Statistics and Computing (1999) **9**, 17–26

On MCMC sampling in hierarchical longitudinal models

SIDDHARTHA CHIB

John M. Olin School of Business, Washington University, One Brookings Drive, St. Louis, Missouri 63130, USA

BRADLEY P. CARLIN

Division of Biostatistics, School of Public Health, University of Minnesota, Box 303 Mayo Building, Minneapolis, Minnesota 55455, USA

Submitted December 1997 and accepted November 1998

Markov chain Monte Carlo (MCMC) algorithms have revolutionized Bayesian practice. In their simplest form (i.e., when parameters are updated one at a time) they are, however, often slow to converge when applied to high-dimensional statistical models. A remedy for this problem is to *block* the parameters into groups, which are then updated *simultaneously* using either a Gibbs or Metropolis-Hastings step. In this paper we construct several (partially and fully blocked) MCMC algorithms for minimizing the autocorrelation in MCMC samples arising from important classes of longitudinal data models. We exploit an identity used by Chib (1995) in the context of Bayes factor computation to show how the parameters in a general linear mixed model may be updated in a *single* block, improving convergence and producing essentially independent draws from the posterior of the parameters of interest. We also investigate the value of blocking in non-Gaussian mixed models, as well as in a class of binary response data longitudinal models. We illustrate the approaches in detail with three real-data examples.

Keywords: Blocking, correlated binary data, convergence acceleration, Gibbs sampler, Metropolis-Hastings algorithm, linear mixed model, panel data, random effects

1. Introduction

Consider the Gaussian linear mixed model (Laird and Ware, 1982),

$$y_i = X_i\beta + W_i b_i + \varepsilon_i$$
$$b_i \sim \mathcal{N}_q(0, D)$$

where the y_i are vectors of length n_i containing the observations on the i^{th} unit, and the ε_i are error vectors of the same length independently distributed as $\mathcal{N}_{n_i}(0, \sigma^2 I_{n_i})$, $i = 1, \ldots, k$. In this mixed model, X_i is an $n_i \times p$ design matrix of covariates and β is a corresponding $p \times 1$ vector of fixed effects. In addition, W_i is a $n_i \times q$ design matrix (q typically less than p), and b_i is a $q \times 1$ vector of subject-specific random effects. The b_i model the subject-specific means, and enable the model to capture marginal dependence among the observations on the i^{th} unit. The hierarchical specification of this model is completed by adding the prior distributions $D^{-1} \sim \mathcal{W}(v_0^{-1}R_0, v_0)$, $\sigma^{-2} \sim \mathcal{G}(v_{00}/2, \delta_{00}/2)$, and $\beta \sim \mathcal{N}_p(\beta_0, B_0)$, where $\mathcal{W}$ denotes the Wishart distribution and $\mathcal{G}$ denotes the gamma distribution. In our parametrization, the Wishart prior has mean R_0 while the gamma prior has mean v_{00}/δ_{00}.

This model lends itself to a full Bayesian analysis by Markov chain Monte Carlo (MCMC) methods. One of the first such algorithms was proposed by Gelfand and Smith (1990) which we summarize as follows:

Algorithm 1

```
1. Sample $\beta$ from $\beta|y, b, \sigma^2, D$
2. Sample $b$ from $\{b_i\}|y, \beta, \sigma^2, D$
3. Sample $D^{-1}$ from $D^{-1}|y, \beta, b, \sigma^2$
4. Sample $\sigma^2$ from $\sigma^2|y, \beta, b, D$
```

5. *Repeat Steps 1–4 using the most recent values of the conditioning variables.*

The Gaussian-linear structure of our model, combined with the conditional conjugacy of our prior specification, means that all of these *full conditional* distributions are easily available in closed form (as normal, normal, Wishart, and inverse gamma, respectively). This Gibbs sampling scheme has been implemented by several authors in longitudinal modeling applications; see for example Lange *et al.* (1992), Carlin (1996), and Carlin and Louis (1996, Sec 8.1).

It is now recognized, however, that Algorithm 1, while relatively easy to implement, can suffer from slow convergence if the parameters are highly correlated *a posteriori*, or if the information in the likelihood and prior is insufficient to completely determine the model parameters. In fact, the latter situation arises automatically in model (1) if the priors on both variance components (D and σ^2) are overly vague, since the data can inform about them only corporately, not independently. Gelfand, Sahu and Carlin (1995, 1996) suggested hierarchical centering of the random effects in such models to reduce serial correlations, later extending the idea to generalized (non-Gaussian) linear mixed models. These authors show the idea to work well when the variance at the second stage (D) dominates that at the first (σ^2), the usual case in hierarchical modeling. Vines, Gilks and Wild (1996) and Gilks and Roberts (1996) instead recommend a "sweeping" reparametrization, to break serial correlations by reducing the dimension of the model space (analogous to the usual frequentist practice of adding identifiability constraints to ANOVA models). Gelfand and Sahu (1999) build on the definition of Dawid (1979) to discuss Bayesian identifiability more formally, and go on to endorse "recentering on the fly," i.e., imposing identifiability constraints numerically at the end of each MCMC iteration, as a simpler but equivalent alternative to sweeping. Such recentering algorithms have long been in use in MCMC analyses of models employing pairwise difference priors, which are identified only up to an additive constant and commonly used in spatial analyses; see Besag *et al.* (1995).

The purpose of this paper is to develop new approaches to the MCMC simulation of longitudinal models that provide significant improvements over Algorithm 1 and its refinements mentioned above. These approaches rely on the use of *blocking*, i.e. updating the parameters in groups, as the means to reduce the serial correlation in the simulation output. The idea of blocking has strong intuitive appeal and theoretical support (Liu, 1994; Liu, Wong, and Kong, 1994), though Roberts and Sahu (1997, Sec. 2.4) give two examples showing blocking is not *guaranteed* to improve the convergence rate of a sampler (such situations appear rare relative to those in which blocking does lead to improvement). In particular, we show that more coarse blocking of model (1) is possible and that, in fact, the parameters of the model may be updated in a *single* block, greatly improving convergence and producing essentially independent draws from the posterior distribution of interest.

