Using a Markov Chain to Construct a Tractable Approximation of an Intractable Probability Distribution

James P. Hobert
Department of Statistics
University of Florida

Galin L. Jones
School of Statistics
University of Minnesota

Christian P. Robert
Université Paris Dauphine
& CREST, INSEE

April 2004 (Revised January 2005, April 2005)

Abbreviated title. Approximating an intractable distribution

Key words and phrases. burn-in, Gibbs sampler, minorization condition, mixture representation, Monte Carlo, regeneration, split chain

Abstract

Let π denote an intractable probability distribution that we would like to explore. Suppose that we have a positive recurrent, irreducible Markov chain that satisfies a minorization condition and has π as its invariant measure. We provide a method of using simulations from the Markov chain to construct a statistical estimate of π from which it is straightforward to sample. We show that this estimate is "strongly consistent" in the sense that the total variation distance between the estimate and π converges to 0 almost surely as the number of simulations grows. Moreover, we use some recently developed asymptotic results to provide guidance as to how much simulation is necessary. Draws from the estimate can be used to approximate features of π or as intelligent starting values for the original Markov chain. We illustrate our methods with two examples.

1 Introduction

Let π be a probability distribution on X that we would like to explore. For example, we might want to know the value of $\mathrm{E}_\pi g := \int_{\mathsf{X}} g(x)\,\pi(dx)$. Suppose that π is intractable in the sense that numerical integration and classical Monte Carlo methods are not viable options for approximating the features of π. Also, assume that we have at our disposal a Markov transition kernel, $P(x, dy)$, that satisfies the usual regularity conditions (see Section 2), has π as its invariant probability measure and is easy to simulate. Write the corresponding Markov chain as $X = \{X_n\}_{n=0}^{\infty}$. As is now well-known, there are many methods for constructing such kernels (see, e.g., Liu, 2001; Robert and Casella, 2004). A Markov chain Monte Carlo (MCMC) solution to the problem is to estimate $\mathrm{E}_\pi g$ using $\overline{g}_n := n^{-1}\sum_{i=0}^{n-1} g(X_i)$, which converges almost surely to $\mathrm{E}_\pi g$ no matter what the distribution of X_0.

Since the complexity of π precludes the use of classical Monte Carlo methods, it will also be difficult, if not impossible, to start the Markov chain in stationarity (by drawing $X_0 \sim \pi$). Hence, the estimate $\overline{g}_n$ will be based on a sequence of random variables that are neither independent nor identically distributed. Two important consequences of this are that $\overline{g}_n$ is a biased estimate of $\mathrm{E}_\pi g$ and that variance estimation cannot be based on standard techniques for independent and identically distributed (iid) data. Note that these two problems do not surface when using classical Monte Carlo methods based on iid draws from π.

Typically, the chain is started by setting $X_0 = x$, where x is just some point from which it is convenient to start the simulation. Let $P^n(x, \cdot)$ represent the distribution of X_n given $X_0 = x$. The basic Markov chain theory underlying MCMC implies that $\|\pi(\cdot) - P^n(x, \cdot)\| \downarrow 0$, where $\|\cdot\|$ denotes the total variation norm. Thus, the marginal distribution of each successive X_n is closer to π than the previous one. Often, in order to reduce the effect of using something other than π as the starting distribution, the first b, say, simulated values are "thrown out" and only the values of $X_b, X_{b+1}, \ldots$ are used to explore π. For example, instead of using $\overline{g}_n$ to estimate $\mathrm{E}_\pi g$, we would use $n^{-1}\sum_{i=b}^{n+b-1} g(X_i)$. This practice is known as *burn-in*. Aside from reducing bias, burn-in may lead to improved performance of variance estimation techniques whose derivations are based on the assumption that the underlying stochastic process is stationary. Ideally, one would like to choose the amount of burn-in by calculating b such that $\|\pi(\cdot) - P^b(x, \cdot)\| < \gamma$ where $\gamma > 0$ is some predetermined constant. Unfortunately, the methods that are currently available for doing this require the user to perform some potentially "difficult theoretical analysis" (Fill, Machida, Murdoch and Rosenthal, 2000) of the Markov chain before they can be applied (Baxendale, 2005; Douc, Moulines and Rosenthal, 2004; Meyn and Tweedie, 1994; Roberts and Tweedie, 1999; Rosenthal, 1995a). To be specific, *minorization* and *drift* conditions must be established for the Markov chain. See Jones and Hobert (2001) for a simple introduction to these concepts.

Hobert and Robert (2004) presented an alternative solution to the problem described above, which, unfortunately, also requires minorization and drift conditions. These authors show if P satisfies a minorization condition of the form $P(x, \cdot) \geq \varepsilon I_C(x)\nu(\cdot)$, where $\varepsilon > 0$,

$C \subset \mathsf{X}$ and $\nu(\cdot)$ is a measure on X, then π can be represented as

$$\pi(A) = \sum_{t=1}^{\infty} Q_t(A)\, p_t \;, \tag{1}$$

where each $Q_t(\cdot)$ is a probability measure (on the same space as π) and $\{p_t\}_{t=1}^{\infty}$ is a sequence of nonnegative numbers that sum to 1. The Q_ts and p_ts, which are formally defined in Section 2, are associated with the hitting times on an accessible atom introduced via the splitting construction of Athreya and Ney (1978) and Nummelin (1978).

Representation (1) is appealing from a simulation point of view because it reveals the potential for drawing from π by randomly drawing an element from the set $\{Q_1, Q_2, Q_3, \dots\}$ according to the probabilities $p_1, p_2, p_3, \dots$ and then making an independent random draw from the chosen Q_t. Of course, the first part of this recipe is equivalent to simulating a discrete random variable, call it T, whose mass function is given by $\Pr(T = t) = p_t$ for $t = 1, 2, 3 \dots$. We shall see later that drawing from $Q_t(\cdot)$ is simple. The challenge is simulating T.

One situation where the distribution of T is simple is when $C = \mathsf{X}$. In this case, the Markov chain X is *uniformly ergodic* and $p_t = \varepsilon(1-\varepsilon)^{t-1}$; that is, the p_ts are geometric probabilities. In this case, it is easy to use (1) to make iid draws from π. This fact has been used either directly or indirectly by many authors including Asmussen, Glynn and Thorisson (1992), Murdoch and Green (1998), Breyer and Roberts (2001) and Wilson (2000). Unfortunately, there is no known method for simulating T outside of the uniformly ergodic case. This is important because most Markov chains underlying practically relevant MCMC algorithms are not uniformly ergodic. (Jones and Hobert (2004) call an MCMC algorithm *practically relevant* when the stationary distribution is complex enough that iid sampling is not straightforward.)

Suppose now that C is a proper subset of X. Let $\{\hat{p}_t\}_{t=1}^{\infty}$ be a second sequence of nonnegative numbers that sum to 1 and let $\hat{T}$ denote the corresponding discrete random variable. Consider an approximation to π of the form $\hat{\pi}(A) = \sum_{t=1}^{\infty} Q_t(A)\, \hat{p}_t$. If one can simulate $\hat{T}$, then one can make draws from $\hat{\pi}$. Furthermore, it is easy to see that the total variation distance between π and $\hat{\pi}$ satisfies

$$\|\pi - \hat{\pi}\| \leq \sum_{t=1}^{\infty} |p_t - \hat{p}_t| \;.$$

Hobert and Robert (2004) show how to use a geometric drift condition on the Markov chain X to construct a sequence $\{\hat{p}_t\}_{t=1}^{\infty}$ such that $\sum_{t=1}^{\infty} |p_t - \hat{p}_t|$ is arbitrarily small *and* simulating $\hat{T}$ is easy. This yields an alternative solution to the burn-in problem. Indeed, one simply constructs $\hat{\pi}$ such that $\|\pi - \hat{\pi}\| < \gamma$ and then $\hat{\pi}$ is used as the starting distribution; i.e., $X_0 \sim \hat{\pi}$.

