Some improved proedures for linear mixed models

Joseph L. Schafer

Contents

1 Maintainer's note

The reference should read[1]

Schafer, J.L. (1998) Some improved proedures for linear mixed models. Dept. of Statistics, The Pennsylvania State University

The marjiuana data as in Table 1 and in the package is reproduced as follows,

	15 minutes			90 minutes		
Subject	Placebo	Low	High	Placebo	Low	High
1	16	20	16	20	-6	-4
2	12	24	12	-6	4	-8
3	8	8	26	-4	4	8
4	20	8	-	-	20	-4
5	8	4	-8	-	22	-8
6	10	20	28	-20	-4	-4
7	4	28	24	12	8	18
8	-8	20	24	-3	8	-24
9	-	20	24	8	12	-

2 Technical Report

The technical report starts from next page.

[1]The technical report was originally obtained from `http://www.stat.psu.edu/~jls/misoftwa.html`

Some improved procedures for linear mixed models

Joseph L. Schafer *

January 12, 1998

Linear mixed-effects models are widely used in the analysis of longitudinal and clustered data. This article presents new expressions for the derivatives of loglikelihood functions for these models with respect to unknown components of variance. These expressions are easy to evaluate and involve the same quantities needed for EM algorithms for maximum-likelihood (ML) and restricted maximum-likelihood (RML) estimation. Three applications for these derivative expressions are developed: (a) new hybrid algorithms for ML and RML estimation that combine the stability and low per-iteration cost of EM with the rapid convergence of Fisher scoring; (b) a new analytic method for correcting traditional empirical Bayes interval estimates for the uncertainty associated with variance components; and (c) a new Markov chain Monte Carlo algorithm for Bayesian posterior simulation that has the low per-iteration cost of a conventional Gibbs sampler but converges rapidly.

Key Words: EM algorithm, Gibbs sampling, mixed-effects model, Metropolis-Hastings algorithm, restricted maximum-likelihood, variance components

*Assistant Professor, Department of Statistics, The Pennsylvania State University, University Park, PA 16802-6202. This research was supported by grant 2R44CA65147-02 from National Institutes of Health, and by grant 1-P50-DA10075-01 from the National Institute on Drug Abuse. Portions of this manuscript were completed while the author was in residence at the Bureau of Labor Statistics as a 1997–98 ASA/NSF/BLS Senior Research Fellow. Thanks to Recai Yucel who participated in the early stages of this research.

1 Introduction

Let y_i be a vector of n_i measurements for sample unit i, $i = 1, \ldots, m$. I assume that y_i follows the general linear mixed model

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \tag{1}$$

where X_i $(n_i \times p)$ and Z_i $(n_i \times q)$ are known covariate matrices, and b_i and ε_i are random errors distributed as

$$b_i \sim N_q(0, \psi), \tag{2}$$

$$\varepsilon_i \sim N_{n_i}(0, \sigma^2 V_i), \tag{3}$$

independently for $i = 1, \ldots, m$. Models of this form were proposed by Hartley and Rao (1967) and became practically useful due to the computational work of Laird and Ware (1982); Jennrich and Schluchter (1986); Laird, Lange, and Stram (1987); Lindstrom and Bates (1988); and others. These models are often applied in longitudinal settings, where y_i represents repeated measurements of a variable over time, and the measurement times are incorporated into X_i and Z_i. Because no particular form is assumed for X_i or Z_i, the model accommodates time-varying covariates, unequally-spaced responses, and incomplete data where some responses are missing for some units. Other applications involve multilevel or clustered data, where y_i represents measurements for subunits nested within unit i (e.g. students within a classroom). Algorithms for fitting these models have been implemented in SAS (Littell *et al.*, 1996), S-PLUS (MathSoft, Inc., 1997), MLn (Multilevel Models Project, 1996) and HLM (Bryk, Raudenbush, and Congdon, 1996).

Depending on the context, various assumptions are made about the covariance matrices ψ and $\sigma^2 V_i$. With clustered data, each V_i $(n_i \times n_i)$ is typically required to be an identity matrix, reflecting an assumption that subunits within clusters have exchangeable errors. In longitudinal applications, exchangeable models are common, as are patterned (e.g. banded or first-order autoregressive) forms where the V_i depend on a small number of free parameters. Often no restrictions are placed on ψ other than positive definiteness ($\psi > 0$). Models with block-diagonal ψ are also popular, because with suitable Z_i this allows units to be grouped into subsamples having different between-unit covariance matrices (e.g. Laird, Lange, and Stram,

1987). For the remainder of this article I assume that $V_1, \ldots, V_m$ are known. Extensions where V_i contains unknown parameters are not difficult and may be addressed in future work. In addition, I will apply an assumption of linearity to the between-unit precision matrix. This linear form is crucial and cannot be done away with easily. The assumption is relatively benign, however, because it covers the usual situations where ψ is unstructured or block-diagonal.

Depending on the context, one may want to draw inferences about the coefficients β common to all units (the "fixed effects"), the unit-specific coefficients b_i (the "random effects"), or the variance components (σ^2, ψ). The dual errors (2) and (3) allow the model to be expressed as a linear regression with patterned covariance,

$$y_i \sim N(X_i\beta, \sigma^2 W_i^{-1}), \tag{4}$$

where $W_i = (Z_i \xi Z_i^T + V_i)^{-1}$ and $\xi = \sigma^{-2}\psi$. Maximum-likelihood (ML) estimates of β, σ^2, and ξ are obtained by maximizing the likelihood function arising from (4),

$$L_0(\beta, \sigma^2, \xi) \propto (\sigma^2)^{-\frac{N}{2}} \prod_{i=1}^{m} |W_i|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - X_i\beta)^T W_i (y_i - X_i\beta)\right\}, \tag{5}$$

where $N = \sum_{i=1}^{m} n_i$. For the variance components, some prefer restricted maximum-likelihood (RML) estimates, which are obtained by maximizing a function equal to the indefinite integral of (5) over β,

$$L_1(\sigma^2, \xi) \propto (\sigma^2)^{-\frac{(N-p)}{2}} \left|\sum_{i=1}^{m} X_i^T W_i X_i\right|^{-\frac{1}{2}} \prod_{i=1}^{m} |W_i|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - X_i\tilde{\beta})^T W_i (y_i - X_i\tilde{\beta})\right\}, \tag{6}$$

where

$$\tilde{\beta} = \left(\sum_{i=1}^{m} X_i^T W_i X_i\right)^{-1} \left(\sum_{i=1}^{m} X_i^T W_i y_i\right) \tag{7}$$

is the generalized least-squares estimate of β implied by (σ^2, ξ). Notice that W_i and $\tilde{\beta}$ are functions of ξ; for notational simplicity, however, the dependence upon ξ is suppressed. ML and RML estimates are discussed from a frequentist perspective by Harville (1977) and from a Bayesian perspective by Dempster, Rubin, and Tsutakawa (1981).

Two approaches are commonly used to maximize L_0 or L_1. Methods based on the EM algorithm are given by Laird and Ware (1982); Jennrich and Schluchter (1986); Laird, Lange, and Stram (1987); and Liu and Rubin (1995). Newton-Raphson and Fisher scoring are discussed by Jennrich and Schluchter (1986)

and Lindstrom and Bates (1988). EM is easy to implement and stable but can be very slow to converge. Progress in speeding EM has recently been made by Meng and van Dyk (1997). Newton-Raphson and scoring typically converge in just a few iterations, but current versions are complicated and have a high per-iteration cost. Moreover, these algorithms require careful implementation because they lack some of the stability of EM; the loglikelihood might not increase at each iteration, estimates might leave the parameter space, and the algorithms may fail if the loglikelihood function is oddly shaped (e.g. non-concave or having maxima on the boundary).

Inferences about random effects are based on the following results which I state without proof. Conditionally upon $y = (y_1, y_2, \ldots, y_m)$ and (β, σ^2, ξ), Bayes's Theorem implies that $b_1, \ldots, b_m$ are independent and normal with posterior moments $E(b_i \mid y, \beta, \sigma^2, \xi) = \hat{b}_i$ and $V(b_i \mid y, \beta, \sigma^2, \xi) = \sigma^2 U_i$, where

$$\hat{b}_i = U_i Z_i^T V_i^{-1}(y_i - X_i\beta), \tag{8}$$

$$U_i = (\xi^{-1} + Z_i^T V_i^{-1} Z_i)^{-1}. \tag{9}$$

Empirical Bayes (EB) point and interval estimates for b_i may be obtained by substituting estimates for (β, σ^2, ξ) into these expressions. Alternatively, if we apply an improper uniform prior density to β, the posterior distribution for β given (σ^2, ξ) is normal with mean $E(\beta \mid y, \sigma^2, \xi) = \tilde{\beta}$ and variance $V(\beta \mid y, \sigma^2, \xi) = \sigma^2\Gamma$, where

$$\Gamma = \left(\sum_{i=1}^{m} X_i^T W_i X_i\right)^{-1}. \tag{10}$$

It follows that the moments for b_i without regard for β are $E(b_i \mid y, \sigma^2, \xi) = \tilde{b}_i$ and $V(b_i \mid y, \sigma^2, \xi) = \sigma^2(U_i + A_i)$, where

$$\tilde{b}_i = U_i Z_i^T V_i^{-1}(y_i - X_i\tilde{\beta}), \tag{11}$$

$$A_i = U_i Z_i^T V_i^{-1} X_i \Gamma X_i^T V_i^{-1} Z_i U_i. \tag{12}$$

Substituting estimates of σ^2 and ξ into (11)–(12) provides another source for EB intervals. This latter method seems preferable because it accounts for uncertainty in estimating β. Neither method, however, acknowledges uncertainty about σ^2 or ξ. As a result, both types of intervals will be artificially precise, tending to have frequentist coverage less than their nominal levels in small to moderate samples.

As an alternative to EB, some prefer a fully Bayesian approach in which knowledge about the unknown parameters is summarized by a posterior distribution. Under a uniform prior for β, the posterior density is

$$P(\beta, \sigma^2, \xi \,|\, y) \;\propto\; L_0(\beta, \sigma^2, \xi)\, \pi(\sigma^2, \xi), \tag{13}$$

where $\pi(\sigma^2, \xi)$ is the prior density function for the variance components. The marginal posterior for the variance components alone can be written as

$$P(\sigma^2, \xi \,|\, y) \;\propto\; L_1(\sigma^2, \xi)\, \pi(\sigma^2, \xi), \tag{14}$$

and Bayesian inferences about b_i are obtained by averaging its conditional posterior

$$b_i \mid y, \sigma^2, \xi \;\sim\; N(\,\tilde{b}_i,\, \sigma^2(U_i + A_i)\,)$$

over (14). These inferences about b_i, and summaries of $P(\sigma^2, \xi \mid y)$ itself, have been difficult to obtain analytically or by traditional numerical methods. It is possible, however, to simulate random draws of θ from $P(\theta \,|\, y)$ using recently developed techniques of Markov chain Monte Carlo (MCMC).

In MCMC, one generates a random sequence of dependent parameter values whose distribution converges to the desired posterior. Applications of MCMC to linear mixed models have been made by Gelfand *et al.* (1990); Zeger and Karim (1991); Liu and Rubin (1995); and Carlin (1996). These applications are similar to EM algorithms in that they rely on simplifications that result if the random effects b_i are assumed known. Augmenting $y_1, \ldots, y_m$ by simulated values of $b_1, \ldots, b_m$ leads to algorithms with an attractive simplicity, but which may require many cycles to converge. Slowly converging MCMC algorithms are troublesome not only because of their computational demands, but because convergence can be notoriously difficult to detect (e.g. Gelman and Rubin, 1992).

