Geometric Ergodicity of Gibbs Samplers

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Alicia A. Johnson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Galin L. Jones, Adviser

July 2009

Acknowledgements

Infinite thanks are owed to the people that have supported me throughout my years at the University of Minnesota. First, I am truly grateful for the opportunity to work with my adviser Galin Jones. There is not a more enthusiastic, personable, and patient mentor. He has been more than generous with his time; reading countless *rough* drafts, offering guidance, and thoughtfully answering my endless questions. From him, I have learned so much about research, teaching, and the importance of taking it all in stride.

I would like to thank Glen Meeden, Charlie Geyer, and Jim Hodges for serving on my committee. I am also grateful to Jim for graciously sharing his data resources and to Glen for his pep talks. The encouragement and opportunities he has given me are truly appreciated. Sandy Weisberg has been another significant source of support. Even during my most humbling experiences in the consulting clinic, he was positive and enthusiastic. He helped me become a better statistician and I am lucky to have worked with him.

Finally, a huge debt of gratitude is owed to my friends and family. Their sense of humor has carried me through this process. Their love has given it meaning. Thank you.

Abstract

Due to a demand for reliable methods for exploring intractable probability distributions, the popularity of Markov chain Monte Carlo (MCMC) techniques continues to grow. In any MCMC analysis, the convergence rate of the associated Markov chain is of practical and theoretical importance. A *geometrically ergodic* chain converges to its target distribution at a geometric rate. In this dissertation, we establish verifiable conditions under which geometric ergodicity is guaranteed for Gibbs samplers in a general model setting. Further, we show that geometric ergodicity of the *deterministic scan* Gibbs sampler ensures geometric ergodicity of the Gibbs sampler under alternative scanning strategies. As an illustration, we consider Gibbs sampling for a popular Bayesian version of the general linear mixed model. In addition to ensuring the rapid convergence required for useful simulation, geometric ergodicity is a key sufficient condition for the existence of central limit theorems and consistent estimators of Monte Carlo standard errors. Thus our results allow practitioners to be as confident in inference drawn from Gibbs samplers as they would be in inference drawn from random samples from the target distribution.

Contents

List of Tables

List of Figures

Chapter 1

Motivation

1.1 Introduction

Markov chain Monte Carlo (MCMC) methods have transformed statistical inference in intractable settings by providing a means for approximately sampling from complicated, high-dimensional probability distributions. For instance, a common use of MCMC is to explore the typically complex posterior distributions corresponding to Bayesian hierarchical models. Though these sophisticated methods provide in-roads to previously intractable problems, they also introduce a unique set of issues that must be addressed in drawing inferences from the Markov chains they produce.

Let π denote a probability measure on $(\mathcal{X}, \mathcal{B})$ where $\mathcal{X}$ denotes the support of π with associated Borel σ-algebra $\mathcal{B}$. Also, let π_u denote the corresponding (possibly unnormalized) probability density with respect to some measure μ, ie. for any $A \in \mathcal{B}$

$$\pi(A) = \frac{\int_A \pi_u(x)\mu(dx)}{\int_{\mathcal{X}} \pi_u(x)\mu(dx)} \, .$$

Suppose we are interested in evaluating

$$\mathrm{E}_\pi g(X) := \int_{\mathcal{X}} g(x)\pi(dx) = \frac{\int_{\mathcal{X}} g(x)\pi_u(x)\mu(dx)}{\int_{\mathcal{X}} \pi_u(x)\mu(dx)}$$

for some function $g : \mathcal{X} \to \mathbb{R}$. When π is complicated, $\mathrm{E}_\pi g$ is the ratio of two intractable integrals. If no closed form solution exists, we might evaluate $\mathrm{E}_\pi g$ using numerical integration techniques. However, these techniques become increasingly difficult to implement in high dimensions. In such cases, data simulation methods provide a more viable approach to this problem.

Examples of data simulation methods include Monte Carlo and Markov chain Monte Carlo. In the classical Monte Carlo setting, an independent and identically distributed (iid) sample $\left\{X^{(0)}, X^{(1)}, \ldots X^{(n-1)}\right\}$ is obtained from π and $\mathrm{E}_\pi g$ is estimated by the sample average $\overline{g}_n := (1/n)\sum_{i=0}^{n-1} g\left(X^{(i)}\right)$. Further inference regarding $\mathrm{E}_\pi g$ is guided by fundamental large-sample statistical theory. First, $\overline{g}_n$ is unbiased for $\mathrm{E}_\pi g$ and the strong law of large numbers guarantees that $\overline{g}_n \to \mathrm{E}_\pi g$ almost surely (ie. with probability one) as $n \to \infty$. Also, if $\mathrm{E}_\pi g(X)^2 < \infty$, the central limit theorem (CLT) guarantees

$$\sqrt{n}\left(\overline{g}_n - \mathrm{E}_\pi g\right) \xrightarrow{d} N\left(0, \sigma_g^2\right) \qquad \text{as } n \to \infty$$

where $\sigma_g^2 = \mathrm{Var}_\pi g(X)$. The sample variance of the $g\left(X^{(i)}\right)$'s, $\hat{\sigma}_g^2$, is *strongly consistent* for σ_g^2. That is, $\hat{\sigma}_g^2 \to \sigma_g^2$ almost surely as $n \to \infty$. Therefore, a valid *Monte Carlo standard error* (MCSE) for $\overline{g}_n$ can be calculated by $\hat{\sigma}_g/\sqrt{n}$. The MCSE provides a measure of the accuracy of $\overline{g}_n$ in estimating $\mathrm{E}_\pi g$. Further, it can be used to determine a sufficient Monte Carlo sample size. For instance, we might choose n for which the confidence interval half-width $t_{\alpha/2,n-1}\hat{\sigma}_g/\sqrt{n}$ is below some prespecified value, say 0.01 (where $t_{\alpha/2,n-1}$ is the appropriate critical value of the Student's t distribution with $n-1$ degrees of freedom).

In some cases, direct simulation from π can be prohibitively difficult if not impossible. However, when π is intractable, MCMC algorithms can often be used to simulate a Markov chain $\left\{X^{(0)}, X^{(1)}, \ldots X^{(n-1)}\right\}$ where (1) the $X^{(i)}$ are drawn from *approx-*

imations of π; and (2) successive draws $X^{(i)}$ and $X^{(i+1)}$ are dependent. Somewhat surprisingly, under certain conditions we can achieve the same level of confidence in estimates based on a dependent Markov chain as we have in estimates based on an iid Monte Carlo sample.

First, a set of regularity conditions guarantees that the Markov chain sample average $\overline{g}_n$, though biased, will still converge to $\mathrm{E}_\pi g$ almost surely as $n \to \infty$ (see Chapter 2.1 for details). These conditions also guarantee that the Markov chain will converge to target distribution π in *total variation distance*. However, it is the *rate* of this convergence that holds considerable weight in drawing further inference from the Markov chain. A Markov chain that converges to its target distribution at a geometric rate is said to be *geometrically ergodic*. Geometric ergodicity is important for at least three reasons: (1) a Markov chain that converges quickly is crucial for achieving effective simulation results in finite time; (2) it is a key sufficient condition for the existence of a central limit theorem (CLT) (Jones, 2004); and (3) it is required for consistent estimation of Monte Carlo standard errors (Flegal et al., 2008; Hobert et al., 2002; Jones et al., 2006). As in the iid case, (2) and (3) are necessary for rigorously assessing the accuracy of $\overline{g}_n$ and determining a sufficient Markov chain sample size n. Geometric ergodicity also guarantees the existence of a (potentially impractical) perfect sampler (Kendall, 2004). The following toy example illustrates the potential impact of the convergence rate.

1.2 A Toy Example

Consider an Exp(1) target distribution with density $\pi(x) = \exp\{-x\}I(x > 0)$. Suppose we are interested in evaluating $\mathrm{E}_\pi X$ (where it is assumed we do not know $\mathrm{E}_\pi X = 1$). We estimate this quantity using both Monte Carlo and Markov chain Monte Carlo techniques. First, Monte Carlo estimation requires a random sample

from π. To this end, we generated an iid sample of size $n = 1000$ from Exp(1). This produced $\overline{x}_n = 0.99$ with a standard error of $\hat{\sigma}/\sqrt{n} = 0.03$ where $\hat{\sigma}$ is the sample standard deviation. From the usual CLT, a 95% confidence interval for $\mathrm{E}_\pi X$ is

$$\overline{x}_n \pm t_{0.025,n-1}\frac{\hat{\sigma}}{\sqrt{n}} \approx 0.99 \pm 1.96(0.03) = (0.93, 1.05) \ .$$

This interval contains 1, the true value of $\mathrm{E}_\pi X$. In addition, n could be increased if a smaller margin of error were desired.

Next, suppose it is not possible to sample directly from π. Instead, consider exploring π using an independence Metropolis sampler with an Exp(θ) proposal distribution. (We say $X \sim \mathrm{Exp}(\theta)$ if it has density $q(y) = \theta \exp\{-\theta y\} I(y > 0)$.) In each iteration, the Markov chain advances from its current state, $X^{(t-1)} = x$, as follows. Draw a *candidate value* $y \sim \mathrm{Exp}(\theta)$. With probability

$$\alpha(x, y) = \min\left\{\frac{\pi(y)q(x)}{\pi(x)q(y)},\ 1\right\} = \min\left\{\exp\{(x-y)(1-\theta)\},\ 1\right\} \ ,$$

'accept' the draw and set $X^{(t)} = y$. Otherwise, set $X^{(t)} = x$. This construction can be accomplished using the following algorithm:

1. Select initial value $X^{(0)}$.

2. On the tth iteration, suppose $X^{(t-1)} = x$ and generate $X^{(t)}$ as follows:

   - Draw candidate value $y \sim \mathrm{Exp}(\theta)$ and $u \sim \mathrm{Uniform}(0, 1)$, independently.
   - If $u < \alpha(x, y)$, set $X^{(t)} = y$.

     Otherwise, set $X^{(t)} = x$.

3. Repeat step 2.

Remark 1.1. Notice that a Markov chain produced by this algorithm has the potential to get "stuck" in the same state for multiple iterations in a row.

As indicated by the dependence of $\alpha(x, y)$ on θ, different proposal distributions produce different Markov chains. Consider, for instance, setting $\theta = 1$. In this case, the independence sampler produces iid samples from Exp(1) since the acceptance probability always equals one:

$$\alpha(x, y) = \min\{\exp\{(x - y)(1 - \theta)\},\ 1\} = \min\{\exp\{0\},\ 1\} = 1\ .$$

On the other hand, it follows from Mengersen and Tweedie (1996) that the chain is geometrically ergodic for any $0 < \theta < 1$ and subgeometric (slower than geometric) for any $\theta > 1$. (See Example 2.3 in Chapter 2.3 for details.) Accordingly, we consider the independence sampler for $\theta \in \{0.5, 1, 4\}$ where $\theta = 1$ is included for comparison.

To gain some intuition into the impact of θ, consider a plot of the densities corresponding to the three distributions of interest (Figure 1.1). The rate of exponential

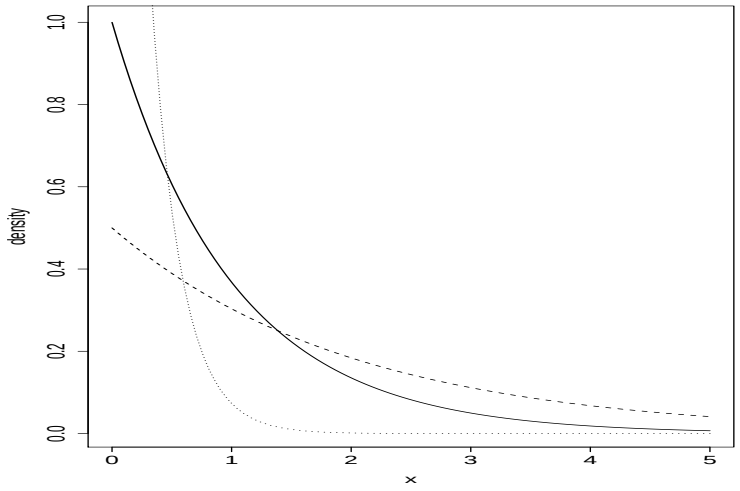


Figure 1.1: Probability densities for Exp(0.5) (dashed line), Exp(1) (solid line), and Exp(4) (dotted line).

decay clearly increases as θ increases. For instance, the Exp(4) tail is much lighter than the Exp(1) tail. As a result, candidate values from Exp(4) will tend to be small

and the corresponding sampler will not thoroughly explore the tail of the target distribution. On the other hand, Exp(0.5) has a heavier tail than Exp(1). The range of candidate values from Exp(0.5) is therefore more likely to be representative of "typical" values of Exp(1).

These suspected behaviors can be observed in short runs of the independence sampler. For each $\theta \in \{0.5, 1, 4\}$ we ran a Markov chain for 1000 iterations started from $X^{(0)} = 1$. The chains are plotted versus iteration number in Figure 1.2.