In this paper we consider the more realistic situation where a geometric drift condition is not available. Our main contribution is a method of using simulations of the Markov chain X to build an explicit sequence $\{\hat{p}_t\}_{t=1}^{\infty}$ in such a way that $\sum_{t=1}^{\infty} |p_t - \hat{p}_t|$ $\big($and hence$\|\pi - \hat{\pi}\|\big)$

is small with high probability. Armed with the numbers $\{\hat{p}_t\}_{t=1}^{\infty}$ and the ability to simulate from Q_t, we can make iid draws from $\hat{\pi}$. While we cannot say for certain that $\hat{\pi}$ is within γ of π in total variation, taking $X_0 \sim \hat{\pi}$ is a much more rigorous method of doing burn-in than the typical *ad hoc* method which involves an essentially subjective choice of b. Of course, if the Markov chain underlying the MCMC algorithm is very complicated, then establishing a viable minorization condition may be difficult.

The rest of the paper is organized as follows. The mixture representation of π is developed in Section 2. The approximation, $\hat{\pi}$, is described in Section 3. In Section 4, we consider a toy example where π is a known univariate exponential distribution. This enables us to compare our results with the truth. In Section 5, we apply our method to a practically relevant MCMC algorithm. Finally, Section 6 contains some discussion about the strengths and limitations of our method.

2 The mixture representation of π

Let $X = \{X_n\}_{n=0}^{\infty}$ be a Markov chain on a general state space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ with Markov transition kernel $P(x, dy)$. Let $P^n(x, dy)$ denote the n-step Markov transition kernel corresponding to P; that is, for $i \in \{0, 1, 2, \dots\}$, $x \in \mathsf{X}$ and a measurable set B, $P^n(x, B) = \Pr(X_{n+i} \in B | X_i = x)$. We assume throughout that X is π-irreducible and positive Harris recurrent where π is the invariant probability measure. Many Markov chains that are the basis of an MCMC algorithm satisfy these basic properties.

Our main additional assumption is that X satisfies a one-step minorization condition; that is, we assume that we have a function $s : \mathsf{X} \to [0, 1]$ satisfying $\int_{\mathsf{X}} s(x)\, \pi(dx) > 0$ and a probability measure ν on $\mathcal{B}(\mathsf{X})$ such that for all $x \in \mathsf{X}$ and all measurable B,

$$P(x, B) \geq s(x)\, \nu(B)\ . \tag{2}$$

This is a more general minorization condition than the one considered in Hobert and Robert (2004). Indeed, these authors assume that $s(x)$ has the specific form $\varepsilon I_C(x)$. There are practical advantages to working with the more general form of minorization (Jones and Hobert, 2001).

While π-irreducibility and positive Harris recurrence do not together imply the existence of a one-step minorization condition, they do imply that a k-step minorization holds; that is, they guarantee the existence of a $k \in \mathbb{N} := \{1, 2, \dots\}$ such that $P^k(x, \cdot) \geq s(x)\, \nu(\cdot)$ where s and ν are as described above. For a given chain, if it is not possible to establish (2), but $P^k(x, \cdot) \geq s(x)\, \nu(\cdot)$ can be established for some $k \in \{2, 3, 4, \dots\}$, then we simply consider the Markov chain corresponding to P^k to be the chain of interest. This is legitimate since π is still invariant for P^k and the k-step chain inherits the basic properties from X. On the other hand, in most standard MCMC settings, P^k will not be available in closed form and this will make the application of our methods more challenging.

It is often straightforward to establish (2). It is especially simple when X is countable since we can just fix a point $\tilde{x} \in \mathsf{X}$ and take $s(x) = I(x = \tilde{x})$ and $\nu(\cdot) = P(\tilde{x}, \cdot)$. Mykland,

Tierney and Yu (1995) describe general methods for establishing (2) in the context of standard MCMC algorithms such as the Gibbs sampler and independence and random walk versions of the Metropolis-Hastings-Green algorithm. We also note that the simulated tempering method of Geyer and Thompson (1995) and Marinari and Parisi (1992) often naturally induces a minorization (see, e.g., Brooks, Fan and Rosenthal, 2004; Møller and Nicholls, 2005). See also Brockwell and Kadane (n.d.).

The minorization allows for the fundamental *splitting construction* of Nummelin (1978, 1984). Specifically, we can use (2) to write $P(x, \cdot)$ as a two-component mixture

$$P(x, dy) = s(x)\,\nu(dy) + [1 - s(x)]\,R(x, dy)\ , \tag{3}$$

where $R(x, dy) := [1 - s(x)]^{-1}[P(x, dy) - s(x)\,\nu(dy)]$ is called the *residual measure*; define $R(x, dy)$ to be 0 if $s(x) = 1$. If X is the basis of an MCMC algorithm, then presumably there is a convenient method of simulating from $P(x, \cdot)$. The mixture representation (3) provides the following alternative method: given $X_n = x$, generate $\delta_n \sim \mathcal{B}er(s(x))$. If $\delta_n = 1$, then draw X_{n+1} from $\nu(\cdot)$, else draw X_{n+1} from $R(x, \cdot)$. In fact, this is a recipe for simulating the *split* chain, $X' = \{(X_n, \delta_n)\}_{n=0}^{\infty}$, which lives on the space $\mathsf{X} \times \{0, 1\}$ and is such that, marginally, the sequence $\{X_n\}_{n=0}^{\infty}$ has the same distribution as the original chain, X. An important property of X' is that $\mathsf{X} \times \{1\}$ is an *accessible atom* and the (random) times at which X' enters $\mathsf{X} \times \{1\}$ are *regeneration times* when the chain stochastically restarts; i.e., the next value has distribution ν. (See Nummelin (1984, Section 4.4) for a thorough development of X' including expressions for its transition kernel and stationary distribution.)

As a practical matter, simulating the split chain in the manner described above may be troublesome since drawing from $R(x, dy)$ can be prohibitively difficult. However, there is a simple method for avoiding this. Specifically, Mykland et al. (1995) suggest simulating from the distribution of $X_{i+1}|X_i$ using the sampler at hand and then "filling in" δ_i by simulating from the distribution of $\delta_i|X_i, X_{i+1}$ with

$$\Pr(\delta_i = 1 \mid X_i, X_{i+1}) = \frac{s(X_i)q(X_{i+1})}{k(X_{i+1}|X_i)} \tag{4}$$

where $q(\cdot)$ and $k(\cdot|x)$ are densities corresponding to $\nu(\cdot)$ and $P(x, \cdot)$. We use this approach in our examples. The development of the split chain is now used to derive (1).

Define τ to be the first return time to the atom; that is,

$$\tau = \min\left\{n \geq 1 : (X_n, \delta_n) \in \mathsf{X} \times \{1\}\right\}\ .$$

Also, let $\Pr^*(\cdot)$ and $\mathrm{E}^*(\cdot)$ denote probability and expectation conditional on $\delta_0 = 1$ (with X_0 chosen arbitrarily); i.e., $X_1 \sim \nu(\cdot)$. Since X' is positive recurrent, it follows that $\mathrm{E}^*(\tau) < \infty$ (Meyn and Tweedie, 1993, Chapter 10). Consequently, we can define a discrete random variable, T, with support $\mathbb{N}$ and probabilities given by

$$p_t = \frac{\Pr^*(\tau \geq t)}{\mathrm{E}^*(\tau)}\ . \tag{5}$$

It is important to recognize that, in general, $p_t \neq \Pr^*(\tau = t)$.

