A simulation approach to convergence rates for Markov chain Monte Carlo algorithms

by

Mary Kathryn Cowles* and Jeffrey S. Rosenthal**

Abstract. Markov chain Monte Carlo (MCMC) methods, including the Gibbs sampler and the Metropolis-Hastings algorithm, are very commonly used in Bayesian statistics for sampling from complicated, high-dimensional posterior distributions. A continuing source of uncertainty is how long such a sampler must be run in order to converge approximately to its target stationary distribution.

Rosenthal (1995b) presents a method to compute rigorous theoretical upper bounds on the number of iterations required to achieve a specified degree of convergence in total variation distance by verifying drift and minorization conditions. We propose the use of auxiliary simulations to estimate the numerical values needed in Rosenthal's theorem.

Our simulation method makes it possible to compute quantitative convergence bounds for models for which the requisite analytical computations would be prohibitively difficult or impossible. On the other hand, although our method appears to perform well in our example problems, it can not provide the guarantees offered by analytical proof.

Keywords. drift condition, Gibbs sampler, Metropolis-Hastings algorithm, Markov chain Monte Carlo, minorization condition, ordinal probit, variance components.

Acknowledgements. We thank Brad Carlin for assistance and encouragement.

* Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A. Internet: cowles@sdac.harvard.edu.

** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Internet: jeff@utstat.toronto.edu.

1. Introduction.

Markov chain Monte Carlo (MCMC) methods, such as the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990; Smith and Roberts, 1993) and the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), have become a very popular means for sampling from complicated, high-dimensional posterior distributions in Bayesian statistics. A continuing source of uncertainty is the rate of convergence of MCMC algorithms. Specifically, how long should they be run until they have approximately converged to their target stationary distribution? That is, how large should the "burn-in" time be?

Rigorous theoretical upper bounds on burn-in times for these algorithms have recently been proposed (see e.g. Frieze, Kannan, and Polson, 1993; Frigessi, Hwang, Sheu, and di Stefano, 1993; Ingrassia, 1994; Meyn and Tweedie, 1994; Rosenthal, 1995b; Baxendale, 1994). However, they have suffered from the difficulties of precise analysis of complicated models, and have largely tended to concentrate on relatively simple problems, and/or to provide impractically large upper bounds.

Consequently, most applied users of MCMC techniques have used convergence diagnostics (see for example, Roberts, 1992; Gelman and Rubin, 1992; Raftery and Lewis, 1992) to assess convergence. These diagnostics often work well in practice; however they are not completely understood and offer no guarantees. See Cowles and Carlin (1996) for a comprehensive review.

In this paper, we present a way to make use of theoretical upper bounds (taken from Rosenthal, 1995b) without doing prohibitively difficult computations. Specifically, we consider the use of auxiliary simulations to numerically verify certain hypotheses (drift and minorization conditions) which are known to provide upper bounds on convergence times. The auxiliary simulations provide numerical values which may then be used in the theoretical results. Our approach is thus an attempt to bridge the gap between theoretical and applied work, making use of the theory while providing a practical method for exploiting it. Details are given in the next section.

After presenting our general method, we apply it to three examples of MCMC. The first (Section 3) is for a model for which upper bounds have already been proven analytically (Rosenthal, 1996). This model thus allows us to check our method against a known answer,

and we find excellent agreement of our method with the theoretical results. Our second example (Section 4) is a variance components model, long advocated (Gelfand and Smith, 1990; Gelfand et al., 1990) as an ideal candidate for the Gibbs sampler. Our third example (Section 5) is a Gibbs sampler for an ordinal probit model, as used in biostatistics contexts (Carlin and Polson, 1992; Albert and Chib, 1993; Cowles, 1996).

In all three of these models, we use our auxiliary simulation method to obtain useful, quantitative bounds on the convergence time of the Markov chain being studied.

We note that, in addition to burn-in, there are other aspects of "convergence" that are relevant to applied use of MCMC methods (but are not directly considered in this paper). These include: determining whether the chain has traversed the entire sample space; and obtaining reasonable estimates of the variances of quantities that are estimated from the dependent samples produced by MCMC algorithms. We discuss these issues briefly in the final section.

We end this section with some notation. We shall consider Markov chains on a general state space $\mathcal{X}$ (usually $\mathcal{X} \subseteq \mathbf{R}^n$), with transition probabilities $P(x, \cdot)$, initial distribution $\nu(\cdot)$, and target stationary distribution $\pi(\cdot)$. In many applications, including all of our examples, $\pi(\cdot)$ is the posterior distribution for a Bayesian statistical model.

We shall concentrate largely on the sequentially-updated Gibbs sampler. There, the state space is a product $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_n$. (Note that the $\mathcal{X}_i$ themselves may be either one- or multi-dimensional.) The Gibbs sampler proceeds by sequentially updating each coordinate from the conditional distribution induced by $\pi(\cdot)$. Specifically, given the state $(x_1^{(k-1)}, x_2^{(k-1)}, \ldots, x_n^{(k-1)})$ at time $k-1$, it chooses

$$\begin{aligned}
x_1^{(k)} &\sim \pi(dx_1 \mid x_2^{(k-1)}, x_3^{(k-1)}, \ldots x_n^{(k-1)}) \,; \\
x_2^{(k)} &\sim \pi(dx_2 \mid x_1^{(k)}, x_3^{(k-1)}, \ldots x_n^{(k-1)}) \,; \\
&\vdots \\
x_n^{(k)} &\sim \pi(dx_n \mid x_1^{(k)}, x_2^{(k)}, \ldots x_{n-1}^{(k)}) \,.
\end{aligned}$$

We shall use this notation in Lemma 2 below.

2. A general method for bounding convergence rates.

For a Markov chain $\{X^{(k)}\}_{k=0}^{\infty}$ on a state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$, we are interested in bounding the total variation distance to stationarity, defined by

$$\|\mathcal{L}(X^{(k)}) - \pi\| \equiv \sup_{S \subseteq \mathcal{X}} |P(X^{(k)} \in S) - \pi(S)| .$$

We begin by taking a result from Rosenthal (1995b), which gives an upper bound on $\|\mathcal{L}(X^{(k)}) - \pi\|$. The following follows easily from Theorem 5 there. [For the special case $m = k_0 = M = 1$, it essentially coincides with Theorem 12 there. In general, it follows immediately by applying Rosenthal's Theorem 5 to the chain P^m, with $j = rk$, with $C = V_d = \{x \in \mathcal{X} \,;\, V(x) \leq d\}$, and with the drift function $h(x, y) = 1 + MV(x) + MV(y)$; or, if $V \geq 1$, with $h(x, y) = \frac{M}{2}\,(V(x) + V(y)) + (1 - M)$.]