Note that our strategy is to expend slightly more analytic and coding effort to obtain a sampler which will produce less highly autocorrelated draws, hence shorter runtimes. Some authors (e.g. Damien *et al.*, 1998) have recommended precisely the opposite strategy, namely *augmenting* the (univariate) sampling order with certain carefully-chosen auxiliary variables, obtaining an algorithm which is easier to state and code but takes longer to converge. While both approaches have their merits, we view ours as more promising for implementation in generalist software for solving the broad class of models we consider; our code may be more difficult to write, but once written, will be faster and easier to use for many different problems.

Section 2 lays out our approach for linear Gaussian mixed models, the kind most commonly occurring in longitudinal data analysis. An extension to non-Gaussian (e.g. Student's t) error distributions and hierarchically centered model formulations is also described. Section 3 examines the value of blocking in the case of discrete longitudinal data. In Section 4 we provide three numerical examples which illustrate the benefit of our blocking schemes and include comparisons with results obtained from the BUGS software package (Spiegelhalter *et al.*, 1995). Finally, Section 5 discusses our findings and offers directions for future research in this area.

2. Blocking for Gaussian mixed models

We begin our investigation into the value of blocking in longitudinal models by considering the distribution of y_i marginalized over the random effects. Due to the conditional Gaussian structure we have that

$$y_i|\beta, \sigma^2, D \sim \mathcal{N}_{n_i}(X_i\beta, V_i)$$

where $V_i = \sigma^2 I + W_i D W_i'$. This implies that the posterior distribution of β conditioned on σ^2 and D (but not on $\{b_i\}$) is (Lindley and Smith, 1972)

$$\beta|y, \sigma^2, D \sim \mathcal{N}_p(\hat{\beta}, B)$$

where

$$\hat{\beta} = B\left(B_0\beta_0 + \sigma^{-2}\sum_{i=1}^{n} X_i' V_i^{-1} y_i\right)$$

and

$$B = \left(B_0 + \sigma^{-2}\sum_{i=1}^{n} X_i' V_i^{-1} X_i\right)^{-1}.$$

As a consequence we immediately note that it is possible to sample the fixed effects β and the random effects $\{b_i\}$ in one block, but retain the essential Gibbs structure, as follows:

Algorithm 2

1. *Sample β and b from $\beta, \{b_i\}|y, \sigma^2, D$ by sampling*
   (a) *β from $\beta|y, \sigma^2, D$*
   (b) *b from $\{b_i\}|y, \beta, \sigma^2, D$*
2. *Sample D^{-1} from $D^{-1}|y, \beta, b, \sigma^2$*
3. *Sample σ^2 from $\sigma^2|y, \beta, b, D$*
4. *Repeat Steps 1–3 using the most recent values of the conditioning variables.*

Except for the change in the sampling of β, this scheme is identical to that in Algorithm 1. This minor refinement is practically quite important, however, and improves the behavior of the MCMC output. Besides, it requires no hierarchical centering because β is sampled without conditioning on the random effects and the entire sampling is still from tractable distributions (albeit with a bit more matrix algebra).

While Algorithm 2 is an improvement on Algorithm 1, it does not address the correlation between D^{-1} and b that can lead to slow mixing for the unique elements of D^{-1} (there is usually no mixing problem for σ^2). To deal with this problem we suggest an approach that allows one to sample all parameters in one block from the joint posterior distribution. That this is possible has not been recognized in the literature before. The idea is to use the following decomposition of the posterior distribution

$$\pi(\sigma^2, D^{-1}, \beta, \{b_i\}|y) = \pi(\sigma^2, D^{-1}|y) \\ \pi(\beta|y, \sigma^2, D)\pi(\{b_i\}|y, \beta, \sigma^2, D)$$

where the last two densities are the same as in Algorithm 2 but the first density is not in closed form but can be updated by the Metropolis-Hastings algorithm (see for example Hastings, 1970, or Chib and Greenberg, 1995). By definition,

$$\pi(\sigma^2, D^{-1}|y) \propto \pi(\sigma^2, D^{-1}) f(y|\sigma^2, D)$$

where

$$\begin{aligned} f(y|\sigma^2, D) &= \int f(y|\beta, \sigma^2, D)\pi(\beta)\mathrm{d}\beta \\ &\propto |\Omega|^{-1/2} \exp\{(y - X\beta_0)'\Omega^{-1}(y - X\beta_0)\}, \end{aligned}$$

$y = (y_1', \ldots, y_n')', X = (X_1', \ldots, X_n')', \Omega = (XB_0^{-1}X' + V)$ and $V = \mathrm{diag}(V_1, \ldots, V_n)$. One way to evaluate this density is to recognize that $f(y|\sigma^2, D)$ is the normalizing constant of $\pi(\beta|y, \sigma^2, D)$. A similar idea is used by Chib (1995) in his approach to find the marginal likelihood of the model. Hence, we may write $f(y|\sigma^2, D)$ as a ratio of three terms

$$f(y|\sigma^2, D) = \frac{\pi(\beta^*) f(y|\beta^*, \sigma^2, D)}{\pi(\beta^*|y, \sigma^2, D)} \tag{2}$$

$$= \frac{\phi_p(\beta^*|\beta_0, B_0) \prod_{i=1}^n \phi_{n_i}(y_i|X_i\beta^*, V_i)}{\phi_p(\beta^*|\hat{\beta}, B)} \tag{3}$$

where β^* is any point (preferably a high density point such as the posterior mean from Algorithm 2) and $\phi_p(t|\mu, \Sigma)$ is density of the p-variate normal distribution with mean vector μ and covariance matrix Σ. Each term in this expression is easy to evaluate. This leads to the following *single block* algorithm for sampling the posterior density of the Gaussian hierarchical model.

