Statistics and Computing **12**: 377–389, 2002
© 2002 Kluwer Academic Publishers. *Manufactured in The Netherlands.*

MCMC sampler convergence rates for hierarchical normal linear models: A simulation approach

MARY KATHRYN COWLES

Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA
kcowles@stat.uiowa.edu

Received May 2000 and accepted June 2001

This paper presents a straightforward method of approximating theoretical bounds on burn-in time for MCMC samplers for hierarchical normal linear models. An extension and refinement of Cowles and Rosenthal's (1998) simulation approach, it exploits Hodges's (1998) reformulation of hierarchical normal linear models. The method is illustrated with three real datasets, involving a one-way variance components model, a growth-curve model, and a spatial model with a pairwise-differences prior. In all three cases, when the specified priors produce proper, unimodal posterior distributions, the method provides very reasonable upper bounds on burn-in time. In contrast, when the posterior distribution for the variance-components model can be shown to be improper or bimodal, the new method correctly identifies convergence failure while several other commonly-used diagnostics provide false assurance that convergence has occurred.

Keywords: agricultural field trials, Bayesian models, Gibbs sampler, total variation distance

1. Introduction

A critical question for users of Markov chain Monte Carlo methods for estimation and inference is determining "burn-in" time—how many initial iterations must be discarded from MCMC sampler output in the hope that the remaining iterates are drawn from a close approximation to the target distribution. Considerable research exists regarding theoretical upper bounds on burn-in time, for example Meyn and Tweedie (1994), Rosenthal (1995), and Roberts and Tweedie (1999). The usefulness of these approaches for real models is limited by the fact that they involve extremely difficult computations even for relatively simple problems and often produce impractically large upper bounds. Jones and Hobert (2001) compute theoretical convergence bounds in a reasonably straightforward manner, but for a single model only. Consequently, most MCMC users apply convergence diagnostics to sampler output. Brooks and Roberts (1998), Cowles and Carlin (1996), and Mengersen, Robert and Guihenneuc-Joyaux (1999) found that diagnostics based on sampler output can be fooled. Furthermore, Cowles, Roberts and Rosenthal (1999) showed that naive use of convergence diagnostics may actually introduce bias into MCMC output analysis.

Two recent approaches to guaranteeing or assessing MCMC convergence are coupling from the past ("CFTP") (see for example Propp and Wilson 1996, Green and Murdoch 1998) and Johnson's (1998) two-chain coupling-regeneration diagnostic. CFTP enables obtaining an exact draw from the target distribution of an MCMC sampler; however, at present it is computationally infeasible for complex samplers with continuous state spaces. Johnson's method, on the other hand, can be applied straightforwardly to the complex hierarchical normal linear models that are the subject of this paper. It requires fairly minimal supplementary coding plus the running of a single auxiliary chain of the same length as the primary chain, the output of which is used for inference and estimation. At intervals of length T, the auxiliary chain is restarted from a chosen point x^* in the state space of the chain. The method enables graphical assessment of geometric ergodicity, estimation of a lower bound on the "effective sample size" of correlated sampler output, and determination of a skip interval k such that, with a specified probability, a subsample consisting of every kth iterate from the sampler constitutes an independent sample from the target distribution.

However, Johnson (1998) states (p. 246): "...when the restart interval T is seriously underestimated and the number of restart intervals is small ... the MCMC algorithm may not reach equilibrium before sampling terminates, and [the procedure] may not provide useful diagnostic information." For example, for a multimodal target distribution, Johnson's diagnostic may incorrectly indicate rapid convergence if one or more major modes is missed by both the primary chain and the auxiliary chain. Furthermore, Johnson's diagnostic may fail to identify convergence failure if no posterior distribution exists. Both cases are illustrated in Section 5.

The limitations of existing convergence-assessment methods suggest a continuing need for a practical method of at least approximating theoretical bounds on burn-in for the kinds of models for which MCMC is used in practice. Cowles and Rosenthal (1998)—hereafter "C&R"— proposed the use of auxiliary simulations to verify numerically certain conditions that are known to provide upper bounds on convergence times but that are difficult or impossible to verify analytically for complex models. Although they successfully applied their method to two models for which theoretical convergence bounds had not previously been determined, they provided little guidance on how to implement their approach in more general cases.

The present paper extends C&R's simulation approach to complex hierarchical normal linear models (HNLMs) and provides explicit direction for applying it to general models of this class. Section 2 reviews a theorem from Rosenthal (1995) and Section 3 describes a reformulation of the HNLM that greatly facilitates C&R's approach. Section 4 details application of C&R's method to HNLMs, using the variance-components model as an example. Real data results are presented in Section 5, under three different prior specifications. When the resulting posterior is proper and unimodal, the simulation method verifies very rapid burn-in. When the priors are altered to produce first an improper posterior and then a bimodal posterior, the new method identifies convergence failure while several popular convergence diagnostics, as well as Johnson's (1998) method, do not. Section 6 applies the method to two more-complicated models—a spatial model with a pairwise-differences prior and a growth-curve model. Section 7 contains discussion.

Although the simulation-based method of convergence verification is more computationally intensive than competing convergence diagnostics, it is far more reliable. It therefore is recommended when correct inference based on models fit with MCMC samplers is crucial.

2. Rosenthal's theorem (1995)

Using coupling theory, Rosenthal (1995) proved a theorem establishing exponential convergence in total variation distance with an explicit rate if an MCMC sampler can be shown to satisfy a *drift* condition and a *minorization* condition. We briefly review these conditions and the result of the theorem. Details are in Rosenthal (1995) and Cowles and Rosenthal (1998).

Let $\{X^{(k)}\}_{k=0}^{\infty}$ be a Markov chain with state space $\mathcal{X}$, stationary distribution π, and transition probabilities $P(X, \cdot)$. Then verifying the drift condition involves specifying a function V mapping the state space $\mathcal{X}$ to the non-negative real line such that:

$$E\big(V\big(X^{(m)}\big) \mid X^{(0)} = x\big) \leq \lambda\, V(x) + \Lambda, \quad x \in \mathcal{X} \qquad (1)$$

where E denotes expectation, $\lambda < 1$, $\Lambda < \infty$, and m is a positive integer (number of iterations). Heuristically the drift condition means that whenever the chain enters a region of the state space that produces large values of V, it tends to move toward regions that produce smaller values. Choosing V such that $V \geq 1$ will produce tighter bounds on convergence in the computations below, and this is assumed in what follows.

The minorization condition states that

$$P^{mk_0}(x, \cdot) \geq \epsilon Q(\cdot), \quad x \in V_d \qquad (2)$$

where d is chosen comfortably larger than $\frac{2\Lambda}{1-\lambda} - 1$, $V_d = \{x \in \mathcal{X};\ V(x) \leq d\}$, $\epsilon > 0$, $Q(\cdot)$ is a probability measure on $\mathcal{X}$, and m and k_0 are positive integers. Intuitively, this means that a subset of the state space V_d, defined by an upper bound on the value of V, exists such that two parallel Markov chains starting at any two different points in V_d would have positive probability of "coupling" (arriving at the same point in the state space) in the next mk_0 iterations.

Two lemmas from Rosenthal (1995) facilitate verification of the minorization condition. Lemma 7 states that the minorization condition need be verified only for those parameters that are updated before being used in computations at the next sampler iteration. Lemma 6b enables determining ϵ when the transition probabilities of the Markov chain are densities. It says that, if a Markov chain satisfies $P^{mk_0}(x, \cdot) = f(x, y)\, dy$, where $f(x, \cdot)$ is a density function and dy is Lebesgue measure, then there exists a probability measure $Q(\cdot)$ satisfying (2), where

$$\epsilon = \int_{\mathcal{X}} \Big(\inf_{x \in V_d} f(x, y) \Big)\, dy \qquad (3)$$

If these conditions can be verified and $V(x) \geq 1$ for all $x \in \mathcal{X}$, then Rosenthal's theorem provides the following bound on total variation distance to stationarity at the kth iteration:

$$\big\| L\big(X^{(k)}\big) - \pi \big\|_{var} \leq (1-\epsilon)^{[rk/mk_0]} + C_0\, (\alpha A)^{-1} \times \big(\alpha^{-(1-rk_0)} A^r\big)^{[k/m]} \qquad (4)$$

where $\alpha^{-1} = \lambda + \frac{M\Lambda + (1-\lambda)(1-M)}{1 + \frac{M}{2}(d-1)}$, $A = M(\lambda d + \Lambda) + (1-M)$, $C_0 = \frac{M}{2}\,(\frac{\Lambda}{1-\lambda} + E_\nu(V(X^{(0)}))) + (1-M)$, and ν is the distribution of the initial value of the Markov chain. Tuning constants r and M may be chosen within the ranges $0 < r < 1$ and $M > 0$ so as to obtain the tightest bounds.

3. Reformulating hierarchical linear models and the Gibbs sampler

Gelman *et al.* (1995) and Hodges (1998) described an approach to reformulating HNLMs as standard linear models, which greatly simplifies C&R's simulation approach to verifying drift and minorization conditions.