In this article, I derive analytic methods for assessing the uncertainty due to the variance components in the general linear mixed model. Section 2 presents new expressions for the first and second derivatives of the logarithms of L_0 and L_1. These expressions are noteworthy for their simplicity and close relationship to quantities appearing in EM algorithms. In Section 3, I use these derivatives to develop new ML and RML algorithms that combine the stability of EM with the rapid convergence of Fisher scoring. When the

dimension of ξ $(q \times q)$ is small, the computational cost per iteration of these new algorithms is similar to that of EM. Section 4 presents an analytic method for incorporating variance-component uncertainty into EB intervals. Unlike MCMC methods, the method does not use simulation or prior distributions for σ^2 and ξ, and it appears to perform well even when sample sizes are quite small. A final application, presented in Section 5, is a new MCMC algorithm for Bayesian inference which has the stability and low per-iteration cost of current Gibbs samplers but converges rapidly.

2 Derivative expressions

Previously, Jennrich and Schluchter (1986) and Lindstrom and Bates (1988) have presented derivative-based methods for maximizing L_0 and L_1. Their methods apply not only to the linear mixed model but to other (e.g. factor analytic) covariance patterns as well. Their derivative formulas are quite general but can be tedious to compute, particularly as the within-unit sample sizes n_i grow. My expressions pertain only to model (1) but are much easier to evaluate. The crucial computations (e.g. inversions) involve matrices of dimension $q \times q$ rather than $n_i \times n_i$.

The only additional requirement for my method is that the inverse of ξ must be a linear function of free parameters. I assume that

$$\xi^{-1} = \sum_{j=1}^{g} \omega_j G_j, \tag{15}$$

where $\omega = (\omega_1, \omega_2, \ldots, \omega_g)^T$ is a vector of unknown covariance parameters and $G_1, \ldots, G_g$ are known symmetric matrices of dimension $q \times q$. The unstructured form for ξ results from taking each G_j to be a matrix containing ones in positions (k, l) and (l, k) and zeros elsewhere, in which case $g = q(q+1)/2$. Clearly (15) also covers any situation where ξ is block-diagonal with unstructured blocks, or where any block is known up to a constant of proportionality.

The expressions below follow the convention that if f is a scalar-valued function and X is an $a \times b$ matrix with typical element x_{ij}, then $\partial f / \partial X$ is the $a \times b$ matrix with corresponding element $\partial f / \partial x_{ij}$. Similarly, the derivative of X with respect to a scalar c is the matrix with elements $\partial x_{ij} / \partial c$. If $x = (x_1, \ldots, x_a)^T$ and $y = (y_1, \ldots, y_b)^T$, then $\partial^2 f / \partial x \partial y^T$ is the $(a \times b)$ matrix with (i, j)th element $\partial^2 f / \partial x_i \partial y_j$. Finally,

if $F(x) = (f_1(y), \ldots, f_a(y))^T$ is a vector-valued function ($a > 1$) of $y = (y_1, \ldots, y_b)^T$, then $\partial F/\partial y$ is the $(a \times b)$ matrix with (i,j)th element $\partial f_i/\partial y_j$.

For convenience, derivatives of the loglikelihood will be taken with respect to the free parameters β, ω, and $\tau = \sigma^{-2}$. Using the relationship $|W_i| = |V_i|^{-1} |\xi|^{-1} |U_i|$, and ignoring constants of proportionality, the logarithm of L_0 may be written as

$$l_0 = \frac{N}{2} \log \tau - \frac{m}{2} \log |\xi| + \frac{1}{2} \sum_{i=1}^{m} \log |U_i| - \frac{\tau}{2} \sum_{i=1}^{m} (y_i - X_i\beta)^T W_i (y_i - X_i\beta).$$

By standard techniques of matrix differentiation (e.g. Schott, 1997) we have $\partial \xi^{-1}/\partial \omega_j = G_j$, $\partial U_i/\partial \omega_j = -U_i G_j U_i$, and

$$\frac{\partial W_i}{\partial \omega_j} = V_i^{-1} Z_i U_i G_j U_i Z_i^T V_i^{-1},$$

which follows from $W_i = V_i^{-1} - V_i^{-1} Z_i U_i Z_i^T V_i^{-1}$. Applying these rules and the identity

$$\sum_{i=1}^{m} (y_i - X_i\beta)^T W_i (y_i - X_i\beta) = \sum_{i=1}^{m} (y_i - X_i\tilde{\beta})^T W_i (y_i - X_i\tilde{\beta}) + (\beta - \tilde{\beta})^T \Gamma^{-1} (\beta - \tilde{\beta}),$$

it can be shown that the first derivatives are $\partial l_0/\partial \beta = -\sigma^{-2}\Gamma^{-1}(\beta - \tilde{\beta})$,

$$\begin{aligned} \frac{\partial l_0}{\partial \omega_j} &= \frac{1}{2} \sum_{i=1}^{m} \operatorname{tr} \left(\xi - U_i - \sigma^{-2} \hat{b}_i \hat{b}_i^T \right) G_j, \\ \frac{\partial l_0}{\partial \tau} &= \frac{N}{2} \sigma^2 - \frac{1}{2} \sum_{i=1}^{m} (y_i - X_i\beta)^T W_i (y_i - X_i\beta), \end{aligned}$$

and the second derivatives are $\partial^2 l_0/\partial \tau^2 = -N\sigma^4/2$, $\partial^2 l_0/\partial \beta \partial \beta^T = -\sigma^{-2}\Gamma^{-1}$,

$$\frac{\partial^2 l_0}{\partial \tau \partial \omega_j} = -\frac{1}{2} \sum_{i=1}^{m} \operatorname{tr} \left(\hat{b}_i \hat{b}_i^T \right) G_j, \tag{16}$$

$$\frac{\partial^2 l_0}{\partial \beta \partial \tau} = -\Gamma^{-1}(\beta - \tilde{\beta}), \tag{17}$$

$$\frac{\partial^2 l_0}{\partial \beta \partial \omega_j} = -\sigma^{-2} \left(\sum_{i=1}^{m} \gamma_i^T U_i G_j U_i \gamma_i \right) (\beta - \tilde{\beta}), \tag{18}$$

$$\frac{\partial^2 l_0}{\partial \omega_j \partial \omega_k} = -\frac{m}{2} \operatorname{tr} \xi G_j \xi G_k + \frac{1}{2} \sum_{i=1}^{m} \operatorname{tr} U_i G_j U_i G_k + \sigma^{-2} \sum_{i=1}^{m} \operatorname{tr} \left(\hat{b}_i \hat{b}_i^T \right) G_j U_i G_k, \tag{19}$$

where $\gamma_i = Z_i^T V_i^{-1} X_i$.

The scoring algorithms of Section 3 will require us to calculate expectations of the second derivatives with respect to the distribution of y for fixed parameters. The expectations of (17) and (18) are zero

because $\tilde{\beta}$ is an unbiased estimate of β. Also, (4) and (8) imply that $E(\hat{b}_i) = 0$ and $E(\hat{b}_i\hat{b}_i^T) = \sigma^2(\xi - U_i)$, from which it can be shown that

$$\begin{aligned} E\left(\frac{\partial^2 l_0}{\partial\tau\partial\omega_j}\right) &= -\frac{\sigma^2}{2}\sum_{i=1}^{m}\mathrm{tr}(\xi - U_i)G_j, \\ E\left(\frac{\partial^2 l_0}{\partial\omega_j\partial\omega_k}\right) &= -\frac{1}{2}\sum_{i=1}^{m}\mathrm{tr}\,(\xi - U_i)G_j(\xi - U_i)G_k. \end{aligned}$$

These expressions can be evaluated quickly by taking into account the sparseness of $G_1, \ldots, G_g$.

The logarithm of L_1, which can be written as

$$l_1 = \frac{(N-p)}{2}\log\tau - \frac{m}{2}\log|\xi| + \frac{1}{2}\sum_{i=1}^{m}\log|U_i| + \frac{1}{2}\log|\Gamma| - \frac{\tau}{2}\sum_{i=1}^{m}(y_i - X_i\tilde{\beta})^T W_i(y_i - X_i\tilde{\beta}),$$

is more tedious to differentiate because Γ and $\tilde{\beta}$ are complicated functions of ξ. The first derivatives reduce to

$$\begin{aligned} \frac{\partial l_1}{\partial\tau} &= \frac{(N-p)}{2}\sigma^2 - \frac{1}{2}\sum_{i=1}^{m}(y_i - X_i\tilde{\beta})^T W_i(y_i - X_i\tilde{\beta}), \\ \frac{\partial l_1}{\partial\omega_j} &= \frac{1}{2}\sum_{i=1}^{m}\mathrm{tr}\left(\xi - U_i - A_i - \sigma^{-2}\tilde{b}_i\tilde{b}_i^T\right)G_j, \end{aligned}$$

where $\tilde{b}_i = U_i Z_i^T V_i^{-1}(y_i - X_i\tilde{\beta})$ and $A_i = U_i\gamma_i\,\Gamma\,\gamma_i^T U_i$. The second derivative with respect to τ is $\partial^2 l_1/\partial\tau^2 = -(N-p)\sigma^4/2$, and the cross-derivative is $\partial^2 l_1/\partial\tau\partial\omega_j = -\sum_{i=1}^{m}\mathrm{tr}\left(\tilde{b}_i\tilde{b}_i^T\right)G_j/2$. Taking expectations gives $E(\tilde{b}_i) = 0$ and $E(\tilde{b}_i\tilde{b}_i^T) = \sigma^2(\xi - U_i - A_i)$, producing

$$E\left(\frac{\partial^2 l_1}{\partial\tau\partial\omega_j}\right) = -\frac{\sigma^2}{2}\sum_{i=1}^{m}\mathrm{tr}(\xi - U_i)G_j + \frac{\sigma^2}{2}\sum_{i=1}^{m}\mathrm{tr}A_iG_j. \tag{20}$$

Notice that the second term in (20) is of a lower order than the first two terms. In an asymptotic sequence where $m \to \infty$, $U_i = O(1)$ and $A_i = O(m^{-1})$; thus the first term in (20) is $O(m)$, whereas the second is $O(1)$. Ignoring the second term gives

$$E\left(\frac{\partial^2 l_1}{\partial\tau\partial\omega_j}\right) \approx -\frac{\sigma^2}{2}\sum_{i=1}^{m}\mathrm{tr}(\xi - U_i)G_j. \tag{21}$$

The second derivatives with respect to ω are complicated, but after taking expectations and dropping lower-order terms they simplify to

$$E\left(\frac{\partial^2 l_1}{\partial\omega_j\partial\omega_k}\right) \approx -\frac{1}{2}\sum_{i=1}^{m}\mathrm{tr}\,(\xi - U_i)G_j(\xi - U_i)G_k. \tag{22}$$

Proofs of these results for l_1 are outlined in the Appendix. Notice that (21) and (22) are identical to the corresponding expressions for the regular loglikelihood l_0.

