is a lower triangular matrix). We introduce the standard priors which are conjugate for the data-augmentation schemes defined below

$$\beta|\sigma^2 \sim N_p(\mu_\beta, \sigma^2\Sigma_\beta), \quad \sigma^2 \sim \frac{\nu\sigma_0^2}{\chi^2_\nu}, \quad \text{and} \quad T \sim \text{inverse Wishart}(\eta, T_0), \qquad (3.2)$$

where the inverse Wishart is parameterized so that $\mathrm{E}(T) = (\eta - q - 1)^{-1}\, T_0$, with η the degrees of freedom. On occasion we replace the inverse Wishart prior for T with

$$\mathrm{vec}_{\mathrm{T}}(L) \sim N_{q(q+1)/2}(\mu_L, \sigma^2\Sigma_L), \qquad (3.3)$$

where $\mathrm{vec}_{\mathrm{T}}(M)$ is a vector containing the elements on or below the diagonal of M; the subscript T stands for triangular. Depending on the data-augmentation scheme one or the other of these priors is conjugate. Although the priors do not exactly coincide, either can be used to incorporate similar prior information. With the second prior, this is facilitated by noting that the diagonal element of σL are prior conditional standard deviations of the components of b_i (i.e., $\mathrm{sd}(b_{ij}|b_{i1}, \dots, b_{i,j-1})$ is the jth diagonal element of σL), while the jkth element of σL $(j > k)$ is the square root of the variance in b_{ij} not explained by $(b_{i1}, \dots, b_{i,k-1})$ that is explained by b_{ik}.

We use an EM-type algorithm to compute the mode of either $p(T, \sigma^2|Y_{\text{obs}})$ or $p(L, \sigma^2|Y_{\text{obs}})$, where $Y_{\text{obs}} = \{(Y_i, X_i, Z_i), i = 1, \dots, m\}$. We are interested in a marginal mode, since working in smaller parameter spaces typically leads to modes with better statistical properties (e.g., Gelman, Carlin, Stern, and Rubin 1995, sec. 9.5). We then use $p(b_1, \dots, b_m, \beta|\zeta, Y_{\text{obs}})$ with $\zeta = (\sigma^2, D)$ as derived below to draw inferences regarding the mixed effects.

Using the working parameter, α, we define $Y_{\text{aug}} = \{(Y_i, X_i, Z_i, L^{-\alpha}b_i), i = 1, \dots, m; \beta\}$ and derive two ECME algorithms to compute the marginal mode. Both of the algorithms dichotomize the parameter $\zeta = (D, \sigma^2)$ into D and σ^2, first updating D by maximizing either $Q_{\alpha=0}(T, \sigma^2|\zeta')$ or $Q_{\alpha=1}(L, \sigma^2|\zeta')$ subject to the constraint that σ^2 is fixed and second updating σ^2 by maximizing either $\ell(T, \sigma^2|Y_{\text{obs}})$ or $\ell(L, \sigma^2|Y_{\text{obs}})$ subject to the constraint that D is fixed. That is, the algorithms do not use data augmentation when updating σ^2 but do require data augmentation to update D. Because of the choice of prior, the first algorithm computes the posterior mode of $p(T, \sigma^2|Y_{\text{obs}})$, while the second computes the posterior mode of $p(L, \sigma^2|Y_{\text{obs}})$, hence the notation $Q_{\alpha=0}(T, \sigma^2|\zeta')$ and $Q_{\alpha=1}(L, \sigma^2|\zeta')$. Since the algorithms differ in the value of the working parameter α, we call them ECME$_0$ which corresponds to $\alpha = 0$, and ECME$_1$ which corresponds to $\alpha = 1$. As is illustrated in Section 4, the relative efficiency of ECME$_1$ and ECME$_0$ depends roughly on the relative size of $(\sigma^2)^\star$ and $\sum Z_i^\top T^\star Z_i$ (see Meng and van Dyk 1998 for details).

We begin with ECME$_0$ which updates D by optimizing

$$\begin{aligned} Q(T, \sigma^2|\zeta^{(t)}) = &-\left(\frac{n+\nu}{2}+1\right)\log(\sigma^2) - \frac{\nu_0\sigma_0^2}{2\sigma^2} \\ &-\frac{1}{2\sigma^2}\sum_{i=1}^{m} \mathrm{E}\left[(Y_i - X_i\beta - Z_ib_i)^\top R_i^{-1}(Y_i - X_i\beta - Z_ib_i)|Y_{\text{obs}}, \zeta^{(t)}\right] \\ &-\frac{m+\eta+q+1}{2}\log|\sigma^2 D| - \frac{1}{2\sigma^2}\left[\mathrm{tr}\left(D^{-1}\left(T_0 + \sum_{i=1}^{m}\mathrm{E}[b_ib_i^\top|Y_{\text{obs}}, \zeta^{(t)}]\right)\right)\right]. \end{aligned} \tag{3.4}$$

To compute the expectations, we use standard calculations to compute the following conditional distributions:

$$\beta|\zeta, Y_{\text{obs}} \sim N\left(\hat\beta(D), \sigma^2\left(\Sigma_\beta^{-1} + \sum_{i=1}^{m} X_i^\top U_i(D) X_i\right)^{-1}\right), \tag{3.5}$$

where $\hat\beta(D) = (\Sigma_\beta^{-1} + \sum_{i=1}^m X_i^\top U_i(D)X_i)^{-1}(\Sigma_\beta^{-1}\mu_\beta + \sum_{i=1}^m X_i^\top U_i(D)Y_i)$ and $U_i(D) = (R_i + Z_iDZ_i^\top)^{-1}$;

$$b_i|\zeta, \beta, Y_{\text{obs}} \sim N\left(\hat b_i(D, \beta), \sigma^2(D - DZ_i^\top U_i(D) Z_i D)\right), \tag{3.6}$$

where $\hat b_i(D, \beta) = DZ_i^\top U_i(D)(Y_i - X_i\beta)$; and

$$b_i|\zeta, Y_{\text{obs}} \sim N\left(\hat b_i(D, \hat\beta(D)), \sigma^2(D - DZ_i^\top P_i(D) Z_i D)\right), \tag{3.7}$$

where $P_i(D) = U_i(D) - U_i(D)X_i\left(\Sigma_\beta^{-1} + \sum_{i=1}^m X_i^\top U_i(D)X_i\right)^{-1} X_i^\top U_i(D)$. Using (3.7) to compute the second expectation in (3.4), it is easy to verify that $Q(T, \sigma^2|\zeta^{(t)})$ is maximized as a function of D by

$$D^{(t+1)} = \frac{1}{(\sigma^2)^{(t)}(m+\eta+q+1)}\left[T_0 + \sum_{i=1}^{m}\hat B_i(\zeta^{(t)})\right], \tag{3.8}$$