Proposition 1. *Let $P(x, \cdot)$ be the transition probabilities for a Markov chain on a state space $\mathcal{X}$, with initial distribution ν and stationary distribution π. Suppose for some non-negative function $V : \mathcal{X} \to \mathbf{R}^{\geq 0}$, some $\lambda < 1$ and $\Lambda < \infty$, some $\epsilon > 0$, some probability measure $Q(\cdot)$ on $\mathcal{X}$, some positive integers m and k_0, and some $d > \frac{2\Lambda}{1-\lambda}$, we have*

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

and also

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

where $V_d = \{x \in \mathcal{X}; V(x) \leq d\}$. Then for any $0 < r < 1$ and $M > 0$, we have

$$\|\mathcal{L}(X^{(k)}) - \pi\| \;\leq\; (1 - \epsilon)^{[rk/mk_0]} \;+\; C_0\,(\alpha A)^{-1} \left(\alpha^{-(1-rk_0)} A^r\right)^{[k/m]} ,$$

where

$$\alpha^{-1} \;=\; \frac{1 + 2M\Lambda + M\lambda d}{1 + Md}; \quad A \;=\; 1 + 2(\lambda M d + M\Lambda) \,; \quad C_0 \;=\; \left(1 + \frac{M\Lambda}{1-\lambda} + M E_\nu(V(X^{(0)}))\right) .$$

If furthermore it is known that $V(x) \geq 1$ for all $x \in \mathcal{X}$, then it suffices that $d > \frac{2\Lambda}{1-\lambda} - 1$, and these values may be improved slightly to

$$\alpha^{-1} \;=\; \lambda + \frac{M\Lambda + (1-\lambda)(1-M)}{1 + \frac{M}{2}(d-1)} \,; \qquad A \;=\; M(\lambda d + \Lambda) + (1 - M) \,;$$

$$C_0 \;=\; \frac{M}{2}\left(\frac{\Lambda}{1-\lambda} + E_\nu(V(X^{(0)}))\right) + (1-M)\,.$$

We note that the hypotheses imply that $\alpha^{-1} < 1$. Hence, choosing $r > 0$ sufficiently small, this proposition provides a quantitative, exponentially-decreasing upper bound on the total variation distance $\|\mathcal{L}(X^{(k)}) - \pi\|$ between the distribution of our Markov chain after k iterations, and the stationary distribution $\pi(\cdot)$. Hence, for a given MCMC algorithm, it is then possible to choose an appropriate value of k to make this distance as small as desired.

We further note that, in principle at least, Proposition 1 and the methods of this paper can be applied to *any* discrete-time Markov chain. In particular, the state space can be finite, countably infinite, or uncountable; the chain can be reversible or not; the chain could arise from a Gibbs sampler, or a Metropolis-Hastings algorithm, or a hybrid Monte Carlo algorithm, or whatever; and so on. Of course, it will still be computationally difficult to apply Proposition 1 to very complicated chains.

To apply Proposition 1, it is necessary to choose a function V, and then to verify the drift condition (1) and the minorization condition (2). We discuss these issues in turn.

The selection of V is non-trivial. Clearly, V need depend only on those parameters which are "remembered" at the next iteration, i.e. on those parameters for which initial values must be supplied. For example, for the sequentially-updated Gibbs sampler, if $X_1^{(k)}, ..., X_n^{(k)}$ are conditionally independent of $X_1^{(k-1)}, ..., X_G^{(k-1)}$, given $X_{G+1}^{(k-1)}, ..., X_n^{(k-1)}$, then V need depend only on $X_{G+1}, ..., X_n$. (This always holds with $G = 1$.) Such observations are used in all of the examples in this paper. Furthermore, V may depend on the data, on any constants associated with the model, and on various numerical values chosen by the user.

The conditions (1) and (2) 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 $V(x) \leq d$. By keeping these two goals in mind, and by qualitatively examining the behavior of the chain, a reasonable choice of V can sometimes be made by inspection. Furthermore, if conditions (1) and (2) can be verified for any function V,

then by Proposition 1 this implies a bound on the total variation distance of the chain to stationarity; we do not need to worry if we have made the "best" choice of V.

On the other hand, verification of equations (1) and (2) (for a given function V) can be quite difficult, especially for complicated, high-dimensional statistical models. Furthermore, to get good values of λ and ϵ it is often desirable to have $k_0 > 1$ or $m > 1$, and in this case analytic verification is often practically impossible. This has tended to limit the effectiveness of theoretical analysis for such models. (We do note that verification of (1) alone is sufficient to establish geometric ergodicity of the Markov chain, but without providing a quantitative bound. See Meyn and Tweedie, 1993; Roberts and Tweedie, 1994; Geyer, 1994.)

Our approach is to *approximately* verify equations (1) and (2) numerically, through auxiliary Monte Carlo computer simulation. This has the disadvantage that it does not provide rigorous proofs of the convergence rates. However, by doing careful Monte Carlo estimation, including computation of standard errors, we provide results which appear to be quite convincing. As a test case, our method works very well on a problem for which analytic results are also available (Section 3). Furthermore, our method is much more straightforward to implement than is theoretical analysis, especially for complicated models or for $mk_0 > 1$.

Our method makes use of the following two results for simplifying the computation of ϵ above. They are taken from Lemmas 6 and 7 of Rosenthal (1995b). The first, specific to the sequentially-updated Gibbs sampler, reduces the computation of the minorization condition on all n variables, to a minorization on only the first $D < n$ variables. The second, for general Markov chains with densities, gives a formula for ϵ in terms of an integral of minimums of densities. (This integration is often not feasible directly, especially for $mk_0 > 1$; however it is the inspiration for step three of our method below.)

Lemma 2. *Consider a sequentially-updated Gibbs sampler, as above. Suppose that for some D, conditional on values for $X_1^{(k)}, \ldots, X_D^{(k)}$, the random variables $X_{D+1}^{(k)}, \ldots, X_n^{(k)}$ are conditionally independent of all $X_i^{(k')}$ for all $k' < k$. (For example, this always holds with $D = n - 1$.) Suppose further that for some $R \subseteq \mathcal{X}$ and $\epsilon > 0$, there is a probability*

measure $Q(\cdot)$ on $\mathcal{X}_1 \times \ldots \times \mathcal{X}_D$ such that

$$\mathcal{L}(X_1^{(mk_0)}, \ldots, X_D^{(mk_0)} \mid (X_1^{(0)}, \ldots, X_n^{(0)}) = x) \;\geq \epsilon\, Q(\cdot)\,, \qquad \text{for all } x \in R\,.$$

Then there is a probability measure $Q'(\cdot)$ on $\mathcal{X}$ such that

$$P^{mk_0}(x, \cdot) \;\geq\; \epsilon\, Q'(\cdot)\,, \qquad \text{for all } x \in R\,.$$

Lemma 3. *Suppose a Markov chain satisfies that $P^{mk_0}(x, \cdot) = f(x, y)dy$, where $f(x, \cdot)$ is a density function and dy is Lebesgue measure (or some other σ-finite reference measure). Then there exists a probability measure $Q(\cdot)$ such that*

$$P^{mk_0}(x, \cdot) \geq \epsilon\, Q(\cdot) \qquad \text{for all } x \in R\,,$$

where

$$\epsilon \;=\; \int_{\mathcal{X}} \left(\inf_{x \in R} f(x, y) \right) dy\,.$$

Remark. Strictly speaking, it is possible (though rare) that the function $\inf_{x \in R} f(x, y)$ may not be integrable. In that case, the definition of ϵ above should be taken to be a *lower integral* (cf. Spivak, 1980, p. 277). (Equivalently, we may take $\epsilon = \int_{\mathcal{X}} g(y)dy$ for any integrable function g satisfying $g(y) \leq f(x, y)$ for all $x \in R$ and $y \in \mathcal{X}$.)

