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

James P. Hobert
Department of Statistics
University of Florida
Gainesville, FL, USA
jhobert@stat.ufl.edu

Galin L. Jones
School of Statistics
University of Minnesota
Minneapolis, MN, USA
galin@stat.umn.edu

Christian P. Robert
Université Paris Dauphine
& CREST, INSEE
Paris, France
xian@ceremade.dauphine.fr

September 9, 2003

AMS 2000 *subject classifications.* Primary 62C15; secondary 60J05

Abbreviated title. Approximating an intractable probability distribution

Key words and phrases. Gibbs sampler, Markov chain, Minorization condition, Monte Carlo, Regeneration, Small function, Split chain

Abstract

Suppose that π is an intractable probability distribution. Let $X = \{X_i : i = 0, 1, \dots\}$ be a positive recurrent Markov chain with Markov transition kernel $P(x, dy)$ and invariant probability measure π. Assume that X satisfies a minorization condition of the form $P(x, \cdot) \geq s(x)\, Q(\cdot)$. We provide a method of using simulations from the Markov chain X to construct a statistical estimate of π, call it $\hat{\pi}$, from which it is straightforward to sample. We show that $\hat{\pi}$ is "strongly consistent" in the sense that the total variation distance between π and $\hat{\pi}$, $\|\pi - \hat{\pi}\|$, 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 in order to make a statement like $\Pr[\|\pi - \hat{\pi}\| < 0.1] \approx 0.90$. Draws from $\hat{\pi}$ can be used to approximate features of $\hat{\pi}$ (and hence of π), or they can be used as intelligent starting values for the original Markov chain. We illustrate our methods via two examples.

1 Introduction

Let π be a probability distribution that we would like to explore. Further 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 straightforward 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; 1999).

We begin by generalizing a result of Hobert and Robert (2004) to show that, if P satisfies a one-step *minorization condition* of the form $P(x, \cdot) \geq s(x)\, Q(\cdot)$, then the probability distribution 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 positive numbers that sum to 1. 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 provide an algorithm for making draws from $Q_t(\cdot)$ that requires only slightly more than the ability to simulate X. Unfortunately, making draws from T^* is often prohibitively difficult. Suppose, however, that $\{\hat{p}_t\}_{t=1}^{\infty}$ is another sequence of positive numbers that sum to 1 and consider an approximation to π of the form $\hat{\pi}(A) = \sum_{t=1}^{\infty} Q_t(A)\, \hat{p}_t$. Then the total variation distance, $\|\cdot\|$, between π and $\hat{\pi}$ satisfies

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

The main contribution of this paper is a method of using simulations of the Markov chain, X, to construct $\{\hat{p}_t\}_{t=1}^{\infty}$ in such a way that $\sum_{t=1}^{\infty} |p_t - \hat{p}_t|$ (and hence $\|\pi - \hat{\pi}\|$) is small with high probability. We are able to accomplish this by first finding a simple upper bound on $\sum_{t=1}^{\infty} |p_t - \hat{p}_t|$ and then taking advantage of some recent results in the probability literature regarding the asymptotic properties of the bound.

Armed with the numbers, $\{\hat{p}_t\}_{t=1}^{\infty}$, and the algorithm for simulating from Q_t, we can straightforwardly make independent and identically distributed (iid) draws from $\hat{\pi}$. These draws can be used to estimate or visualize features of π. Another interesting use of $\hat{\pi}$ is as a starting distribution for the original Markov chain; that is, since $\|\pi(\cdot) - \hat{\pi}(\cdot)\|$ is small, we can start the Markov chain X with $X_0 \sim \hat{\pi}$ thereby eliminating the need for *burn-in*.

Let $P^n(x, \cdot)$ represent the distribution of X_n given $X_0 = x$ and suppose that $\varepsilon > 0$. The basic Markov chain theory underlying Markov chain Monte Carlo (MCMC) implies that

there exists an $n^* = n^*(\varepsilon, x)$ such that $\|\pi(\cdot) - P^{n^*}(x, \cdot)\| < \varepsilon$. Furthermore, it is easy to sample from the distribution $P^{n^*}(x, \cdot)$ - just start the chain at x and simulate n^* iterations. Thus, we have an approximation, $P^{n^*}(x, \cdot)$, from which we can make iid draws and that is guaranteed to be within ε of π in total variation. So is $P^{n^*}(x, \cdot)$ a better approximation of π than $\hat{\pi}$? Yes, but finding n^* is much more difficult than constructing $\hat{\pi}$. Indeed, establishing a minorization condition is only a (frequently minor) part of the analysis that is required to find n^* (Roberts and Tweedie; 1999; Rosenthal; 1995a). Moreover, even when it is possible to calculate n^*, it often turns out to be too large to be of any practical value.