Algorithm 3

Setup:

- *Run Algorithm 2 for $G = 500$ iterations (say) and let $\beta^* = G^{-1}\sum_{g=1}^G \beta^{(g)}$. Also let $\mu = G^{-1}\sum_{g=1}^G \theta^{(g)}$ and $\Sigma = G^{-1}\sum_{g=1}^G (\theta^{(g)} - \mu)(\theta^{(g)} - \mu)'$, where $\theta = (\sigma^2, \psi)$, and $\psi = \mathrm{vech}(D^{-1})$ denotes the unique elements of D^{-1}.*

Start of algorithm:

1. *Sample θ, β and b from $[\theta, \beta, b|y]$ by sampling*

   (a) *θ from $\pi(\theta|y)$ using the Metropolis-Hastings algorithm with proposal density given by $q(\theta) = f_{MVT}(\theta|\mu, \tau^2\Sigma, v)$, where f_{MVT} is the multivariate-t density with v degrees of freedom, and τ^2 and v are tuning parameters. Given the current value θ^c, first draw θ^t from $q(\theta)$ and move to the point θ^t with probability given by*

$$\alpha(\theta^c, \theta^t) = \min\left\{1, \frac{f(y|\sigma^{2t}, D^t)\pi(\sigma^{2t}, D^t)q(\sigma^{2c}, D^c)}{f(y|\sigma^{2c}, D^c)\pi(\sigma^{2c}, D^c)q(\sigma^{2t}, D^t)}\right\}.$$

   (b) *Sample β from $\mathcal{N}_p(\hat{\beta}, B)$ where $\hat{\beta} = B(B_0\beta_0 + \sigma^{-2}\sum_{i=1}^n X_i'V_i^{-1}y_i)$ and $B = (B_0 + \sigma^{-2}\sum_{i=1}^n X_i'V_i^{-1}X_i)^{-1}$.*

   (c) *Sample b_i independently from $\mathcal{N}_q(\hat{b}_i, C_i)$ where $\hat{b}_i = C_i(\sigma^{-2}W_i'(y_i - X_i\beta))$ and $C_i = (D^{-1} + \sigma^{-2}W_i'W_i)$.*

2. *Repeat Step 1 using the most recent values of the conditioning variables.*

Turning to extensions of our approach, we first note that the above approach can be extended to several symmetric but nonnormal error distributions using the idea of normal scale mixtures (Andrews and Mallows, 1974). Several such alternative error densities are available in this way; see e.g. Carlin and Louis (1996, p.210). For example, to obtain errors that are Student's t with v degrees of freedom, we simply replace the $\mathcal{N}(0, \sigma^2)$ specifications for the ε_{ij} with the two-part specification

$$\varepsilon_{ij}|\lambda_{ij} \sim \mathcal{N}(0, \lambda_{ij}^{-1}\sigma^2), \ \lambda_{ij} \sim \mathcal{G}(v/2, v/2). \tag{4}$$

Thus conditional on $\boldsymbol{\lambda} = \{\lambda_{ij}\}$, $f(y_i|\beta, b, \sigma^2, D, \boldsymbol{\lambda})$ still emerges as normal, hence so does the marginal density $f(y_i|\beta, \sigma^2, D, \boldsymbol{\lambda})$. Further, the full conditional distributions of the λ_{ij} are gamma, so implementation of Algorithm 2 above is straightforward. Regarding Algorithm 3, the approach of equation (2) now produces a closed form for $f(y|\sigma^2, D, \boldsymbol{\lambda})$ but the high dimension of $\boldsymbol{\lambda}$ likely renders infeasible a single multivariate Metropolis-Hastings update for $(\sigma^2, \psi), \boldsymbol{\lambda})$. Instead, we might use the multivariate t M-H update proposed above for $(\sigma^2, \psi|\boldsymbol{\lambda})$, and univariate M-H updates for the $(\lambda_{ij}|\sigma^2, D)$ – say, Hastings independence chains based on Gamma proposals centered at one, roughly the center of the mixing distribution in (4).

As a final extension, we observe that hierarchically centered Gaussian structures of the kind advocated by Gelfand *et al.* (1995) are easily handled within our framework. For example, in the case of two-stage models this centering takes the form

$$y_i = X_i b_i + \varepsilon_i$$
$$b_i \sim \mathcal{N}_p(A_i\beta, D),$$

with similar prior structure as above. Analysis proceeds in much the same way as described above; we omit the details.

3. Longitudinal binary probit models

In this section we consider various blocking schemes for the class of probit longitudinal binary random effects models. A Bayesian analysis of these models using a version of Algorithm 1 is provided by Albert and Chib (1996), and by Zeger and Karim (1991) under the logit link.

Consider a sequence of binary measurements $Y_i = (y_{i1}, \ldots, y_{in_i})', y_{it} \in \{0, 1\}$ on the ith unit taken at n_i specific time points. Let the probability of obtaining a positive response on the ith unit at occasion $t(1 \le i \le n, 1 \le t \le n_i)$ be given by the function

$$\Pr(y_{it} = 1|b_i) = \Phi(x_{it}'\beta + w_{it}'b_i), \tag{5}$$

where Φ is the standard normal cdf, x_{it}' and w_{it}' are the tth rows of X_i and W_i, respectively, and b_i as before is $\mathcal{N}_q(0, D)$. For this model, the likelihood contribution $f(y_i|\beta, D)$ is given by

$$\int \left\{ \prod_{t=1}^{n_i} [\Phi(x_{it}'\beta + w_{it}'b_i)]^{y_{it}} [1 - \Phi(x_{it}'\beta + w_{it}'b_i)]^{1-y_{it}} \right\} \phi_q(b_i|0, D)\, \mathrm{d}b_i \tag{6}$$

which is expensive to evaluate when b_i is multi-dimensional. One way to deal with this problem is via a latent variables approach (Albert and Chib, 1993, 1996; Carlin and Polson, 1992). Let z_{it} denote independent latent variables such that

$$z_{it}|b_i \sim \mathcal{N}(x_{it}'\beta + w_{it}'b_i, 1),\ \ 1 \le t \le n_i;\ 1 \le i \le n,$$

and let the observed response y_{it} be given by

$$y_{it} = \begin{cases} 1 & \text{if} \quad z_{it} > 0 \\ 0 & \text{if} \quad z_{it} \le 0 \end{cases}.$$