where $\hat B_i(\zeta) = \mathrm{E}[b_ib_i^\top|Y_{\text{obs}}, \zeta] = \hat b_i(D, \hat\beta(D))\hat b_i^\top(D, \hat\beta(D)) + \sigma^2(D - DZ_i^\top P_i(D)Z_iD)$. To update σ^2, we write,

$$\ell(\sigma^2, T|Y_{\text{obs}}) = \log p(\sigma^2, T) - \frac{1}{2}\log\left|\sigma^{-2}\left(\sum_{i=1}^{m} X_i^\top U_i(D) X_i + \Sigma_\beta^{-1}\right)\right|$$
$$-\frac{1}{2}\sum_{i=1}^{m}\log\left|\sigma^2 U_i^{-1}(D)\right| - \frac{p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}$$
$$\times\left[\sum_{i=1}^{m}(Y_i - X_i\hat{\beta}(D))^\top U_i(D)(Y_i - X_i\hat{\beta}(D)) + (\mu_\beta - \hat{\beta}(D))^\top\Sigma_\beta^{-1}(\mu_\beta - \hat{\beta}(D))\right],$$

which is maximized as a function of σ^2 with D fixed at $D^{(t+1)}$ by

$$(\sigma^2)^{(t+1)} = \frac{1}{n+\nu+2}\left[\sigma_0^2\nu\right.$$
$$+\sum_{i=1}^{m}\left(Y_i - X_i\hat{\beta}\left(D^{(t+1)}\right)\right)^\top U_i\left(D^{(t+1)}\right)\left(Y_i - X_i\hat{\beta}\left(D^{(t+1)}\right)\right)$$
$$\left.+\left(\mu_\beta - \hat{\beta}\left(D^{(t+1)}\right)\right)^\top\Sigma_\beta^{-1}\left(\mu_\beta - \hat{\beta}\left(D^{(t+1)}\right)\right)\right]. \tag{3.9}$$

This completes a single integration of ECME$_0$. We note the relationship between this ECME algorithm and the one given by Liu and Rubin (1994) for maximum likelihood estimation. Although they used the same data-augmentation scheme (i.e., $\alpha = 0$), Liu and Rubin's update for σ^2 is not in closed form because they use the parameterization, (β, σ^2, T) in place of (β, σ^2, D) in the constraint functions. The required numerical optimization slows down the algorithm substantially (see van Dyk and Meng 1997). The Liu and Rubin algorithm also does not include prior information and does not integrate out the fixed effects. Although their algorithm could be adapted to the Bayesian setting, it is unlikely to be fruitful since the numerical optimization in the update for σ^2 is slow. The utility of the parameterization (β, σ^2, D) was also noted by Schafer (1998) (see also Lindstrom and Bates 1988).

We now turn our attention to $\alpha = 1$ in the conditional-augmentation scheme and use the alternative prior given in (3.3) in order to maintain a closed-form algorithm. We rescale the random effects by L^{-1} and consider $\{c_i \equiv L^{-1}b_i, i = 1, \ldots, m\}$ to be the missing data. In order to update D, we rewrite (3.1) in terms of $(c_1, \ldots, c_m)$,

$$Y_i = X_i\beta + Z_iLc_i + e_i = X_i\beta + \sum_{j=1}^{q}\sum_{k=j}^{q} c_{ij}Z_{ik}l_{kj} + e_i, \tag{3.10}$$

where $c_i \sim N(0, \sigma^2 I)$, $e_i \sim N(0, \sigma^2 R_i)$, $L = (l_{kj})$, and $Z_i = (Z_{i1}, \ldots, Z_{iq})$ with Z_{ik} an $n_i \times 1$ vector to obtain

$$Q(L, \sigma^2|\zeta^{(t)})$$
$$= -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{m}\mathrm{E}\left[(Y_i - X_i\beta - Z_iLc_i)^\top R_i^{-1}(Y_i - X_i\beta - Z_iLc_i)\middle|Y_{\text{obs}}, \zeta^{(t)}\right]\right.$$
$$\left.+ (L - \mu_L)^\top\Sigma_L^{-1}(L - \mu_L)\right)$$

as a function of L. Thus,

$$\mathrm{vec}_{\mathrm{T}}(L^{(t+1)}) = \left(\sum_{i=1}^{n} \tilde{C}_i(\zeta^{(t)}) + \Sigma_L^{-1}\right)^{-1} \times \left(\mathrm{E}\left[\sum_{i=1}^{m} \tilde{X}_i^{\top} R_i^{-1}(Y_i - X_i\beta) \middle| Y_{\mathrm{obs}}, \zeta^{(t)}\right] + \Sigma_L^{-1}\mu_L\right), \qquad (3.11)$$

where $\tilde{X}_i$ is an $n_i \times (q(q+1)/2)$ matrix with columns $c_{ij}Z_{ik}$ for $j = 1, \ldots, q$ and $k = j, \ldots, q$ in the ordering that corresponds to $\mathrm{vec}_{\mathrm{T}}(L)$ and $\tilde{C}_i(\zeta) = \mathrm{E}[\tilde{X}_i^{\top} R_i^{-1} \tilde{X}_i | Y_{\mathrm{obs}}, \zeta]$, the elements of which are calculated using

$$\mathrm{E}\left[c_{ij} Z_{ik}^{\top} R_i^{-1} Z_{ik'} c_{ij'} | Y_{\mathrm{obs}}, \zeta\right] = [L^{-1}\hat{B}_i(\zeta)(L^{-1})^{\top}]_{jj'} Z_{ik}^{\top} R_i^{-1} Z_{ik'}, \qquad (3.12)$$

where $[M]_{jj'}$ if the (j, j')th element of the matrix M. To compute the expectation in (3.11) we note

$$\begin{aligned}&\mathrm{E}[c_{ij} Z_{ik}^{\top} R_i^{-1}(Y_i - X_i\beta) | Y_{\mathrm{obs}}, \zeta] \\ &\quad = [L^{-1}\hat{b}_i(D, \hat{\beta}(D))]_j Z_{ik}^{\top} R_i^{-1} Y_i - Z_{ik}^{\top} R_i^{-1} X_i^{\top} [\hat{H}_i(\zeta)]_j^{\top}, \qquad (3.13)\end{aligned}$$

where $[v]_j$ is the jth component of the vector v and $[\hat{H}_i(\zeta)]_j$ is the jth row of $\mathrm{E}[c_i\beta^{\top} | Y_{\mathrm{obs}}, \zeta]$,

$$\begin{aligned}\hat{H}_i(\zeta) &= L^{-1}\hat{b}_i(D, \hat{\beta}(D))[\hat{\beta}(D)]^{\top} \\ &\quad -\sigma^2 L^{\top} Z_i^{\top} U_i(D) X_i \left[\sum_{i=1}^{m} X_i^{\top} U_i(D) X_i + \Sigma_{\beta}^{-1}\right]^{-1}. \qquad (3.14)\end{aligned}$$