Our method is designed to approximately verify conditions (1) and (2), after the selection of a function V has been made. It consists of three steps.

First, we find a lower bound on Λ, as follows. For each point $x \in \mathcal{X}$ such that $V(x) = 0$, we simulate N_0 draws from $\mathcal{L}(X^{(m)}|X^{(0)} = x)$, and thus estimate $E(V(X^{(m)})|X^{(0)} = x)$ as the mean of $V(X^{(m)})$ over the N_0 draws. N_0 is chosen to obtain a standard error of this mean that is less than or equal to some desired tolerance. The maximum of these estimated expected values, over different choices of x, provides a lower bound $\widehat{\Lambda}$. (If we know that $V \geq 1$, then we instead apply this procedure to $V - 1$, and then add 1 to our resulting lower bound.)

Second, for a given value of $\widehat{\Lambda}$ (at least as large as the previously computed lower bound), we estimate a corresponding value for λ as follows. We generate N_1 different

initial values $x \in \mathcal{X}$ randomly, from some appropriate scheme designed to make them cover all potentially "bad" parts of the space. (In practice, we generate them from various normal distributions with a variety of variances.) For each such initial value x, we simulate N_2 (again chosen to obtain satisfactory standard errors) draws from $\mathcal{L}(X^{(m)}|X^{(0)} = x)$, and thus estimate $e(x) = E(V(X^{(m)})|X^{(0)} = x)$. The maximum of

$$(e(x) - \widehat{\Lambda})/V(x), \tag{3}$$

over different choices of x, provides an estimate $\widehat{\lambda}$ corresponding to the given $\widehat{\Lambda}$. (Note that the estimate is less stable, and requires a larger value of N_2, if $V(x)$ is close to 0. This is a motivation for choosing functions V satisfying $V \geq 1$.) If $\widehat{\lambda} < 1$, then we have found evidence for a useful drift condition. If not, then we increase our value of $\widehat{\Lambda}$ and try again.

Third, for an appropriate value of d chosen to be comfortably larger than $2\widehat{\Lambda}/(1 - \widehat{\lambda})$, we estimate a corresponding value for ϵ. To do this, we divide our state space (or at least those coordinates over which a minorization is required, according to Lemma 2) into a large number of little "bins," designed to be small enough so that transition probabilities have densities which are roughly constant over each bin. We then generate a set of initial values x, each in $V_d = \{x \in \mathcal{X}\,;\, V(x) \leq d\}$, designed so that transition probabilities from these different initial values have minimal overlap among all choice of $x \in V_d$. (In practice, we do this by inspection, choosing starting values from all of the "corners" of the set V_d.) For each initial value x, we generate N_3 different samples from $\mathcal{L}(X^{(mk_0)}|X^{(0)} = x)$, and keep track of what fraction of them land in each of our little bins. We then compute an estimate $\widehat{\epsilon}$ by summing, over all little bins, the *minimum* over different choices of x, of the fraction of samples landing in that bin. This approximates the sum of $\int\limits_{B_j} \left(\inf\limits_{x \in R} f(x, y) \right) dy$, summed over all the little bins B_j. [Formally, it is necessary *first* to ensure that the bins are sufficiently small to avoid fluctuation in the densities, and *then* to ensure that N_3 is sufficiently large for that particular choice of bin size. In practice, this means that for a given bin size, we should choose N_3 larger and larger until the results appear stable. We should then repeat this process for smaller and smaller bin sizes, choosing larger and larger N_3 for each new bin size, until the resulting estimate $\widehat{\epsilon}$ appears to be stable as the bin size decreases. In sum, appropriate size of bins and of N_3 is a delicate question and can require

some experimentation.]

(We note that the bound of Proposition 1 improves with *smaller* values of Λ and λ, but with *larger* values of ϵ. Thus, to be conservative, we round up our estimates $\widehat{\Lambda}$ and $\widehat{\lambda}$, and round down our estimates $\widehat{\epsilon}$, by amounts at least comparable to the observed standard error of the estimate. We thus obtain bounds which are protected against numerical errors in the auxiliary simulations.)

Having found estimates $\widehat{\Lambda}$, $\widehat{\lambda}$, and $\widehat{\epsilon}$, we then obtain an estimate for a bound on the convergence rate of our chain, by using Proposition 1. In applying the estimate, we are free to choose r and M as we wish (subject to $0 < r < 1$ and $M > 0$); some experimenting with different values is recommended. Generally speaking, we will say that k iterations suffice to achieve convergence if, for some r and M, the bound can be made to be less than 0.01 for the particular choice of k. We emphasize that such a k is an *upper bound* on the time to convergence. Thus, running k iterations to achieve burn-in is sufficient, but may be overly conservative.

Remark. Examining the proof of Rosenthal (1995b, Theorem 12), we see that condition (1) above is used to bound exponential moments of the return times of the chain to the set V_d. In theory, it might be possible to forget about condition (1), and instead use auxiliary simulation to estimate these exponential moments directly. However, in practice, the resulting estimates would be extremely unstable due to the heavy-tail behavior. Thus, we do not consider them further here.

3. Example: A model related to James-Stein estimators.

We first try our method on a model related to James-Stein estimators, taken from Rosenthal (1996), which followed a suggestion of Jun Liu. This model is similar to the full variance components model, but is simpler in that one of the components of variance is fixed. Since Rosenthal (1996) analytically obtained numerical convergence bounds for this problem with $m = k_0 = 1$, we can use it as a check of our method.