Remark 1. *The random variable T is related to the (discrete) delayed renewal process $S_n = \sum_{i=0}^{n} Y_i$ where $Y_1, Y_2, \ldots$ are iid copies of τ and Y_0 is an independent, nonnegative discrete random variable. Indeed, if $Y_0 \stackrel{d}{=} T - 1$, then S_n is an equilibrium renewal process; i.e., the equilibrium distribution is that of $T - 1$ (Ross, 1983, p.76).*

Now, for any $t \in \mathbb{N}$ and any measurable B, we define

$$Q_t(B) = \Pr^*(X_t \in B | \tau \geq t) \; ; \tag{6}$$

i.e., Q_t is the conditional distribution of X_t given that $(X_0, \delta_0) \in \mathsf{X} \times \{1\}$ *and* that there are no regenerations in the split chain before time t. We now formally state an extension of Hobert and Robert's (2004) Theorem 1.

Theorem 1. *Let X be a Markov chain on a general state space $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ with Markov transition kernel P. Assume that X is π-irreducible and positive Harris recurrent where π is the invariant probability measure. Assume further that* (2) *holds. Then for any $B \in \mathcal{B}(\mathsf{X})$, we have*

$$\pi(B) = \sum_{t=1}^{\infty} Q_t(B)\, p_t \; , \tag{7}$$

where p_t and Q_t are defined in terms of the split chain at (5) *and* (6)*.*

Proof. See the proof of Hobert and Robert's (2004) Theorem 1, which still goes through with the more general minorization condition. □

Remark 2. *Asmussen et al. (1992) provide general methods for constructing stationary versions of certain regenerative stochastic processes. Representation* (7) *can also be obtained by applying their methods to the split chain, which possesses the necessary regenerative properties.*

Equation (7) demonstrates the possibility of simulating a random variable from π using a sequential sampling mechanism. That is, a draw from π can be made by first drawing T, call the result t, and then making an independent draw from Q_t. In fact, it is always possible to simulate from Q_t using a simple accept-reject algorithm that we call Algorithm I. All that is required is the ability to simulate the split chain. Note that $Q_1(\cdot) \equiv \nu(\cdot)$ so in the statement of the algorithm, it is assumed that $t \geq 2$.

Algorithm I:

1. `Take` $(x_0, \delta_0) \in \mathsf{X} \times \{1\}$ `and simulate the split chain for` t `iterations.`
2. `If` $\delta_1 = \cdots = \delta_{t-1} = 0$`, then take` x_t`; otherwise, repeat step 1.`

Assume now that $s(x)$ in (2) is not bounded away from zero. This will be the case whenever the Markov chain X is not uniformly ergodic. There is no known method for

simulating T in this case and hence we focus on using (7) to build an approximation of π from which it is straightforward to sample. Our approximation takes the form

$$\hat{\pi}(\cdot) = \sum_{t=1}^{\infty} Q_t(\cdot)\, \hat{p}_t \ ,$$

where $\{\hat{p}_t\}_{t=1}^{\infty}$ are nonnegative numbers that sum to one. It is easy to show that $\|\pi - \hat{\pi}\| \leq \sum_{t=1}^{\infty} |p_t - \hat{p}_t|$; that is, the total variation distance between the distributions π and $\hat{\pi}$ is bounded above by twice the total variation distance between the distributions of T and $\hat{T}$, where $\hat{T}$ is the discrete random variable on $\mathbb{N}$ with probabilities $\{\hat{p}_t\}_{t=1}^{\infty}$. (Note that we are using the version of total variation which does not have the factor of 2; that is, $\|\pi - \hat{\pi}\|$ is defined to be the supremum over measurable B of $|\pi(B) - \hat{\pi}(B)|$.) In the next section we will show that given any $\gamma > 0$, it is possible to construct a sequence $\{\hat{p}_t\}_{t=1}^{\infty}$ such that $\sum_{t=1}^{\infty} |p_t - \hat{p}_t| < \gamma$ with high probability..

3 Approximating π

The key to our argument is that making iid draws from the distribution of τ is trivial; just take $X_1 \sim \nu(\cdot)$, run the split chain, and count how many iterations until the first regeneration. This unlimited supply of iid copies of τ can be used to construct a statistical estimate of $p_t = \Pr^*(\tau \geq t)/\mathrm{E}^*(\tau)$. Indeed, let $\tau_1, \ldots, \tau_m$ denote a random sample of size m and let $F_m(t)$ denote the corresponding empirical distribution function. We estimate p_t with

$$\hat{p}_t = \frac{1 - F_m(t-1)}{\overline{\tau}} \tag{8}$$

where $\overline{\tau}$ is the sample mean. Since $\sum_{t=1}^{\infty} \hat{p}_t = 1$, $\{\hat{p}_t\}_{t=1}^{\infty}$ is a legitimate mass function on $\mathbb{N}$ from which we can sample.

We now use asymptotic arguments to show that $\{\hat{p}_t\}_{t=1}^{\infty}$ enjoys a type of "strong consistency" and to get a handle on the error of $\{\hat{p}_t\}_{t=1}^{\infty}$. These results allow us to develop a method of choosing an appropriate value for m. For obvious reasons, we use $\sum_{t=1}^{\infty} |\hat{p}_t - p_t|$ as our measure of error. Let G_1 and G_2 denote two univariate distribution functions. The L_1-*Wasserstein distance* between the probability distributions corresponding to G_1 and G_2 is defined as (Shorack and Wellner, 1986, Chapter 2)

$$d_1(G_1, G_2) = \int_{-\infty}^{\infty} |G_1(x) - G_2(x)|\, dx.$$

The following result shows that (at least asymptotically) $\{\hat{p}_t\}_{t=1}^{\infty}$ is a reasonable estimate of the mass function of T.

Theorem 2. *For $\{\hat{p}_t\}_{t=1}^{\infty}$ as defined above, we have*

$$\sum_{t=1}^{\infty} |p_t - \hat{p}_t| \leq 2 d_1(F_m, F).$$

Hence, $\sum_{t=1}^{\infty} |p_t - \hat{p}_t| \to 0$ a.s. as $m \to \infty$.

Proof. First

$$\begin{aligned}
|\hat{p}_t - p_t| &= \left| \frac{1 - F_m(t-1)}{\overline{\tau}} \pm \frac{1 - F(t-1)}{\overline{\tau}} - \frac{1 - F(t-1)}{\mathrm{E}^*(\tau)} \right| \\
&\leq \frac{|F_m(t-1) - F(t-1)|}{\overline{\tau}} + \frac{[1 - F(t-1)]\,|\overline{\tau} - \mathrm{E}^*(\tau)|}{\overline{\tau}\mathrm{E}^*(\tau)} \\
&\leq |F_m(t-1) - F(t-1)| + \frac{[1 - F(t-1)]\,|\overline{\tau} - \mathrm{E}^*(\tau)|}{\mathrm{E}^*(\tau)} \\
&\leq |F_m(t-1) - F(t-1)| + \frac{[1 - F(t-1)]\sum_{s=1}^{\infty} |F_m(s) - F(s)|}{\mathrm{E}^*(\tau)}
\end{aligned}$$

and hence

$$\sum_{t=1}^{\infty} |\hat{p}_t - p_t| \leq 2 \sum_{t=1}^{\infty} |F_m(t) - F(t)| = 2 \int_{-\infty}^{\infty} |F_m(t) - F(t)|\, dt = 2 d_1(F_m, F).$$

