Honest Exploration of Intractable Probability Distributions Via Markov Chain Monte Carlo

Galin L. Jones
School of Statistics
University of Minnesota

James P. Hobert*
Department of Statistics
University of Florida

July 2001

* This research partially supported by NSF Grant DMS-00-72827.

Abstract

Two important questions that must be answered whenever a Markov chain Monte Carlo (MCMC) algorithm is used are (Q1) What is an appropriate *burn-in*? and (Q2) How long should the sampling continue after burn-in? Developing rigorous answers to these questions presently requires a detailed study of the convergence properties of the underlying Markov chain. Consequently, in most practical applications of MCMC, exact answers to (Q1) and (Q2) are not sought. The goal of this paper is to demystify the analysis that leads to honest answers to (Q1) and (Q2). The authors hope that this article will serve as a bridge between those developing Markov chain theory and practitioners using MCMC to solve practical problems.

The ability to formally address (Q1) and (Q2) comes from establishing a *drift condition* and an associated *minorization condition*, which together imply that the underlying Markov chain is *geometrically ergodic*. In this paper, we explain exactly what drift and minorization are as well as how and why these conditions can be used to form rigorous answers to (Q1) and (Q2). The basic ideas are as follows. The results of Rosenthal (1995) and Roberts and Tweedie (1999) allow one to use drift and minorization conditions to construct a *formula* giving an analytic upper bound on the distance to stationarity. A rigorous answer to (Q1) can be calculated using this formula. The desired characteristics of the target distribution are typically estimated using ergodic averages. Geometric ergodicity of the underlying Markov chain implies that there are central limit theorems available for ergodic averages (Chan and Geyer 1994). The regenerative simulation technique (Mykland, Tierney and Yu 1995, Robert 1995) can be used to get a consistent estimate of the variance of the asymptotic normal distribution. Hence, an asymptotic standard error can be calculated, which provides an answer to (Q2) in the sense that an appropriate time to stop sampling can be determined. The methods are illustrated using a Gibbs sampler for a Bayesian version of the one-way random effects model and a data set concerning styrene exposure.

Key words and phrases: Central limit theorem; Convergence rate; Coupling inequality; Drift condition; General state space; Geometric ergodicity; Gibbs sampler; Hierarchical random effects model; Metropolis algorithm; Minorization condition; Regeneration; Splitting; Uniform ergodicity.

1 Introduction

1.1 The questions

During the decade or so since the appearance of the seminal paper by Gelfand and Smith (1990), Markov chain Monte Carlo (MCMC) methods have revolutionized statistical computing. While the Bayesians have certainly made the most use of MCMC, applications have popped up in many different areas of statistics. For example, MCMC techniques can be used to calculate p-values in exact conditional inference (Diaconis and Sturmfels 1998) and to maximize intractable likelihood functions associated with generalized linear mixed models (McCulloch 1997). Furthermore, the popularity of the BUGS (Bayesian inference Using Gibbs Sampling) software package (Spiegelhalter, Thomas and Best 1999) indicates that MCMC is used routinely in applied work. An excellent introduction to MCMC is the book edited by Gilks, Richardson and Spiegelhalter (1996).

MCMC methods allow us to circumvent many of the difficulties associated with drawing random samples from complex, high-dimensional probability distributions. Unfortunately, using a Markov chain instead of a random sample creates new problems that must be solved before we can use MCMC methodology with the same level of confidence that we have in classical Monte Carlo methods. Our goal in this paper is to spell out exactly what these new problems are and to explain how they can be solved.

We begin with classical Monte Carlo integration. Suppose we want to know the value of $E_\pi g := \int g(\boldsymbol{x})\,\pi(\boldsymbol{x})\,d\boldsymbol{x}$, where π is a probability density and g is some real-valued function, but this integral cannot be evaluated analytically. Classical Monte Carlo integration requires an independent and identically distributed (iid) sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ from π. By the Strong Law of Large Numbers, with probability 1,

$$\bar{g}_n := \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i) \to E_\pi g \quad \text{as } n \to \infty. \tag{1}$$

Moreover, if we know that $E_\pi g^2$ is finite, there is a central limit theorem (CLT) for $\bar{g}_n$; that is, $\sqrt{n}\,(\bar{g}_n - E_\pi g) \xrightarrow{d} \mathrm{N}(0, \sigma^2)$ and, of course, the usual sample variance of the $g(\boldsymbol{X}_i)$'s is an unbiased estimate of σ^2. Hence, it is simple to construct a Monte Carlo standard error for $\bar{g}_n$ and this can

be used to decide upon a reasonable value of n.

The fundamental idea underlying MCMC is that even when obtaining iid draws from π is prohibitively difficult, it may still be feasible to simulate a Markov chain, $\boldsymbol{X}_0^*, \boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \ldots$ that has stationary density π. Let f_n denote the density of $\boldsymbol{X}_n^*$. Under a few simple regularity conditions (described in Section 2) $\boldsymbol{X}_n^*$ converges in *total variation* to a random variable from π; that is,

$$\frac{1}{2}\int |\, f_n(\boldsymbol{x}) - \pi(\boldsymbol{x})\,|\, d\boldsymbol{x} \;\downarrow\; 0 \quad \text{as } n \to \infty\,.$$

(This is much stronger than convergence in distribution.) More importantly for us, these same regularity conditions imply that the Ergodic Theorem holds and hence, with probability 1,

$$\bar{g}_n^* := \frac{1}{n}\sum_{i=0}^{n-1} g(\boldsymbol{X}_i^*) \to E_\pi g \quad \text{as } n \to \infty\,. \tag{2}$$

Unfortunately, $E_\pi g^2 < \infty$ is no longer enough to guarantee a CLT (Meyn and Tweedie 1993, Chapter 17). Indeed, sufficient conditions for a CLT involve the convergence rate (or *mixing* properties) of the Markov chain and, as we will see, these conditions can be quite difficult to verify in practice.

Whenever the MCMC method is employed, the user should give serious thought to the following two questions:

> (Q1) When should sampling begin? That is, how long does it take the Markov chain to get sufficiently close to the stationary distribution; i.e., what is an appropriate *burn–in*?
>
> (Q2) How long should the sampling continue after burn-in? That is, how do we know when the estimates based on the output are sufficiently accurate or, put another way, what are the standard errors of the estimates?

Observe that having to deal with (Q1) and (Q2) is a "new" problem in the sense that, when it is possible to make iid draws from π, (Q1) is moot and (Q2) is easy. (Actually, there is an ongoing debate in the MCMC community over the usefulness of burn-in; see Subsection 2.1 for some discussion.) In most practical applications of MCMC, (Q1) and (Q2) are not rigorously

addressed. Instead, a mixture of intuition, experience, and *ad hoc* methods are used to determine the amount of burn-in and the accuracy of the resulting estimates. One has to wonder how this affects the quality of any subsequent inferences. In this paper, we will explain how to develop rigorous answers to (Q1) and (Q2).

1.2 Honest answers

In this paper we consider (what is currently) the most straightforward method of developing rigorous answers to (Q1) and (Q2) for Markov chains on general state spaces. This method is applicable only when the underlying chain converges to its stationary distribution at a geometric rate; i.e., when the chain is *geometrically ergodic* (Meyn and Tweedie 1993, Chapter 15). Generally speaking, geometrically ergodic chains are "good" in the sense that they can be expected to quickly produce output that is similar to what one would get by sampling directly from the target distribution. The following example, introduced by Roberts and Rosenthal (1998a), is intended to illustrates the potential difference between geometric and subgeometric (slower than geometric) convergence.

EXAMPLE 1. Suppose the target distribution is Exp(1); that is, $\pi(x) = e^{-x} I(x > 0)$. Consider an independence Metropolis sampler with an Exp(θ) proposal; i.e., the proposal density is $p(x) = \theta e^{-\theta x} I(x > 0)$. The chain evolves as follows: Let the current state be $X_n = x$. Draw $y \sim \text{Exp}(\theta)$ and set $X_{n+1} = y$ with probability

$$\min\left\{\frac{\pi(y)p(x)}{\pi(x)p(y)}, 1\right\} = \exp\{(x-y)(1-\theta)\} \wedge 1 \ ;$$

otherwise, set $X_{n+1} = x$. A more algorithmic way to think of this is as follows: Draw $y \sim \text{Exp}(\theta)$ and independently draw $u \sim \text{Uniform}(0,1)$. If $u < \exp\{(x-y)(1-\theta)\}$ then set $X_{n+1} = y$, otherwise set $X_{n+1} = x$. (See Chib and Greenberg (1995) for a nice introduction to the Metropolis-Hastings algorithm and Billera and Diaconis (2001) for an interesting geometric interpretation of the algorithm.)

Note that if $\theta = 1$, this algorithm provides iid draws from the target distribution. Results in Mengersen and Tweedie (1996) can be used to show that the chain is geometrically ergodic

if $0 < \theta < 1$ and subgeometric if $\theta > 1$. (See Subsection 3.2 for more details.) The problem in the $\theta > 1$ case is that the tails of the proposal density are too light relative to the target density. This makes it difficult for the chain to reach larger values in the state space and, when it does, it tends to "get stuck" there for long periods. See J. S. Rosenthal's web page at `http://markov.utstat.toronto.edu/jeff/java/exp.html` for a graphical illustration of this Markov chain for several different values of θ.

Let $X_0, X_1, X_2, \ldots$ denote this independence Metropolis sampler with $\theta = 0.5$ and starting value $X_0 = 1$. We ran 10,000 independent copies of this chain for 15 iterations in order to see how close the distribution of X_{15} is to Exp(1). (It is actually easy to get an upper bound on the total variation distance between X_{15} and Exp(1); see Subsection 3.2.) The top plot in Figure 1 is a histogram of the 10,000 iid copies of X_{15} along with (an appropriately scaled version of) the Exp(1) density. This plot suggests (and theory confirms) that the distribution of X_{15} is very close to Exp(1) and hence this Markov chain converges quickly. We performed the same experiment for the subgeometric chain corresponding to $\theta = 4$, and the results are shown in the middle plot of Figure 1. Judging from this histogram, this Markov chain is still quite far from stationarity after 15 iterations. The spike in the histogram at the value 1 is due to the fact that in many of the 10,000 runs, the chain was stuck at the starting value for all 15 iterations. In the third and last performance of the experiment, we ran the subgeometric chain for 1,000 iterations instead of just 15. The results are shown in the bottom plot of Figure 1. While it appears that there is reasonable agreement between the histogram and the invariant density, the distribution of X_{1000} is actually still quite far from Exp(1). In particular, there are serious discrepancies near 0 and in the tail. For example, only 2 of the 10,000 X_{1000}'s were larger than 4. In a random sample of size 10,000 from the Exp(1) distribution, we expect to see about 180 observations larger than 4.

Of course, not all geometrically ergodic chains converge as quickly as the $\theta = 0.5$ chain, and not all subgeometric chains behave as badly as the $\theta = 4$ chain. Indeed, when θ is just a little larger than 1, this independence Metropolis sampler works pretty well. However, in practical applications of MCMC, the user does not have the luxury of comparing the distribution of $\boldsymbol{X}_n$ to the stationary distribution for various values of n. Establishing geometric ergodicity provides the

user with "peace of mind" concerning the mixing rate of the Markov chain, and, as we will see, allows the user to formally answer (Q1) and (Q2).

||

The method described herein for developing exact answers to (Q1) and (Q2) requires that one establish a *drift condition* and an associated *minorization condition* for the underlying Markov chain, which together imply geometric ergodicity (Meyn and Tweedie 1993, Chapters 15 & 16). It is difficult to discuss drift and minorization before describing some basic ideas and notation from Markov chain theory (see Subsection 2.1), but we can describe how they are used to give rigorous answers to (Q1) and (Q2).

Once drift and minorization have been established, the results of Rosenthal (1995) or Roberts and Tweedie (1999) can be employed to calculate a bound on exactly how many iterations are necessary to get within a prespecified (total variation) distance of the target distribution. In other words, we can find an n' such that

$$\frac{1}{2} \int | f_{n'}(\boldsymbol{x}) - \pi(\boldsymbol{x}) | \, d\boldsymbol{x} < 0.01, \quad \text{say.}$$

The value n' is an "honest" answer to (Q1).

Typically, the characteristics of the target distribution that we desire are estimated using ergodic averages like (2). As mentioned above, a finite second moment guarantees a CLT in the iid case, but is *not* sufficient when using a Markov chain. Chan and Geyer (1994) have shown, however, that geometric ergodicity of the Markov chain together with (a bit more than) a finite second moment guarantees the following CLT

$$\sqrt{n} \, (\bar{g}_n^* - E_\pi g) \xrightarrow{d} \mathrm{N}(0, \sigma_*^2) \tag{3}$$

where σ_*^2 is an appropriate if complex variance. It is important to recognize that $\bar{g}_n$ and $\bar{g}_n^*$ are quite different estimates of $E_\pi g$ and hence typically $\sigma_*^2 \neq \sigma^2$. Moreover, the usual sample variance of the $g(\boldsymbol{X}_i^*)$'s will not generally be a consistent estimate of σ_*^2. Fortunately, a consistent estimate of σ_*^2 can be formed using the *regenerative simulation* technique, which requires a minorization condition (Mykland et al. 1995, Robert 1995). Hence, we are able to calculate asymptotic standard

errors for the estimates based on, say, the first n iterations, and then continue to run the chain if these standard errors are unacceptably large. Thus, we are able to provide an "honest" answer to (Q2).

1.3 An example

In Section 6 we provide an example of a realistic application of the methods described in this paper by conducting a detailed study of a Gibbs sampler for a Bayesian hierarchical model. We now describe this model and give the reader a taste of what can be accomplished using these methods.

The model, which we refer to as $(\mathcal{M})$, is a Bayesian version of the standard, normal theory one-way random effects model with conjugate priors. It has three levels. First, conditional on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ and λ_e, the data values, Y_{ij}, are independent with

$$Y_{ij}|\theta_i, \lambda_e \sim \mathrm{N}(\theta_i, \lambda_e^{-1})$$

where $i = 1, \ldots, K$ and $j = 1, \ldots, m$. At the second stage, conditional on μ and λ_θ, $\boldsymbol{\theta}$ and λ_e are independent with

$$\boldsymbol{\theta}|\mu, \lambda_\theta \sim \mathrm{N}(\mu\mathbf{1}, \lambda_\theta^{-1}\mathbf{I}) \quad \text{and} \quad \lambda_e \sim \mathrm{Gamma}(a_2, b_2),$$

where $\mathbf{1}$ is a $K \times 1$ column vector of ones, $\mathbf{I}$ is a $K \times K$ identity matrix, and a_2 and b_2 are known positive constants. (We say $W \sim \mathrm{Gamma}(\alpha, \beta)$ if its density is proportional to $w^{\alpha-1}e^{-w\beta}I(w > 0)$.) Finally, at the third stage, μ and λ_θ are assumed independent with

$$\mu \sim \mathrm{N}(\mu_0, \lambda_0^{-1}) \quad \text{and} \quad \lambda_\theta \sim \mathrm{Gamma}(a_1, b_1)$$

where μ_0, λ_0, a_1 and b_1 are known constants; all but μ_0 are assumed to be strictly positive so that all of the distributions are proper. Also, we assume that $K \geq 3$ and that $m \geq 2$. Note that the standard, normal theory one-way random effects model (Searle, Casella and McCulloch 1992, Chapter 3) corresponds to viewing μ, λ_e and λ_θ as fixed and unknown. Let $\bar{y}_i = m^{-1}\sum_{j=1}^{m} y_{ij}$ where the y_{ij} are the observed values of the Y_{ij}.