3 Fast algorithms for ML and RML estimation

The EM algorithm (Dempster, Laird, and Rubin, 1977) is a well known technique for parameter estimation in incomplete-data problems. Traditional applications of EM to the linear mixed model treat the random effects as missing data, relying on simplifications that occur when y is augmented by assumed values for $b_1, \ldots, b_m$. The likelihood function based on the augmented data $(y, b_1, \ldots, b_m)$ can be written as

$$\begin{aligned} L_A(\beta, \sigma^2, \psi) \quad \propto \quad & |\psi|^{-\frac{m}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr} \psi^{-1} \left(\sum_{i=1}^{m} b_i b_i^T \right) \right\} \\ & \times (\sigma^2)^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - X_i\beta - Z_i b_i)^T V_i^{-1} (y_i - X_i\beta - Z_i b_i) \right\}, \end{aligned} \tag{23}$$

where $\psi = \sigma^2 \xi$. Because L_A factors into distinct functions of ψ and (β, σ^2), the overall maximum can be found by maximizing each factor separately. In particular, the factor involving ψ is maximized at $\hat{\psi} = m^{-1} \sum_{i=1}^{m} b_i b_i^T$. Each cycle of EM maximizes the expected logarithm of L_A, where the expectation is taken with respect to the distribution of $b_1, \ldots, b_m$ given y with the parameters fixed at their most recent estimates. This expectation can be found by noting that $\log L_A$ is linear in b_i and $b_i b_i^T$, whose expectations are $\hat{b}_i$ and $\hat{b}_i \hat{b}_i^T + \sigma^2 U_i$, respectively.

An interesting variation on EM arises by noting that when ξ is held constant, the original likelihood L_0 becomes proportional to

$$(\sigma^2)^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - X_i\beta)^T W_i (y_i - X_i\beta) \right\},$$

where W_i is now fixed. For any given ξ, L_0 thus achieves a conditional maximum at $\beta = \tilde{\beta}$ and

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{m} (y_i - X_i\tilde{\beta})^T W_i (y_i - X_i\tilde{\beta}).$$

For fixed β and σ^2, however, L_0 does not reduce to a convenient function of ξ that can be maximized in closed form. The additional information contained in the "missing data" $b_1, \ldots, b_m$ is helpful for estimating ψ but is not really needed for (β, σ^2). This suggests a strategy in which we alternately (a) maximize L_0

with respect to (β, σ^2), holding ξ fixed at its current estimate; and (b) maximize $E(\log L_A)$ with respect to ξ, holding β and σ^2 fixed at their current estimates. This modified algorithm is no longer EM, but belongs to a more general class of procedures which Liu and Rubin (1995) have called ECME. This ECME algorithm for ML estimation, which does not seem to have been published before, updates the current estimates $\beta^{(t-1)}$, $\sigma^{2(t)}$, and $\xi^{(t)}$ by the following steps:

$$U_i^{(t)} = \left(\xi^{(t)^{-1}} + Z_i^T V_i^{-1} Z_i\right)^{-1}, \tag{24}$$

$$W_i^{(t)} = V^{-1} - V^{-1} Z_i U_i^{(t)} Z_i^T V_i^{-1}, \tag{25}$$

$$\beta^{(t)} = \left(\sum_{i=1}^m X_i^T W_i^{(t)} X_i\right)^{-1} \left(\sum_{i=1}^m X_i^T W_i^{(t)} y_i\right), \tag{26}$$

$$\sigma^{2(t+1)} = \frac{1}{N} \sum_{i=1}^m (y_i - X_i \beta^{(t)})^T W_i^{(t)} (y_i - X_i \beta^{(t)}), \tag{27}$$

$$\hat{b}_i^{(t)} = U_i^{(t)} Z_i^T V_i^{-1} (y_i - X_i \beta^{(t)}), \tag{28}$$

$$\xi^{(t+1)} = \frac{1}{m} \sum_{i=1}^m \left(\sigma^{-2(t)} \hat{b}_i^{(t)} \hat{b}_i^{(t)T} + U_i^{(t)}\right). \tag{29}$$

Slightly faster convergence may result if we replace $\sigma^{2(t)}$ by the updated estimate $\sigma^{2(t+1)}$ in (29), but for the developments below it notationally convenient not to do so. Note that V_i^{-1}, $Z_i^T V_i^{-1}$, and $Z_i^T V_i^{-1} Z_i$ may be calculated once and stored for all iterations. An attractive feature of this algorithm is that the loglikelihood function at each cycle can be evaluated after step (27) with almost no additional cost. Ignoring constant terms, the loglikelihood at cycle t reduces to

$$l_0(\beta^{(t)}, \sigma^{2(t)}, \xi^{(t)}) = -\frac{N}{2} \log \sigma^{2(t)} - \frac{m}{2} \log |\xi^{(t)}| + \frac{1}{2} \sum_{i=1}^m \log |U_i^{(t)}| - \frac{N}{2} \left(\frac{\sigma^{2(t+1)}}{\sigma^{2(t)}}\right).$$

The determinants of $\xi^{(t)}$ and $U_i^{(t)}$ may be computed along with their inverses during step (24).

Comparing these steps to the derivatives of l_0 presented in Section 2, we see that (26), (27), and (29) are precisely equivalent to setting $\partial l_0/\partial\beta$, $\partial l_0/\partial\tau$, and $\partial l_0/\partial\omega_j$ $(j = 1, \ldots, g)$ to zero. The algorithm (24)–(29) can thus be viewed as a simple fixed-point iteration that operates solely on first derivatives. It also suggests that the convergence rate can be substantially improved by using second-derivative information.

In the well known Newton-Raphson procedure, a function $l(\theta)$ is maximized by repeatedly solving the linear system $C\theta^{(t+1)} = d$, where $C = -\partial^2 l/\partial\theta\partial\theta^T$, $d = C\theta^{(t)} + \partial l/\partial\theta$, and the derivatives are evaluated

at $\theta = \theta^{(t)}$. In well behaved statistical applications where l is a loglikelihood function and θ represents unknown parameters, Newton-Raphson converges to an ML estimate $\hat{\theta}$, and first-order asymptotic theory allows the final value of C^{-1} to be used as an estimate of $V(\hat{\theta} - \theta)$. In practice it is not necessary to use the exact second derivatives of l; the asymptotic properties still hold if $C = -\partial^2 l / \partial\theta\partial\theta^T$ is replaced by $C = -\partial^2 l / \partial\theta\partial\theta^T + R$, provided that the remainder R is $o_p(n)$ where n is proportional to sample size (e.g. Cox and Hinkley, 1974, Ch. 9). When $C = -E(\partial^2 l / \partial\theta\partial\theta^T)$ the technique is called Fisher scoring. Using expected rather than observed second derivatives may simplify the calculations, and scoring typically converges about as quickly as Newton-Raphson.

To apply large-sample results to the present model, I assume an asymptotic sequence in which the the number of units m grows but the within-unit sample sizes n_i may remain bounded. In principle one could allow either or both of these measures to approach infinity. Many applications of the linear mixed model involve a large number of units but a modest number of observations per unit, so arguments requiring the n_i to be large are often unrealistic. If $m \to \infty$ but the n_i remain bounded, one cannot obtain consistent estimates of the b_i (Neyman and Scott, 1948). ML or RML estimates for the fixed effects and variance components will be consistent, however, differing from the true parameters by terms of size $O_p(m^{-1/2})$. In this framework, any terms smaller than $O_p(m)$ in the second derivatives of the loglikelihood may be ignored.

The results of Section 2 lead to a simple scoring algorithm for maximizing l_0. The cross-derivatives (17) and (18) have zero expectation, causing the scoring step to separate into independent linear systems for β and the variance parameters. One may easily verify that the scoring step to calculate $\beta^{(t)}$ from $(\beta^{(t-1)}, \sigma^{2(t)}, \xi^{(t)})$ is equivalent to (26), so the ECME algorithm already performs scoring for β. Denote the scoring step for the variance parameters as $C\eta = d$, where $\eta = (\eta_0, \eta_1, \ldots, \eta_g)^T = (\tau, \omega_1, \ldots, \omega_g)^T$, and the elements of C and d are c_{jk} and d_j, respectively $(j, k = 0, 1, \ldots, g)$. Applying results from Section 2, we obtain $c_{00} = N{\sigma^4}^{(t)}/2$,

$$c_{0j} = \frac{{\sigma^2}^{(t)}}{2} \sum_{i=1}^{m} \mathrm{tr}(\xi^{(t)} - U_i^{(t)}) G_j, \tag{30}$$

$$c_{jk} = \frac{1}{2}\sum_{i=1}^{m} \mathrm{tr}(\xi^{(t)} - U_i^{(t)})G_j(\xi^{(t)} - U_i^{(t)})G_k, \tag{31}$$

$$d_0 = N\sigma^{2\,(t)} - \frac{N}{2}\,\sigma^{2\,(t+1)}_{ECME} + \sum_{l=1}^{g} c_{0l}\omega_l^{(t)}, \tag{32}$$

$$d_j = \frac{m}{2}\,\mathrm{tr}\Big(\xi^{(t)} - \xi^{(t+1)}_{ECME}\Big)G_j + c_{0j}\sigma^{-2\,(t)} + \sum_{l=1}^{g} c_{jl}\omega_l^{(t)} \tag{33}$$

for $j = 1, \ldots, g$ and $k = 1, \ldots, g$, where $\sigma^{2\,(t+1)}_{ECME}$ and $\xi^{(t+1)}_{ECME}$ are the ECME-updated estimates given by (27) and (29). Upon solution ($\eta = C^{-1}d$), the scoring-updated estimates are $\sigma^{2\,(t+1)}_{SCORE} = 1/\eta_0$ and $\xi^{(t+1)}_{SCORE} = \left(\sum_{j=1}^{g}\eta_j G_j\right)^{-1}$.

Scoring-updated variance components are usually much closer to the ML estimates than their ECME counterparts. Unlike ECME, however, scoring is not guaranteed to increase the loglikelihood at each cycle. For this reason, I perform the scoring step concurrently with (28)–(29), tentatively setting $\sigma^{2\,(t+1)} = \sigma^{2\,(t+1)}_{SCORE}$ and $\xi^{(t+1)} = \xi^{(t+1)}_{SCORE}$ but storing the ECME estimates as well. After step (27) of the next cycle I evaluate the loglikelihood as

$$l_0(\beta^{(t+1)}, \sigma^{2\,(t+1)}, \xi^{(t+1)}) = -\frac{N}{2}\log\sigma^{2\,(t+1)} - \frac{m}{2}\log|\xi^{(t+1)}| + \frac{1}{2}\sum_{i=1}^{m}\log|U_i^{(t+1)}| - \frac{N}{2}\left(\frac{\sigma^{2\,(t+2)}_{ECME}}{\sigma^{2\,(t+1)}}\right).$$

If l_0 has decreased I reject the scoring estimates, replacing them with $\sigma^{2\,(t+1)} = \sigma^{2\,(t+1)}_{ECME}$ and $\xi^{(t+1)} = \xi^{(t+1)}_{ECME}$, and recalculate (24)–(27). Note that it is possible for the solution to $C\eta = d$ to lie outside the parameter space; the variance estimates implied by $\eta = C^{-1}d$ might not be positive definite. When this happens I use a step-halving procedure to move η back to an allowable value. Finally, if C is not positive definite, I abort the scoring procedure for that cycle and use the ECME estimates, returning a warning message that the loglikelihood at that cycle is not concave.

