The Annals of Statistics
1995, Vol. 23, No. 3, 740–761

RATES OF CONVERGENCE FOR GIBBS SAMPLING FOR VARIANCE COMPONENT MODELS

BY JEFFREY S. ROSENTHAL

University of Minnesota

This paper analyzes the Gibbs sampler applied to a standard variance component model, and considers the question of how many iterations are required for convergence. It is proved that for K location parameters, with J observations each, the number of iterations required for convergence (for large K and J) is a constant times $(1 + \log K/\log J)$. This is one of the first rigorous, a priori results about time to convergence for the Gibbs sampler. A quantitative version of the theory of Harris recurrence (for Markov chains) is developed and applied.

1. Introduction. In the past several years there has been a lot of attention given to the Gibbs sampler algorithm for sampling from posterior distributions. This Markov chain Monte Carlo algorithm, popularized by Geman and Geman [11] and summarized in [8], has its roots in the Metropolis–Hastings algorithm ([17], [14]). It is closely related to the data augmentation algorithm of Tanner and Wong [24]. It exploits the simplicity of certain conditional distributions to define a Markov chain that converges in law to the posterior distribution under consideration. Once the Markov chain has converged, the values of the Markov chain provide samples from the posterior. This facilitates sampling from the posterior, even though it may be very difficult to compute directly. Thus, certain computational problems normally associated with Bayesian inference can be overcome. Gibbs sampling has recently been applied in non-Bayesian contexts as well; see [13].

One obvious question is how long the Markov chain must be run "until it converges." In most actual implementations of Gibbs sampling, this question is answered heuristically, as in "Let's run it 1000 times" (see, e.g., [5], page 6). This may be risky since Gibbs sampling sometimes converges very slowly; see for example [16]. Now, it may be possible to use convergence diagnostics to check if the distribution after (say) 1000 steps is indeed close to the distribution to which the chain appears to converge; see [12] and [20]. On the other hand, see [10] for warnings about possible problems. In any case, it would be comforting to have theoretical results regarding how many iterations are required before the chain has in fact converged.

There has been limited analysis of this question to date (though it can be expected that there wil be more in the future). In [23] and [15], general theorems about the functional form of the convergence are obtained, and it is

Received March 1992; revised October 1993.
AMS 1991 *subject classifications*. Primary 62M05; secondary 60J05.
Key words and phrases. Gibbs sampler, Markov chain Monte Carlo, rate of convergence, variance component model, Harris recurrence.

shown that the convergence will often be geometric. However, no quantitative results regarding the convergence rate are given. (Perhaps this point should be stressed: It is one thing to say the variation distance to the true posterior distribution after k steps will be less than $A\alpha^k$ for some $\alpha < 1$ and $A > 0$. It is quite another to give some idea of how much less than 1 this α will be, and how large A is, or equivalently to give a quantitative estimate of how large k should be to make the variation distance less than some ε.) In [23] several simple models are analyzed exactly, facilitating convergence results for these cases. In [22], quantitative convergence rates are obtained for data augmentation for a two-step hierarchical model involving Bernoulli random variables. Also, see [1] for an interesting analysis of a related "discretization" algorithm.

In this paper we analyze the convergence rate of the variance component models as described in [8], Section 3.4, and defined herein in Section 3. (See also [6] and [9].) Briefly, this model involves an overall location parameter μ, and K different parameters $\theta_1, \ldots, \theta_K$ which are normally distributed around μ. For each θ_i there are J different observations $Y_{i1}, \ldots, Y_{iJ}$, normally distributed around θ_i. The point of view is that μ, the θ_i and the two variances involved are all unknown and are to be estimated. We focus our attention on the case when K and J are both fairly large.

This is a model in which Gibbs sampling may be very useful. Thus, it would appear to be particularly important to know how long to run the Gibbs sampler Markov chain until it converges to the desired posterior distribution. Specifically, one can ask how many iterations must be run until the variation distance between the law of the Markov chain and the true posterior is appropriately small.

This paper provides the following answer (Theorem 1). If we consider a model with K different location parameters and with J observed data for each parameter, then the variation distance in question after k iterations of the Gibbs sampler is less than

$$1.1 \exp\left(-Bk \Big/ \left(1 + \frac{\log K}{\log J} \right) \right)$$

plus a small correction term, provided K and J are not too small. Here B is a positive number independent of J, K and k, although it depends on the priors and on the nature of the data being studied. To the extent that one is willing to ignore the correction term, the result is in some sense sharp up to constants; see the remarks following the statement of the theorem.

Theorem 1 therefore shows that the Gibbs sampler will converge relatively quickly for the variance component models case, even for fairly large J and K. This is an encouraging result, and runs contrary to the warning in [8], page 401, that Gibbs sampling tends to converge fairly slowly with many parameters. A partial explanation is that in variance component models as we shall study them, the K location parameters all act as a unit. In particular, conditional on the values of the other three parameters, the K

location parameters are all conditionally independent. (We note that it may have been intended in [8] that these K different parameters were to be thought of as a single vector parameter, so that their warning is not contradicted.) Thus, while the results herein are encouraging, one should not expect similar results for models with more complicated interdependencies, such as those arising in image processing.

The proof of Theorem 1 employs a coupling argument (Lemma 2) related to the notion of Harris recurrence (see [2]–[4], [18]). This lemma is quite general, and reduces the study of convergence rates for Markov chains to the question of how much "overlap" there is between the multistep transition probabilities starting from different points. The lemma produces upper bounds on the variation distance between a Markov chain after k steps and its stationary distribution.

This paper is organized as follows. In Section 2 we review the Gibbs sampler algorithm. In Section 3 we define the variance component models we shall study, and discuss how Gibbs sampling is applied to them. We also state our main theorem (Theorem 1). In Section 4 we state and prove Lemma 2, the key lemma in the proof of Theorem 1. Finally, in Section 5 we use Lemma 3 (a specialization of Lemma 2), together with some careful computation, to prove Theorem 1. We close with an Appendix that discusses variation distance and coupling.

2. The Gibbs sampler. The Gibbs sampler algorithm was popularized by Geman and Geman [11]. It is related to the data augmentation algorithm of Tanner and Wong [24]. A good review of these and other related algorithms may be found in [8].

Suppose we have random variables $U_1, U_2, \ldots, U_n$ and we wish to sample from their joint distribution $\mathscr{L}(U_1, \ldots, U_n)$. Suppose this joint distribution is complicated and therefore difficult to sample from. The Gibbs sampler algorithm proceeds as follows.

We first guess initial values $U_1^{(0)}, \ldots, U_n^{(0)}$ for the random variables. (These may themselves be chosen from some initial distribution.) We then update U_1 conditional on the initial values of the other variables:

$$U_1^{(1)} \sim \mathscr{L}\big(U_1 | U_2 = U_2^{(0)}, U_3 = U_3^{(0)}, \ldots, U_n = U_n^{(0)}\big).$$

We continue in this way, updating each of the random variables conditional on the most recent values of the others:

$$\begin{aligned} U_2^{(1)} &\sim \mathscr{L}\big(U_2 | U_1 = U_1^{(1)}, U_3 = U_3^{(0)}, \ldots, U_n = U_n^{(0)}\big) \\ &\vdots \\ U_n^{(1)} &\sim \mathscr{L}\big(U_n | U_1 = U_1^{(1)}, U_2 = U_2^{(1)}, \ldots, U_{n-1} = U_{n-1}^{(1)}\big). \end{aligned}$$

This completes one iteration of the Gibbs sampler; it requires n different updates. Continuing in the same manner, we again update the variables $U_1, \ldots, U_n$ in order, conditional on their most recent values, to obtain $U_1^{(2)}, \ldots, U_n^{(2)}$, and so on. After k iterations we have generated the random

variables $U_1^{(k)}, \ldots, U_n^{(k)}$. (In some implementations, the order in which $U_1, \ldots, U_n$ are updated is not fixed; see [11] and [8], but we do not consider that here.)

This algorithm may be thought of as a Markov chain with $x_k = (U_1^{(k)}, \ldots, U_n^{(k)})$. It is easily seen that the probability distribution $\pi(\cdot) = \mathscr{L}(U_1, \ldots, U_n)$ is invariant for the chain. Geman and Geman [11] and Schervish and Carlin [23] have proved that under certain (usually satisfied) positivity conditions (roughly, that conditional on the values of U_j for $j \neq i$, anything is possible for U_i), this Markov chain will converge to its invariant distribution $\pi(\cdot)$ as $k \to \infty$. Thus, for large values of k, $\mathscr{L}(U_1^{(k)}, \ldots, U_n^{(k)})$ will be close to $\pi(\cdot) = \mathscr{L}(U_1, \ldots, U_n)$, so that $(U_1^{(k)}, \ldots, U_n^{(k)})$ can be thought of as a sample from $\pi(\cdot)$.

An obvious and practical question to ask is how large must k be so that $\mathscr{L}(U_1^{(k)}, \ldots, U_n^{(k)})$ is indeed close to $\pi(\cdot)$? There have been very few theoretical results about this question. In this paper, we will get quantitative bounds on the variation distance

$$\left\| \mathscr{L}\left(U_1^{(k)}, \ldots, U_n^{(k)}\right) - \pi(\cdot) \right\|_{\text{var}}$$

between these two distributions, in the particular case of variance component models (Theorem 1). This will allow us to say how large k must be to make the variation distance less than a given $\varepsilon > 0$.