For an HNLM, let $\mathbf{y}$ denote the observed data and $\mathbf{\Theta_1}$ denote those parameters required to specify the expectation of $\mathbf{y}$, i.e. $E(\mathbf{y}) = \bar{\mathbf{X}}_1\boldsymbol{\Theta}_1$ for $\mathbf{X}_1$ a known design matrix. Similarly at the second level, let $\mathbf{Z}_1$ and $\mathbf{Z}_2$ specify the mean of $\boldsymbol{\Theta}_1$ conditional on parameters in $\boldsymbol{\Theta}_2$. Finally, let matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ and vector $\mathbf{M}$ specify the prior means for parameters in $\boldsymbol{\Theta}_2$ and parameters in $\boldsymbol{\Theta}_1$ that are not modeled at the second level of the hierarchy. Then the entire means structure may be expressed as:

$$\left(\begin{array}{c} \mathbf{y} \\ \hline \mathbf{0} \\ \hline \mathbf{M} \end{array}\right) = \left(\begin{array}{c|c} \mathbf{X}_1 & 0 \\ \hline \mathbf{Z}_1 & \mathbf{Z}_2 \\ \hline \mathbf{W}_1 & \mathbf{W}_2 \end{array}\right) \left(\begin{array}{c} \boldsymbol{\Theta}_1 \\ \hline \boldsymbol{\Theta}_2 \end{array}\right) + \left(\begin{array}{c} \boldsymbol{\psi} \\ \boldsymbol{\delta} \\ \boldsymbol{\xi} \end{array}\right)$$

or

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E}$$

where $\mathbf{Y}$ consists of known values, $\mathbf{X}$ is a known design matrix, $\boldsymbol{\Theta}$ is a vector of unknown parameters, and $\boldsymbol{\psi}$, $\boldsymbol{\delta}$, and $\boldsymbol{\xi}$ are error vectors.

The covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{E}$ is block diagonal. The upper lefthand block is the covariance matrix of $\boldsymbol{\psi}$, the middle block the covariance matrix of $\boldsymbol{\delta}$, and the lower righthand block the known prior covariance matrix.

For example, in the variance-components model, the three stages of the standard hierarchical form are:

$$Y_{ij} \mid \theta_i, \sigma_y^2 \sim N(\theta_i, \sigma_y^2), \quad i = 1, \ldots, K, j = 1, \ldots, J$$
$$\theta_i \mid \mu, \sigma_\theta^2 \sim N(\mu, \sigma_\theta^2), \quad \mu \sim N(\mu_0, \sigma_0^2)$$

which may be written as

$$\left(\begin{array}{c} \mathbf{y}_{\mathbf{n}\times\mathbf{1}} \\ \hline \mathbf{0}_{\mathbf{K}\times\mathbf{1}} \\ \hline \mu_{0,1\times 1} \end{array}\right) = \left(\begin{array}{c|c} \mathrm{X}_{n\times K} & 0_{n\times 1} \\ \hline -\mathrm{I}_{K\times K} & 1_{K\times 1} \\ \hline 0_{1\times K} & 1_{1\times 1} \end{array}\right) \left(\begin{array}{c} \boldsymbol{\theta} \\ \hline \mu \end{array}\right) + \left(\begin{array}{c} \boldsymbol{\psi} \\ \boldsymbol{\delta} \\ \xi \end{array}\right)$$

The covariance matrix of the error terms is diagonal, with the first n diagonal entries equal to σ_y^2 and the next K diagonal entries equal to to σ_θ^2. The final diagonal entry is equal to σ_0^2.

The Bayesian variance-components model is completed with priors on σ_y^2 and σ_θ^2.

$$\sigma_y^2 \sim IG(a_1, b_1), \quad \sigma_\theta^2 \sim IG(a_2, b_2)$$

where a_1, b_1, a_2 and b_2 are known constants and IG denotes the inverse gamma p.d.f.

Hodges (1998) and Sargent, Hodges and Carlin (2000) point out that the reformulation suggests an improved MCMC sampler for HNLMs, in which all the means parameters are generated as a block. (The variance/covariance parameters are generated as usual from their conjugate full conditional distributions.) For the variance-components example, the full conditionals used at iteration k of the sampler are:

$$\sigma_y^{2(k)} \mid \boldsymbol{\theta}^{(k-1)} \sim IG\left(a_1 + \frac{JK}{2}, b_1 + \frac{\sum_{i=1}^K \sum_{j=1}^J \left(Y_{ij} - \theta_i^{(k-1)}\right)^2}{2}\right) \qquad (5)$$

$$\sigma_\theta^{2(k)} \mid \boldsymbol{\theta}^{(k-1)}, \mu^{(k-1)} \sim IG\left(a_2 + \frac{K}{2}, b_2 + \frac{\sum_{i=1}^K \left(\theta_i^{(k-1)} - \mu^{(k-1)}\right)^2}{2}\right) \qquad (6)$$

$$\boldsymbol{\Theta}^{(\mathbf{k})} = \left[\boldsymbol{\theta}^{(\mathbf{k})}, \mu^{(\mathbf{k})}\right] \mid \sigma_{\mathbf{y}}^{\mathbf{2(k)}}, \sigma_\theta^{\mathbf{2(k)}} \sim N([\mathbf{X}^{\mathbf{T}}\Sigma^{-\mathbf{1}}\mathbf{X}]^{-1}\mathbf{X}^{\mathbf{T}}\Sigma^{-\mathbf{1}}\mathbf{Y}, [\mathbf{X}^{\mathbf{T}}\Sigma^{-\mathbf{1}}\mathbf{X}]^{-1}) \qquad (7)$$

When the multivariate normal in (7) is high-dimensional, an efficient computing algorithm based on the Choleski decomposition of $\mathbf{X}^{\mathbf{T}}\Sigma^{-\mathbf{1}}\mathbf{X}$ speeds up generation enormously.

4. Bounding burn-in time for an MCMC sampler for the reformulated model

4.1. *Specifying the V function*

C&R gave little guidance on the choice of the function V, saying only that

> The conditions [drift and minorization] ... imply the following informal goals for the function V: (a) if the chain is 'far away', then the value of V should tend to decrease on the next iteration; and (b) the transition probabilities $P(x, \cdot)$ should have reasonably large 'overlap' from all points x with $V(x) \leq d$.

For a reformulated HNLM, choosing an appropriate mathematical form for the V function is easy. The function V needs to "control" only those parameters whose values are used in the *next* sampler iteration after that in which they are generated. Inspection of (5)–(7) reveals that for the variance-components model, these are $\boldsymbol{\theta}$ and μ. (In contrast, new values of σ_y^2 and σ_θ^2 are generated at the beginning of each iteration, and these new values are used in the *same* iteration.) For general reformulated HNLMs, V must control the means parameters designated Θ in Section 3. Let $\mathbf{T}$ denote the vector of all variance-covariance parameters and $p(\mathbf{T})$ its prior. Then the unnormalized joint posterior distribution of all parameters is:

$$p(\mathbf{T}, \Theta \mid \mathbf{Y}) \propto p(\mathbf{T}) \times \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{[\mathbf{Y} - \mathbf{X}\Theta]^T \boldsymbol{\Sigma}^{-1}[\mathbf{Y} - \mathbf{X}\Theta]}{2}\right)$$

For any fixed value of $\mathbf{T}$, $V^* = [\mathbf{Y} - \mathbf{X}\Theta]^T \boldsymbol{\Sigma}^{-1}[\mathbf{Y} - \mathbf{X}\Theta]$ (the *weighted sum of squared residuals*) is exactly what is needed: a non-negative function of $\boldsymbol{\Theta}$ that takes on large values for low-posterior-probability regions of the parameter space (and thus of the state space of the Markov chain) and small values for high-posterior-probability regions. For the variance-components model, this is:

$$V^* = \frac{\sum_i \sum_j (Y_{ij} - \theta_i)^2}{\sigma_y^2} + \frac{\sum_i (\theta_i - \mu)^2}{\sigma_\theta^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2} \qquad (8)$$

Because σ_y^2 and σ_θ^2 are unknown parameters treated as random in our Bayesian model, (σ_0^2 in contrast is a known prior variance), appropriate constants must be substituted into (8). To choose such constants, the maximum likelihood estimate $\hat{\Theta} = \left[\mathbf{X}^{\mathbf{T}}\Sigma^{-1}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathbf{T}}\Sigma^{-1}\mathbf{Y}$ may be plugged into V*, producing a profile joint posterior that is a function of σ_y^2 and σ_θ^2 alone. Since the dimension is only two, their posterior modes or posterior means may be computed by nonlinear optimization or numeric integration respectively. Alternatively, the means or medians of the posterior distributions of the variance-components from a pilot sampler, preferably based on a different MCMC algorithm, may be used.

Values obtained by one of these methods may be substituted into (8), producing

$$V^{**} = \frac{\sum_i \sum_j (Y_{ij} - \theta_i)^2}{\hat{\sigma}_y^2} + \frac{\sum_i (\theta_i - \mu)^2}{\hat{\sigma}_\theta^2} + \frac{(\mu - \mu_0)^2}{\sigma_0^2}, \quad (9)$$

Because tighter bounds on burn-in can be obtained with a V function that takes on values ≥ 1, a final useful step is to minimize V^{**} over $\boldsymbol{\theta}$ and μ, obtaining a minimum value, say v, and to set

$$V = \frac{V^{**}}{v} \geq 1 \quad (10)$$

4.1.1. *Analytic evidence of the appropriateness of this V function*

Jones and Hobert (2001) analytically established drift and minorization conditions for the reformulated version of the one-way variance components model. Their bounds on total variation distance are not tight, in part because only the case $m = k_0 = 1$ could be considered due to difficulty of the analytic computations required. In addition, they have not yet obtained analytic results for any HNLMs more complicated than the one-way variance-components model. For their V function, Jones and Hobert used

$$V_{JH} = \phi \sum_{i=1}^{K} (\theta_i - \mu)^2 + \sum_{i=1}^{K} (\theta_i - \bar{Y}_i)^2$$

where ϕ is chosen such that $\frac{\phi K}{2a_1+JK-2} + \max(\frac{1}{2a_2+K-2}, \frac{K+1}{2a_1+JK-2}) < 1$. Let SSE denote $\sum_{i=1}^{K}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_i)^2$. Then $V_{JH} = V^{**} - \frac{SSE}{J}$ with $\frac{1}{\hat{\sigma}_0^2} = 0$ (a locally uniform prior on μ), $\frac{1}{\hat{\sigma}_\theta^2} = \phi$, and $\frac{1}{\hat{\sigma}_y^2} = \frac{1}{J}$. That is, V_{JH} differs from V^{**} only by an additive constant, which is absorbed into Jones and Hobert's expression for Λ in (1). This provides analytic evidence that a function in the form (9) will enable verification of the drift condition. It also suggests bounds (defined in terms of constants from the prior and the data) on the ratio $\frac{\hat{\sigma}_\theta^2}{\hat{\sigma}_y^2}$ in order for a drift condition to be verified.