The rest of the paper is organized as follows. Some basic Markov chain background material is given in Section 2. The mixture representation of π upon which our approximation is based is described in Section 3. Our method of estimating the sequence $\{p_t\}_{t=1}^{\infty}$ is described in Section 4. In Section 5, we argue that $\hat{\pi}$ can be used to eliminate the need for burn-in. Finally, Sections 6 and 7 contain examples illustrating our methods.

2 Minorization and the Split Chain

Let $X = \{X_i : i = 0, 1, 2, \dots\}$ 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 $\pi(s) := \int_{\mathsf{X}} s(x)\, \pi(dx) > 0$ and a measure ν on $\mathcal{B}(\mathsf{X})$ such that for all $x \in \mathsf{X}$ and all measurable B,

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

Following Nummelin, we call s a *small function* and ν a *small measure*. This set-up is more general than that of Hobert and Robert (2004), who assume that $s(x)$ has the specific form $\varepsilon I_C(x)$ where $\varepsilon > 0$ and $C \in \mathcal{B}(\mathsf{X})$. There are often practical advantages to working with the more general minorization (c.f., Jones and Hobert; 2001).

Remark 1. *While the basic properties do not guarantee the existence of a one-step minorization condition, they do guarantee 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. Of course, the k-step chain inherits the basic properties from X. Finally, note that if X is countable, then it is easy to establish* (2) *by fixing a point $\tilde{x} \in \mathsf{X}$ and taking $s(x) = I(x = \tilde{x})$ and $\nu(\cdot) = P(\tilde{x}, \cdot)$.*

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, $X' = \{(X_i,\delta_i) : i = 0,1,\ldots\}$, which lives on the space $\mathsf{X}\times\{0,1\}$ and is such that, marginally, the sequence $\{X_i : i = 0,1,\ldots\}$ has the same distribution as the original chain, X. An important property of X' is that $A := \mathsf{X}\times\{1\}$ is a *proper atom* for the chain X' and the (random) times at which X' enters A 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, Tierney and Yu (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

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

where $q(\cdot)$ and $k(\cdot|x)$ are the densities corresponding to $\nu(\cdot)$ and $P(x,\cdot)$. We will use this approach in our examples. Jones and Hobert (2001) and Mykland et al. (1995) provide further practical advice on simulating the split chain. In the next section we use the development of the split chain to derive an identity for π.

3 Approximating π

Define τ_A to be the first return time to A; that is,

$$\tau_A = \min\{n \geq 1 : (X_n,\delta_n) \in A\}\ .$$

Also, let $\Pr_A(\cdot)$ and $\mathrm{E}_A(\cdot)$ 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 $\mathrm{E}_A(\tau_A) < \infty$. Consequently, we can define a discrete random variable, T^*, with support $\mathbb{N}$ and probabilities defined by

$$p_t = \frac{\Pr_A(\tau_A \geq t)}{\mathrm{E}_A(\tau_A)}\ . \tag{4}$$

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

$$Q_t(B) = \Pr_A(X_t \in B | \tau_A \geq t)\ ; \tag{5}$$

i.e., Q_t is the conditional distribution of X_t given that $(X_0, \delta_0) \in A$ and that there are no regenerations in the split chain before time t.

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*

$$\pi(B) = \sum_{t=1}^{\infty} Q_t(B)\, p_t \ , \tag{6}$$

where Q_t and p_t are defined in terms of the split chain at (4) and (5).

Proof. See Appendix A. □

Clearly, (6) is appealing from a simulation point of view because it shows that it is possible to simulate a random variable from π using a sequential sampling mechanism. That is, a draw from π can be made by first drawing from the distribution T^*, call the result t^*, and then making a 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 algorithm, it is assumed that $t \geq 2$.

Algorithm I:

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

Hobert and Robert (2004) show that if $s(x) = \varepsilon > 0$, then T^* has a geometric distribution. Therefore, in that case, one can use iid draws from the geometric distribution in conjunction with Algorithm I to make iid draws from π. Unfortunately, in most cases where X is the basis of a practically relevant MCMC algorithm, it is difficult, at best, to make draws from the distribution of T^*. (Jones and Hobert (2004) call an MCMC algorithm *practically relevant* when the stationary distribution is complex enough that iid sampling is not straightforward.)

Alternatively, suppose that we could find probabilities, call them $\{\hat{p}_t\}_{t=1}^{\infty}$, that are "close" to the probabilities $\{p_t\}_{t=1}^{\infty}$. Then we could approximate π with

$$\hat{\pi}(\cdot) = \sum_{t=1}^{\infty} Q_t(\cdot)\, \hat{p}_t \ , \tag{7}$$

from which it is straightforward to sample. Furthermore, note that

$$\|\pi(\cdot) - \hat{\pi}(\cdot)\| = \left\| \sum_{t=1}^{\infty} Q_t(\cdot)\, p_t - \sum_{t=1}^{\infty} Q_t(\cdot)\, \hat{p}_t \right\| \leq \sum_{t=1}^{\infty} |p_t - \hat{p}_t| \ . \tag{8}$$

Thus, 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 $\{\hat{p}_t\}_{t=1}^{\infty}$.