Finally, the fact that $\mathrm{E}^*(\tau) < \infty$ implies that $d_1(F_m, F) \to 0$ a.s. as $m \to \infty$ (Shorack and Wellner, 1986, p. 65). □

Obviously, no matter how large m is, we can never say for certain that $d_1(F_m, F) < \gamma$. However, we can use asymptotic results to make statements like $\Pr[d_1(F_m, F) < \gamma] \approx 1 - \alpha$. Indeed, Del Barrio, Gine and Matran (1999) have recently described the first-order asymptotics for the L_1-Wasserstein distance between the empirical and true distribution functions. In particular, their results imply that if

$$\sum_{t=1}^{\infty} \sqrt{\Pr^*(\tau \geq t)} < \infty\,, \tag{9}$$

then

$$\sqrt{m}\; d_1(F_m, F) \xrightarrow{d} \sum_{t=1}^{\infty} |B(F(t))| \tag{10}$$

where $B(s)$, $0 \leq s \leq 1$, denotes a Brownian bridge process. Condition (9) is very close to a finite second moment condition. Indeed, (9) implies that $\mathrm{E}^*(\tau^2) < \infty$, while if $\mathrm{E}^*(\tau^{2+\varepsilon}) < \infty$ for some $\varepsilon > 0$ then (9) holds.

We now explain how (10) can be used to come up with a reasonable value of m. Suppose that $\tau_1, \ldots, \tau_{m'}$ is an initial sample of iid τ's with corresponding empirical distribution function $F_{m'}$. Let $u_{m'}$ denote the number of unique values in this sample. Also, let L denote the random variable $\sum_{t=1}^{\infty} |B(F_{m'}(t))|$, which, if m' is large, should have a distribution quite similar to that of $\sum_{t=1}^{\infty} |B(F(t))|$. Simulating the random variable L is quite simple. Indeed, all that is required is $u_{m'}$ values of one realization of standard Brownian motion in $(0, 1)$, which can be done sequentially using only univariate normal draws. Hence, it is easy to find c such that

$$\Pr[L < c] \approx 1 - \alpha.$$

Then if we take $m = 4c^2/\gamma^2$, we can say that $\Pr[2 d_1(F_m, F) < \gamma] \approx 1 - \alpha$ and hence that $\|\pi - \hat{\pi}\| < \gamma$ with probability approximately equal to $1 - \alpha$.

If (9) fails then (10) fails (Del Barrio et al., 1999) and our method for choosing m is not applicable. This situation is analogous to one where we have iid random variables $W_1, W_2, \ldots$ such that $\mathrm{E}|W_1|^p$ is finite when $p = 1$ but is infinite when $p = 2$. In this case, $n^{-1}\sum_{i=1}^{n} W_i$ can be used to estimate $\mathrm{E}(W)$ since the strong law holds, but the central limit theorem (CLT) cannot be used to choose an appropriate value of n. On the other hand, the condition (9) is closely related to the mixing properties of the Markov chain and is a weak condition. In fact, if (9) were to fail, the Markov chain would probably not mix sufficiently well to be of any practical use anyway.

Remark 3. *It appears that a weak form of polynomial ergodicity (of X) is enough to guarantee that $E^*\left(\tau^{2+\varepsilon}\right) < \infty$ for some $\varepsilon > 0$ (Jarner and Roberts, 2002), but we are unaware of any clean statements in the literature connecting polynomial ergodicity and the moments of τ. On the other hand, if X is geometrically ergodic, then τ has a moment generating function (see, e.g. Hobert, Jones, Presnell and Rosenthal, 2002).*

In the next two sections, we illustrate the construction of $\hat{\pi}$ with toy and realistic examples, respectively.

4 A toy example

Suppose that $\pi(x) = e^{-x} I_{\mathbb{R}^+}(x)$. This distribution is clearly not intractable in any sense, but using a simple, univariate distribution allows us to evaluate our approximations by comparing them directly to the truth. The Markov chain we consider is the independence Metropolis sampler with an $\mathcal{E}xp(\theta)$ proposal; that is, the proposal density is $q(x) = \theta e^{-\theta x} I_{\mathbb{R}^+}(x)$. The chain evolves as follows: Given $X_n = x$, draw $y \sim \mathcal{E}xp(\theta)$ and independently draw $u \sim \mathcal{U}ni(0,1)$. If $u < \exp\{(x-y)(1-\theta)\}$ then set $X_{n+1} = y$, otherwise set $X_{n+1} = x$. The case $\theta = 1$ is not of interest to us since in this case the algorithm yields iid draws from the target distribution. Results in Mengersen and Tweedie (1996) can be used to show that when $0 < \theta < 1$, the chain is uniformly ergodic and hence $\mathrm{E}^*\left(\tau^{2+\varepsilon}\right) < \infty$ for any $\varepsilon > 0$. Moreover, it is easy to verify the conditions of Theorem 5.3 in Jarner and Roberts (2002) which shows that this sampler is polynomially ergodic, but apparently this is not sufficient to guarantee $\mathrm{E}^*\left(\tau^2\right) < \infty$ when $1 < \theta$.

Finding a minorization condition is simple. Let $w(x) = \theta^{-1} e^{x(\theta-1)}$. Applying results in Mykland et al. (1995, p. 236) shows that (2) is satisfied with

$$s(x) = \left\{ \frac{a}{w(x)} \wedge 1 \right\}$$

and ν having density proportional to

$$q(y) \left\{ \frac{w(y)}{a} \wedge 1 \right\}$$

for any $a > 0$. Mykland et al. also give an expression for the probability of regeneration that does not require the normalizing constant for the density of ν.

We constructed three approximations to π: The first was based on a uniformly ergodic sampler with $\theta = 0.75$; the second was based on a subgeometric sampler with $\theta = 1.5$; and the third used a subgeometric sampler with $\theta = 2.5$. (Actually, we suspect that (9) fails in the $\theta = 2.5$ case.) In all cases, after some trial and error, we chose $a = 1.5$. For each value of θ, an initial sample of $m' = 2.5 \times 10^5$ iid τ's was drawn. The results are reported in Table 1 which gives the number of unique values observed ($u_{m'}$), the maximum value observed (max), and the 99th percentile (99%).

Then, for each value of θ, we simulated 5×10^4 values of L and the results are given in Table 2. In particular, Table 2 gives the number (m) of τ's necessary to ensure that $\hat{\pi}$ is within γ of the stationary distribution in total variation distance with approximate probability $1-\alpha$. The values of m in Table 2 clearly reflect the fact that the sampler enjoys superior mixing for smaller values of θ.

For each value of θ, we constructed $\hat{\pi}$ using the values of m given in Table 2. Then, for each value of θ, we compared a density estimate based on a random sample of size 5×10^4 from $\hat{\pi}$ with the $\mathcal{E}xp(1)$ density and the two curves were essentially coincidental. Figure 1 suggests that $\hat{\pi}$ is an excellent approximation to π even when $\theta = 2.5$.