The iteration is completed by setting $D^{(t+1)} = L^{(t+1)}(L^{(t+1)})^{\top}$ and computing

$$\begin{aligned}(\sigma^2)^{(t+1)} = \frac{1}{n + \nu + 2 + q(q+1)/2} \Bigg[&\sigma_0^2 \nu + (L - \mu_L)^{\top} \Sigma_L^{-1} (L - \mu_L) \\ &+ \left(\mu_{\beta} - \hat{\beta}\left(D^{(t+1)}\right)\right)^{\top} \Sigma_{\beta}^{-1} \left(\mu_{\beta} - \hat{\beta}\left(D^{(t+1)}\right)\right) \\ &+ \sum_{i=1}^{m} \left(Y_i - X_i\hat{\beta}\left(D^{(t+1)}\right)\right)^{\top} U_i\left(D^{(t+1)}\right) \left(Y_i - X_i\hat{\beta}\left(D^{(t+1)}\right)\right) \Bigg], \end{aligned} \qquad (3.15)$$

which adjusts (3.9) for the prior. The matrix inversions here and in the various algorithms can be facilitated with the SWEEP operator (Beaton 1964; Little and Rubin 1987; Meng and van Dyk 1998).

We conclude this section with one final data-augmentation scheme. Above we considered the reparameterization $LL^{\top} = D$, where L is a lower triangular matrix. If instead we allow L to be an arbitrary invertible matrix, we can derive an even more general class of algorithms (see also Foulley and van Dyk in press). Unfortunately, since there is no ready interpretation of the elements of L in this case, establishing a meaningful prior distribution is difficult. Nonetheless, we derive the resulting algorithm, since it is a simple modification of ECME$_1$ and serves as a useful building block for algorithms with improper priors.

This algorithm, ECME$_2$, can be expressed easily if in (3.11)–(3.14) we consider L to be a $q \times q$ invertible matrix with prior $\text{vec}(L) \sim N_{q^2}(\mu_L, \Sigma_L)$, where $\text{vec}(M)$ is a vector containing all the elements of M. We also substitute $\tilde{X}_i$ with $\check{X}_i$, an $n_i \times q^2$ matrix with columns $c_{ij} z_{ik}$ for $j = 1, \ldots, q$ and $k = 1, \ldots, q$ in the ordering corresponding to $\text{vec}(L)$. With these changes in notation, $\text{vec}(L^{(t+1)})$ is given in (3.11)–(3.14) and $(\sigma^1)^{(t+1)}$ in (3.15) with the denominator replaced by $n + \nu + 2 + q^2$ to account for the change in prior.

3.2 REML Calculation and Improper Priors

Here we again consider model (3.1) but replace the prior for T with $p(T) \propto 1$. Although this corresponds to setting $T_0 = 0$ and $\eta = -(q+1)$ in (3.2), and therefore can be fit with ECME$_0$, in this important special case we can use the data-augmentation scheme used for ECME$_2$ to derive a parameter-expanded ECME$_2$ algorithm that is more efficient than ECME$_0$, ECME$_1$, or ECME$_2$. (An ECME$_1$ or ECME$_2$ algorithm can be used if we consider the prior $p(L) \propto 1$.) The algorithms derived here assume the improper prior

$$p(\beta, \sigma^2, T) \propto (\sigma^2)^{-(1+(p+\nu)/2)} \exp\left\{ -\frac{1}{2\sigma^2} [\nu\sigma_0^2 + (\beta - \mu_\beta)^\top \Sigma_\beta^{-1} (\beta - \mu_\beta)] \right\}.$$

If in addition, we set $\sigma_0^2 = 0$, $\nu = -(2+p)$ and $\Sigma_\beta^{-1} = 0$, the posterior mode of (σ^2, D) corresponds to the REML estimate (Laird and Ware 1982).

To derive the parameter-expanded ECME$_2$ algorithm, we define $Y_{\text{aug}} = \{(Y_i, X_i, Z_i, L^{-1} b_i),\ i = 1, \ldots, m\}$; here and in the remainder of the article, L is an arbitrary invertible $q \times q$ working parameter. We emphasize that L is not a transformation of T, but a free working parameter. We set $L_0 = I$ at each iteration, thus the E-step is the same for ECME$_0$ and the parameter expanded ECME$_2$ algorithm. We update the parameters using the same conditioning scheme as in the fully Bayesian setting. That is, we update (T, L) by maximizing $Q_{\text{px}}(T, \sigma^2, L | \zeta^{(t)}, L_0)$ with σ^2 fixed and update σ^2 by maximizing $\ell(T, \sigma^2 | Y_{\text{obs}})$ with D fixed. In particular, if we replace b_i with $L c_i$, $Q_{\text{px}}(T, \sigma^2, L | \zeta^{(t)}, L_0)$ is given by (3.4) which we maximize jointly as a function of D and L to update these parameters. This is accomplished via the transformation from (D, L) to $(\tilde{D}, L)$, where $\tilde{D} = L^{-1} D (L^{-1})^\top$. In particular, we maximize $Q_{\text{px}}(T, \sigma^2, L | \zeta^{(t)}, L_0)$ by computing $L^{(t+1)}$ using (3.11) and setting $\tilde{D}^{(t+1)}$ to the right side of (3.8). In these calculations we set $\Sigma_L^{-1} = 0$, $T_0 = 0$, and $\eta = -(q+1)$ throughout, and set $L = L_0 = I$ and $\tilde{X}_i$ to $\check{X}_i$ in (3.11)–(3.14). Finally, we update D with $D^{(t+1)} = L^{(t+1)} \tilde{D}^{(t+1)} [L^{(t+1)}]^\top$ and complete the iteration by updating σ^2 with (3.9).