4.2. *Verifying the drift condition for reformulated models*

C&R proposed a two-step simulation process to find lower bounds on λ and Λ in (1). First, for each point $x_0 \in \mathcal{X}$ such that $V(x_0) = 0$, they run N_0 m-iteration chains with x_0 as initial value and thus estimate $E(V(X^{(m)}) \mid X^{(0)} = x_0)$ as the mean of $V(X^{(m)})$ over the N_0 draws. The largest of these estimated expected values, over all values of x_0, provides a lower bound $\hat{\Lambda}$. If $V(x)$ has been chosen greater than or equal to 1, they carry out this process with $V - 1$ and add 1 to the resulting $\hat{\Lambda}$. N_0 is specified large enough that the standard error of this mean is less than or equal to a chosen tolerance, and m is chosen by trial and error to produce a small final bound. In our experience with the types of models described in this paper, values of m between 3 and 12 usually are optimal.

Next, a value of λ corresponding to $\hat{\Lambda}$ must be estimated. C&R generate N_1 (determined by the complexity of the state space) different sets of initial values $x_0^* \in \mathcal{X}$. As many sets as are required to represent all finite boundaries of the state space are specified deterministically. Additional sets of initial values are generated randomly from interior regions of the state space and/or to approach infinite boundaries. For each such initial value x_0^*, C&R run N_2 m-iteration chains from which they estimate $e(x) = E(V(X^{(m)}) \mid X^{(0)} = x_0^*)$. An estimate $\hat{\lambda}$ corresponding to the given $\hat{\Lambda}$ is obtained as the maximum of $\hat{\lambda}_j = (e(x) - \hat{\Lambda})/V(x)$ over all choices of x_0^*. The number of replicate chains N_2 is chosen to produce acceptably small standard errors of all $\hat{\lambda}_j$s. If $\hat{\lambda} \geq 1$, they repeat the procedure with a larger value of $\hat{\Lambda}$ or of m.

For the variance-components model, initial values are needed only for $\boldsymbol{\theta}$ and μ because values of σ_y^2 and σ_θ^2 from the previous iterations are not used in any full conditionals. When the V function is constructed as above, the single set of initial values x_0 such that $V(x_0) = 1$ is found automatically when minimizing V.

Regarding initial values x_0^* for the second step, inspection of (5)–(7) indicates that initial values affect the output of the first iteration of this sampler only through the sums of squares in (5) and (6). Thus sets of initial values that produce the minima of these sums of squares, as well as representative values from their full range of possible values, must be selected. Clearly, the lower bound on $\sum_i \sum_j (Y_{ij} - \theta_i)^2$ is $\sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2$. All terms in $\sum_{i=1}^{K}(\theta_i - \mu)^2$ are identically 0 if each $\theta_i = \mu, i = 1, \ldots, K$. Thus, sets of initial values x_0^* may be generated as follows. The first set of initial values x_{01}^* is generated by setting each $\theta_i = \bar{Y}_i, i = 1, \ldots, K$ and $\mu = \bar{\bar{Y}}$ (the overall mean of the observed values). A second set x_{02}^* is generated by setting not only μ but also all $\theta_i, i = 1, \ldots, K$, equal to $\bar{\bar{Y}}$. Additional sets of initial values may be generated randomly from a sequence of normal distributions centered at x_{01}^* and with larger and larger variances.

4.3. *Verifying the minorization condition*

When the MCMC sampler's transition probabilities are densities, C&R suggest the following procedure to estimate ϵ in (2) and (3). They divide those coordinates of the state space over which Lemma 7 requires a minorization into a large number of little "bins," which must be small enough that the densities of the transition probabilities are nearly constant over each bin. They choose a value of d comfortably larger than $\frac{2\hat{\Lambda}}{1-\hat{\lambda}} - 1$ and define $V_d \equiv \{x \in \mathcal{X}; V(x) \leq d\}$. They then determine the "extremes of V_d"—that is, the sets of initial values x_0^{**} in V_d from which transition probabilities have minimum overlap among all choices of $x \in V_d$. From each initial value x_0^{**} they run N_3 different chains of length mk_0 and record what proportion of the final iterations land in each little bin. They approximate the integral in (3) by summing, over all little bins, the *minimum* over different choices of x_0^{**}, of the fraction of samples landing in that bin.

4.3.1. Selecting the number of replications and the bin size

C&R provided little guidance on how to choose N_3 and the bin size. We propose the following semi-automatic procedure, which has produced reasonable results in our work. Here *dim* denotes the number of dimensions over which the minorization condition must be verified.

1. Choose an initial value of N_3 such that $\frac{N_3}{5}$ is comfortably larger than 10^{dim} and run N_3 replicate samplers from each set of initial values x_0^{**}.
2. For each coordinate, find the minimum and maximum value obtained from all samplers from all starting values. Create the first set of bins by partitioning this range for each coordinate into 10 equal-length intervals. Estimate ϵ three times as described above, using the following three different fractions of the samples run: $\frac{N_3}{5}$, $2 \times \frac{N_3}{5}$, and $3 \times \frac{N_3}{5}$.
3. Halve the bin size by repartitioning each coordinate into $2^{1/dim} \times 10$ (rounded to an integer) equal-length intervals. Re-estimate ϵ using $2 \times \frac{N_3}{5}$, $3 \times \frac{N_3}{5}$, and $4 \times \frac{N_3}{5}$ replicate samples from each x_0^{**}.
4. Halve the bin size again and re-estimate ϵ using $3 \times \frac{N_3}{5}$, $4 \times \frac{N_3}{5}$, and N_3 replicate samples.

The binning method for estimating ϵ is based on histogram density estimation, and criteria for evaluating the results of the process come from that field. Stability of the estimate of ϵ as the sample size increases for a fixed bin size suggests that that bin size is small enough. Thus at the least, the three estimates obtained in step 4 should be approximately equal. Stability as the bin size decreases for a fixed sample size suggests that the sample size is large enough, so the two estimates obtained with $4 \times \frac{N_3}{5}$ replicates should be approximately equal. Finally, in order for a histogram estimator to be consistent, the number of observations per bin must get larger as the the bin size gets smaller. Thus, the first estimates from each of steps 2 and 3, and the last two estimates from step 4, should all be approximately equal. If these criteria are not met, the process is repeated after increasing either N_3 or the initial number of bins or both. A lower bound for ϵ is required. Thus, once the criteria are met, $\hat{\epsilon}$ should be chosen less than or equal to the *smallest* estimate considered.

4.3.2. Verifying the minorization condition for reformulated hierarchical normal models

The reformulation of HNLMs simplifies bounding burn-in time by reducing the number of parameters over which the minorization condition must be verified. According to Lemma 7 in Rosenthal (1995), in the reformulated variance-components model, this condition must be verified only for σ_y^2 and σ_θ^2. In the standard Gibbs sampler for this model, in which μ is generated separately from $\boldsymbol{\theta}$, it must be verified for μ as well.

4.4. *Finding the extremes of V_d*

"The extremes of V_d" are upper and lower bounds, subject to the constraint $V \leq d$, on any quantities that appear in the full conditionals for those parameters for which the minorization condition must be verified. If not all extremes of V_d are found, the method may fail to detect slow convergence. Conversely, if the extremes are estimated too conservatively, the estimated bounds on burn-in time will be unnecessarily large.

Fortunately, computing the extremes of V_d exactly is straightforward for reformulated HNLMs. For the variance-components model, the two quantities that must be bounded in (5) and (6) are $\sum_i(\theta_i - \mu)^2$ and $\sum_i \sum_j (Y_{ij} - \theta_i)^2$, and a little algebra shows that:

$$0 \leq \sum_i (\theta_i - \mu)^2 \leq \hat{\sigma}_\theta^2 \left(vd - \frac{\sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2}{\hat{\sigma}_y^2} \right)$$

$$\sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2 \leq \sum_i \sum_j (Y_{ij} - \theta_i)^2 \leq \hat{\sigma}_y^2 vd$$

5. Numerical example of a variance components model

5.1. *Peak discharge data*

For illustration we used the dataset from Table 4-4 of Montgomery (1991) consisting of $J = 6$ measurements of peak discharge (in cubic feet per second) taken by each of $K = 4$ different methods at the same watershed. As recommended by Montgomery, we applied the square root transformation to stabilize variance, obtaining:

Method	Square-root transformed observed values					
1	0.5830952	0.3464102	1.109054	0.836660	1.322876	0.3464102
2	0.9539392	1.7146428	1.462874	1.536229	1.691153	2.1330729
3	2.5119713	2.8930952	3.122499	2.467793	3.133688	2.6907248
4	4.1412558	3.4380227	3.309078	4.147288	3.788139	4.1012193

5.2. *Results for a proper, unimodal posterior*

We first specified the following priors:

$$\mu \sim N(0, \infty)$$
$$\sigma_y^2 \sim IG(0, 0)$$
$$\sigma_\theta^2 \sim IG(3, 4)$$

The priors on μ and σ_y^2 are improper, but the proper prior on σ_θ^2 guarantees a proper posterior (see Hobert and Casella 1996). In addition, this prior specification in combination with the likelihood for these data can be shown to produce a unimodal posterior (see Liu and Hodges 2002).