This model is defined by

$$Y_i \,|\, \theta_i \;\sim\; N(\theta_i, v) \qquad (1 \le i \le K)$$

$$\theta_i \,|\, \mu, A \;\sim\; N(\mu, A) \qquad (1 \le i \le K)$$

Here $Y_1, \ldots, Y_K$ are observed data, μ has a flat prior, v is an (empirically estimated) constant, and A has prior $IG(a, b)$ for fixed constants a and b. We are interested in the posterior distribution

$$\pi(\cdot) \;=\; \mathcal{L}(A, \mu, \theta_1, \ldots, \theta_K \,|\, Y_1, \ldots, Y_K)\,.$$

The Gibbs sampler acts on the $(K+2)$-dimensional space $(A, \mu, \theta_1, \ldots, \theta_K)$, conditional on data $Y_1, \ldots, Y_K$ and constants a and b, as follows. After choosing initial values $A^{(0)}, \mu^{(0)}, \theta_1^{(0)}, \ldots, \theta_K^{(0)}$ from some initial distribution, it updates these variables repeatedly (for iterations $k = 1, 2, 3, \ldots$) by the conditional distributions

$$A^{(k)} \;\sim\; \mathcal{L}(A \,|\, \theta_i = \theta_i^{(k-1)}, Y_i) \;=\; IG\left(a + \frac{K-1}{2},\, b + \frac{1}{2}\sum(\theta_i^{(k-1)} - \overline{\theta}^{(k-1)})^2\right)\,; \quad (4)$$

$$\mu^{(k)} \;\sim\; \mathcal{L}(\mu \,|\, A = A^{(k)}, \theta_i = \theta_i^{(k-1)}, Y_i) \;=\; N(\overline{\theta}^{(k-1)}, A^{(k)}/K)\,; \quad (5)$$

$$\theta_i^{(k)} \;\sim\; \mathcal{L}(\theta_i \,|\, A = A^{(k)}, \mu = \mu^{(k)}, Y_i) \;=\; N\left(\frac{\mu^{(k)} v + Y_i A^{(k)}}{v + A^{(k)}}, \frac{A^{(k)} v}{v + A^{(k)}}\right)\,; \quad (6)$$

where $\overline{\theta}^{(k)} = \frac{1}{K}\sum \theta_i^{(k)}$. For derivation of these conditional distributions and further details, see Rosenthal (1996). In particular, note that the updating order was chosen so that we may take $D = 2$ in Lemma 2.

For the data $\{Y_i\}$, we use the baseball data presented in Morris (1983, Table 1). This data has $K = 18$, $v = 0.00434$, and $\Delta = \sum_i (Y_i - \overline{Y})^2 = 0.0822$. We further choose prior values $a = -1$ and $b = 2$.

This model was analyzed rigorously in Rosenthal (1996). There, it was shown that for this data, with $m = k_0 = 1$ and $d = 1$, we may satisfy equations (1) and (2) with

$$V(A, \mu, \theta_1, \ldots, \theta_K) \;=\; \sum_{i=1}^{K} (\theta_i - \overline{Y})^2\,; \quad (7)$$

$$\lambda = 0.000289; \qquad \Lambda = 0.161; \qquad \epsilon = 0.0656\,. \quad (8)$$

This led to the useful bound

$$\|\mathcal{L}(X^{(k)}) \;-\; \pi(\cdot)\| \;\le\; (0.967)^k + (0.935)^k \left(1.17 + E\left(\sum(\theta_i^{(0)} - \overline{Y})^2\right)\right)\,,$$

which equals 0.009 if (say) $k = 140$ and $\theta_i^{(0)} \equiv \overline{Y}$. Thus, it was shown that approximately 140 iterations suffice to achieve convergence of this Gibbs sampler.

To compare these results with a commonly-used convergence diagnostic, Figure 1 shows the traces of A, μ, θ_1, and θ_2 from three Gibbs sampler chains, each run for 500 iterations. The median and .975 quantile of Gelman and Rubin's (1992) convergence diagnostic are shown above each plot. Certainly the visual impression is that the chains indeed are drawing from the same target distribution well before the 200th iteration, but Gelman and Rubin's diagnostic suggests that more iterations are needed before we can have any confidence that all the chains have traversed the entire state space. Thus, the theoretical bounds appear to be sensible, but do not give all the information needed for estimation purposes.

We proceed to apply our simulation method to the same problem.

Choosing the V() function. With the conditional distributions given in (4) - (6), initial values are needed only for $\theta_1, \ldots, \theta_K$. Thus we adopted the V function given in (7).

Obtaining a lower bound for Λ. With $V(x)$ defined as in (7), $V(x_0) = 0$ only if x_0 corresponds to $\theta_i^{(0)} = \overline{Y}$ for all $i = 1, \ldots, K$. Thus, to determine a lower bound on Λ in (1), we ran $N_0 = 30,000$ single-iteration chains, all started with all $\theta_i^{(0)} = \overline{Y}$, and obtained $\widehat{\Lambda} = 0.157$, the mean of $V(X^{(1)}|X^{(0)} = x_0)$, with a standard error of .00045. We rounded this mean up to 0.16, which agreed well with the value obtained analytically.

Obtaining a lower bound for λ. We generated 5 sets of initial values $\theta_i^{(0)}, i = 1, \ldots, K$, from each of 4 normal distributions, all centered at $\overline{Y}$, and with standard deviations ranging from 0.05 to 0.50. To keep the expression (3) from "blowing up," any set of initial values that produced $V(X^{(0)}) < .01$ was rejected and redrawn. From each set of initial values $x_0^{(l)}, l = 1, \ldots, N_1 = 20$, we ran a minimum of $N_2 = 500$ single-iteration chains and computed the mean $V(X^{(1)}|X^{(0)} = x_0^{(l)})$. When $V(x_0^{(l)}) < 1$, N_2 was multiplied by a suitable constant so that the standard errors of these means were approximately constant, with none larger than .0021. The largest value of (3) gave $\widehat{\lambda} = .0011$. If there had been a trend toward larger values of (3) with larger standard deviations of the normal generating distribution for the starting values, we would have repeated the process with still larger standard deviations. Because instead the mean $V(X^{(1)}|X^{(0)} = x_0^{(l)})$ did not vary greatly

as a function of $x_0^{(l)}$, we concluded that our choices of starting places had covered the space satisfactorily.

Estimating ϵ in the minorization condition. Our estimates $\widehat{\Lambda} = 0.16$ and $\widehat{\lambda} = 0.0011$ enabled us to choose $d = 1$ as in Rosenthal (1996), to ensure that $d > \frac{2\widehat{\Lambda}}{1-\widehat{\lambda}}$. By Lemma 2, we need establish overlap in the transition probabilities only for A and μ. By inspection of the full conditionals (4) and (5) and computations similar to those in Rosenthal (1996), we find that the starting points with the least overlap, subject to the constraint that $V(x_0) \leq d$, are those such that the pair $\left(\sum(\theta_i^{(0)} - \overline{\theta}^{(0)})^2,\ \overline{\theta}^{(0)}\right)$ equals one of the four choices

$$\left(d, \overline{Y}\right); \qquad \left(0, \overline{Y}\right); \qquad \left(0,\ \overline{Y} - \sqrt{d/K}\right); \qquad \left(0,\ \overline{Y} + \sqrt{d/K}\right).$$

We ran $N_3 = 40000$ single-iteration Gibbs sampler chains from each of these four starting points. We then used an S-Plus routine to construct a grid of 2-dimensional bins spanning the range of the $A's$ and the $\mu's$ in the 4 samples combined, to compute what fraction of the points in each sample fell into each of the bins, and finally to estimate ϵ by summing over all bins the minimum fraction falling into that bin from each of the 4 samples.