The posterior density corresponding to model $(\mathcal{M})$ is characterized by

$$\pi(\boldsymbol{\theta}, \mu, \lambda_e, \lambda_\theta|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta}, \lambda_e)\, f(\boldsymbol{\theta}|\mu, \lambda_\theta)\, f(\lambda_e)\, f(\mu)\, f(\lambda_\theta) \tag{4}$$

where $\boldsymbol{y}$ is a vector containing all of the data and f denotes a generic density. The integrals required for inferences through this posterior are not available in closed form. Thus, we might resort to MCMC techniques like the Gibbs sampler. Indeed, variance component models have been advocated as an ideal application for the Gibbs sampler (Gelfand and Smith 1990, Gelfand, Hills, Racine-Poon and Smith 1990).

The data in Table 1 were simulated according to model ($\mathcal{M}$) with $K = 6$, $m = 8$, $a_1 = a_2 = b_1 = b_2 = \lambda_0 = 1$ and $\mu_0 = 0$. We now pretend that the origin of the data is unknown and that we desire the posterior expectations of λ_θ and λ_e under three different priors; that is, under three different hyperparameter settings. The three settings are given in Table 2. Note that setting #1 agrees exactly with the values used to simulate the data; i.e., it is the "correct" prior. Setting #2 is a "diffuse" prior, and setting #3 is the result of some experimentation to find a setting that yields a particularly short burn-in.

Consider using the block Gibbs sampler (described in Section 6) to estimate the six posterior expectations. For starting values, we will use $\theta_i = \bar{y}_i$ and $\mu = \bar{y}$. Due to the structure of the block Gibbs sampler, starting values for λ_θ and λ_e are not required.

Table 1: Simulated Data

Cell	1	2	3	4	5	6
$\bar{y}_i$	-0.22795	-1.1913	0.030547	0.48428	0.036639	-0.026581
$M_T = mK = 48$						
$\bar{y} = M_T^{-1} \sum_{i=1}^{6} \sum_{j=1}^{8} y_{ij} = -0.14906$						
$\text{SSE} = \sum_{i=1}^{6} \sum_{j=1}^{8} (y_{ij} - \bar{y}_i)^2 = 23.251$						

We first address (Q1) by finding an n' such that $\frac{1}{2} \int \mid f_{n'} - \pi \mid < 0.01$, where f_n denotes the marginal density of the nth iterate of the block Gibbs sampler and π denotes the posterior density in (4). The results are given in Table 3. For example, under the first prior, after 4300 iterations of the block Gibbs sampler, the total variation distance to stationarity is at most 0.0092. (The formula used to find n' is given in Section 4.) These burn-ins are quite manageable considering

Table 2: Three Different Prior Specifications

Hyperparameter Setting	a_1	b_1	a_2	b_2	μ_0	λ_0
1	1	1	1	1	0	1
2	0.1	0.1	0.1	0.1	0	0.1
3	3	7	6	3	0	1

that it takes only about 2 minutes to run 1 million iterations on a standard PC. Unfortunately, as we show in Section 6, things do not always work out this nicely. Now on to (Q2).

Table 3: Total Variation Bounds for the Simulated Data

Hyperparameter Setting	Iterations (n')	Bound
1	4300	0.0092
2	150000	0.0093
3	900	0.0086

Table 4 contains point estimates and asymptotic 95% confidence intervals for the posterior expectations of λ_θ and λ_e. These were obtained via the regenerative method, which requires *no burn-in*. Authors of standard textbooks on simulation (e.g. Bratley, Fox and Schrage 1987) view regenerative simulation as the preferred method for obtaining confidence intervals from simulation output. (A detailed explanation of the regenerative simulation method is given in Section 5.) Also reported in Table 4 are the $\hat{\sigma}_*^2$'s, the number of regenerations upon which the variance estimates are based, and the mean number of iterations per regeneration. Roughly speaking, the more often a Markov chain regenerates, the faster it converges. Hence, it appears that the chain corresponding to setting #3 mixes the fastest and the chain associated with setting #2 mixes the slowest. Note that this is exactly what we would have guessed based on Table 3.

Table 4: Point Estimates and Asymptotic 95% Confidence Intervals for $E(\lambda_\theta|\boldsymbol{y})$ and $E(\lambda_e|\boldsymbol{y})$

Setting	Parameter	Estimate	$\hat{\sigma}_*^2$	95% CI	Number of regenerations	Mean number of iter/regen
1	λ_θ	2.065	0.378	(2.052, 2.078)	9000	4.7
	λ_e	1.754	0.038	(1.749, 1.759)		
2	λ_θ	4.229	4.543	(4.210, 4.248)	50000	7.4
	λ_e	1.790	0.027	(1.789, 1.791)		
3	λ_θ	0.711	0.026	(0.708, 0.714)	14000	3.8
	λ_e	1.856	0.040	(1.853, 1.859)		

The remainder of this paper is organized as follows. Section 2 contains some basic Markov chain background which is illustrated using a toy Gibbs sampler. In Section 3, drift and minorization are defined and we demonstrate the type of calculations required to establish these conditions using our toy Gibbs sampler. Section 3 also contains an heuristic explanation of the theoretical connection between geometric ergodicity and drift and minorization. During this development, we derive the *coupling inequality*, which is the key result for deriving convergence rate bounds for Markov chains on general state spaces. A theorem of Rosenthal (1995) that allows one to use drift and minorization to get exact upper bounds on the distance to stationarity is stated in Section 4. Rosenthal's result is applied to our toy Gibbs sampler for illustration. In Section 5, we explain how to use regenerative simulation to calculate Monte Carlo standard errors. Section 6 contains another analysis like the one in Subsection 1.3. However, in this case, real data are used and all the details are given. The data used in Section 6 concern styrene exposure of laminators at a boat manufacturing plant (Lyles, Kupper and Rappaport 1997). Some final comments are given in Section 7.

2 Markov Chain Background

This section consists of two subsections. In Subsection 2.1, we develop some necessary notation, briefly describe the basic convergence results for ergodic Markov chains, and introduce a toy example that will be used for illustration throughout the paper. Geometric ergodicity and its consequences are described in Subsection 2.2. More general accounts of the material in this section can be found in Nummelin (1984), Meyn and Tweedie (1993), Tierney (1994), or Robert and Casella (1999).

2.1 Basics

Let $\mathcal{X} \subseteq \mathbb{R}^p$ for $p \geq 1$ and let $\mathcal{B}$ denote the associated Borel σ-algebra. Suppose that

$$\Phi = \{\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \dots\}$$

is a discrete time Markov chain with state space $\mathcal{X}$ and *Markov transition kernel* P; that is, for $\boldsymbol{x} \in \mathcal{X}$ and $A \in \mathcal{B}$,

$$P(\boldsymbol{x}, A) = \Pr(\boldsymbol{X}_{i+1} \in A | \boldsymbol{X}_i = \boldsymbol{x}) \, .$$

In order to simplify the exposition, we will often assume that the probability measure $P(\boldsymbol{x}, \cdot)$ has a (conditional) density, $k(\cdot|\boldsymbol{x})$, with respect to Lebesgue measure. That is,

$$P(\boldsymbol{x}, A) = \int_A k(\boldsymbol{u}|\boldsymbol{x}) \, d\boldsymbol{u} \, .$$

We call k the *Markov transition density*. (All the Gibbs samplers that we discuss in this paper have Markov transition densities.) For $n \in \mathbb{N} := \{1, 2, 3, \dots\}$, let P^n denote the n-step transition kernel; i.e., $P^n(\boldsymbol{x}, A) = \Pr(\boldsymbol{X}_{i+n} \in A | \boldsymbol{X}_i = \boldsymbol{x})$ so, in particular, $P \equiv P^1$. Note that $P^n(\boldsymbol{x}, \cdot)$ is the probability measure corresponding to the random variable $\boldsymbol{X}_n$, conditional on starting the chain at $\boldsymbol{X}_0 = \boldsymbol{x}$.

If π is a density such that

$$\pi(\boldsymbol{x}) = \int_{\mathcal{X}} k(\boldsymbol{x}|\boldsymbol{x}') \, \pi(\boldsymbol{x}') \, d\boldsymbol{x}' \, , \tag{5}$$

then π is called an *invariant (or stationary) density* for the Markov chain Φ. Consider the significance of equation (5). Imagine drawing $\boldsymbol{x}'$ from π and then making a single transition $\boldsymbol{x}' \to \boldsymbol{x}$

according to the Markov chain. The joint density of $(\boldsymbol{x}', \boldsymbol{x})$ induced by this recipe is exactly the integrand in (5). Hence, (5) implies that if the current state of the chain was drawn from π, then the marginal density of the next state is also π. Consequently, if the Markov chain Φ is started by taking $\boldsymbol{X}_0 \sim \pi$ (which is usually impossible in the MCMC context), then Φ is simply a sequence of *dependent* observations from π. In other words, the Markov chain is *stationary*.

Abusing notation slightly, let π also denote the probability measure associated with the density π so $\pi(A) = \int_A \pi(\boldsymbol{x})\, d\boldsymbol{x}$. The Markov chain Φ is called π-*irreducible* if for every $\boldsymbol{x} \in \mathcal{X}$ and every A with $\pi(A) > 0$, there exists an $n \in \mathbb{N}$ such that $P^n(\boldsymbol{x}, A) > 0$. In words, Φ is π-irreducible if any set with positive π-measure is *accessible* from any point in the state space. We now describe a toy Gibbs sampler that will be used for illustration several times throughout the paper.

EXAMPLE 2. Let $Y_1, \ldots, Y_m$ be iid $\mathrm{N}(\mu, \theta)$ and let the prior for (μ, θ) be proportional to $1/\sqrt{\theta}$. The posterior density is characterized by

$$\pi(\mu, \theta | \boldsymbol{y}) \propto \theta^{-\frac{m+1}{2}} \exp\left\{ -\frac{1}{2\theta} \sum_{j=1}^{m} (y_j - \mu)^2 \right\} \tag{6}$$

where $\boldsymbol{y} = (y_1, \ldots, y_m)^T$. It is easy to check that this posterior is proper as long as $m \geq 3$ and we assume this throughout. Using the Gibbs sampler to make draws from (6) requires the full conditional densities, $f(\mu|\theta, \boldsymbol{y})$ and $f(\theta|\mu, \boldsymbol{y})$, which are as follows:

$$\begin{aligned} \mu|\theta, \boldsymbol{y} &\sim \mathrm{N}(\bar{y}, \theta/m)\,, \\ \theta|\mu, \boldsymbol{y} &\sim \mathrm{IG}\left(\frac{m-1}{2}, \frac{s^2 + m(\bar{y} - \mu)^2}{2} \right)\,, \end{aligned}$$

where $\bar{y}$ is the sample mean and $s^2 = \sum (y_i - \bar{y})^2$. (We say $W \sim \mathrm{IG}(\alpha, \beta)$ if its density is proportional to $w^{-(\alpha+1)} e^{-\beta/w} I(w > 0)$.) Consider the Gibbs sampler (or data augmentation algorithm) that updates θ then μ; that is, if we let (θ', μ') denote the current state and (θ, μ) denote the future state, the transition looks like $(\theta', \mu') \to (\theta, \mu') \to (\theta, \mu)$. The state space in this case is $\mathcal{X} = \mathbb{R}^+ \times \mathbb{R}$ and the Markov transition density is

$$k(\theta, \mu | \theta', \mu') = f(\theta|\mu', \boldsymbol{y})\, f(\mu|\theta, \boldsymbol{y})\,. \tag{7}$$

In other words, the density of the new value (θ, μ) given the current state (θ', μ') is $k(\theta, \mu|\theta', \mu')$. Simulating a random variable from this density can be done sequentially by first taking $\theta \sim$

$f(\theta|\mu', \boldsymbol{y})$ followed by $\mu \sim f(\mu|\theta, \boldsymbol{y})$ (the usual Gibbs updating strategy). Note that, as is almost always the case, P^n does *not* have a closed form. Obviously (6) is not intractable in any sense, so this Gibbs sampler would never actually be used. However, its simplicity makes it ideal for demonstrating the calculation of drift and minorization conditions (see Subsection 3.1).

By construction, the posterior density is invariant for the Gibbs Markov chain; that is,

$$\pi(\mu, \theta|\boldsymbol{y}) = \int_{\mathbb{R}^+} \int_{\mathbb{R}} k(\theta, \mu|\theta', \mu')\, \pi(\mu', \theta'|\boldsymbol{y})\, d\mu'\, d\theta' \,. \tag{8}$$

The reader is invited to verify that (8) follows directly from (7). Now, if $A \in \mathcal{B}$ is such that $\pi(A) > 0$, then A must have positive Lebesgue measure. Thus, for any $(\theta', \mu') \in \mathcal{X}$, we have

$$P\big((\theta', \mu'), A\big) = \int_A k(\theta, \mu|\theta', \mu')\, d(\theta, \mu) > 0$$

since k is strictly positive on $\mathcal{X}$. Thus, the probability of moving from *any* point in the state space to *any* set with positive π-measure *in one step* is positive. We conclude that this Gibbs Markov chain is π-irreducible and *aperiodic*. (See Tierney (1994, p. 1711) for a general definition of aperiodicity.) We will return to this example later.

‖

Under simple regularity conditions, a Markov chain will "converge" to its invariant distribution no matter how it is started. We will say that Φ satisfies assumption $(\mathcal{A})$ if it:

(i) possesses an invariant density (or probability measure), π;
(ii) is π-irreducible;
(iii) is aperiodic; and
(iv) is Harris recurrent.

Harris recurrence (Meyn and Tweedie 1993, Chapter 9) is a technical condition that is usually easy to verify when (i), (ii), and (iii) are satisfied. Specific results concerning the Harris recurrence of Gibbs samplers and Metropolis–Hastings chains can be found in Tierney (1994, Corollaries 1 and 2). From a practical point of view, assumption $(\mathcal{A})$ implies that the starting value is irrelevant and that the chain will thoroughly explore the state space as the number of iterations grows large.

Under assumption $(\mathcal{A})$, for every $\boldsymbol{x} \in \mathcal{X}$ we have

$$\|P^n(\boldsymbol{x}, \cdot) - \pi(\cdot)\| \downarrow 0 \quad \text{as} \quad n \to \infty, \tag{9}$$

where

$$\|P^n(\boldsymbol{x}, \cdot) - \pi(\cdot)\| := \sup_{A \in \mathcal{B}} |P^n(\boldsymbol{x}, A) - \pi(A)|$$

is the *total variation* distance between the probability measures $P^n(\boldsymbol{x}, \cdot)$ and $\pi(\cdot)$. (When the probability measures have densities, the total variation distance can be expressed as one half the integrated absolute difference between the densities.) In words, (9) says that no matter what the starting value, the random variables in the Markov chain look more and more like a random variable from π as n gets large. (See Rosenthal (2001) for an accessible proof of this result.)

Assumption $(\mathcal{A})$ also guarantees that the Ergodic Theorem holds. Specifically, if $E_\pi|h| := \int |h(\boldsymbol{x})|\, \pi(d\boldsymbol{x}) < \infty$, then for any starting value

$$\frac{1}{n} \sum_{i=0}^{n-1} h(\boldsymbol{X}_i) \to E_\pi h \quad \text{as } n \to \infty$$

with probability 1. Thus, letting $B \in \{0, 1, \dots\}$ denote the burn-in,

$$\bar{h}_{n,B} := \frac{1}{n} \sum_{i=B}^{B+n-1} h(\boldsymbol{X}_i) \tag{10}$$

is a strongly consistent estimator of $E_\pi h$. In this notation, our original two questions become: (Q1) How large should we take B? and (Q2) How large an n is required?