3.3 Maximum Likelihood Calculations

Computing the maximum likelihood estimate is similar to computing the posterior mode with an improper flat prior, (as described in Section 3.2) except we regard β as a model parameter and seek $\theta^\star$ that maximizes $\ell(\theta|Y_{\rm obs})$, with $\theta = (\sigma^2, T, \beta)$, rather than the mode of the marginal posterior (e.g., $\ell(\sigma^2, T|Y_{\rm obs})$). This simplifies calculations somewhat since β is regarded as a constant rather than a random variable in the E-step. In particular, we update L and $\tilde{D}$ to compute $D^{(t+1)}$ and then compute $\beta^{(t+1)}$ and $(\sigma^2)^{(t+1)}$ with two separate conditional maximizations. (In all formulas we replace $\tilde{X}_i$ with $\check{X}_i$ and set $\Sigma_L^{-1} = \Sigma_\beta^{-1} = T_0 = \sigma_0^2 = 0$, $\nu = -(p+2)$, and $\eta = -(q+1)$ and in (3.12) we fix $L = L_0 = I$.) We begin by computing $L^{(t+1)}$ using (3.11) with two simplifications. First $P(D)$ is replaced with $U(D)$ in the definition of $\hat{B}_i(\zeta)$ used to compute $\tilde{C}_i(\zeta^{(t)})$ (this change reflects the difference between (3.6) and (3.7)). Second, the expectation is computed conditional on $\theta^{(t)} = (\zeta^{(t)}, \beta^{(t)})$ using

$$\mathrm{E}[c_{ij} Z_{ik}^{\top} R_i^{-1}(Y_i - X_i\beta)|Y_{\rm obs}, \theta] = [\hat{b}_i(D,\beta)]_j Z_{ik}^{\top} R_i^{-1}(Y_i - X_i\beta).$$

We then compute $\tilde{D}^{(t+1)}$ using (3.8) again substituting $U(D)$ for $P(D)$ in the definition of $\hat{B}_i(\zeta)$ and set $D^{(t+1)} = L^{(t+1)}\tilde{D}^{(t+1)}(L^{(t+1)})^{\top}$. Next we maximize the observed data likelihood as a function of β with D fixed at $D^{(t+1)}$ and σ^2 fixed at $(\sigma^2)^{(t)}$, with $\beta^{(t+1)} = \hat{\beta}(D^{(t+1)})$. Likewise, in a final conditional maximization, we update σ^2 with

$$(\sigma^2)^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i\beta^{(t+1)}\right)^{\top} U_i\left(D^{(t+1)}\right)\left(Y_i - X_i\beta^{(t+1)}\right).$$

This completes the parameter-expanded ECME$_2$ iteration.

4. SIMULATIONS

In this section we illustrate the relative computational efficiency of ECME$_0$, ECME$_1$, ECME$_2$, and parameter-expanded ECME$_1$ and ECME$_2$. (Parameter expanded ECME$_1$ is analogous to parameter-expanded ECME$_2$, but with a lower triangular working parameter.) We fit a REML model to a number of datasets generated from the model

$$Y_i = X\beta + Z_i b_i + e_i, \quad \text{for} \quad i = 1, \ldots, 30, \tag{4.1}$$

where Y_i is 3×1 for each i, $X = (1,1,1)^{\top}$, $\beta = 1$, Z_i is a 3×3 matrix with elements generated as independent standard normals, $b_i \overset{\text{iid}}{\sim} N(0, \text{variance} = \text{diag}(1,4,9))$, and $e_i \overset{\text{iid}}{\sim} N(0, \sigma^2 I)$, with b_i and e_i independent.

As will be seen in the simulation results (see also Meng and van Dyk 1998 for theoretical arguments), the relative efficiency of ECME$_0$ and ECME$_1$ depends on the relative sizes of the fitted values of the variance of $Z_i b_i$ and the residual variance σ^2, which we quantify via a measure of the overall coefficient of determination,

$$\Delta^{\star} = \frac{\sum_{i=1}^{m} \text{tr}(Z_i T^{\star} Z_i^{\top})/m}{(\sigma^2)^{\star} + \sum_{i=1}^{m} \text{tr}(Z_i T^{\star} Z_i^{\top})/m}.$$

In order to vary $\Delta^{\star}$ in the simulations, 50 datasets were generated with each of several values of σ^2 (.25, 1, 4, 9, 16, 25, 36, 49, 64, and 81). For each of these 500 datasets,