Now consider the closely related problem of RML estimation. Traditional EM algorithms for RML treat all coefficients $(\beta, b_1, \ldots, b_m)$ as missing data. Each cycle of EM maximizes the expected logarithm of L_A, where the expectation is now taken over the distribution of $(\beta, b_1, \ldots, b_m)$ given y under a uniform prior density for β. Instead of formulating an EM algorithm *per se*, let us consider an ECME procedure analogous to (24)–(29) for RML. Notice that in L_1, the quantities W_i and $\tilde{\beta}$ depend on ξ but not σ^2. If ξ

were known, L_1 would be maximized at

$$\sigma^2 = \frac{1}{(N-p)} \sum_{i=1}^{m} (y_i - X_i\tilde{\beta})^T W_i (y_i - X_i\tilde{\beta}).$$

Fixing σ^2, however, does not really simplify the estimation of ξ. Thus it makes sense to alternately (a) maximize L_1 with respect to σ^2, fixing ξ at its current estimate; and (b) maximize $E(\log L_A)$ with respect to ξ, fixing σ^2 at its current estimate. To carry out (b), notice that $\log L_A$ is a linear function of the terms $b_i b_i^T$ whose expectations follow from (11) and (12). The ECME cycle for RML becomes

$$U_i^{(t)} = \left({\xi^{(t)}}^{-1} + Z_i^T V_i^{-1} Z_i \right)^{-1}, \tag{34}$$

$$W_i^{(t)} = V^{-1} - V^{-1} Z_i U_i^{(t)} Z_i^T V_i^{-1}, \tag{35}$$

$$\Gamma^{(t)} = \left(\sum_{i=1}^{m} X_i^T W_i^{(t)} X_i \right)^{-1} \tag{36}$$

$$\tilde{\beta}^{(t)} = \Gamma^{(t)} \left(\sum_{i=1}^{m} X_i^T W_i^{(t)} y_i \right), \tag{37}$$

$$\sigma^{2\,(t+1)}_{ECME} = \frac{1}{(N-p)} \sum_{i=1}^{m} (y_i - X_i\tilde{\beta}^{(t)})^T W_i^{(t)} (y_i - X_i\tilde{\beta}^{(t)}), \tag{38}$$

$$\tilde{b}_i^{(t)} = U_i^{(t)} Z_i^T V_i^{-1} (y_i - X_i\tilde{\beta}^{(t)}), \tag{39}$$

$$A_i^{(t)} = U_i^{(t)} \gamma_i \Gamma_i^{(t)} \gamma_i^T U_i^{(t)}, \tag{40}$$

$$\xi_{ECME}^{(t+1)} = \frac{1}{m} \sum_{i=1}^{m} \left(\sigma^{-2\,(t)} \tilde{b}_i^{(t)} \tilde{b}_i^{(t)T} + U_i^{(t)} + A_i^{(t)} \right), \tag{41}$$

where $\gamma_i = Z_i^T V_i^{-1} X_i$. Applying results from Section 2, the RML scoring step solves $C\eta = d$ where $c_{00} = (N-p)\sigma^{4\,(t)}/2$,

$$c_{0j} = \frac{\sigma^{2\,(t)}}{2} \sum_{i=1}^{m} \mathrm{tr}(\xi^{(t)} - U_i^{(t)}) G_j, \tag{42}$$

$$c_{jk} = \frac{1}{2} \sum_{i=1}^{m} \mathrm{tr}(\xi^{(t)} - U_i^{(t)}) G_j (\xi^{(t)} - U_i^{(t)}) G_k, \tag{43}$$

$$d_0 = (N-p)\sigma^{2\,(t)} - \frac{(N-p)}{2} \sigma^{2\,(t+1)}_{ECME} + \sum_{l=1}^{g} c_{0l}\omega_l^{(t)}, \tag{44}$$

$$d_j = \frac{m}{2} \mathrm{tr}\left(\xi^{(t)} - \xi_{ECME}^{(t+1)} \right) G_j + c_{0j}\sigma^{-2\,(t)} + \sum_{l=1}^{g} c_{jl}\omega_l^{(t)}, \tag{45}$$

$j = 1, \ldots, g$, $k = 1, \ldots, g$. As before, I set $\sigma^{(t+1)}$ and $\xi^{(t+1)}$ equal to their scoring estimates unless the loglikelihood decreases, in which case I use the ECME versions. The loglikelihood at each cycle is available

following (38) as

$$l_1(\sigma^{2(t)}, \xi^{(t)}) = -\frac{(N-p)}{2}\log\sigma^{2(t)} - \frac{m}{2}\log|\xi^{(t)}| + \frac{1}{2}\sum_{i=1}^{m}\log|U_i^{(t)}| + \frac{1}{2}\log|\Gamma^{(t)}| - \frac{(N-p)}{2}\left(\frac{\sigma^{2(t+1)}_{ECME}}{\sigma^{2(t)}}\right).$$

Although I have presented scoring in terms of $\eta = (\tau, \omega_1, \ldots, \omega_g)^T$, one may also apply the method to other parameterizations of the variance components. If η^* is a one-to-one transformation of η, the scoring step for η^* solves $C^*\eta^{*(t+1)} = d^*$ where $C^* = (\partial\eta/\partial\eta^*)^T C(\partial\eta/\partial\eta^*)$ and $d^* = C^*\eta^{*(t)} + (\partial\eta/\partial\eta^*)^T(\partial l/\partial\eta)$. One parameterization that seems to work well applies a log transformation to τ and to each ω_j corresponding to a diagonal element of ξ^{-1}, but leaves the off-diagonal elements unchanged. In several data examples, scoring converged substantially faster on this new scale, particularly when the ML and RML estimates were close to the boundary of the parameter space.

These new ML and RML algorithms have been implemented as functions for S-PLUS (MathSoft, Inc., 1997) using subroutines written in Fortran-77. The functions currently assume that ψ unstructured, but extensions to other (e.g. block-diagonal) forms may be made available in the future. They can be downloaded from my website at `http://www.stat.psu.edu/~jls/`. I now illustrate their performance with a small data example.

The data in Table 1, reported by Weil, Zinberg, and Nelson (1968), come from a pilot study of the clinical and psychological effects of marijuana. Nine male subjects were given three treatments in the form of low-dose, high-dose, and placebo cigarettes. The order of treatments within subjects was balanced in a replicated 3×3 Latin square, but because the order for each subject was not reported in the article, I shall proceed as if the order effects are negligible. Changes in heart rate were recorded 15 and 90 minutes after marijuana use, and five of the 54 data values are missing.

Letting y_{ijk} denote the response for subject $i = 1, \ldots, 9$, treatment j (1=placebo, 2=low dose, 3=high dose), and time k (1=15 minutes, 2=90 minutes), I fit the compound symmetry model $y_{ijk} = \mu_{jk} + b_i + \varepsilon_{ijk}$, where the μ_{jk} are fixed effects for treatment $\times$ time, $b_i \sim N(0, \psi)$ are random effects for subjects, and $\varepsilon_{ijk} \sim N(0, \sigma^2)$ are independent experimental errors. Without additional constraints on the μ_{jk}, this model

Table 1: Change in heart rate recorded 15 and 90 minutes after marijuana use, measured in beats per minute above baseline

	15 minutes			90 minutes		
Subject	Placebo	Low	High	Placebo	Low	High
1	16	20	16	2	−6	−4
2	12	24	12	−6	4	−8
3	8	8	26	−4	4	8
4	20	8	—	—	20	−4
5	8	4	−8	—	22	−8
6	10	20	28	−20	−4	−4
7	4	28	24	12	8	18
8	−8	20	24	−3	8	−24
9	—	20	24	8	12	—
mean	8.8	16.9	18.2	1.0	7.6	−3.2

has $p = 6$ fixed effects. Starting values for parameters were obtained by the procedure of Laird, Lange, and Stram (1987), and the algorithms were stopped when the relative change in all parameters from one cycle to the next dropped below 0.01%. The ML algorithm converged to $(\hat{\sigma}^2, \hat{\psi}) = (87.88, 3.089)$ and $\hat{\mu} = (\mu_{11}, \mu_{21}, \mu_{31}, \mu_{21}, \mu_{22}, \mu_{23}) = (8.838, 16.89, 18.30, -1.640, 7.556, -3.162)$ in 8 cycles, and the RML version converged to $(\hat{\sigma}^2, \hat{\psi}) = (100.2, 3.477)$ and $\hat{\mu} = (8.837, 16.89, 18.30, -1.640, 7.556, -3.163)$ in 10 cycles. In contrast, the ECME algorithms for ML and RML took 221 and 247 cycles, respectively, and required about 10 times as much processing time. Each cycle of the new algorithms took only about 50% longer than the corresponding cycles of ECME.

Similar improvements over EM-type algorithms have been seen in a variety of small and large datasets; time savings of 90% or more are common. Scoring typically converges by 10–15 cycles, whereas EM may require hundreds or thousands. Unless the number of variance parameters is unusually large, the per-iteration cost of the new methods tends to be roughly 1.5 times that of EM. When the loglikelihood is oddly shaped and the scoring procedure fails, EM takes over and the user is warned of the anomaly.

4 Corrections to empirical Bayes interval estimates

Let us now consider the problem of inferences about individual random effects b_i. If values of (β, σ^2, ψ) were known, the conditional posterior distribution of b_i would be

$$b_i \mid y, \beta, \sigma^2, \psi \sim N(\hat{b}_i, \sigma^2 U_i). \tag{46}$$

One can show that if $m \to \infty$ but the n_i remain bounded, the point estimate $\hat{b}_i$ is not a consistent estimate of b_i because $U_i = O(1)$ is stable. Nonetheless, many statisticians would agree that (46) provides a sound basis for point and interval estimation for functions of b_i when (β, σ^2, ψ) are known. Substituting consistent estimates for (β, σ^2, ψ) into (46) leads to traditional empirical Bayes (EB) inferences for b_i. This method may work well with large samples when the fixed effects and variance components are well estimated. In small to moderate samples, however, the procedure tends to be artificially precise and may substantially understate the actual uncertainty. Here I present a new method for interval estimation which involves a full correction of the conditional variance estimate $\sigma^2 U_i$ up to terms of size $O(m^{-1})$ to account for uncertainty due to β, σ^2 and ψ.

The new method can be motivated with frequentist arguments, but it is logically simpler when presented as a first-order approximation to fully Bayesian inference. The basic idea is to use the loglikelihood and its derivatives to construct an approximate joint posterior distribution for β and the variance components. Using Taylor linearization, this approximate posterior is then combined with (46) to approximate the unconditional posterior mean and variance of b_i. Despite the Bayesian motivation, the user is not required to specify prior distributions for the unknown parameters. Moreover, the method seems to perform well by frequentist criteria; in simulations the interval estimates have frequency coverage close to nominal levels in even in samples that are quite small.