We used three different grid sizes to assess our estimate of ϵ. When each dimension was chopped into 16 equal-length intervals, $\widehat{\epsilon} = .0958$. When each dimension was chopped into 32 intervals, $\widehat{\epsilon} = .0726$, and with 64×64 bins, $\widehat{\epsilon} = .0715$. We conclude that estimating ϵ at approximately .071 is reasonable, since the estimate did not change substantially with the final repartitioning.

We see that our simulation-based estimates of $\widehat{\Lambda}$, $\widehat{\lambda}$, and $\widehat{\epsilon}$ correspond closely to the values (8) obtained analytically by Rosenthal (1996). This suggests that our method works well in this model, and may justify applying it to a related model, the convergence properties of which have not been as precisely determined analytically.

4. Example: Variance components model.

We next consider the full variance components model. This model was used by Gelfand and Smith (1990) as an example which is difficult to analyze without the help of the Gibbs sampler. The resulting Gibbs sampler was analyzed heuristically by Gelfand, Hills, et al. (1990), who reported good success. Asymptotic convergence rates were derived by Rosenthal (1995a). In a very complicated calculation, Geyer (1994) analytically verified a drift condition of the form (1), for a simplified version of this Gibbs sampler. However, he did not provide a minorization condition or convergence rate estimate. This model has remained a standard application of the Gibbs sampler, with no clear understanding of its convergence rate or properties.

The model is defined by

$$\theta_i | \sigma_\theta^2, \sigma_e^2, \mu \;\sim\; N(\mu, \sigma_\theta^2); \qquad (1 \le i \le K)$$

$$Y_{ij} | \sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K \;\sim\; N(\theta_i, \sigma_e^2); \qquad (1 \le i \le K; \quad 1 \le j \le J)$$

where σ_θ^2 and σ_e^2 correspond respectively to A and v in the model in Section 3. Here, however, the simplifying assumption that v is known is removed, and σ_e^2 is an additional unknown parameter. Here Y_{ij} are observed data, and σ_θ^2, σ_e^2, and μ have the conjugate prior distributions

$$\sigma_\theta^2 \sim IG(a_1, b_1); \qquad \sigma_e^2 \sim IG(a_2, b_2); \qquad \mu \sim N(\mu_0, \sigma_0^2);$$

where $a_1, b_1, a_2, b_2, \mu_0$, and σ_0^2 are fixed constants.

The Gibbs sampler proceeds on the $(K+3)$-dimensional space $(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K)$, conditional on data $\{Y_{ij}\,;\, 1 \le i \le K,\ 1 \le j \le J\}$. After choosing initial values, it repeatedly updates them (for iterations $k = 1, 2, 3, \ldots$) by the conditional distributions

$$\begin{aligned}
\sigma_\theta^{2(k)} \;&\sim\; \mathcal{L}(\sigma_\theta^2 \mid \mu^{(k-1)}, \sigma_e^{2(k-1)}, \theta_1^{(k-1)}, \ldots, \theta_K^{(k-1)}, Y_{ij}) \\
&=\; IG\left(a_1 + \frac{1}{2}K,\ b_1 + \frac{1}{2}\sum_i (\theta_i^{(k-1)} - \mu^{(k-1)})^2\right); \\
\sigma_e^{2(k)} \;&\sim\; \mathcal{L}(\sigma_e^2 \mid \mu^{(k-1)}, \sigma_\theta^{2(k)}, \theta_1^{(k-1)}, \ldots, \theta_K^{(k-1)}, Y_{ij})
\end{aligned}$$

$$= IG\left(a_2 + \frac{1}{2}KJ,\ b_2 + \frac{1}{2}\sum_{i,j}(Y_{ij} - \theta_i^{(k-1)})^2\right);$$

$$\mu^{(k)} \sim \mathcal{L}(\mu \mid \sigma_\theta^{2(k)}, \sigma_e^{2(k)}, \theta_1^{(k-1)}, \ldots, \theta_K^{(k-1)}, Y_{ij})$$

$$= N\left(\frac{\sigma_\theta^{2(k)}\mu_0 + \sigma_0^2\sum_i \theta_i^{(k-1)}}{\sigma_\theta^{2(k)} + K\sigma_0^2},\ \frac{\sigma_\theta^{2(k)}\sigma_0^2}{\sigma_\theta^{2(k)} + K\sigma_0^2}\right);$$

$$\theta_i^{(k)} \sim \mathcal{L}(\theta_i \mid \mu^{(k)}, \sigma_\theta^{2(k)}, \sigma_e^{2(k)}, \theta_1^{(k)}, \ldots, \theta_{i-1}^{(k)}, \theta_{i+1}, \ldots, \theta_K, Y_{ij})$$

$$= N\left(\frac{J\sigma_\theta^{2(k)}\overline{Y}_i + \sigma_e^{2(k)}\mu^{(k)}}{J\sigma_\theta^{2(k)} + \sigma_e^{2(k)}},\ \frac{\sigma_\theta^{2(k)}\sigma_e^{2(k)}}{J\sigma_\theta^{2(k)} + \sigma_e^{2(k)}}\right) \qquad (1 \le i \le K).$$

[Here $\overline{Y}_i = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$.] We have chosen the updating order so that we may take $D = 3$ in Lemma 2.

We proceed to apply our method to two datasets illustrating the one-way variance components model, which are analyzed in Chapter 5 of Box and Tiao (1973). One, taken from Davies (1967), involves between- and within- batch variation in yield of dyestuff. The other is simulated data, to which Gelfand, Hills et al. (1990) applied the Gibbs sampler. In both cases, K = 6 and J = 5.

For both datasets we specified the following flat prior on σ_e^2, and weak but proper priors on μ and σ_θ^2:

$$\mu_0 = 0; \qquad \sigma_0^2 = 10^{12};$$

$$a_1 = 0.5; \qquad b_1 = 1.0;$$

$$a_2 = 0.0; \qquad b_2 = 0.0.$$

The proper prior on σ_θ^2 ensured parameter identifiability and prevented the Gibbs sampler from "getting stuck" due to values of σ_θ^2 too close to 0.