Hobert and Robert (2004) show that, given any $\gamma > 0$, it is possible to use a *geometric drift condition* on the Markov chain X to construct $\{\hat{p}_t\}_{t=1}^{\infty}$ such that $\sum_{t=1}^{\infty} |p_t - \hat{p}_t| < \gamma$. In the next section, we show that even without a drift condition, it is possible to do the same thing except that, instead of being able to say with certainty that $\sum_{t=1}^{\infty} |p_t - \hat{p}_t| < \gamma$, we will only be able to say that this inequality holds with high probability.

4 Estimating the Mass Function of T^*

The key to our construction is that making iid draws from the distribution of τ_A is straightforward; 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_A can be used to construct a statistical estimate of $\hat{p}_t$ via (4). First note that

$$p_t = \frac{\Pr_A(\tau_A \geq t)}{\mathrm{E}_A(\tau_A)} = \frac{\Pr_A(\tau_A \geq t)}{\sum_{s=1}^{\infty} \Pr_A(\tau_A \geq s)} = \frac{1 - F(t-1)}{1 + \sum_{s=1}^{\infty}[1 - F(s)]}$$

where $F(t) := \Pr_A(\tau_A \leq t)$ is the distribution function of τ_A conditional on $(X_0, \delta_0) \in A$. Let $\tau_{A,1}, \ldots, \tau_{A,m}$ denote an iid sample of size m from the distribution of τ_A, and let $F_m(t)$ denote the corresponding empirical distribution function. To estimate p_t, we plug-in F_m in place of F; i.e.,

$$\hat{p}_t = \frac{1 - F_m(t-1)}{1 + \sum_{s=1}^{\infty}[1 - F_m(s)]} = \frac{1 - F_m(t-1)}{\overline{\tau}_A} \tag{9}$$

where $\overline{\tau}_A$ is the sample mean. Note that $\sum_{t=1}^{\infty} \hat{p}_t = 1$. Hence, $\{\hat{p}_t\}$ will always be a legitimate mass function on $\mathbb{N}$ from which we can sample.

We now use asymptotic arguments to show that $\{\hat{p}_t\}_{t=0}^{\infty}$ enjoys a type of "strong consistency" and to get a handle on the error of $\{\hat{p}_t\}_{t=0}^{\infty}$. These results allow us to develop a method of choosing an appropriate value for m. In light of (8), 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 $\{\hat{p}_t\}_{t=0}^{\infty}$ is an asymptotically reasonable estimate of the mass function of T^*.

Theorem 2. *For $\{\hat{p}_t\}$ as defined in (9) 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. See Appendix B. □

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 $\mathrm{E}_A\left[\tau_A^{2+\varepsilon}\right] < \infty$ for some $\varepsilon > 0$, then

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

where $B(s)$, $0 \le s \le 1$, denotes a Brownian bridge process. We can use this result to choose an appropriate value for m in (9).

Remark 2. *The assumption that $E_A\left[\tau_A^{2+\varepsilon}\right] < \infty$ is a weak condition that is closely related to the mixing properties of the Markov chain (see e.g. Roberts and Tweedie; 1999). In fact, if this condition were to fail, it is difficult to imagine that the Markov chain would mix sufficiently well to be of any practical use.*

As we now describe, (10) can be used to come up with a reasonable value of m. Suppose that $\tau_{A,1}, \ldots, \tau_{A,m'}$ is an initial sample of τ_A'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 $\Pr[2d_1(F_m, F) < \gamma] \approx 1 - \alpha$, and hence that $\{\hat{p}_t\}$ is within γ of $\{p_t\}$ with probability approximately equal to $1 - \alpha$. We may then conclude that

$$\|\pi(\cdot) - \hat{\pi}(\cdot)\| < \gamma$$

with probability approximately equal to $1 - \alpha$. Before describing how this approximation can be used to attack the burn-in problem, we briefly mention one possible avenue for improving upon these asymptotic approximations.