It is important to recognize that burn-in is not strictly necessary; that is, using $B = 0$ in (10) still results in a strongly consistent estimator of $E_\pi h$. However, most variance estimation techniques (e.g. batch means and spectral analysis) are more effective when the Markov chain is stationary (Bratley et al. 1987, p. 94). Hence, if one of these techniques is to be employed, then it may be necessary to use a non-zero burn-in. In contrast, the regenerative simulation method of estimating variance (discussed in Section 5) does not require a stationary chain. In fact, $\boldsymbol{X}_0$ is drawn from a prescribed distribution *not equal to* π! (For more on the burn-in debate, see C. J. Geyer's web page at `http://www.stat.umn.edu/~charlie/`.)

2.2 Geometric convergence to π and its connection to (Q1) and (Q2)

Assumption $(\mathcal{A})$ gets us the convergence in (9), but does not tell us anything about the *rate* of convergence. In fact, rigorous answers to (Q1) and (Q2) can be developed if it can be established that the convergence in (9) takes place at a geometric rate. More specifically, the Markov chain Φ satisfying assumption $(\mathcal{A})$ is said to be *geometrically ergodic* if there exists a constant $0 < t < 1$ and a function $M : \mathcal{X} \mapsto \mathbb{R}^+$ such that for any $\boldsymbol{x} \in \mathcal{X}$,

$$\|P^n(\boldsymbol{x}, \cdot) - \pi(\cdot)\| \leq M(\boldsymbol{x})\, t^n \tag{11}$$

for $n = 1, 2, \ldots$. (If there exists a bounded M satisfying (11), then Φ is called *uniformly ergodic*, and if $\mathcal{X}$ has a finite number of elements, then M is necessarily bounded.)

Obviously, if (11) holds and we have (or can bound) $M(\cdot)$ and t, then for a given starting value, $\boldsymbol{X}_0 = \boldsymbol{x}$, we can calculate *exactly* how many iterations are necessary to get the total variation distance below some prespecified value. As we will see later, establishing a drift condition and an associated minorization condition allows us to use the results of Rosenthal (1995) or Roberts and Tweedie (1999) to form an upper bound on the right-hand side of (11). This takes care of (Q1).

We should point out that several techniques for bounding the right-hand side of (11) have been developed specifically for cases where $\mathcal{X}$ is finite (but very large) (see e.g. Diaconis and Stroock 1991). Applications of such techniques in MCMC contexts include Frigessi, di Stefano, Hwang and Sheu (1993) and Ingrassia (1994). Unfortunately, these methods are not directly applicable to chains on general state spaces (but see Yuen 2000).

We know that for any fixed $B \in \{0, 1, \ldots\}$, $\bar{h}_{n,B}$ is a strongly consistent estimator of $E_\pi h$. We now seek a reliable measure of its accuracy. Suppose that the following CLT holds

$$\sqrt{n}\left(\bar{h}_{n,B} - E_\pi h\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma_h^2\right) . \tag{12}$$

Then given an estimate of σ_h^2, we could get an asymptotic standard error for $\bar{h}_{n,B}$. Indeed, Chan and Geyer (1994) show that if Φ satisfies assumption $(\mathcal{A})$, is geometrically ergodic, and $E_\pi|h|^{2+\epsilon} < \infty$ for some $\epsilon > 0$, then (12) holds with

$$\sigma_h^2 = \mathrm{Var}_\pi(h(\boldsymbol{X}_0)) + 2\sum_{i=1}^{\infty} \mathrm{Cov}_\pi(h(\boldsymbol{X}_0), h(\boldsymbol{X}_i)).$$

The subscript "π" means that the variance and covariances are calculated under stationarity; i.e., assuming that $\boldsymbol{X}_0 \sim \pi$.

Mykland et al. (1995) and Robert (1995) show that when the CLT holds it is possible to obtain a consistent estimate of σ_h^2 by uncovering *regeneration times*; i.e., times at which the Markov chain stochastically restarts itself. This technique is called regenerative simulation (RS) and is closely related to the regenerative method that has been discussed extensively in the operations research literature (see e.g. Glynn and Iglehart 1987). In contrast to other variance estimation methods (see e.g. Geyer 1992), RS does not require Φ to be stationary or reversible. (While standard, fixed scan Gibbs samplers like the ones we analyze in this paper are not reversible, some flavors of the Gibbs sampler are reversible (Besag, Green, Higdon and Mengersen 1995).) The use of RS for the purpose of constructing Monte Carlo standard errors is described in Section 5 and illustrated in Subsection 6.4.

Geometric ergodicity is not necessary for CLTs (see e.g. Jarner and Roberts 2001), but a CLT may fail to hold even in very simple applications of subgeometric MCMC. For example, Roberts (1999) shows that for the independence Metropolis algorithm of Example 1, a CLT (for all functions h that are bounded away from zero at ∞) will *not* hold if $\theta > 2$. In the next section, we define drift and minorization and describe how they can be used to establish (11), and to formally address (Q1) and (Q2).

3 Geometric Ergodicity via Drift and Minorization

This section is broken up into three subsections. In Subsection 3.1, we define drift and minorization and illustrate the required calculations with our toy Gibbs sampler. Subsection 3.2 begins with a description of how a minorization condition can be used to *split* the Markov transition density into a mixture of two densities. This is an important concept for understanding both convergence rate bounds and regenerative simulation. Subsection 3.2 also contains a recipe for using this mixture representation to construct two copies of Φ that eventually *couple*; i.e., become the same chain. This construction leads to the *coupling inequality* which is the key result for deriving convergence

rate bounds. In Subsection 3.3, the coupling inequality is used to show how drift and minorization together imply that the Markov chain converges at a geometric rate.

3.1 Definitions and examples

Throughout this section we assume that the Markov chain Φ satisfies assumption $(\mathcal{A})$. We say a *drift condition* holds if for some function $V : \mathcal{X} \mapsto \mathbb{R}^+$, some $0 < \lambda < 1$, and some $b < \infty$

$$E\left[V(\boldsymbol{X}_{i+1})|\boldsymbol{X}_i = \boldsymbol{x}\right] \leq \lambda V(\boldsymbol{x}) + b \qquad \forall\ \boldsymbol{x} \in \mathcal{X}. \tag{13}$$

Note that this expectation is with respect to the Markov transition kernel and *not* π. It is useful to think of V as a potential energy surface. When (13) holds, the chain tends to "drift" towards states of lower energy in expectation. In this context, V is called an *energy function*. Here is an example of establishing (13).

EXAMPLE 2 CONTINUED. Assume that $m \geq 5$. We shall establish a drift condition using the function $V(\mu, \theta) = (\mu - \bar{y})^2$. The form of the Markov transition density (7) implies that (i) given μ', (μ, θ) is conditionally independent of θ'; and (ii) given θ, μ is conditionally independent of μ'. It follows that

$$E[V(\mu,\theta)|\theta', \mu'] = E[V(\mu,\theta)|\mu'] = E\big\{E[V(\mu,\theta)|\theta\,]|\mu'\big\}\,.$$

Since $\mu|\theta, \boldsymbol{y} \sim \mathrm{N}(\bar{y}, \theta/m)$, the innermost expectation yields

$$E[V(\mu,\theta)|\theta] = E[(\mu-\bar{y})^2|\theta] = \mathrm{Var}(\mu|\theta) = \frac{\theta}{m}\,.$$

Similarly,

$$E(\theta|\mu') = \frac{s^2 + m(\mu' - \bar{y})^2}{m-3}\,.$$

Therefore,

$$\begin{aligned} E[V(\mu,\theta)|\theta', \mu'] &= \frac{1}{m-3}(\mu' - \bar{y})^2 + \frac{s^2}{m(m-3)} \\ &\leq \lambda\, V(\mu', \theta') + b \end{aligned}$$

for $b = s^2/[m(m-3)]$ and any $\lambda \geq \frac{1}{m-3}$. This establishes (13) since $m \geq 5$.

$\|$

A *minorization condition* holds if for some probability measure Q on $\mathcal{B}$, some set C for which $\pi(C) > 0$, and some $\varepsilon > 0$

$$P(\boldsymbol{x}, A) \geq \varepsilon Q(A) \quad \forall\ \boldsymbol{x} \in C,\ A \in \mathcal{B}. \tag{14}$$

The set C is called a *small set*. Note that plugging $\mathcal{X}$ into (14) shows that $\varepsilon \leq 1$.

One way to verify that Φ is geometrically ergodic is to show that Φ satisfies both a drift condition and an *associated* minorization condition. Specifically, the chain is geometrically ergodic if it satisfies (13) *and* (14) with $C = \{\boldsymbol{x} \in \mathcal{X} : V(\boldsymbol{x}) \leq d\}$ and any d larger than $2b/(1-\lambda)$ (Rosenthal 1995). We now demonstrate how to establish a minorization condition, again using our toy Gibbs sampler.

EXAMPLE 2 CONTINUED. We use a technique that is based on Rosenthal's (1995) Lemma 6b. Let $C = \{(\theta, \mu) : (\mu - \overline{y})^2 < d\}$ where $d > 0$. Suppose that we can find a density $q(\theta, \mu)$ on $\mathcal{X} = \mathbb{R}^+ \times \mathbb{R}$ and an $\varepsilon > 0$ such that whenever $(\theta', \mu') \in C$

$$f(\theta|\mu', \boldsymbol{y}) f(\mu|\theta, \boldsymbol{y}) \geq \varepsilon q(\theta, \mu) \quad \forall\ (\theta, \mu) \in \mathcal{X}\ . \tag{15}$$

Let $Q(\cdot)$ be the probability measure associated with the density q. Then for any set A and any $(\theta', \mu') \in C$ we have

$$P\big((\theta', \mu'), A\big) = \int_A f(\theta|\mu', \boldsymbol{y}) f(\mu|\theta, \boldsymbol{y})\ d(\theta, \mu) \geq \varepsilon \int_A q(\mu, \theta)\ d(\theta, \mu) = \varepsilon Q(A),$$

and hence (14) is established. We now construct a $q(\theta, \mu)$ and an $\varepsilon > 0$ that satisfy (15).

Let $C_\mu = \{\mu : (\mu - \overline{y})^2 < d\}$ and note that for any $\mu' \in C_\mu$ we have

$$f(\theta|\mu', \boldsymbol{y}) f(\mu|\theta, \boldsymbol{y}) \geq f(\mu|\theta, \boldsymbol{y}) \inf_{\mu \in C_\mu} f(\theta|\mu, \boldsymbol{y}).$$

Recall that $f(\theta|\mu, \boldsymbol{y})$ is just an IG density. In fact, $g(\theta) := \inf_{\mu \in C_\mu} f(\theta|\mu, \boldsymbol{y})$ can be written in closed form. (The infimum does actually depend on the data and this is being suppressed in the notation.) Let $\mathrm{IG}(\alpha, \beta; w)$ denote the value of the $\mathrm{IG}(\alpha, \beta)$ density evaluated at the point $w > 0$.

A calculation similar to one done in Rosenthal (1996) yields

$$g(\theta) = \inf_{\mu \in C_\mu} \mathrm{IG}\left(\frac{m-1}{2}, \frac{s^2}{2} + \frac{m}{2}(\mu - \overline{y})^2; \theta\right) = \begin{cases} \mathrm{IG}\left(\frac{m-1}{2}, \frac{s^2}{2} + \frac{md}{2}; \theta\right) & \theta < \theta^* \\ \mathrm{IG}\left(\frac{m-1}{2}, \frac{s^2}{2}; \theta\right) & \theta \geq \theta^* \end{cases}$$

where $\theta^* = md\left[(m-1)\log(1 + md/s^2)\right]^{-1}$. (See Jones and Hobert (2001) for more details about this.) Figure 2 shows $g(\theta)$ for the case where $m = 5$, $s^2 = 10$, and $d = 22/5$. Now put

$$\varepsilon = \int_{\mathbb{R}^+} \int_{\mathbb{R}} g(\theta) f(\mu|\theta, \boldsymbol{y})\, d\mu\, d\theta = \int_{\mathbb{R}^+} g(\theta)\, d\theta.$$

Then (15) is satisfied with this ε and the density $q(\theta, \mu) = \varepsilon^{-1} g(\theta) f(\mu|\theta, \boldsymbol{y})$. Note that ε can be calculated with two evaluations of the incomplete gamma function. Since our minorization holds for any $d > 0$, we may conclude that this Gibbs sampler is geometrically ergodic as long as $m \geq 5$. We will return to this example in Section 4.

||

Of course, (13) and (14) may be difficult (if not impossible) to establish in realistic settings. Indeed, there is no guarantee that convergence occurs at a geometric rate even in simple applications of MCMC (recall Example 1). Examples of the use of drift and/or minorization for analyzing MCMC algorithms include Robert (1995), Rosenthal (1995, 1996), Roberts and Rosenthal (1998b), Hobert and Geyer (1998), Jones and Hobert (2001), and Hobert (2001) who considered Gibbs samplers; Meyn and Tweedie (1994), Mengersen and Tweedie (1996), Roberts and Tweedie (1996), and Jarner and Hansen (2000) who worked on Metropolis-Hastings algorithms; and Roberts and Rosenthal (1999) and Mira and Tierney (2001) who examined *slice sampler* Markov chains. In the remainder of this section, we explain the theoretical connection between geometric ergodicity and drift and minorization.

3.2 Minorization and coupling

The concepts of *splitting* and *coupling* are explained in this subsection. We begin with splitting. Recall that $P(\boldsymbol{x}, A) = \int_A k(\boldsymbol{u}|\boldsymbol{x})\, d\boldsymbol{u}$. For the time being, we require only a minorization condition. Hence, assume that we have established (14) by finding a set C, a density $q(\cdot)$ on $\mathcal{X}$, and an $\varepsilon > 0$

such that whenever $\boldsymbol{x} \in C$,

$$k(\boldsymbol{u}|\boldsymbol{x}) \geq \varepsilon q(\boldsymbol{u}) \quad \forall\ \boldsymbol{u} \in \mathcal{X}. \tag{16}$$

Note that for each fixed $\boldsymbol{x} \in C$,

$$r(\boldsymbol{u}|\boldsymbol{x}) := \frac{k(\boldsymbol{u}|\boldsymbol{x}) - \varepsilon q(\boldsymbol{u})}{1-\varepsilon}$$

is a density in $\boldsymbol{u}$ and is called the *residual density*. It follows that, whenever $\boldsymbol{x} \in C$, we can *split* $k(\boldsymbol{u}|\boldsymbol{x})$ into a mixture of two densities as follows

$$k(\boldsymbol{u}|\boldsymbol{x}) = \varepsilon\, q(\boldsymbol{u}) + (1-\varepsilon)\, r(\boldsymbol{u}|\boldsymbol{x})\ . \tag{17}$$

Whenever $\boldsymbol{X}_i \in C$, (17) can be used to generate $\boldsymbol{X}_{i+1}$ sequentially as follows. Given $\boldsymbol{X}_i \in C$, generate $\delta_i \sim$ Bernoulli(ε). If $\delta_i = 1$, then draw $\boldsymbol{X}_{i+1} \sim q(\cdot)$, else draw $\boldsymbol{X}_{i+1} \sim r(\cdot|\boldsymbol{X}_i)$.