The V function was constructed as outlined in Section 4.1. The third term $\frac{(\mu-\mu_0)^2}{\sigma_0^2}$ was omitted because it would always evaluate to 0 due to the specification of $\sigma_0^2 = \infty$ in the prior on μ. As the required constants, we used REML estimates $\hat{\sigma}_y^2 = 0.134$ and $\hat{\sigma}_\theta^2 = 1.793$ obtained from *proc mixed* in SAS version 6.12 (SAS Institute, Cary NC), We used Maple (Waterloo Maple, Inc.) to minimize the resulting V^{**} over $\boldsymbol{\theta}$ and μ, obtaining $v = 23.06368862$, and defined $V \geq 1$ as in (10).

We used the simulation methods of Section 4.2 to obtain the constants needed to verify the drift condition (1). The one set of values x_0 of $\boldsymbol{\theta}$ and μ for which V attains its minimum, 1, was obtained from the minimization of V^{**} in Maple. Running $N_0 = 10000$ chains of length $m = 3$ estimated Λ as 1.2034 with a standard error of .0015, so we conservatively set $\hat{\Lambda} = 1.21$. Computer run time for this step was 3 seconds. (All reported run times are for programs coded in C and run on a Hewlett-Packard C180 Unix workstation.)

In preparation for estimating λ, we computed the initial values denoted x_{01}^* and x_{02}^* in Section 4.2. We randomly generated 50 additional sets of initial values from normal distributions with means x_{01}^* and standard deviations ranging from .25 to 9.0. These initial values produced a wide range of values for the sums of squares in (5) and (6). Furthermore, estimates of λ became *smaller* with larger standard deviations, so trying still larger standard deviations would not have changed our final estimate. From each set of initial values, we ran $N_2 = 5000$ 3-iteration samplers (run time 2 minutes and 27 seconds). With this choice of N_2, the standard errors of all resulting estimates of λ were ≤ 0.015. We conservatively chose $\hat{\lambda} = .04$, the largest value of any estimate of λ plus twice its standard error.

Based on $\hat{\lambda}$ and $\hat{\Lambda}$ from the drift condition, we specified $d = 2.5$ and calculated the extremes of V_d as described in Section 4.2. To estimate ϵ in the minorization condition (2), we ran $N_3 = 10000$ 3-iteration samplers ($m = 3$ and $k_0 = 1$) from each of the 4 combinations of extreme values (run time 13 seconds). We constructed two-dimensional bins of different sizes by partitioning each dimension into first 10, then 14, then 20 intervals, producing totals of 100, 196, and 400 bins respectively (i.e. approximately doubling the number of bins each time). Resulting estimates of ϵ are as follows.

	Number of Bins		
Number of replicates	100	196	400
2000	[✓ .89]		
4000	.90	[✓ .87]	
6000	.91	[✓ .89]	[✓ .88]
8000		.90	[✓ .89]
10000			[✓ .90]

Applying the criteria from Section 4.3.1, we very conservatively set $\hat{\epsilon} = 0.85$, decidedly smaller than the smallest estimate from the circled cells.

To compute the final bound on total variation distance to stationarity in (4), by trial and error we chose the tuning constants $r = .231$ and $M = 5.8$. If the initial values of the sampler are those that minimize V, then

$$\left\| L\left(X^{(k)}\right) - \pi \right\|_{var} \leq (0.15)^{[0.077k]} + (0.3077)(0.7333)^{[\frac{k}{3}]}$$

If $k = 39$, this bound is .0089. Thus, if the first 39 iterations are discarded, subsequent draws are from a distribution that differs from the true target distribution in total variation distance by well under 0.01.

We compared these results with similar bounds obtained from Johnson's (1998) method. As a restart point x^*, we used REML estimates of σ_y^2 and σ_θ^2, the MLE of μ, and empirical Bayes estimates of the θs, all obtained from *proc mixed* in SAS version 6.12 (SAS Institute, Cary, NC). We ran a two-chain coupler with restart interval 12 in which the primary chain was initialized with $\mu = \bar{\bar{y}}$ and $\theta_i = \bar{Y}_i,\ i = 1, \ldots, K$. In a 4812-iteration run, the chains coupled during each of the 400 restart intervals. The resulting SCQ plot and plot of the log of estimated ρ_m versus m are shown in row 1 of Fig. 1. Johnson's weighted least squares approach suggests that the distribution of a subsample of size 100 taken at skip intervals of 7 from the primary chain would differ in total variation distance from the distribution of a random sample of size 100 from the true posterior distribution by less than 0.0037. Thus our method, though producing a very small lower bound on burn-in time, is slightly more conservative than Johnson's for this problem. Jones and Hobert's analytic approach is much more conservative, suggesting that 800 burn-in iterations are required to get a bound on total variation distance less than 0.01.

5.3. *Results when the posterior does not exist*

As discussed by Hobert and Casella (1996), if the prior on σ_θ^2 is improper, i.e.

$$\sigma_\theta^2 \sim IG(0, 0)$$

then the posterior distribution is also improper. That is, no stationary distribution exists to which the MCMC sampler can converge. Neither Johnson's method nor the most commonly-used single-chain convergence diagnostics are guaranteed to detect this problem.

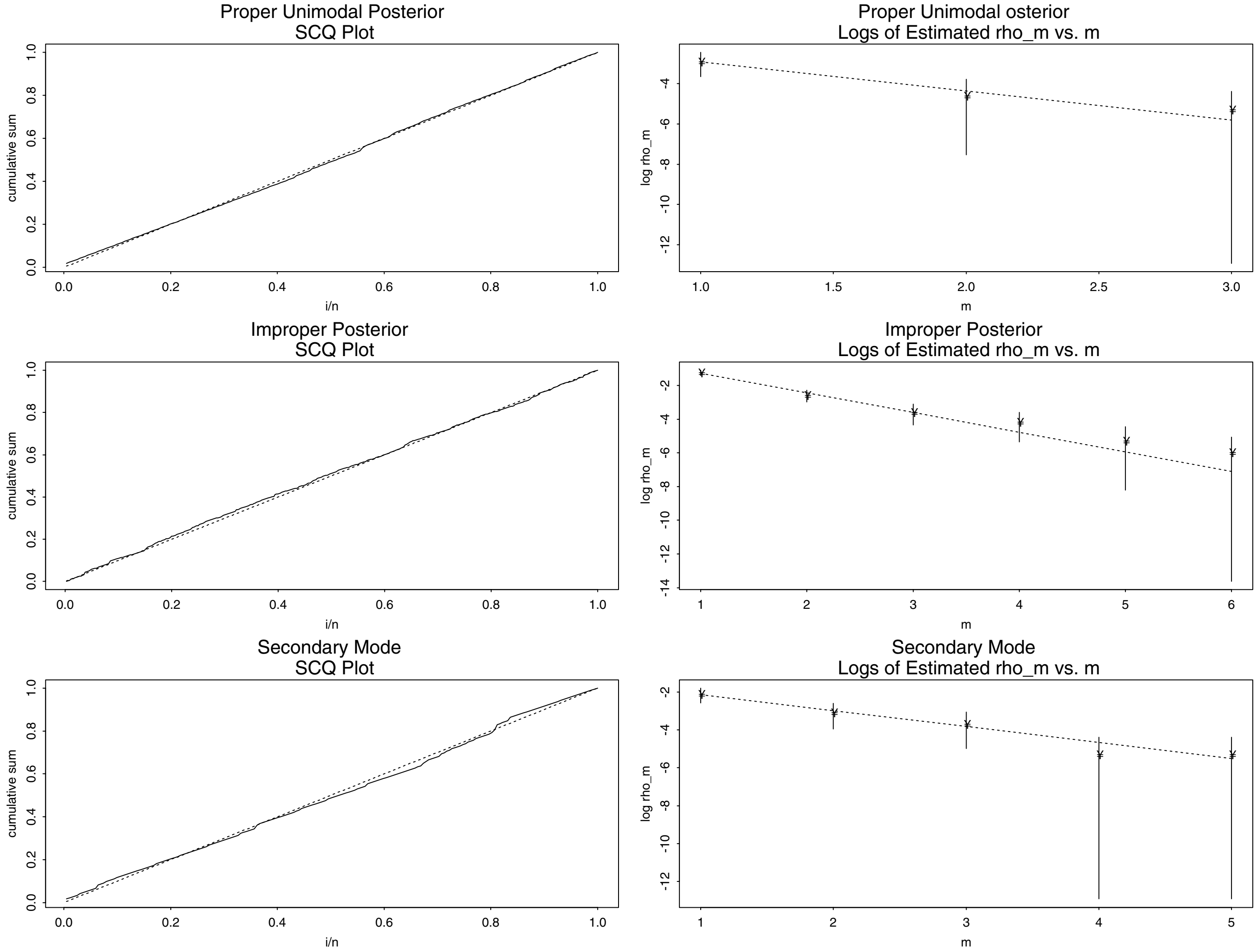

Fig. 1. *Application of Johnson's method to discharge-rate data*

We ran a two-chain coupler with the same initial values, coupling point, restart interval, and number of iterations as described in the preceding section but with this improper posterior. Again, the chains coupled within every restart interval. As shown in row 2 of Fig. 1, the SCQ plot deviates only slightly from the diagonal line, and the plot of the estimated ρ_m versus m is nearly linear. Johnson's weighted least squares computation suggested a burn-in time and skip interval of 10. Similarly, all the single-chain convergence diagnostics implemented in the software package BOA (Smith 2000) suggested very rapid convergence. The largest magnitude Z-score obtained with Geweke's diagnostic was 0.794. The sample path for every parameter passed Heidelberger and Welch's stationarity test without any initial iterations being discarded. Raftery and Lewis's diagnostic indicated that more than enough iterations had been run to estimate the 0.025 quantile of all parameters except θ_4 to within ±0.005 with probability 0.95.