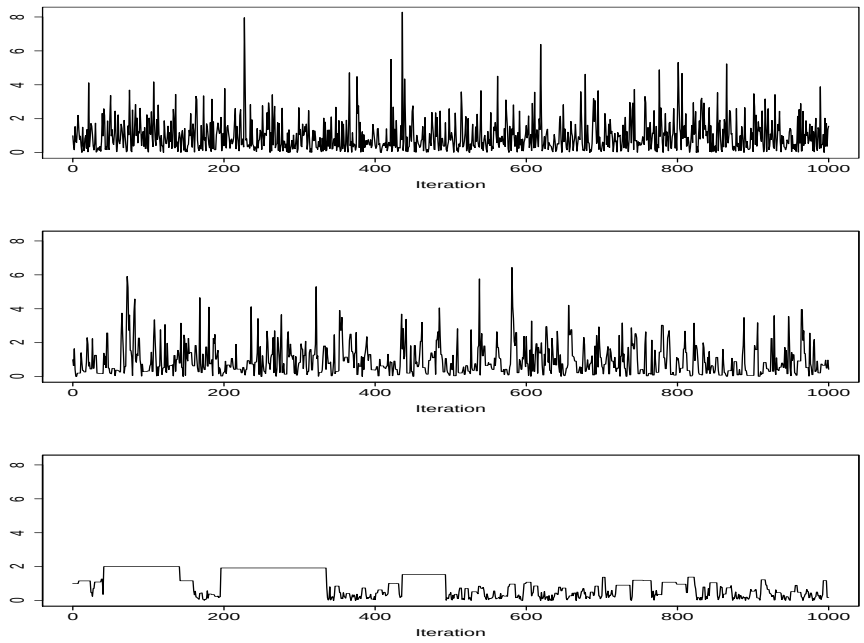


Figure 1.2: 1000 iterations of the Markov chains starting from $X^{(0)} = 1$ for $\theta = 1$ (top plot), $\theta = 0.5$ (middle plot), and $\theta = 4$ (bottom plot).

The independence sampler with $\theta = 0.5$ mixes quickly (middle plot). The relative frequency with which the chain reaches the tail of Exp(1) is also similar to that for the iid sampler (top plot). On the other hand, the independence sampler with $\theta = 4$ (bottom plot) only visits states in the approximate range (0,2). Further, the chain gets "stuck" when it visits the upper end of this range. This phenomenon follows

from the fact that when $\theta = 4$, the acceptance probability

$$\alpha(x, y) = \min\left\{\exp\{-3(x - y)\},\ 1\right\}$$

tends to be small when current state x is large in comparison to candidate value y.

Though each independence sampler will eventually converge to the target density, Figure 1.2 suggests the $\theta = 4$ chain will converge more slowly than the $\theta = 0.5$ chain. To illustrate the different convergence properties, for each $\theta \in \{0.5, 1, 4\}$ we ran 1000 independent Markov chains starting from $X^{(0)} = 1$ for 1000 iterations each. From every chain, we picked off and stored the values of $X^{(15)}$ and $X^{(1000)}$. Figure 1.3 displays histograms of the 1000 independent values of $X^{(15)}$ for each independence sampler. To informally assess the distance of the samplers from stationarity after

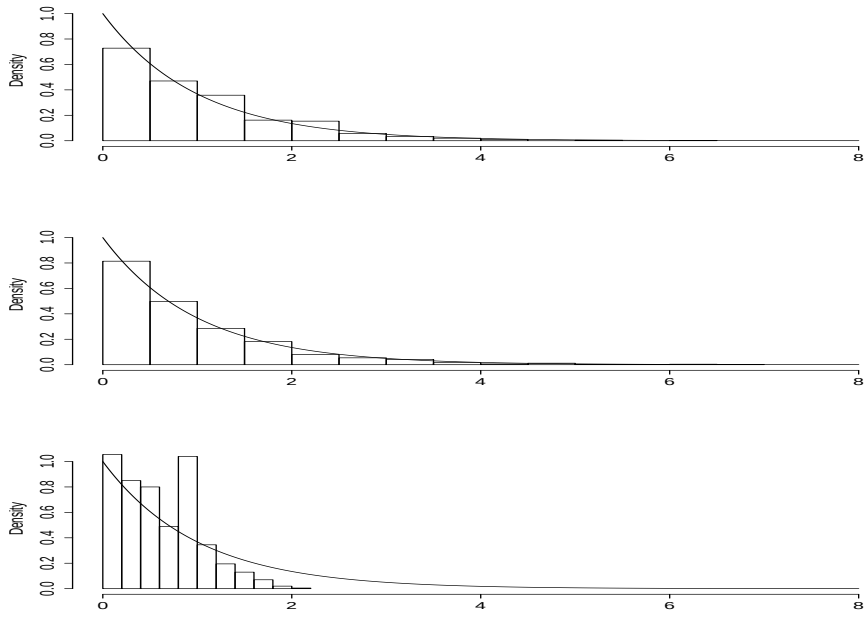


Figure 1.3: Histograms of the 1000 independent copies of $X^{(15)}$ for $\theta = 1$ (top plot), $\theta = 0.5$ (middle plot), and $\theta = 4$ (bottom plot). The Exp(1) density is super-imposed on each histogram (solid line).

15 iterations, these distributions can be compared to the Exp(1) density included in each plot. Indeed, there is strong agreement between the density and histograms for both $\theta = 1$ and $\theta = 0.5$ (top plot and middle plot, respectively). This is of course expected when $\theta = 1$. When $\theta = 0.5$, the agreement suggests the corresponding sampler converges very quickly (a notion supported by theory). On the other hand, there is a much larger discrepancy between the density and histogram for $\theta = 4$ (bottom plot). This suggests that convergence to stationarity requires more than 15 iterations in this case. Further, the spike at $x = 1$ reflects the large number of chains that were still stuck at the starting value after 15 iterations.

Histograms of the 1000 independent values of $X^{(1000)}$ for each sampler are displayed in Figure 1.4. Each sampler appears closer to stationarity after 1000 iterations

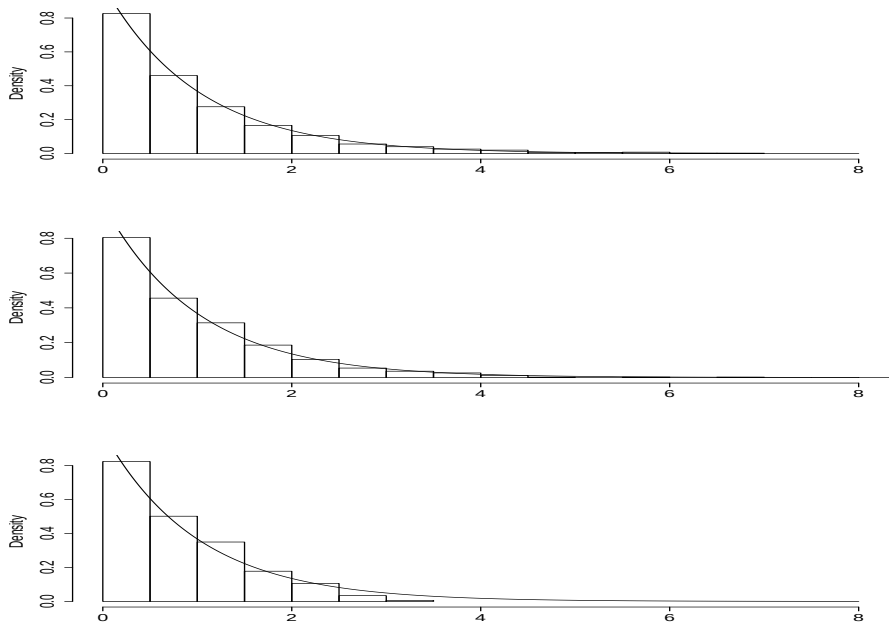


Figure 1.4: Histograms of the 1000 independent copies of $X^{(1000)}$ for $\theta = 1$ (top plot), $\theta = 0.5$ (middle plot), and $\theta = 4$ (bottom plot). The Exp(1) density is super-imposed on each histogram (solid line).

than after 15. However, there is still a notable discrepancy between π and the distri-

bution of $X^{(1000)}$ for $\theta = 4$ (bottom plot). For instance, the largest observed value of $X^{(1000)}$ was less than 3.40. In contrast, we would expect nearly 30 values in a random sample of 1000 from Exp(1) to be *greater* than 3.40 since $\Pr_\pi(X > 3.40) \approx 0.03$. Therefore, even after 1000 iterations, the independence sampler with $\theta = 4$ largely ignores the tail of the target distribution.

To consider the impact of convergence rate on the estimation of $\mathrm{E}_\pi X$, we calculated the Markov chain average

$$\overline{x}_{1000} = \frac{1}{1000} \sum_{i=1}^{1000} X^{(i)}$$

for each of the above Markov chains. This produced a collection of 1000 independent estimates of $\mathrm{E}_\pi X$ for each $\theta \in \{0.5, 1, 4\}$. Histograms of the averages are displayed in Figure 1.5 and approximate the sampling distributions of $\overline{x}_{1000}$ corresponding to the

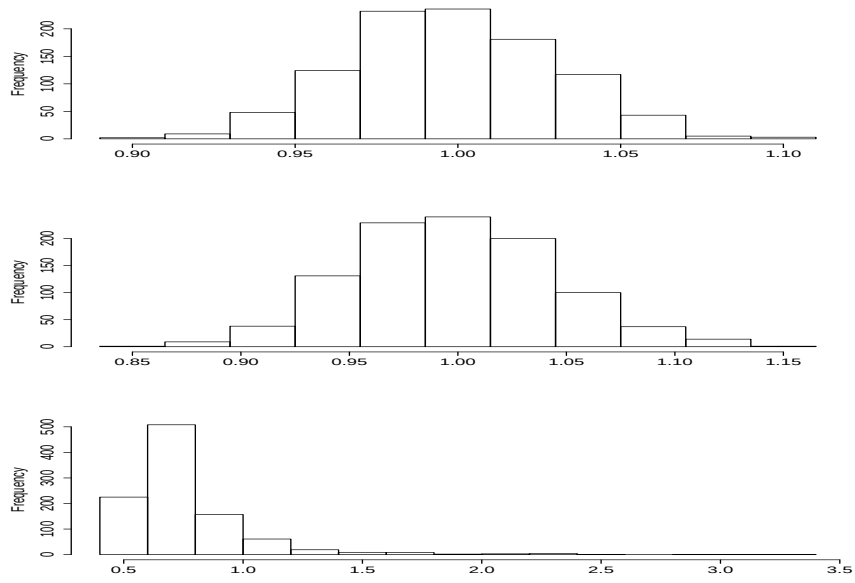


Figure 1.5: Histograms of the 1000 independent sample averages $\overline{x}_{1000}$ for $\theta = 1$ (top plot), $\theta = 0.5$ (middle plot), and $\theta = 4$ (bottom plot).

three independence samplers.

The histogram for $\theta = 1$ (top plot) supports what is already known in the iid case, that the CLT holds. Though the distribution of estimates is somewhat more variable, it also appears that a CLT holds when $\theta = 0.5$. Indeed, since the corresponding sampler is geometrically ergodic, the existence of a CLT is guaranteed by theory (see Chapter 3.2.1). Unfortunately, a CLT does not appear to hold when $\theta = 4$ (bottom plot). Further, the Markov chain averages systematically underestimate $\mathrm{E}_\pi X = 1$. (This is hardly surprising given that the right tail of Exp(1) is under-represented in Markov chain samples when $\theta = 4$.) This phenomenon is not a result of the relatively small Markov chain sample sizes ($n = 1000$). It is known that a CLT does not exist when $\theta = 4$ (see Roberts (1999) for details).

We conclude by completing Markov chain inference for $\mathrm{E}_\pi X$ using an independence sampler with $\theta = 0.5$. In this case, geometric ergodicity gives us the tools we need to be as confident in the resulting MCMC estimate as we are in estimates based on iid samples from Exp(1). First, we ran the sampler for 1000 iterations starting from $X^{(0)} = 1$. This produced $\overline{x}_{1000} = 1.05$ with a standard error of $\hat{\sigma}/\sqrt{1000} = 0.04$ where $\hat{\sigma}$ was calculated using consistent batch means (Jones et al., 2006). By the existence of a CLT and a consistent estimator of the standard error, we can construct an asymptotically valid 95% confidence interval for $\mathrm{E}_\pi X$:

$$\overline{x}_{1000} \pm t_{0.025,999} \frac{\hat{\sigma}}{\sqrt{1000}} \approx 1.05 \pm 1.96(0.04) = (0.97, 1.13) \; .$$

Positive correlation among the Markov chain observations leads to a wider interval than that constructed from the iid sample. In other words, the Markov chain estimate has a larger standard error than the estimate based on an iid sample of the same size. However, as in the iid case, the interval half-width can be decreased by increasing Markov chain sample size n (if desired).