3. Variance component models and main result. We will follow the definition of variance component models given in [8]. (See also [6] and [9].) We suppose that there is some overall parameter μ, and that the location parameters $\theta_1, \ldots, \theta_K$ are independently normally distributed around μ:

$$\theta_i \sim N\left(\mu, \sigma_\theta^2\right), \qquad 1 \leq i \leq K.$$

We further suppose that for each θ_i, there are J data points Y_{ij} independently normally distributed around θ_i:

$$Y_{ij} \sim N\left(\theta_i, \sigma_e^2\right), \qquad 1 \leq i \leq K, 1 \leq j \leq J.$$

We suppose that the Y_{ij} are observed, but that σ_θ^2, σ_e^2, μ and $\theta_1, \ldots, \theta_K$ are to be estimated.

For example, $\theta_1, \ldots, \theta_K$ might be K different extractions from a lake. For each extraction i, $Y_{i1}, \ldots, Y_{iJ}$ might be J different measurements of the concentration of a certain pollutant. We wish to estimate the overall concentration of the pollutant μ, the concentration θ_i in each extaction and the variances σ_θ^2 between extractions and σ_e^2 between measurements.

The Bayesian approach to this problem involves putting priors on σ_θ^2, σ_e^2 and μ, and then computing (or estimating) the posterior distribution on $(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K)$. For simplicity we stick to conjugate priors, so we set

$$\sigma_\theta^2 \sim \text{IG}(a_1, b_1), \qquad \sigma_e^2 \sim \text{IG}(a_2, b_2), \qquad \mu \sim N\left(\mu_0, \sigma_0^2\right).$$

We take a_1, b_1, a_2, b_2, μ_0 and σ_0^2 as known constants. [Here N stands for the normal distribution; also IG stands for inverse gamma, and means that

the reciprocal of the random variable has a gamma distribution. IG(a, b) has density $b^a e^{-b/x}/(\Gamma(a)x^{a+1})$ for $x > 0$, and has mean $b/(a-1)$ and variance $b^2/((a-1)^2(a-2))$ if $a > 2$.] This defines a probability model for $\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K, Y_{ij}$. The desired posterior distribution is then the law of

$$\left(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K\right)$$

conditional on the observed values of the Y_{ij}.

We note that it is possible to work with more general priors. While the computations and arguments become somewhat more complicated, the basic ideas and the main results appear to be similar. However, we do not consider that further here.

This posterior distribution is difficult to compute directly for large J and K, essentially because there are no satisfactory algorithms for doing the necessary high-dimensional numberical integration. Instead we shall consider using the Gibbs sampler algorithm to sample from the posterior. For the variance component models, this works as follows. The state space is

$$\mathscr{X} = \left\{\left(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K\right) \in \mathbb{R}^{K+3} | \sigma_\theta^2, \sigma_e^2 > 0\right\}.$$

For each iteration, we update first σ_θ^2, then σ_e^2, then μ and then $\theta_1, \ldots, \theta_K$, each conditional on the most recent values of the other $K+2$ variables. The conditional distributions of these variables are easily computed [8] to be

$$\mathscr{L}\left(\sigma_\theta^2 | \mu, \sigma_e^2, \theta_1, \ldots, \theta_K, Y_{ij}\right) = \mathrm{IG}\left(a_1 + \tfrac{1}{2}K, b_1 + \tfrac{1}{2}\sum_i (\theta_i - \mu)^2\right),$$

$$\mathscr{L}\left(\sigma_e^2 | \mu, \sigma_\theta^2, \theta_1, \ldots, \theta_K, Y_{ij}\right) = \mathrm{IG}\left(a_2 + \tfrac{1}{2}KJ, b_2 + \tfrac{1}{2}\sum_{i,j} (Y_{ij} - \theta_i)^2\right),$$

$$\mathscr{L}\left(\mu | \sigma_\theta^2, \sigma_e^2, \theta_1, \ldots, \theta_K, Y_{ij}\right) = N\left(\frac{\sigma_\theta^2\mu_0 + \sigma_0^2\Sigma_i\theta_i}{\sigma_\theta^2 + K\sigma_0^2}, \frac{\sigma_\theta^2\sigma_0^2}{\sigma_\theta^2 + K\sigma_0^2}\right),$$

$$\mathscr{L}\left(\theta_i | \mu, \sigma_\theta^2, \sigma_e^2, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_K, Y_{ij}\right)$$
$$= N\left(\frac{J\sigma_\theta^2\overline{Y}_i + \sigma_e^2\mu}{J\sigma_\theta^2 + \sigma_e^2}, \frac{\sigma_\theta^2\sigma_e^2}{J\sigma_\theta^2 + \sigma_e^2}\right), \qquad 1 \le i \le K.$$

[Here $\overline{Y}_i = (1/J)\Sigma_{j=1}^J Y_{ij}$.] Thus, the Gibbs sampler algorithm involves guessing $(\sigma_\theta^{2(0)}, \sigma_e^{2(0)}, \mu^{(0)}, \theta_1^{(0)}, \ldots, \theta_K^{(0)})$ and then generating from these conditional distributions, in turn,

$$\sigma_\theta^{2(1)}, \sigma_e^{2(1)}, \mu^{(1)}, \theta_1^{(1)}, \ldots, \theta_K^{(1)}, \sigma_\theta^{2(2)}, \sigma_e^{2(2)}, \ldots, \sigma_\theta^{2(k)}, \sigma_e^{2(k)}, \mu^{(k)}, \theta_1^{(k)}, \ldots, \theta_K^{(k)}.$$

This completes k iterations of the Gibbs sampler.

Write $x^{(k)}$ for $(\sigma_\theta^{2(k)}, \sigma_e^{2(k)}, \mu^{(k)}, \theta_1^{(k)}, \ldots, \theta_K^{(k)})$ and write $\pi(\cdot)$ for the true posterior distribution $\mathscr{L}(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K)$. We are interested in the total variation distance

$$\left\|\mathscr{L}\left(x^{(k)}\right) - \pi(\cdot)\right\|_{\mathrm{var}}.$$

In particular, we wish to know how large k should be, as a function of J and K, to make this variation distance small. This question is answered by the following theorem. To state it, recall that $\bar{Y}_i = (1/J)\Sigma_j Y_{ij}$ and set

$$\bar{\bar{Y}} = \frac{1}{KJ}\sum_{i,j} Y_{ij} = \frac{1}{K}\sum_i \bar{Y}_i,$$

$$v_1 = \frac{1}{KJ}\sum_{i,j}\left(Y_{ij} - \bar{Y}_i\right)^2,$$

$$v_2 = \frac{1}{K}\sum_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2.$$

THEOREM 1. *For the Gibbs sampler algorithm defined above, with a starting distribution concentrated within the subset R_* defined below, there are positive numbers J_0, B, B', A_1, A_2 and A_3, depending only on v_1, v_2 and the priors, but otherwise independent of J, K, k and the data Y_{ij}, such that for $J \geq J_0$,*

(a) $$\left\|\mathscr{L}\left(x^{(k)}\right) - \pi(\cdot)\right\|_{\text{var}} \leq 1.1\exp\left(-Bk\Big/\left(1 + \frac{\log K}{\log J}\right)\right) + kA_1\exp\left(-A_2K\right);$$

(b) $$\left\|\mathscr{L}\left(x^{(k)}\right) - \pi(\cdot)\right\|_{\text{var}} \leq A_3\exp\left(-B'\sqrt{k}\,/\left((K/J)\log k + K^6/J^2\right)\right).$$

REMARKS.

1. The main thrust of this theorem is providing quantitative bounds on the total variation distance (see Appendix) to the stationary distribution after running the Gibbs sampler for k iterations. It thus provids estimates of how long the Gibbs sampler should be run before samples from the Gibbs sampler can be regarded as good approximations to samples from the true posterior.
2. Even for this particular variance components model, there are many factors which affect the rate of convergence, including the number of parameters K, the amount of data per parameter J, the "spread" of the data as measured by v_1 and v_2 and the prior distributions as specified by a_1, b_1, a_2, b_2, μ_0 and σ_0^2. (Note that from a Bayesian perspective the data and prior are fixed throughout, so none of these quantities is a random

variable.) Although the results of this paper in principle give rates of convergence in terms of all of these quantities, we have chosen to emphasize how the rate depends specifically on the values of K and J, with all of the other quantities held fixed. We have further concentrated particularly on the case in which K and J are both relatively large. If a different dependence is to be emphasized, the same general arguments would apply, but it might be necessary to reinterpret the results somewhat. In particular, in the case of small K and J, it may be necessary to modify the choices of R_*, etcetera, in the details of the proof.