This mixture representation of $k(\boldsymbol{u}|\boldsymbol{x})$ allows for the (joint) construction of two Markov chains, $\Phi_x = \{\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \dots\}$ and $\Phi_y = \{\boldsymbol{Y}_0, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots\}$, that, marginally, are identical copies of Φ, but which are not independent. In fact, Φ_x and Φ_y are constructed in such a way that they eventually *become the same Markov chain*; that is, they eventually *couple*. What makes this possible is the fact that the density q in (17) does not depend on $\boldsymbol{x}$. Here are the details.

Let $\boldsymbol{X}_0 = \boldsymbol{x}_0$ be an arbitrary, fixed starting value and draw $\boldsymbol{Y}_0$ from the invariant probability distribution; i.e., $\boldsymbol{Y}_0 \sim \pi$. (Note that we will not actually have to simulate from π). The construction involves two different methods of simulating $(\boldsymbol{X}_{i+1}, \boldsymbol{Y}_{i+1})$ conditional on $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, and which of the two is used depends on whether or not $(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in C \times C$. First, if $(\boldsymbol{X}_i, \boldsymbol{Y}_i) \notin C \times C$, then we draw $\boldsymbol{X}_{i+1} \sim k(\cdot|\boldsymbol{X}_i)$ and *independently* draw $\boldsymbol{Y}_{i+1} \sim k(\cdot|\boldsymbol{Y}_i)$. If, on the other hand, $(\boldsymbol{X}_i, \boldsymbol{Y}_i) \in C \times C$, then we use (17) as follows. We draw $\delta_i \sim$ Bernoulli(ε). If $\delta_i = 0$, then we draw $\boldsymbol{X}_{i+1} \sim r(\cdot|\boldsymbol{X}_i)$ and *independently* draw $\boldsymbol{Y}_{i+1} \sim r(\cdot|\boldsymbol{Y}_i)$. But if $\delta_i = 1$, then we draw $\boldsymbol{X}_{i+1} = \boldsymbol{Y}_{i+1} \sim q(\cdot)$ *and* all future draws are made in such a way that the two chains remain equal.

The *coupling time*, T, is defined to be the (random) time at which coupling occurs; that is, the time at which the two chains become the same. The key result that is used to bound convergence rates of (general state space) Markov chains is the so-called *coupling inequality* which is now

derived. Recall that the total variation distance between the probability measures $P^n(\boldsymbol{x}_0, \cdot)$ and $\pi(\cdot)$ is defined as

$$\|P^n(\boldsymbol{x}_0, \cdot) - \pi(\cdot)\| := \sup_{A \in \mathcal{B}} |P^n(\boldsymbol{x}_0, A) - \pi(A)|.$$

Consider the right hand side and note that, using the construction above, we have

$$\begin{aligned}
| P^n(\boldsymbol{x}_0, A) - \pi(A) | &= | \Pr(\boldsymbol{X}_n \in A) - \Pr(\boldsymbol{Y}_n \in A) | \\
&= | \Pr(\boldsymbol{X}_n \in A, \boldsymbol{X}_n = \boldsymbol{Y}_n) + \Pr(\boldsymbol{X}_n \in A, \boldsymbol{X}_n \neq \boldsymbol{Y}_n) \\
&\qquad - \Pr(\boldsymbol{Y}_n \in A, \boldsymbol{X}_n = \boldsymbol{Y}_n) - \Pr(\boldsymbol{Y}_n \in A, \boldsymbol{X}_n \neq \boldsymbol{Y}_n) | \\
&= | \Pr(\boldsymbol{X}_n \in A, \boldsymbol{X}_n \neq \boldsymbol{Y}_n) - \Pr(\boldsymbol{Y}_n \in A, \boldsymbol{X}_n \neq \boldsymbol{Y}_n) | \\
&\leq \max \{\Pr(\boldsymbol{X}_n \in A, \boldsymbol{X}_n \neq \boldsymbol{Y}_n), \Pr(\boldsymbol{Y}_n \in A, \boldsymbol{X}_n \neq \boldsymbol{Y}_n)\} \\
&\leq \Pr(\boldsymbol{X}_n \neq \boldsymbol{Y}_n) \\
&\leq \Pr(T > n) \, .
\end{aligned}$$

Hence, the coupling inequality:

$$\|P^n(\boldsymbol{x}_0, \cdot) - \pi(\cdot)\| \leq \Pr(T > n).$$

In the case where the entire state space is small; i.e., when $C = \mathcal{X}$, the coupling inequality immediately yields a bound on the total variation distance to stationarity. To see this, note that when $C = \mathcal{X}$, $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ is *always* in $C \times C$, which means that a Bernoulli(ε) is drawn at every step. Thus, $T \sim$ Geometric(ε); that is, $\Pr(T = n) = \varepsilon(1-\varepsilon)^{n-1}$ for $n \in \mathbb{N}$. It follows that $\Pr(T > n) = (1-\varepsilon)^n$, and hence we have the following result.

Theorem 1. *(Meyn and Tweedie 1993, p392) Suppose the Markov chain Φ satisfies assumption (A) as well as the minorization condition (14) with $C = \mathcal{X}$. Then*

$$\|P^n(\boldsymbol{x}_0, \cdot) - \pi(\cdot)\| \leq (1-\varepsilon)^n.$$

Observe that the bound does not depend on the starting value, $\boldsymbol{x}_0$, and hence Φ is uniformly ergodic. Uniform ergodicity is not common in MCMC, but there are a few instances (Robert and Casella 1999, Chapter 9). One is the independence Metropolis algorithm with fat tail proposals which we now discuss.

EXAMPLE 1 CONTINUED. Consider a general independence Metropolis sampler with target density $\pi(\boldsymbol{x})$ and proposal density $p(\boldsymbol{x})$. Suppose that both of these densities are continuous and strictly positive on $\mathcal{X}$. In this case, the Markov transition kernel, P, does not have a density with respect to Lebesgue measure (Tierney 1998), but the chain does satisfy assumption $(\mathcal{A})$. Mengersen and Tweedie (1996) show that if there exists a $\kappa > 0$ such that

$$\frac{\pi(\boldsymbol{x})}{p(\boldsymbol{x})} \leq \kappa \quad \forall\ \boldsymbol{x} \in \mathcal{X}\,, \tag{18}$$

then (14) holds with $C = \mathcal{X}$ and $\varepsilon = \kappa^{-1}$. Thus, by Theorem 1 the Markov chain is uniformly ergodic and

$$\|P^n(\boldsymbol{x}, \cdot) - \pi(\cdot)\| \leq \left(1 - \frac{1}{\kappa}\right)^n .$$

Mengersen and Tweedie (1996) also show that if for every $\kappa > 0$ there is a set of positive measure where (18) fails to hold, then the chain converges at a subgeometric rate. Thus, there is an "all or nothing" aspect to the independence sampler. For the independence sampler of Example 1 we have $\pi(\boldsymbol{x})/p(\boldsymbol{x}) = \theta^{-1}\exp\{x(\theta - 1)\}$. Consequently, when $\theta \in (0, 1)$ the chain is uniformly ergodic and if $\theta > 1$ it is subgeometric.

Note that the existence of κ satisfying (18) is *exactly* what is required to implement rejection sampling with proposal density p (Robert and Casella 1999, p.49). However, unlike the rejection sampler, the independence sampler can be implemented without knowing the value of κ. (See Tierney (1994) and Caffo, Booth and Davison (2001) for more on this.)

||

Unfortunately, even when the hypotheses of Theorem 1 hold, it is often the case that the value of ε is too small for Theorem 1 to be of any practical value. Indeed, there is typically a trade-off between the size of the small set and the magnitude of ε. When C is a proper subset of $\mathcal{X}$, the distribution of T is quite complicated. This case is addressed in the next subsection.

3.3 Connecting drift and minorization to geometric convergence

We now explain how drift and minorization can be used to establish geometric convergence to the invariant distribution when the set C in (14) is not the entire state space. We do not intend

this argument to be rigorous. We are striving only to convey the nature of the connection. For a completely rigorous approach, the reader should consult Lindvall (1992), Meyn and Tweedie (1993), Rosenthal (1995), and Roberts and Tweedie (1999).

We have seen that if the whole space is small, then the number of steps until we successfully couple has a geometric distribution. Suppose now that (14) holds for some set C that is a proper subset of $\mathcal{X}$. Then each time we reach C, we can draw a Bernoulli(ε) and if we get a "success" we couple and we can apply the coupling inequality.

Thus, the next step is to consider how long it takes between visits to C. Let τ_C denote the (random) number of steps it takes the chain to return to the set C; that is, $\tau_C = \min\{n \geq 1 : \boldsymbol{X}_n \in C\}$. Obviously, the distribution of τ_C will depend on the starting value, $\boldsymbol{x}_0$. Suppose we could show that, for every $\boldsymbol{x}_0 \in C$, τ_C has a moment generating function. It would then follow that the time to a successful coupling, T, is a geometric sum of (random) excursion times each of which has a "thin tailed" distribution; so overall T itself would have a thin tail and hence a moment generating function (Roberts and Tweedie 1999, Theorem 2.1). Thus, there would exist a $\beta > 1$ such that $E(\beta^T) < \infty$ and from the coupling inequality we would have

$$\beta^n \|P^n(\boldsymbol{x}, \cdot) - \pi(\cdot)\| \leq \beta^n \Pr(T > n) \leq E[\beta^T I(T > n)] \to 0 \tag{19}$$

as $n \to \infty$ by dominated convergence. Therefore, we would be able to conclude that $\|P^n(\boldsymbol{x}, \cdot) - \pi(\cdot)\| = o(\beta^{-n})$; that is, the convergence to stationarity occurs at a geometric rate. (Some might refer to this as *exponential* convergence, but see Lindvall (1992, p.30).)

The role of the drift condition is to ensure that the return time, τ_C, has the required tail behavior. To see this, we need to introduce a second sufficient condition for geometric convergence that involves a slight variation on (13). Specifically, suppose that for some function $V : \mathcal{X} \mapsto [1, \infty)$, some $0 < \lambda < 1$, and some $b < \infty$ the Markov chain Φ satisfies

$$E[V(\boldsymbol{X}_{i+1}) | \boldsymbol{X}_i = \boldsymbol{x}] \leq \lambda V(\boldsymbol{x}) + b I_C(\boldsymbol{x}) \qquad \forall\ \boldsymbol{x} \in \mathcal{X} \tag{20}$$

where $C = \{\boldsymbol{x} \in \mathcal{X} : V(\boldsymbol{x}) \leq d\}$ and d is any number larger than $\frac{b}{2(1-\lambda)} - 1$. If (20) holds and a minorization condition is satisfied on C, then Φ is geometrically ergodic (Roberts and Tweedie

1999). Suppose now that (20) holds and note that we may rewrite it as

$$\Delta V(\boldsymbol{x}) := E[V(\boldsymbol{X}_{i+1})|\boldsymbol{X}_i = \boldsymbol{x}] - V(\boldsymbol{x}) \leq -(1-\lambda)V(\boldsymbol{x}) + bI_C(\boldsymbol{x}).$$

Therefore, the expected change in the value of V is negative when $\boldsymbol{x}$ is not in C; and indeed, not only does $V(\boldsymbol{x})$ "drift" in to the set where V is small in this sense, but it does so in such a way that from points $\boldsymbol{x}$ with larger V values, the drift is faster. Conceptually, this suggests that once the chain leaves C it should return quickly.

Mathematically, it turns out that V gives the required tail behavior in a very explicit sense. Indeed, it can be shown that (20) implies that for any $\boldsymbol{x}_0 \in C$

$$E\left[\lambda^{-\tau_C}\right] \leq d + \frac{b}{\lambda}$$

(Meyn and Tweedie 1994, Lund and Tweedie 1996). So overall, drift covers the tail behavior of the return times to C and minorization ensures that only a geometric number of those times are needed; and together they lead to the existence of a $\beta > 1$ satisfying (19). In the next section, we state a theorem of Rosenthal (1995) that gives explicit upper bounds on the distance to stationarity in terms of the drift and minorization conditions.

4 How much burn–in?

Suppose that the Markov chain Φ satisfies assumption $(\mathcal{A})$. Here is a slightly simplified version of Rosenthal's (1995) result:

Theorem 2. *(Rosenthal 1995) Suppose that Φ satisfies the drift condition (13). Further suppose that Φ satisfies the minorization condition (14) on $C = \{\boldsymbol{x} : V(\boldsymbol{x}) \leq d\}$ where d is any number larger than $2b/(1-\lambda)$. Let $\boldsymbol{X}_0 = \boldsymbol{x}_0$ and define two constants as follows*

$$\alpha = \frac{1+d}{1+2b+\lambda d} \qquad \text{and} \qquad U = 1 + 2(\lambda d + b)\,.$$

Then for any $0 < r < 1$

$$\|P^n(\boldsymbol{x}_0, \cdot) - \pi(\cdot)\| \leq (1-\varepsilon)^{rn} + \left(\frac{U^r}{\alpha^{1-r}}\right)^n \left(1 + \frac{b}{1-\lambda} + V(\boldsymbol{x}_0)\right)\,.$$

In applying this result, the user must specify the values of d and r. It makes sense to do so in such a way that $\frac{U^r}{\alpha^{1-r}} < 1$, otherwise the bound may not decrease in n. Furthermore, in our experience, slight changes in d and r can lead to wildly different results, so it pays to experiment. We use Theorem 2 in a realistic application in Section 6. Here is a simpler example of its use.

EXAMPLE 2 CONTINUED. Suppose that $m = 5$ and $s^2 = 10$. Then the drift condition established in Subsection 3.1 holds with $b = 1$ and any $\lambda \geq 1/2$. If we take $\lambda = 1/2$, then d can be any number larger than $2b/(1-\lambda) = 4$. If we take $d = 6$, then $\varepsilon \approx 0.35$. Then taking $r = 0.05$ and starting the chain with $\mu_0 = \bar{y}$, we have

$$\|P^n((\theta_0, \mu_0), \cdot) - \pi(\cdot)\| \leq (0.9785)^n + 3\,(0.9641)^n \ .$$

Hence, after 220 iterations, the total variation distance is less than 0.01. (Of course, the total variation distance could be less than 0.01 much sooner - for an extreme example of this, see the Rejoinder of van Dyk and Meng (2001).)

$\|$

An alternative to Rosenthal's bound is given in Roberts and Tweedie (1999, 2001). These authors prove that their bound is better than Rosenthal's as the number of iterations tends to infinity, but some of their examples show little practical difference between the two bounds. The main difference between the hypotheses of Theorem 2 and those of Roberts and Tweedie is the form of the drift condition. Specifically, Theorem 2 requires (13) while Roberts and Tweedie require (20). In our experience, (13) is easier to establish than (20). There is a "conversion" formula that can be used to construct a drift condition of the form (20) given one of the form (13) (Jones and Hobert 2001). Unfortunately, it appears that whatever one would gain by using the Roberts and Tweedie result rather Theorem 2 is negated somewhere in the conversion. This is the reason that we do not apply the Roberts and Tweedie result in this paper. See Jones and Hobert (2001) for more on this. In the next section, a rigorous answer to (Q2) is formulated.