To construct the approximate posterior distribution, let $\hat{\eta} = (\hat{\tau}, \hat{\omega}_1, \ldots, \hat{\omega}_g)^T$ denote RML estimates of the variance-component parameters. Let $\tilde{\beta}$ denote the generalized least-squares estimate (7), which is implicitly a function of η, and let $\hat{\tilde{\beta}}$ denote the value of $\tilde{\beta}$ obtained by setting $\eta = \hat{\eta}$. Finally, let $\hat{C}^{-1}$ denote the final value of C^{-1} upon convergence of the RML scoring procedure. The approximation $(\hat{\eta} - \eta) \sim N(0, \hat{C}^{-1})$ can be justified from either a frequentist or a Bayesian perspective (e.g. DeGroot, 1970, ch. 10; Gelman et al., 1995, Appendix B). Adopting the Bayesian view, we can regard

$$\eta \mid y \sim N(\hat{\eta}, \hat{C}^{-1}) \tag{47}$$

as an approximate marginal posterior distribution for η. The conditional posterior for β given η is

$$\beta \mid y, \eta \sim N(\tilde{\beta}, \sigma^2 \Gamma), \tag{48}$$

from which it follows that the unconditional moments are $E(\beta \mid y) = E(\tilde{\beta} \mid y)$ and $V(\beta \mid y) = E(\sigma^2\Gamma \mid y) + V(\tilde{\beta} \mid y)$. Calculating these moments is difficult because $\tilde{\beta}$ and $\sigma^2\Gamma$ are nonlinear functions of η. Using the first-order Taylor expansion

$$(\tilde{\beta} - \hat{\tilde{\beta}}) = \left(\frac{\partial \tilde{\beta}}{\partial \eta}\right)_{\eta=\hat{\eta}} (\eta - \hat{\eta}) + O_p(m^{-1}), \tag{49}$$

however, we obtain the approximations $E(\beta \mid y) \approx \hat{\tilde{\beta}}$ and

$$V(\tilde{\beta} \mid y) \approx \left(\frac{\partial \tilde{\beta}}{\partial \eta}\right)_{\eta=\hat{\eta}} \hat{C}^{-1} \left(\frac{\partial \tilde{\beta}}{\partial \eta}\right)^T_{\eta=\hat{\eta}}.$$

Similarly, a first-order expansion of $\sigma^2\Gamma$ about $\eta = \hat{\eta}$ leads to $E(\sigma^2\Gamma \mid y) \approx \hat{\sigma}^2\hat{\Gamma}$, where $\hat{\sigma}^2$ and $\hat{\Gamma}$ are calculated by setting $\eta = \hat{\eta}$. Finally, it follows from (48) and (49) that

$$Cov(\beta, \eta \mid y) \approx \left(\frac{\partial \tilde{\beta}}{\partial \eta}\right)_{\eta=\hat{\eta}} \hat{C}^{-1}.$$

Combining these results, we obtain a limiting posterior distribution for $(\beta^T, \eta^T)^T$ that is normal with mean $(\hat{\tilde{\beta}}^T, \hat{\eta}^T)^T$ and covariance matrix

$$\begin{bmatrix} \hat{\sigma}^2\hat{\Gamma} + \left(\frac{\partial \tilde{\beta}}{\partial \eta}\right)_{\eta=\hat{\eta}} \hat{C}^{-1} \left(\frac{\partial \tilde{\beta}}{\partial \eta}\right)^T_{\eta=\hat{\eta}} & \left(\frac{\partial \tilde{\beta}}{\partial \eta}\right)_{\eta=\hat{\eta}} \hat{C}^{-1} \\ \hat{C}^{-1} \left(\frac{\partial \tilde{\beta}}{\partial \eta}\right)^T_{\eta=\hat{\eta}} & \hat{C}^{-1} \end{bmatrix} = O_p(m^{-1}). \tag{50}$$

The derivatives of $\tilde{\beta}$ with respect to the elements of $\eta = (\tau, \omega_1, \ldots, \omega_g)^T$ are $\partial\tilde{\beta}/\partial\tau = 0$ and $\partial\tilde{\beta}/\partial\omega_j = \Gamma\left(\sum_{i=1}^m \gamma_i^T U_i G_j \tilde{b}_i\right)$.

Now let us use this posterior to obtain approximate Bayesian inferences for the random effects. Using (46) and (48), it is straightforward to show that the distribution of b_i given η (but not β) is normal with mean $\tilde{b}_i$ and variance $\sigma^2(U_i + A_i)$, both of which are nonlinear functions of η. Expanding these functions about $\eta = \hat{\eta}$, and using $E(b_i \mid y) = E(\tilde{b}_i \mid y)$ and $V(b_i \mid y) = E(\sigma^2(U_i + A_i) \mid y) + V(\tilde{b}_i \mid y)$, it follows that the limiting posterior distribution of b_i is normal with moments

$$E(b_i \mid y) \approx \hat{\tilde{b}}_i, \tag{51}$$

$$V(b_i \mid y) \approx \hat{\sigma}^2(\hat{U}_i + \hat{A}_i) + \left(\frac{\partial \tilde{b}_i}{\partial \eta}\right)_{\eta=\hat{\eta}} \hat{C}^{-1} \left(\frac{\partial \tilde{b}_i}{\partial \eta}\right)^T_{\eta=\hat{\eta}}, \tag{52}$$

where $\hat{\tilde{b}}_i$, $\hat{U}_i$, and $\hat{A}_i$ are obtained by setting $\eta = \hat{\eta}$. The derivatives of $\tilde{b}_i$ with respect to the elements of η are $\partial\tilde{b}_i/\partial\tau = 0$ and $\partial\tilde{b}_i/\partial\omega_j = -U_i G_j \tilde{b}_i - U_i\gamma_i(\partial\tilde{\beta}/\partial\omega_j)$.

In practice, one may want to draw inferences about linear combinations of b_i and β, e.g. the subject-level means $X_i\beta + Z_i b_i$. Let X and Z be arbitrary known matrices with p and q columns, respectively. Approximate confidence regions for $X\beta + Zb_i$ may be obtained from a normal distribution with mean $E(X\beta + Zb_i \mid y) \approx X\hat{\tilde{\beta}} + Z\hat{\tilde{b}}_i$ and variance

$$\begin{aligned} V(X\beta + Zb_i \mid y) &= XV(\beta \mid y)X^T + ZV(b_i \mid y)Z^T \\ &\quad + X\,Cov(b_i, \beta \mid y)^T Z^T + Z\,Cov(b_i, \beta \mid y)X^T. \end{aligned} \tag{53}$$

Approximations to the variance terms in (53) are given by (52) and the upper-left portion of (50). For the covariances, it can be shown that

$$Cov(b_i, \beta \mid y) = -E(\sigma^2 U_i \gamma_i \Gamma \mid y) + Cov(\tilde{b}_i, \tilde{\beta} \mid y),$$

and by expanding $(\tilde{b}_i, \tilde{\beta})$ about $\eta = \hat{\eta}$ we obtain $-E(\sigma^2 U_i \gamma_i \Gamma \mid y) \approx -\hat{\sigma}^2 \hat{U}_i \gamma_i \hat{\Gamma}$ and

$$Cov(\tilde{b}_i, \tilde{\beta} \mid y) \approx \left(\frac{\partial \tilde{b}_i}{\partial \eta}\right)_{\eta=\hat{\eta}} \hat{C}^{-1} \left(\frac{\partial \tilde{\beta}}{\partial \eta}\right)^T_{\eta=\hat{\eta}}.$$

These approximations have been implemented in S-PLUS and Fortran-77 as part of RML procedure of Section 3. Upon convergence, the user is provided with $\hat{\tilde{b}}_i$ and estimates of $V(b_i \mid y)$, $V(\beta \mid y)$, and $Cov(b_i, \beta \mid y)$ for $i = 1, \ldots, m$, which can be used to compute point and interval estimates for arbitrary linear combinations of b_i and β.

Returning to the data of Table 1, this procedure was used to obtain 95% interval estimates for the random subject effects b_i in the model $y_{ijk} = \mu_{jk} + b_i + \epsilon_{ijk}$. Intervals were calculated as $\hat{\tilde{b}}_i \pm 2[V(b_i \mid y)]^{1/2}$, with $V(b_i \mid y)$ approximated using (52). The results are shown in Table 2, along with conventional EB intervals $\hat{\tilde{b}}_i \pm 2(\hat{\sigma}^2 \hat{U}_i)^{1/2}$ which ignore uncertainty in β, σ^2 and ψ. The new intervals are from 2% to 141% wider than the conventional ones. The relative expansion depends on the magnitude of the estimated random effect; larger estimates for b_i produce greater increases.

To assess the performance of the new method, I conducted a small simulation designed to mimic the data of this example. Responses for $m = 9$ subjects were drawn according to model $y_{ijk} = \mu_{jk} + b_i + \epsilon_{ijk}$ with $\mu = (10, 15, 20, 0, 0, 0)$, $\sigma^2 = 90$, and $\psi = 10$, producing an intra-subject correlation of $10/(10 + 90) = 0.1$.

Table 2: Random effect estimates and 95% intervals calculated by conventional and new methods, with relative increase in width under new method

Subject	Est.	Conventional	New	Increase (%)
1	−0.080	(−3.47, 3.31)	(−3.55, 3.39)	2
2	−0.252	(−3.64, 3.14)	(−3.97, 3.46)	9
3	0.092	(−3.30, 3.49)	(−3.38, 3.56)	2
4	0.423	(−3.07, 3.92)	(−3.86, 4.70)	22
5	−0.900	(−4.34, 2.54)	(−7.08, 5.29)	80
6	−0.482	(−3.87, 2.91)	(−4.83, 3.87)	28
7	1.356	(−2.04, 4.75)	(−6.81, 9.52)	141
8	−0.855	(−4.25, 2.54)	(−6.69, 4.98)	72
9	0.698	(−2.80, 4.19)	(−4.68, 6.07)	54

Missing values were introduced completely at random at an average rate of 10%. The sampling procedure was repeated 1000 times, with new b_i and ϵ_{ijk} drawn each time. For each sample, I calculated nominal 95% interval estimates by the new and old methods and noted whether the intervals covered the true values of b_i. For 245 of the 1000 samples, neither method could be used because the RML estimate of ψ fell on the boundary (which was assumed to have happened if the RML procedure converged to an estimate below 10^{-4}). Among the remaining 755 samples, the intervals calculated by the new method had an average coverage rate of 94.1%, compared to 87.2% for the old method. The new intervals were on average 35% wider than the old, suggesting that uncertainty in estimating β, σ^2, and ψ plays a substantial role in inferences about $b_1, \ldots, b_m$.

Despite the small sample size and proximity of ψ to the boundary, the performance of the new procedure in this simulation is very encouraging. The new intervals essentially attained their nominal coverage, whereas the old intervals had an error rate of more than 2.5 times their stated value. To improve the behavior of the RML estimation procedure, I repeated the experiment after raising the number of subjects to $m = 15$ and the intra-subject correlation to 0.5. Under these new conditions, only one of the 1000 estimates of ψ fell on the boundary, and the simulated coverage of the new method was 94.1% versus 89.2% for the old.

5 A rapidly converging MCMC algorithm

My final application is a new Markov chain Monte Carlo (MCMC) algorithm for simulating draws of parameters from a Bayesian posterior distribution. Conventional MCMC algorithms for linear mixed models regard (β, σ^2, ψ) as part of a larger system of unknown quantities that also includes $B = (b_1, \ldots, b_m)$; draws from the joint posterior distribution of $(\beta, \sigma^2, \psi, B)$ are then obtained by repeatedly drawing from various conditional posterior distributions in turn. For example, consider an iterative algorithm in which the current parameter values $(\beta^{(t)}, \sigma^{2\,(t)}, \psi^{(t)})$ and random effects $B^{(t)}$ are updated in four steps by drawing from the conditional posterior distributions

$$\sigma^{2\,(t+1)} \sim P(\sigma^2 \mid y, \beta^{(t)}, \psi^{(t)}, B^{(t)}), \tag{54}$$

$$\beta^{(t+1)} \sim P(\beta \mid y, \sigma^{2\,(t+1)}, \psi^{(t)}, B^{(t)}), \tag{55}$$

$$B^{(t+1)} \sim P(B \mid y, \beta^{(t+1)}, \sigma^{2\,(t+1)}, \psi^{(t)}), \tag{56}$$

$$\psi^{(t+1)} \sim P(\psi \mid y, \beta^{(t+1)}, \sigma^{2\,(t+1)}, B^{(t+1)}), \tag{57}$$

in a slight abuse of notation. The cycle (54)–(57) defines a discrete-time, continuous-state space Markov chain called a Gibbs sampler. Given starting values in the support of the parameter space, the distribution of $(\beta^{(t)}, \sigma^{2\,(t)}, \psi^{(t)}, B^{(t)})$ converges to $P(\beta, \sigma^2, \psi, B \mid y)$ as $t \to \infty$. Technical details on the convergence of Gibbs samplers and related methods are provided by Liu, Wong, and Kong (1994) and Tierney (1996). Overviews of MCMC are given by Gelfand and Smith (1990); Smith and Roberts (1993); Tanner (1993); and Gilks, Richardson, and Spiegelhalter (1996). Gibbs samplers for linear mixed models are described by Gelfand *et al.* (1990); Zeger and Karim (1991); Liu and Rubin (1995); and Carlin (1996).