3. To the extent that one is willing to ignore terms of the form $k \exp(-(\text{const})K)$ (which are very small if K is large and k is moderate), part (a) of the theorem gives a very pleasing answer. It states that $O(1 + \log K/\log J)$ iterations are required (provided J is not too small) to make the variation distance small. (In other words, if k is large compared to $1 + \log K/\log J$, then the variation distance is small.) In particular, if $\log K/\log J$ remains fixed, the number of iterations required does not grow with J and K. This may be somewhat surprising, since as J increases, the posterior becomes more peaked, and as K increases, the posterior becomes more complicated.
4. The quantity $O(1 + \log K/\log J)$ of the previous remark is in some sense the best possible. Specifically, it is seen (see the remark on *lower bounds* at the end of the paper) that if k is very small compared to $1 + \log K/\log J$, then the Gibbs sampler cannot possibly have converged. Thus, to the extent that one ignores the $k \exp(-(\text{const})K)$ term, the quantity $O(1 + \log K/\log J)$ is "correct," so that the result of part (a) is "sharp up to constants."
5. Despite the previous remarks, the fact remains that the second term in the bound in part (a) of the theorem, of the form $k \exp(-(\text{const})K)$, is not going to 0 as a function of k (in fact it is going to infinity). This unfortunate situation arises because of the difficulty in controlling the (rare) occurrences when the Gibbs sampler escapes from the set R_* defined below. The problem is remedied in part (b) of the theorem, which is proved by the unusual method of allowing the set R to grow as a function of k. The bound in part (b) ensures that the variation distance does indeed go to 0 as function of k. As a penalty, however, part (b) gives too slow a rate of convergence; if $K \geq J$, we need k to be large compared to $(K^6/J^2)^2$ for the variation distance to be small. We take the point of view that part (a) of Theorem 1 shows that the variation distance gets fairly small when k is of order $1 + \log K/\log J$, while part (b) shows that for even larger k the variation distance does indeed go to 0. [Furthermore, the bound in part (b) goes to 0 at a superpolynomial but subexponential rate. Thus, it does not quite establish that this Gibbs sampler is geometrically ergodic.]
6. While the theorem's aim is to provide quantitative bounds on the time to convergence of this Gibbs sampler, it is still stated in terms of the unspecified numbers B, B', A_1, A_2 and A_3. However, the proof of the

theorem (Section 5) does explain (in a necessarily complicated and multistep way) how these numbers are computed, and we have tried to indicate this as explicitly as possible. Thus, a researcher with a given data set could compute values for these numbers, and use Theorem 1 to obtain precise upper bounds for how many iterations of the Gibbs sampler will be required. A general formula for these numbers could be given, but unfortunately it would be very awkward and also not optimal, especially for small data sets. We are presently working on getting sharper values of the numbers in these cases.

7. The theorem requires that we use an appropriate starting distribution. Specifically, the starting distribution should be supported entirely in the set R_* defined in Section 5. However, for reasonably large J (which is our emphasis) this is a very large set, so this requirement is not very severe. (For small values of J, however, the set R_* could even be empty, hence our requirement that $J \geq J_0$. For such small J, the proof could be modified to produce a bound using an alternative, nonempty set R_*. Indeed, any bounded subset R_* could in principle be used, though of course the quantitative bounds would be affected.) We note that our theorem applies for *any* starting distribution supported in R_*, including a point mass.

The main lemma used to prove Theorem 1 is stated in Section 4, and Theorem 1 is then proved in Section 5.

4. The main lemma. It is difficult to approach the proof of Theorem 1 directly. This is because both the law of $x^{(k)}$ (the Gibbs sampler after k iterations) and the true posterior distribution are difficult to compute, and so the variation distance between them is also difficult to compute.

Our approach instead will be to use the following lemma. It gives a bound on the variation distance of a Markov chain to its stationary distribution in terms of the amount of "overlap" of the transition probabilities starting from different places. The lemma is closely related to the notion of Harris recurrence; see [2]–[4] and [18]. A special case of this lemma was described in [22]. We wish to emphasize that the lemma is valid for any Markov chain, and may be useful in situations quite different from Gibbs sampling.

We need the following notation. If $Q_1(\cdot)$ and $Q_2(\cdot)$ are probability measures and $\varepsilon > 0$, then we will write $Q_1(\cdot) \geq \varepsilon Q_2(\cdot)$ to mean that $Q_1(A) \geq \varepsilon Q_2(A)$ for all measurable sets A. If $Q_1(\cdot)$ and $Q_2(\cdot)$ have densities $q_1(x)$ and $q_2(x)$ with respect to Lebesgue measure, then this is equivalent to saying that $q_1(x) \geq \varepsilon q_2(x)$ for almost all x.

LEMMA 2. *Let $P(x, \cdot)$ be the transition probabilities for a time-homogeneous Markov chain on a state space $\mathcal{X}$. Suppose that for some measurable subset $R \subseteq \mathcal{X}$, some probability distribution $Q(\cdot)$ on $\mathcal{X}$, some positive integer*

k_0 and some $\varepsilon > 0$,

$$P^{k_0}(x, \cdot) \geq \varepsilon Q(\cdot) \quad \text{for all } x \in R,$$

where P_0^k represents the k_0-step transition probabilities. Let $\pi(\cdot)$ be any stationary distribution for the Markov chain on $\mathscr{X}$. Then for any initial distribution $\pi_0(\cdot)$ supported entirely in R, the distribution $\pi_k(\cdot)$ of the Markov chain after k steps satisfies

$$\|\pi_k(\cdot) - \pi(\cdot)\|_{\text{var}} \leq (1 - \varepsilon)^{\lfloor k/k_0 \rfloor} + a + 2\lfloor k/k_0 \rfloor b,$$

where $\|\cdot\|_{\text{var}}$ is total variation distance, $\lfloor r \rfloor$ is the greatest integer not exceeding r, $a = \pi(R^C) = 1 - \pi(R)$ and

$$b = \sup_{x \in R} P^{k_0}(x, R^C) = 1 - \inf_{x \in R} P^{k_0}(x, R).$$

PROOF. The proof shall be by a coupling argument. (For background on coupling, see the Appendix.) We first note that, replacing $P(x, \cdot)$ by $P^{k_0}(x, \cdot)$ if necessary and using the fact that the variation distance to a stationary distribution is (weakly) monotonically decreasing, it suffices to consider the case $k_0 = 1$.

We let $\{X_t\}$ begin in the distribution π_0 and let $\{Y_t\}$ begin in the distribution π. We let them progress as follows. Given the values X_m and Y_m, where $0 \leq m \leq k$, we choose X_{m+1} and Y_{m+1} by:

(i) If $X_m \in R$ and $Y_m \in R$, then flip a coin with probability of heads equal to ε. Then:
  (a) If it is heads, choose $x \in \mathscr{X}$ according to $Q(\cdot)$ and set $X_{m+1} = Y_{m+1} = x$.
  (b) If it is tails, choose X_{m+1} and Y_{m+1} independently according to the distributions $(1/(1 - \varepsilon))(P(X_m, \cdot) - \varepsilon Q(\cdot))$ and $(1/(1 - \varepsilon))(P(Y_m, \cdot) - \varepsilon Q(\cdot))$, respectively.

(ii) If $X_m \notin R$ or $Y_m \notin R$, then simply choose X_{m+1} and Y_{m+1} independently according to the distributions $P(X_m, \cdot)$ and $P(Y_m, \cdot)$, respectively.

This defines a prescription for choosing X_m and Y_m for $0 \leq m \leq k$. It is easily checked that these values are chosen with probabilities consistent with the transition probabilities $P(x, \cdot)$.

Let T be the first time we choose option (i)(a) above [with $T = \infty$ if we never choose option (i)(a)]. Let Z_m be equal to Y_m for $m \leq T$ and to X_m for $m > T$. Then (X_m, Z_m) is easily seen to be a coupling with coupling time T.

The coupling inequality (see Appendix) then gives that

$$\|\mathscr{L}(X_k) - \mathscr{L}(Y_k)\|_{\text{var}} = \|\pi_k(\cdot) - \pi(\cdot)\|_{\text{var}} \leq \text{Prob}(T > k).$$

We now observe that conditional on X_m and Y_m remaining in R, the coupling time T will be a geometric random variable with parameter ε. Thus

$$\begin{aligned}
&\text{Prob}(T > k) \\
&\quad \le \text{Prob}(X_m \notin R \text{ or } Y_m \notin R \text{ for some } 0 \le m \le k) \\
&\qquad + \text{Prob}(X_m \in R \text{ and } Y_m \in R \text{ for all } 0 \le m \le k, \text{ and } T > k) \\
&\quad \le \text{Prob}(X_0 \notin R) + \text{Prob}(Y_0 \notin R) \\
&\qquad + \sum_{m=1}^{k} \left(\text{Prob}(X_m \notin R | X_{m-1} \in R) + \text{Prob}(Y_m \notin R | Y_{m-1} \in R)\right) \\
&\qquad + \text{Prob}(X_m \in R \text{ and } Y_m \in R \text{ for all } 0 \le m \le k, \text{ and } T > k) \\
&\quad \le 0 + a + \left(\sum_{m=1}^{k} (2b)\right) + (1 - \varepsilon)^{\lfloor k/k_0 \rfloor} \\
&\quad = a + 2kb + (1 - \varepsilon)^{k}.
\end{aligned}$$

□

REMARKS.