5 Monte Carlo standard errors

In this section, we discuss the use of regenerative simulation (RS) for calculating standard errors of ergodic averages. Basically, the regenerative method involves breaking simulation output up into iid pieces that can be analyzed using standard results for iid data. A complete development of this subject can be found in Ripley (1987, Chapter 6) or Bratley et al. (1987, Chapter 3). Mykland et al. (1995), Robert (1995), and Robert and Casella (1999, Chapter 8) discuss RS in the MCMC context. Regenerative methods of analyzing simulation-based output have a rich history in the operations research literature (see e.g. Crane and Iglehart, 1975; Glynn, 1985; and Glynn and Iglehart, 1987, 1993).

This section consists of three subsections. A generalization of (14) is introduced in Subsection 5.1. This more general minorization condition can be used in the same manner as (14) to represent the Markov transition density as a mixture of two densities. It is this mixture that is used to break the output up into iid pieces. The details are given in Subsection 5.2. In Subsection 5.3, we explain exactly how RS is used to calculate Monte Carlo standard errors. The method of *batch means*, which is an alternative to RS, is also briefly discussed.

5.1 A more general minorization condition

A generalization of (14) is as follows. For some function $s : \mathcal{X} \mapsto \mathbb{R}^+$ such that $E_\pi s > 0$ and some probability measure Q on $\mathcal{B}$

$$P(\boldsymbol{x}, A) \geq s(\boldsymbol{x})\, Q(A) \quad \forall\ \boldsymbol{x} \in \mathcal{X},\ A \in \mathcal{B}. \tag{21}$$

Clearly, (14) is the special case of (21) where $s(\boldsymbol{x}) = \varepsilon I(\boldsymbol{x} \in C)$. Note that plugging $\mathcal{X}$ into (21) shows that $s(\boldsymbol{x}) \leq 1$. Mykland et al. (1995) show that it is often easy to establish (21) for Gibbs samplers and Metropolis–Hastings algorithms. (Our reasons for introducing this generalization will be spelled out later in this section.)

Recall that $P(\boldsymbol{x}, A) = \int_A k(\boldsymbol{u}|\boldsymbol{x})\, d\boldsymbol{u}$. We can establish (21) by finding a non-negative function

$s(\cdot)$ and a density $q(\cdot)$ on $\mathcal{X}$ such that

$$k(\boldsymbol{u}|\boldsymbol{x}) \geq s(\boldsymbol{x})\, q(\boldsymbol{u}) \quad \forall\ \boldsymbol{x}, \boldsymbol{u} \in \mathcal{X}\ . \tag{22}$$

We now establish (22) for our toy Gibbs sampler using a technique described in Mykland et al. (1995).

EXAMPLE 2 CONTINUED. We will construct a density $q(\theta, \mu)$ on $\mathcal{X} = \mathbb{R}^+ \times \mathbb{R}$ and a function $s(\theta', \mu')$ such that

$$k(\theta, \mu|\theta', \mu') \geq s(\theta', \mu')\, q(\theta, \mu)$$

for all $(\theta, \mu), (\theta', \mu') \in \mathcal{X}$. To this end, let $(\tilde{\theta}, \tilde{\mu})$ be a "distinguished point" in $\mathcal{X}$ and let D be a set in $\mathcal{X}$. Note that

$$\begin{aligned} k(\theta, \mu|\theta', \mu') &= f(\theta|\mu', \boldsymbol{y}) f(\mu|\theta, \boldsymbol{y}) \\ &= \left[\frac{f(\theta|\mu', \boldsymbol{y})}{f(\theta|\tilde{\mu}, \boldsymbol{y})}\right] f(\theta|\tilde{\mu}, \boldsymbol{y})\, f(\mu|\theta, \boldsymbol{y}) \\ &\geq \left[\inf_{(\theta,\mu)\in D} \frac{f(\theta|\mu', \boldsymbol{y})}{f(\theta|\tilde{\mu}, \boldsymbol{y})}\right] f(\theta|\tilde{\mu}, \boldsymbol{y})\, f(\mu|\theta, \boldsymbol{y})\, I\left[(\theta, \mu) \in D\right] \end{aligned}$$

for all $(\theta, \mu), (\theta', \mu') \in \mathcal{X}$. Let

$$\varepsilon = \int_D f(\theta|\tilde{\mu}, \boldsymbol{y})\, f(\mu|\theta, \boldsymbol{y})\, d(\theta, \mu)\ .$$

Now simply take $q(\theta, \mu) = \varepsilon^{-1} f(\theta|\tilde{\mu}, \boldsymbol{y}) f(\mu|\theta, \boldsymbol{y})\, I\left[(\theta, \mu) \in D\right]$ and take

$$s(\theta', \mu') = \varepsilon \inf_{(\theta,\mu)\in D} \frac{f(\theta|\mu', \boldsymbol{y})}{f(\theta|\tilde{\mu}, \boldsymbol{y})}\ .$$

As a specific example, take the distinguished point to be $(\tilde{\theta}, \tilde{\mu}) = (1, \overline{y})$ and $D = [d_1, d_2] \times \mathbb{R}$ where $0 < d_1 < d_2 < \infty$. Then

$$\inf_{(\theta,\mu)\in D} \frac{f(\theta|\mu', \boldsymbol{y})}{f(\theta|\tilde{\mu}, \boldsymbol{y})} = \left[1 + \frac{m(\overline{y} - \mu')^2}{s^2}\right]^{\frac{m-1}{2}} \exp\left\{-\frac{m(\overline{y} - \mu')^2}{2d_1}\right\}$$

and calculating ε again boils down to evaluating the incomplete gamma function. In practice, the distinguished point is often set at a preliminary estimate of the mean of the stationary distribution and D is centered about that point; see Subsection 6.4.

‖

5.2 The split chain

The analogue of (17) for our new minorization condition is

$$k(\boldsymbol{u}|\boldsymbol{x}) = s(\boldsymbol{x})\, q(\boldsymbol{u}) + (1 - s(\boldsymbol{x}))\, r(\boldsymbol{u}|\boldsymbol{x}) \tag{23}$$

where the residual density $r(\boldsymbol{u}|\boldsymbol{x})$ is now defined as

$$r(\boldsymbol{u}|\boldsymbol{x}) := \frac{k(\boldsymbol{u}|\boldsymbol{x}) - s(\boldsymbol{x})\, q(\boldsymbol{u})}{1 - s(\boldsymbol{x})} \, .$$

The mixture (23) can be used to generate $\boldsymbol{X}_{i+1}$ sequentially as follows. Given $\boldsymbol{X}_i = \boldsymbol{x}$, generate $\delta_i \sim \text{Bernoulli}(s(\boldsymbol{x}))$. If $\delta_i = 1$, then draw $\boldsymbol{X}_{i+1} \sim q(\cdot)$, else draw $\boldsymbol{X}_{i+1} \sim r(\cdot|\boldsymbol{x})$. What we are actually doing here is simulating the so-called *split chain*

$$\Phi' = \{(\boldsymbol{X}_0, \delta_0), (\boldsymbol{X}_1, \delta_1), (\boldsymbol{X}_2, \delta_2), \dots\} \, ,$$

which has state space $\mathcal{X} \times \{0, 1\}$ (Athreya and Ney, 1978; Nummelin, 1978, 1984). The times at which $\delta_i = 1$ are *regeneration times* when Φ' probabilistically restarts itself. More specifically, suppose we start Φ' with $\boldsymbol{X}_0 \sim q(\cdot)$. Then each time that $\delta_i = 1$, $\boldsymbol{X}_{i+1} \sim q(\cdot)$ and we are, in effect, starting over again. Moreover, the *tours* in between regeneration times are iid. Observe that as we make $s(\boldsymbol{x})$ larger, we expect the average tour *length* to decrease.

In order to use RS to get standard errors, we must be able to simulate Φ'. The most straightforward way to do this is as described above. This is problematic, however, because drawing from $r(\cdot|\boldsymbol{x})$ can be quite difficult in practice (see, e.g., Robert 1995). (Note that (17) was used only for the theoretical argument leading to the coupling inequality, and hence the issue of drawing from r never came up in Subsection 3.2.)

Fortunately, Mykland et al. (1995) provide a simple and clever way of avoiding r altogether. If we write the transition as $\boldsymbol{X}_i \to \delta_i \to \boldsymbol{X}_{i+1}$, we need to generate from $(\delta_i, \boldsymbol{X}_{i+1})|\boldsymbol{X}_i$. Above, we suggested doing this by first drawing from $\delta_i|\boldsymbol{X}_i$ and then drawing from $\boldsymbol{X}_{i+1}|\delta_i, \boldsymbol{X}_i$, which, if $\delta_i = 0$, entails simulation from $r(\cdot|\boldsymbol{X}_i)$. Mykland et al. (1995) note that simulating from the residual density can be avoided by first drawing from $\boldsymbol{X}_{i+1}|\boldsymbol{X}_i$ (in the usual way) and then drawing from $\delta_i|\boldsymbol{X}_i, \boldsymbol{X}_{i+1}$. A straightforward calculation shows that

$$\Pr(\delta_i = 1|\boldsymbol{X}_i, \boldsymbol{X}_{i+1}) = \frac{s(\boldsymbol{X}_i)\, q(\boldsymbol{X}_{i+1})}{k(\boldsymbol{X}_{i+1}|\boldsymbol{X}_i)} \, , \tag{24}$$

which is often easy to calculate.

There is actually another important advantage to drawing from $(\delta_i, \boldsymbol{X}_{i+1})|\boldsymbol{X}_i$ in this way. Mykland et al.'s (1995) method of establishing (22) (for Gibbs samplers) entails first showing that $k(\boldsymbol{u}|\boldsymbol{x}) \geq s'(\boldsymbol{x})q'(\boldsymbol{u})$, where q' is an unnormalized density, and then letting $q = q'/\int q'$ and $s = s' \int q'$. Note, however, that drawing from $\delta_i|\boldsymbol{X}_i, \boldsymbol{X}_{i+1}$ requires only the product $s(\boldsymbol{X}_i)\, q(\boldsymbol{X}_{i+1})$. Consequently, there is no need to calculate the normalizing constant!

It *is* possible to use the original minorization condition (14) for RS. However, in our experience, minorization conditions of the form (21) typically lead to many more regenerations than those of the form (14). This is important because if regenerations happen very infrequently, it may take an inordinate amount of time to observe enough regenerations so that the approximations (described below) are reasonable. We now explain how the ideas in Subsections 5.1 and 5.2 are applied in RS.

5.3 Regenerative simulation

Assume that the Markov chain Φ satisfies assumption $(\mathcal{A})$. We know that $E_\pi|h| < \infty$ implies that $\bar{h}_n := \bar{h}_{n,0} = n^{-1}\sum_{i=0}^{n-1} h(\boldsymbol{X}_i)$ is a strongly consistent estimator of $E_\pi h$ regardless of the starting value. Assume further that Φ is geometrically ergodic and that $E_\pi|h|^{2+\epsilon} < \infty$ for some $\epsilon > 0$, so that we have the following CLT

$$\sqrt{n}(\bar{h}_n - E_\pi h) \xrightarrow{d} \mathrm{N}\left(0, \sigma_h^2\right). \tag{25}$$

Estimation of σ_h^2 is difficult because the $\boldsymbol{X}_i$'s constituting $\bar{h}_n$ are not independent. By using the split chain, we can rewrite $\bar{h}_n$ as a function of iid bivariate random vectors. This trick allows us to use standard techniques from iid theory to calculate a valid Monte Carlo standard error. Here are the details.

Suppose that we have established (22) so we can simulate Φ' using Mykland et al.'s (1995) technique. Let $\tau_0 < \tau_1 < \cdots$ be the (random) regeneration times; i.e., $\tau_{t+1} = \min\{i > \tau_t : \delta_{i-1} = 1\}$. Assume that $\tau_0 = 0$ so the chain is started with a regeneration; that is, $\boldsymbol{X}_0 \sim q(\cdot)$. (Mykland et al. (1995) show that starting with a regeneration is quite easy for standard MCMC algorithms.)

Also assume that Φ is run for a fixed number, R, of tours; that is, the simulation is stopped the Rth time that a $\delta_i = 1$. Thus, the total length of the simulation, N, is random. Let N_t be the length of the tth tour; that is, $N_t = \tau_t - \tau_{t-1}$ and define

$$S_t = \sum_{j=\tau_{t-1}}^{\tau_t - 1} h(\boldsymbol{X}_j)$$

for $t = 1, \ldots, R$. The (N_t, S_t) pairs are iid since each is based on a different tour. Assume that N_t and S_t have finite second moments. Let $\bar{N}$ be the average tour length; that is, $\bar{N} = R^{-1} \sum_{t=1}^{R} N_t$ and, analogously, let $\bar{S} = R^{-1} \sum_{t=1}^{R} S_t$. By the Strong Law of Large Numbers

$$\bar{h}_R = \frac{\sum_{t=1}^{R} S_t}{\sum_{t=1}^{R} N_t} = \frac{\bar{S}}{\bar{N}} = \frac{1}{N} \sum_{j=0}^{N-1} h(\boldsymbol{X}_j) \to E_\pi h \tag{26}$$

with probability 1 as $R \to \infty$. Furthermore, by the CLT

$$\sqrt{R}\left(\bar{h}_R - E_\pi h\right) \xrightarrow{d} \mathrm{N}\left(0, \gamma_h^2\right) . \tag{27}$$

Moreover, γ_h^2 may be consistently estimated with

$$\hat{\gamma}_h^2 = \frac{\sum_{t=1}^{R} (S_t - \bar{h}_R N_t)^2}{R \bar{N}^2} . \tag{28}$$

Given this estimate, we can form an asymptotic confidence interval for $E_\pi h$ using the formula

$$\bar{h}_R \pm z \left(\frac{\hat{\gamma}_h^2}{R} \right)^{1/2}$$

where z denotes the appropriate standard normal quantile. Mykland et al. (1995) recommend using (28) only when the estimated coefficient of variation (CV) of $\bar{N}$ is less than 0.01. We illustrate the use of the RS method in a realistic situation in Section 6.

There is little discussion in the literature about how to actually establish that N_t and S_t have finite second moments. However, Hobert, Jones, Presnell and Rosenthal (2001) have recently shown that geometric ergodicity of Φ combined with $E_\pi |h|^{2+\epsilon} < \infty$ implies that these moments are finite. Furthermore, the variances in (25) and (27) are not necessarily the same, which is why we used two different symbols.

Despite its attractive theoretical properties, there seem to be few substantive applications of RS in the MCMC literature. Four examples are Geyer and Thompson (1995), who use regeneration

to calculate Monte Carlo standard errors in the context of their simulated tempering algorithm, Gilks, Roberts and Sahu (1998), who employ regeneration to create adaptive MCMC algorithms, Guihenneuc-Jouyaux and Robert (1998), who use renewal theory as an approach to assessing convergence, and Levine and Casella (2000) who employ RS in the context of the Markov chain Monte Carlo EM algorithm.

We now briefly describe the method of *batch means*, which is an alternative method of calculating Monte Carlo standard errors. This technique is a special case of a methodology called *standardized time series*, and is the method used by the popular software package BUGS. (For more details, see Ripley (1987, Chapter 6), Bratley et al. (1987, Chapter 3) or Geyer (1992).)