Like EM algorithms, Gibbs samplers that simulate $b_1, \ldots, b_m$ may converge very slowly. The worst performance occurs when m is large and the random effects are poorly estimated, i.e. where the within-unit precision matrices $Z_i^T V_i^{-1} Z_i$ are small relative to the between-unit precision ξ^{-1}. Nevertheless, these Gibbs samplers are stable and easy to implement. I will use the derivative approximations of Section 2 to construct a new MCMC algorithm that tends to converge more rapidly. Before presenting the new algorithm, however, I will describe a modified Gibbs sampler that is closely related to the ECME procedure

for RML estimation from Section 3.

Several authors, including Gelman (1992) and Tierney (1994), have noted that one need not sample from exact conditional distributions at each cycle of a Gibbs sampler. Any of the steps (54)–(57), for example, may be replaced by one or more cycles of another MCMC algorithm that converges to the respective conditional, and the stationary distribution overall algorithm will be maintained. This idea leads to an MCMC algorithm that nests one Gibbs sampler inside another. Consider a two-step Gibbs sampler in which we draw from (a) the conditional posterior distribution of σ^2 given ξ, and (b) the conditional posterior distribution of ξ given σ^2. Alternating between (a) and (b) eventually produces a draw from $P(\sigma^2, \xi \mid y)$. Step (b) is difficult because the likelihood L_1 is a complicated function of ξ. Inferences about ξ become much easier, however, if y is augmented by simulated values of β and B. Suppose we replace (b) by one or more cycles of another two-step Gibbs sampler

$$(\beta^{(t)}, B^{(t)}) \sim P(\beta, B \mid y, \sigma^2, \xi^{(t)}), \tag{58}$$

$$\xi^{(t+1)} \sim P(\xi \mid y, \sigma^2, \beta^{(t)}, B^{(t)}). \tag{59}$$

Alternating between (58) and (59) eventually produces a draw of ξ from $P(\xi \mid y, \sigma^2)$. Imbedding a cycle of (58)–(59) into the Gibbs sampler for σ^2 and ξ leads to

$$\sigma^{2(t+1)} \sim P(\sigma^2 \mid y, \xi^{(t)}) \tag{60}$$

$$\beta^{(t)} \sim P(\beta \mid y, \sigma^{2(t)}, \xi^{(t)}) \tag{61}$$

$$B^{(t)} \sim P(B \mid y, \beta^{(t)}, \sigma^{2(t)}, \xi^{(t)}), \tag{62}$$

$$\xi^{(t+1)} \sim P(\xi \mid y, \sigma^{2(t)}, B^{(t)}), \tag{63}$$

where $\beta^{(t)}$ has dropped out of (63) because it carries no information about ξ once B is known. Notice that (61) and (62) are equivalent to simply drawing B from its marginal distribution $P(B \mid y, \sigma^{2(t)}, \xi^{(t)})$. Simulation of B is most conveniently accomplished in two steps, however, because given β the random effects $b_1, \ldots, b_m$ are conditionally independent. Slightly faster convergence may result if we condition on $\sigma^{2(t+1)}$ rather than $\sigma^{2(t)}$ in (61)–(63), but formulas that follow remain simpler if we do not.

Steps (60) and (63) require prior distributions for σ^2 and ξ. Following common practice, let us suppose

that σ^2 and $\psi = \sigma^2\xi$ are independently distributed as $\sigma^2 \sim a\chi_b^{-2}$ and $\psi^{-1} \sim W(c, D)$, where a, b, c, and D are user-specified hyperparameters and $W(c, D)$ denotes a Wishart distribution with degrees of freedom c and scale matrix D. Values for a and b may be chosen by regarding a/b as a rough prior guess for σ^2 and b as an imaginary degrees of freedom on which this guess is based. Similarly, $c^{-1}D^{-1}$ may be regarded as a guess for ψ with $c \geq q$ degrees of freedom. When little or no prior information is available, one may choose a/b and $c^{-1}D^{-1}$ to be near the RML estimates of σ^2 and ψ, respectively, and take the degrees of freedom to be weak so that the inference will be dominated by the shape of the likelihood rather than the prior. Notice that the inverse-Wishart prior is appropriate for models where ψ is unstructured. Block-diagonal situations can be handled by applying independent Wishart distributions to the non-zero blocks of ψ^{-1}, which introduces only minor modifications to the following procedures.

Under this joint prior for (σ^2, ψ), it is a straightforward exercise in transformation to show that the implied conditional priors for σ^2 and ξ are

$$\sigma^2 \mid \xi \sim \left(a + \mathrm{tr}D^{-1}\xi^{-1}\right)\chi_{b+cq}^{-2}, \tag{64}$$

$$\xi^{-1} \mid \sigma^2 \sim W(c, \sigma^2 D). \tag{65}$$

Combining (64) with L_1 leads to $\sigma^2 \mid (y, \xi) \sim \left(a' + \mathrm{tr}D^{-1}\xi^{-1}\right)\chi_{b'+cq}^{-2}$, where $b' = b + N - p$ and $a' = a + \sum_{i=1}^{m}(y_i - X_i\tilde{\beta})^T W_i (y_i - X_i\tilde{\beta})$, and combining (65) with L_A leads to $\xi^{-1} \mid (y, \sigma^2, B) \sim W(c', \sigma^2 D')$, where $c' = c + m$ and $D' = \left(D^{-1} + \sum_{i=1}^{m} b_i b_i^T\right)^{-1}$. The modified Gibbs sampler can thus be implemented as

$$U_i^{(t)} = \left(\xi^{(t)^{-1}} + Z_i^T V_i^{-1} Z_i\right)^{-1}, \tag{66}$$

$$W_i^{(t)} = V^{-1} - V^{-1} Z_i U_i^{(t)} Z_i^T V_i^{-1}, \tag{67}$$

$$\Gamma^{(t)} = \left(\sum_{i=1}^{m} X_i^T W_i^{(t)} X_i\right)^{-1} \tag{68}$$

$$\tilde{\beta}^{(t)} = \Gamma^{(t)}\left(\sum_{i=1}^{m} X_i^T W_i^{(t)} y_i\right), \tag{69}$$

$$a'^{(t)} = a + \sum_{i=1}^{m}(y_i - X_i\tilde{\beta}^{(t)})^T W_i^{(t)}(y_i - X_i\tilde{\beta}^{(t)}), \tag{70}$$

$$\sigma^{2(t+1)} \sim \left(a'^{(t)} + \mathrm{tr}D^{-1}\xi^{(t)^{-1}}\right)\chi_{b+N-p+cq}^{-2}, \tag{71}$$

$$\beta^{(t)} \sim N(\,\tilde{\beta}^{(t)},\, \sigma^{2(t)}\Gamma^{(t)}\,), \tag{72}$$

$$b_i^{(t)} \sim N(U_i^{(t)} Z_i^T V_i^{-1}(y_i - X_i\beta^{(t)}),\, \sigma^{2(t)} U_i^{(t)}), \tag{73}$$

$$\xi^{(t+1)^{-1}} \sim W\left(c + m,\, \sigma^{2(t)}\left(D^{-1} + \sum_{i=1}^{m} b_i^{(t)} b_i^{(t)T}\right)^{-1}\right). \tag{74}$$

This algorithm bears a strong resemblance to the ECME method for RML estimation described in Section 3; it may be regarded as a stochastic version of the ECME procedure, with expectation and maximization procedures replaced by simulation. Moreover, just as second derivatives were useful in speeding the convergence of ECME, they are also useful in speeding the convergence of this algorithm.

My technique for speeding convergence is based on another popular MCMC method known as the Metropolis-Hastings (MH) algorithm (Metropolis *et al.*, 1953; Hastings, 1970). Overviews of MH are given by Gelman *et al.* (1995) and Gilks, Richardson, and Spiegelhalter (1996). Given a current simulated value $\eta^{(t)}$ of the variance parameters $\eta = (\tau, \omega_1, \ldots, \omega_g)^T$, suppose that we sample a candidate value $\eta^\dagger$ from a density function $h(\eta^\dagger)$ meant to approximate the marginal posterior $P(\eta \,|\, y) \propto L_1(\eta)\,\pi(\eta)$, where $\pi(\eta)$ is the prior density applied to η. We then calculate the acceptance ratio

$$R^{(t)} = \frac{P(\eta^\dagger \,|\, y)}{P(\eta^{(t)} \,|\, y)} \, \frac{h(\eta^{(t)})}{h(\eta^\dagger)}$$

and set

$$\eta^{(t+1)} = \begin{cases} \eta^\dagger & \text{if } u \le R^{(t)}, \\ \eta^{(t)} & \text{if } u > R^{(t)}, \end{cases}$$

where $u \sim U(0,1)$ is a uniform random variate. This defines a special type of MH algorithm called an independence sampler (Tierney, 1994). Executing these steps repeatedly creates a Markov chain whose stationary distribution is $P(\eta \,|\, y)$, provided that h is nonzero over the support of $P(\eta \,|\, y)$. The algorithm converges rapidly if h is a good approximation to $P(\eta \,|\, y)$, producing acceptance ratios close to one. It is wise to choose h to have somewhat heavier tails than $P(\eta \,|\, y)$, so that the candidate values cover the region of appreciable posterior density, and to reduce the chance of "getting stuck" in the tails of $P(\eta \,|\, y)$.

My approximation to $P(\eta \,|\, y)$ is based on a multivariate t distribution. If $Z \sim N_d(0, S)$ and $X \sim \chi^2_\nu$ are independent, then $T = Z\sqrt{\nu/X} + \mu$ is said to have a multivariate t distribution centered at μ with scale matrix S $(d \times d)$ and ν degrees of freedom; its density function is

$$h(T) \propto \left[1 + \frac{1}{\nu}\,(T-\mu)^T S^{-1} (T-\mu)\right]^{-\left(\frac{\nu+d}{2}\right)}. \tag{75}$$

I approximate $P(\eta \mid y)$ by a multivariate t distribution centered at the mode of $P(\eta \mid y)$, which can be calculated by a slight modification of the scoring procedure for RML estimation given in Section 3; this modification is described in the Appendix. The scale matrix S is found by equating the final value of $-C$ from scoring to the second derivative of the logarithm of the t density, which gives $S = \left(\frac{\nu+g+1}{\nu}\right) C^{-1}$. Regarding the degrees of freedom, Gelman *et al.* (1995) suggest that $\nu = 4$ is often reasonable for approximating posterior distributions; my experience also indicates that $\nu = 4$ does tend to be a good choice here.