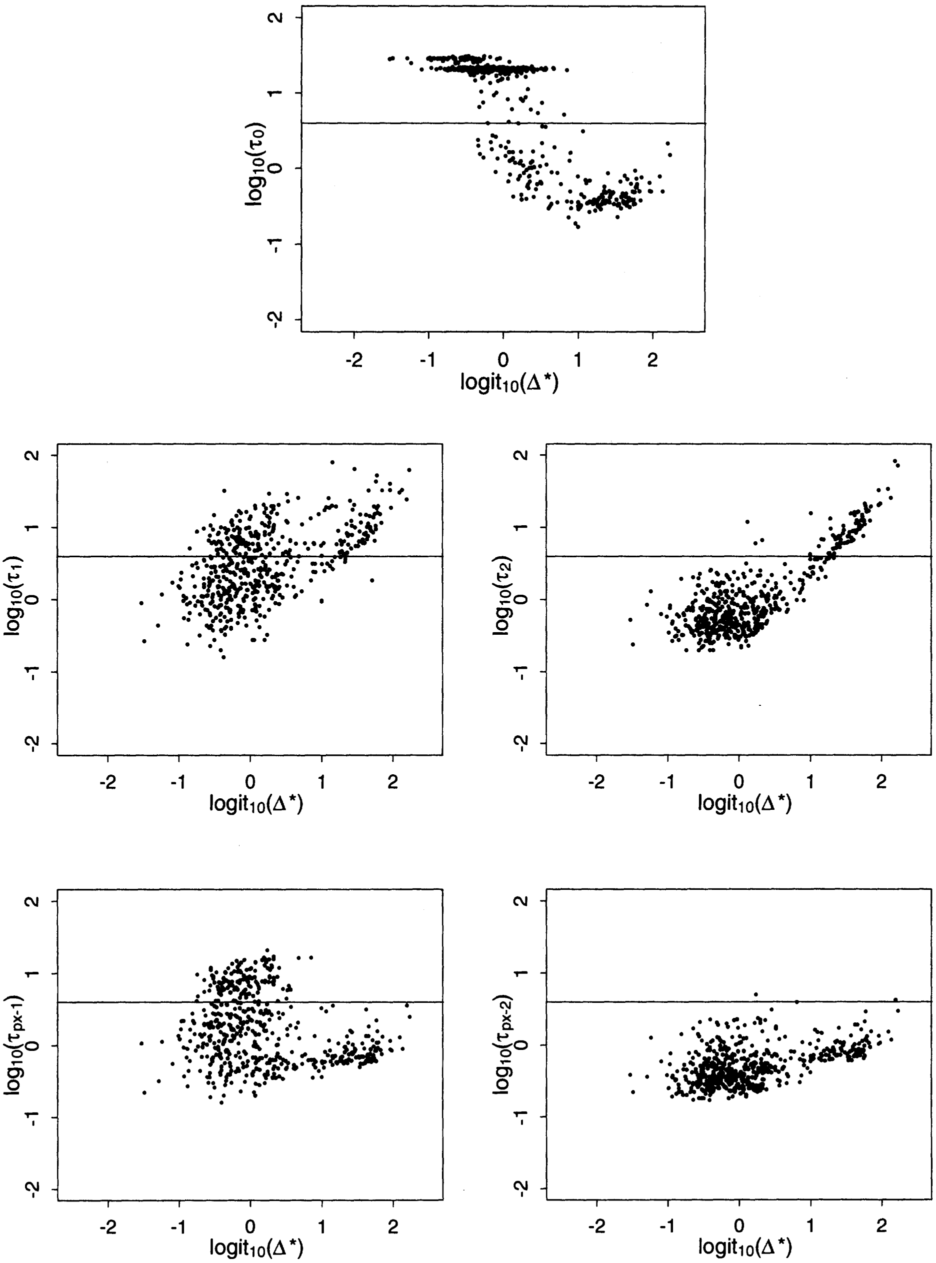

Figure 2. The time required by ECME$_i$ (τ_i), $i = 1, 2, 3$ and parameter-expanded ECME$_i$ ($\tau_{\text{px}-i}$), $i = 1, 2$, to fit (4.1) to each of the 500 simulated datasets. While ECME$_0$, ECME$_1$, and ECME$_2$ can be slow to converge for extreme values of Δ^*, the parameter-expanded ECME algorithms perform well for all values of Δ^* in the simulation. Increasing the dimension of the working parameter further improves computational performance (e.g., parameter-expanded ECME$_2$). The horizontal line in each plot represents the approximate time required by the `lme` *routine in S-Plus. It is clear that parameter-expanded ECME$_2$ performs well relative to* `lme`*, with superior convergence properties (e.g., monotone convergence in log posterior).*

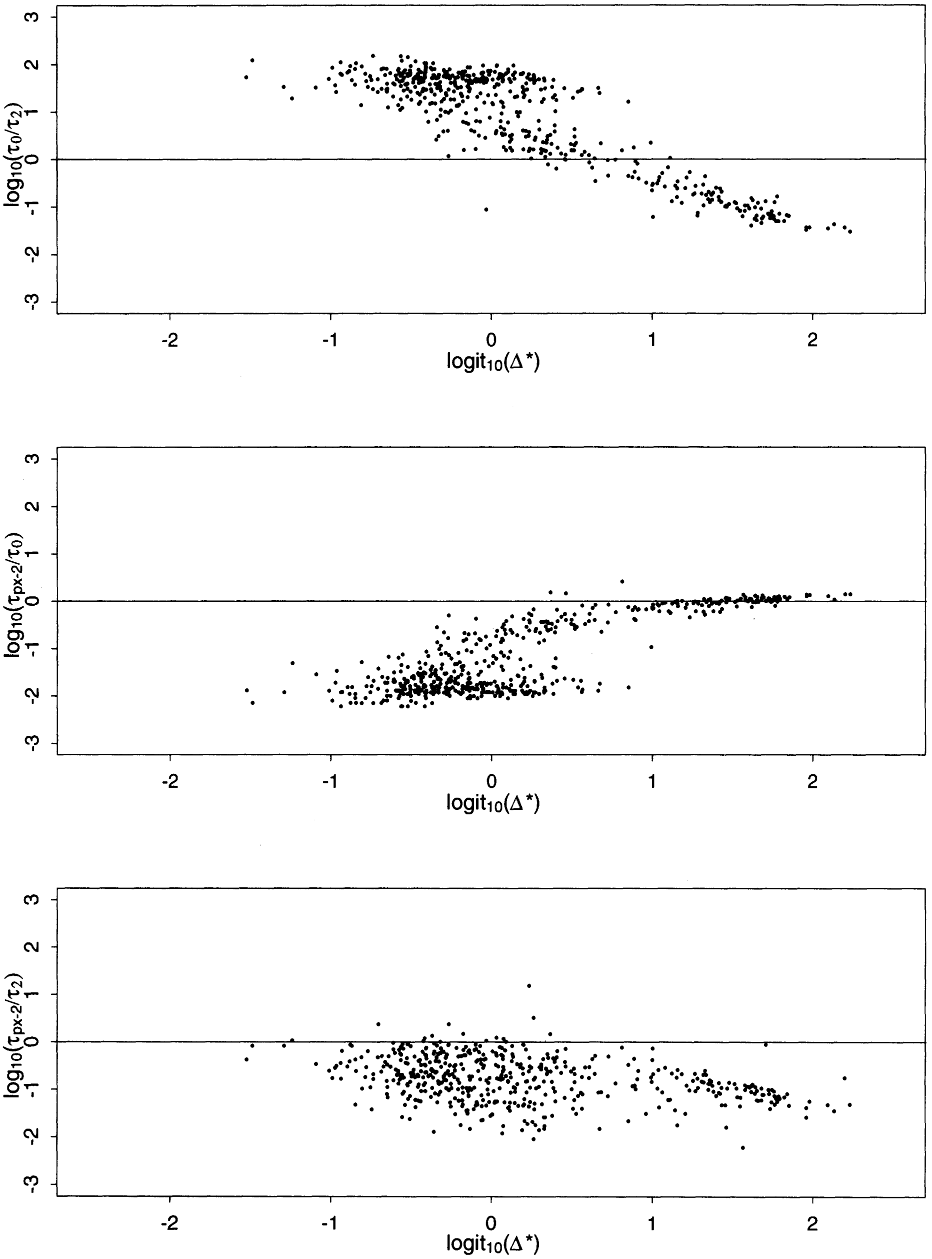

Figure 3. The relative time required by pairs of the three algorithms (ECME$_0$, ECME$_2$, and parameter-expanded ECME$_2$). The choice between ECME$_0$ and ECME$_2$ depends on Δ^* while parameter-expanded ECME$_2$ typically outperforms either algorithm.

REML estimates were computed using each of five algorithms, ECME$_i$, $i = 0, 1, 2$, and parameter-expanded ECME$_i$, $i = 1, 2$ using flat priors. The starting value $(\sigma^2)^{(0)}$ was obtained by fitting (4.1) ignoring the random effects, and $T^{(0)}$ was set to the identity matrix. Each algorithm was run until $\ell((\sigma^2)^{(t)}, T^{(t)}|Y_{\text{obs}}) - \ell((\sigma^2)^{(t-1)}, T^{(t-1)}|Y_{\text{obs}}) < 10^{-7}$.