Choosing the V() function. For V() we needed a function that would control both μ and θ, since those are the parameters for which initial values are required. We reasoned that, since the priors were so weak, the marginal posterior distributions would be almost

entirely driven by the data. Accordingly, we chose the following V() function, which incorporates data-based estimates of σ_θ^2 and σ_e^2.

$$V(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K) \;=\; \frac{1}{K}\sum_{i=1}^{K}\left(\theta_i \;-\; \frac{Jv_1\overline{Y}_i + v_2\overline{\overline{Y}}}{Jv_1 + v_2}\right)^2 \;+\; (\mu - \overline{\overline{Y}})^2 \qquad (9)$$

where $v_1 = \frac{1}{KJ}\sum_{i,j}(Y_{ij} - \overline{Y}_i)^2$ and $v_2 = \frac{1}{K}\sum_{i}(\overline{Y}_i - \overline{\overline{Y}})^2$.

Obtaining a lower bound for Λ. Here as in the James-Stein model, there is only one configuration of starting values for which $V(x_0) = 0$. For each of the two datasets, we ran $N_0 = 10000$ single-iteration chains (i.e., $m = 1$) from this configuration and computed the mean value of $V(X^{(1)})$. For the simulated data, this was 1.60, and for the dyestuff data it was 818.

We decided that, for this problem, we would add 1.0 to the V function to prevent problems in estimating λ and to enable us to use the improved bounds mentioned in Proposition 1. Accordingly, we added 1.0 to these respective initial estimates of Λ.

Obtaining a lower bound for λ. We again chose initial values by specifying dispersion parameters for generating $\mu^{(0)}$ and the $\theta_i^{(0)}$, i = $1, \ldots, K$, from normal distributions centered at $\overline{\overline{Y}}$. For the simulated data, we chose 5 values of standard deviations ranging from 0.125 to 3.0, and generated 10 sets of initial values using each. From each of those 50 sets of initial values we ran 5000 single-iteration chains. With the V() function defined as 1.0 + the expression in (9), the largest value of $\widehat{\lambda}$ computed as in (3) was 0.54.

For the dyestuff data, from each of 20 sets of initial values generated at each of 5 dispersions ranging from 1 to 625, we ran 20000 single-iteration chains. Our largest value of $\widehat{\lambda}$ was 0.65.

Estimating ϵ *in the minorization condition.* Our estimates Λ and λ enabled us to choose $d = 15$ for the simulated dataset and $d = 5,000$ for the dyestuff dataset. By Lemma 2, we need establish overlap in the transition probabilities only for σ_e^2, σ_θ^2, and μ. We specified expressions for conservative upper and lower bounds for the three quantities required in the full conditionals for these parameters – $\sum_i(\theta_i - \mu)^2$, $\sum_{ij}(Y_{ij} - \theta_i)^2$, and $\overline{\theta}$ – subject to the constraint that $V(X^{(0)}) \leq d$. For each dataset, we then plugged the appropriate numbers into these expressions to obtain 8 sets of initial values.

For the simulated data, we constructed 3-dimensional bins with two grid sizes, corresponding to 25 intervals per dimension and 50 intervals per dimension. After some experimentation, we chose $k_0 = 10$, and ran 30,000 ten-iteration chains from each of the 8 starting points. The sum over all bins of the minimum fraction of points falling into that bin from each of the 8 samples was .749 for the coarser grid and .709 for the finer. We conclude that .70 is a reasonable choice for $\widehat{\epsilon}$ for these data.

For the dyestuff data, a similar procedure (but with $k_0 = 50$) led to $\widehat{\epsilon} = .28$. The difference between these results and those for the simulated data is due to the much larger numerical values in the dyestuff dataset. These caused the upper and lower bounds for the initial values for the samples used in estimating ϵ to be very far apart, which in turn caused the means of the resulting full conditionals to be widely separated. Hence many iterations (large k_0) were required for the chains to reach overlapping parts of the parameter space.

Bounding the convergence to stationarity. Having found estimates $\widehat{\Lambda}$, $\widehat{\lambda}$, and $\widehat{\epsilon}$, we now use these values in the bound provided by Proposition 1.

For the simulated data, we have $\widehat{\Lambda} = 2.6$, $\widehat{\lambda} = 0.54$, and $\widehat{\epsilon} = 0.70$, with $d = 15$, $m = 1$, and $k_0 = 10$. After some experimenting, we choose $r = 0.042$ and $M = 0.1$. Recalling that $V \geq 1$, and assuming that we start with $V(X^{(0)}) = 1$, we obtain from Proposition 1 that

$$\|\mathcal{L}(X^{(k)}) - \pi(\cdot)\| \leq (0.30)^{\lfloor 0.0042\, k \rfloor} + (0.586)(0.990)^k .$$

For example, if $k = 955$, this bound is equal to 0.00816.

For the dyestuff data, we have $\widehat{\Lambda} = 820$, $\widehat{\lambda} = 0.65$, and $\widehat{\epsilon} = 0.28$, with $d = 5,000$, $m = 1$, and $k_0 = 50$. We choose $r = 0.0076$ and $M = 0.0001$. Again using that $V \geq 1$ and assuming $V(X^{(0)}) = 1$, we obtain from Proposition 1 that

$$\|\mathcal{L}(X^{(k)}) - \pi(\cdot)\| \leq (0.72)^{\lfloor 0.000152\, k \rfloor} + (0.7905)(0.99985)^k .$$

For example, if $k = 98,750$, this bound is equal to 0.0072. (This value of k is, of course, overly conservative. Tighter upper and lower bounds for the starting values used in estimating ϵ most likely would help.)

5. Example: Ordinal probit model.

Our final example comes from an ordinal probit model, common in biostatistical and econometric applications (Carlin and Polson, 1992; Albert and Chib, 1993; Cowles, 1996).

We take the simplest case of K observations of a response variable $w_i, i = 1, \ldots, K$ that can take on three ordered values arbitrarily labeled -1, 0, and 1 corresponding to, say, "worse," "no change", and "improved." In addition a single continuous covariate x_i is observed for each subject. Following the formulation of Albert and Chib (1993) we introduce latent continuous variables y_i^* underlying the ordinal w_i's and a set of cutpoints $-\infty$, 0, γ, and ∞ that divide the real line into three intervals such that $w_i = -1$ corresponds to $y_i^* < 0$, $w_i = 0$ corresponds to $0 < y_i^* < \gamma$, and $w_i = 1$ corresponds to $\gamma < y_i^*$. We assume that the $y*_i$'s are distributed $N(\beta_0 + \beta_1 x_i, 1)$ for some β_0 and β_1.

The unknown parameters in this model are thus the intercept β_0, the coefficient β_1 of the covariate, and the cutpoint γ. For $\beta = (\beta_0, \beta_1)^T$ we specify a flat prior, and for γ we specify a flat prior constrained to the set $\{\gamma > 0.05\}$.