In contrast, the proposed simulation method flagged the improper posterior in the second step of attempting to verify the drift condition. We again used REML estimates for $\hat{\sigma}_y^2$ and $\hat{\sigma}_\theta^2$ in the V function. In the first step, we estimated $\hat{\Lambda} = 1.27$, not very different from the 1.21 obtained with a proper posterior. However, when, in trying to obtain a bound for λ, we attempted to run a sampler from the initial values denoted x_{02}^* in Section 4.2 in which $\theta_i = \mu, i = 1, \ldots, K$, the sampler stopped with an error message at its first attempt to generate σ_θ^2. This occurred because the full conditional (6) was improper (the scale parameter $b_2 + \frac{\sum_{i=1}^K(\theta_i^{(k-1)} - \mu^{(k-1)})^2}{2} = 0$). If any full conditional is improper, it follows that the joint posterior distribution is improper. Jones and Hobert's method similarly flagged the improper full conditional with a divide-by-zero error in a computation.

5.4. *Results when the posterior has two well-separated modes*

Liu and Hodges (2002) developed an analytic method for determining whether the posterior distribution in a Bayesian balanced one-way variance-components model is unimodal or bimodal. The result depends on the number of groups (K), the number of observations per group (J), the prior hyperparameters a_1, b_1, a_2, and b_2, and the data values $\sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2$ ("SS_W") and $K \sum_i (\bar{Y}_i - \bar{\bar{Y}})^2$ ("SS_B").

For our dataset, K = 4, J = 6, $SS_B = 32.684$, $SS_W = 2.688$, and if the prior on σ_θ^2 is changed to

$$\sigma_\theta^2 \sim IG(4.0, 0.01)$$

then Liu and Hodges's method reveals a bimodal posterior distribution. For each model parameter, Fig. 2 shows traces of the output from two samplers, one initialized near each mode. Clearly the samplers do not jump readily from one mode to the other (in 2000 iterations, one chain switches modes once and the other chain never).

To illustrate that Johnson's diagnostic could fail to detect convergence failure if a sampler was stuck in one of the two modes, we ran a two-chain coupler with the restart point selected by setting μ and all the θs equal to $\bar{\bar{y}}$. The primary chain was initialized with μ and all the θs set equal to 0. We ran 2412 iterations with a restart interval of 12, obtaining 201 couplings. The plots shown in row 3 of Fig. 1 again do not show serious problems with convergence. The weighted least squares computation suggested that a burn-in time and skip interval of 12 would be required to produce a sample whose distribution differed from the distribution of a random sample of size 100 from the true posterior by a total variation distance less than 0.0040. The largest magnitude Z-score obtained with Geweke's diagnostic was –1.466, and each parameter passed Heidelberger and Welch's stationarity test without any initial iterations being discarded. Raftery and Lewis's diagnostic indicated that more than enough iterations had been run to estimate the 0.025 quantile of all parameters except σ_θ^2 to within ±0.0075 with probability 0.95, and that a total of 4524 iterations would be sufficient for σ_θ^2. Thus none of these four diagnostic methods detected the fact that a major mode had been missed altogether by the samplers.

To implement the simulation method, we set $\hat{\sigma}_y^2 = 1.6321$ and $\hat{\sigma}_\theta^2 = 0.0037$, the means of the sampler output from the

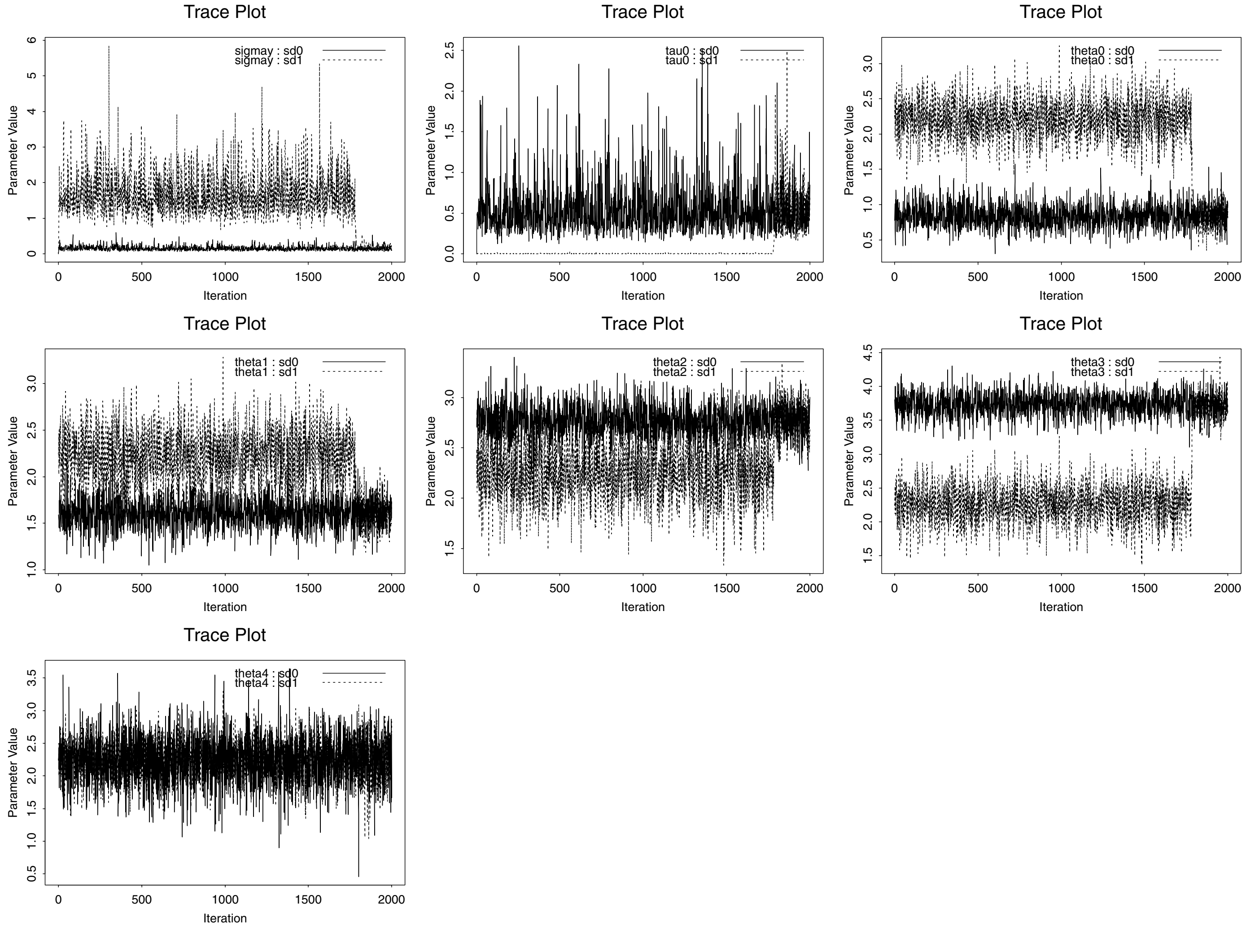

Fig. 2. *Trace plots from parallel chains, bimodal posterior*

primary chain in the Johnson coupler. The first step, preliminary estimation of Λ, produced $\hat{\Lambda} = 1.22$, very similarly to the other models. However, at the second step, we had to increase the estimate of Λ to $\hat{\Lambda} = 25$ in order to obtain an estimate of λ that was less than 1.0 with a reasonable number of iterations m. We finally settled on $\hat{\Lambda} = 25, \hat{\lambda} = 0.98$, and $m = 10$. Based on these values we chose $d = 3000$ which as required is comfortably larger than $\frac{2\hat{\Lambda}}{1-\hat{\lambda}} - 1$. The estimate of ϵ from the binning procedure was 0.0065. With tuning constants $r = 0.001$ and $M = 0.001$, we obtained

$$\left\| L\left(X^{(k)}\right) - \pi \right\|_{var} \leq (.9935)^{[0.0001k]} + (0.4090)(0.9994)^{[\frac{k}{10}]}$$

If $k = 5{,}000{,}000$, this bound is equal to .038. Although this bound could be tightened with improved choices of m, r, and M, clearly the simulation method has flagged slow convergence due to bimodality.

6. More complex models

6.1. *A spatial model*

Spatial models with intrinsic priors based on Markov random fields are a more complicated class of models to which the simulation-based method of convergence assessment is easy to apply. As an example, we consider an agricultural field experiment for which data and frequentist and Bayesian analyses appeared previously (Besag and Kempton 1986, Besag and Higdon 1993). The trial was an unreplicated 2×3^3 factorial design. The 54 treatment combinations were planted in three columns of 18 plots each. Although the primary purpose of the trial was to identify differences among effects of the levels of the four factors, an appropriate statistical model must also control for local variation in soil fertility. Like the previously-cited authors, we assumed that each plot $i, i = 1, \ldots, 54$ had its own unobservable fertility level F_i and that columns of plots were sufficiently separated that fertility levels in different columns may be considered independent. Within columns, spatial correlation was modeled by the assumption that plot fertility levels formed a Gaussian random walk, i.e. that pairwise differences were independent normals. Then stages one and two of the Bayesian model may be expressed as

$$Y_i \mid \boldsymbol{\beta},\ \boldsymbol{F}, \sigma_y^2 \sim N\left(\boldsymbol{z_i}^T\boldsymbol{\beta} + F_i, \sigma_y^2\right), \quad i = 1, \ldots, 54$$

$$F_i \mid F_{i-1}, \sigma_f^2 \sim N\left(F_{i-1}, \sigma_f^2\right), \quad i = 2, \ldots, 18;\ i = 20, \ldots 36;$$
$$i = 38, \ldots 54$$

where Y_i is the yield of the ith plot, $\mathbf{z_i}$ is a vector of indicator variables for the factor levels in the ith plot (main effects only), $\boldsymbol{\beta}$ is a vector of coefficients for the factors, and σ_y^2 and σ_f^2 are variances. The overall mean yield, as well as the block effects for the three columns of plots, are absorbed into the fertilities F_i.