1.3 Our Goal

The above toy example illustrates the potential impact of convergence rate on the quality of Markov chain inference. In practice, the discrepancy between Markov chains with different convergence rates might not be so acute. For instance, subgeometric chains do not always converge as slowly or produce inference as misleading as the $\theta = 4$ chain. Unfortunately, complete knowledge of the true nature of the target distribution is unavailable in practice. (If it were, inference would not require MCMC!) Therefore, it is typically impossible to determine if a particular subgeometric chain is one of the "lucky" ones. On the other hand, geometric ergodicity *guarantees* good behavior of the Markov chain (at least asymptotically).

Due to the "peace of mind" it provides, geometric ergodicity has received substantial attention in the literature (the references are too many to list here). However, convergence rate depends on both the choice of MCMC algorithm and the target distribution. Therefore, it is difficult to treat geometric ergodicity on a truly general level. We address some of the resulting gaps in the literature by focusing on geometric ergodicity for a particular MCMC algorithm, the Gibbs sampler. Considered in the seminal article by Geman and Geman (1984), the use of this sampler has increased with the popularity of data augmentation methods and the introduction of the BUGS software (Bayesian inference Using Gibbs Sampling) (Spiegelhalter et al., 2005).

We begin in Chapter 2 with an overview of basic MCMC theory and techniques for establishing geometric ergodicity. In Chapter 3 we discuss the implications of geometric ergodicity in rigorously addressing the following critical questions: (1) When has the chain converged to its target distribution?; and (2) What is a sufficient Markov chain simulation length? In Chapter 4 we derive conditions under which geometric ergodicity is guaranteed for Gibbs samplers in a general model setting. Finally, we illustrate our results with a series of simulations and applications in Chapter 5.

Chapter 2

Geometric Ergodicity

2.1 Markov chain Monte Carlo Background

Let π denote a probability measure on measurable space $(\mathcal{X}, \mathcal{B})$ where $\mathcal{B}$ denotes the σ-algebra associated with state space $\mathcal{X}$. In a slight abuse of notation, also let π denote the corresponding probability density with respect to some measure μ. That is, for any set $A \in \mathcal{B}$, $\pi(A) = \int_A \pi(x)\mu(dx)$. For ease of exposition, we will often assume that μ is the Lebesgue measure. However, all methods herein extend beyond the Lebesgue setting.

Let $\Phi = \left\{X^{(0)}, X^{(1)}, \ldots\right\}$ denote a discrete time Markov chain on $(\mathcal{X}, \mathcal{B})$. The construction of Φ begins from starting value $X^{(0)}$. This might be drawn from some *initial distribution* π_0. However, $X^{(0)}$ is often set to some chosen value. From $X^{(0)}$, the Markov chain evolves according to some Markov *transition kernel* P. We assume throughout that P has corresponding *transition density* k. Specifically, for $i \in \{0, 1, 2, \cdots\}$, $X^{(i+1)} \sim P\left(X^{(i)}, \cdot\right)$ where for any $A \in \mathcal{B}$

$$P(x, A) = Pr\left(X^{(i+1)} \in A \,\middle|\, X^{(i)} = x\right) = \int_A k(x, y)dy \; .$$

The chain constructed from π_0 and P is Markov since it satisfies the *Markov property*.

That is, conditional on the "present", the "future" is independent of the "past":

$$\Pr\left(X^{(i+1)} \,\middle|\, X^{(i)}, X^{(i-1)}, \ldots, X^{(0)}\right) = \Pr\left(X^{(i+1)} \,\middle|\, X^{(i)}\right) .$$

Further, the Markov chain is *time invariant* since P does not depend on iteration number.

For any fixed x, $P(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$. The transition kernel P also represents two linear operators. First, for any probability distribution λ on $(\mathcal{X}, \mathcal{B})$, define λP as follows:

$$\lambda P(A) = \int_{\mathcal{X}} \lambda(x) P(x, A) dx$$

for $A \in \mathcal{B}$. Therefore, if $X^{(i)} \sim \lambda$, λP represents the marginal distribution of $X^{(i+1)}$. Next, for any nonnegative measurable function f on $(\mathcal{X}, \mathcal{B})$, define Pf as

$$Pf(x) = \int_{\mathcal{X}} P(x, dy) f(y) = \mathrm{E}_P\left[f\left(X^{(i+1)}\right) \,\middle|\, X^{(i)} = x\right]$$

where E_P denotes expectation with respect to the transition kernel.

Extending the above definitions, let P^m denote the m-step transition kernel corresponding to the m-step transition density k^m. That is, for any $i \in \{0, 1, 2, \cdots\}$,

$$P^m(x, A) = Pr\left(X^{(i+m)} \in A \,\middle|\, X^{(i)} = x\right) = \int_A k^m(x, y) dy$$

where $A \in \mathcal{B}$ and k^m can be defined iteratively as

$$k^m(x, y) = \int_{\mathcal{X}} k(x, z) k^{m-1}(z, y) dz .$$

Therefore, $P^m(x, A)$ is the probability the Markov chain moves from state x to set A in m iterations. Also, setting $i = 0$, $P^m(x, A)$ represents the distribution of the

Markov chain at the mth iteration given the chain started in state x.

If P preserves draws from π, ie. $\pi = \pi P$, π is the *invariant* or *stationary density* for the chain. Equivalently, π is invariant for the chain if

$$\pi(x) = \int_{\mathcal{X}} \pi(y)k(y,x)dy$$

since for any $A \in \mathcal{B}$, integrating over both sides gives

$$\pi(A) = \int_A \pi(x)dx = \int_A \int_{\mathcal{X}} \pi(y)k(y,x)dydx = \int_{\mathcal{X}} \pi(y)P(y,A)dy = \pi P(A) \ .$$

If $X^{(m-1)} \sim \pi$ and $X^{(m)} \sim P\left(X^{(m-1)}, \cdot\right)$, the invariance of π guarantees $X^{(m)} \sim \pi$. There are several techniques for ensuring π is the invariant distribution. One of these is to construct a Markov chain that satisfies the following *detailed balance* condition:

$$\pi(x)k(x,y) = \pi(y)k(y,x) \quad \text{for all } x, y \in \mathcal{X} \ .$$

If this property holds, the Markov chain is *reversible* with respect to π. The invariance of π follows by integrating over both sides of the detailed balance condition:

$$\int_{\mathcal{X}} \pi(y)k(y,x)dy = \int_{\mathcal{X}} \pi(x)k(x,y)dy = \pi(x) \ .$$

(NOTE: Reversibility is not required for π to be invariant.)

Suppose a Markov chain has invariant distribution π. Then if $X^{(0)} \sim \pi$, $X^{(m)} \sim \pi$ for all $m = 0, 1, \ldots$ (where the draws themselves are dependent). When $X^{(i)} \sim \pi$ for all i, we say the Markov chain is *stationary*. Unfortunately, Markov chains are typically not stationary (as this requires $X^{(0)} \sim \pi$). However, when the Markov chain satisfies certain regularity conditions, it is guaranteed to *converge* to stationarity. Discussing this result requires some Markov chain definitions.

Definition 2.1. A Markov chain is *ϕ-irreducible* for some measure ϕ on $(\mathcal{X}, \mathcal{B})$ if for all $x \in \mathcal{X}$ and $A \in \mathcal{B}$ for which $\phi(A) > 0$, there exists n for which $P^n(x, A) > 0$. That is, a Markov chain is ϕ-irreducible if every ϕ-positive set is *accessible* from any state $x \in \mathcal{X}$.

Definition 2.2. A ϕ-irreducible Markov chain has *period* d if state space $\mathcal{X}$ can be partitioned into disjoint sets $N, D_1, \ldots, D_d \in \mathcal{B}$ for which $\phi(N) = 0$, $\Pr(x, D_{i+1}) = 1$ for all $x \in D_i$ and $i = 0, \ldots, d-1$, and $\Pr(x, D_1) = 1$ for $x \in D_d$. Markov chain Φ is *periodic* if $d \geq 2$, that is, if $\mathcal{X}$ can be partitioned in a way so that Φ makes a regular tour through the partition. Otherwise, if $d = 1$, Φ is *aperiodic*.

Definition 2.3. A ϕ-irreducible Markov chain is *Harris recurrent* if for any starting value x and any set A for which $\phi(A) > 0$, the chain reaches set A with probability one. Equivalently, a Markov chain is Harris recurrent if for any starting value x the chain visits set A infinitely often with probability one, ie. $\Pr\left(X^{(n)} \in A \text{ i.o. } |X^{(0)} = x\right) = 1$.

Definition 2.4. A Markov chain is *Harris ergodic* if it is ϕ-irreducible, aperiodic, Harris recurrent, and possesses invariant distribution π for some measures ϕ and π.

Definition 2.5. Let $\| \mu(\cdot) - \nu(\cdot) \|$ denote the *total variation distance* between two measures $\mu(\cdot)$ and $\nu(\cdot)$ on $(\mathcal{X}, \mathcal{B})$ where

$$\| \mu(\cdot) - \nu(\cdot) \| := \sup_{A \in \mathcal{B}} |\mu(A) - \nu(A)|.$$

Unless noted otherwise, we assume throughout that Markov chain Φ is *Harris ergodic*. Regardless of the initial distribution, a chain satisfying this property is guaranteed to explore the entire state space without getting "stuck," at least asymptotically. Harris ergodicity also guarantees strong consistency of the Markov chain average and convergence of the Markov chain to stationarity in total variation distance.

Theorem 2.1 (Ergodic Theorem)**.**

Suppose Markov chain Φ *is Harris ergodic with invariant distribution* π*. Also, suppose* $E_\pi|g(X)| < \infty$ *for some function* $g : \mathcal{X} \to \mathbb{R}$*. Then for any starting value* $x \in \mathcal{X}$*,*

$$\overline{g}_n = \frac{1}{n}\sum_{i=0}^{n-1} g\left(X^{(i)}\right) \to E_\pi g(X) \quad \textit{almost surely as } n \to \infty\ .$$

Theorem 2.2.

Suppose Markov chain Φ *is Harris ergodic with invariant distribution* π*. Then for any starting value* $x \in \mathcal{X}$*,* Φ *will converge to* π *in total variation distance. That is,*

$$\parallel P^n(x,\cdot) - \pi(\cdot) \parallel \to 0 \quad \textit{as} \quad n \to \infty\ . \tag{2.1}$$

Further, $\parallel P^n(x,\cdot) - \pi(\cdot) \parallel$ *is monotonically nonincreasing in* n*.*

Remarks:

1. The monotonic nonincreasing nature of the total variation distance of the chain to stationarity is guaranteed by Proposition 13.3.2 of Meyn and Tweedie (1993).

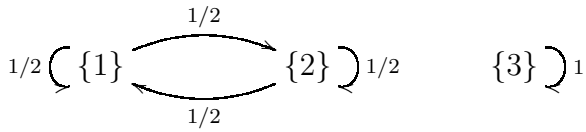
2. Convergence in total variation distance is much stronger than convergence "in distribution." Let y be a continuity point of $F(y) := \pi((-\infty, y])$. Then convergence in distribution only requires $|P^n(x, A) - \pi(A)| \to 0$ as $n \to \infty$ for any set A of the form $A = (-\infty, y]$.

3. If Φ is *not* Harris recurrent but *is* ϕ-irreducible and aperiodic with invariant distribution π, (2.1) still holds for π-*almost every* $x \in \mathcal{X}$. Specifically, let A denote the set of starting values x for which (2.1) does not hold. Then $\pi(A) = 0$.

A proof of Theorem 2.2 is outlined in Chapter 2.3. Next, we illustrate some of the intuitive connections between Harris ergodicity and convergence to stationarity with the following example adapted from Roberts and Rosenthal (2004).

Example 2.1.

Define distribution π on state space $\mathcal{X} = \{1, 2, 3\}$ with $\pi\{1\} = \pi\{2\} = \pi\{3\} = 1/3$. Consider Markov chains Φ_1 and Φ_2 for π corresponding to two different transition kernels, P_1 and P_2, respectively. First, suppose $P_1(1, \{1\}) = P_1(1, \{2\}) = P_1(2, \{1\}) = P_1(2, \{2\}) = 1/2$ and $P_1(3, \{3\}) = 1$:

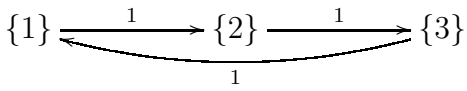


In this case, Φ_1 corresponding to P_1 has invariant distribution π since

$$\pi P(\{y\}) = \sum_{i=1}^{3} \pi(i) P(i, \{y\}) = \pi(y) \; .$$

However, Φ_1 is clearly reducible since $P_1^n(1, \{3\}) = 0$ for all n. That is, if started in state $\{1\}$, Φ_1 will *never* visit state $\{3\}$. It follows that $P_1^n(1, \{3\}) \nrightarrow \pi(3)$ as $n \to \infty$. Therefore, the reducibile Markov chain will not converge to stationarity.