The results of the simulation appear in Figures 2 and 3. Figure 2 contains five plots which record the time required for each of the five algorithms (τ_i for ECME$_i$, $i = 0, 1, 2$, and $\tau_{\text{px}-i}$ for parameter-expanded ECME$_i$, $i = 1, 2$, in seconds on the $\log_{10}$ scale). All computations were run on a Sun UltraSparc computer. Judging from the first plot in Figure 2, ECME$_0$ performs well when $\Delta^\star$ is large, but often required more than 20 seconds to converge when $\Delta^\star$ was relatively small. Conversely, ECME$_1$ and ECME$_2$ perform well when $\Delta^\star$ are small, but occasionally are very slow when $\Delta^\star$ is large. Finally, the parameter-expanded ECME algorithms converge quickly for all values of $\Delta^\star$ in the simulation. Nonetheless, increasing the dimension of the working parameter results in further advantage; compare parameter-expanded ECME$_1$ with parameter-expanded ECME$_2$. Parameter-expanded ECME$_2$ performs very well overall, requiring more than four seconds only twice and less than one second in 84% of the replications.

Figure 3 compares ECME$_0$, ECME$_2$ and parameter-expanded ECME$_2$ by recording the $\log_{10}$ relative time required by each pair of two algorithms (ECME$_2$ versus ECME$_0$, parameter-expanded ECME$_2$ versus ECME$_0$, and parameter-expanded ECME$_2$ versus ECME$_2$). Theory in Meng and van Dyk (1998) suggested ECME$_0$ is faster than ECME$_1$ only when $\Delta^* > 2/3$ or $\text{logit}_{10}(\Delta^*) > .30$ (approximately); the simulation verifies the same relationship between ECME$_0$ and ECME$_2$. It is also clear that ECME$_2$ can be much faster than ECME$_0$, while ECME$_0$ shows only small relative gains over ECME$_2$. Thus, we recommend using ECME$_2$ over ECME$_0$ (e.g., for fully Bayesian analysis) unless it is known a priori that Δ^* is large. [Meng and van Dyk (1998) discussed an adaptive strategy which approximates $\Delta^\star$ after several initial iterations to choose between $\alpha = 0$ and $\alpha = 1$; see also Section 5.] The final two plots show that parameter-expanded ECME$_2$ tends to be about as fast as the faster of ECME$_0$ and ECME$_2$. Thus, we recommend using parameter-expanded ECME$_2$ whenever a flat prior is used for T (e.g., REML and ML calculations).

The horizontal line in each plot of Figure 2 corresponds to four seconds, the approximate time required to fit this model using the `lme` routine in S-Plus on the same computer. The computation time required by parameter-expanded ECME$_2$ is generally less than four seconds, but with the additional important advantage of computational stability (e.g., monotone convergence in log-likelihood; see Section 5). Meng and van Dyk (1998) compared EM-type algorithms similar to ECME$_0$ and ECME$_1$ with the `lme` routine in S-Plus and the `xtreg` routine in STATA. Both comparisons were favorable for the EM-type algorithms in terms of both computational time and stability. In their investigations `lme` did not always converge to a mode of the log-likelihood. As a further comparison, we randomly selected one dataset generated with each of the ten values of σ^2 and fit the mixed-effects model with `lme`, parameter-expanded ECME$_2$, and the `proc mixed` routine in SAS. Table 1 displays the value of the restricted log-likelihood at the point of convergence for each algorithm applied to each of the ten datasets. Unfortunately, in six of the ten replications, `proc mixed` did not converge at all or converged to a value of

Table 1. The Value of the Restricted Loglikelihood at the Point of Convergence for Three Algorithms applied to Ten Datasets Selected Randomly from the Simulation. The SAS and S-Plus routines exhibited proper convergence in only 40% and 70% of these simulations respectively.

	Algorithm		
σ^2	*lme (S-Plus)*	*proc mixed (SAS)*	*PX-ECME$_2$*
.25	−184.11	−184.06	−184.06
1	−218.38	−218.39	−218.39
4	−227.98	‡	−227.98
9	−252.93	−252.93	−252.93
16	−269.75	‡	−269.75
25	−289.70†	−280.62	−280.62
36	−301.14	−301.03*	−301.14
49	−304.69	‡	−304.69
64	−321.41†	−319.85*	−320.64
81	−327.40†	‡	−325.44

† Did not converge to a mode.
‡ Did not converge, even after adjusting the `maxfunc` parameter to increase the maximum number of likelihood evaluations allowed.
* Converged to a point outside the parameter space.

$T^\star$ that was not positive semidefinite. The `lme` routine again did not always converge to a mode of the log-likelihood. (There is now a new version of the `lme` routine (`lme` 3.0) that may perform better. S-Plus 5.0 for Windows includes `lme` 3.0—however, it is not expected to be included in UNIX/Linux S-Plus until Release 6.0.) Parameter-expanded ECME$_2$ (and the other EM-type algorithms) exhibited no convergence difficulties.

5. DISCUSSION

The simulation results in Section 4 agree with theoretical arguments given elsewhere (e.g., Meng and van Dyk 1998). It is clear parameter-expanded ECME$_2$ is a general purpose, reliable, and efficient algorithm for REML and ML calculations with mixed-effects models. For Bayesian calculation with a proper prior on the random-effects variance, there is unfortunately no known efficient parameter-expanded algorithm. Thus, we recommend using either ECME$_0$ or ECME$_1$. The choice between these algorithms is based on the size of $\Delta^\star$, which can be approximated by replacing $(\sigma^2)^\star$ and $T^\star$ with a priori values, values computed with initial iterations from one of the ECME algorithms, or perhaps REML estimates computed using parameter-expanded ECME$_2$. This final strategy is especially attractive when the prior distributions are very diffuse since REML estimates should also serve as very good starting values for ECME$_0$ or ECME$_1$.

Based on the simulations, these algorithms are comparable to commercially available software in terms of the computation time required for convergence (see also Meng and van Dyk 1998) but with important advantages. The EM-type algorithms are guaranteed to increase $\ell(\theta|Y_{\text{obs}})$ at each iteration and to converge to estimates within the parameter space. Other methods (e.g., `lme` and `proc mixed`) require special monitoring and still may exhibit poor behavior. The `lme` routine, for example, was in general release for years before it was discovered that without special user intervention lme can converge to a point that is far from a mode. This was discovered by comparing lme with EM-type algorithms which converged correctly (Meng and van Dyk 1998). (This lme routine remains in the current version of S-Plus for UNIX/Linux.)

A third advantage of the EM-type algorithms is that they can be modified to handle more sophisticated models that cannot be fit using standard software. For example, missing values among $\{(Y_i, X_i, Z_i), i = 1, \ldots, m\}$ can be accommodated using a model for the missing data and a revised (perhaps Monte Carlo) E-step. Certain generalized linear mixed models can also be fit using efficient new data-augmentation methods (e.g., parameter expansion and nesting, see van Dyk 2000). For example, probit hierarchical regression models can be viewed as an extension of (3.1) in which we observe only the sign of each component of Y_i, $i = 1, \ldots, m$. The Y_i themselves are considered to be missing data. We can also extend the model to a hierarchical t model by replacing the normal distributions for either or both of e_i and b_i with t distributions. This is accomplished by writing the t variable as the scaled ratio of a normal variable and the square root of an independent chi-square variable and treating the chi-square variable as missing data (e.g, Dempster, Laird, and Rubin 1977). Thus, by finding very efficient EM-type algorithms for the mixed-effects model, we add an important building block for a variety of important models.