As recommended by Besag and Higdon (1993, 1999), we standardized the yields Y_i to have sample variance 1.0, giving a range of values from 2.29 to 7.83. The same authors point out that *proper* priors on the variances σ_y^2 and σ_f^2 are necessary in order to obtain a proper joint posterior. We specified fairly informative inverse gamma priors on both variance components by setting $a_1 = 1.0, b_1 = 1.0, a_2 = 1.0, b_2 = 1.0$. These priors can be considered to provide information equivalent to two previously-observed yield values with average squared deviation of 1.0 from their true means, and two previously-observed first differences in fertilities with average squared values of 1.0. With vague but proper independent normal priors on $\boldsymbol{\beta}$, our priors were:

$$\beta_j \sim N(0, 10000), \quad j = 1, \ldots, 7$$
$$\sigma_y^2 \sim IG(a_1 = 1, b_1 = 1)$$
$$\sigma_f^2 \sim IG(a_2 = 1, b_2 = 1)$$

The means portion of the above model may be expressed as a linear model as follows.

$$\left(\begin{array}{c} \mathrm{Y}_{\mathbf{54\times1}} \\ \hline 0_{\mathbf{51\times1}} \\ \hline 0_{\mathbf{7\times1}} \end{array}\right) = \left(\begin{array}{c|c} \mathrm{Z}_{\mathbf{54\times7}} & \mathrm{I}_{\mathbf{54\times54}} \\ \hline 0_{\mathbf{51\times7}} & \Delta_{51\times54} \\ \hline \mathrm{I}_{7\times7} & 0_{\mathbf{7\times54}} \end{array}\right) \left(\begin{array}{c} \beta_1 \\ \vdots \\ \beta_7 \\ \hline F_1 \\ \vdots \\ F_{54} \end{array}\right) + \left(\begin{array}{c} \epsilon \\ \boldsymbol{\delta} \\ -\boldsymbol{\psi} \end{array}\right)$$

where

$$\Delta = \begin{pmatrix} D & 0 & 0 \\ 0 & D & 0 \\ 0 & 0 & D \end{pmatrix}$$

with each block

$$D_{17\times18} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 \\ & \vdots & & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

Like the variance-components model, this model may be expressed as $\mathbf{Y} = \mathbf{X}\Theta + \mathbf{E}$. The covariance matrix Σ of the errors is diagonal with the first 54 diagonal entries equal to σ_y^2, the next 51 equal to σ_f^2, and the last 7 equal to 10,000, the prior variance of the βs.

The full conditionals for a Gibbs sampler to fit this model are:

$$\sigma_y^{2(k)} \mid \boldsymbol{\beta}^{(k-1)},$$
$$\mathbf{F^{(k-1)}} \sim IG\left(a_1 + \frac{54}{2}, b_1 + \frac{\sum_{i=1}^{54}\left(Y_i - \boldsymbol{z_i}^T\boldsymbol{\beta}^{(k-1)} - F_i^{(k-1)}\right)^2}{2}\right) \tag{11}$$

$$\sigma_f^{2(k)} \big| \mathbf{F^{(k-1)}}$$
$$\sim IG\left(a_2 + \frac{51}{2}, b_1 + \frac{\sum_{l=1}^{3}\sum_{j=2}^{18}\left(F_{18(l-1)+j}^{(k)} - F_{18(l-1)+j-1}^{(k)}\right)^2}{2}\right) \tag{12}$$

$$\Theta^{(\mathbf{k})} = \left[\beta^{(k)}, \mathbf{F}^{(k)}\right] \mid \sigma_{\mathbf{y}}^{2(\mathbf{k})},$$
$$\sigma_{\mathbf{f}}^{2(\mathbf{k})} \sim N\left(\left[\mathbf{X^T}\Sigma^{(\mathbf{k})-\mathbf{1}}\mathbf{X}\right]^{-1}\mathbf{X^T}\Sigma^{(\mathbf{k})-\mathbf{1}}\mathbf{Y}, \left[\mathbf{X^T}\Sigma^{(\mathbf{k})-\mathbf{1}}\mathbf{X}\right]^{-1}\right) \tag{13}$$

The simulation method of convergence assessment is easy for this model because both variance components are scalars and there are only two parameters over which the minorization must be established. In particular,

$$V^{**} = \frac{\sum_i \left(Y_i - \mathbf{z_i}^T\boldsymbol{\beta} - F_i\right)^2}{\hat{\sigma}_y^2} + \frac{\sum_{k=1}^{3}\sum_{j=2}^{18}\left(F_{18(k-1)+j} - F_{18(k-1)+j-1}\right)^2}{\hat{\sigma}_f^2} + \frac{\sum_{j=1}^{7}\beta_j^2}{10{,}000}$$

and for the extremes of V_d

$$0 \le \sum_i \left(Y_i - \mathbf{z_i}^T\boldsymbol{\beta} - F_i\right)^2 \le \hat{\sigma}_y^2 vd \tag{14}$$

$$0 \le \sum_{k=1}^{3}\sum_{j=2}^{18}\left(F_{18(k-1)+j} - F_{18(k-1)+j-1}\right)^2 \le \hat{\sigma}_f^2 vd \tag{15}$$

Constants $\hat{\sigma}_y^2 = \frac{1}{6.42}, \hat{\sigma}_f^2 = \frac{1}{4.16}$ were obtained from a 15,000-iteration run of BUGS, from which the initial 5,000 iterations were discarded. Minimization over $\boldsymbol{\beta}$ and $\mathbf{F}$ in Maple produced $v = 27.4827$, so

$$V = \left(\frac{\sum_i \left(Y_i - \mathbf{z_i}^T\boldsymbol{\beta} + F_i\right)^2}{\hat{\sigma}_y^2} + \frac{\sum_{k=1}^{3}\sum_{j=2}^{18}\left(F_{18(k-1)+j} - F_{18(k-1)+j-1}\right)^2}{\hat{\sigma}_f^2} + \frac{\sum_{j=1}^{7}\beta_j^2}{10{,}000}\right)\Big/ 27.4827$$

Using the methods in Section 4.2, we verified (1) with $\hat{\lambda} = 0.051$, and $\hat{\Lambda} = 3.42$ with $m = 8$, $N_0 = 1000$ and $N_2 = 500$. Initial values for the chains used in estimating Λ were chosen to give the sums of squares in (11) and (12) values that ranged from 0 to 100,000. Run times for the two required steps were 2 and 1/2 minutes and 25 minutes respectively.

Choosing $d = 10$ and using the extremes of V_d given above, we obtained $\hat{\epsilon} = .75$ in the minorization condition, with $k_0 = 1$. Because the number of dimensions over which the minorization condition must be verified is only two, $N_3 = 16{,}000$ replicate chains from each of four sets of initial values were sufficient. Although there were 61 means parameters in this model, the run time for this step was only two hours.

Finally, we used Rosenthal's theorem with tuning constants $r = .138$ and $M = .675$ to conclude that, if the initial values of the sampler are those that minimize V, then

$$\left\| L\left(X^{(k)}\right) - \pi \right\|_{var} \le (0.25)^{[0.01725k]} + (0.4411)(0.8539)^{[\frac{k}{4}]}$$

If $k = 232$ then this bound is 0.0084, less than our chosen criterion of 0.01.

6.2. *A random-coefficients model*

In the growth-curve or random-coefficients model, the three stages of the standard hierarchical formulation are given as follows:

$$Y_{ij} \mid \alpha_{i0}, \quad \alpha_{i1}, \sigma_y^2 \sim N\left(\alpha_{i0} + \alpha_{i1}x_{ij}, \sigma_y^2\right), \quad i = 1, \ldots, n$$

$$\begin{bmatrix} \alpha_{i0} \\ \alpha_{i1} \end{bmatrix} \Bigg| \begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \quad \Sigma_\alpha \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \Sigma_\alpha = \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix}\right)$$

$$\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix} \Bigg| \begin{bmatrix} M_0 \\ M_1 \end{bmatrix}, \quad \Sigma_0 \sim N\left(\begin{bmatrix} M_0 \\ M_1 \end{bmatrix}, \Sigma_0 = \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}\right)$$

which may be written as

$$\left(\begin{array}{c} \mathrm{Y}_{\boldsymbol{n}\times\mathbf{1}} \\ \hline \mathrm{0}_{\mathbf{2}\boldsymbol{n}\times\mathbf{1}} \\ \hline \mathrm{M}_{\mathbf{2}\times\mathbf{1}} \end{array}\right) = \left(\begin{array}{c|c} \mathrm{X}_{\boldsymbol{n}\times\mathbf{2}\boldsymbol{n}} & \mathrm{0}_{\boldsymbol{n}\times\mathbf{2}} \\ \hline -\mathrm{I}_{\mathbf{2}\boldsymbol{n}\times\mathbf{2}\boldsymbol{n}} & \mathrm{Z}_{\mathbf{2}} \\ \hline \mathrm{0}_{\mathbf{2}\times\mathbf{2}\boldsymbol{n}} & \mathrm{I}_{\mathbf{2}\times\mathbf{2}} \end{array}\right) \left(\begin{array}{c} \alpha \\ \hline \mu \end{array}\right) + \begin{pmatrix} \psi \\ \delta \\ \xi \end{pmatrix}$$

The covariance matrix of the error terms is block diagonal rather than diagonal. The first block is diagonal with all diagonal entries equal to σ_y^2. The next block is block diagonal with each block equal to Σ_α. The final block is equal to Σ_0.