1. The conclusion of the lemma is unsatisfying in that the upper bound given does not approach 0 as $k \to \infty$. One can remedy this by letting the set R get larger and larger (so that a and b get smaller and smaller) as a function of k. This idea is used in the proof of Theorem 1(b).
2. It is easily seen that we can bound the quantity b above by $k_0 b_1$, where $b_1 = \sup_{x \in R} P(x, R^C)$ is the one-step analog of the k_0-step b. This shows that

$$\|\pi_k(\cdot) - \pi(\cdot)\|_{\text{var}} \le (1 - \varepsilon)^{\lfloor k/k_0 \rfloor} + a + 2kb_1,$$

which may be easier to apply in some cases.
3. It is not necessary that the Markov chain under consideration be time-homogeneous; it is easily seen that the proof still goes through, even with the simplification of Remark 2, as long as $P^{t,t+k_0}(x, \cdot) > \varepsilon Q(\cdot)$ for all $x \in R$ and for all times t, provided we redefine b and b_1 as

$$b = \sup_t \sup_{x \in R} P^{t,t+k_0}(x, R^C), \qquad b_1 = \sup_t \sup_{x \in R} P^{t,t+1}(x, R^C).$$

4. Lemma 2 is similar in appearance to the strong stopping times of Aldous and Diaconis (see [7], Chapter 4A). However, in Lemma 2 the probability measure $Q(\cdot)$ is arbitrary, while in the case of strong stopping times, $Q(\cdot)$ is required to be a stationary distribution for the chain. This difference is significant since in many cases the stationary distribution is unknown or difficult to work with. Also, the conclusion is slightly weaker: With strong stopping times one can bound the *separation* distance to stationarity, while with Lemma 2 it is easy to construct counterexamples to show that only the *variation* distance is so bounded.

We shall actually require Lemma 2 in a slightly more specialized form. For clarity we record it here.

LEMMA 3. *Let $\mathscr{X}$, $P(\cdot,\cdot)$ and let $\pi(\cdot)$ be as in Lemma 2. Suppose there are measurable subsets $R_1, R_2 \subseteq \mathscr{X}$, some probability distribution $Q(\cdot)$ on $\mathscr{X}$, some positive integer k_0 and some $\varepsilon_1, \varepsilon_2 > 0$, such that*

$$P^{k_0}(x, R_2) \geq \varepsilon_1 \quad \text{for all } x \in R_1$$

and

$$P(x, \cdot) \geq \varepsilon_2 Q(\cdot) \quad \text{for all } x \in R_2.$$

Then for any initial distribution $\pi_0(\cdot)$ supported entirely in R_1, the distribution $\pi_k(\cdot)$ of the Markov chain after k steps satisfies

$$\|\pi_k(\cdot) - \pi(\cdot)\|_{\text{var}} \leq (1 - \varepsilon_1\varepsilon_2)^{\lfloor k/(k_0+1)\rfloor} + a + 2\lfloor k/(k_0+1)\rfloor b,$$

where $a = \pi(R_1^C) = 1 - \pi(R_1)$ and

$$b = \sup_{x \in R_1} P^{k_0+1}(x, R_1^C) + 1 - \inf_{x \in R_1} P^{k_0+1}(x, R_1).$$

PROOF. This follows immediately from Lemma 2, since the hypotheses imply that

$$P^{k_0+1}(x, \cdot) \geq \varepsilon_1\varepsilon_2 Q(\cdot) \quad \text{for all } x \in R_1. \qquad \square$$

REMARKS.

1. It is useful to think of the set R_2 above as being very small, so that the transition probabilities from R_2 are all pretty much the same (so ε_2 is reasonably large). Thus, it is most difficult to show that the Markov chain will jump from R_1 to R_2 after k_0 steps with probability $\geq \varepsilon_1$.
2. As in Remark 2 following Lemma 2 above, we can simplify Lemma 3 to state that

$$\|\pi_k(\cdot) - \pi(\cdot)\|_{\text{var}} \leq (1 - \varepsilon_1\varepsilon_2)^{\lfloor k/(k_0+1)\rfloor} + a + 2kb_1,$$

where $b_1 = \sup_{x \in R} P(x, R^C)$. We shall use this in Section 5 in the proof of Theorem 1(a).

5. The proof of Theorem 1. In this section we prove Theorem 1, making use of Lemma 3. Our plan will be as follows. We shall choose an appropriate "small" set R_2 and "large" set R_1 such that beginning in the set R_1, with large probability (i.e., with probability bounded below independently of J and K), the Markov chain will get to the set R_2 after some k_0 steps, *and* such that the transition probabilities from R_2 have large overlap (i.e., overlap bounded below independently of J and K). Here k_0 will be $O(1 + \log K/\log J)$. We will then use Lemma 3 with $\varepsilon_1, \varepsilon_2$ chosen independent of J and K to conclude that the Markov chain converges in $O(k_0/\varepsilon_1\varepsilon_2) = O(1 + \log K/\log J)$ steps.

We now proceed to make this more precise. Recall the definitions of $\bar{\bar{Y}}$, v_1 and v_2 from Section 3. We let R_2 be the subset of $\mathscr{X}$, where μ and $\theta_1, \ldots, \theta_K$ satisfy

$$\left|\left(\frac{1}{K}\sum_i (\theta_i - \mu)^2\right) - \left(v_2 - \left(\frac{1}{J-1}\right)v_1\right)\right| \leq \frac{1}{\sqrt{K}},$$

$$\left|\left(\frac{1}{JK}\sum_{ij} (\theta_i - Y_{ij})^2\right) - \left(\left(\frac{J}{J-1}\right)v_1\right)\right| \leq \frac{1}{\sqrt{JK}}$$

and

$$\left|\left(\frac{1}{K}\sum_i \theta_i\right) - \bar{\bar{Y}}\right| \leq \frac{1}{\sqrt{K}}.$$

Note that R_2 will be nonempty for sufficiently large values of J. The following lemma states that the transition probabilities from R_2 have large overlap.

LEMMA 4. *There is a probability measure $Q(\cdot)$ on $\mathscr{X}$ and an $\varepsilon_2 > 0$ independent of J and K (although it may depend on v_1, v_2 and the priors) such that*

$$P(x, \cdot) \geq \varepsilon_2 Q(\cdot) \quad \text{for all } x \in R_2.$$

PROOF. We define $Q(\cdot)$ to be the measure which chooses ${\sigma_\theta}^2$, ${\sigma_e}^2$, μ, $\theta_1, \ldots, \theta_K$ as follows. Choose ${\sigma_\theta}^2$, ${\sigma_e}^2$ and μ independently, with ${\sigma_\theta}^2$ uniform on the set

$$I_{{\sigma_\theta}^2} = \left[v_2 - \left(\frac{1}{J-1}\right)v_1 - \frac{1}{\sqrt{K}}, v_2 - \left(\frac{1}{J-1}\right)v_1 + \frac{1}{\sqrt{K}}\right],$$

${\sigma_e}^2$ uniform on the set

$$I_{{\sigma_e}^2} = \left[\left(\frac{J}{J-1}\right)v_1 - \frac{1}{\sqrt{JK}}, \left(\frac{J}{J-1}\right)v_1 + \frac{1}{\sqrt{JK}}\right]$$

and μ uniform on the set

$$I_\mu = \left[\bar{\bar{Y}} - \frac{1}{\sqrt{K}}, \bar{\bar{Y}} + \frac{1}{\sqrt{K}}\right].$$

Then choose the θ_i according to their "correct" conditional distributions, that is, chosen independently from the normal distribution

$$N\left(\frac{J{\sigma_\theta}^2 \bar{Y}_i + {\sigma_e}^2 \mu}{J{\sigma_\theta}^2 + {\sigma_e}^2}, \frac{{\sigma_\theta}^2 {\sigma_e}^2}{J{\sigma_\theta}^2 + {\sigma_e}^2}\right).$$

Thus, $Q(\cdot)$ picks ${\sigma_\theta}^2$, ${\sigma_e}^2$ and μ independently, but then picks $\theta_1, \ldots, \theta_K$ in a dependent fashion.