Then, it can be seen that the y_{it} satisfy model (5). With the introduction of the latent data, the probit model is similar to the Gaussian model above and the posterior distribution of the parameters (β, D) may be sampled in parallel fashion. Let $Z = (Z_1, \ldots, Z_n)$ and $Z_i = (z_{i1}, \ldots, z_{in_i})$ then a MCMC scheme analogous to Algorithm 1 is defined as follows:

Algorithm 4

1. *Sample β from $\beta|Z, b, D$*
2. *Sample b from $\{b_i\}|Z, \beta, D$*
3. *Sample D^{-1} from $D^{-1}|b$*
4. *Sample $\{z_{it}\}$ from $z_{it}|y_{it}, \beta, b, D$*
5. *Repeat Steps 1–4 using the most recent values of the conditioning variables.*

The first three conditional distributions follow the same form as those given above. The last is given by a sequence of independent truncated normal distributions, namely $\mathcal{TN}_{(0,\infty)}(x_{it}'\beta + w_{it}'b_i, 1)$ if $y_{it} = 1$, or $\mathcal{TN}_{(-\infty,0)}(x_{it}'\beta + w_{it}'b_i, 1)$ if $y_{it} = 0$. Albert and Chib (1996) provide further details.

The first refinement to this algorithm is based on marginalizing the distribution of Z_i over the random effects b_i. Then,

$$Z_i \sim \mathcal{N}_{n_i}(X_i\beta, V_i)$$

where now $V_i = I_{n_i} + W_i D W_i'$ and the model becomes a special case of the multivariate probit model analyzed by Chib and Greenberg (1998). The resulting algorithm is similar to Algorithm 4 except that β is sampled from $\beta|Z, D$, and Z_i from the *multivariate normal* distribution $\mathcal{N}_{n_i}(X_i\beta, V_i)$ truncated to the region implied by the *vector* y_i. We follow Chib and Greenberg (1998) and sample this truncated multivariate normal vector from a sequence of (full conditional) univariate truncated normal distributions. Thus, in this case, integrating out the random effects does not lead to a reduction in the number of blocks in the sampling (relative to Algorithm 4). Nonetheless, marginalization over the b_i can be expected to improve the sampling of the fixed effects for the reasons outlined above. We summarize this algorithm as follows:

Algorithm 5

1. *Sample β and $\{Z_i\}$ from $[\beta, \{Z_i\}|y, D]$ by sampling*
   (a) *β from $\beta|y, Z, D$*
   (b) *$\{Z_i\}$ from $Z_i|y_i, \beta, D$*
2. *Sample b from $\{b_i\}|y, Z, \beta, D$*
3. *Sample D^{-1} from $D^{-1}|b$*
4. *Repeat Steps 1–3 using the most recent values of the conditioning variables.*

We can refine this algorithm by sampling D^{-1} from the distribution $\pi(D^{-1}|y, Z)$ by employing the same technique that was used in connection with Algorithm 3. The resulting algorithm, **Algorithm 6**, is then based on the conditional distributions $[D^{-1}|y, Z]$, $[\beta|y, Z, D]$, $[Z_i|y_i, \beta, D]$, and $[b|y, Z, \beta, D]$. The details are similar to those above and are suppressed.

A more interesting refinement of Algorithm 5 (which we refer to as **Algorithm 7**) works with just the single distribution $[\beta, D|y]$. In this case β and D are sampled by marginalizing over *b and Z*. We are now down to just one block in the sampling. The main problem in implementing this algorithm is that the density $[\beta, D|y]$ requires the computation of the likelihood contribution in (6). One way to compute this contribution is by a method that is called the Geweke-Keane-Hajivassiliou method in the econometrics literature [see Chib and Greenberg (1998)]. From the latent variable representation, $f(y_i|\beta, D)$ can be written as

$$\int_{B_{in_i}} \cdots \int_{B_{i1}} \phi_{n_i}(Z_i|X_i\beta, V_i)\,\mathrm{d}Z_i, \tag{7}$$

where B_{it} is the interval $(0, \infty)$ if $y_{it} = 1$, and the interval $(-\infty, 0)$ if $y_{it} = 0$. Let $V_i = LL'$ and make a change of variable from Z_i to ε_i, where L is the lower triangular Choleski factorisation and $Z_i = X_i\beta + L\varepsilon_i$. Then

$$f(y_i|\beta, \Sigma) = \int_{c^*_{in_i}}^{d^*_{in_i}} \cdots \int_{c^*_{i1}}^{d^*_{i1}} \phi_J(t|0, I)\,\mathrm{d}t, \tag{8}$$

where

$$c^*_{it} = \frac{c_{it} - x'_{it}\beta - \sum_{k=1}^{t-1} l_{tk}\varepsilon_{ik}}{l_{tt}},$$
$$d^*_{it} = \frac{d_{it} - x'_{it}\beta - \sum_{k=1}^{t-1} l_{tk}\varepsilon_{ik}}{l_{tt}},$$

and (c_{it}, d_{it}) denotes the lower and upper endpoints of $B_{it}(j \le n_i)$. Now compute the quantity

$$p_i = \prod_{t=1}^{n_i} \left(\Phi(d^*_{it}) - \Phi(c^*_{it})\right)$$

where the end-points are based on ε_{ik} drawn from a $N(0,1)$ distribution truncated to the interval (c^*_{ik}, d^*_{ik}). The quantity p_i is computed afresh for a new sequence of end-points and the calculations are repeated a large number of times. The average of the resulting p_i is the Monte Carlo estimate of $f(y_i|\beta, \Sigma)$.