5 Hierarchical linear mixed models

Consider the usual frequentist general linear mixed model

$$Y = X\beta + Zu + \varepsilon \ ,$$

where Y is an $n \times 1$ vector of observations, X is a known $n \times p$ matrix, Z is a known $n \times q$ matrix, β is a $p \times 1$ vector of parameters, u is a $q \times 1$ vector of random variables, and ε is an $n \times 1$ vector of residual errors. We assume that X is of full column rank so that $X^T X$ is invertible. A Bayesian version of this model may be expressed as the following conditionally independent hierarchical model

$$\begin{aligned} Y|\beta, u, R, D &\sim \ \mathrm{N}_n(X\beta + Zu, R^{-1}) \\ \beta|u, R, D &\sim \ \mathrm{N}_p(\beta_0, B^{-1}) \\ u|D, R &\sim \ \mathrm{N}_q(0, D^{-1}) \end{aligned} \tag{11}$$

with as yet unspecified priors $f(R)$ and $f(D)$. Here β_0 and B^{-1} are assumed to be known. The posterior density of (β, u, R, D) given the data, y, is characterized by

$$\pi(\beta, u, R, D|y) \propto f(y|\beta, u, R, D) f(\beta|u, R, D) f(u|D, R) f(R) f(D) \ . \tag{12}$$

We assume that the priors on R and D are such that the resulting posterior (12) is proper. Even if proper conjugate priors are chosen, the integrals required for inference through this posterior can not be evaluated in closed form. Thus, exploring the posterior in order to make inferences might require MCMC.

5.1 A block Gibbs sampler and a minorization condition

In this section, we consider a block Gibbs sampler with components R, D and $\xi = (u^T, \beta^T)^T$. The full conditional densities for R and D are given by

$$\begin{aligned}\pi(R|\xi, D, y) &= C_R^{-1}(\xi)|R|^{1/2}\exp\{-0.5(y - X\beta - Zu)^T R(y - X\beta - Zu)\}f(R)\\ \pi(D|\xi, R, y) &= C_D^{-1}(\xi)|D|^{1/2}\exp\{-0.5u^T Du\}f(D)\end{aligned}$$

where

$$C_R(\xi) = \int |R|^{1/2}\exp\{-0.5(y - X\beta - Zu)^T R(y - X\beta - Zu)\}f(R)\,dR$$

and

$$C_D(\xi) = \int |D|^{1/2}\exp\{-0.5u^T Du\}f(D)\,dD\ .$$

The density $\pi(\xi|R, D, y)$ is a $(p + q)$-variate normal with mean ξ_0 and covariance matrix Σ^{-1} where

$$\Sigma = \begin{pmatrix} Z^T RZ + D & Z^T RX \\ X^T RZ & X^T RX + B\end{pmatrix} \qquad \text{and} \qquad \Sigma\xi_0 = \begin{pmatrix} Z^T Ry \\ X^T Ry + B\beta_0\end{pmatrix}\ . \tag{13}$$

Consider the block Gibbs sampler corresponding to the following updating scheme:

$$(D', R', \xi')\ \to\ (D, R', \xi')\ \to\ (D, R, \xi')\ \to\ (D, R, \xi)\ .$$

Conditional on ξ, D and R are independent and hence the order in which they are updated is irrelevant. That is, we are effectively dealing with a two-variable Gibbs sampler. Suppressing dependence on the data, the transition density is given by

$$k(D, R, \xi|D', R', \xi') = \pi(D|\xi')\,\pi(R|\xi')\,\pi(\xi|R, D)\ .$$

We now develop a minorization condition of the form (2) for this block Gibbs sampler. Fix a point $\tilde{\xi}$ and sets $\mathbb{M}_R \subset \mathbb{R}^{n(n+1)/2}$ and $\mathbb{M}_D \subset \mathbb{R}^{q(q+1)/2}$ so that when $R \in \mathbb{M}_R$ and $D \in \mathbb{M}_D$ we have

$$\begin{aligned}k(D, R, \xi|D', R', \xi') &= \frac{\pi(D|\xi')\pi(R|\xi')}{\pi(D|\tilde{\xi})\pi(R|\tilde{\xi})}\pi(D|\tilde{\xi})\pi(R|\tilde{\xi})\pi(\xi|R, D)\\ &\geq \left[\inf_{R\in\mathbb{M}_R}\frac{\pi(R|\xi')}{\pi(R|\tilde{\xi})}\right]\left[\inf_{D\in\mathbb{M}_D}\frac{\pi(D|\xi')}{\pi(D|\tilde{\xi})}\right]\pi(D|\tilde{\xi})\pi(R|\tilde{\xi})\pi(\xi|R, D)\ .\end{aligned}$$

Then the minorization condition will follow by taking

$$s(\xi', \tilde{\xi}) = c_q\left[\inf_{R\in\mathbb{M}_R}\frac{\pi(R|\xi')}{\pi(R|\tilde{\xi})}\right]\left[\inf_{D\in\mathbb{M}_D}\frac{\pi(D|\xi')}{\pi(D|\tilde{\xi})}\right] \tag{14}$$

and

$$q(D, R, \xi) = c_q^{-1}\pi(D|\tilde{\xi})\pi(R|\tilde{\xi})\pi(\xi|R, D)I(R \in \mathbb{M}_R)I(D \in \mathbb{M}_D)$$

where

$$c_q = \int\int \pi(D|\tilde{\xi})\pi(R|\tilde{\xi})I(R \in \mathbb{M}_R)I(D \in \mathbb{M}_D)\, dR\, dD\ .$$

Let S denote the space in which R lives; that is, the set of points in $\mathbb{R}^{n(n+1)/2}$ corresponding to symmetric, positive definite $n \times n$ matrices. Note that $\mathbb{M}_R$ must be chosen so that $\mathbb{M}_R \cap S$ has positive measure. Otherwise, c_q will be zero and, from a practical standpoint, R will never land in $\mathbb{M}_R$. Similar comments apply to the choice of $\mathbb{M}_D$.

Using equation (4), it is easy to see that when $R \in \mathbb{M}_R$ and $D \in \mathbb{M}_D$ the probability of regeneration is given by

$$\Pr(\delta = 1|D', R', \xi', D, R, \xi) = \left[\inf_{R\in\mathbb{M}_R} \frac{\pi(R|\xi')}{\pi(R|\tilde{\xi})}\right]\left[\inf_{D\in\mathbb{M}_D} \frac{\pi(D|\xi')}{\pi(D|\tilde{\xi})}\right]\frac{\pi(D|\tilde{\xi})\pi(R|\tilde{\xi})}{\pi(D|\xi')\pi(R|\xi')}\ . \qquad (15)$$

Thus we have to calculate the infima in (14) and plug into (15). Let $a_{1ij} \le a_{2ij}$ for $i = 1, \ldots, q$ and $j = 1, \ldots, q$ be constants and define $\mathbb{M}_D = \{M_{q\times q} : a_{1ij} \le m_{ij} \le a_{2ij}\}$. Then

$$\begin{aligned}
\inf_{D\in\mathbb{M}_D} \frac{\pi(D|\xi')}{\pi(D|\tilde{\xi})} &= \frac{C_D(\tilde{\xi})}{C_D(\xi')} \inf_{D\in\mathbb{M}_D} \frac{\exp\{-0.5u'^T D u'\}}{\exp\{-0.5\tilde{u}^T D\tilde{u}\}} \\
&= \frac{C_D(\tilde{\xi})}{C_D(\xi')} \inf_{D\in\mathbb{M}_D} \exp\left\{-0.5\sum_i\sum_j (u'_i u'_j - \tilde{u}_i\tilde{u}_j)d_{ij}\right\} \\
&= \frac{C_D(\tilde{\xi})}{C_D(\xi')} \exp\left\{-0.5\sum_i\sum_j (u'_i u'_j - \tilde{u}_i\tilde{u}_j)g_{ij}\right\} \\
&= \frac{C_D(\tilde{\xi})}{C_D(\xi')} g(u', \tilde{u})
\end{aligned}$$

where

$$g_{ij} = \begin{cases} a_{1ij} & \text{if } u'_i u'_j - \tilde{u}_i\tilde{u}_j \le 0, \\ a_{2ij} & \text{if } u'_i u'_j - \tilde{u}_i\tilde{u}_j > 0. \end{cases}$$