Now, if $(\sigma_\theta^{2(k)}, \sigma_e^{2(k)}, \mu^{(k)}, \theta_1^{(k)}, \ldots, \theta_K^{(k)}) \in R_2$, then the Markov chain will proceed as follows. First $\sigma_\theta^{2(k+1)}$ will be chosen from the inverse gamma distribution

$$\mathscr{L}\left(\sigma_\theta^{2(k+1)} | x^{(k)}\right) = \mathrm{IG}\left(a_1 + \tfrac{1}{2}K, b_1 + \tfrac{1}{2}\sum_i T\left(\theta_i^{(k)} - \mu^{(k)}\right)^2\right).$$

[Recall that $x^{(k)}$ stands for $(\sigma_\theta^{2(k)}, \sigma_e^{2(k)}, \mu^{(k)}, \theta_1^{(k)}, \ldots, \theta_K^{(k)})$.] Now, recalling that the distribution $\mathrm{IG}(a, b)$ has mean $b/(a-1)$ and variance $b^2/((a-1)^2(a-2))$, and that $(1/K)\sum_i(\theta_i^{(k)} - \mu^{(k)})^2$ is within $1/\sqrt{K}$ of $v_2 - (1/(J-1))v_1$, we see that $\mathscr{L}(\sigma_\theta^{2(k+1)} | x^{(k)})$ has mean within $O(1/\sqrt{K})$ of $v_2 - (1/(J-1))v_1$ and variance which is $O(1/K)$. Thus, the standard deviation will be $O(1/\sqrt{K})$. Now, it is easily seen that such an inverse gamma distribution will have large overlap with the uniform distribution on the set $I_{\sigma_\theta^2}$. Specifically, write $\mathrm{IG}(a, b; x)$ for the density function of the distribution $\mathrm{IG}(a, b)$ and set

$$\varepsilon_{\sigma_\theta^2} = \min\left\{\left(\sqrt{K}/2\right)^{-1} \mathrm{IG}\left(a_1 + \frac{1}{2}K, b_1 + \frac{1}{2}Kt; x\right) \middle| x \in I_{\sigma_\theta^2}, \left|t - \left(v_2 - \left(\frac{1}{J-1}\right)v_1\right)\right| \le \frac{1}{\sqrt{K}}\right\}.$$

Then $\varepsilon_{\sigma_\theta^2}$ is easily seen to be bounded below independently of J and K. [Indeed, this just amounts to saying that the IG distribution with variance $O(K)$ has density uniformly of $O(\sqrt{K})$ everywhere within $O(\sqrt{K})$ of its mean.] Also, by construction, if $x^{(k)} \in R_2$, then

$$\mathscr{L}\left(\sigma_\theta^{2(k+1)} | x^{(k)}\right) \ge \varepsilon_{\sigma_\theta^2} \mathscr{U}_{I_{\sigma_\theta^2}}$$

as measures (where $\mathscr{U}_S$ stands for the uniform distribution on the set S).

Second, $\sigma_e^{2(k+1)}$ will be chosen from the inverse gamma distribution

$$\mathscr{L}\left(\sigma_e^{2(k+1)} | x^{(k)}\right) = \mathrm{IG}\left(a_2 + \tfrac{1}{2}KJ, b_2 + \tfrac{1}{2}\sum_{ij}\left(Y_{ij} - \theta_i^{(k)}\right)^2\right),$$

with mean within $O(1/\sqrt{JK})$ of $(J/(J-1))v_1$ and with standard deviation which is $O(1/\sqrt{JK})$. By a virtually identical argument to the above, there is $\varepsilon_{\sigma_e^2} > 0$ bounded below independently of J and K such that if $x^{(k)} \in R_2$, then

$$\mathscr{L}\left(\sigma_e^{2(k+1)} | x^{(k)}\right) \ge \varepsilon_{\sigma_e^2} \mathscr{U}_{I_{\sigma_e^2}}.$$

Specifically, we set

$$\varepsilon_{\sigma_e^2} = \min\left\{\left(\frac{\sqrt{JK}}{2}\right)^{-1} \mathrm{IG}\left(a_2 + \frac{1}{2}KJ, b_2 + \frac{1}{2}KJt; x\right) \middle| x \in I_{\sigma_e^2}, \left|t - \left(\frac{J}{J-1}\right)v_2\right| \le \frac{1}{\sqrt{JK}}\right\},$$

similar to the above.

Third, $\mu^{(k+1)}$ will be chosen from the normal distribution

$$\mathscr{L}\left(\mu^{(k+1)}|\sigma_\theta^{2(k+1)}, \theta_i^{(k)}\right) = N\left(\frac{\sigma_\theta^{2(k+1)}\mu_0 + \sigma_0^2\Sigma_i\theta_i^{(k)}}{\sigma_\theta^{2(k+1)} + K\sigma_0^2}, \frac{\sigma_\theta^{2(k+1)}\sigma_0^2}{\sigma_\theta^{2(k+1)} + K\sigma_0^2}\right).$$

Write $N(a, b; x)$ for the density function of the normal distribution and set

$$\varepsilon_\mu = \min\left\{\left(\frac{\sqrt{K}}{2}\right)^{-1} N\left(\frac{s\mu_0 + \sigma_0^2 Kt}{s + K\sigma_0^2}, \frac{s\sigma_0^2}{s + K\sigma_0^2}; x\right)\middle| x \in I_\mu, \right.$$
$$\left. s \in I_{\sigma_\theta^2}, |t - \bar{\bar{Y}}| \le \frac{1}{\sqrt{K}}\right\}.$$

Then it is easily checked that ε_μ is bounded below independently of J and K and that conditional on $\sigma_\theta^{2(k+1)} \in I_{\sigma_\theta^2}$ and $x^{(k)} \in R_2$, we will have

$$\mathscr{L}\left(\mu^{(k+1)}|\sigma_\theta^{2(k+1)}, \theta_i^{(k)}\right) \ge \varepsilon_\mu \mathscr{U}_{I_\mu}.$$

Finally, $\theta_1^{(k+1)}, \ldots, \theta_K^{(k+1)}$ will be chosen precisely from the distribution

$$\mathscr{L}\left(\theta_i^{(k+1)}|\sigma_\theta^{2(k+1)}, \sigma_e^{2(k+1)}, \mu^{(k+1)}\right)$$
$$= N\left(\frac{J\sigma_\theta^{2(k+1)}\bar{Y}_i + \sigma_e^{2(k+1)}\mu^{(k+1)}}{J\sigma_\theta^{2(k+1)} + \sigma_e^{2(k+1)}}, \frac{\sigma_\theta^{2(k+1)}\sigma_e^{2(k+1)}}{J\sigma_\theta^{2(k+1)} + \sigma_e^{2(k+1)}}\right),$$

which is precisely the same as under $Q(\cdot)$.

Combining all of this information, we conclude that if $x^{(k)} \in R_2$, then $\mathscr{L}(x^{(k+1)}|x^{(k)}) \ge \varepsilon_{\sigma_\theta^2}\varepsilon_{\sigma_e^2}\varepsilon_\mu Q(\cdot)$, as measures. Furthermore, while $\varepsilon_{\sigma_\theta^2}\varepsilon_{\sigma_e^2}\varepsilon_\mu$ may depend on J and K, it is bounded below by (say) $\varepsilon_2 > 0$ independent of J and K. This completes the proof. □

We wish to use this R_2 and $Q(\cdot)$ as in Lemma 3. We shall show that with uniform probability (i.e., with probability bounded below independently of J and K), the Markov chain will get from any starting point in some large set R_1 to the set R_2, and that this will happen in some k_0 steps, where k_0 will be $O(1 + \log K/\log J)$. This will allow us to use Lemma 3 to conclude (once we obtain bounds on a and b) that the Markov chain will converge in $O(k_0) = O(1 + \log K/\log J)$ steps.

The argument is trickiest when $K \gg J$. In this case, there are problems if σ_θ^2 gets "stuck" too close to 0. It is worthwhile to keep this case in mind to fully appreciate the difficulties involved.

We begin by letting R_* be the subset of $\mathscr{X}$ on which $(1/K)\Sigma_i(\theta_i - \mu)^2 \ge (v_2 - (1/(J - 1))v_1)/10$, $(1/JK)\Sigma_{ij}(Y_{ij} - \theta_i)^2 \le 10(v_1 + v_2)$ and $|((1/K)\Sigma_i\theta_i) - \bar{\bar{Y}}| \le 10$.

LEMMA 5. *There are $\varepsilon_*, J_0, c_* > 0$, all independent of J and K, such that assuming $J \ge J_0$, there is $k_* \le c_*(1 + \log K/\log J)$ such that if $(\sigma_\theta^{2(0)}, \sigma_e^{2(0)}, \mu^{(0)}, \theta_1^{(0)}, \ldots, \theta_K^{(0)}) \in R_*$, then with probability greater than or equal to ε_*, we will have $(\sigma_\theta^{2(k_*)}, \sigma_e^{2(k_*)}, \mu^{(k_*)}, \theta_1^{(k_*)}, \ldots, \theta_K^{(k_*)}) \in R_2$.*

PROOF. The Markov chain can be analyzed as follows. We use the notation $r = (1/K)\Sigma_i(\theta_i - \mu)^2$, $s = (1/JK)\Sigma_i(Y_{ij} - \theta_i)^2$ and $t = (1/K)\Sigma_i\theta_i$. The set R_2 then corresponds to the set $|r - (v_2 - (1/(J-1))v_1)| \le 1/\sqrt{K}$, $|s - (J/(J-1))v_1)| \le 1/\sqrt{JK}$ and $|t - \bar{\bar{Y}}| \le 1/\sqrt{K}$.