Consider estimating $E_\pi h$ with $\bar{h}_n$ and suppose it is known that a CLT of the form (25) holds. The run of the sampler is broken up into batches of equal size that are assumed to be *approximately* independent. Specifically, suppose the algorithm is run for a total of $n = ab$ iterations where b is large enough so that the quantities

$$S_k = \frac{1}{b} \sum_{i=(k-1)b}^{kb-1} h(\boldsymbol{X}_i)$$

are approximately independently $\mathrm{N}\left(E_\pi h, \frac{\sigma_h^2}{b}\right)$ for $k = 1, \ldots, a$. The batch means estimate of σ_h^2 is

$$\hat{\sigma}_h^2 = \frac{b}{a-1} \sum_{k=1}^{a} \left(S_k - \bar{h}_n\right)^2 \ . \tag{29}$$

Bratley et al. (1987, Chapter 3) recommend forming an approximate confidence interval for $E_\pi h$ using

$$\bar{h}_n \pm t_{a-1} \left(\frac{\hat{\sigma}_h^2}{ab}\right)^{1/2}$$

where t_{a-1} is the appropriate quantile from the t-distribution with $a-1$ degrees of freedom. The estimator (29) is not a consistent estimator of σ_h^2 (Glynn and Iglehart 1990). Furthermore, Geyer (1992) points out that the batch means method will be effective only if the size of the batches, b, is much larger than the mixing time for the chain. In most practical applications, the only way to get a handle on the mixing time is by analyzing empirical autocorrelations.

We view the batch means method as an *ad hoc* version of the RS method. In particular, both methods break the run of the sampler up into pieces. The difference is that in the RS method,

this is done in a way that guarantees that the pieces are truly independent. Consequently, it is not necessary to analyze empirical autocorrelations before applying RS. (Of course, constructing a useful minorization condition is usually much harder than examining empirical autocorrelation plots.) Standard errors produced using RS are compared with those produced using batch means in the next section, which contains a realistic application of all the techniques described so far in this paper.

6 A Realistic Application

The methods described in Sections 3, 4 and 5 are now used to develop rigorous answers to (Q1) and (Q2) for the block Gibbs sampler for model ($\mathcal{M}$). This section has four subsections. In Subsection 6.1 we discuss the data set that will be analyzed, the priors that will be considered, and the "posterior quantities of interest" that will be estimated. A detailed description of the block Gibbs sampler is given in Subsection 6.2. Honest answers to (Q1) and (Q2) are provided in Subsections 6.3 and 6.4, respectively.

6.1 The data, the priors, and the posterior quantities of interest

Lyles et al. (1997) described an experiment in which laminators working at a boat manufacturing plant were measured for styrene exposure. Specifically, 13 workers were randomly selected from a group within the plant and each one's styrene exposure was measured on 3 separate occasions. The data are summarized in Table 5.

Lyles et al. (1997) performed a frequentist analysis of these data using a random effects model. We consider a Bayesian analysis using model ($\mathcal{M}$) from Subsection 1.3. Specifying the prior is tantamount to choosing values for the six hyperparameters: a_1, b_1, a_2, b_2, μ_0 and λ_0. While we are actually interested in the performance of the block Gibbs sampler for all possible choices of the hyperparameters, we will settle for studying six different hyperparameter settings that are shown in Table 6.

The first two settings in Table 6 represent priors that are consistent with the data. They

Table 5: Styrene Exposure Data

Worker	1	2	3	4	5	6	7
$\bar{y}_i$	3.302	4.587	5.052	5.089	4.498	5.186	4.915
Worker	8	9	10	11	12	13	
$\bar{y}_i$	4.876	5.262	5.009	5.602	4.336	4.813	

$$M_T = Km = 39$$

$$\bar{y} = M_T^{-1} \sum_{i=1}^{13} \sum_{j=1}^{3} y_{ij} = 4.809$$

$$\text{SSTR} = 3 \sum_{i=1}^{13} (\bar{y}_i - \bar{y})^2 = 11.430$$

$$\text{SSE} = \sum_{i=1}^{13} \sum_{j=1}^{3} (y_{ij} - \bar{y}_i)^2 = 14.711$$

are based on the ANOVA estimates of μ, λ_e^{-1} and λ_θ^{-1} as we now describe. Define MSE = SSE/$(M_T - K)$ and MSTR = SSTR/$(K-1)$. The ANOVA estimates of μ, λ_e^{-1} and λ_θ^{-1} are $\bar{y}$, MSE, and (MSTR $-$ MSE)/m, respectively (Searle et al. 1992, Chapter 3). We set the prior expectations for λ_θ and λ_e equal to the obvious values

$$E(\lambda_\theta) = \frac{a_1}{b_1} = \frac{m}{\text{MSTR} - \text{MSE}} = 7.76 \qquad \text{and} \qquad E(\lambda_e) = \frac{a_2}{b_2} = \frac{1}{\text{MSE}} = 1.77$$

and then solved for a_1, b_1, a_2, and b_2 by setting the prior variances both equal to $c \in \{0.1, 1\}$. As for μ, the prior mean was set equal to $\bar{y}$ and we considered a couple of different prior variances. Setting #3 is a so-called "diffuse" prior; that is, all of the prior variances are large. As we will see below, settings #4 - #6 were selected to illustrate certain points about how our answer to (Q1) depends upon the hyperparameters. We now describe the posterior quantities of interest.

Recall from (4) that the posterior density is characterized by

$$\pi(\boldsymbol{\theta}, \mu, \lambda_e, \lambda_\theta | \boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta}, \lambda_e) f(\boldsymbol{\theta}|\mu, \lambda_\theta) f(\lambda_e) f(\mu) f(\lambda_\theta) \, .$$

Due to the conjugacy in model $(\mathcal{M})$,

$$g(\lambda_\theta, \lambda_e) := \int_{\mathbb{R}} \int_{\mathbb{R}^K} f(\boldsymbol{y}|\boldsymbol{\theta}, \lambda_e) f(\boldsymbol{\theta}|\mu, \lambda_\theta) f(\lambda_e) f(\mu) f(\lambda_\theta) \, d\boldsymbol{\theta} \, d\mu$$

Table 6: Hyperparameter Settings

Setting	a_1	b_1	a_2	b_2	μ_0	λ_0	c
1	60.176	7.7573	3.1237	1.7674	4.809	1	1
2	601.76	77.573	31.237	17.674	4.809	0.1	0.1
3	0.1	0.1	0.1	0.1	4.809	0.1	–
4	1	5	1	1	3.6	1	–
5	0.6	1	120	16	4.809	1	–
6	4	80	40	100	4	1	–

has a closed form. However, the normalizing constant for $\pi(\boldsymbol{\theta}, \mu, \lambda_e, \lambda_\theta | \boldsymbol{y})$ given by

$$c_\pi := \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} g(\lambda_\theta, \lambda_e)\, d\lambda_e\, d\lambda_\theta$$

does not have an analytic solution. We will take the posterior quantities of interest to be the posterior expectations of λ_θ and λ_e; that is,

$$E(\lambda_\theta | \boldsymbol{y}) = \frac{\int\int \lambda_\theta\, g(\lambda_\theta, \lambda_e)\, d\lambda_e\, d\lambda_\theta}{c_\pi} \quad \text{and} \quad E(\lambda_e | \boldsymbol{y}) = \frac{\int\int \lambda_e\, g(\lambda_\theta, \lambda_e)\, d\lambda_e\, d\lambda_\theta}{c_\pi}\,.$$

Each of these is a ratio of intractable two-dimensional integrals, and can be computed (more or less exactly) using numerical integration. On the other hand, if we were interested in the posterior expectation of a complex function of $(\boldsymbol{\theta}, \mu, \lambda_e, \lambda_\theta)$, then the dimension of one of the intractable integrals would likely be much higher than two and could be as high as $K + 3 = 16$. Fortunately, the RS procedure that we develop and apply here can be used to get a standard error for *any* posterior expectation (as long as the appropriate CLT holds). Thus, there is really no loss of generality in looking at such simple posterior expectations. In the next subsection, we describe the block Gibbs sampler that will be applied.

6.2 The block Gibbs sampler

The two obvious fixed scan Gibbs samplers that could be employed to sample from the posterior π in (4) are (i) the ordinary "one-at-a-time" version that updates each component sequentially,

and (ii) a block version in which all of the normal components $(\theta_1, \ldots, \theta_K, \mu)$ are updated simultaneously. Given the work of Liu, Wong and Kong (1994) and Roberts and Sahu (1997), it seems likely that the block Gibbs sampler will mix faster than the ordinary Gibbs sampler. Generally speaking, blocking is effective when the constituent parts of the block are highly correlated. Also, programming the block Gibbs sampler is only slightly more difficult than programming the ordinary Gibbs sampler. Henceforth, we confine our attention to the block Gibbs sampler that is now formally defined.

Let $\boldsymbol{\xi} = (\theta_1, \ldots, \theta_K, \mu)^T$ and $\boldsymbol{\lambda} = (\lambda_\theta, \lambda_e)^T$ and define

$$V_1(\boldsymbol{\xi}) = \sum_{i=1}^{K} (\theta_i - \mu)^2 \quad \text{and} \quad V_2(\boldsymbol{\xi}) = m \sum_{i=1}^{K} (\theta_i - \overline{y}_i)^2 .$$

The full conditionals for the variance components are

$$\lambda_\theta | \boldsymbol{\xi}, \lambda_e, \boldsymbol{y} \sim \text{Gamma}\left(\frac{K}{2} + a_1, \frac{V_1(\boldsymbol{\xi})}{2} + b_1 \right)$$

and

$$\lambda_e | \boldsymbol{\xi}, \lambda_\theta, \boldsymbol{y} \sim \text{Gamma}\left(\frac{M_T}{2} + a_2, \frac{V_2(\boldsymbol{\xi}) + \text{SSE}}{2} + b_2 \right) .$$

Hobert and Geyer (1998) show that $\boldsymbol{\xi} | \boldsymbol{\lambda}, \boldsymbol{y} \sim \text{N}(\boldsymbol{\xi}^*, \boldsymbol{\Sigma})$ and give the specific forms of $\boldsymbol{\xi}^* = \boldsymbol{\xi}^*(\boldsymbol{\lambda}, \boldsymbol{y})$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\lambda}, \boldsymbol{y})$.

One cycle of the block Gibbs sampler consists of updating λ_θ, λ_e, and $\boldsymbol{\xi}$ in some order. Note that λ_θ and λ_e are conditionally independent given $\boldsymbol{\xi}$, and hence the order in which they are updated is irrelevant. Thus, in effect, the block Gibbs sampler is a data augmentation algorithm (Tanner and Wong 1987), the two components being $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$. If we update $\boldsymbol{\lambda}$ first, a one-step transition looks like $(\boldsymbol{\lambda}', \boldsymbol{\xi}') \to (\boldsymbol{\lambda}, \boldsymbol{\xi}') \to (\boldsymbol{\lambda}, \boldsymbol{\xi})$, and the corresponding Markov transition density is

$$k(\boldsymbol{\lambda}, \boldsymbol{\xi} | \boldsymbol{\lambda}', \boldsymbol{\xi}') = f(\lambda_\theta | \boldsymbol{\xi}', \boldsymbol{y}) f(\lambda_e | \boldsymbol{\xi}', \boldsymbol{y}) f(\boldsymbol{\xi} | \boldsymbol{\lambda}, \boldsymbol{y}). \tag{30}$$

It is easy to show that this Markov chain satisfies assumption $(\mathcal{A})$. Jones and Hobert (2001) established drift and minorization conditions for this block Gibbs sampler and these are stated in Appendix I.

6.3 Honest burn-in

For each of the six hyperparameter settings in Table 6, we used Theorem 2 in conjunction with the drift and minorization conditions in Appendix I to find a value of n such that

$$\|P^n((\boldsymbol{\lambda}_0, \boldsymbol{\xi}_0), \cdot) - \pi(\cdot)\| \leq 0.01 \tag{31}$$

where $(\boldsymbol{\lambda}_0, \boldsymbol{\xi}_0)$ is the starting value. In each case, we used $\boldsymbol{\xi}_0 = (\bar{y}_1, \ldots, \bar{y}_K, \ \bar{y})^T$. As is clear from (30), a starting value for $\boldsymbol{\lambda}$ is not required. (This convergence criterion (≤ 0.01) has become fairly standard (Rosenthal 1996, Cowles and Rosenthal 1998, Roberts and Rosenthal 1999).)

Table 7 contains the results. For example, under the first hyperparameter setting, after 3×10^8 iterations of the block Gibbs sampler, the total variation distance to stationarity is at most 0.00429. It takes about 2 minutes to run 1 million iterations of our block Gibbs sampler on a standard PC. Consequently, the only settings that result in unmanageable burn-in's are settings #3 and #4. Recall that setting #3 is the diffuse prior. The result for setting #4 is typical of those settings in which μ_0 is far from $\bar{y}$ and we included setting #4 specifically to demonstrate this problem. Settings #5 and #6 are the result of "playing around" with the hyperparameters. Indeed, with #5 we were trying to find a setting that would give a very short burn-in. With #6, we were trying to see if it was possible to find a setting in which $|\mu_0 - \bar{y}|$ is not small, but the resulting burn-in is still manageable. It is interesting to note that the value of λ_0 does not appear in the drift condition or in the minorization condition and hence its value has no bearing on the results in Table 7.

There are two possible reasons why our results suggest that such a long burn-in is necessary for settings #3 and #4: (i) The results merely reflect the fact that the block Gibbs sampler mixes very slowly under these hyperparameter settings, or (ii) For these particular hyperparameter settings, Theorem 2 (combined with the drift and minorization from Jones and Hobert (2001)) results in very conservative bounds. The results in the next subsection suggest that the real reason is probably (ii).

Table 7: Total Variation Bounds for the Styrene Exposure Data

Setting	λ	b	d	r	ε	Iterations	Bound
1	0.4810	2.9872	12.311	0.0107	4.00×10^{-12}	3×10^{8}	0.00429
2	0.2112	3.5925	10.708	0.0426	2.74×10^{-6}	10,000	0.00997
3	0.0504	2.5097	11.212	0.0158	5.39×10^{-14}	2×10^{12}	0.00324
4	0.4265	25.380	93.20	0.0059	1.06×10^{-15}	7×10^{17}	0.00638
5	0.1279	1.6620	4.8113	0.08550	4.48×10^{-3}	4,600	0.00921
6	0.1450	22.737	61.283	0.0275	1.11×10^{-4}	95,000	0.00515

6.4 Honest standard errors

The drift and minorization conditions established by Jones and Hobert (2001) show that our block Gibbs sampler is geometrically ergodic (see also Hobert and Geyer 1998). Furthermore, it is easy to show that $E_\pi(\lambda_\theta^p|\boldsymbol{y})$ and $E_\pi(\lambda_e^p|\boldsymbol{y})$ are both finite for any $p > 0$. Thus, as discussed above, there are CLTs for the Monte Carlo estimators of our posterior quantities of interest.

A minorization condition of the form (22) is constructed for our block Gibbs sampler in Appendix II. In order to use this minorization condition for RS, we had to choose the distinguished point $\tilde{\boldsymbol{\xi}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_K, \tilde{\mu})^T \in \mathbb{R}^{K+1}$ and the d_i's that determine the set $D = [d_1, d_2] \times [d_3, d_4] \subseteq \mathbb{R}^+ \times \mathbb{R}^+$. These choices were made as follows. The block Gibbs sampler was run for an initial 10,000 iterations from the starting value $\boldsymbol{\xi}_0 = (\bar{y}_1, \ldots, \bar{y}_K, \bar{y})^T$. Let $\theta_1^{(b)}, \ldots, \theta_K^{(b)}, \mu^{(b)}, \lambda_e^{(b)}$ and $\lambda_\theta^{(b)}$ denote the estimates of the posterior expectations of the corresponding parameters based on these initial 10,000 iterations. For the distinguished point, we set $\tilde{\boldsymbol{\xi}} = (\theta_1^{(b)}, \ldots, \theta_K^{(b)}, \mu^{(b)})^T$. As for D, we set $[d_1, d_2] = \lambda_\theta^{(b)} \pm 1.1s$ where s is the (usual) sample standard deviation of the 10,000 values of λ_θ, and the interval $[d_3, d_4]$ was constructed similarly. (We used 1.1 as it seems to result in a reasonable regeneration rate.)