With the likelihood function computed in this manner, the actual updating of β and $\psi = \mathrm{vech}(D^{-1})$ is then through a Metropolis step with a proposal density that is found from a preliminary set-up run similar to that in Algorithm 3. This procedure is computationally demanding because the likelihood contribution must be computed for each new value of β and D. Nonetheless, in terms of blocking, this is the best that one can do in this context. Algorithm 7 thus provides a benchmark for judging the value of integrating out random effects and latent variables in non-linear longitudinal models.

4. Numerical Illustrations

4.1. *Longitudinal continuous observations*

We illustrate the basic algorithmic approach of Section 2 using continuous longitudinal data from a clinical trial comparing the effectiveness of two antiretroviral drugs (didanosine, *ddI*, and zalcitabine, *ddC*) in 467 persons with advanced HIV infection. The response variable y_{ij} for patient i at time j is the patient's CD4 count, a seriological measure of immune system health and prognostic factor for AIDS-related illness and mortality. These data were originally presented and analyzed by Abrams *et al.* (1994), and subsequently subjected to Bayesian reanalysis by Goldman *et al.* (1996), Carlin (1996), and Carlin and Louis (1996, Sec. 8.1.2). The dataset records patient CD4 counts at study entry and again at 2, 6, 12, and 18 months after entry, though a great many of these observations are missing for many patients (the sample sizes at the five time points for the two drug groups are (230, 182, 153, 102, 22) and (236, 186, 157, 123, 14), for the ddI and ddC groups, respectively).

Following the aforementioned work, we seek to fit model (1) where the j^{th} row of patient i's design matrix W_i takes the form

$$w_{ij} = (1, t_{ij}, (t_{ij} - 2)^+),$$

where $t_{ij} \in \{0, 2, 6, 12, 18\}$ and $z^+ = \max(z, 0)$. Thus the three columns of W_i correspond to individual-level intercept, slope, and possible change in slope after the two month visit (by which time the drugs are expected to produce a detectable benefit). We further account for the effect of two covariates by including them in the fixed effect design matrix X_i. These covariates are d_i, a binary variable indicating whether patient i received ddI ($d_i = 1$) or ddC ($d_i = 0$) and a_i, a binary variable telling whether the patient was diagnosed as having AIDS at baseline ($a_i = 1$) or not ($a_i = 0$). Thus we set

$$X_i = (W_i | d_i W_i | a_i W_i),$$

so that $p = 3q = 9$.

Boxplots of the individual CD4 counts for the two drug groups (not shown) indicate a high degree of skewness toward high CD4 values. This, combined with the count nature of the data, suggests a square root transformation for each group. We complete our model specification with minimally informative priors, taking care to ensure that they do not lead to improper posterior distributions for the variance components σ^2 and D. Following previous work, we set $\nu_0 = 24$ and $R_0 = \mathrm{Diag}(2^2, (.25)^2, (.25)^2)$, which should preserve identifiability while still allowing the random effects a reasonable amount of freedom. For the prior on σ^{-2}, we take a $\mathcal{G}(1, 100)$, so that σ^{-2} has both mean and standard deviation equal to $(1/10)^2$. This specification implies a relatively high error variance, which we expect for our relatively noisy data. Finally, for the prior on β we set

$$\beta_0 = (10, 0, 0, 0, 0, 0, -3, 0, 0), \quad \text{and}$$
$$B_0 = \mathrm{Diag}(2^2, 1^2, 1^2, (.1)^2, 1^2, 1^2, 1^2, 1^2, 1^2),$$

a prior biased strongly away from 0 only for the baseline intercept, β_1, and the intercept adjustment for a positive

AIDS diagnosis, β_7. This prior also forces the drug group intercept (i.e., the effect at baseline) β_4 to be very small, since patients were assigned to drug group at random.

Running our various MCMC algorithms for these data and model for 5000 iterations each produces the correlation summaries in Table 1. The table shows the lag 1 sample autocorrelations for Algorithms 1–3 above, where the results for Algorithm 1 are computed using the BUGS (Spiegelhalter *et al.*, 1995) programming language. We note that in Algorithm 1 the correlations are rather high for D, σ^2, and most elements of β. Fully multivariate hierarchical centering (as recommended for this dataset by Gelfand *et al.*, 1995) is not possible within the current version of BUGS, though the language does support some simpler, univariate centering forms.

Algorithm 2 provides a dramatic reduction in the serial correlation of the components of β. Effectively, we are now sampling iid draws for this parameter. Algorithm 3 shows further improvement in the autocorrelations for σ^2 and the components of D.

While lag 1 autocorrelations are a good predictor of MCMC algorithm performance, they of course do not tell the whole story. To summarize the autocorrelations at all lags and their overall rate of decay, Table 2 gives the *autocorrelation time* $\kappa = 1 + 2\sum_{k=1}^{\infty}\rho(k)$ for each parameter in Table 1, where $\rho(k)$ is the autocorrelation at lag k for the parameter of interest. We estimated κ using the sample autocorrelations estimated from the MCMC chain, cutting off the summation when these dropped below 0.1 in magnitude. The κ values can in turn be used to define *effective sample sizes* (Kass *et al.*, 1998, p. 99) as the MCMC sample size, G, divided by κ. Thus κ can be interpreted as the relative increase in run length necessitated by the Markov dependence.