We use the following facts: (1) If X is a random variable with mean m and variance v, then the expected value $E(X^2) = m^2 + v$; (2) If $\mathscr{L}(X) = N(a, b)$, then $\mathscr{L}(X - l) = N(a - l, b)$; and (3) $(1/n)\Sigma_{i=1}^n(\alpha_i - \beta)^2 = (\beta - \bar{\alpha})^2 + (1/n)\Sigma_{i=1}^n(\alpha_i - \bar{\alpha})^2$. Using these facts it is straightforward to show that

$$E(r^{(k+1)}|r^{(k)}, s^{(k)}, t^{(k)}) = \left(\frac{Jr^{(k)}}{Jr^{(k)} + s^{(k)}}\right)^2\left(v_2 + \left(\bar{\bar{Y}} - t^{(k)}\right)^2\right) + \left(\frac{r^{(k)}s^{(k)}}{Jr^{(k)} + s^{(k)}}\right) + O\left(\frac{1}{\sqrt{K}}\right),$$

$$(*)\quad E(s^{(k+1)}|r^{(k)}, s^{(k)}, t^{(k)}) = v_1 + \left(\frac{s^{(k)}}{Jr^{(k)} + s^{(k)}}\right)^2\left(v_2 + \left(\bar{\bar{Y}} - t^{(k)}\right)^2\right) + \left(\frac{r^{(k)}s^{(k)}}{Jr^{(k)} + s^{(k)}}\right) + O\left(\frac{1}{\sqrt{JK}}\right),$$

$$E(t^{(k+1)}|r^{(k)}, s^{(k)}, t^{(k)}) = \frac{Jr^{(k)}\bar{\bar{Y}} + s^{(k)}t^{(k)}}{Jr^{(k)} + s^{(k)}} + O\left(\frac{1}{\sqrt{K}}\right).$$

For example, for $r^{(k+1)}$, we proceed as follows. Given $x^{(k)}$, we see that $\sigma_\theta^{2(k+1)}$ will have mean within $O(1/K)$ of $r^{(k)}$ and variance $O(1/K)$, that $\sigma_e^{2(k+1)}$ will have mean within $O(1/K)$ of $s^{(k)}$ and variance $O(1/JK)$, and that $\mu^{(k+1)}$ will have mean within $O(1/K)$ of $t^{(k)}$ and variance $O(1/K)$. Thus, $(\theta_i^{(k+1)} - \mu^{(k+1)})$ will be a random variable with mean within $O(1/\sqrt{K})$ of $(Jr^{(k)}/(Jr^{(k)} + s^{(k)}))(\bar{Y}_i - t^{(k)})$ and with variance $r^{(k)}s^{(k)}/(Jr^{(k)} + s^{(k)}) + O(1/K)$. Hence,

$$E(r^{(k+1)}|x^{(k)}) = \frac{1}{K}\sum_I\left(\frac{Jr^{(k)}}{Jr^{(k)} + s^{(k)}}\left(\bar{Y}_i - t^{(k)}\right)\right)^2 + \frac{r^{(k)}s^{(k)}}{Jr^{(k)} + s^{(k)}} + O\left(\frac{1}{\sqrt{K}}\right).$$

The result for $r^{(k)}$ now follows from noting that

$$\sum_i\left(\bar{Y}_i - t^{(k)}\right)^2 = \left(\bar{\bar{Y}} - t^{(k)}\right)^2 + \sum_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 = \left(\bar{\bar{Y}} - t^{(k)}\right)^2 + v_2 \cdot .$$

Furthermore, it is easily checked that

$$\begin{aligned}\mathrm{Var}(r^{(k+1)}|r^{(k)}, s^{(k)}, t^{(k)}) &= O(1/K),\\ \mathrm{Var}(s^{(k+1)}|r^{(k)}, s^{(k)}, t^{(k)}) &= O(1/JK),\\ \mathrm{Var}(t^{(k+1)}|r^{(k)}, s^{(k)}, t^{(k)}) &= O(1/K).\end{aligned}$$

We conclude that with high probability $r^{(k+1)}$, $s^{(k+1)}$ and $t^{(k+1)}$ will be close to their expected values as given above.

Now, it follows from this that after one iteration, for appropriate C_1 independent of J and K, with uniform probability we will have $r^{(1)}$, $s^{(1)}$, $t^{(1)}$ in the compact set R_{**} defined by $(v_1 - (1/(J-1))v_2)/10 \le r \le C_1(v_1 - (1/(J-1))v_2)$, $0 \le s \le 10(v_1 + v_2)$, and in addition $|t - \bar{\bar{Y}}| \le C_1/J$.

Once *this* is true, let us again consider equations $(*)$ but let us omit the $O(1/\sqrt{K})$ and $O(1/\sqrt{JK})$ error terms. Equations $(*)$ then define a *dynamical system* in the variables r, s and t. This dynamical system is seen (by directly checking) to have a fixed point

$$r^{(k)} = r_* \equiv v_2 - \left(\frac{1}{J-1}\right)v_1,$$

$$s^{(k)} = s_* \equiv \left(\frac{J}{J-1}\right)v_1,$$

$$t^{(k)} = t_* \equiv \bar{\bar{Y}}.$$

Furthermore, we see by direct computation (recalling that $|t^{(k)} - \bar{\bar{Y}}| \le C_1/J$) that all partial derivatives of the form $\partial u^{(k+1)}/\partial v^{(k)}$, where $u, v \in \{r, s, t\}$, will be $O(1/J)$ uniformly in the set R_{**}. This means that the fixed point (r_*, s_*, t_*) will have a rate of attraction which is $O(1/J)$ inside R_{**}. Thus, there is $C_2 > 0$ independent of J and K, such that the distance from $(r^{(k)}, s^{(k)}, t^{(k)})$ to (r_*, s_*, t_*) will be multiplied by a factor less than or equal to C_2/J on each iteration of the dynamical system. Hence, for $J > C_2$, after $\log(1/\sqrt{K})/\log(C_2/J) = \frac{1}{2}\log K/(\log J - \log C_2)$ iterations, the dynamical system will be within $1/\sqrt{K}$ of its fixed point.

Now, the Markov chain itself does not follow a deterministic procedure. However, it "nearly" does, in the following sense. If we set $C_3 = 2C_2$, then with very high probability the vaues $r^{(k)}$, $s^{(k)}$, $t^{(k)}$ will get closer to their fixed point by a factor less than or equal to C_3/J. Hence, provided $J \ge J_0 \equiv 2C_3$ (say), the values of $(r^{(k)}, s^{(k)}, t^{(k)})$ will with high probability get geometrically closer and closer to (r_*, s_*, t_*). This will continue until the values get to within $O(1/\sqrt{K})$ of this fixed point. However, from that close, there is uniform probability that the Markov chain will jump into the set R_2 in a single step. Furthermore, the probabilities that $(r^{(k)}, s^{(k)}, t^{(k)})$ will *fail* to get geometrically closer to (r_*, s_*, t_*) are summable and uniformly bounded above by something less than 1. Thus, there is uniform probability that $(r^{(k)}, s^{(k)}, t^{(k)})$ will proceed geometrically to the set R_2. (A similar argument is presented in greater detail as Lemma 4 of [22].) We conclude that if we set $k_* = \lceil \frac{1}{2}\log K/(\log J - \log C_3)\rceil + 2$ and set

$$\varepsilon' = \min\{P^{k_*}(x, R_2) | x \in R_1\},$$

then ε' can be bounded below independently of J and K. The lemma follows. □

REMARK. If $J \ge O(K)$, then the proof of Lemma 5 can be simplified greatly. Indeed, in that case it suffices to take $k_* = 2$ and it is not necessary to consider the iterative argument at all.

We are now in a position to prove Theorem 1, using Lemma 3. We shall make use of some technical lemmas (Lemmas 6, 7 and 8), whose statements and proofs we defer until the end.

For Theorem 1(a), we use Lemma 3 with $R_1 = R_*$, $k_0 = k_*$ and $\varepsilon_1 = \varepsilon_*$ [and with R_2, ε_2 and $Q(\cdot)$ as in Lemma 5]. Lemma 6 below shows that $\pi(R_*^C)$ and $\sup_{x \in R_*} P(x, R_*^C)$ are both bounded by expressions of the form $c_1 \exp(-c_2 k)$. Theorem 1(a) then follows directly from Lemmas 3 (with Remark 2 following it), 4, 5 and 6, and a little bit of rearranging, with $B = -\log(1 - \varepsilon_1 \varepsilon_2)$, $A_1 = 2c_1 + 1$ and $A_2 = c_2$. The factor of 1.1 is included simply to avoid reference to the greatest integers less than certain values; the extra 0.1 leeway, together with appropriate choices of the constants, takes care of this.

For Theorem 1(b), we fix a number of iterations k and let R_1 be the subset of $\mathscr{X}$ on which $|\bar{\bar{Y}} - \sum_i \theta_i| \leq k^{1/4}$. (Note that we are letting the set R_1 depend on the total number of iterations; see the first remark after Lemma 2 above.) Lemma 7 below then states that if $x^{(0)} \in R_1$, then with probability greater than or equal to $\delta_1 > 0$ independent of J, K and k, we will have $x^{(k_1)} \in R_*$, where k_1 is $O(K^3 \sqrt{k} \log k)$. Combining this with Lemma 5, we conclude that with probability greater than or equal to $\varepsilon_* \delta_1$, we will have $x^{(k_0)} \in R_2$, where $k_0 = k_1 + k_*$. Lemma 8 below states that $\pi(R_1^C)$ and $\sup_{x \in R_1} P^{k_0+1}(x, R_1^C)$ are both bounded by expressions of the form $c_3 \exp(-c_4 \sqrt{k})$ for c_3, c_4 independent of J, K and k. Using all of this information, Theorem 1(b) now follows from Lemma 3, with ε_2, $Q(\cdot)$ and R_2 as in Lemma 4, with $k_0 = k_1 + k_*$, with R_1 as above, with $\varepsilon_1 = \delta_1 \varepsilon_*$ and with $B' = \min(c_4, -\log(1 - \varepsilon_1 \varepsilon_2))$ and $A_3 = 2.1 + 2c_3$.