The Bayesian growth-curve model is completed with priors on σ_y^2 and Σ_α.

$$\sigma_y^2 \sim IG(a_1, b_1)$$
$$\begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix} \sim IW((\rho R)^{-1}, \rho)$$

where a_1, b_1 and ρ are known constants, R is a known matrix, and IG and IW denote the inverse gamma and inverse Wishart p.d.f.s respectively.

The full conditional distributions used at iteration k of the blocked sampler are:

$$\sigma_y^{2(k)} \mid \boldsymbol{\alpha_0}^{(\mathbf{k-1})}, \quad \boldsymbol{\alpha_1}^{(\mathbf{k-1})} \sim IG\left(a_1 + \frac{JK}{2}, \; b_1 + \frac{\sum_{i=1}^{K}\sum_{j=1}^{J}\left(Y_{ij} - \alpha_{i0}^{(k-1)} - \alpha_{i0}^{(k-1)}x_{ij}\right)^2\right)}{2}\right) \tag{16}$$

$$\begin{bmatrix} \tau_0^{2(k)} & \tau_{01}^{(k)} \\ \tau_{01}^{(k)} & \tau_1^{2(k)} \end{bmatrix} \Bigg| \alpha^{(k-1)}, \quad \mu^{(k-1)} \sim IW \times \left(\left(\rho R + \sum_{i=1}^{K}\left(\alpha_i^{k-1} - \mu^{k-1}\right)\left(\alpha_i^{k-1} - \mu^{k-1}\right)^T\right)\right)^{-1}, \rho + K\right) \tag{17}$$

$$\Theta^{(\mathbf{k})} = \left[\alpha_{\mathbf{0}}{}^{(\mathbf{k})}, \alpha_{\mathbf{1}}{}^{(\mathbf{k})}, \mu_{\mathbf{0}}^{(\mathbf{k})}, \mu_{\mathbf{1}}^{(\mathbf{k})}\right] \mid \sigma_{\mathbf{y}}^{\mathbf{2(k)}}, \tau_{\mathbf{0}}^{\mathbf{2(k)}}, \tau_{\mathbf{01}}^{(\mathbf{k})}, \tau_{\mathbf{1}}^{\mathbf{2(k)}}$$
$$\sim N([\mathbf{X}^{\mathbf{T}}\Sigma^{-\mathbf{1}}\mathbf{X}]^{-1}\mathbf{X}^{\mathbf{T}}\Sigma^{-\mathbf{1}}\mathbf{Y}, [\mathbf{X}^{\mathbf{T}}\Sigma^{-\mathbf{1}}\mathbf{X}]^{-1}) \tag{18}$$

The V function is specified as:

$$\begin{aligned} V^{**} = {} & \frac{\sum_i \sum_j (Y_{ij} - \alpha_{i0} - \alpha_{i1}x_{ij})^2}{\hat{\sigma}_y^2} + \frac{\sum_i (\alpha_{i0} - \mu_0)^2}{\hat{\tau}_0^2(1 - \hat{\gamma}^2)} \\ & - \frac{2\hat{\gamma}\sum_i (\alpha_{i0} - \mu_0)(\alpha_{i1} - \mu_1)}{\hat{\tau}_0\hat{\tau}_1(1 - \hat{\gamma}^2)} + \frac{\sum_i (\alpha_{i1} - \mu_1)^2}{\hat{\tau}_1^2(1 - \hat{\gamma}^2)} \\ & + \frac{(\mu_0 - \eta_0)^2}{\sigma_0^2} + \frac{(\mu_1 - \eta_1)^2}{\sigma_1^2} \end{aligned} \tag{19}$$

where $\hat{\gamma}^2 = \frac{\hat{\tau}_{01}^2}{\hat{\tau}_0^2\hat{\tau}_1^2}$, and $V = V^{**}/v$, where v is obtained by minimizing V^{**} over $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$.

For illustration we used the dataset from a CIBA-GEIGY study consisting of weights in grams of $K = 30$ baby rats measured at one-week intervals for $J = 5$ weeks, with the prior specifications from the analysis of Gelfand *et al.* (1990). This model has 66 unknown parameters.

Running $N_0 = 1000$ chains of length $m = 12$ (run time 1 minute 23 seconds) produced an estimate $\hat{\Lambda} = 1.443$.

In preparation for estimating λ, we used "proc reg" (SAS version 6.12, SAS Institute, Cary, NC) to compute an individual slope and intercept for each rat and "proc mixed" to estimate the population intercept and slope. For the first set of initial values x_{01}^* we set each α_{i0} and α_{i1} equal to the individual intercept and slope for the corresponding rat and μ_0 and μ_1 equal to the estimated group intercept and slope. The second set of initial values x_{02}^* had not only μ_0 and μ_1 but also α_{i0} and α_{i1} for all i set equal to the estimated group intercept and slope. We randomly generated 20 additional sets of initial values from normal distributions with means x_{01}^* and standard deviations ranging from .025 to 80. From each set of initial values, we ran $N_2 = 500$ 12-iteration samplers (run time 23-1/2 minutes). The resulting estimate was $\hat{\lambda} = .12$.

Finding the "extremes of V_d" for the random-coefficients model is complicated by the fact that the covariance matrix of the error terms is block diagonal rather than diagonal. Details are supplied in the Appendix. Furthermore, the minorization condition in (2) must be verified for all four variance/covariance parameters—σ_y^2, τ_0^2, τ_1^2, and τ_{01}. Based on $\hat{\lambda}$ and $\hat{\Lambda}$ from the drift condition, we specified $d = 4$. To get a sufficiently stable estimate of ϵ required $N_3 = 80{,}000$ replicates. Running 80,000 12-iteration samplers ($m = 12$ and $k_0 = 1$) from each of the 6 combinations of extreme values was the only computationally-burdensome step in the process; it was run overnight and took 11 hours. We constructed four-dimensional bins of different sizes by partitioning each dimension into first 10, then 12, then 14 intervals, producing totals of 10000, 20736, and 38416 bins respectively. Applying the criteria from Section 4.3.1, we conservatively set $\hat{\epsilon} = 0.69$.

To compute the final bound on total variation distance to stationarity in (4) after some trial and error we chose the tuning constants $r = .313$ and $M = 2.5$. Assuming that the initial values of the sampler are those that minimize V, we obtained

$$\left\| L\left(X^{(k)}\right) - \pi \right\|_{var} \leq (0.31)^{[0.0261k]} + (0.3252)(0.7493)^{[\frac{k}{12}]}$$

If $k = 192$, this bound equals .0061, again less than our chosen criterion of 0.01. This bound is only slightly more conservative than that of Johnson (1998), who recommended a burn-in time and skip interval of 150 for this dataset.

7. Discussion

The simulation approach presented here is feasible for approximating theoretical bounds on burn-in times for MCMC samplers for HNLMs. It requires far less *analytic* computation than non-simulation-based theoretical methods and can be applied when m and/or $k_0 > 1$, circumstances under which analytic computations are likely to be impossible. The method produces realistic bounds on required burn-in—tens or hundreds of iterations when sampler convergence is rapid. Furthermore, as illustrated in Section 5, it is far more trustworthy than convergence diagnostics, particularly when the target distribution of the sampler is multimodal or when there are regions of the parameter space in which the sampler may get stuck.

Several quantities—m, N_3, bin sizes, and the tuning constants r and M—are chosen by trial and error. It is reassuring that poor choices of m, r, and/or M will lead to unnecessarily *conservative* bounds, which, while wasteful of computing time, do not increase the risk of incorrect inference. By contrast, it is critical that the bin sizes be chosen small enough and the number of replicate samples N_3 large enough as failure to do so would lead to over-estimation of ϵ and thus to anti-conservative bounds.

In fairness, it must be stated that, with careful thought to producing dispersed initial values, both Johnson's (1998) method and Gelman and Rubin's (1992) multi-chain convergence diagnostic also would have detected convergence failure in the simple examples with improper or bimodal posteriors. However, an advantage of the simulation method as applied to reformulated HNLMs is the ease of defining the most extreme sets of initial values in terms of the resulting sums of squares in full conditionals for covariance parameters.

The convergence results obtained using the simulation method for HNLMs apply only to samplers using the blocked algorithm described in Section 3. As demonstrated in Sargent, Hodges and Carlin (2000), samplers that generate means parameters from different levels of the hierarchical model in separate blocks generally converge far more slowly.

As illustrated in the random-coefficients example, the limiting factor in practical application of this method is the number of dimensions over which the minorization condition must be verified. For reformulated hierarchical normal models, this is the number of variance/covariance parameters. The number of sets of initial values that constitute the extremes of V_d increases with this dimension. Even worse, the number of bins increases exponentially with the number of dimensions, requiring a larger number of replicate chains from each starting value. With the computing speed of desktop workstations or PCs today, run times for verifying the minorization condition would be prohibitive for models with more than five or six covariance parameters. Thus this simulation method is particularly applicable to hierarchical normal spatial and spatio-temporal models involving many means parameters but few covariance parameters.