There are three forms of approximation used in the above argument: (i) the asymptotic approximation in (10); (ii) the use of $\sum_{t=1}^{\infty} |B(F_{m'}(t))|$ in place of $\sum_{t=1}^{\infty} |B(F(t))|$; and (iii) the estimation of the quantile of L. The most bothersome of these is certainly (ii). It may be possible to dispense with approximations (ii) and (iii), and simplify the method at the same time. To be specific, let $\mathcal{G}$ denote the class of distribution functions corresponding to discrete random variables with support $\mathbb{N}$ that have a finite $2 + \varepsilon$ moment. Consider the random variables

$$L_G = \sum_{t=1}^{\infty} |B(G(t))|$$

as G ranges over $\mathcal{G}$. Suppose there exists a G^* such that L_{G^*} is stochastically larger then any other L_G. Then we could simply use L_{G^*} in place of $\sum_{t=1}^{\infty} |B(F(t))|$. This would remove approximation (ii) and would alleviate the need for the initial sample of m' τ_A's. Furthermore, the quantiles of L_{G^*} could be tabulated and this would obviate approximation (iii). All of this suggests that an investigation into the possible existence of a "dominating" G^* could be well worth the effort. However, we do not explore this possibility any further here.

5 An Application to Burn-in

Our assumptions about P imply that for every initial probability measure $\lambda(\cdot)$ on $\mathcal{B}(\mathsf{X})$ we have

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

where $P^n(\lambda, A) := \int_{\mathsf{X}} P^n(x, A)\, \lambda(dx)$ is the probability distribution of X_n given that $X_0 \sim \lambda$. Typically, the MCMC user has no particular starting distribution in mind. Indeed, $\lambda(\cdot)$ is usually taken to be a point mass at some point from which it is convenient to start the simulation. An important problem in the implementation of MCMC algorithms is *burn-in* (time), which is formally described as follows. Given $\lambda(\cdot)$ and $\gamma > 0$, we want to find an n^* such that

$$\|P^{n^*}(\lambda, \cdot) - \pi(\cdot)\| < \gamma \,. \tag{11}$$

If (11) holds, then the marginal distribution of X_n (conditional on $X_0 \sim \lambda$) is within γ of π for all $n \geq n^*$. Hence, n^* may be regarded as a reasonable time to start sampling the Markov chain.

Several authors have recently shown that drift and minorization conditions on the Markov chain can be used to derive computable upper bounds on $\|P^n(\lambda, \cdot) - \pi(\cdot)\|$ that decrease geometrically fast in n (Douc, Moulines and Rosenthal; 2002; Meyn and Tweedie; 1994; Roberts and Tweedie; 1999; Rosenthal; 1995a). These upper bounds can be used to find an n^* that satisfies (11). The phrase "difficult theoretical analysis" is used by Fill, Machida, Murdoch and Rosenthal (2000) to describe this method. Furthermore, when this strategy is used in the context of a practically relevant MCMC algorithm, it is not unusual for the resulting n^* to be too large to be of any practical value (see e.g. Jones and Hobert; 2004).

A much simpler, but slightly less rigorous approach to burn-in is to simply start the Markov chain, X, by taking $X_0 \sim \hat{\pi}$. To be specific, suppose that the methods of Subsection 4 are used to construct $\{\hat{p}_t\}$ such that

$$\Pr\left[\sum_{t=1}^{\infty} |\hat{p}_t - p_t| < \gamma\right] \approx 1 - \alpha$$

where α and γ are small. Then we can say that $\Pr(\|\pi - \hat{\pi}\| < \gamma) \approx 1 - \alpha$. Then if X is started with $X_0 \sim \hat{\pi}(\cdot)$, one can begin sampling the Markov chain right away. In the next two sections, we illustrate the construction of $\hat{\pi}$ with toy and realistic examples, respectively.

6 A Toy Example

Suppose that $\pi(x) = e^{-x}I(x > 0)$. 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 sampler with an Exp(θ) proposal; that is, the proposal density is $q(x) = \theta e^{-\theta x}I(x > 0)$. The chain evolves as follows: Given $X_n = x$, 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 $X_{n+1} = x$. The mixing behavior of this sampler (as a function of θ) is well known and this allows us to evaluate the importance of regularity conditions. 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 the chain is uniformly ergodic if $0 < \theta < 1$ which guarantees that τ_A has a moment generating function and hence $E_A[\tau_A^{2+\epsilon}] < \infty$. The rate of convergence is known to be subgeometric for $\theta > 1$. Furthermore, the results of Roberts (1999) suggest that $E_A[\tau_A^2]$ is finite for $1 < \theta \leq 2$ and possibly infinite when $\theta > 2$.

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

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

and ν has 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 $\theta = 2.5$. 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_A'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%).

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

Then, for each value of θ,we then 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_A'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.