We now proceed to the missing lemmas.

LEMMA 6. *Let $\pi(\cdot)$ be the true posterior distribution for the variance component model. Let $P(\cdot, \cdot)$ be the transition probabilities for the Gibbs sampler. Let R_* be as above and let*

$$a = \pi(R_*^C), \qquad b = \max_{x \in R_*} P(x, R_*^C).$$

Then there are $J_0, c_1, c_2 > 0$ independent of J and K, such that $a \leq c_1 \exp(-c_2 K)$ and $b \leq c_1 \exp(-c_2 K)$ provided $J \geq J_0$.

PROOF. We examine b first. We again recall equations $(*)$ from the proof of Lemma 5. Those equations imply that

$$E(r^{(k+1)} | x^{(k)}) \geq \left(\frac{J r^{(k)}}{J r^{(k)} + s^{(k)}} \right)^2 v_2.$$

Hence, if $x^{(k)} \in R_*$, then we will have

$$E(r^{(k+1)} | x^{(k)}) \geq \left(\frac{J(v_2/10)}{J(v_2/10) + 10(v_1 + v_2)} \right)^2 v_2.$$

For sufficiently large J, this expression will be greater than $v_2/5$ (say), which is "safely inside R_*." Furthermore, the variance of $r^{(k+1)}$ given $x^{(k)}$ will be $O(1/K)$. In addition, all the possible choices for the random variables involved [i.e., σ_θ^2, σ_e^2, μ, $\Sigma\theta_i$, $\Sigma(\theta_i - \mu)^2$ and $\Sigma(\theta_i - Y_{ij})^2$] are made from distributions which fall off at least as fast as $e^{-(\text{const})K_y}$ or $y^{-(\text{const})K}$ away from their modes. We conclude that $\text{Prob}(r^{(k+1)} < v_2/10 | x^{(k)}) \leq (\text{const})e^{-(\text{const})K}$ for $x^{(k)} \in R_*$. (Of course, here "const" may depend on v_1, v_2 and the priors, but it is independent of J and K.)

In an entirely similar manner, we conclude that $\text{Prob}(s^{(k+1)} > 10(v_1 + v_2) | x^{(k)}) \leq (\text{const})e^{-(\text{const})K}$ and that $\text{Prob}(|t^{(k+1)} - \bar{\bar{Y}}| > 10 | x^{(k)}) \leq (\text{cosnt})e^{-(\text{const})K}$, for sufficiently large J. The conclusion about b follows.

To understand a, we take a direct approach. Recall that $\pi(\cdot)$ is the posterior distribution for the model in question. Thus, the density for $\pi(\cdot)$ is proportional to

$$\mathscr{L}\left(Y_{ij}, \sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K\right) = \text{IG}\left(a_1, b_1; \sigma_\theta^2\right) \times \text{IG}\left(a_1, b_2; \sigma_e^2\right) \times N\left(\mu_0, \sigma_0^2; \mu\right) \times \prod_{i=1}^{K} \left(N\left(\mu, \sigma_\theta^2; \theta_i\right) \prod_{j=1}^{J} N\left(\theta_i, \sigma_e^2; Y_{ij}\right) \right).$$

For fixed Y_{ij}, this density is easily seen to be largest when μ is within $O(1/\sqrt{K})$ of $\bar{\bar{Y}}$, θ_i is within $O(1/\sqrt{J})$ of $\bar{Y}_i$, σ_θ^2 is within $O(1/\sqrt{K})$ of v_2 and σ_e^2 is within $O(1/\sqrt{JK})$ of v_1. Furthermore, it is easily seen to fall off at least as fast as $(\text{const})e^{-(\text{const})Ky}$ or $y^{-(\text{const})K}$ away from this mode. The conclusion about a now follows from straightforward bounding, and amounts to observing that the ratio of the density for $\pi(\cdot)$ near its mode and far from its mode (in R_*) is sufficiently small. □

LEMMA 7. *Let k be a positive integer and let R_1 be the subset of $\mathscr{X}$ defined by $|\bar{\bar{Y}} - \Sigma_i \theta_i| \leq k^{1/4}$. Then there are $J_0, c, \delta_1 > 0$ independent of J, K and k, such that*

$$\inf_{x \in R_1} P\left(x^{(k_1)} \in R_1 | x^{(0)} = x\right) \geq \delta_1,$$

for some $k_1 \leq c\sqrt{k}\,((K/J)\log k + K^6/J^2)$, provided $J \geq J_0$.

PROOF. The idea of the proof is that $|\bar{\bar{Y}} - \Sigma_i \theta_i|$ may be very large at the beginning, but for "reasonable" values of σ_θ^2 and σ_e^2, it will tend to get smaller by a factor $O(1/J)$ at each iteration, so that the Markov chain will approach R_* rather rapidly. The only problem is that if σ_θ^2 is "stuck" at a very small value, then special care must be taken. (We repeat again that this is only an issue if $K \gg J$; if not, then σ_θ^2 returns to "reasonable" values at most iterations.)

Set $t^{(n)} = \Sigma_i \theta_i^{(n)}$ as before and assume that $x^{(0)} \in R_1$. We proceed as follows. On the first iteration, with uniform probability we will have $\mu^{(1)}$ close to $t^{(0)}$ and $|\bar{\bar{Y}} - t^{(1)}| \leq k^{1/4}$. After that, referring to equations $(*)$ from the proof of Lemma 5, we see directly that with uniform probability $s^{(2)}$ will be

less than or equal to (const)$\sqrt{k}$ and $|\overline{\overline{Y}} - \mu^{(2)}|$ will be less than or equal to (const)$k^{1/4}$. (Here "const" means independent of J, K and k.) If $\sigma_\theta^{2(2)}$ were bounded away from 0, then it would be clear that $t^{(n)}$ would rapidly approach $\overline{\overline{Y}}$. To handle the general case, we observe that the mean of σ_θ^2 at each iteration will be at least $b_1/(a_1 + K/2)$. Thus, with uniform probability we will have $\sigma_\theta^2 \geq$ (const)$/K$ in at least half (say) of the iterations.

Now, recall that the mean of θ_i at each iteration is $(J\sigma_\theta^2 \overline{Y}_i + \sigma_e^2\mu)/(J\sigma_\theta^2 + \sigma_e^2)$. Hence, the mean of $\Sigma_i \theta_i$ at each iteration is $(J\sigma_\theta^2 \overline{\overline{Y}} + \sigma_e^2\mu)/(J\sigma_\theta^2 + \sigma_e^2)$. Hence, $|\overline{\overline{Y}} - \Sigma_i \theta_i|$ is roughly $(\sigma_e^2/(J\sigma_\theta^2 + \sigma_e^2))|\mu - \overline{\overline{Y}}| = (1/(1 + J\sigma_\theta^2/\sigma_e^2))|\mu - \overline{\overline{Y}}|$. Recalling that μ has mean within $O(1/K)$ of $(1/K)\Sigma_i \theta_i$, we see that, up to $O(1/K)$, with uniform probability $|\overline{\overline{Y}} - \Sigma\theta_i|$ gets multiplied by about $1/(1 + J\sigma_\theta^2/\sigma_e^2)$ at each iteration. If $\sigma_\theta^2 \geq$ (const)$/K$ and $\sigma_e^2 \leq$ (const)$\sqrt{k}$, then this factor is less than or equal to $1/$ (1 + (const)$J/K\sqrt{k}$), which for large k is less than or equal to $\exp(-$(const)$J/K\sqrt{k})$.

We conclude from all of this that after $k' =$ (const)$K\sqrt{k}(\log k)/J$ iterations, with uniform probability, the value of $|\overline{\overline{Y}} - \Sigma_i \theta_i|$ will have become less than 10 (say). From then on, the equations $(*)$ imply that the values of $s^{(n)}$ and $|\overline{\overline{Y}} - \Sigma_i \theta_i|$ will tend to remain "reasonable." Thus, we would be done except for the lingering problem that σ_θ^2 may be "stuck" too close to 0.

To handle this problem, we refer again to equations $(*)$ from the proof of Lemma 5. Direct computation implies that regardless of the values of $s^{(n)}$ and $t^{(n)}$, we have $E(r^{(n+1)}|x^{(n)}) \geq r^{(n)}$ for small $r^{(n)}$ (although it is very close). Thus, if $r^{(n)}$ is small, then $r^{(n+1)} - r^{(n)}$ has nonnegative mean. Also, it is easily seen to have variance at least $(J(b_1/K^3)/(J(b_1/K^3) + s^{(n)}))^2$, even if $\theta_i^{(n)} = \mu^{(n)}$ for all i. [Here the factor (b_1/K^3) is from the variance of $\sigma_\theta^{2(n+1)}$. The rest of the expression comes from the way the law of θ_i depends on σ_θ^2.] Now, as long as $s^{(n)}$ is bounded independently of J, K and k, we conclude that this variance is greater than or equal to $O(J^2/K^6)$.