Now, consider transition kernel P_2 where $P(1, \{2\}) = P(2, \{3\}) = P(3, \{1\}) = 1$:



The corresponding Markov chain Φ_2 has invariant distribution π and is clearly irreducible. In fact, Φ_2 is Harris recurrent since it will visit $\{i\}$ infinitely often for any $i \in \{1, 2, 3\}$, regardless of the starting value. However, Φ_2 has period $d = 3$. For instance, suppose Φ_2 is started in state $\{1\}$. Then $P^n(1, \{1\}) = 1$ if n is any multiple of 3 and $P^n(1, \{1\}) = 0$ otherwise. In this case, the periodic Markov chain will not converge to stationarity since $P^n(1, \{1\}) \nrightarrow \pi(1)$.

Theorem 2.2 establishes convergence to stationarity for Harris ergodic Markov chains. However, it does not guarantee anything about the *rate* at which this conver-

gence occurs. A Markov chain is *geometrically ergodic* if there exists some function $M : \mathcal{X} \to \mathbb{R}$ and some constant $t \in (0, 1)$ that satisfy

$$\| P^n(x, \cdot) - \pi(\cdot) \| \leq M(x)\, t^n \quad \text{for any } x \in \mathcal{X} . \tag{2.2}$$

If M is bounded, the Markov chain is *uniformly ergodic*.

Remarks:

1. So long as x is not some bad starting value (ie. $M(x)$ is not large), geometric ergodicity guarantees quick convergence for the Markov chain.

2. Geometric ergodicity (in fact, uniform ergodicity) holds for every irreducible and aperiodic Markov chain on a finite state space. However, this is *not* true for Markov chains on general state spaces. In the next section we will discuss techniques for establishing geometric ergodicity in this more general setting.

2.2 Establishing Geometric Ergodicity

A constructive technique for establishing geometric ergodicity is to derive *drift* and *minorization* conditions.

Definition 2.6. A *Type I drift condition* holds if there exists some non-negative function $V : \mathcal{X} \to \mathbb{R}_{\geq 0}$ and constants $0 < \gamma < 1$ and $L < \infty$ for which

$$PV(x) \leq \gamma V(x) + L \quad \text{for any } x \in \mathcal{X} . \tag{2.3}$$

Further, we call V a *drift function* and γ a *drift rate*.

Definition 2.7. A *minorization condition* holds on set $C \in \mathcal{B}$ if there exist some

positive integer m, $\varepsilon > 0$, and probability measure Q on $(\mathcal{X}, \mathcal{B})$ for which

$$P^m(x, A) \geq \varepsilon Q(A) \quad \text{for all } x \in C \text{ and } A \in \mathcal{B} \,. \tag{2.4}$$

If (2.4) holds for $m = 1$, we call this a *one-step minorization condition.*

Definition 2.8. If a minorization condition holds on set C, we say C is *small.*

Remark 2.1. If P has corresponding transition density k, (2.4) holds on set C if there exist some positive integer m, $\varepsilon > 0$, and probability density q on $\mathcal{X}$ for which

$$k^m(x, y) \geq \varepsilon q(y) \quad \text{for all } x \in C \text{ and } y \in \mathcal{X}. \tag{2.5}$$

To see this, suppose (2.5) holds and let Q be the probability measure corresponding to density q. Then integrating over both sides of (2.5) establishes (2.4): For all $x \in C$ and $A \in \mathcal{B}$,

$$P^m(x, A) = \int_A k^m(x, y)dy \geq \int_A \varepsilon q(y)dy = \varepsilon Q(A) \,.$$

The sufficiency of drift and minorization for geometric ergodicity is established in the following proposition.

Proposition 2.1.

Suppose Markov chain Φ *is irreducible and aperiodic with invariant distribution* π*. Then* Φ *is geometrically ergodic if the following two conditions are met:*

1. *A Type I drift condition* (2.3) *holds for some non-negative function* $V : \mathcal{X} \to \mathbb{R}_{\geq 0}$ *and constants* $0 < \gamma < 1$ *and* $L < \infty$*.*

2. *There exists some constant* $d > 2L/(1 - \gamma)$ *for which a minorization condition* (2.4) *holds with* $m = 1$ *on set* $C = \{x : V(x) \leq d\}$*.*

Remarks:

1. Proposition 2.1 is a corollary of Theorem 12 of Rosenthal (1995a). See the remarks following Theorem 3.1 in Chapter 3.1.1 for details.

2. Under the assumptions of Proposition 2.1, $M(x)$ in (2.2) can be taken to be proportional to $V(x) + 1$. Therefore, starting the chain from the state x which minimizes $V(x)$ is a natural choice.

3. Suppose Φ is *Feller*, ie. $P(\cdot, O)$ is *lower semicontinuous* for any open set $O \in \mathcal{B}$:

$$\liminf_{y \to x} P(y, O) \geq P(x, O) \quad \text{for } x, y \in \mathcal{X} .$$

   In this case, Proposition 2.1 also follows from drift condition (2.3) if the support of π has non-empty interior and the drift condition is *unbounded off compact sets*, ie. if $C = \{x : V(x) \leq d\}$ is compact for any $d > 0$. When a minorization condition is difficult to construct, this might provide an alternative approach for establishing geometric ergodicity. A proof of this claim follows from Lemma 15.2.8 and Theorems 6.0.1 and 15.0.1 of Meyn and Tweedie (1993).

4. The conditions of this proposition also guarantee Harris recurrence. An argument can be given by combining Theorem 9.1.8 with Lemmas 15.2.2 and 15.2.8 of Meyn and Tweedie (1993).

Though we will typically refer to Proposition 2.1, we also include an alternative (but similar) result that will come in handy. A proof follows directly from Theorem 15.0.1 of Meyn and Tweedie (1993) and requires the following drift condition.

Definition 2.9. A *Type II drift condition* holds if there exists some function $W : \mathcal{X} \to [1, \infty)$ finite at some $x \in \mathcal{X}$, some set $D \in \mathcal{B}$, and constants $0 < \rho < 1$ and $b < \infty$ for which

$$PW(x) \leq \rho W(x) + bI_D(x) \quad \text{for all } x \in \mathcal{X} . \tag{2.6}$$

Theorem 2.3.

Suppose Markov chain Φ is aperiodic and ϕ-irreducible with invariant distribution π. Then Φ is geometrically ergodic if there exists some small set D, drift function $W : \mathcal{X} \to [1, \infty)$, and constants $0 < \rho < 1$ and $b < \infty$ for which a Type II drift condition (2.6) *holds.*

We typically find it easier to work with and establish (2.3) in comparison to (2.6). However, the two conditions are clearly related. In fact, (2.6) holds if and only if (2.3) holds.

Lemma 2.1.

1. *Suppose a Type II drift condition* (2.6) *holds. Then a Type I drift condition* (2.3) *holds for $V(x) = W(x)$, $\gamma = \rho$, and $L = b$.*

2. *Suppose a Type I drift condition* (2.3) *holds. Then a Type II drift condition* (2.6) *holds for $W(x) = V(x) + 1$, $\rho = (\gamma + 1)/2$, $b = L + (1 - \gamma)$, and $D = \{x : W(x) \leq b/(1 - \rho)\}$.*

Proof.

The proof of statement 1 is obvious. To prove statement 2, suppose (2.3) holds. Notice that $W : \mathcal{X} \to [1, \infty)$ and $0 < \rho < 1$. Also, it follows from (2.3) that

$$\begin{aligned} PW(x) &\leq \gamma W(x) + b \\ &= (2\rho - 1)W(x) + b \\ &= \rho W(x) - (1 - \rho)W(x) + b \, . \end{aligned}$$

Hence for any $x \in D$, $1 < W(x) \leq b/(1 - \rho)$ and $PW(x) \leq \rho W(x) + b$. On the other hand, for any $x \notin D$, $W(x) > b/(1 - \rho)$ so that

$$PW(x) \leq \rho W(x) - (1 - \rho)\left(\frac{b}{1 - \rho}\right) + b = \rho W(x) \, .$$

Therefore, $PW(x) \leq \rho W(x) + bI_D(x)$. □

To conclude this section, we demonstrate the construction of drift and minorization conditions for a toy example considered by Jones and Hobert (2001).

Example 2.2.

For $m \geq 5$, let $Y_1, \ldots, Y_m$ be independent and identically distributed (iid) $N(\mu, \theta)$ where the joint prior density for (μ, θ) is $f(\mu, \theta) \propto 1/\sqrt{\theta}$. Also, let $y = (y_1, \ldots, y_m)$ denote the sample data with mean $\overline{y}$ and (scaled) variance $s^2 = \sum(y_i - \overline{y})^2$. This model yields posterior density

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

and full conditional distributions

$$\theta|\mu, y \sim IG\left(\frac{m-1}{2}, \frac{s^2 + m(\mu - \overline{y})^2}{2}\right)$$
$$\mu|\theta, y \sim N(\overline{y}, \theta/m)$$

where we say $X \sim IG(a, b)$ if it has density proportional to $x^{-(a+1)}e^{-b/x}I(x > 0)$.

Consider exploring $\pi(\mu, \theta|y)$ using a deterministic-scan Gibbs sampler (DUGS) with update scheme

$$(\theta', \mu') \to (\theta, \mu') \to (\theta, \mu).$$

In each iteration, this sampler consecutively updates θ' and μ' by drawing from full conditional distributions $\theta|\mu', y$ and $\mu|\theta, y$, respectively. The corresponding transition density can be written as

$$k((\mu', \theta'), (\mu, \theta)) = \pi(\theta|\mu', y)\pi(\mu|\theta, y)$$

where $\pi(\theta|\mu, y)$ and $\pi(\mu|\theta, y)$ represent the probability densities corresponding to the full conditional distributions $\theta|\mu, y$ and $\mu|\theta, y$, respectively. Also, let P denote the transition kernel corresponding to k.

Geometric ergodicity for the DUGS follows by establishing Conditions 1 and 2 of Proposition 2.1. We begin by constructing a Type I drift condition (2.3). Define $V(\mu, \theta) = (\mu - \overline{y})^2$ and notice that by the construction of the Gibbs sampler,

$$\mathrm{E}\left[V(\mu,\theta)|\mu',\theta'\right] = \mathrm{E}\left[V(\mu,\theta)|\mu'\right] = \mathrm{E}\left[\mathrm{E}\left[V\left(\mu,\theta\right)|\theta\right]|\mu'\right]$$

where

$$\mathrm{E}\left[V(\mu,\theta)|\theta\right] = \mathrm{E}\left[(\mu-\overline{y})^2|\theta\right] = \mathrm{Var}[\mu|\theta] = \theta/m\ .$$

Therefore,

$$\begin{aligned}\mathrm{E}\left[V(\mu,\theta)|\mu',\theta'\right] = \mathrm{E}\left[\theta/m|\mu'\right] &= \frac{1}{m}\frac{s^2+m(\mu'-\overline{y})^2}{m-3}\\ &= \frac{(\mu'-\overline{y})^2}{m-3} + \frac{s^2}{m(m-3)}\\ &= \frac{1}{m-3}V\left(\mu',\theta'\right) + \frac{s^2}{m(m-3)}\end{aligned}$$

where the second equality follows from the fact that if $X \sim \mathrm{IG}(a, b)$, then $\mathrm{E}(X) = b/(a-1)$. Notice that $m \geq 5$ guarantees $1/(m-3) < 1$. Hence, the following drift condition holds:

$$PV\left(\mu',\theta'\right) = \mathrm{E}\left[V(\mu,\theta)|\mu',\theta'\right] \leq \gamma V(\mu',\theta') + L$$

where $\gamma \in (1/(m-3), 1)$ and $L = s^2/(m(m-3))$. In general, this drift condition guarantees the DUGS will drift toward states (μ, θ) for which $V(\mu, \theta)$ is small, that is, states for which μ is close to the sample mean $\overline{y}$. (General intuition for drift conditions is provided in the next section.)

An *associated* minorization condition will hold on set $C = \{(\mu, \theta) : \ V(\mu, \theta) \leq d\}$ for $d > 2L/(1-\gamma)$ if there exist density q and $\varepsilon > 0$ for which

$$k\left((\mu', \theta'), (\mu, \theta)\right) \geq \varepsilon q(\mu, \theta) \quad \text{for all } \ (\mu', \theta') \in C \text{ and } (\mu, \theta) \in \mathbb{R} \times \mathbb{R}_+. \tag{2.7}$$

To this end, first notice that for any $(\mu', \theta') \in C$ and $(\mu, \theta) \in \mathbb{R} \times \mathbb{R}_+$,

$$k\left((\mu', \theta'), (\mu, \theta)\right) = \pi(\mu|\theta, y)\pi(\theta|\mu', y) \geq \pi(\mu|\theta, y) \inf_{(\mu', \theta') \in C} \pi(\theta|\mu', y) \ .$$

Let $\mathrm{IG}(a, b; x)$ denote the value of the $\mathrm{IG}(a, b)$ density at the point $x > 0$. Some work shows that

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

where $\theta^* = md[(m-1)\log(1 + md/s^2)]^{-1}$. It follows that for any $(\mu', \theta') \in C$ and $(\mu, \theta) \in \mathbb{R} \times \mathbb{R}_+$,

$$k\left((\mu', \theta'), (\mu, \theta)\right) \geq \pi(\mu|\theta, y)g(\theta) = \varepsilon q(\mu, \theta)$$

where $q(\mu, \theta) = \varepsilon^{-1}\pi(\mu|\theta, y)g(\theta)$ and

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

Therefore, a minorization condition (2.7) holds.