The Gibbs sampler for this model runs on the $(K+3)$-dimensional space $(\beta_0, \beta_1, \gamma, y_1^*, \ldots, y_K^*)$. For notational ease, we set

$$M_0 = \max\{y_i^* \,;\, w_i = 0\}\,; \quad m_1 = \min\{y_i^* \,;\, w_i = 1\}\,; \quad \mathbf{B} = (B_0, B_1)^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\,,$$

where $\mathbf{X}$ is a $K \times 2$ matrix consisting of a column of 1's and a column of the x_i's. The Gibbs sampler has (conditional) updating distributions

$$\mathcal{L}(\beta \mid \mathbf{w}, \mathbf{Y}) \;=\; N\left(\mathbf{B},\; (\mathbf{X}^T\mathbf{X})^{-1}\right)\,;$$

$$\mathcal{L}(\gamma \mid \mathbf{w}, \mathbf{Y}, \beta) \;=\; \mathcal{U}\left[\max(M_0, 0.05),\; m_1\right]\,;$$

$$\mathcal{L}(y_i^* \mid \beta, \gamma, \mathbf{w}, y_{-i}^*) \;=\; TN_{w_i,\gamma}(\beta_0 + \beta_1 x_i,\, 1), \qquad (1 \le i \le K)\,.$$

[Here $TN_{w_i,\gamma}$ is a normal distribution truncated to the appropriate interval, i.e. to $(-\infty, 0)$ if $w_i = -1$; to $(0, \gamma)$ if $w_i = 0$; or to (γ, ∞) if $w_i = 1$.] Again, we have chosen the updating order intentionally so that we may take $D = 3$ in Lemma 2.

For the data $\{x_i, w_i\}$, we take the first 50 observations from the simulated data set reported in Cowles (1996); thus, $K = 50$. We now proceed to apply our method. The

procedure is similar to the previous examples, and will be described in somewhat less detail.

Choosing the $V()$ function. We need our function to "control" the values of B_0, B_1, M_0, and m_1. Accordingly, we set

$$V(\beta_0, \beta_1, \gamma, y_1^*, \ldots, y_K^*) = (B_0 - 0.4302)^2 + (B_1 - 2.3361)^2 + (M_0 - 1.8175)^2 + (m_1 - 1.9347)^2 ,$$

where the numerical values were chosen empirically, based on maximum likelihood estimates obtained from the SAS module *proc logistic* (SAS Institute, 1990).

Obtaining a lower bound for Λ. We chose $m = 3$, and generated $N_0 = 5000$ different 3-iteration chains, from the unique (aside from the unimportant value β) starting point which has $V() = 0$. This led to the (rather small) estimate $\widehat{\Lambda} = 0.2$. As in the previous example, we then added 1.0 to the function $V()$, and hence also to $\widehat{\Lambda}$.

Obtaining an estimate for λ. We generated 5 sets of initial values from each of 6 normal distributions with different standard deviations. From each such initial value, we ran $N_2 = 1000$ different 3-iteration chains. Using (3) gave an estimate of $\widehat{\lambda} = 0.70$.

Obtaining an estimate for ϵ. Our values $\widehat{\Lambda}$ and $\widehat{\lambda}$ require that $d > \frac{2\widehat{\Lambda}}{1-\widehat{\lambda}} - 1 = 7.0$, so we chose $d = 10$. To find starting distributions with the least overlap subject to the constraint $V() \leq d$, we considered the following 9 different choices for the starting quadruple (B_0, B_1, M_0, m_1):

$$(-0.6698, -3.4361, 0.05, 0.06); \quad (-1.1198, -2.3361, 0.05, 0.06); \quad (1.9802, -2.3361, 0.05, 0.06);$$

$$(1.5302, -1.2361, 0.05, 0.06); \quad (-1.6930, -4.4593, 1.8175, 1.9347);$$

$$(-1.6930, -0.2129, 1.8175, 1.9347); \quad (2.5534, -4.4593, 1.8175, 1.9347);$$

$$(2.5534, -0.2129, 1.8175, 1.9347); \quad (0.4302, -2.3361, 4.00, 4.01) .$$

We created little bins by dividing up the state space into 40 intervals for each of the 3 different parameters. Taking $k_0 = 10$, and running $40,000$ different chains from each of these 9 starting points, with $mk_0 = 30$ iterations each, and computing $\widehat{\epsilon}$ as before, we obtained the estimate 0.182, which we rounded down to $\widehat{\epsilon} = 0.18$.

Bounding the convergence to stationarity. We use the values $\widehat{\Lambda} = 1.2$, $\widehat{\lambda} = 0.70$, and $\widehat{\epsilon} = 0.18$, with $d = 10$, $m = 3$, and $k_0 = 10$. We choose $r = 0.035$ and $M = 0.05$. Recalling again that $V \geq 1$, and assuming again that we start with $V(X^{(0)}) = 1$, we obtain from Proposition 1 that

$$\|\mathcal{L}(X^{(k)}) - \pi(\cdot)\| \leq (0.82)^{\lfloor 0.00117\, k \rfloor} + (0.7759)\,(0.9987)^{\lfloor k/3 \rfloor}.$$

For $k = 21,000$, this bound is equal to 0.00863.

6. Discussion and conclusion.

We have presented a method for bounding the burn-in time for complicated Markov chains to converge to their stationary distribution. We consider our method to be a middle ground between ad-hoc convergence diagnostics (which offer little in the way of guarantees), and rigorous theoretical analysis (which is often difficult to apply). We make use of a theoretical result (Proposition 1) for bounding the distance to stationarity, but we supplement this by a method for estimating the drift and minorization conditions which Proposition 1 requires. In this sense, our work is similar in spirit to related works by Garren and Smith (1995) and by Geyer (1992). Furthermore, as in the approach of Garren and Smith (1995), our numerical estimates are all taken from preliminary, auxiliary simulations, so they do not in any way bias the results of the final MCMC run.

We have applied our method to several realistic examples of the Gibbs sampler. In each case we have obtained upper bounds on the time required to approximately converge to stationarity. In some cases these bounds were probably overly conservative, but in all cases they required less than $100,000$ iterations and thus were feasible to implement. This is to be compared with some theoretical results (which may require billions of iterations to be of use), and with convergence diagnostics (which may be overly optimistic and suggest too few iterations to properly achieve burn-in).

Our method has advantages and disadvantages when compared with either non-simulation-based theoretical bounds or convergence diagnostics. We avoid the traditional limitations of theoretical bounds on convergence, both by making the computations feasible and by allowing for the superior results which may be obtained for $mk_0 > 1$ (which is

nearly impossible to compute analytically for large examples). In addition, our method is almost certainly less easily fooled than are convergence diagnostics applied to the output of an MCMC sampler.

On the other hand, our method is not as automatic as are the convergence diagnostics of Gelman and Rubin (1992), Raftery and Lewis (1992), and others; analytic work is required to choose a useful V function and to identify the extremes of V_d. In addition, our method is computer-intensive, requiring substantial additional auxiliary simulation in addition to the actual MCMC run.