APPENDIX

There are a number of examples of Gaussian mixed-effects models being fit with improper priors despite the fact that the resulting posteriors are also improper (Gelfand, Hills, Racine-Poon, and Smith 1990; Geyer 1992; Wang, Rutledge, and Gianola 1993; and discussed in Hobert and Casella 1996, 1998). This is actually not surprising since it can be quite difficult to verify posterior propriety in a complicated model when the prior is not proper. Nonetheless, improper priors are often attractive since we may have little prior information or may be unable or unwilling to quantify what prior information we do have. Moreover, there are many who recommend improper priors as defaults, perhaps because improper priors can be "less influential" on inference than proper priors. Thus, it is especially important to determine precise sufficient conditions for posterior propriety in this context. In this Appendix, we extend Hobert and Casella's (1996) result for the special case where the random effects are assumed to be a priori independent to the more general case of correlated random effects.

We consider the Gaussian mixed-effects model given by (3.1) using the parameterization $(\sigma^2, \beta, T = \sigma^2 D)$. We may rewrite the model as

$$Y = X\beta + Zb + e, \quad \text{with} \quad b \sim N(0, V) \quad \text{and} \quad e_i \sim N(0, I_n \sigma^2),$$

where $Y^\top = (Y_1^\top, \ldots, Y_m^\top)$, $n = \sum_{i=1}^m n_i$, $X^\top = (X_1^\top, \ldots, X_m^\top)$, $b^\top = (b_1^\top, \ldots, b_m^\top)$, $Z = \text{diag}(Z_1, \ldots, Z_m)$, and $V = \text{diag}(T, \ldots, T)$. We consider an improper prior of the form, $p(\beta, \sigma^2, T) \propto (\sigma^2)^{-(\nu/2+1)} |T|^{-(\eta+q+1)/2}$. In this case, we have the following result. (We emphasize that the conditions of the theorem are sufficient but may be stronger than necessary.)

Theorem 1. *Let* $P_X = (I - X(X^\top X)^{-1})X^\top)$ *and suppose* $r \equiv qm = \mathit{rank}(P_X Z)$, *then the following are sufficient for propriety of the posterior: (a)* $\eta < 1 - q$, *(b)* $m > q - 1 - \eta$, *and (c)* $n > p - q\eta - \nu$.

Proof: We wish to show that $\int p(\theta, b|Y)dbd\theta < \infty$, where $p(\theta, b|Y) \propto p(Y|b,\theta)$ $p(b|T)p(\theta)$, with $\theta = (\beta, \sigma^2, T)$. Standard REML calculations show (Harville 1977; Laird and Ware 1982; Laird, Lang, and Stram 1987; and Lindstrom and Bates 1988) that

$$\int p(\theta, b|Y)dbd\beta$$
$$= k_0 \frac{\exp\{-\frac{1}{2}Y^\top[M_1(V)XM_2(V)X^\top M_1(V) - M_1(V)]Y\}}{(\sigma^2)^{(n-r-p)/2}|V|^{1/2}|\sigma^2 V^{-1} + Z^\top P_X Z|^{1/2}} p(\sigma^2, T), \quad \text{(A.1)}$$

where $M_1(V) = (ZVZ^\top + I\sigma^2)^{-1}$, $M_2(V) = (X^\top M_1(V)X)^{-1}$, and k_0 does not depend on T or σ^2. We bound the numerator and the denominator of (A.1) as functions of V. We start with the exponential term in the numerator and use the fact (proven by Hobert and Casella 1996) that if V_0 is diagonal,

$$\begin{aligned} f(V_0, Z) &\equiv \exp\{-\frac{1}{2}Y^\top[M_1(V_0)XM_2(V_0)X^\top M_1(V_0) - M_1(V_0)]Y\} \\ &\leq \exp\{-\frac{1}{2\sigma^2}Y^\top[P_Z - P_Z X(X^\top P_Z X)^{-1}X^\top P_Z]Y\}, \end{aligned} \quad \text{(A.2)}$$

where $P_Z = I - Z(Z^\top Z)^{-1}Z^\top$. Writing $V = QV_0Q^\top$, where V_0 is a diagonal matrix with diagonal elements equal to the eigenvalues of V and Q is an invertible matrix, we have $f(V, Z) = f(V_0, \tilde{Z})$ with $\tilde{Z} = ZQ$. Since $P_Z = P_{\tilde{Z}}$, we have that the exponential factor in (A.1) is bounded above by (A.2).

In order to bound the denominator of (A.1), we note (using an inequality due to Fiedler 1971; see, e.g., Marshall and Olkin 1979, G.3.a)

$$|\sigma^2 V^{-1} + Z^\top P_X Z| \geq \prod_{i=1}^{q}\left[\frac{\sigma^2}{\tau_i} + \lambda_{\text{min}}\right]^m,$$

where $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_q$ are the ordered eigenvalues of T and λ_{min} is the smallest eigenvalue of $Z^\top P_X Z$. Using the two bounds we obtain, by averaging (A.1) over T,

$$\int p(\theta, b|Y)dbd\beta dT$$
$$\leq \frac{k_0 p(\sigma^2)\exp\{-\frac{1}{2\sigma^2}Y^\top[P_Z - P_Z X(X^\top P_Z X)^{-1}X^\top P_Z]Y\}}{(\sigma^2)^{(n-r-p)/2}}$$
$$\times \int \frac{p(T)}{|V|^{1/2}\prod_{i=1}^{q}\left[\frac{\sigma^2}{\tau_i} + \lambda_{\text{min}}\right]^{m/2}} dT. \quad \text{(A.3)}$$