2.3 Understanding Drift and Minorization

From their definitions alone, it is not entirely obvious that drift and minorization guaranteee convergence to stationarity, let alone geometric ergodicity. Here we present a brief discussion to provide some intuition into the connections between these concepts. Some other useful references include Jones and Hobert (2001), Lindvall (1992), Meyn and Tweedie (1993), Roberts and Tweedie (1999), and Rosenthal (1995a). This discussion will also serve as an outline for a proof of Theorem 2.2.

Assume throughout that Markov chain Φ is Harris ergodic with transition kernel P and invariant distribution π. From Theorem 2.2 it follows that Φ converges to stationarity. That is, for any $x \in \mathcal{X}$

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

A *coupling argument* provides insight into this result. This requires the following observation from Jain and Jamison (1967).

Theorem 2.4 (Jain and Jamison (1967))**.**
Let Φ be a ϕ-irreducible Markov chain on $(\mathcal{X}, \mathcal{B})$. Then there exists some small set $C \in \mathcal{B}$ for which $\phi(C) > 0$. Furthermore, the corresponding minorization measure $Q(\cdot)$ can be defined so that $Q(C) > 0$.

Let $C \in \mathcal{B}$ be some small set (which exists by Theorem 2.4). In this case, there exists m, $\varepsilon > 0$, and probability measure Q such that for any $A \in \mathcal{B}$,

$$P^m(x, A) \geq \varepsilon Q(A) \quad \text{for all } x \in C .$$

For simplicity we will assume $m = 1$. (For an argument when m is any positive integer, see Roberts and Rosenthal (2004).) In this case we can write the transition

kernel P as

$$P(x, A) = \varepsilon Q(A) + (1 - \varepsilon)R(x, A) \quad \text{for all } x \in C \text{ and } A \in \mathcal{B} \tag{2.8}$$

where $R(x, A) = (1 - \varepsilon)^{-1}(P(x, A) - \varepsilon Q(A))$. Further, for all $x \in C$, $R(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$ since $R(x, \mathcal{X}) = 1$ and the minorization condition guarantees $R(x, A) \geq 0$ for all $A \in \mathcal{B}$. Therefore, we can *split* $P(x, \cdot)$ into a mixture of probability measures $Q(\cdot)$ and $R(x, \cdot)$ where only $R(x, \cdot)$ depends on x.

The mixture representation of P can be used to construct two separate but dependent chains that eventually *couple* with probability one. Let $\Phi_X = \left\{X^{(0)}, X^{(1)}, \cdots\right\}$ and $\Phi_Y = \left\{Y^{(0)}, Y^{(1)}, \cdots\right\}$ and update the chains from $\left(X^{(n)}, Y^{(n)}\right)$ to $\left(X^{(n+1)}, Y^{(n+1)}\right)$ using the following algorithm:

1. While $X^{(n)} \neq Y^{(n)}$,

   If $\left(X^{(n)}, Y^{(n)}\right) \notin C \times C$:

   - Draw $X^{(n+1)} \sim P\left(X^{(n)}, \cdot\right)$ and $Y^{(n+1)} \sim P\left(Y^{(n)}, \cdot\right)$ independently.

   If $\left(X^{(n)}, Y^{(n)}\right) \in C \times C$:

   - Draw $\delta_n \sim \text{Bern}(\varepsilon)$.
   - If $\delta_n = 0$, draw $X^{(n+1)} \sim R\left(X^{(n)}, \cdot\right)$ and $Y^{(n+1)} \sim R\left(Y^{(n)}, \cdot\right)$ independently.
   - Otherwise, if $\delta_n = 1$, draw $X^{(n+1)} = Y^{(n+1)} \sim Q(\cdot)$.

2. Once $X^{(n)} = x = Y^{(n)}$, draw $X^{(n+1)} = Y^{(n+1)} \sim P(x, \cdot)$.

Remark 2.2. The coupling construction maintains the structure of the transition kernel P for both chains Φ_x and Φ_y.

Define *coupling time,* T, to be the random time at which the chains couple or become equal. That is, T is the first time n for which $\left(X^{(n-1)}, Y^{(n-1)}\right) \in C \times C$ and $\delta_{n-1} = 1$. Using the above algorithm, once the chains couple, they remain equal.

The distribution of T provides an upper bound on the total variation distance of Φ to stationarity. First, assume $X^{(0)} = x$ and $Y^{(0)} \sim \pi$ and let $\Pr_x$ denote probability with respect to starting value x. Then Φ_y is stationary and for any $A \in \mathcal{B}$

$$\begin{aligned}
\left| P^n(x, A) - \pi(A) \right| &= \left| \Pr_x\left(X^{(n)} \in A\right) - \Pr_x\left(Y^{(n)} \in A\right) \right| \\
&= \Big| \Pr_x\left(X^{(n)} \in A, X^{(n)} = Y^{(n)}\right) + \Pr_x\left(X^{(n)} \in A, X^{(n)} \neq Y^{(n)}\right) \\
&\quad - \Pr_x\left(Y^{(n)} \in A, X^{(n)} = Y^{(n)}\right) - \Pr_x\left(Y^{(n)} \in A, X^{(n)} \neq Y^{(n)}\right) \Big| \\
&= \left| \Pr_x\left(X^{(n)} \in A, X^{(n)} \neq Y^{(n)}\right) - \Pr_x\left(Y^{(n)} \in A, X^{(n)} \neq Y^{(n)}\right) \right| \\
&\leq \max\left\{ \Pr_x\left(X^{(n)} \in A, X^{(n)} \neq Y^{(n)}\right), \Pr_x\left(Y^{(n)} \in A, X^{(n)} \neq Y^{(n)}\right) \right\} \\
&\leq \Pr_x\left(X^{(n)} \neq Y^{(n)}\right) \\
&= \Pr_x(T > n) \, .
\end{aligned}$$

It follows that

$$\begin{aligned}
\| \, P^n(x, \cdot) - \pi(\cdot) \, \| &= \sup_{A \in \mathcal{B}} | \, P^n(x, A) - \pi(A) \, | \\
&\leq \sup_{A \in \mathcal{B}} \Pr_x(T > n) \\
&= \Pr_x(T > n) \, .
\end{aligned}$$

That is, the following *coupling inequality* is satisfied:

$$\| \, P^n(x, \cdot) - \pi(\cdot) \, \| \leq \Pr_x(T > n) \, . \tag{2.9}$$

To consider the implications of (2.9), we begin with the simple case when (2.4) holds on the entire state space, ie. $C = \mathcal{X}$. In this case, $\left(X^{(n)}, Y^{(n)}\right) \in C \times C$ for all n so that T is the first time n for which $\delta_{n-1} = 1$. Since $\Pr(\delta_{n-1} = 1) = \varepsilon$, $T \sim \text{Geo}(\varepsilon)$ for any starting value x where we say $W \sim \text{Geo}(p)$ if $\Pr(W > w) = (1 - p)^w$.

Therefore, when $C = \mathcal{X}$, the chain is uniformly ergodic since

$$\| P^n(x, \cdot) - \pi(\cdot) \| \leq \Pr_x(T > n) = (1 - \varepsilon)^n \ .$$

It is easy to show that $C = \mathcal{X}$ (and therefore uniform ergodicity) holds for any Harris ergodic Markov chain on a finite state space. This property is less common for Markov chains on general state spaces. The following is a special case.

Example 2.3.

Consider the independence Metropolis sampler with proposal density $q(\cdot)$ and stationary density $\pi(\cdot)$. The corresponding Markov chain is constructed as follows:

1. Select initial value $X^{(0)}$.
2. On the tth iteration, suppose $X^{(t-1)} = x$ and generate $X^{(t)}$ as follows:
   - Draw candidate value $y \sim q(\cdot)$ and $u \sim \text{Uniform}(0, 1)$, independently.
   - Calculate
   $$\alpha(x, y) = \min \left\{ \frac{\pi(y)q(x)}{\pi(x)q(y)}, \ 1 \right\} \ .$$
   - If $u < \alpha(x, y)$, set $X^{(t)} = y$.

     Otherwise, set $X^{(t)} = x$.
3. Repeat step 2.

Mengersen and Tweedie (1996) show this sampler is uniformly ergodic if the tails of q are sufficiently fat in comparison to those of π; that is, if there exists some $\kappa > 0$ for which

$$\frac{\pi(x)}{q(x)} \leq \kappa \quad \text{for any } x \in \mathcal{X} \ . \tag{2.10}$$

In this case, the following one-step minorization condition holds:

$$P(x, A) \geq \kappa^{-1}\pi(A) \quad \text{for any } x \in \mathcal{X}$$

since for all $x \in \mathcal{X}$,

$$\begin{aligned} P(x, A) &\geq \int_A q(y)\alpha(x, y)dy \\ &= \int_A \min\left\{\pi(y)\frac{q(x)}{\pi(x)},\ q(y)\right\} dy \\ &\geq \int_A \min\left\{\kappa^{-1}\pi(y),\ q(y)\right\} dy \\ &= \int_A \kappa^{-1}\pi(y)dy \\ &= \kappa^{-1}\pi(A)\ . \end{aligned}$$

In conjunction with the coupling inequality, this gives

$$\parallel P^n(x, \cdot) - \pi(\cdot) \parallel \leq (1 - \kappa^{-1})^n\ .$$

On the other hand, Mengersen and Tweedie (1996) show the chain is *subgeometric* if for all $\kappa > 0$ there exists $A \in \mathcal{B}$ with positive π-measure on which (2.10) does not hold. That is, there exists set A for which $\pi(A) > 0$ and

$$\frac{\pi(x)}{q(x)} > \kappa \quad \text{for all } x \in A\ .$$

As an illustration, we return to the independence sampler considered in Chapter 1.2 with an Exp(1) target and Exp(θ) proposal distribution. In this case, for any $x > 0$

$$\frac{\pi(x)}{q(x)} = \frac{\exp\{-x\}}{\theta \exp\{-\theta x\}} = \theta^{-1}\exp\{x(\theta - 1)\}\ .$$

If $\theta \in (0,1)$,

$$\frac{\pi(x)}{q(x)} \leq \theta^{-1} \quad \text{for any } x > 0$$

so that (2.10) holds with $\kappa = \theta^{-1}$. Therefore, the chain is uniformly ergodic if $\theta \in (0,1)$. On the other hand, the chain is subgeometric if $\theta > 1$. To this end, let κ be any positive value and define set $A = \{x : x > \log\{\theta\kappa\}/(\theta - 1)\}$. Then if $\theta > 1$, $\pi(A) > 0$ and for any $x \in A$

$$\frac{\pi(x)}{q(x)} = \theta^{-1} \exp\{x(\theta - 1)\} > \theta^{-1} \exp\{\log\{\theta\kappa\}\} = \kappa \, .$$

When $C \neq \mathcal{X}$, the distribution of T is typically complicated. Therefore, it is not always possible to derive $\Pr_x(T > n)$. However, Harris ergodicity guarantees T is finite almost surely for any starting value x. That is, for any $x \in \mathcal{X}$, $\Pr_x(T < \infty) = 1$ and $\Pr_x(T > n) \to 0$ as $n \to \infty$. (For discussions of this result, see Chapter 4.6 of Roberts and Rosenthal (2004) or Theorem 11.2 of Lindvall (1992).) In conjunction with the coupling inequality, this guarantees that for any $x \in \mathcal{X}$,

$$\| \, P^n(x, \cdot) - \pi(\cdot) \, \| \leq \Pr_x(T > n) \to 0 \quad \text{as } n \to \infty \, .$$

Furthermore, the convergence to stationarity is *geometric* if T has a thin-tailed distribution. Specifically, when T has a thin-tailed distribution, there exists some $\beta > 1$

for which $\mathrm{E}(\beta^T) < \infty$. Then by the coupling inequality and dominated convergence,

$$\begin{aligned}
\lim_{n\to\infty} \beta^n \parallel P^n(x,\cdot) - \pi(\cdot) \parallel &\leq \lim_{n\to\infty} \beta^n \mathrm{Pr}(T > n) \\
&= \lim_{n\to\infty} \mathrm{E}[\beta^n I(T > n)] \\
&\leq \lim_{n\to\infty} \mathrm{E}[\beta^T I(T > n)] \\
&= \mathrm{E}[\beta^T \lim_{n\to\infty} I(T > n)] \\
&= 0 \, .
\end{aligned}$$

It follows that the chain is geometrically ergodic with $\parallel P^n(x,\cdot) - \pi(\cdot) \parallel = o(\beta^{-n})$.