Table 1. *Lag 1 sample autocorrelations for MCMC algorithms in ddI/ddC data model*

Parameter	Algorithm 1 (BUGS)	Algorithm 2	Algorithm 3
β_1	0.798	−0.006	0.012
β_2	0.194	−0.002	−0.011
β_3	0.207	−0.001	−0.014
β_4	0.204	0.013	−0.005
β_5	0.436	0.009	−0.012
β_6	0.408	0.004	−0.020
β_7	0.811	−0.008	0.006
β_8	0.134	−0.020	−0.006
β_9	0.154	−0.008	−0.010
σ^2	0.530	0.549	0.500
D_{11}	0.388	0.283	0.654
D_{21}	0.942	0.932	0.790
D_{22}	0.891	0.934	0.795
D_{31}	0.934	0.924	0.791
D_{32}	0.967	0.951	0.799
D_{33}	0.918	0.945	0.787

Table 2. *Autocorrelation times $\kappa = 1 + 2\sum_{k=1}^{\infty}\rho(k)$, where $\rho(k)$ is the autocorrelation at lag k for the parameter of interest, for MCMC algorithms in the ddI/ddC data model*

Parameter	Algorithm 1 (BUGS)	Algorithm 2	Algorithm 3
β_1	20.42	1.00	1.00
β_2	1.59	1.00	1.00
β_3	1.41	1.00	1.00
β_4	1.67	1.00	1.00
β_5	3.27	1.00	1.00
β_6	2.52	1.00	1.00
β_7	20.36	1.00	1.00
β_8	1.27	1.00	1.00
β_9	1.31	1.00	1.00
σ^2	4.32	4.23	4.81
D_{11}	3.88	3.57	4.26
D_{21}	60.13	28.11	10.87
D_{22}	40.55	27.20	11.53
D_{31}	55.33	26.40	9.20
D_{32}	53.05	30.05	11.55
D_{33}	42.46	28.64	8.71

The autocorrelation times in Table 2 reveal an essentially similar pattern to that for the lag 1 autocorrelation in Table 1. Algorithm 1 has substantial autocorrelation times for almost all parameters. Algorithms 2–3 offer dramatic improvements as mentioned above, with both producing near perfect output for β and Algorithm 3 emerging as the most efficient for D.

4.2. *Longitudinal binary observations, single random effect*

Our numerical illustration for this model considers a subset of data from the Six Cities study, a longitudinal study of the health effects of air pollution (see e.g. Fitzmaurice and Laird (1993) for the data and a likelihood-based analysis). The data consist of repeated binary measurements y_{ij} of the wheezing status (1 = yes, 0 = no) of child i at time j, $i = 1, \ldots, I, : j = 1, \ldots, J$, for each of $I = 537$ children living in Stuebenville, Ohio at $J = 4$ timepoints. We are given two predictor variables: a_{ij}, the age of child i in years at measurement point j (7, 8, 9, or 10 years), and s_i, the smoking status of child i's mother (1 = yes, 0 = no). Following the Bayesian analysis of Chib and Greenberg (1998), we adopt the conditional response model

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ p_{ij} &\equiv Pr(Y_{ij} = 1) = g^{-1}(\mu_{ij}) \\ \mu_{ij} &= \beta_0 + \beta_1 z_{ij1} + \beta_2 z_{ij2} + \beta_3 z_{ij3} + b_i, \end{aligned} \tag{9}$$

where $z_{ijk} = (x_{ijk} - \bar{x}_{..k}), : k = 1, 2, 3$, and $x_{ij1} = a_{ij}$, $x_{ij2} = s_i$, and $x_{ij3} = a_{ij}s_i$, a smoking-age interaction term. The b_i are individual-specific random effects, initially given an exchangeable $N(0, D)$ specification, which allow for dependence among the longitudinal responses for child i. We adopt the probit link for $g(.)$, i.e.,

$$g(p_{ij}) = \text{probit}(p_{ij}) = \Phi^{-1}(p_{ij}),$$

so that sampling for the latent data may proceed as described in Section 3 above.

This time we used runs of $G = 10\,000$ iterations from each algorithm, after placing flat priors on the β_k and a vague $\mathcal{G}(.001, .001)$ prior on D^{-1}. Table 3 gives the lag 1 sample correlation summaries, while Table 4 does the same for lag 5. Here, "BUGS 0" is a BUGS run using the uncentered covariates (the x_{ijk}), while "BUGS 1" uses the centered covariates (the z_{ijk}). The correlations are marginally smaller under the centered BUGS parametrization, but not much; the sample autocorrelations for D and β_0 persist almost to lag 30. This is also reflected in the rather high autocorrelation times for these two parameters shown in Table 5.

The results for Algorithms 5, 6, and 7 are also shown in Table 3, 4, and 5. Algorithm 7 performs the best, as expected, though the benefit it offers is much more apparent at lag 5 than at lag 1. Given its longer runtimes, perhaps Algorithm 5 is actually preferable in this case. We also note that the improvements offered by our blocked algorithms over BUGS 1 are not as dramatic as in the previous example. The reason for this is that BUGS is actually updating a smaller (marginalized, or "collapsed") parameter space, namely (β, D, b) instead of (β, D, b, Z). That is, despite the probit structure, BUGS collapses over the missing data and uses adaptive rejection sampling (Gilks and Wild, 1992), rather than conjugacy, to draw the necessary samples. In this way BUGS can handle the logit or complementary log-log link as easily as the probit, changes which destroy the conjugate structure for the missing data formulation. However, the sampling in BUGS is much more complicated, and requires the adaptive tuning provided by the software to be feasible. Also, if the response variable had more than two categories (ordinal response, rather than binary), the problem could not be handled by the current version of BUGS at all.