Furthermore, like all diagnostic techniques, our method does not come with guarantees. Specifically, there are no guarantees that we have chosen enough starting values, or small enough bins, in estimating λ and ϵ. This question must always be handled with care. However, our current investigations, including the comparison to analytic work (Section 3) and the varying of the various parameters involved, suggest that the method is fairly stable and is working well.

Similarly, like all other analytical approaches but unlike some diagnostics, our method becomes more prohibitive in high dimensions. This "curse of dimensionality" is somewhat unavoidable. However, we do have the advantage that, for the sequentially-updated Gibbs sampler, our little bins need only cover D of the n dimensions, and this can often result in great savings. Furthermore, the ease of implementation, and allowance for $mk_0 > 1$, suggests that the method will work well for "moderate" dimensional models, and will also allow for exploratory work in moderate dimensions which could offer insight into higher-dimensional situations.

Unlike the convergence diagnostics of Gelman and Rubin (1992), Raftery and Lewis (1992) and Geweke (1992), and the work of Geyer (1992), our method addresses only burn-in and not the issues of whether the chain has traversed the entire sample space and whether the variances of estimates are reasonable.

Although these disadvantages make our method currently too unwieldy to be used by applied statisticians for every data analysis involving MCMC techniques, we believe that our approach can be of practical value. At minimum, it could be used by mathematical statisticians to screen out candidate V functions that are not worth pursuing analytically.

At best, it might be extendible to compute approximate convergence bounds for whole classes of models. It is even possible that it could suggest approximate formulas into which applied users could simply plug constants such as number of observations, number of variance components, number of categories, values of hyperparameters, etc. in order to determine the number of burn-in iterations required for specific analyses.

[We would recommend that, after running the sample for the number of iterations so computed, the applied user turn to methods such as those of Geweke (1992) and Geyer (1992) to determine the number of subsequent iterations required to obtain the desired precision of estimation of quantities of interest from the dependent samples.]

We are particularly encouraged by our results for the variance components model, the simplest random effects model. We look forward to extending our method to the more complex random effects models that form the basis of many biostatistical analyses for which MCMC methods are used. We also plan to use our method to more fully compare different models, different data sets, and different choices of V functions.

An additional area for future research is developing an adaptive method for constructing the bins used in verifying the minorization condition that will enable automated assessment of the accuracy of the estimate $\widehat{\epsilon}$.

In conclusion, we are cautiously optimistic that our simulation method for computing convergence bounds can make a useful contribution toward bridging the gap between theoretical analysis and applied MCMC usage.

REFERENCES

Albert, J.H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, **88**, 669–679.

Baxendale, P.H. (1994), "Uniform Estimates for Geometric Ergodicity of Recurrent Markov Chains." Tech. Rep., Dept. of Mathematics, University of Southern California.

Box, G.E.P. and Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis*, New York: Wiley.

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

et al., eds.), 577–586.

Cowles, M.K. (1996), "Accelerating Monte Carlo Markov Chain Convergence for Cumulative-Link Generalized Linear Models." *Statistics and Computing* **6**, 101–111.

Cowles, M.K. and Carlin, B.P. (1996), "Markov Chain Monte Carlo Convergence Diagnostics: a Comparative Review." *Journal of the American Statistical Association*, to appear.

Davies, O.L. (1967), *Statistical Methods in Research and Production*, 3rd ed., London: Oliver and Boyd.

Frieze, A., Kannan, R. and Polson, N.G. (1993), "Sampling from Log-concave Distributions." *Annals of Applied Probability*, **4**, 812–837.

Frigessi, A., Hwang, C.-R., Sheu, S.J. and Di Stefano, P. (1993), "Convergence Rates of the Gibbs Sampler, the Metropolis Algorithm, and Other Single-Site Updating Dynamics." *Journal of the Royal Statistical Society, Series B*, **55**, 205–220.

Garren, S.T. and Smith, R.L. (1995), "Estimating the Second Largest Eigenvalue of a Markov Transition Matrix." Research Report #95-18, Statistical Laboratory, University of Cambridge.

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

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

Gelman, A. and Rubin, D.B. (1992), "Inference from Iterative Simulation using Multiple Sequences." *Statistical Science*, Vol. **7**, No. **4**, 457-472.

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

Geweke, J. (1992), "Evaluating the Accuracy of Sampling Based Approaches to the Calculation of Posterior Moments." In *Bayesian Statistics 4* (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith), pp. 169-193. Oxford University Press.

Geyer, C.J. (1992), "Practical Markov Chain Monte Carlo." *Statistical Science*, Vol.

7, No. **4**, 473-483.

Geyer, C.J. (1994), "Geometric Ergodicity of the Block Gibbs Sampler for a Simple Hierarchical Model." Unpublished manuscript.

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

Ingrassia, S. (1994), "On the Rate of Convergence of the Metropolis Algorithm and Gibbs Sampler by Geometric Bounds." *Annals of Applied Probability*, **4**, 347–389.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087-1091.

Meyn, S.P. and Tweedie, R.L. (1993), *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

Meyn, S.P. and Tweedie, R.L. (1994), "Computable Bounds for Convergence Rates of Markov Chains." *Annals of Appled Probability*, **4**, 981-1011.

Morris, C. (1983), "Parametric Empirical Bayes Confidence Intervals." *Scientific Inference, Data Analysis, and Robustness*, 25-50.

Raftery, A.E. and Lewis, S. (1992), "How Many Iterations in the Gibbs Sampler?" In *Bayesian Statistics 4* (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith), pp. 763-773. Oxford University Press.

Roberts, G.O. (1992), "Convergence Diagnostics of the Gibbs Sampler." In Bayesian Statistics **4** (J.M. Bernardo et al., eds.), 777-784. Oxford University Press.

Roberts, G.O. and Tweedie, R.L. (1994), "Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms." *Biometrika*, to appear.

Rosenthal, J.S. (1995a), "Rates of Convergence for Gibbs Sampling for Variance Components Models." *Annals of Statistics* **23** (1995), 740–761.

Rosenthal, J.S. (1995b), "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo." *Journal of the American Statistical Association,* **90**, 558–566. Correction, p. 1136.

Rosenthal, J.S. (1996), "Analysis of the Gibbs Sampler for a Model Related to James-

Stein Estimators." *Statistics and Computing* **6** (1996), 269-275.

SAS Institute (1990), The Logistic Procedure. SAS/STAT User's Guide, Vol 2, Version 6, 4th edition, pp. 1071-1126. SAS Institute, Cary, NC.

Smith, A.F.M. and Roberts, G.O. (1993), "Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods" (with discussion). *Journal of the Royal Statistical Society*, Series B **55**, 3-24.

Spivak, M. (1980), Calculus, 2nd ed. Publish or Perish, Inc. Wilmington, Delaware.