Though Harris ergodicity guarantees T is almost surely finite, it is *not* a sufficient condition for T to have a thin-tailed distribution. To this end, it is enough to establish a drift and associated minorization condition. First, suppose the following Type II drift condition holds:

$$PW(x) \leq \rho W(x) + bI_C(x) \quad \text{for all } x \in \mathcal{X} \tag{2.11}$$

where $C = \{x : W(x) \leq d\}$. Also suppose a one-step minorization condition (hence (2.8)) holds on C. Then C represents the "center" of the state space and coupling occurs when Φ_X and Φ_Y both reach this set *and* a success is drawn from $\mathrm{Bern}(\varepsilon)$. For T to have a thin-tailed distribution (ie. for coupling to occur reasonably quickly) it is necessary that Φ_X and Φ_Y make *frequent* visits to C. This behavior is guaranteed by the drift condition. First, it follows from (2.11) that

$$\Delta W(x) \leq -(1-\rho)W(x) + bI_C(x) \quad \text{for all } x \in \mathcal{X} \tag{2.12}$$

where $\Delta W(x) = PW(x) - W(x)$. Therefore, when the chain is in some state x outside set C, $\Delta W(x) \leq -(1-\rho)W(x)$ which implies $PW(x) \leq \rho W(x)$. Thus, geometric

drift guarantees that when either chain leaves the center of the state space (set C) it tends to drift back quickly. Further, the rate of the drift is controlled by ρ. The closer ρ is to 1, the slower the drift. Also, drift occurs more quickly for states x further outside of C (those for which $W(x)$ is large).

Remark 2.3. By Condition 1 of Lemma 2.1, (2.12) implies drift condition (2.3) holds for $V(x) = W(x)$, $\gamma = \rho$, and $L = b$. In this translation, a value of γ close to 1 also reflects a slow drift of the Markov chain to the center of the state space.

Next, the minorization condition guarantees that every time Φ_X and Φ_Y are both in set C, they couple with probability ε. Therefore, they need only drift back to C (concurrently) a geometric number of times before coupling occurs. Since the drift condition ensures that visits to set C are frequent, minorization and drift together guarantee a thin-tailed distribution for coupling time T.

Finally, notice the trade-off between the size of set C and the magnitude of ε in the coupling time. By the one-step minorization condition,

$$P(x, A) \geq \varepsilon Q(A) \quad \text{for all } x \in C \, .$$

It typically follows that the larger the set C, the smaller ε must be in order for the minorization condition to hold. Similarly, the smaller the set C, the larger ε will typically be. Therefore, when C is large, Φ_X and Φ_Y have a higher probability of concurrently visiting set C, yet a smaller probability ε of coupling once there. On the other hand, when C is small, Φ_X and Φ_Y have a *smaller* probability of concurrently visiting set C, yet a *larger* probability ε of coupling once there.

Chapter 3

Implications of Geometric Ergodicity

Two fundamental questions that arise in the implementation of MCMC algorithms and subsequent use of Markov chain output for inference are

(Q1) When has the Markov chain converged to stationarity?; and

(Q2) For how long should we run the Markov chain? That is, how much simulation effort is required for the Markov chain estimates to achieve pre-specified levels of accuracy?

(Q1) concerns the distance of the chain from the target distribution whereas (Q2) concerns the quality of Markov chain estimates (or the distance of these estimates from "the truth"). Though related, these questions require separate treatment.

(Q1) and (Q2) are often addressed on an ad-hoc basis, drawing from a combination of one's own experience and popular practice. Indeed, due to the dependence of (Q1) and (Q2) on the target distribution and choice of MCMC algorithm, formal answers are typically either not readily available or not widely known. Establishing geometric ergodicity allows us to construct rigorous answers to these questions, thus eliminating the reliance on heuristic and sometimes specious arguments. (Q1) and (Q2) are addressed in Chapters 3.1 and 3.2, respectively.

3.1 Assessing Convergence

For a Markov chain sample to be representative of target distribution π, it is important that sampling continue beyond convergence to stationarity. Unfortunately, in practice we have no choice but to stop simulation after a finite number of iterations. Therefore, even if the chain is geometrically ergodic, we risk terminating the simulation before the asymptotics have had a chance to kick in. We illustrate this worst-case scenario using a toy example.

Example 3.1. Witch's Hat

Consider the target distribution on $\mathcal{X} = [0, 1]$ with unnormalized density $\pi_u(x) = b + I_{[a,a+b]}(x)$ where $b = 10^{-100}$ and $0 < a < 1 - b$. This target distribution is aptly referred to as a "Witch's Hat":

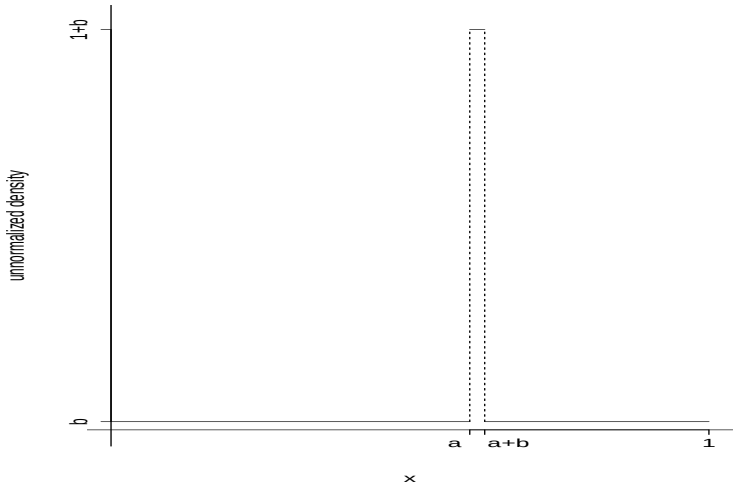


Figure 3.1: The unnormalized witch's hat density $\pi_u(x)$.

Notice that the normalized density can be written as $\pi(x) = (2b)^{-1}\pi_u(x)$. It follows that $\pi([a, a + b]) \approx 1/2$. In other words, approximately half of the mass of the target distribution is contained in the very narrow interval $[a, a + b]$. A typical Metropolis-

Hastings algorithm for π whose proposal distribution has support on $[0, 1]$ will be geometrically ergodic. However, unless we are very lucky, the chain will never visit interval $[a, a + b]$ in a fixed number of iterations. Moreover, the chain will appear to converge to the uniform distribution on $[0, 1]$. Without rigorous answers to (Q1), this phenomenon would likely cause the user to stop the chain with a false sense of confidence in the output.

This example illustrates the importance of obtaining rigorous answers to (Q1). To this end, it suffices to find n for which

$$\| P^n(x, \cdot) - \pi(\cdot) \| < \omega \tag{3.1}$$

for some chosen $\omega > 0$. In other words, it suffices to find n for which the chain is "close enough" to stationarity after n iterations. If n satisfies (3.1), the chain *remains* within ω of π since the distance to stationarity is monotone nonincreasing.

In some special cases, n for which (3.1) holds is easily derived. For instance, suppose a one-step minorization condition holds on the whole space: $P(x, A) \geq \varepsilon Q(A)$ for all $x \in \mathcal{X}$ and $A \in \mathcal{B}$. In this case, the chain is uniformly ergodic with

$$\| P^n(x, \cdot) - \pi(\cdot) \| < (1 - \varepsilon)^n$$

and it is straightforward to find n for which $(1-\varepsilon)^n < \omega$ (see Chapter 2.3 for details). Unfortunately, Markov chains on general state spaces are typically not uniformly ergodic. Further, decisions regarding Markov chain convergence are often based on intuition, experience, or so-called *convergence diagnostics*. For a review of convergence diagnostics see Cowles and Carlin (1996). At worst, these methods are known to fail for even simple toy examples (again, see Cowles and Carlin (1996)). At best, they do

not produce n for which (3.1) is *guaranteed* to hold.

Example 3.1 illustrates that geometric ergodicity is an *asymptotic* property. It does *not* ensure that a sampler is well-behaved in a finite number of iterations nor does it always ensure improved performance. However, under geometric ergodicity, there *do* exist rigorous approaches to answering (Q1). First, geometric ergodicity guarantees the existence of some $M : \mathcal{X} \to \mathbb{R}$ and $t \in (0, 1)$ for which

$$\| P^n(x, \cdot) - \pi(\cdot) \| \leq M(x)t^n \quad \text{for any } x \in \mathcal{X} .$$

When M and t are available, it is straightforward to derive n for which (3.1) holds. However, M and t are seldom known. Here we discuss two alternative techniques for assessing convergence that utilize drift and minorization conditions. (Therefore, establishing geometric ergodicity is not our only motivation for their construction!) Specifically, drift and minorization can be used to construct upper bounds on the distance of the chain to stationarity (Rosenthal, 1995a) and to construct a high quality approximation to the target distribution (Hobert and Robert, 2004).

In addition to deriving formal answers to (Q1), these techniques can be used to rigorously address the so-called *burn-in* problem. Many MCMC practitioners prefer to defer sampling until the chain is sufficiently close to stationarity. A common remedy is to allow the chain to *burn in*, tossing the samples at the start of the chain which do not meet the convergence criterion. Specifically, if n is the smallest integer that satisfies (3.1), then $\left\{X^{(0)}, X^{(1)}, \cdots X^{(n-1)}\right\}$ is the burn-in sample. Therefore, burn-in can be seen as a method for finding an initial distribution of the form P^n. However, the ergodic theorem holds for *any* initial distribution. We will also see that the same is true of Markov chain central limit theorems. Hence, it can be argued that as long as we choose a reasonable starting value, there is no theoretical value in tossing out burn-in samples. In fact, burn-in is viewed by some as wasteful. By tossing out

valuable information with the burn-in samples, this practice sacrifices a certain degree of accuracy for a reduction in bias. For more on this see Charlie Geyer's website, "Burn-in is Unnecessary," at *http://www.stat.umn.edu/∼charlie/mcmc/burn.html*.

3.1.1 Rosenthal's bound

Rosenthal (1995a) derives an upper bound on the total variation distance of the chain to stationarity using drift and minorization. This bound can be used to derive n for which (3.1) is guaranteed to hold. In turn, this provides an answer to (Q1).

Theorem 3.1. *Rosenthal (1995a)*
Let Φ be an aperiodic and irreducible Markov chain with invariant distribution π. Suppose Φ satisfies the following drift and associated one-step minorization condition: For some drift function $V : \mathcal{X} \to \mathbb{R}_+$, drift rate $0 < \gamma < 1$, and $L < \infty$,

$$PV(x) \leq \gamma V(x) + L \quad \text{for all } x \in \mathcal{X} \, .$$

Also, for some probability measure Q on $\mathcal{B}$ and some $\varepsilon > 0$,

$$P(x, A) \geq \varepsilon Q(A) \quad \text{for all } x \in C \text{ and } A \in \mathcal{B}$$

where $C = \{x : V(x) \leq d\}$ for some $d > 2L/(1-\gamma)$. Let $X^{(0)} = x_0$ be the starting value and define

$$\alpha = \frac{1+d}{1+2L+\gamma d} \quad \text{and} \quad U = 1 + 2(\gamma d + L).$$

Then for any $0 < r < 1$,

$$\| \, P^n(x_0, \cdot) - \pi(\cdot) \, \| \leq (1-\varepsilon)^{rn} + \left(\frac{U^r}{\alpha^{1-r}} \right)^n \left(1 + \frac{L}{1-\gamma} + V(x_0) \right). \tag{3.2}$$

Remarks:

1. Proposition 2.1 follows directly from Theorem 3.1. First, under the assumptions of Proposition 2.1, Rosenthal's result guarantees that

$$\| P^n(x_0, \cdot) - \pi(\cdot) \| \leq \left((1-\varepsilon)^r \vee \frac{U^r}{\alpha^{1-r}} \right)^n \left(2 + \frac{L}{1-\gamma} + V(x_0) \right).$$

   From (2.2), geometric ergodicity follows if $U^r/\alpha^{1-r} < 1$. To this end, notice that $d > 2L/(1-\gamma)$ guarantees $\alpha^{-1} < 1$. Therefore, setting r sufficiently close to 0 guarantees

$$\frac{U^r}{\alpha^{1-r}} = \alpha^{-1} (U\alpha)^r < 1 .$$

2. Recall from the remarks following Proposition 2.1 that starting the Markov chain at x for which $V(x)$ is minimized will hasten convergence. This is also reflected by Rosenthal's bound since (3.2) is minimized at this same value.

3. In our experience, (3.2) is sensitive to the values of r and d. Appropriate choices can be made by searching over a sensible range of values for the two variables. When possible, (r, d) should at least guarantee $U^r/\alpha^{1-r} < 1$ so that the upper bound is guaranteed to decrease as n increases.