Table 5. *Autocorrelation times κ, algorithms for the Six Cities data model*

Parameter	BUGS 0	BUGS 1	Algorithm 5	Algorithm 7
β_0	19.81	16.02	12.63	5.02
β_1	3.18	2.21	5.56	4.87
β_2	9.43	5.59	3.56	5.10
β_3	3.40	2.38	5.21	5.12
D	21.55	15.39	18.49	6.30

4.3. *Longitudinal binary observations, multiple random effects*

Our final numerical illustration is taken from the University of Michigan's Panel Survey of Income Dynamics, a sample of 520 households observed over the period 1976–1982. Here the response y_{ij} is the labor force participation decision for woman i at survey point j, where all women surveyed were between the ages of 25 and 62. The two predictor variables are x_{ij1}, the woman's education in number of grades completed, and x_{ij2}, total family income excluding the woman's earnings in thousands of dollars. As in the previous subsection we adopt the longitudinal probit model (9), but now using two random effects, namely

$$\mu_{ij} = \beta_0 + \beta_1 z_{ij1} + \beta_2 z_{ij2} + b_{i1} + b_{i2} z_{ij2},$$

where again $z_{ijk} = (x_{ijk} - \bar{x}_{..k}), : k = 1, 2$, and we include family-specific random intercepts b_{i1} and income slopes b_{i2}, respectively.

Table 3. *Lag 1 sample autocorrelations, algorithms for the Six Cities data model*

Parameter	BUGS 0	BUGS 1	Algorithm 5	Algorithm 6	Algorithm 7
β_0	0.873	0.807	0.663	0.529	0.533
β_1	0.573	0.385	0.741	0.731	0.531
β_2	0.827	0.722	0.450	0.356	0.473
β_3	0.530	0.486	0.683	0.767	0.541
D	0.882	0.857	0.879	0.811	0.796

Table 4. *Lag 5 sample autocorrelations, algorithms for the Six Cities data model*

Parameter	BUGS 0	BUGS 1	Algorithm 5	Algorithm 6	Algorithm 7
β_0	0.570	0.448	0.445	0.357	0.134
β_1	0.006	−0.006	0.118	0.219	0.041
β_2	0.370	0.189	0.118	0.082	0.115
β_3	0.079	0.015	0.032	0.215	0.039
D	0.632	0.567	0.660	0.559	0.507

Table 6. *Lag 1 sample autocorrelations, algorithms for the labor force participation data model*

Parameter	Algorithm 4	Algorithm 5	Algorithm 7
β_0	0.971	0.384	0.490
β_1	0.974	0.429	0.491
β_2	0.993	0.519	0.494
D_{11}	0.928	0.930	0.925
D_{12}	0.890	0.886	0.892
D_{22}	0.812	0.806	0.791

Table 7. *Lag 5 sample autocorrelations, algorithms for the labor force participation data model*

Parameter	Algorithm 4	Algorithm 5	Algorithm 7
β_0	0.860	0.155	0.046
β_1	0.873	0.192	0.055
β_2	0.965	0.254	0.071
D_{11}	0.808	0.765	0.751
D_{12}	0.725	0.641	0.656
D_{22}	0.663	0.542	0.529

Placing a bivariate $N_2(0, D)$ mixing distribution on the b_i vectors, we run 5000 iterations of our Section 3 algorithms, displaying the autocorrelation results for β and D at lags 1, 5, and 10 in Tables 6, 7, and 8, respectively. As expected, Algorithm 4 performs very poorly, with cripplingly high autocorrelations even at lag 10. Algorithm 5 offers substantial improvement, especially for β, but even here the autocorrelation is still quite apparent at lag 10. Our fully blocked Algorithm 7 appears to offer little further improvement at lag 1, but by lag 5 the improvement for β is noticeable; by lag 10, so is the improvement for D, and the β samples are effectively uncorrelated. We speculate that as the dimension of the random effects increases, the advantage of Algorithm 7 would become even more apparent.

5. Discussion

In this paper we have described several reduced (partially and fully blocked) MCMC algorithms for minimizing the autocorrelation in MCMC samples arising from the important classes of longitudinal continuous and binary data models. Our approaches have been shown to offer considerable advantages over existing methods, including the form of Gibbs sampler used by the leading generalist Bayesian software package (BUGS), while still being relatively straightforward to code. In the continuous case, the fully blocked Algorithm 3 can be recommended, since it is easy to use and performed well in our head-to-head comparison. In the binary case, however, the partially blocked Algorithm 5 may be best overall, in terms of both case of use and quality of the output. In our two data examples of this type, the further reduced Algorithm 7 improved autocorrelation performance, but not in relation to the coding and execution time expended. Still, this algorithm acts as a benchmark for the kind of improvement that is possible in binary data models when when one works with the likelihood function directly. In both the continuous and binary settings, however, one common conclusion is that the fixed effects should be simulated only after the random effects are marginalized out.

Table 8. *Lag 10 sample autocorrelations, algorithms for the labor force participation data model*

Parameter	Algorithm 4	Algorithm 5	Algorithm 7
β_0	0.738	0.099	0.002
β_1	0.760	0.135	0.002
β_2	0.934	0.169	0.009
D_{11}	0.693	0.631	0.604
D_{12}	0.584	0.491	0.473
D_{22}	0.560	0.447	0.395

Another possibility to improving the convergence of the matrix D might be a group move of the form suggested by Liu and Sabatti (1998). These authors' *simulated sintering* approach is reminiscent of simulated tempering (Geyer and Thompson, 1995), but instead of varying a user-selected "temperature" parameter (the "coldest" of which corresponds to the true posterior), it involves varying the accuracy used in describing the underlying problem. Our work on blocking for longitudinal models is closely related to similar work for state space models. Carter and Kohn (1994, 1996) showed how blocking can considerably improve convergence in linear Gaussian state space model settings over the univariate updating algorithm of Carlin, Polson and Stoffer (1992). More recently, Kim, Shephard and Chib (1998) provide a detailed analysis of the value of blocking in non-linear state space models of stochastic volatility. In the context of general state space models, Shephard and Pitt (1997) discuss the use of random block sizes for sampling state vectors via the Metropolis algorithm while Knorr-Held (1998) provides an alternative implementation of the Metropolis step. Finally, the advantages of blocking are also demonstrated by Chib (1996) for hidden Markov (or Markov mixture) models.