Combining these two facts and writing

$$r^{(k'+n)} = (r^{(k'+1)} - r^{(k')}) + (r^{(k'+2)} - r^{(k'+1)}) + \cdots + (r^{(k'+n)} - r^{(k'+n-1)}),$$

we see that $r^{(k'+n)} - r^{(k')}$ will have nonnegative mean and variance greater than or equal to (const)$nJ^2/K^6\sqrt{k}$. It follows that for $n = k'' =$ (const)$K^6\sqrt{k}/J^2$, with uniform probability we will have $r^{(k''+k')} \geq {}_2/10$ (say). Also, with uniform probability $s^{(k'+k'')}$ and $t^{(k'+k'')}$ will have stayed "reasonable," and we will have $x^{(k'+k'')} \in R_*$.

Putting all of this together, the lemma follows with $k_1 = k' + k''$. □

LEMMA 8. *Let $\pi(\cdot)$ be the true posterior, let k_* be as in Lemma 5 and let k, R_1 and k_1 be as in Lemma 7. Let $t = k_1 + k_* + 1$ and let $a = \pi(R_1^C)$ and $b_t = \inf_{x \in R_1} P^t(x, R_1^C)$. Then a and b_t are bounded above by expressions of the form $c_3 \exp(-c_4\sqrt{k})$, with $c_3, c_4 > 0$ independent of J, K and k.*

PROOF. The assertion about a folows from a similar argument to that used in Lemma 6 and is omitted. For the assertion about b_t, arguing as in Lemma 7 we note that if the Markov chain begins inside R_1, then the value

of $|\bar{\bar{Y}} - \Sigma_i \theta_i|$ after t steps will tend to be small. It is straightforward to argue (by considering the tails of the normal distribution) that the probability that it will be greater than $k^{1/4}$ will be less than or equal to $c_3 \exp(-c_4\sqrt{k})$ for appropriate $c_3, c_4 > 0$ independent of J, K and k. □

REMARKS.

1. *On lower bounds.* Theorem 1 provides only *upper bounds* on how many iterations are required for the Gibbs sampler Marov chain to converge. However, the upper bound of $O(1 + \log K/\log J)$, gotten from part (a) of Theorem 1 by ignoring the (small) second term, is easily seen to be "sharp up to constants." That is, if the number of iterations done is small compared to $1 + \log K/\log J$, then the distance to stationarity will be close to 1, for all sufficiently large K and J. Indeed, if $J \geq O(K)$, then $O(1 + \log K/\log J) = O(1)$. However, we obviously need at least one iteration, so this rate is clearly correct up to a constant. Also, for $K \gg J$, $O(1 + \log K/\log J) = O(\log K/\log J)$, and we claim the quantity $O(\log K/\log J)$ is also necessary to get close to stationarity. Indeed, if we do a number of iterations which is small compared to $\log K/\log J$, then arguing as in Lemma 5, we see that the probability will be quite small that we will be within $1/\sqrt{K}$ of the fixed point (r_*, s_*, t_*) (unless we started exactly there). However, it is also easily seen (arguing as in Lemma 6) that $\pi(\cdot)$ has most of its mass in this range. Thus, we conclude that *the variation distance to* $\pi(\cdot)$ *will be quite close to* 1 *if the number of iterations done is small compared to* $\log K/\log J$.
2. In principle, Lemma 2 can be used to get rates of convergence for any Markov chain. The computations, of course, will vary from chain to chain, but the idea of Lemma 3, in which with large probability the Markov chain will go to a certain small "good" set R_2 within a certain number of iterations k_0, would appear to be quite applicable to Gibbs sampling situations in wich there is lots of data. In such situations, the data will "swamp" the conditional distributions, and they will tend to pile up on certain particular values (roughly corresponding to the mode of the posterior). Choosing $Q(\cdot)$ appropriately should allow Lemma 3 to give good rates of convergence for quite a variety of Gibbs sampler problems.

APPENDIX

Variation distance and coupling. Lemma 2 above provides a bound on the variation distance between two measures, using the coupling inequality. Coupling is widely used in Markov chain theory (see, e.g., [19] or Chapter 4E of [7]), but it may be less familiar to statisticians. For completeness, we review it briefly here.

Given probability measures ν_1 and ν_2 defined on the same probability space, the *variation distance* between them is defined to be

$$\|\nu_1 - \nu_2\|_{\text{var}} \equiv \sup_A |\nu_1(A) - \nu_2(A)|,$$

where the supremum is taken over all measurable subsets A. This distance gives a good idea of how much the measure ν_1 differs from the measure ν_2.

Given a Markov chain $P(\cdot, \cdot)$ with stationary distribution $\pi(\cdot)$, suppose we are able to define random variables $\{X_k\}$ and $\{Y_k\}$ such that:

1. $\mathcal{L}(X_{k+1}|X_k) = P(X_k, \cdot)$.
2. $\mathcal{L}(Y_{k+1}|Y_k) = P(Y_k, \cdot)$.
3. $\mathcal{L}(Y_0) = \pi(\cdot)$.

Conditions 1 and 2 say that each of X_k and Y_k marginally follow the transition law $P(\cdot, \cdot)$, so that condition 3 then implies that $\mathcal{L}(Y_k) = \pi(\cdot)$ for all k. The variables $\{(X_k, Y_k)\}$ are then a *coupling* if there is a random time T such that $X_k = Y_k$ for all $k \geq T$.

The *coupling inequality* then says that the variation distance between $\mathcal{L}(X_k)$ and $\pi(\cdot)$ is bounded above by the probability that $T > k$:

$$\|\mathcal{L}(X_k) - \pi(\cdot)\| \leq \text{Prob}(T > k).$$

To prove this, we simply note that for any subset A,

$$\begin{aligned}
&|\mathcal{L}(X_k)(A) - \pi(A)| \\
&\quad = |\text{Prob}(X_k \in A) - \text{Prob}(Y_k \in A)| \\
&\quad = |\text{Prob}(X_k \in A, X_k = Y_k) + \text{Prob}(X_k \in A, X_k \neq Y_k) \\
&\qquad - \text{Prob}(Y_k \in A, X_k = Y_k) - \text{Prob}(Y_k \in A, X_k \neq Y_k)| \\
&\quad = |\text{Prob}(X_k \in A, X_k \neq Y_k) - \text{Prob}(Y_k \in A, X_k \neq Y_k)| \\
&\quad \leq \text{Prob}(X_k \neq Y_k) \leq \text{Prob}(T > k).
\end{aligned}$$

This completes the proof.

Coupling thus provides a simple method for bounding the variation distance to stationarity for a Markov chain. The trick then becomes how to define the random variables $\{X_k\}$ and $\{Y_k\}$ in such a way that they are a coupling with a useful coupling time T. Lemma 2 explains how to do this under the additional hypothesis that $P^{k_0}(x, \cdot) \geq \varepsilon Q(\cdot)$ for all $x \in R$.

Acknowledgments. This work was part of the author's doctoral dissertation [21] at Harvard University. I am very grateful to Persi Diaconis, my Ph.D. advisor, for introducing me to this area and for many helpful discussions. I thank the referee for helpful comments.

REFERENCES

[1] APPLEGATE, D., KANNAN, R. and POLSON, N. (1990). Random polynomial time algorithms for sampling from joint distributions. Technical Report 500, School of Computer Science, Carnegie-Mellon Univ.

[2] ASMUSSEN, S. (1987). *Applied Probability and Queues*. Wiley, New York.

[3] ATHREYA, K. B. and NEY, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.* **245** 493–501.

[4] ATHREYA, K. B., MCDONALD, D. and NEY, P. (1978). Limit theorems for semi-Markov processes and renewal theory for Markov chains. *Ann. Probab.* **6** 788–797.
[5] BESAG, J., YORK, J. and MOLLIÉ, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43** 1–21.
[6] BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis* Chap. 5. Addison-Wesley, Reading, MA.
[7] DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. IMS, Hayward, CA.
[8] GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.
[9] 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. *J. Amer. Statist. Soc.* **85** 972–985.
[10] GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
[11] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6** 721–741.
[12] GEYER, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* **7** 473–503.
[13] GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B.* **54** 657–699.
[14] HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.
[15] LIU, J., WONG, W. H. and KONG, A. (1991). Correlation structure and convergence rate of the Gibbs sampler I, II. Technical Reports, Dept. Statistics, Univ. Chicago.
[16] MATTHEWS, P. (1993). A slowly mixing Markov chain with implications for Gibbs sampling. *Statist. Probab. Lett.* **17** 231–236.
[17] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1091.
[18] NUMMELIN, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge Univ. Press.
[19] PITMAN, J. W. (1976). On coupling of Markov chains. *Z. Wahrsch. Verw. Gebiete* **35** 315–322.
[20] ROBERTS, G. O. (1992). Convergence diagnostics of the Gibbs sampler. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 777–784. Oxford Univ. Press.
[21] ROSENTHAL, J. S. (1992). Rates of convergence for Gibbs sampler and other Markov chains. Ph.D. dissertation, Dept. Mathematics, Harvard Univ.
[22] ROSENTHAL, J. S. (1993). Rates of convergence for data augmentation on finite sample spaces. *Ann. Appl. Probab.* **3** 819–839.
[23] SCHERVISH, M. J. and CARLIN, B. P. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics* **1** 111–127.
[24] TANNER, M. and WONG, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.
[25] TIERNEY, L. (1994). Markov Chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO
CANADA M5S 1A1