Table 2: Results from Simulating L

θ	α	c	γ	m
0.75	0.20	10.237	0.05	1.68×10^5
			0.25	6.70×10^3
	0.10	12.877	0.05	2.65×10^5
			0.25	1.06×10^4
1.5	0.20	48.093	0.05	3.70×10^6
			0.25	1.48×10^5
	0.10	60.167	0.05	5.79×10^6
			0.25	2.32×10^5
2.5	0.20	197.09	0.05	6.22×10^7
			0.25	2.49×10^6
	0.10	245.59	0.05	9.65×10^7
			0.25	3.86×10^6

For each value of θ, we constructed $\hat{\pi}$ using the m corresponding to the row in Table 2 with $\alpha = 0.20$ and $\gamma = 0.25$. In Figure 1 we present two density estimates. One estimate is based on iid samples from the target while the other estimate is based on a random sample of draws from $\hat{\pi}$. The density estimates were made using the `density` function available in the R software package (Ihaka and Gentleman; 1996). Examination of Figure 1 reveals that when $\theta = 0.75$ we obtain a good approximation to π. However, when $\theta = 1.5$ or $\theta = 2.5$, the quality of the approximation is only slightly worse.

7 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 also 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 a conditionally independent hierarchical model

$$Y|\beta, u, R, D \sim \ \mathrm{N}_n(X\beta + Zu, R^{-1})$$

$$\beta|u, R, D \sim \ \mathrm{N}_p(\beta_0, B^{-1})$$

$$u|D,R \sim \ \mathrm{N}_q(0,D^{-1}) \tag{12}$$

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

$$\pi(\beta,u,R,D|y) \propto f(y|\beta,u,R,D)f(\beta|u,R,D)f(u|D,R)f(R)f(D)\ . \tag{13}$$

We assume that the priors on R and D are such that the resulting posterior (13) 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.

7.1 A block Gibbs sampler and a minorization condition

In this section, we consider a block Gibbs sampler with components R, D and $\xi = (\beta^T, u^T)^T$. 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 $(p+q)$–variate Normal with mean ξ_0 and covariance matrix Σ^{-1} where

$$\Sigma = \begin{pmatrix} Z^TRZ+D & Z^TRX \\ X^TRZ & X^TRX+B \end{pmatrix} \qquad \text{and} \qquad \Sigma\xi_0 = \begin{pmatrix} Z^TRy \\ X^TRy+B\beta_0 \end{pmatrix}\ . \tag{14}$$

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

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

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 "distinguished 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

$$s(\xi', \tilde{\xi}) = c_q \left[\inf_{R \in \mathbb{M}_R} \frac{\pi(R|\xi')}{\pi(R|\tilde{\xi})} \right] \left[\inf_{D \in \mathbb{M}_D} \frac{\pi(D|\xi')}{\pi(D|\tilde{\xi})} \right] \tag{15}$$

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 that R lives in; 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 results from Nummelin (1984) and Mykland et al. (1995) 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

$$\Pr(\delta = 1 | D', R', \xi', D, R, \xi) = \left[\inf_{R \in \mathbb{M}_R} \frac{\pi(R|\xi')}{\pi(R|\tilde{\xi})} \right] \left[\inf_{D \in \mathbb{M}_D} \frac{\pi(D|\xi')}{\pi(D|\tilde{\xi})} \right] \frac{\pi(D|\tilde{\xi}) \pi(R|\tilde{\xi})}{\pi(D|\xi') \pi(R|\xi')} \; . \tag{16}$$

Thus we have to calculate the infima in (15) and plug into (16). Let $a_{1ij} \leq a_{2ij}$ for $i = 1, \ldots, q$ and $j = 1, \ldots, q$ be constants and define $\mathbb{M}_D = \{M_{q \times q} : a_{1ij} \leq m_{ij} \leq a_{2ij}\}$. Then

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

where

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

Let $v' = y - X\beta' - Zu'$ and $\tilde{v} = y - X\tilde{\beta} - Z\tilde{u}$. Also, let $b_{1ij} \leq b_{2ij}$ for $i = 1, \ldots, n$ and

$j = 1, \ldots, n$ be constants and define $\mathbb{M}_R = \{M_{n\times n} : b_{1ij} \leq m_{ij} \leq b_{2ij}\}$. Then

$$\begin{aligned}
\inf_{R\in\mathbb{M}_R} \frac{\pi(R|\xi')}{\pi(R|\tilde{\xi})} &= \frac{C_R(\tilde{\xi})}{C_R(\xi')} \inf_{R\in\mathbb{M}_R} \frac{\exp\{-0.5(y - X\beta' - Zu')^T R(y - X\beta' - Zu')\}}{\exp\{-0.5(y - X\tilde{\beta} - Z\tilde{u})^T R(y - X\tilde{\beta} - Z\tilde{u})\}} \\
&= \frac{C_R(\tilde{\xi})}{C_R(\xi')} \inf_{R\in\mathbb{M}_R} \frac{\exp\{-0.5 v'^T R v'\}}{\exp\{-0.5\tilde{v}^T R\tilde{v}\}} \\
&= \frac{C_R(\tilde{\xi})}{C_R(\xi')} \exp\left\{-0.5 \sum_i \sum_j (v'_i v'_j - \tilde{v}_i\tilde{v}_j) h_{ij}\right\} \\
&= \frac{C_R(\tilde{\xi})}{C_R(\xi')} h(v', \tilde{v})
\end{aligned}$$

where

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

Thus the probability of regeneration is given by

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

7.2 A Numerical Example

In this section, we identify a specific example of the model (12), 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

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

Set $\xi = (u^T, \beta)^T$ and $\lambda = (\lambda_D, \lambda_R)^T$. The data in Table 3 were simulated according to this model with $n = 5$, $r_1 = r_2 = d_1 = d_2 = 1$ and covariate $X \sim \text{N}(0, I_5)$.