Notice that drawing candidates from a t distribution may occasionally produce variance components outside the parameter space ($\sigma^2 < 0$ or $\xi < 0$). When this happens, we may simply discard the bad candidate and draw again. In effect we are sampling from a t distribution that has been truncated outside the allowable region. Truncation introduces an unknown normalizing constant into the approximate density h, but this constant cancels out in the acceptance ratio and may be ignored.

The quality of the approximation may be improved by applying the t distribution to some nonlinear transformation of η rather than to η itself. If η^* is a one-to-one function of η, the scale matrix for η^* can be obtained by replacing C with $C^* = (\partial\eta/\partial\eta^*)^T C (\partial\eta/\partial\eta^*)$. The candidate for η is obtained by back-transforming the candidate for η^*, and the approximation density must take into account the Jacobian $|\partial\eta^*/\partial\eta|$. Good results have been obtained by applying a log transformation to τ and to each ω_j corresponding to a diagonal element of ξ^{-1}, while leaving the off-diagonal elements unchanged. This transformation with $\nu = 4$ degrees of freedom produced acceptance rates of 40–60% in a variety of examples.

The acceptance ratio $R^{(t)}$ requires evaluation of $P(\eta \mid y)$ at each new candidate value. The density at any particular value $\eta^{(t)}$ can be calculated as

$$\begin{aligned} \log P(\eta^{(t)} \mid y) \quad = \quad & -\left(\frac{N-p+b+cq-2}{2}\right)\left[\log \sigma^{2(t)} + \frac{\hat{\sigma}^2}{\sigma^{2(t)}}\right] \\ & -\left(\frac{m+c-q-1}{2}\right)\log|\xi^{(t)}| + \frac{1}{2}\sum_{i=1}^{m}\log|U_i^{(t)}| + \frac{1}{2}\log|\Gamma^{(t)}|, \end{aligned} \tag{76}$$

where

$$\hat{\sigma}^2 = \frac{a + \mathrm{tr} D^{-1}\xi^{(t)^{-1}} + \sum_{i=1}^{m}(y_i - X_i\tilde{\beta}^{(t)})^T W_i^{(t)}(y_i - X_i\tilde{\beta}^{(t)})}{N-p+b+cq-2},$$

and $U_i^{(t)}$, $\Gamma_i^{(t)}$, $W_i^{(t)}$ and $\tilde{\beta}^{(t)}$ are given by (66)–(69).

Notice that evaluating $P(\eta \mid y)$ requires many of the steps needed for a cycle of the modified Gibbs sampler (66)–(74). This suggests that, if the candidate value from MH is rejected, we can still update η inexpensively by completing the Gibbs cycle. More specifically, given the current simulated values $\sigma^{2(t)}$ and $\xi^{(t)}$, suppose we perform a single cycle of (66)–(74) to obtain Gibbs-updated versions $\sigma^{2\,(t+1)}_{GIBBS}$ and $\xi_{\,GIBBS}^{(t+1)}$. After step (69) of this cycle, the posterior density $P(\eta^{(t)} \mid y)$ corresponding to $\sigma^{2\,(t)}$ and $\xi^{(t)}$ becomes available via (76). At the same time we draw candidate values for MH, calling them $\sigma^{2(t+1)}_{MH}$ and $\xi_{\,MH}^{(t+1)}$. We tentatively set $\sigma^{2(t+1)} = \sigma^{2(t+1)}_{MH}$ and $\xi^{(t+1)} = \xi_{\,MH}^{(t+1)}$ and proceed to the next cycle of Gibbs. When the value of $P(\eta^{(t+1)} \mid y)$ corresponding to the tentative $\sigma^{2(t+1)}$ and $\xi^{(t+1)}$ becomes available after step (69), we complete the calculation of $R^{(t)}$ and decide whether to retain the tentative values. If they are retained, we finish the cycle to obtain $\sigma^{2\,(t+2)}_{GIBBS}$, $\xi_{\,GIBBS}^{(t+2)}$, $\sigma^{2(t+2)}_{MH}$ and $\xi_{\,MH}^{(t+2)}$. If they are rejected, we set $\sigma^{2(t+1)} = \sigma^{2\,(t+1)}_{GIBBS}$ and $\xi^{(t+1)} = \xi_{\,GIBBS}^{(t+1)}$ and re-do (66)–(69).

Setting $(\sigma^{2(t+1)}, \xi^{(t+1)})$ to $(\sigma^{2\,(t+1)}_{GIBBS}, \xi_{\,GIBBS}^{(t+1)})$ rather than to $(\sigma^{2(t)}, \xi^{(t)})$ upon rejection does not alter the stationary distribution, because the Gibbs algorithm has the same stationary distribution as MH. One advantage of doing this is that we will never "get stuck" at a single state; if the approximation used in MH is so poor that all candidate values are rejected, the algorithm still progresses at the same rate as the Gibbs algorithm (66)–(74). If the approximation is good, however, the convergence can be substantially faster. This hybrid combination of MH and Gibbs is especially attractive because the performance of MH tends to improve precisely where the Gibbs sampler deteriorates. Conventional Gibbs samplers converge more slowly as the number of units m grows, because the augmented-data likelihood function L_A becomes tightly concentrated about the current value of ψ, causing $\psi^{(t)}$ and $\psi^{(t+1)}$ to be highly correlated. As m grows, however, the actual posterior $P(\eta \mid y)$ more closely resembles a normal distribution, leading to higher MH acceptance rates.

I now demonstrate the performance of this algorithm on three data examples. In each I used $\nu = 4$ degrees of freedom for the t approximation and applied log transformations to σ^2 and the diagonal elements of ξ^{-1}. The first example uses the data from Table 1 and the compound-symmetry model described in

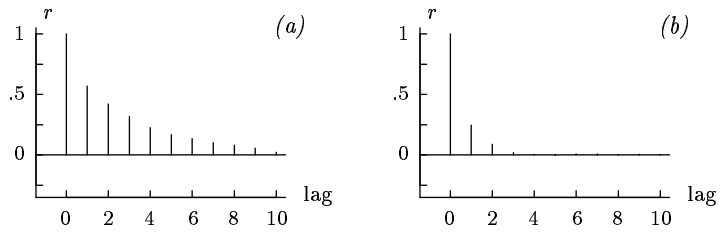

Figure 1: *Sample autocorrelation r for* $\log\psi$ *under (a) the modified Gibbs sampler and (b) the new MCMC method, using data from Table 1.*

Section 3, with prior distributions $\sigma^2 \sim 300\chi_3^{-2}$ and $\psi^{-1} \sim W(3, 1/15) = \chi_3^2/15$. For comparison, I ran both the modified Gibbs sampler and the new MCMC algorithm for 5,000 cycles each, storing the simulated values of σ^2 and ψ for all iterations. Time-series plots and autocorrelation functions (ACFs) revealed that both chains converged rapidly, but the new method was noticeably faster. Sample ACFs for $\log\psi$, shown in Figure 1, suggest that the serial dependence dies down in 8–10 cycles with the Gibbs sampler and 2–3 cycles with the new method. Total processing time for the new method was about 50% longer than for Gibbs (25 versus 16 seconds on a 133 Mhz Pentium machine). In the new method, candidates were accepted at an average rate of 49%.

My second example uses the growth data of Pothoff and Roy (1964). This well known dataset has four measurements taken at two-year intervals on 11 girls and 16 boys. I applied a linear growth model with random slopes, random intercepts, and fixed effects for gender and gender $\times$ time. The prior distributions for σ^2 and ψ (2×2) were centered near their RML estimates with 3 and 4 degrees of freedom, respectively. Figure 2 shows ACFs pertaining to the log-variance of the random intercept estimated from 5,000 cycles of each algorithm (plots for other variance parameters were similar to these). Dependence appears to die down by about 4 cycles with the new method, compared to more than 10 cycles with Gibbs. Again, the average per-iteration processing time for the new algorithm was about 50% longer than for Gibbs, with an average acceptance rate of 38%.

The final example simulates a situation where conventional Gibbs samplers perform poorly. I generated responses under a one-way random-effects model $y_{ij} = \mu + b_i + \epsilon_{ij}$, $b_i \sim N(0, \psi)$, $\epsilon_{ij} \sim N(0, \sigma^2)$ for $m = 200$ units with $n_i = 2$ observations per unit, with variance components $\psi = 10$ and $\sigma^2 = 90$. The

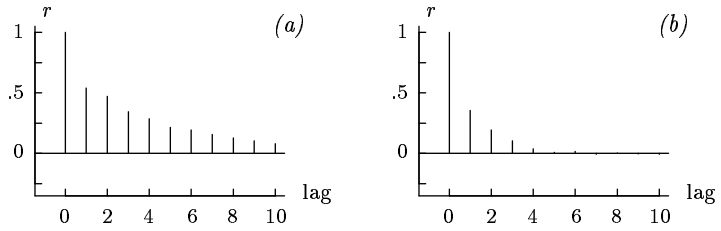

Figure 2: *Sample autocorrelation r for the log-variance of the random intercept under (a) the modified Gibbs sampler and (b) the new MCMC method, using data from Pothoff and Roy (1964).*

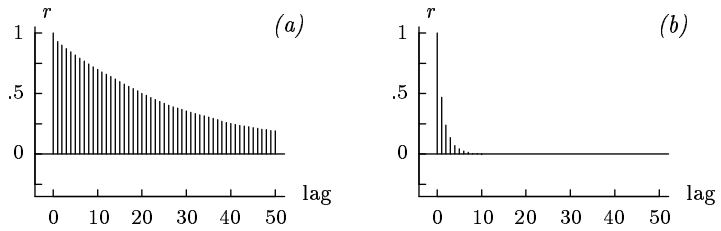

Figure 3: *Sample autocorrelation r for* $\log\psi$ *under (a) the modified Gibbs sampler and (b) the new MCMC method, using simulated data.*

priors for ψ and σ^2 were centered near their true values with three degrees of freedom apiece, $\psi \sim 30\chi_3^{-2}$ and $\sigma^2 \sim 270\chi_3^{-2}$. ACFs for $\log\sigma^2$ and $\log\psi$ estimated from 10,000 cycles of each algorithm are shown in Figure 3. Under the new algorithm the ACF dies down in 8–10 cycles, but Gibbs shows dependence well beyond lag 50. The average MH acceptance rate was 56%.

6 Extensions and future work

The formulas in Section 2, especially the derivatives of $\log L_1$, lead to appealing procedures for this limited but important class of models. With some additional work, I believe that these methods can be extended in a number of useful ways. For example, I have assumed that the matrices $V_1, \ldots, V_m$ are known. Models where V_i depends on unknown parameters are also important, particularly in longitudinal applications with a large number of observations per unit. Derivatives of the loglikelihood with respect to these additional parameters are not difficult to obtain and can be easily incorporated into scoring and MH algorithms, as well as the corrected empirical Bayes procedures of Section 4.

The methods can also be extended to more complicated models with random effects for additional levels of nesting. For example, it is quite common for longitudinal data to be collected on units nested within larger units—e.g. repeated measures for students within classrooms. Adding more random terms to the linear model introduces additional variance parameters which require additional derivatives. When investigating the large-sample properties of these models, it is especially important to carefully define an appropriate asymptotic sequence.