Let $v' = y - X\beta' - Zu'$ and $\tilde{v} = y - X\tilde{\beta} - Z\tilde{u}$. Also, let $b_{1ij} \le b_{2ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n$ be constants and define $\mathbb{M}_R = \{M_{n\times n} : b_{1ij} \le m_{ij} \le b_{2ij}\}$. A calculation similar to the one above shows that

$$\inf_{R\in\mathbb{M}_R} \frac{\pi(R|\xi')}{\pi(R|\tilde{\xi})} = \frac{C_R(\tilde{\xi})}{C_R(\xi')} h(v', \tilde{v})$$

where

$$h_{ij} = \begin{cases} b_{1ij} & \text{if } v'_i v'_j - \tilde{v}_i\tilde{v}_j \le 0, \\ b_{2ij} & \text{if } v'_i v'_j - \tilde{v}_i\tilde{v}_j > 0. \end{cases}$$

Thus the probability of regeneration is given by

$$\Pr(\delta = 1|D', R', \xi', D, R, \xi) = g(u', \tilde{u})h(v', \tilde{v}) \exp\left\{-0.5[(\tilde{u}^T D\tilde{u} - u'^T D u') + (\tilde{v}^T R\tilde{v} - v'^T R v')]\right\}\ .$$

We end this section by noting that the minorization condition established above is quite different than the one derived in Jones and Hobert (2004) for the simpler Bayesian hierarchical version of the one-way random effects model. One major difference is that the one-way model contains only two univariate variance components, whereas the general linear mixed model considered here contains two unstructured covariance matrices. Moreover, Jones and Hobert (2004) established a minorization condition of the form $P(x,\cdot) \geq \varepsilon\, I_C(x)\, \nu(\cdot)$ while the one we have derived here is of the form $P(x,\cdot) \geq s(x)\, \nu(\cdot)$ with an s that cannot be expressed as a constant multiple of an indicator.

5.2 A numerical example

In this subsection, we identify a specific example of the model (11), simulate some data from that model and then use the block Gibbs sampler described above to form an approximation of the resulting intractable posterior density.

Suppose that $p = 1$ so that $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ and that $q = n$ with $Z = I_n$. Fix $\beta_0 = 0$ and $B^{-1} = 1$. Assume that $R^{-1} = \lambda_R^{-1} I_n$ and $D^{-1} = \lambda_D^{-1} I_n$ where λ_R^{-1} and λ_D^{-1} are scalar variance components whose reciprocals are assigned the following conjugate priors

$$\lambda_R \sim \mathcal{G}am(r_1, r_2) \qquad \text{and} \qquad \lambda_D \sim \mathcal{G}am(d_1, d_2)\,.$$

Set $\xi = (u^T, \beta)^T$ and $\lambda = (\lambda_D, \lambda_R)^T$. We simulated data according to this model with $n = 6$, $r_1 = r_2 = d_1 = d_2 = 1$ and the vector of covariates drawn from the $\mathrm{N}(0, I_6)$ distribution. The resulting (x, y) pairs were: $(-0.65201, 3.05577)$, $(0.46053, -0.84096)$, $(-0.39088, -3.21066)$, $(-0.64953, -0.47085)$, $(-0.65276, 2.23286)$, $(0.75399, -0.02815)$. We will construct $\hat{\pi}$ corresponding to the posterior that results from these data and from setting $r_1 = d_1 = 1$ and $r_2 = d_2 = 2$. Recall that the block Gibbs sampler from the previous section uses the sampling scheme: $(\lambda', \xi') \to (\lambda, \xi') \to (\lambda, \xi)$. The full conditionals for the precision parameters are given by

$$\lambda_R|\xi, y \sim \mathcal{G}am\left(4\,,\, 2 + \frac{1}{2}(y - X\beta - u)^T(y - X\beta - u)\right)\,,$$
$$\lambda_D|\xi, y \sim \mathcal{G}am\left(4\,,\, 2 + \frac{1}{2}u^T u\right)\,.$$

Now $\xi|\lambda_R, \lambda_D, y \sim \mathrm{N}_7(\xi_0, \Sigma^{-1})$ where

$$\Sigma = \begin{pmatrix} (\lambda_R + \lambda_D) I_6 & \lambda_R X \\ \lambda_R X^T & 1 + \lambda_R X^T X \end{pmatrix} \qquad \text{and} \qquad \Sigma\xi_0 = \lambda_R \begin{pmatrix} y \\ X^T y \end{pmatrix}\,.$$

Routine calculations show that $\Sigma = LL^T$ where

$$L^{-1} = \begin{pmatrix} a^{-1} I_6 & 0 \\ -b(ac)^{-1} X^T & c^{-1} \end{pmatrix}\,,$$

and $a = \sqrt{\lambda_R + \lambda_D}\,$, $b = \lambda_R / a$ and $c = \sqrt{1 + (\lambda_R \lambda_D / a^2) X^T X}$.

While some work has been done analyzing block Gibbs samplers for simpler hierarchical linear models (Hobert and Geyer, 1998; Jones and Hobert, 2004; Rosenthal, 1995b), none of these results apply to our block Gibbs sampler. That is, little is known about the mixing properties of our Markov chain. We simply assume that it satisfies $\mathrm{E}^*\left(\tau^{2+\varepsilon}\right) < \infty$ for some $\varepsilon > 0$.

To use the minorization condition developed in the previous subsection, we must fix a point $\tilde{\xi}$ and sets $\mathbb{M}_D = [a_1, a_2]$ and $\mathbb{M}_R = [b_1, b_2]$ where $0 < a_1 < a_2$ and $0 < b_1 < b_2$. To this end, we ran the block Gibbs sampler for 1×10^4 iterations starting from $\xi = \bar{y}1$ where 1 is a vector of ones. Let $\tilde{u}_1, \ldots \tilde{u}_6, \tilde{\beta}, \tilde{\lambda}_D, \tilde{\lambda}_R$ be the estimated posterior expectations of the associated parameters. We set $\tilde{\xi} = (\tilde{u}_1, \ldots \tilde{u}_6, \tilde{\beta})^T$, $[a_1, a_2] = \tilde{\lambda}_D \pm w\, s_{\lambda_D}$ and $[b_1, b_2] = \tilde{\lambda}_R \pm w\, s_{\lambda_R}$ where $w > 0$ and s_{λ_D}, s_{λ_R} are the usual sample standard deviations of the sample of λ_D's and λ_R's, respectively. Note that the choice of w controls the trade-off between the size of $\mathbb{M}_D$ and $\mathbb{M}_R$ and the magnitude of the probability of regeneration. In our example, we used $w = 1.5$.

We simulated an initial sample of $m' = 1 \times 10^4$ iid τs. The number of unique values in the sample was 209, the maximum was 368 and the 99th percentile was 155. We then simulated 1×10^4 values of L. Using this sample together with the formula from Section 3 leads to the conclusion that a random sample of size 1,623,331 τ's is necessary to ensure that $\hat{\pi}$ is within 0.10 of the stationary distribution in total variation distance with approximate probability 0.90. (The value of c was 63.7.) Using these results we constructed $\hat{\pi}$ and subsequently simulated 1×10^4 iid draws from it and estimated the marginal density functions of λ_D, λ_R and β using the `density` command available in the R software package (R Development Core Team, 2004). This density estimates are shown in Figure 1. Unlike the toy example studied in the previous section, here we cannot compare our density estimate with the truth since we do not have the marginal posterior density at our disposal. Instead, we ran 1×10^4 independent Gibbs chains each started from $\xi = \bar{y}1$ for 1×10^4 iterations and collected the last state from each chain to form an iid sample. We then used this iid sample to estimate the marginal posterior densities and these estimates are also shown in Figure 1. (Our experience analyzing similar Markov chains suggests that the block Gibbs sampler is probably quite close to stationarity after 1×10^4 iterations.) Note that the two density estimates for each parameter are nearly coincidental.