We will construct $\hat{\pi}$ corresponding to the posterior that results from the data in Table 3 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: $(\lambda', \xi') \to (\lambda, \xi)$. The full conditionals for the precision parameters are given by

$$\begin{aligned}
\lambda_R|\xi, y &\sim \text{Gamma}\left(1 + \frac{n}{2},\ 2 + \frac{1}{2}(y - X\beta - u)^T(y - X\beta - u)\right) , \\
\lambda_D|\xi, y &\sim \text{Gamma}\left(1 + \frac{n}{2},\ 2 + \frac{1}{2}u^T u\right) .
\end{aligned}$$

Table 3: Simulated Data

y	x
3.05577	-0.65015
-0.84096	0.46053
-3.21066	-0.39088
-0.47085	-0.64953
2.23286	-0.65276

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

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

To simulate from this multivariate normal distribution we require the Cholesky decomposition of Σ and this is reported in Appendix C.

While some work has been done analyzing block Gibbs samplers for 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 and hence we simply assume that it satisfies $\mathrm{E}_A \tau_A^{2+\epsilon} < \infty$.

To use the minorization condition developed in the previous section 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 $0 < b_1 < b_2$. We ran the block Gibbs sampler for 5×10^5 iterations starting from $\xi_0 = \bar{y}1$ where 1 is a vector of ones. Let $\tilde{u}_1, \ldots \tilde{u}_5, \tilde{\beta}, \tilde{\lambda}_D, \tilde{\lambda}_R$ be the estimated posterior expectations of the associated parameters. We set $\tilde{\xi} = (\tilde{u}_1, \ldots \tilde{u}_5, \tilde{\beta})^T$, $[a_1, a_2] = \tilde{\lambda}_D \pm w\, s_{\lambda_D}$ 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.

We simulated an initial sample of $m' = 5 \times 10^5$ iid τ_A's. The results are reported in Table 4. We then simulated 1×10^5 values of L and the results are given in Table 5. Using these results we constructed $\hat{\pi}$ and subsequently simulated 5×10^4 iid draws from it and estimated the marginal density functions of λ_D, λ_D and β using the R `density` function. In Figure 2, we compare these estimated densities with the corresponding estimated densities based on 5×10^4 draws from the block Gibbs sampler after discarding the first 5×10^6 iterations.

In looking at Figure 2, it is clear that both sets of density estimates largely agree. Apparently, drawing a starting value from $\hat{\pi}$ would produce a starting value that is (marginally) similar to that obtained from a long burn-in period.

Table 4: Initial Sample Results

w	u'_m	max	99%
0.75	205	248	86

Table 5: Results from Simulating L

α	c	δ	m
0.20	208.27	0.25	2.78×10^6

Appendices

A Proof of Theorem 1

The proof closely follows the proof of Theorem 1 in Hobert and Robert (2004). Let π' denote the invariant measure for X'. Applying Meyn and Tweedie's (1993) Theorem 10.2.1 to X' and using the fact that $\pi'(A \times \{0,1\}) = \pi(A)$, we have

$$\pi(A) = \frac{1}{\mathrm{E}_\alpha(\tau_\alpha)} \sum_{t=1}^{\infty} \Pr_\alpha(X_t \in A, \tau_\alpha \geq t) = \sum_{t=1}^{\infty} \Pr_\alpha(X_t \in A | \tau_\alpha \geq t)\, p_t \; .$$

B Proof of Theorem 2

First

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

and hence

$$\sum_{t=1}^{\infty} |\hat{p}_t - p_t| \leq 2 \sum_{t=1}^{\infty} |F_m(t) - F(t)| = 2 \int_{-\infty}^{\infty} |F_m(t) - F(t)|\, dt = 2 d_1(F_m, F). \qquad (17)$$

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

C Cholesky Decomposition of Σ

The Cholesky decomposition of Σ is given by $\Sigma = LL^T$, where L is lower triangular with positive elements on the diagonal. Let

$$L = \begin{pmatrix} l_1 & 0 \\ l_2 & l_3 \end{pmatrix}$$

solving for L we obtain

$$l_1 = aI_n, \quad l_2 = bX^T \quad \text{and} \quad l_3 = c$$

where $a = \sqrt{\lambda_R + \lambda_D}$, $b = \lambda_R / a$ and $c = \sqrt{1 + (\lambda_R \lambda_D / a^2) X^T X}$. It is easy to see that

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

and hence

$$\Sigma^{-1} = \begin{pmatrix} a^{-2} I_n + (b/ac)^2 X X^T & -b/ac^2 X \\ -b/ac^2 X^T & c^{-2} \end{pmatrix}.$$

Acknowledgments

Hobert's research partially supported by NSF Grant DMS-00-72827.