In future work we hope to extend our methods to other longitudinal data models, including generalized linear and nonlinear response models, such as pharmacokinetic models.

Acknowledgements

The research of the second author was supported in part by a National Institute of Allergy and Infectious Diseases (NIAID) Grant 1-R01-AI41966.

References

Abrams, D. I., Goldman, A. I., Launer, C., Korvick, J. A., Neaton, J. D., Crane, L. R., Grodesky, M., Wakefield, S., Muth, K., Kornegay, S., Cohn, D. L., Harris, A., Luskin-Hawk, R., Markowitz, N., Sampson, J. H., Thomson, M., Deyton, L., and the Terry Beirn Community Programs for Clinical Research on AIDS (1994) Comparative trial of didanosine and zalcitabine in patients with human immunodeficiency virus infection who are intolerant of or have failed zidovudine therapy. *New England Journal of Medicine,* **330**, 657–62.

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.*, **88**, 669–79.

Albert, J. and Chib, S. (1996) Bayesian modeling of binary repeated measures data with application to crossover trials. In *Bayesian Biostatistics*, D. A. Berry and D. K. Stangl, eds., New York: Marcel Dekker, pp. 577–99

Andrews, D. F. and Mallows, C. L. (1974) Scale mixtures of normality. *J. Roy. Statist. Soc., Ser. B*, **36**, 99–102.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3–66.

Carlin, B. P. (1996) Hierarchical longitudinal modeling. In *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 303–19.

Carlin, B. P. and Louis, T. A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman and Hall.

Carlin, B. P. and Polson, N. G. (1992) Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds., Oxford: Oxford University Press, pp. 577–86.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992) A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Amer. Statist. Assoc.*, **87**, 493–500.

Carter, C. K. and Kohn, R. (1994) On Gibbs sampling for state space models. *Biometrika*, **81**, 541–53.

Carter, C. K. and Kohn, R. (1996) Markov chain Monte Carlo in conditionally Gaussian state space models. *Biometrika*, **83**, 589–601.

Chib, S. (1995) Marginal likelihood from the Gibbs output. *Journal of the Americal Statistical Association*, **90**, 1313–21.

Chib, S. (1996) Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models. *Journal of Econometrics*, **75**, 79–97.

Chib, S. and Greenberg E. (1995) Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**, 327–35.

Chib, S. and Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika*, **85**, 347–61.

Damien, P., Wakefield, J. and Walker, S. (1998) Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxilliary variables. To appear *J. Roy. Statist. Soc.*, Ser. B.

Dawid, A. P. (1979) Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc.* Ser. B, **41**, 1–31.

Fitzmaurice, G. F. V. and Laird, N. M. (1993) A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–51.

Gelfand, A. E. and Sahu, S. K. (1999) Identifiability, improper priors and Gibbs sampling for generalized linear models. To appear *J. Amer. Statist. Assoc.*

Gelfand, A. E. Sahu, S. K. and Carlin, B. P. (1995) Efficient parametrizations for normal linear mixed models. *Biometrika*, **82**, 479–88.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996) Efficient parametrizations for generalized linear mixed models (with discussion). In *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford: Oxford University Press, pp. 165–80.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398–409.

Geyer, C. J. and Thompson, E. A. (1995) Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.*, **90**, 909–20.

Gilks, W. R. and Roberts, G. O. (1996) Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, D. J., eds, London: Chapman and Hall, pp. 89–114.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *J. Roy. Statist. Soc., Ser. C* (*Applied Statistics*), **41**, 337–48.

Goldman, A. I., Carlin, B. P., Crane, L. R., Launer, C., Korvick, J. A., Deyton, L. and Abrams, D. I. (1996) Response of $CD4^+$ and clinical consequences to treatment using ddI or ddC in patients with advanced HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **11**, 161–69.

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

Kass, R. E., Carlin, B. P., Gelman, A. and Neal, R. (1998) Markov chain Monte Carlo in practice: A round table discussion. *The American Statistician*, **52**, 93–100.

Kim, S., Shephard, N. and Chib, S. (1998) Stochastic Volatility, Likelihood Inference and Comparison with ARCH Models. *The Review of Economic Studies*, **65**, 361–394.

Knorr-Held, L. (1998) Conditional prior proposals in dynamic models. To appear *Scandinavian Journal of Statistics*.

Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–74.

Lange, N., Carlin, B. P., and Gelfand, A. E. (1992) Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T- cell numbers (with discussion). *J. Amer. Statist. Assoc.*, **87**, 615–32.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc.,* Ser. B, **34**, 1–41.

Liu, J. S. (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 958–66.

Liu, J. S. and Sabatti, C. (1998) Simulated sintering: Markov chain Monte Carlo with spaces of varying dimension. To appear *Bayesian Statistics 6*, J. M. Bernardo, J. O. Berger,

A. P. Dawid and A. F. M. Smith, eds., Oxford: Oxford University Press.

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

Roberts, G. O. and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking, and parameterization for the Gibbs sampler. *J. Roy. Statist. Soc.*, Ser. B, **59**, 291–17.

Shephard, N. and Pitt, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653–67.

Spiegelhalter, D. J., Thomas, A., Best, N. and Gilks, W. R. (1995) BUGS 0.5: Bayesian inference using Gibbs sampling manual. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University.

Vines, S. K., Gilks, W. R. and Wild, P. (1996) Fitting Bayesian multiple random effects models. *Statistics and Computing*, **6**, 337–46.

Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.*, **86**, 79–86.