Suppose that t is a "large" integer and consider using Algorithm I to get a draw from Q_t. Algorithm I is successful; that is, returns a draw from Q_t only if $\delta_1 = \cdots = \delta_{t-1} = 0$. Thus, we must run the split chain over and over again until we get a realization in which there are no regenerations before time t. It may seem as if this could take an impractically large amount of time. However, if large values of T are observed, this suggests that the mass function of T has a heavy tail, which in turn suggests that the split chain is prone to long stretches without a regeneration. Hence, it is not unlikely for a long consecutive string of δs to be 0 and this means that Algorithm I will be viable. As an illustration, consider the example described above. The largest value of T that was observed while simulating the random sample of size 1×10^4 from $\hat{\pi}$ was 333. We performed an experiment in which we used Algorithm I to get 10 draws from Q_{333} and the number of iterations of Algorithm

I that were required ranged from 656 to 19,995 and the entire experiment took only about 15 minutes on a slow workstation.

6 Discussion

Let $P(x, dy)$ be a Markov transition kernel with invariant probability measure π that satisfies the minorization condition $P(x, \cdot) \geq s(x)\,\nu(\cdot)$. We have shown how to use simulations from the corresponding split chain to build a strongly consistent statistical approximation to π from which it is easy to sample. Furthermore, we have shown how to take advantage of asymptotic results (that hold under a weak condition on P) to construct the approximation, $\hat{\pi}$, in such a way that $\|\pi - \hat{\pi}\|$ is small with high (asymptotic) probability.

In our view, an important practical use for $\hat{\pi}$ is as a starting distribution for the original Markov chain. Suppose it is necessary to make sure the that chain is close to stationarity before sampling begins. The best way to accomplish this is via the exact methods mentioned in Section 1 which yield a b such that $\|P^b(x, \cdot) - \pi(\cdot)\| < \gamma$. Unfortunately, the exact methods cannot even be implemented until both drift and minorization conditions have been established for the underlying Markov chain. Our method is less rigorous than the exact method, but it requires far less analysis of the underlying chain. That said, most users of MCMC would not be willing to develop a viable minorization condition just to get a reasonable starting value. On the other hand, as we mention in Section 2, general minorization conditions are already available for many standard MCMC algorithms. Also, it is possible to develop minorization conditions for very general models as in Section 5.1.

While our method is less rigorous than the exact method, it has a much firmer theoretical grounding than most *ad hoc* convergence diagnostics that involve running the chain and observing the behavior of some univariate statistics. In fact, in some cases, our approximation can be used to improve convergence diagnostics. For example, the widely-used Gelman and Rubin (1992) diagnostic involves running independent (parallel) Markov chains whose starting points are drawn from an "overdispersed starting distribution." This starting distribution is based on an approximation of π that is a mixture of multivariate normal distributions whose components are centered at the modes of π. Finding the modes of π requires some potentially tedious and time-consuming numerical analysis and, if the target distribution is complex and high dimensional, there is no guarantee of finding all of the important modes. Our approximation to π is an attractive alternative to Gelman and Rubin's (1992) mixture of normals since it requires no direct numerical analysis of π.

It is important to recognize that our method (and more generally burn-in) is not a way to "fix" a poorly mixing Markov chain. Indeed, such chains are not very useful even when started at stationarity. In particular, for chains with good mixing properties, regardless of the starting distribution, $n^{-1}\sum_{i=0}^{n-1} g(X_i)$ converges almost surely to $\int_{\mathsf{X}} g(x)\,\pi(dx)$ *and* there is a corresponding CLT that can be used to assess Monte Carlo error (Jones, 2004). Unfortunately, the CLT fails to hold when the chain converges too slowly. Moreover, it is well known that if the CLT holds for *any* initial distribution then it holds for *all* initial

distributions (Meyn and Tweedie, 1993, Proposition 17.1.6). Thus, starting a poorly mixing chain at stationarity cannot help it to enjoy a CLT. Generally speaking, a non-stationary chain with good mixing properties is much more useful than a stationary version of a poorly mixing chain.

Although we have not emphasized it, $\hat{\pi}$ can also be used to visualize important features of π. The existing MCMC methods for such visualization (Geyer, 1994, 1996; Henderson and Glynn, 2001; Sköld and Roberts, 2003) have problems that are partly due to the fact that they are based on dependent data. For sufficiently small α and γ our method will produce iid samples from a distribution that is close to π, thus circumventing any problems due to the use of dependent data. Of course, using a small α and γ will likely require substantial computational resources.

Acknowledgments

The authors are grateful to three anonymous reviewers for helpful comments and suggestions. Hobert's research was partially supported by NSF Grant DMS-00-72827.

References

Asmussen, S., Glynn, P. W. & Thorisson, H. (1992). Stationarity detection in the initial transient problem. *ACM Trans. Model. Comput. Simul.* **2**, 130–157.

Athreya, K. B. & Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.* **245**, 493–501.

Baxendale, P. H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.*, to appear.

Breyer, L. A. & Roberts, G. O. (2001). Catalytic perfect simulation. *Methodol. Comput. Appl. Probab.* **3**, 161–177.

Brockwell, A. E. & Kadane, J. B. (n.d.). Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *J. Comput. and Graph. Statist*, to appear.

Brooks, S. P., Fan, Y. & Rosenthal, J. S. (2004). Perfect forward simulation via simulated tempering. *Technical report*. University of Cambridge.

Del Barrio, E., Gine, E. & Matran, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.* **27**, 1009–1071.

Douc, R., Moulines, E. & Rosenthal, J. S. (2004). Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Ann. Appl. Probab.* **14**, 1643–1665.

Fill, J. A., Machida, M., Murdoch, D. J. & Rosenthal, J. S. (2000). Extension of Fill's perfect rejection sampling algorithm to general chains. *Random Structures and Algorithms* **17**, 290–316.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7**, 457–472.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. Roy. Statist. Soc. Ser. B* **56**, 261–274.

Geyer, C. J. (1996). Estimation and optimization of functions. *in* W. R. Gilks, S. Richardson & D. J. E. Spiegelhalter (eds), *Markov chain Monte Carlo in practice*. Chapman & Hall. Boca Raton. pp. 241–258.

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

Henderson, S. G. & Glynn, P. W. (2001). Computing densities for Markov chains via simulation. *Math. Oper. Res.* **26**, 375–400.

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

Hobert, J. P. & Robert, C. P. (2004). A mixture representation of π with applications in Markov chain Monte Carlo and perfect sampling. *Ann. Appl. Probab.* **14**, 1295–1305.

Hobert, J. P., Jones, G. L., Presnell, B. & Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika* **89**, 731–743.

Jarner, S. F. & Roberts, G. O. (2002). Polynomial convergence rates of Markov chains. *Ann. Appl. Probab.* **12**, 224–47.

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys* **1**, 299–320.

Jones, G. L. & Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.* **16**, 312–334.

Jones, G. L. & Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *Ann. Statist.* **32**, 784–817.

Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. Springer. New York.

Marinari, E. & Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* **19**, 451–458.

Mengersen, K. & Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24**, 101–121.

Meyn, S. P. & Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Springer-Verlag. London.

Meyn, S. P. & Tweedie, R. L. (1994). Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.* **4**, 981–1011.

Møller, J. & Nicholls, G. K. (2005). Perfect simulation for sample-based inference. *Stat. Comput.*, to appear.

Murdoch, D. J. & Green, P. J. (1998). Exact sampling from a continuous state space. *Scand. J. Statist.* **25**, 483–502.