To evaluate the integral on the right hand side of (A.3) we use Theorem 2 of Hsu (1938) to accomplish a change of variable, namely,

$$\begin{aligned}
&\int \frac{|T|^{-(\eta+q+1)/2}}{|V|^{1/2}\prod_{i=1}^{q}\left[\frac{\sigma^2}{\tau_i}+\lambda_{\min}\right]^{m/2}}dT \\
&= k_1 \int \frac{\left(\prod_{i=1}^{q}\prod_{j=i+1}^{q}(\tau_i-\tau_j)\right)|T|^{-(\eta+q+1)/2}}{|V|^{1/2}\prod_{i=1}^{q}\left[\frac{\sigma^2}{\tau_i}+\lambda_{\min}\right]^{m/2}}d\tau_1\cdots d\tau_q \\
&\leq k_1 \prod_{i=1}^{q}\int \frac{\tau_i^{-(\eta-q+1+2i)/2}}{(\sigma^2+\tau_i\lambda_{\min})^{m/2}}d\tau_i, \qquad \text{(A.4)}
\end{aligned}$$

where k_1 does not depend on T or σ^2. Equation (A.4) is integrable if both $(\eta+2i-q-1)/2<0$ and $m>-(\eta+2i-q-1)$ for $i=1,\ldots,q$. This is satisfied if conditions (a) and (b) of the theorem hold, in which case (A.4) is proportional to $(\sigma^2)^{-q(\eta+m)/2}$ as a function of σ^2. Substituting this last expression into (A.3), we see that (A.3) is an integrable function of σ^2 if (c) holds. This completes the proof. □

ACKNOWLEDGMENTS

The author's research was supported in part by NSF grant DMS 97-05157 and in part by the U.S. Census Bureau. The author thanks the referees and editors of the *Journal of Computational and Graphical Statistics,* for their helpful comments and J. L. Foulley for several suggestions that have greatly improved the article, in particular the use of an arbitrary invertible working parameter matrix.

[Received August 1998. Revised June 1999.]

REFERENCES

Beaton, A. E. (1964), "The Use of Special Matrix Operations in Statistical Calculus," *Education Testing Service Research Bulletin*, RB-64-51.

Callanan, T. P., and Harville, D. A. (1991), "Some New Algorithms for Computing Restricted Maximum Likelihood Estimates of Variance Components," *Journal of Statistical Computation and Simulation*, 38, 239–259.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation from Incomplete-Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

Dempster, A. P., Selwyn, M. R., Patel C. M., and Roth, A. J. (1984), "Statistical and Computational Aspects of Mixed Model Analysis," *Applied Statistics*, 33, 203–214.

Fessler, J. A., and Hero, A. O. (1994), "Space-Alternating Generalized Expectation-Maximization Algorithm," *IEEE Transactions on Signal Processing*, 42, 2664–2677.

——— (1995), "Penalized Maximum-Likelihood Image Reconstruction Using Space-Alternating Generalized EM Algorithm," *IEEE Transactions on Image Processing*, 4, 1417–1438.

Fiedler, M. (1971), "Bounds for the Determinant of the Sum of Hermitian Matrices," in *Proceedings of the American Mathematical Society*, 30, pp. 27–31.

Foulley, J.-L., and Quaas, R. L. (1995), "Heterogeneous Variance in Gaussian Linear Mixed Models," *Genetics Selection Evolution*, 27, 211–228.

Foulley, J.-L., and van Dyk, D. A. (in press), "The PX-EM Algorithm for Fast Stable Fitting of Henderson's Mixed Model," *Genetics, Selection, Evolution.*

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association*, 85, 972–985.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London: Chapman and Hall.

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo," *Statistical Science*, 7, 473–511.

Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72, 320–240.

Hobert, J. P., and Casella, G. (1996), "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models," *Journal of the American Statistical Association*, 91, 1461–1473.

——— (1998), "Functional Compatibility, Markov Chains, and Gibbs Sampling With Improper Posteriors," *Journal of Computational and Graphical Statistics*, 7, 42–60.

Hsu, P. L. (1938), "On the Distribution of Roots of Certain Determinantal Equations," *The Annals of Eugenics*, 8, 376–386.

Laird, N. M. (1982), "Computation of Variance Components Using The E-M Algorithm," *Journal of Statistical Computations and Simulation*, 14, 295–303.

Laird, N., Lange, N., and Stram, D. (1987), "Maximizing Likelihood Computations With Repeated Measures: Application of the EM Algorithm," *Journal of the American Statistical Association*, 82, 97–105.

Laird, N. M., and Ware, J. H. (1982), "Random Effects Models for Longitudinal Data," *Biometrics*, 38, 967–974.

Lindstrom, M. J., and Bates, D. M. (1988), "Newton–Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measure Data," *Journal of the American Statistical Association*, 83, 1014–1022.

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

Liu, C., and Rubin, D. B. (1994), "The ECME Algorithm: a Simple Extension of EM and ECM With Fast Monotone Convergence," *Biometrika*, 81, 633–648.

——— (1995), "ML Estimation of the T-Distribution Using EM and its Extensions, ECM and ECME," *Statistica Sinica*, 5, 19–40.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter Expansion for EM Acceleration—the PX-EM Algorithm," *Biometrika*, 85, 755–770.

Liu. J. S., and Wu, Y. N. (1999), "Parameter Expansion Scheme for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

Marshall, A. W., and Olkin, I. (1979), *Inequalities: Theory of Majorization and Its Applications*, New York: Academic Press.

Meng, X.-L., and Pedlow, S. (1992), "EM: A Bibliographic Review With Missing Articles," in *Proceedings of the Statistical Computation Section*, Alexandria, VA: American Statistical Association, 24–27.

Meng, X.-L., and Rubin, D. B. (1994), "On the Global and Component-Wise Rates of Convergence of the EM Algorithm," *Linear Algebra and its Applications (Special Issue Honoring Ingram Olkin)*, 199, 413–425.

Meng, X.-L., and van Dyk, D. A. (1997), "The EM Algorithm—an Old Folk Song Sung to a Fast New Tune" (with discussion), *Journal of the Royal Statistical Society*, Series B, 59, 511–567.

——— (1998), "Fast EM Implementations for Mixed-Effects Models," *Journal of the Royal Statistical Society*, Series B, 60, 559–578.

——— (1999), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation," *Biometrika*, 86, 301–320.

Schafer, J. L. (1998), "Some Improved Procedures for Linear Mixed Models," Technical Report 98-28, The Methodological Center, The Pennsylvania State University, available at http://methcenter.psu.edu/techrpts.html.

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, 805–811.

Thompson, R., and Meyer, K. (1986), "Estimation of Variance Components: What is Missing in the EM

Algorithm?" *Journal of Statistical Computation and Simulation*, 24, 215–230.

van Dyk, D. A. (2000), "Fast New EM-type Algorithms with Applications in Astrophysics," unpublished manuscript.

——— (2000), "Nesting EM Algorithms for Computational Efficiency", *Statistica Sinica*, 10, 203–225.

van Dyk, D. A., and Meng, X.-L. (in press), "The Art of Data Augmentation" (with discussion), *Journal of Computational and Graphical Statistics*.

Wang, C. S., Rutledge, J. J., and Gianola, D. (1993), "Marginal Inference About Variance Components in a Mixed Linear Model Using Gibbs Sampling," *Genetics Selection Evolution*, 25, 41–62.

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