The algorithms in this article seem to perform well because each is a hybrid combination of two distinct methods; one method is rapidly converging but possibly unstable, whereas the other may converge slowly but is very reliable. At any cycle, the ML and RML algorithms of Section 3 revert to EM if Fisher scoring fails. Similarly, the MCMC algorithm of Section 5 reverts to Gibbs sampling whenever a Metropolis-Hastings candidate is rejected. The crucial feature of linear mixed-effects models that allows us to create these attractive hybrid algorithms is that the loglikelihood function can be evaluated at each cycle without much difficulty. Hybrid algorithms would be difficult to create for mixed-effects models with non-normal error distributions (e.g. logistic regression with random effects) because the likelihood for those models is much harder to compute. Other models where the loglikelihood is available include traditional factor analysis and latent-class models for categorical data. Because EM and MCMC for these models can be notoriously slow to converge, it may well be worthwhile to develop hybrid algorithms for these situations as well.

7 References

Bryk, A.S., Raudenbush, S.W., and Congdon, R.T. (1996) *Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs*, Scientific Software International, Inc., Chicago.

Carlin, B.P. (1996) Hierarchical longitudinal modelling. *Markov Chain Monte Carlo in Practice* (eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter), 303–319, Chapman & Hall, London.

Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman & Hall, New York.

DeGroot, M.H. (1970) *Optimal Statistical Decisions*, McGraw-Hill, New York.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1–38.

Dempster, A.P., Rubin, D.B., and Tsutakawa, R.K. (1981) Estimation in covariance components models. *Journal of the American Statistical Association*, **76**, 341–353.

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.

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.

Gelman, A. (1992) Iterative and non-iterative simulation algorithms. *Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface*, 433–438, Interface Foundation of North America, Fairfax, VA.

Gelman, A. and Rubin, D.B. (1992) A single series from the Gibbs sampler provides a false sense of security. *Bayesian Statistics* (eds. J.M. Bernardo, J.O. Bergen, A.P. Dawid and A.F.M. Smith), 627–633, Oxford University Press.

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

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., eds. (1996), *Markov-Chain Monte Carlo in Practice*. Chapman & Hall, London.

Hartley, H.O. and Rao, J.N.K. (1967) Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**, 93–108.

Harville, D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–340.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Jennrich, R.I. and Schluchter, M.D. (1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **38**, 967–974.

Laird, N.M., Lange, N. and Stram, D. (1987) Maximum 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**, 963–974.

Lange, K. (1995) A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, **5**, 1–18.

Lindstrom, M. J. and Bates, D.M. (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, **83**, 1014–1022.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996) *SAS System for Mixed Models.* SAS Institute, Inc., Cary, NC.

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

Liu, J., Wong, W.H. and Kong, A. (1994) Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and sampling schemes. *Biometrika*, **81**, 27–40.

Liu, C. and Rubin, D.B. (1995) Application of the ECME algorithm and the Gibbs sampler to general linear mixed models. *Proceedings of the* 17*th International Biometric Conference*, **1**, 97–107.

MathSoft, Inc. (1997) *S-PLUS User's Guide*, Data Analysis Products Division, MathSoft, Seattle, WA.

Meng, X.L. and van Dyk, D. (1997) Fast EM-type implementations for mixed-effects models. *Journal of the Royal Statistical Society*, to appear.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087–1092.

Multilevel Models Project (1996) *Multilevel Modeling Applications - a Guide for Users of MLn.* (ed. Geoff Woodhouse) Institute of Education, University of London.

Neyman, J. and Scott, E.L. (1948) Consistent estimates based on partially consistent observations, *Econometrica*, **16**, 1–32.

Pothoff, R. and Roy, S.N. (1964) A generalized multivariate analysis of variance model especially useful for growth curve problems, *Biometrika*, **51**, 313–326.

Roberts, G.O. (1996) Markov chain concepts related to sampling algorithms. *Markov Chain Monte Carlo in Practice* (eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter), 45–57, Chapman & Hall, London.

Schott, J.R. (1997) *Matrix Analysis for Statistics.* J. Wiley & Sons, New York.

Smith, A.F.M. and Roberts, G.O. (1993) Bayesian Computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B*, **55**, 3–23.

Tanner, M.A. (1993) *Tools for Statistical Inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions.* (Second Edition) Springer-Verlag, New York.

Tierney, L. (1994) Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701–1762.

Tierney, L. (1996) Introduction to general state-space Markov chain theory. *Markov Chain Monte Carlo in Practice* (eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter), 59–74, Chapman & Hall, London.

Weil, A.T., Zinberg, N.E. and Nelson, J.M. (1968) Clinical and psychological effects of marihuana in man. *Science*, **162**, 1234–1242.

Zeger, S.L. and Karim, M.R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

Appendix

Differentiating $l_1 = \log L_1$ is straightforward except for calculating the derivatives of $\log|\Gamma|$ and $\sum_i (y_i - X_i\tilde{\beta})^T W_i (y_i - X_i\tilde{\beta})$ with respect to $\omega_1, \ldots, \omega_g$. Using $\Gamma = \left(\sum_i X_i^T W_i X_i\right)^{-1}$, $W_i = V_i^{-1} - V_i^{-1} Z_i U_i Z_i^T V_i^{-1}$, and $\partial U_i / \partial \omega_j = -U_i G_j U_i$, it follows that $\partial \Gamma / \partial \omega_j = -\Gamma \left(\sum_i \gamma_i^T U_i G_j U_i \gamma_i\right) \Gamma$ and

$$\begin{aligned} -\frac{\partial}{\partial \omega_j} \log|\Gamma| &= \mathrm{tr}\left(\sum_i \gamma_i^T U_i G_j U_i \gamma_i\right) \Gamma \\ &= \mathrm{tr}\left(\sum_i U_i \gamma_i \Gamma \gamma_i^T U_i\right) G_j \\ &= \sum_i \mathrm{tr} A_i G_j, \end{aligned}$$

where $\gamma_i = Z_i^T V_i^{-1} X_i$. Differentiating again with respect to ω_k gives

$$\frac{\partial^2}{\partial \omega_j \partial \omega_k} \log|\Gamma| = 2\sum_i \mathrm{tr} A_i G_j U_i G_k + \mathrm{tr}\left(\sum_i \gamma_i^T U_i G_j U_i \gamma_i\right) \Gamma \left(\sum_i \gamma_i^T U_i G_k U_i \gamma_i\right) \Gamma,$$

but because $\Gamma = O(m^{-1})$, these terms are $O(1)$ and may be ignored.

To differentiate the quadratic form, use the identity

$$\sum_i (y_i - X_i\tilde{\beta})^T W_i (y_i - X_i\tilde{\beta}) = \left(\sum_i y_i^T W_i y_i\right) - \left(\sum_i X_i^T W_i y_i\right)^T \tilde{\beta},$$

which gives

$$\begin{aligned}\frac{\partial}{\partial\omega_j}\textstyle\sum_i(y_i-X_i\tilde\beta)^TW_i(y_i-X_i\tilde\beta) &= \left(\textstyle\sum_i\lambda_i^TU_iG_jU_i\lambda_i\right)-\left(\textstyle\sum_i\gamma_i^TU_iG_jU_i\lambda_i\right)^T\tilde\beta \\ &\quad -\left(\textstyle\sum_i X_i^TW_iy_i\right)^T\frac{\partial\tilde\beta}{\partial\omega_j}, \end{aligned} \tag{77}$$

where $\lambda_i = Z_i^TV_i^{-1}y_i$. But

$$\begin{aligned}\frac{\partial\tilde\beta}{\partial\omega_j} &= -\Gamma\left(\textstyle\sum_i\gamma_i^TU_iG_jU_i\gamma_i\right)\tilde\beta+\Gamma\left(\textstyle\sum_i\gamma_i^TU_iG_jU_i\lambda_i\right) \\ &= \Gamma\left(\textstyle\sum_i\gamma_i^TU_iG_j\tilde b_i\right),\end{aligned}$$

and with algebraic manipulation (77) reduces to

$$\frac{\partial}{\partial\omega_j}\textstyle\sum_i(y_i-X_i\tilde\beta)^TW_i(y_i-X_i\tilde\beta) = \operatorname{tr}\left(\textstyle\sum_i\tilde b_i\tilde b_i^T\right)G_j.$$

For the second derivative, note that

$$\frac{\partial}{\partial\omega_k}\operatorname{tr}\left(\textstyle\sum_i\tilde b_i\tilde b_i^T\right)G_j = 2\displaystyle\sum_i\tilde b_i^TG_j\left(\frac{\partial\tilde b_i}{\partial\omega_k}\right). \tag{78}$$

But $\tilde b_i = \hat b_i - U_i\gamma_i(\tilde\beta-\beta)$ where $\hat b_i = U_iZ_i^TV_i^{-1}(y_i-X_i\beta)$, which gives

$$\begin{aligned}\frac{\partial\tilde b_i}{\partial\omega_k} &= -U_iG_k\hat b_i+U_iG_kU_i\gamma_i(\tilde\beta-\beta)-U_i\gamma_i\left(\frac{\partial\tilde\beta}{\partial\omega_k}\right) \\ &= -U_iG_k\tilde b_i+-U_i\gamma_i\left(\frac{\partial\tilde\beta}{\partial\omega_k}\right).\end{aligned} \tag{79}$$

Moreover, it can be shown that $E(\tilde b_i\tilde b_i^T) = \sigma^2(\xi-U_i-A_i)$ and $E(\tilde b_{i'}\tilde b_i^T) = \sigma^2U_{i'}\gamma_{i'}\Gamma\gamma_iU_i$ for $i\neq i'$. Substituting (79) into (78), taking expectations, and dropping terms smaller than $O(m)$ produces

$$E\left(\frac{\partial^2}{\partial\omega_j\partial\omega_k}\textstyle\sum_i(y_i-X_i\tilde\beta)^TW_i(y_i-X_i\tilde\beta)\right) \approx -2\sigma^2\displaystyle\sum_i\operatorname{tr}(\xi-U_i)G_jU_iG_k.$$

The RML estimation procedure of Section 3 is easily modified to find the posterior mode for η, which is equivalent to maximizing the density $P(\sigma^{-2},\xi^{-1}\mid y)$. To do this, replace (38) by

$$\sigma^{2\,(t+1)}_{ECME} = \frac{a'^{(t)}+\operatorname{tr}D^{-1}\xi^{(t)^{-1}}}{N-p+b+cq-2}, \tag{80}$$

where $a'^{(t)} = a+\sum_{i=1}^m(y_i-X_i\tilde\beta^{(t)})^TW_i^{(t)}(y_i-X_i\tilde\beta^{(t)})$, and replace (41) by

$$\xi^{(t+1)}_{ECME} = \left(\frac{1}{m+c-q-1}\right)\left[\sigma^{-2\,(t)}D^{-1}+\sum_{i=1}^m\left(\sigma^{-2\,(t)}\tilde b_i^{(t)}\tilde b_i^{(t)T}+U_i^{(t)}+A_i^{(t)}\right)\right]. \tag{81}$$

For the scoring step, replace $(N-p)$ with $(N-p+b+cq-2)$ in the expressions for c_{00} and d_0, m with $(m+c-q-1)$ in the expression for d_j, and (42)–(43) with

$$\begin{aligned} c_{0j} &= \frac{\sigma^{2(t)}}{2}\sum_{i=1}^{m}\mathrm{tr}(\xi^{(t)}-U_i^{(t)})G_j+\frac{1}{2}\,\mathrm{tr}D^{-1}G_j, \\ c_{jk} &= \frac{1}{2}\sum_{i=1}^{m}\mathrm{tr}(\xi^{(t)}-U_i^{(t)})G_j(\xi^{(t)}-U_i^{(t)})G_k+\left(\frac{c-q-1}{2}\right)\mathrm{tr}\xi G_j\xi G_k. \end{aligned}$$

At each cycle, the log-posterior density may be evaluated using (76).

.