Appendix

Inspection of (16)–(18) reveals that the four quantities that must be bounded for the growth-curve example are $\sum_i \sum_j (Y_{ij} - \alpha_{i0} - \alpha_{i1}x_{ij})^2$, $\sum_i (\alpha_{i0} - \mu_0)^2$, $\sum_i (\alpha_{i0} - \mu_0)(\alpha_{i1} - \mu_1)$ and $\sum_i (\alpha_{i1} - \mu_1)^2$. The lower bound on the first is the sum of the error sums of squares from individual subject-specific linear regressions, henceforth denoted *ESS*. Note that the second through fourth terms on the right side of (22) constitute a quadratic form. Therefore

$$0 \le \frac{\sum_i (\alpha_{i0} - \mu_0)^2}{\hat{\tau}_0^2 (1 - \hat{\gamma}^2)} - \frac{2\hat{\gamma} \sum_i (\alpha_{i0} - \mu_0)(\alpha_{i1} - \mu_1)}{\hat{\tau}_0 \hat{\tau}_1 (1 - \hat{\gamma}^2)} + \frac{\sum_i (\alpha_{i1} - \mu_1)^2}{\hat{\tau}_1^2 (1 - \hat{\gamma}^2)}$$

If it is assumed that the lower bounds on $(\mu_0 - \eta_0)^2$ and $(\mu_1 - \eta_1)^2$ are also zero, then the upper bound on $\sum_i \sum_j (Y_{ij} - \alpha_{i0} - \alpha_{i1}x_{ij})^2$ is $\hat{\sigma}_y^2 vd$, where v is defined in (22).

The components of the quadratic form in (22) must be considered together. The lower bound for both $\sum_i (\alpha_{i0} - \mu_0)^2$ and $\sum_i (\alpha_{i1} - \mu_1)^2$ is 0. Suppose that $\sum_i (\alpha_{i0} - \mu_0)^2 = 0$. Then $\alpha_{i0} - \mu_0 = 0$ for each i; therefore $\sum_i (\alpha_{i0} - \mu_0)(\alpha_{i1} - \mu_1)$ must also equal 0, and $\sum_i (\alpha_{i1} - \mu_1)^2 \le [vd - \frac{ESS}{\hat{\sigma}_y^2}][\hat{\tau}_1^2(1 - \hat{\gamma}^2)]$. Similarly, if $\sum_i (\alpha_{i1} - \mu_1)^2 = 0$, then $\sum_i (\alpha_{i0} - \mu_0)(\alpha_{i1} - \mu_1) = 0$ and $\sum_i (\alpha_{i0} - \mu_0)^2 \le [vd - \frac{ESS}{\hat{\sigma}_y^2}][\hat{\tau}_0^2(1 - \hat{\gamma}^2)]$.

Finally, the largest-magnitude possible negative and positive values for $\sum_i (\alpha_{i0} - \mu_0)(\alpha_{i1} - \mu_1)$, subject to $V^{**} \le vd$, are required. By the Schwarz inequality,

$$\left| \sum_i (\alpha_{i0} - \mu_0)(\alpha_{i1} - \mu_1) \right| \le \sqrt{\sum_i (\alpha_{i0} - \mu_0)^2 \sum_i (\alpha_{i1} - \mu_1)^2}$$

A straightforward application of Lagrange multipliers to maximize $\sum_i (\alpha_{i0} - \mu_0)^2 \sum_i (\alpha_{i1} - \mu_1)^2$ subject to the constraint, $\frac{\sum_i (\alpha_{i0} - \mu_0)^2}{\hat{\tau}_0^2 (1 - \hat{\gamma}^2)} - \frac{2\hat{\gamma}\sqrt{\sum_i (\alpha_{i0} - \mu_0)^2 \sum_i (\alpha_{i1} - \mu_1)^2}}{\hat{\tau}_0 \hat{\tau}_1 (1 - \hat{\gamma}^2)} + \frac{\sum_i (\alpha_{i1} - \mu_1)^2}{\hat{\tau}_1^2 (1 - \hat{\gamma}^2)} = vd - \frac{ESS}{\hat{\sigma}_y^2}$, yields a positive upper bound

$$\sum_i (\alpha_{i0} - \mu_0)(\alpha_{i1} - \mu_1) \le \frac{\hat{\tau}_0 \hat{\tau}_1 \left(vd - \frac{ESS}{\hat{\sigma}_y^2}\right)(1 - \hat{\gamma}^2)}{2(1 - \hat{\gamma})}$$

which can be attained only if

$$\sum_i (\alpha_{i0} - \mu_0)^2 = \frac{\hat{\tau}_0^2 \left(vd - \frac{ESS}{\hat{\sigma}_y^2}\right)(1 - \hat{\gamma}^2)}{2(1 - \hat{\gamma})}$$

and

$$\sum_i (\alpha_{i1} - \mu_1)^2 = \frac{\hat{\tau}_1^2 \left(vd - \frac{ESS}{\hat{\sigma}_y^2}\right)(1 - \hat{\gamma}^2)}{2(1 - \hat{\gamma})}$$

The same procedure, but with a plus sign before the second term in the constraint, produces a negative lower bound

$$\sum_i (\alpha_{i0} - \mu_0)(\alpha_{i1} - \mu_1) \ge -\frac{\hat{\tau}_0 \hat{\tau}_1 \left(vd - \frac{ESS}{\hat{\sigma}_y^2}\right)(1 - \hat{\gamma}^2)}{2(1 + \hat{\gamma})}$$

which can be attained only if

$$\sum_i (\alpha_{i0} - \mu_0)^2 = \frac{\hat{\tau}_0^2 \left(vd - \frac{ESS}{\hat{\sigma}_y^2}\right)(1 - \hat{\gamma}^2)}{2(1 + \hat{\gamma})}$$

and

$$\sum_i (\alpha_{i1} - \mu_1)^2 = \frac{\hat{\tau}_1^2 \left(vd - \frac{ESS}{\hat{\sigma}_y^2}\right)(1 - \hat{\gamma}^2)}{2(1 + \hat{\gamma})}$$

Consequently, six combinations of quantities constitute the "extremes of V_d" for the growth-curve model.

Acknowledgments

This research was supported in part by NSF/EPA grant number 1 3802400. The author thanks Drs. Jim Hodges and Jeff Rosenthal for helpful input.

References

Besag J. and Higdon D.M. 1999. Bayesian analysis of agricultural field experiments. Journal of the Royal Statistical Society, Series B 61: 691–717.

Besag J. and Kempton R.A. 1986. Statistical analysis of field experiments using neighbouring plots. Biometrics 42: 231–251.

Brooks S.P. and Roberts G.O. 1998. Convergence assessment techniques for Markov chain Monte Carlo. Statistics and Computing 8: 319–335.

Cowles M.K. and Carlin B.P. 1996. Markov chain Monte Carlo convergence diagnostics: A comparative review. Journal of the American Statistical Assocation 91: 883–904.

Cowles M.K., Roberts G.O., and Rosenthal J.S. 1999. Possible biases induced by MCMC convergence diagnostics. Journal of Statistical Computation and Simulation 64: 87–104.

Cowles M.K. and Rosenthal J.S. 1998. A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. Statistics and Computing 8: 115–124.

Gelfand A.E., Hills S.E., Racine-Poon A., and Smith A.F.M. 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. Journal of the American Statistical Association 85: 972–985.

Gelman A. and Rubin D.J. 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7: 457–511.

Gelman A., Carlin J.B., Stern H.S., and Rubin D.B. 1995. Bayesian Data Analysis, Chapman and Hall, London.

Green P.J. and Murdoch D.J. 1999. Exact sampling for Bayesian inference: Towards general purpose algorithms. In: Bernardo J.M., Berger J.O., Dawid A.P., and Smith A.F.M. (Eds.), Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting, Oxford University Press, Oxford.

Hobert J.P. and Casella G. 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. Journal of the American Statistical Association 91: 1461–1473.

Hodges J.S. 1998. Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). J. Roy. Stat. Soc., Series B 60: 497–536.

Johnson V.E. 1998. A Coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. Journal of the American Statistical Association 93: 238–248.

Jones G.L. and Hobert J.P. 2001. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. Technical report, University of Florida.

Liu J. and Hodges J.S. 2002. Posterior bimodality in the balanced one-way random effects model. Journal of the Royal Statistical Society, Series B: to appear.

Mengersen K.L., Robert C.P., and Guihenneuc-Joyaux C. 1999. Markov chain Monte Carlo convergence diagnostics: A review. In: Bernardo J.M., Berger J.O., Dawid A.P., and Smith A.F.M. (Eds.), Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting, Oxford University Press, Oxford.

Meyn S.P. and Tweedie R.L. 1994. Computable bounds for convergence rates of Markov chains. Annals of Appled Probability 4: 981–1011.

Montgomery D.C. 1991. Design and Analysis of Experiments, 3rd ed., Wiley, New York.

Propp J.G. and Wilson B.M. 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Structures and Algorithms 9(1/2): 223–252.

Roberts G.O. and Tweedie R.L. (1999). Bounds on regeneration times and convergence rates for Markov chains. Stochastic Processes and their Applications 80: 211–229; Correction 91: 337–338.

Rosenthal J.S. 1995. Minorization conditions and convergence rates for Markov chain Monte Carlo. Journal of the American Statistical Association 90: 558–566; Correction p. 1136.

Sargent D.J., Hodges J.S., and Carlin B.P. 2000. Structured Markov chain Monte Carlo. Journal of Computational and Graphical Statistics 9(2): 217–234.

Smith B.J. 2000. Bayesian Output Analysis Program (BOA), Version 0.5.0 for S-PLUS and R. Available at http://www.public-health.uiowa.edu/BOA.

Spiegelhalter D., Thomas A., Best N, and Gilks W. 1995. BUGS: Bayesian inference Using Gibbs Sampling, Version 0.5, Vol. 1. MRC Biostatistics Unit, Cambridge, England.