Mykland, P., Tierney, L. & Yu, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.* **90**, 233–241.

Nummelin, E. (1978). A splitting technique for Harris recurrent Markov chains. *Z. Wahr. Verw. Geb.* **43**, 309–318.

Nummelin, E. (1984). *General irreducible Markov chains and non-negative operators*. Cambridge University Press. London.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-00-3.

Robert, C. P. & Casella, G. (2004). *Monte Carlo statistical methods*. 2nd edn. Springer. New York.

Roberts, G. O. & Tweedie, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Process. Appl.* **80**, 211–229.

Rosenthal, J. S. (1995a). Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **90**, 558–566.

Rosenthal, J. S. (1995b). Rates of convergence for Gibbs sampling for variance component models. *Ann. Statist.* **23**, 740–761.

Ross, S. (1983). *Stochastic processes*. John Wiley and Sons. New York.

Shorack, G. R. & Wellner, J. A. (1986). *Empirical processes with applications to statistics*. John Wiley and Sons. New York.

Sköld, M. & Roberts, G. O. (2003). Density estimation for the Metropolis-Hastings algorithm. *Scand. J. Statist.* **31**, 699 – 718.

Wilson, D. B. (2000). How to couple from the past using a read-once source of randomness. *Random Structures Algorithms* **16**, 85–113.

Table 1: Initial Sample Results

θ	$u_{m'}$	max	99%
0.75	11	11	5
1.5	48	141	9
2.5	193	2472	16

Table 2: Results from Simulating L

θ	α	c	γ	m
0.75	0.10	4.80	0.10	9.22×10^3
1.5	0.10	23.17	0.10	2.15×10^5
2.5	0.10	95.52	0.10	3.65×10^6

The $\mathcal{E}xp(1)$ Density and an Estimate Based on $\hat{\pi}$

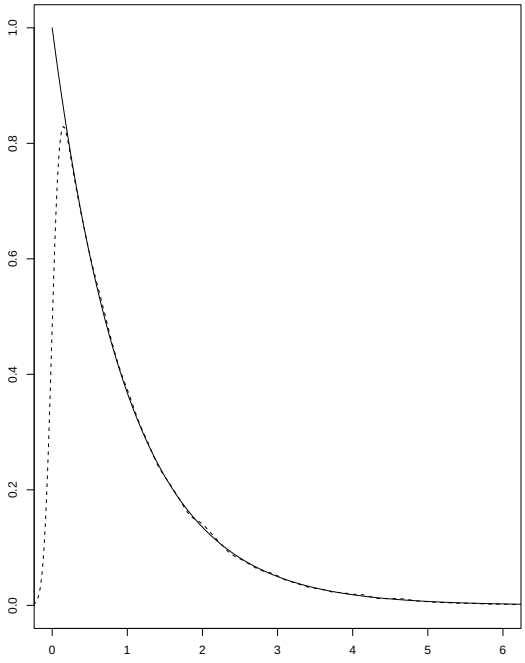

Figure 1: The solid line is the $\mathcal{E}xp(1)$ density and the dashed line is a density estimate based on a random sample of size 5×10^4 from $\hat{\pi}$ when $\theta = 2.5$. The density estimate was made using the `density` function in R with the default settings.

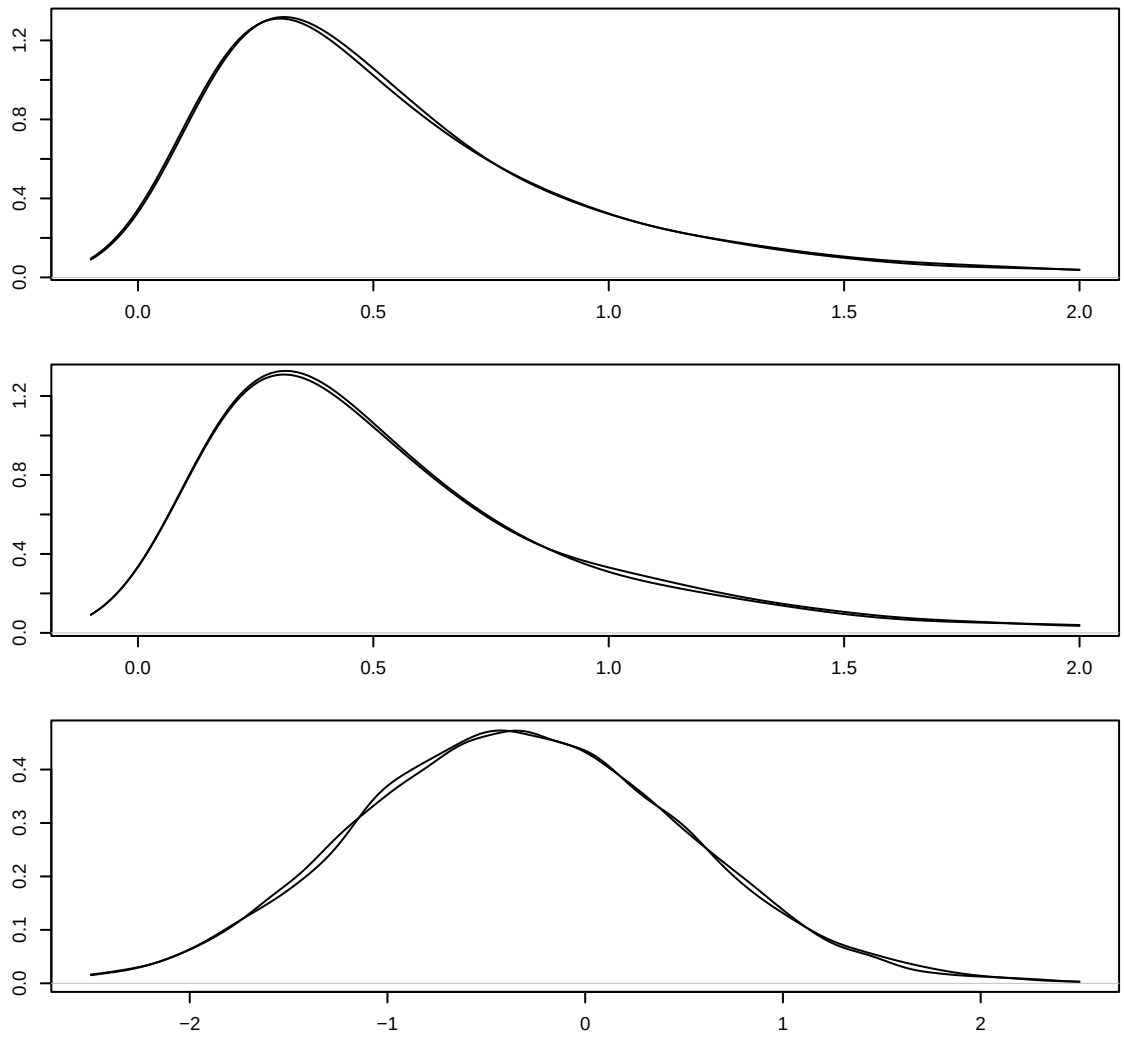

Figure 2: From top to bottom, the three plots correspond to λ_D, λ_R and β. In each plot, the solid and dashed lines are density estimates constructed using samples from $\hat{\pi}$ and the Gibbs sampler, respectively. Each density estimate is based on a random sample of size 1×10^4. The random sample from the Gibbs sampler was constructed by running 1×10^4 independent chains (each started from $\xi = \bar{y}1$) for 1×10^4 iterations and collecting the last state from each chain. All six density estimates were made using the `density` function in R with the bandwidth parameter set to 0.12.