4. Rosenthal's bound tends to be conservative. See Rosenthal (1995a) and Jones and Hobert (2004) for examples of when (3.2) is not sharp. A potentially tighter bound is given by Roberts and Tweedie (1999). However, Roberts and Tweedie require a drift condition of the form (2.6) which we find to be slightly more difficult to establish than (2.3).

We conclude this section with a simple toy example to illustrate the calculation of Rosenthal's bounds.

Example 2.2 (Continued)

Consider Example 2.2 in Chapter 2.2 with $m = 5$, $\overline{y} = 4$, and $s^2 = 10$. Then a Type I drift condition holds with $V(\mu, \theta) = (\mu - \overline{y})^2$, $\gamma = 1/2$, and $L = 1$:

$$PV\,(\mu, \theta) \leq \frac{1}{2} V\,(\mu, \theta) + 1\ .$$

In addition, an associated minorization condition holds on set $C = \{(\mu, \theta) : V(\mu, \theta) \leq d\}$ for any $d > 2L/(1-\gamma) = 4$. Setting $d = 6$, the minorization condition holds with $\varepsilon = \int_{\mathbb{R}_+} g(\theta) d\theta$ where

$$g(\theta) = \begin{cases} \text{IG}(2, 20; \theta) & \text{if } \theta < \theta^* \\ \text{IG}(2, 5; \theta) & \text{if } \theta \geq \theta^* \end{cases}$$

for $\theta^* = 30[4\log(4)]^{-1}$. From R, we find that $\varepsilon \approx 0.35$:

```
> tstar <- 30*(4*log(4))^(-1)
> invgam <- function(theta,a,b){
  #The inverse gamma density with parameters a and b
  b^a/gamma(a) * theta^(-a-1) * exp(-b/theta)
  }
> integrate(invgam, lower=0, upper=tstar, a=2, b=20)$value
  + integrate(invgam, lower=tstar, upper=Inf, a=2, b=5)$value
[1] 0.3528772
```

It also follows from $d = 6$ that $\alpha = 7/6$ and $U = 9$. In addition, (3.2) is minimized by starting from $\mu_0 = \overline{y}$ since this minimizes $V(\mu_0, \theta_0)$ at $V(\overline{y}, \theta_0) = 0$ for any θ_0. To construct (3.2), it only remains to choose $r \in (0, 1)$. In a grid search over a range of r values, $r = 0.05$ produced relatively small bounds in comparison to other values of

r (for fixed n). Then (3.2) with $d = 6$, $r = 0.05$, and $\mu_0 = \overline{y}$ gives

$$\| P^n((\mu_0, \theta_0), \cdot) - \pi(\cdot) \| \leq (0.9787)^n + 3(0.9641)^n$$

which guarantees

$$\begin{aligned} \| P^n((\mu_0, \theta_0), \cdot) - \pi(\cdot) \| &\leq 0.01 \qquad \text{for } n \geq 220; \text{ and} \\ \| P^n((\mu_0, \theta_0), \cdot) - \pi(\cdot) \| &\leq 0.001 \qquad \text{for } n \geq 325 \, . \end{aligned}$$

That is, the total variation distance between the chain and $\pi(\mu, \theta|y)$ is guaranteed to be less than 0.01 after 220 iterations and less than 0.001 after 325 iterations. As an illustration, we ran 1000 independent Gibbs samplers for 325 iterations each, each starting from $\mu_0 = \overline{y} = 4$. Histograms of the 1000 corresponding copies of $\theta^{(220)}$ and $\theta^{(325)}$ are given in Figure 3.2. For better viewing, the horizontal axes only extend to 60. However, values above this threshold were observed.

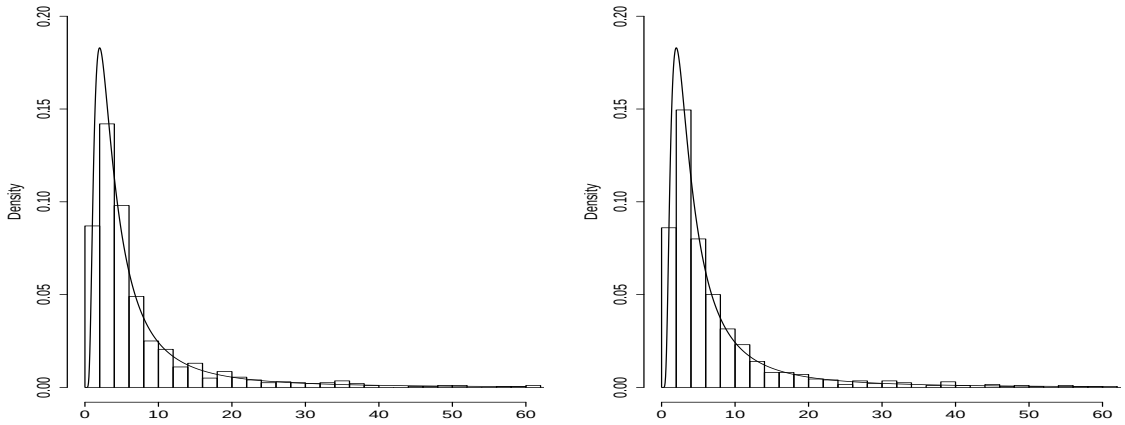


Figure 3.2: Histograms of the 1000 independent copies of $\theta^{(220)}$ (left) and $\theta^{(325)}$ (right). The IG$(1.5, 5)$ density is super-imposed on each histogram (solid line).

Some work shows that $\theta|y \sim \text{IG}((m-2)/2, s^2/2)$ where, for this example, $(m-2)/2 = 1.5$ and $s^2/2 = 5$. The corresponding density $\pi(\theta|y)$ is included in both plots.

The results support what is known from Rosenthal's bounds. That is, there is little discrepancy between $\pi(\theta|y)$ and the distributions of $\theta^{(220)}$ and $\theta^{(325)}$. Further, this discrepancy is smaller for $\theta^{(325)}$ than for $\theta^{(220)}$.

3.1.2 Hobert and Robert's approximation

Consider a Markov chain with invariant distribution $\pi(\cdot)$. For any $\omega > 0$, Hobert and Robert (2004) show that drift and minorization conditions can be used to construct λ so that

$$\| \lambda(\cdot) - \pi(\cdot) \| \leq \omega \, . \tag{3.3}$$

Then if $x_0 \sim \lambda(\cdot)$,

$$\| P^n(x_0, \cdot) - \pi(\cdot) \| \leq \omega \quad \text{for all } n \geq 1 \, .$$

In this case, (Q1) is moot since the chain is started "close enough" to stationarity. We provide a brief overview of Hobert and Robert's result here.

First, suppose the following one-step minorization condition holds on set $C \in \mathcal{B}$:

$$P(x, A) \geq \varepsilon Q(A) \quad \text{for all } x \in C \text{ and } A \in \mathcal{B} \tag{3.4}$$

where $Q(\cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$. Recall from Chapter 2.3 that (3.4) implies $P(x, \cdot)$ can be split into the following mixture distribution:

$$P(x, A) = \varepsilon Q(A) + (1 - \varepsilon) R(x, A) \quad \text{for all } x \in \mathcal{C} \text{ and } A \in \mathcal{B}$$

where $R(x, A) = (1-\varepsilon)^{-1}[P(x, A) - \varepsilon Q(A)]$. Then the *split chain* algorithm produces a Markov chain with transition kernel P.

The Split Chain Algorithm:

1. Select initial value $\left(X^{(0)}, \delta_0\right)$.

2. On the tth iteration, generate $X^{(t)}$ and δ_{t-1} as follows:

   - If $X^{(t-1)} \notin C$, draw $X^{(t)} \sim P\left(X^{(t-1)}, \cdot\right)$.
   - If $X^{(t-1)} \in C$, draw $\delta_{t-1} \sim \text{Bern}(\varepsilon)$.

     If $\delta_{t-1} = 0$, draw $X^{(t)} \sim R\left(X^{(t-1)}, \cdot\right)$.

     If $\delta_{t-1} = 1$, draw $X^{(t)} \sim Q\left(\cdot\right)$.

3. Repeat step 2.

Remark 3.1. See Chapter 3.2.2 for an equivalent split chain algorithm that does not require sampling from $Q(\cdot)$.

Let $\left\{\left(X^{(t)}, \delta_t\right)\right\}$ be the corresponding *split chain*. Also, define $\alpha = C \times \{1\}$. The split chain reaches α with positive probability and the values $n+1$ for which $\left(X^{(n)}, \delta_n\right) \in \alpha$ are called *regeneration times*. At each regeneration time, the chain probabilistically restarts itself since conditional on $\left(X^{(n)}, \delta_n\right) \in \alpha$, $X^{(n+1)} \sim Q(\cdot)$ (ie. $X^{(n+1)}$ is independent of all past states). Further, suppose $\left(X^{(0)}, \delta_0\right) \in \alpha$ and let τ_α denote the time of the first return to α. Specifically,

$$\tau_\alpha = \inf\left\{n \geq 1 : \left(X^{(n)}, \delta_n\right) \in \alpha\right\}.$$

Hobert and Robert (2004) show that a mixture distribution for π can be derived from the distribution of τ_α.

Theorem 3.2. *Hobert and Robert (2004)*
Suppose Markov chain Φ is Harris ergodic and satisfies minorization condition (3.4). Also, let Pr_α and E_α denote probability and expectation conditional on $\left(X^{(0)}, \delta_0\right) \in \alpha$,

respectively. Then for any $A \in \mathcal{B}$,

$$\pi(A) = \sum_{t=1}^{\infty} U_t(A) p_t \tag{3.5}$$

where

$$U_t(A) = Pr_\alpha\left(X^{(t)} \in A \,\middle|\, \tau_\alpha \geq t\right) \qquad \textit{and} \qquad p_t = \frac{Pr_\alpha(\tau_\alpha \geq t)}{E_\alpha(\tau_\alpha)}.$$

In theory, the following algorithm can be used to produce iid samples from π:

1. Randomly select index $t \in \{1, 2, \ldots\}$ with probability p_t.
2. Draw $x \sim U_t(\cdot)$.

The implementation of this algorithm requires knowledge of the distribution of τ_α. As was seen for coupling time T in Chapter 2.3, this distribution is easy to define when (3.4) holds on the entire state space (ie. $C = \mathcal{X}$). In this case, τ_α denotes the first time $n \geq 1$ for which $\delta_n = 1$. Therefore, $\tau_\alpha \sim \text{Geo}(\varepsilon)$. This gives geometric probabilities $p_t = \varepsilon(1-\varepsilon)^{t-1}$ since $\Pr_\alpha(\tau_\alpha \geq t) = (1-\varepsilon)^{t-1}$ and $\mathrm{E}_\alpha(\tau_\alpha) = \varepsilon^{-1}$. Also, $C = \mathcal{X}$ and $\tau_\alpha \geq t$ together imply $\delta_1, \ldots, \delta_{t-1} = 0$. Hence for any $A \in \mathcal{B}$,

$$\begin{aligned}
U_t(A) &= \Pr_\alpha\left(X^{(t)} \in A | \tau_\alpha \geq t\right) \\
&= \Pr\left(X^{(t)} \in A | X^{(1)} \sim Q \text{ and } \delta_1, \ldots, \delta_{t-1} = 0\right) \\
&= \Pr\left(X^{(t)} \in A | X^{(1)} \sim Q,\, X^{(2)} \sim R(Q,\cdot),\, X^{(3)} \sim R^2(Q,\cdot), \ldots,\, X^{(t)} \sim R^{t-1}(Q,\cdot)\right) \\
&= \Pr\left(X^{(t)} \in A | X^{(t)} \sim R^{t-1}(Q,\cdot)\right) \\
&= R^{t-1}(Q, A)
\end{aligned}$$

where $R^{t-1}(Q, A)$ denotes the distribution that evolves by drawing $X^{(1)} \sim Q(\cdot)$ and applying R, $t-1$ times. Finally, the following algorithm for sampling from π when $C = \mathcal{X}$ follows from Theorem 3.2.

Hobert and Robert's Algorithm for sampling from π when $C = \mathcal{X}$:

1. Draw $X^{(1)} \sim Q(\cdot)$ and $t \sim \text{Geo}(\varepsilon)$ independently.
2. If $t = 1$, set $x = X^{(1)}$. Otherwise, simulate the transition $X^{(n+1)} \sim R\left(X^{(n)}, \cdot\right)$ for $n = 1, \ldots, t-1$ and set $x = X^{(t)}$.