For each of the six hyperparameter settings, we performed RS as described in Section 5. Specifically, we started with a regeneration; that is, $(\boldsymbol{\lambda}_0, \boldsymbol{\xi}_0) \sim q(\boldsymbol{\lambda}, \boldsymbol{\xi})$, and then ran the chain for as many regenerations as we needed to get the Monte Carlo error down to a reasonable level. Table 8

contains the results. For example, we are (approximately) 95% confident that the true value of $E(\lambda_\theta|\boldsymbol{y})$ under the first prior is in the interval $(7.753, 7.765)$. (In each case, the estimated CV of $\bar{N}$ was less than 0.01.) Aside from point estimates and confidence intervals, Table 8 also contains the estimates of the asymptotic variances calculated via (28), the total number of regenerations (R) for which the chain was run, and the mean number of iterations per regeneration.

Table 8: Point Estimates and Asymptotic 95% Confidence Intervals for $E(\lambda_\theta|\boldsymbol{y})$ and $E(\lambda_e|\boldsymbol{y})$

Setting	Parameter	Estimate	$\hat{\gamma}_h^2$	95% CI	Number of regenerations	Mean number of iter/regen
1	λ_θ	7.759	0.2003	(7.753, 7.765)	25000	5.68
	λ_e	1.779	0.0435	(1.776, 1.782)		
2	λ_θ	7.758	0.0305	(7.755, 7.761)	12000	3.39
	λ_e	1.769	0.0227	(1.766, 1.772)		
3	λ_θ	7.363	7.9731	(7.349, 7.377)	150000	24.4
	λ_e	1.793	0.0161	(1.792, 1.794)		
4	λ_θ	0.958	0.0251	(0.955, 0.961)	10000	7.43
	λ_e	1.756	0.0453	(1.752, 1.760)		
5	λ_θ	2.438	0.3036	(2.427, 2.449)	10000	5.04
	λ_e	5.699	0.0537	(5.694, 5.704)		
6	λ_θ	0.118	0.0003	(0.1176, 0.1184)	6000	4.55
	λ_e	0.498	0.0012	(0.497, 0.499)		

It is important to recognize that we did not use any burn-in to obtain the results in Table 8. Indeed, as noted by Ripley (1987, Chapter 6), Bratley et al. (1987, Chapter 3) and Mykland et al. (1995), one of the best features of the RS method is that burn-in is simply not an issue. On the other hand, you could consider the initial 10,000 iterations used to construct the distinguished point and the set D as a form of burn-in.

Recall the results of Theorem 2 given in Table 7: In order to be certain that the total variation distance to stationarity is less than 0.01, the chains corresponding to settings #3 and #4 need

to be run for 2×10^{12} iterations and 7×10^{17} iterations, respectively. On the other hand, we see from Table 8 that chains #3 and #4 needed to be run for only about 3.7 million iterations and 75,000 iterations, respectively, in order to get the Monte Carlo error down to a reasonable level. This suggests that, in these two cases, Theorem 2 (combined with the drift and minorization from Jones and Hobert (2001)) is extremely conservative.

Based on Table 8, it seems that the chain associated with setting #3 is by far the slowest mixing of the 6. Recall that setting #3 is the diffuse prior. This is consistent with the findings of Natarajan and McCulloch (1998) whose empirical results suggest that the mixing rate of the Gibbs sampler becomes slower as the priors become more diffuse (but see van Dyk and Meng 2001).

Table 9: Batch Means Estimates and Asymptotic Confidence Intervals

Setting	Parameter	Estimate	$\hat{\sigma}_h^2$	95% CI	Batch Size	Iterations
1	λ_θ	7.756	0.9959	(7.751, 7.761)	4733	141990
	λ_e	1.778	0.1920	(1.776, 1.780)		
2	λ_θ	7.757	0.0966	(7.754, 7.760)	1359	40770
	λ_e	1.770	0.0659	(1.767, 1.773)		
3	λ_θ	7.352	179.07	(7.338, 7.367)	122000	3660000
	λ_e	1.793	0.3817	(1.792, 1.794)		
4	λ_θ	0.957	0.2027	(0.954, 0.960)	2476	74280
	λ_e	1.760	0.2808	(1.756, 1.764)		
5	λ_θ	2.435	1.4942	(2.424, 2.446)	1680	50400
	λ_e	5.702	0.1850	(5.698, 5.706)		
6	λ_θ	0.118	0.0015	(0.1175, 0.1185)	911	27330
	λ_e	0.498	0.0045	(0.497, 0.499)		

For the sake of comparison, we also calculated approximate confidence intervals for the posterior quantities of interest using the method of batch means. To make the comparison with RS fair, we used a burn-in of 10,000 iterations and ran the block Gibbs sampler for roughly the same overall number of iterations. An examination of the empirical autocorrelation function indicated

that, in each case, using 30 batches results in sufficiently large batch sizes. This is consistent with the recommendations of Schmeiser (1982). Table 9 contains the results. The variance estimate reported in Table 9 is (29). The confidence intervals in Tables 8 and 9 are quite similar.

7 Concluding remarks

Our hope is that this paper will serve as a bridge between those developing theoretical Markov chain theory and practitioners who would like exact answers to (Q1) and (Q2) for their particular MCMC algorithms. Obviously, the MCMC samplers that we have considered in this paper are relatively simple compared to most of those being used in realistic settings. For example, if one (or more) of the full conditionals in a Gibbs sampler is a nonstandard density, then establishing drift and minorization conditions is sure to be much harder than it was for the Gibbs samplers studied here. Furthermore, replacing the Gibbs update of the nonstandard conditional by a Metropolis–Hastings update will usually further complicate the calculations. There is clearly a great deal of work to be done before rigorously addressing (Q1) and (Q2) becomes standard practice when applying MCMC.

If nothing else, the results in this paper show that forming honest answers to (Q1) and (Q2) can be hard work. Moreover, recall that when using classical Monte Carlo methods based on independent samples, (Q1) is moot and (Q2) is easy. Thus, before resorting to MCMC, one should try the Monte Carlo methods based on independent samples; e.g., rejection sampling or importance sampling. In other words, MCMC should not be the default approach when one is confronted with analytically difficult integrals. This may sound obvious, but so does wearing a seat-belt.

Appendix I: Drift and minorization from Jones and Hobert (2001)

Jones and Hobert (2001) established drift and minorization conditions for the block Gibbs sampler for model ($\mathcal{M}$) which we now state. We begin with the drift condition. Define two constants as follows

$$\delta_1 = \frac{1}{2a_1 + K - 2} \quad \text{and} \quad \delta_2 = \frac{1}{2a_2 + M_T - 2}\,.$$

Also, let $\delta_3 = K\,\delta_2$ and $\delta_4 = (K+1)\,\delta_2$. It follows from our assumptions about K and m that δ_1, δ_2, δ_3, and δ_4 are all in $(0,1)$. Let $\delta = \max\{\delta_1, \delta_4\}$. Choose $\lambda \in (\delta, 1)$ and then choose $\phi > 0$ such that $\phi\,\delta_3 + \delta < \lambda$. Define

$$V(\boldsymbol{\xi}) = \phi\, V_1(\boldsymbol{\xi}) + m^{-1}\, V_2(\boldsymbol{\xi})$$

where V_1 and V_2 are defined in Subsection 6.2. Jones and Hobert (2001) show that

$$E\left[V(\boldsymbol{\xi})|\boldsymbol{\lambda}', \boldsymbol{\xi}'\right] \leq \lambda\, V(\boldsymbol{\xi}') + b \tag{32}$$

where

$$b = 2\phi b_1 \delta_1 + \delta_2 \left[\frac{\phi K + K + 1}{m}\right](2b_2 + \text{SSE}) + K(\phi + 1)(\bar{y} - \mu_0)^2\,.$$

We now state the minorization condition.

Let $d > \frac{2b}{1-\lambda}$ and recall that, in order to apply Theorem 2, we must have a minorization condition that holds on $C = \{(\boldsymbol{\lambda}, \boldsymbol{\xi}) : V(\boldsymbol{\xi}) \leq d\}$. Jones and Hobert (2001) establish that the Markov transition density for the block Gibbs sampler satisfies

$$k(\boldsymbol{\lambda}, \boldsymbol{\xi}|\boldsymbol{\lambda}', \boldsymbol{\xi}') \geq \varepsilon\, q(\boldsymbol{\lambda}, \boldsymbol{\xi})$$

for all $(\boldsymbol{\lambda}', \boldsymbol{\xi}') \in C$ where $q(\boldsymbol{\lambda}, \boldsymbol{\xi})$ is a density on $(\mathbb{R}^+ \times \mathbb{R}^+) \times \mathbb{R}^{K+1}$ defined by

$$q(\boldsymbol{\lambda}, \boldsymbol{\xi}) = \left[\frac{h_1(\lambda_\theta)}{\int_{\mathbb{R}^+} h_1(\lambda_\theta)\, d\lambda_\theta}\right]\left[\frac{h_2(\lambda_e)}{\int_{\mathbb{R}^+} h_2(\lambda_e)\, d\lambda_e}\right] \pi(\boldsymbol{\xi}|\boldsymbol{\lambda}, \boldsymbol{y})$$

and $\varepsilon = \left[\int_{\mathbb{R}^+} h_1(\lambda_\theta)\, d\lambda_\theta\right]\left[\int_{\mathbb{R}^+} h_2(\lambda_e)\, d\lambda_e\right]$. The functions h_1 and h_2 are defined as follows. Let $\text{Gamma}(\alpha, \beta; w)$ denote a $\text{Gamma}(\alpha, \beta)$ density evaluated at $w > 0$. Then

$$h_1(\lambda_\theta) = \begin{cases} \text{Gamma}\left(\frac{K}{2} + a_1, b_1; \lambda_\theta\right) & \lambda_\theta < \lambda_\theta^* \\ \text{Gamma}\left(\frac{K}{2} + a_1, \frac{d}{2\phi} + b_1; \lambda_\theta\right) & \lambda_\theta \geq \lambda_\theta^* \end{cases}$$

for

$$\lambda_\theta^* = \frac{\phi(K+2a_1)}{d}\log\left(1+\frac{d}{2b_1\phi}\right)$$

and

$$h_2(\lambda_e) = \begin{cases} \text{Gamma}\left(\frac{M_T}{2}+a_2, \frac{\text{SSE}}{2}+b_2; \lambda_e\right) & \lambda_e < \lambda_e^* \\ \text{Gamma}\left(\frac{M_T}{2}+a_2, \frac{\text{SSE}+d}{2}+b_2; \lambda_e\right) & \lambda_e \geq \lambda_e^* \end{cases}$$

for

$$\lambda_e^* = \frac{M_T+2a_2}{d}\log\left(1+\frac{d}{2b_2+\text{SSE}}\right).$$

Note that ε can be computed with four calls to the incomplete gamma function.

Appendix II: Minorization for the block Gibbs sampler

In this appendix, we construct a minorization condition of the form (22) for the block Gibbs sampler introduced in Subsection 6.2. We will use the same technique that was used in the example of Subsection 5.1.

Fix a distinguished point $\tilde{\boldsymbol{\xi}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_K, \tilde{\mu})^T \in \mathbb{R}^{K+1}$. Now let $0 < d_1 < d_2 < \infty$ and $0 < d_3 < d_4 < \infty$ and define $D = [d_1, d_2] \times [d_3, d_4] \subseteq \mathbb{R}^+ \times \mathbb{R}^+$. For notational convenience, let $V_i' = V_i(\boldsymbol{\xi}')$ and $\tilde{V}_i = V_i(\tilde{\boldsymbol{\xi}})$ for $i = 1, 2$. (Note that V_1 and V_2 are defined in Subsection 6.2.) Then we have

$$f(\boldsymbol{\lambda}|\boldsymbol{\xi}', \boldsymbol{y}) f(\boldsymbol{\xi}|\boldsymbol{\lambda}, \boldsymbol{y}) \geq s(\boldsymbol{\xi}')\, q(\boldsymbol{\lambda}, \boldsymbol{\xi})$$

where q is a density on $D \times \mathbb{R}^{K+1}$ given by

$$q(\boldsymbol{\lambda}, \boldsymbol{\xi}) = \frac{1}{c} f(\boldsymbol{\lambda}|\tilde{\boldsymbol{\xi}}, \boldsymbol{y})\, f(\boldsymbol{\xi}|\boldsymbol{\lambda}, \boldsymbol{y})\, I(\boldsymbol{\lambda} \in D)$$

for $c = \int_D f(\boldsymbol{\lambda}|\tilde{\boldsymbol{\xi}}, \boldsymbol{y})\, d\boldsymbol{\lambda}$ and

$$\begin{aligned} s(\boldsymbol{\xi}') &= c \inf_{\boldsymbol{\lambda} \in D} \left[\frac{f(\boldsymbol{\lambda}|\boldsymbol{\xi}', \boldsymbol{y})}{f(\boldsymbol{\lambda}|\tilde{\boldsymbol{\xi}}, \boldsymbol{y})}\right] \\ &= c \inf_{\boldsymbol{\lambda} \in D} \left[\frac{\text{Gamma}\left(\frac{K}{2}+a_1, b_1+\frac{1}{2}V_1'; \lambda_\theta\right) \text{Gamma}\left(\frac{M_T}{2}+a_2, b_2'+\frac{1}{2}V_2'; \lambda_e\right)}{\text{Gamma}\left(\frac{K}{2}+a_1, b_1+\frac{1}{2}\tilde{V}_1; \lambda_\theta\right) \text{Gamma}\left(\frac{M_T}{2}+a_2, b_2'+\frac{1}{2}\tilde{V}_2; \lambda_e\right)}\right] \end{aligned}$$

where $b_2' = b_2 + \mathrm{SSE}/2$. Now straightforward calculations reveal that

$$s(\boldsymbol{\xi}') = c \left[\frac{2b_1 + V_1'}{2b_1 + \tilde{V}_1}\right]^{\frac{K}{2}+a_1} \left[\frac{2b_2' + V_2'}{2b_2' + \tilde{V}_2}\right]^{\frac{M_T}{2}+a_2} \exp\left\{\frac{g_\theta}{2}(\tilde{V}_1 - V_1') + \frac{g_e}{2}(\tilde{V}_2 - V_2')\right\}$$

where

$$g_\theta = \begin{cases} d_1 & \text{if } \tilde{V}_1 > V_1' \\ d_2 & \text{if } \tilde{V}_1 < V_1' \end{cases} \qquad \text{and} \qquad g_e = \begin{cases} d_3 & \text{if } \tilde{V}_2 > V_2' \\ d_4 & \text{if } \tilde{V}_2 < V_2' \end{cases}.$$

Finally, for the transition $(\boldsymbol{\lambda}', \boldsymbol{\xi}') \to \delta \to (\boldsymbol{\lambda}, \boldsymbol{\xi})$, we have

$$\Pr\left[\delta = 1 | (\boldsymbol{\lambda}', \boldsymbol{\xi}'), (\boldsymbol{\lambda}, \boldsymbol{\xi})\right] = I(\boldsymbol{\lambda} \in D) \exp\left\{\frac{1}{2}(g_\theta - \lambda_\theta)(\tilde{V}_1 - V_1') + \frac{1}{2}(g_e - \lambda_e)(\tilde{V}_2 - V_2')\right\}. \quad (33)$$

Observe that the value of the normalizing constant c was not required for this calculation.