References

Del Barrio, E., Gine, E. and Matran, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions, *The Annals of Probability* **27**: 1009–1071.

Douc, R., Moulines, E. and Rosenthal, J. S. (2002). Quantitative bounds for geometric convergence rates of Markov chains, *Technical report*, University of Toronto, Department of Statistics.

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

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. and Robert, C. P. (2004). A mixture representation of π with applications in Markov chain Monte Carlo and perfect sampling, *The Annals of Applied Probability* (to appear).

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**: 299–314.

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo, *Statistical Science* **16**: 312–334.

Jones, G. L. and Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model, *The Annals of Statistics* (to appear).

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*, Springer, New York.

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.

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

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.

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 Tweedie, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains, *Stochastic Processes and their Applications* **80**: 211–229. Corrigendum (2001) **91**: 337-338.

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

Rosenthal, J. S. (1995b). Rates of convergence for Gibbs sampling for variance component models, *The Annals of Statistics* **23**: 740–761.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*, John Wiley and Sons, New York.

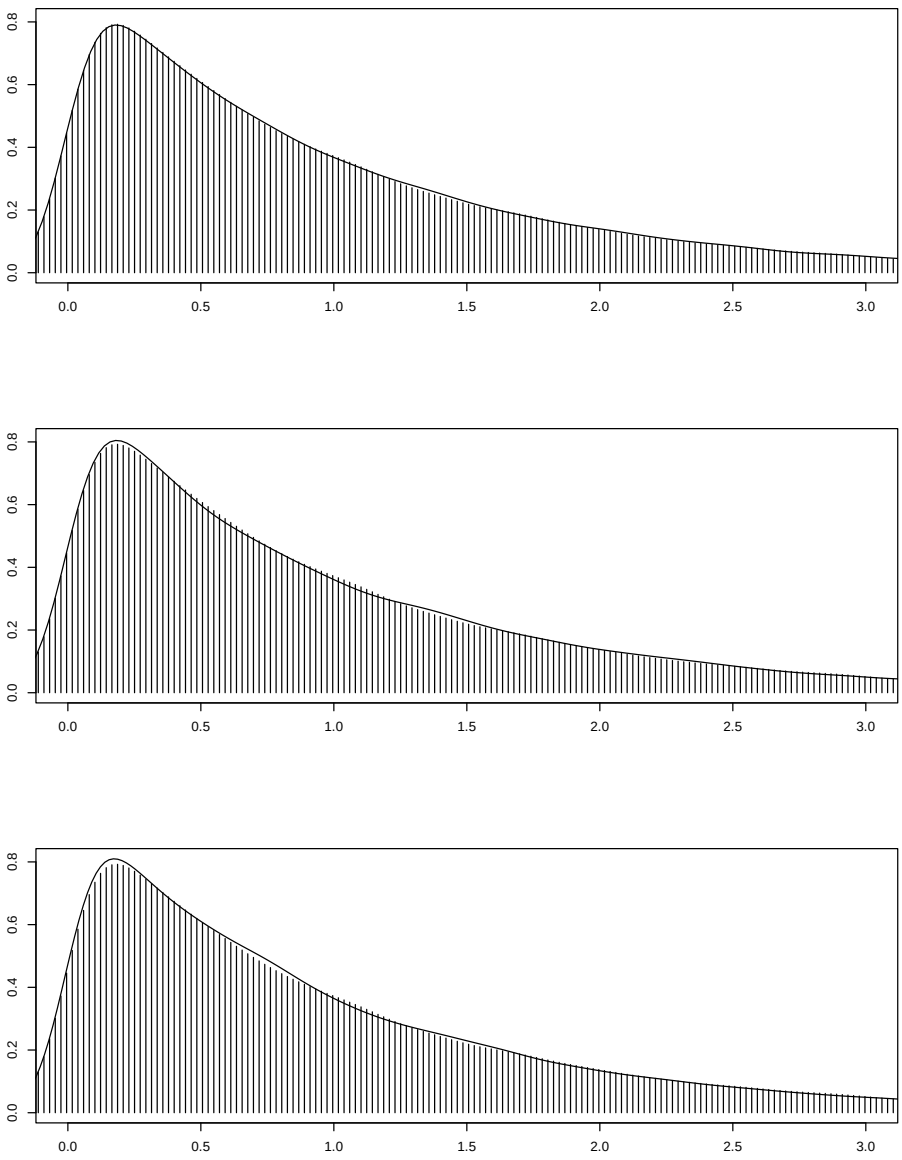

Figure 1: Each plot contains two estimates of the target Exp(1) density: The shaded region is based on a random sample of size 5×10^4 drawn from an Exp(1) distribution while the curve is based on a random sample of size 5×10^4 drawn from $\hat{\pi}$. In the top plot $\hat{\pi}$ was produced using an independence sampler with Exp(0.75) candidate while in the middle plot $\hat{\pi}$ was based on an independence sampler with Exp(1.5) candidate and the bottom plot was constructed using $\hat{\pi}$ generated by an independence sampler with Exp(2.5) candidate.

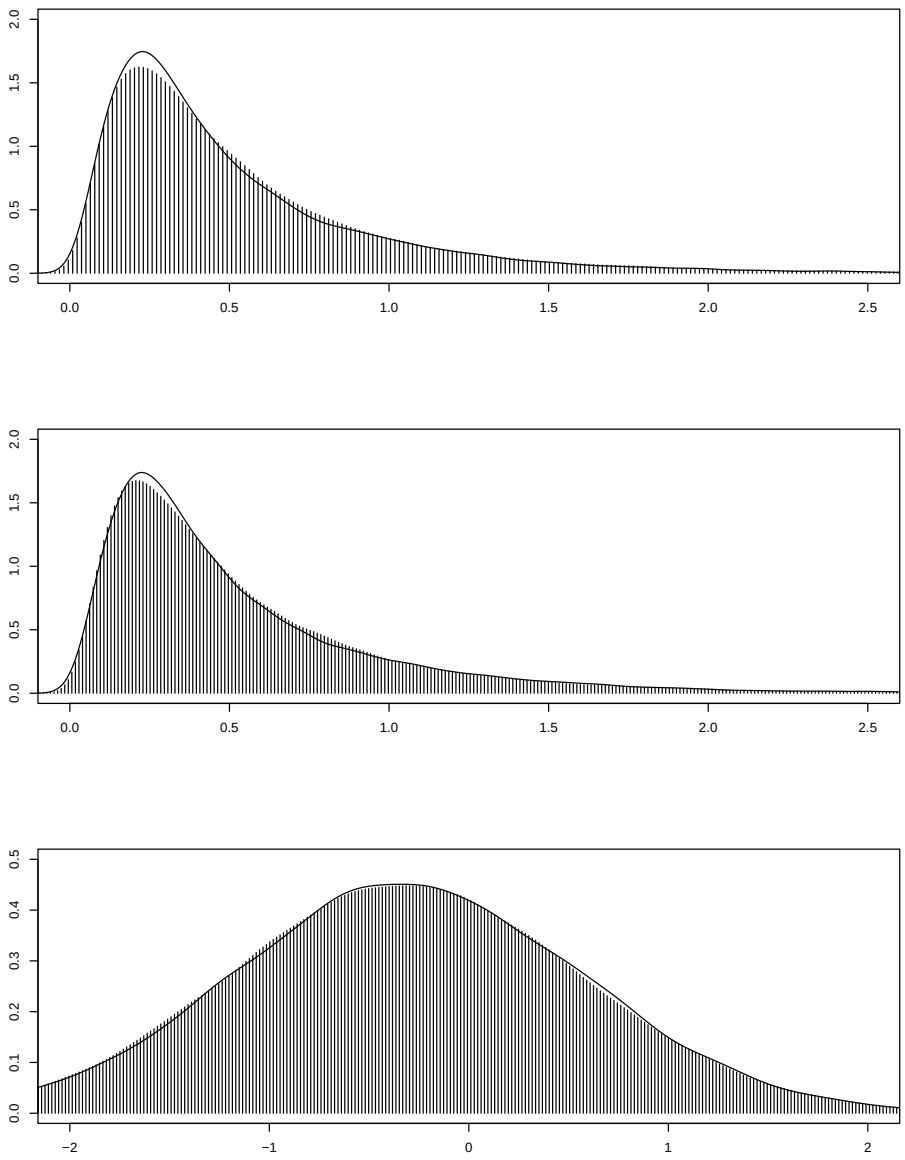

Figure 2: Each plot presents two estimates of a marginal density: The top plot corresponds to the density of λ_D, the middle plot corresponds to the density of λ_R and the bottom plot corresponds to the density of β. In each plot the curve is based on a random sample of size 5×10^4 drawn from $\hat{\pi}$ while the shaded region is based on 5×10^4 Gibbs draws after a burn-in of 5×10^6.