Acknowledgment. The authors are grateful to Brian Caffo, Leon Gleser, Brett Presnell, Jeff Rosenthal, Richard Tweedie, an anonymous Editor, and an anonymous referee for constructive comments and suggestions.

References

Athreya, K. B. and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains, *Transactions of the American Mathematical Society* **245**: 493–501.

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

Billera, L. J. and Diaconis, P. (2001). A geometric interpretation of the Metropolis algorithm, *Technical report*, Cornell University.

Bratley, P., Fox, B. L. and Schrage, L. E. (1987). *A Guide to Simualtion*, Springer–Verlag, New York.

Caffo, B. S., Booth, J. G. and Davison, A. C. (2001). Empirical sup rejection sampling, *Technical report*, University of Florida.

Chan, K. S. and Geyer, C. J. (1994). Comment on "Markov chains for exploring posterior distributions", *The Annals of Statistics* **22**: 1747–1758.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm, *The American Statistician* **49**: 327–335.

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

Crane, M. A. and Iglehart, D. L. (1975). Simulating stable stochastic systems III: Regenerative processes and discrete-event simulations, *Operations Research* **23**: 33–45.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains, *The Annals of Applied Probability* **1**: 36–61.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions, *The Annals of Statistics* **26**: 363–397.

Frigessi, A., di Stefano, P., Hwang, C.-R. and Sheu, S.-J. (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**: 205–219.

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.

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.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion), *Statistical Science* **7**: 473–511.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference, *Journal of the American Statistical Association* **90**: 909–920.

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

Gilks, W. R., Roberts, G. O. and Sahu, S. K. (1998). Adaptive Markov chain Monte Carlo through regeneration, *Journal of the American Statistical Association* **93**: 1045–1054.

Glynn, P. W. (1985). Regenerative structure of Markov chains simulated via common random numbers, *Operations Research Letters* **4**: 49–53.

Glynn, P. W. and Iglehart, D. L. (1987). A joint central limit theorem for the sample mean and regenerative variance estimator, *The Annals of Operations Research* **8**: 41–55.

Glynn, P. W. and Iglehart, D. L. (1990). Simulation output analysis using standardized time series, *Mathematics of Operations Research* **15**: 1–16.

Guihenneuc-Jouyaux, C. and Robert, C. P. (1998). Discretization of continuous Markov chains and Markov chain Monte Carlo convergence assessment, *Journal of the American Statistical Association* **93**: 1055–1067.

Hobert, J. P. (2001). Discussion of "The art of data augmentation", *Journal of Computational and Graphical Statistics* **10**: 59–68.

Hobert, J. P. and Geyer, C. J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model, *Journal of Multivariate Analysis* **67**: 414–430.

Hobert, J. P., Jones, G. L., Presnell, B. and Rosenthal, J. S. (2001). On the applicability of regenerative simulation in Markov chain Monte Carlo, *Technical report*, University of Florida.

Ingrassia, S. (1994). On the rate of convergence of the Metropolis algorithm and Gibbs sampler by geometric bounds, *The Annals of Applied Probability* **4**: 347–389.

Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms, *Stochastic Processes and Their Applications* **85**: 341–361.

Jarner, S. F. and Roberts, G. O. (2001). Polynomial convergence rates of Markov chains, *The Annals of Applied Probability* (to appear).

Jones, G. L. and Hobert, J. P. (2001). Upper bounds on the distance to stationarity for the block Gibbs sampler for a hierarchical random effects model, *Technical report*, University of Florida.

Levine, R. A. and Casella, G. (2000). Implementations of the Monte Carlo EM algorithm, *Journal of Computational and Graphical Statistics* (to appear).

Lindvall, T. (1992). *Lectures on the Coupling Method*, Wiley-Interscience, New York.

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

Lund, R. B. and Tweedie, R. L. (1996). Geometric convergence rates for stochastically ordered Markov chains, *Mathematics of Operations Research* **20**: 182–194.

Lyles, R. H., Kupper, L. L. and Rappaport, S. M. (1997). Assessing regulatory compliance of occupational exposures via the balanced one-way random effects ANOVA model, *Journal of Agricultural, Biological, and Environmental Statistics* **2**: 64–86.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association* **92**: 162–170.

Mengersen, K. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms, *The Annals of Statistics* **24**: 101–121.

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 geometric convergence rates of Markov chains, *The Annals of Applied Probability* **4**: 981–1011.

Mira, A. and Tierney, L. (2001). On the use of auxiliary variables in Markov chain Monte Carlo sampling, *Scandinavian Journal of Statistics* (to appear).

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers, *Journal of the American Statistical Association* **90**: 233–241.

Natarajan, R. and McCulloch, C. E. (1998). Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference?, *Journal of Computational and Graphical Statistics* **7**: 267–277.

Nummelin, E. (1978). A splitting technique for Harris recurrent Markov chains, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **43**: 309–318.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*, Cambridge University Press, London.

Ripley, B. D. (1987). *Stochastic Simulation*, John Wiley and Sons, New York.

Robert, C. P. (1995). Convergence control methods for Markov chain Monte Carlo algorithms, *Statistical Science* **10**: 231–253.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer, New York.

Roberts, G. O. (1999). A note on acceptance rate criteria for CLTs for Metropolis-Hastings algorithms, *Journal of Applied Probability* **36**: 1210–1217.

Roberts, G. O. and Rosenthal, J. S. (1998a). Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion), *Canadian Journal of Statistics* **26**: 5–31.

Roberts, G. O. and Rosenthal, J. S. (1998b). On convergence rates of Gibbs samplers for uniform distributions, *The Annals of Applied Probability* **8**: 1291–1302.

Roberts, G. O. and Rosenthal, J. S. (1999). Convergence of slice sampler Markov chains, *Journal of the Royal Statistical Society,* Series B **61**: 643–660.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parametrization for the Gibbs sampler, *Journal of the Royal Statistical Society,* Series B **59**: 291–317.

Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms, *Biometrika* **83**: 95–110.

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

Roberts, G. O. and Tweedie, R. L. (2001). Corrigendum to "Bounds on regeneration times and convergence rates for Markov chains", *Stochastic Processes and their Applications* **91**: 337–338.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo, *Journal of the American Statistical Association* **90**: 558–566.

Rosenthal, J. S. (1996). Analysis of the Gibbs sampler for a model related to James-Stein estimators, *Statistics and Computing* **6**: 269–275.

Rosenthal, J. S. (2001). A review of asymptotic convergence for general state space Markov chains, *Far East Journal of Theoretical Statistics* **5**: 37–50.

Schmeiser, B. (1982). Batch size effects in the analysis of simulation output, *Operations Research* **30**: 556–568.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Componenets*, John Wiley and Sons, New York.

Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1999). WinBUGS Version 1.2, MRC Biostatistics Unit, Cambridge: UK.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion), *Journal of the American Statistical Association* **82**: 528–550.

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

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces, *The Annals of Applied Probability* **8**: 1–9.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion), *Journal of Computational and Graphical Statistics* **10**: 1–111.

Yuen, W. K. (2000). Applications of geometric bounds to the convergence rate of Markov chains on $\mathbb{R}^n$, *Stochastic Processes and Their Applications* **87**: 1–23.

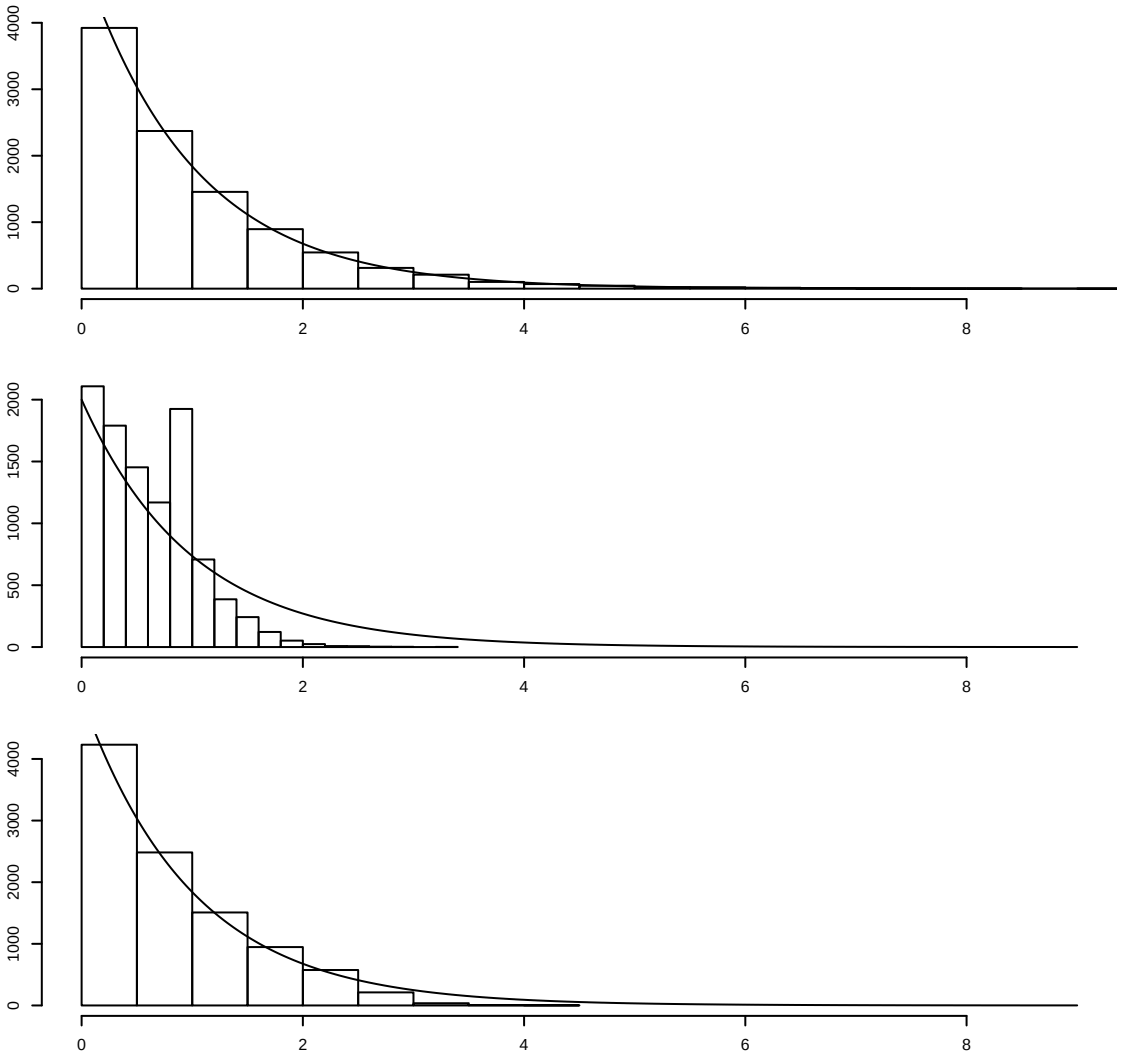

Figure 1: The independence Metropolis algorithm with $\theta = 0.5$ was run for 15 iterations 10,000 times. In each case, the starting value was $X_0 = 1$. The histogram at the top shows the 10,000 iid copies of X_{15}. The experiment was repeated with $\theta = 4$ and the middle histogram again shows the 10,000 iid copies of X_{15}. Finally, the independence Metropolis algorithm with $\theta = 4$ was run for 1,000 iterations 10,000 times. The bottom histogram shows the 10,000 iid copies of X_{1000}. In each plot, the solid line is an appropriately scaled version of the stationary density.

A graphical illustration of $g(\theta) := \inf_{\mu \in C_\mu} f(\theta|\mu, \boldsymbol{y})$

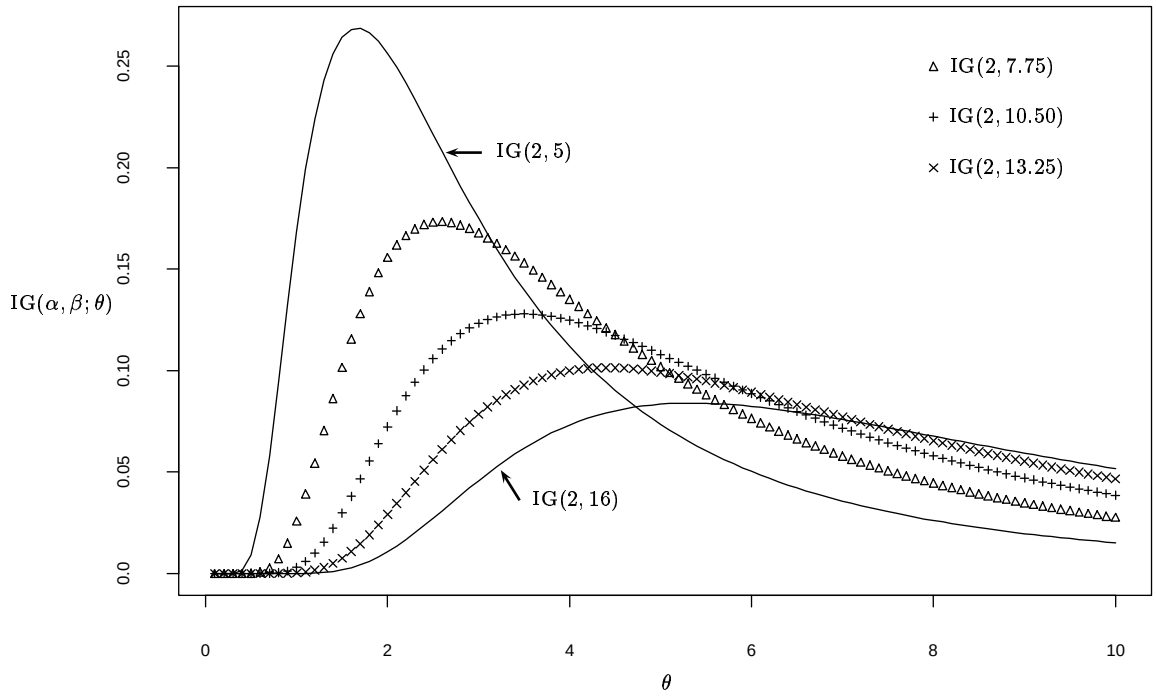

Figure 2: Consider the family of densities IG $\left(\frac{m-1}{2}, \frac{s^2}{2} + \frac{m}{2}(\mu - \overline{y})^2\right)$ as μ ranges over the set C_μ; that is, as $(\mu - \overline{y})^2$ ranges between 0 and d. Suppose, for example, that $m = 5$, $s^2 = 10$, and $d = 22/5$. Then the shape parameter is 2 and the scale parameter ranges between 5 and 16. Five of these densities, including the extremes, are pictured above. The point of intersection of the two extremes is $\theta^* = 4.73$. It is clear that one of the extremes is always the minimum. Specifically, when $\theta \in (0, \theta^*)$, IG$(2, 16)$ is below all the others, while for values above θ^*, IG$(2, 5)$ is the smallest.