As previously mentioned, it is uncommon for a minorization condition to hold on $C = \mathcal{X}$ in the general state space setting. Unfortunately, when $C \neq \mathcal{X}$ the distribution of τ_α can be complicated thus precluding the use of Theorem 3.2 for direct sampling from π. Specifically, it is difficult to derive $\text{E}_\alpha(\tau_\alpha)$ (hence p_t) when $C \neq \mathcal{X}$. However, Hobert and Robert (2004) show that for any $\omega > 0$, drift and minorization conditions can be used to construct probabilities $\tilde{p}_t$ for which

$$\sum_{t=1}^{\infty} |\tilde{p}_t - p_t| \leq \omega \ .$$

In turn, defining $\lambda(\cdot) = \sum_{t=1}^{\infty} U_t(\cdot)\tilde{p}_t$ guarantees $\lambda(\cdot)$ satisfies (3.3) since

$$\| \lambda(\cdot) - \pi(\cdot) \| = \left\| \sum_{t=1}^{\infty} U_t(\cdot)\tilde{p}_t - \sum_{t=1}^{\infty} U_t(\cdot)p_t \right\| \leq \sum_{t=1}^{\infty} |\tilde{p}_t - p_t| \ .$$

Theorem 3.3. *Hobert and Robert (2004)*
Let Φ be a Harris ergodic Markov chain and suppose Φ satisfies the following Type II drift condition. For $W : \mathcal{X} \to [1, \infty)$, $0 < \rho < 1$, set $C \in \mathcal{B}$, and $b < \infty$,

$$PW(x) \leq \rho W(x) + bI_C(x) \quad \textit{for all } x \in \mathcal{X} \ .$$

Also, suppose Φ satisfies the following minorization condition. For some probability measure Q on $(\mathcal{X}, \mathcal{B})$ and some $\varepsilon > 0$,

$$P(x, A) \geq \varepsilon Q(A) \quad \textit{for all } x \in C \textit{ and } A \in \mathcal{B}.$$

Let $d = \sup_{x \in C} W(x)$, $D = \sup_{x \in C} PW(x)$, $J = (D - \varepsilon)/\rho$, *and*

$$\eta^* = \begin{cases} \rho^{-1} & \text{if } J < 1 \\ \exp\left\{ \frac{\log \rho \log(1-\varepsilon)}{\log J - \log(1-\varepsilon)} \right\} & \text{if } J \geq 1 \end{cases}.$$

Furthermore, let $\phi(\eta) = \log \eta / \log \rho^{-1}$ *and*

$$g(\eta, \varepsilon, J) = \eta \left[\frac{b}{\varepsilon(1-\rho)} \right]^{\phi(\eta)} \left[\frac{1 - \eta(1-\varepsilon)}{1 - (1-\varepsilon)(J/(1-\varepsilon))^{\phi(\eta)}} \right].$$

Fix $\omega > 0$ *and* $\eta \in (1, \eta^*)$, *then*

$$\| \lambda(\cdot) - \pi(\cdot) \| \leq \omega$$

where

$$\lambda(\cdot) = \sum_{t=1}^{M} U_t(\cdot) \tilde{p}_t \qquad \text{for} \qquad \tilde{p}_t = \frac{Pr_\alpha(\tau_\alpha \geq t)}{\sum_{i=1}^{M} Pr_\alpha(\tau_\alpha \geq i)}, \tag{3.6}$$

U_t *as in Theorem 3.2, and any integer* M *such that*

$$M > (\log \eta)^{-1} \log \left[\frac{2g(\eta, \varepsilon, J)}{\omega(\eta - 1)} \right].$$

Remark 3.2. By Lemma 2.1, the Type II drift condition required by Theorem 3.3 can be constructed by one of the form (2.3) (if necessary).

Sampling from λ in Theorem 3.3 requires the generation of a random variable $\tilde{T}$ with $\Pr(\tilde{T} = t) = \tilde{p}_t$. To this end, let Z be uniform on $\{1, \ldots, M\}$ and notice that

$\tilde{p}_t = \Pr_\alpha(Z = t \,|\, \tau_\alpha \geq Z)$:

$$\begin{aligned}
\Pr_\alpha(Z = t \,|\, \tau_\alpha \geq Z) &= \frac{\Pr_\alpha(Z = t,\ \tau_\alpha \geq Z)}{\Pr_\alpha(\tau_\alpha \geq Z)} \\
&= \frac{\Pr_\alpha(Z = t)\Pr_\alpha(\tau_\alpha \geq t)}{\sum_{i=1}^{M} \Pr_\alpha(\tau_\alpha \geq i)\Pr_\alpha(Z = i)} \\
&= \frac{\Pr_\alpha(\tau_\alpha \geq t)}{\sum_{i=1}^{M} \Pr_\alpha(\tau_\alpha \geq i)} \\
&= \tilde{p}_t \ .
\end{aligned}$$

Then the following algorithm produces draws from $\lambda(\cdot)$.

Hobert and Robert's Algorithm for sampling from λ:

1. Take $\left(X^{(0)}, \delta_0\right) \in \alpha$.

2. Generate t with probability $\tilde{p}_t$ as follows:

   - Draw $z \sim \text{Unif}\{1, \ldots, M\}$ and w from the distribution of τ_α.
   - If $w \geq z$, set $t = z$; otherwise, repeat (ie. draw $z \sim \text{Unif}\{1, \ldots, M\}$ and w from the distribution of τ_α until $w \geq z$).

3. Simulate from $U_t(\cdot)$ given t as follows.

   - Starting from $\left(X^{(0)}, \delta_0\right) \in \alpha$, simulate the split chain for t iterations.
   - If $t = 1$, take $X^{(1)} \sim Q(\cdot)$.
   - If $t > 1$ and $\left(X^{(i)}, \delta_i\right) \notin \alpha$ for $i = 1, \ldots, t-1$, take $X^{(t)}$. Otherwise, repeat.

Remark 3.3. The above algorithm requires that the chain start from a regeneration, ie. $X^{(1)} \sim Q(\cdot)$. This can be accomplished in one of two ways: (i) start from an arbitrary point and discard the simulation up to the time of the first regeneration; or (ii) directly draw $X^{(1)} \sim Q(\cdot)$. Method (ii) will typically not be difficult. See Mykland et al. (1995) for examples.

Example 2.2 (Continued)

We again return to Example 2.2 in Chapter 2.2 with $m = 5$, $\overline{y} = 4$, and $s^2 = 10$. In Chapter 3.1.1 we saw that a Type I drift condition holds with $V(\mu, \theta) = (\mu - \overline{y})^2$, $\gamma = 1/2$, and $L = 1$:

$$PV(\mu, \theta) \leq \frac{1}{2} V(\mu, \theta) + 1 \, .$$

Applying Lemma 2.1 gives the following Type II drift condition (of the form required by Theorem 3.3). Let $W(\mu, \theta) = V(\mu, \theta) + 1$. Then for any $(\mu, \theta) \in \mathbb{R} \times \mathbb{R}_+$,

$$PW(\mu, \theta) \leq \frac{3}{4} W(\mu, \theta) + \frac{3}{2} I_C(\mu, \theta)$$

where $C = \{(\mu, \theta) : W(\mu, \theta) \leq 6\} = \{(\mu, \theta) : V(\mu, \theta) \leq 5\}$.

From the work in Chapter 2.2, a minorization condition holds on C with $\varepsilon = \int_{\mathbb{R}_+} g(\theta) d\theta$ where

$$g(\theta) = \begin{cases} \text{IG}(2, 17.5; \theta) & \text{if } \theta < \theta^* \\ \text{IG}(2, 5; \theta) & \text{if } \theta \geq \theta^* \end{cases}$$

for $\theta^* = 25[4 \log(3.5)]^{-1}$. From R, we find that $\varepsilon \approx 0.40$ (see the example in Chapter 3.1.1 for the relevant code). Now, in the notation of Theorem 3.3, $d = D = 6$ and $J = 7.47$. Therefore,

$$\eta^* = \exp\left\{ \frac{\log \rho \log(1 - \varepsilon)}{\log J - \log(1 - \varepsilon)} \right\} > 1.05 \, .$$

Set $\eta \in (1, \eta^*)$ to $\eta = 1.05$. In this case, $\phi(\eta) = 0.17$ and

$$g(\eta, \varepsilon, J) \approx 1.05[4.69] \, [15]^{0.17} \approx \ 7.80.$$

Define $\lambda(\cdot)$ as in (3.6), that is,

$$\lambda(\cdot) = \sum_{t=1}^{M} U_t(\cdot)\tilde{p}_t \qquad \text{for} \qquad \tilde{p}_t = \frac{\Pr_\alpha(\tau_\alpha \geq t)}{\sum_{i=1}^{M} \Pr_\alpha(\tau_\alpha \geq i)} \ .$$

Then $\| \lambda(\cdot) - \pi(\cdot) \| \leq \omega$ if

$$M > 20.50 \log\left[312\omega^{-1}\right] \ .$$

This guarantees

$$\begin{aligned} \| \lambda(\cdot) - \pi(\cdot) \| &\leq 0.01 \qquad \text{for } M \geq 214; \text{ and} \\ \| \lambda(\cdot) - \pi(\cdot) \| &\leq 0.001 \qquad \text{for } M \geq 261 \ . \end{aligned}$$

Simulation from $\lambda(\cdot)$ becomes more computationally expensive as M increases. However, for this example, the simulation effort required when $M = 214$ or $M = 261$ is not prohibitive. For both values of M, we used Hobert and Robert's algorithm to obtain a random sample $\{(\mu_i, \theta_i)\}_{i=1}^{1000}$ from

$$\lambda(\cdot) = \sum_{t=1}^{M} U_t(\cdot)\tilde{p}_t \ .$$

Each draw required simulating the split chain $\left\{\left(\left(\mu^{(t)}, \theta^{(t)}\right), \delta_t\right)\right\}$ starting from a regeneration. To this end, in each repetition we started the Gibbs sampler from $\mu_0 = \overline{y} = 4$ and set $(\mu^{(1)}, \theta^{(1)})$ to the value of the chain at the first regeneration.

For $M = 214$, obtaining 1000 random draws from $\lambda(\cdot)$ took 29.5 seconds on a Mac OS X (Version 10.5.7). For $M = 261$, the simulation took approximately 34.2 seconds. Histograms of the 1000 corresponding values of θ_i are given in Figure 3.3 for $M = 214$ (left) and $M = 261$ (right). Again, superimposed on each plot is the IG$(1.5, 5)$ density corresponding to the marginal posterior $\theta|y$. In addition, the

horizontal axes are truncated at 60 though values above this threshold were observed. These histograms illustrate the quality of $\lambda(\cdot)$ as an approximation of posterior density $\pi(\mu, \theta|y)$ where the discrepancy between the two densities decreases as M increases.

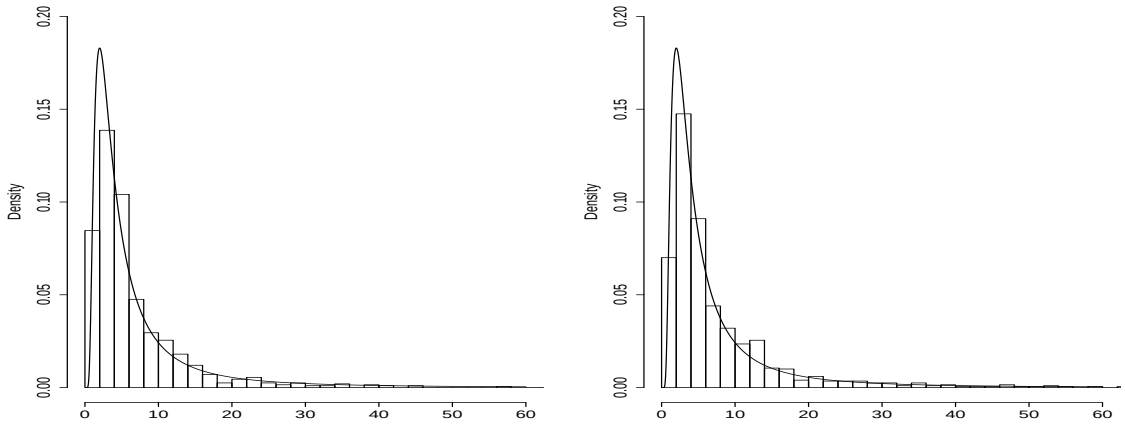


Figure 3.3: Histograms of the 1000 independent values of θ drawn from $\lambda(\cdot)$ corresponding to $M = 214$ (left) and $M = 261$ (right). The IG$(1.5, 5)$ density is superimposed on each histogram (solid line).

3.2 Assessing the Accuracy of Markov Chain Estimates

Let π denote some target distribution and suppose we are interested in evaluating

$$\mathrm{E}_\pi g := \int_{\mathcal{X}} g(x)\pi(dx)$$

for some $g : \mathcal{X} \to \mathbb{R}$. When $\mathrm{E}_\pi g$ is intractable, we can estimate it using the Markov chain average

$$\overline{g}_n = \frac{1}{n}\sum_{i=0}^{n-1} g\left(X^{(i)}\right) \ .$$

For every initial distribution, $\overline{g}_n$ is strongly consistent for $\mathrm{E}_\pi g$ (Theorem 2.1). This guarantees the estimate approaches the truth as the chain evolves. However, it does not provide insight into one obvious question: How long is "long enough"? That is, how long do we need to run the chain for the Monte Carlo error $\overline{g}_n - \mathrm{E}_\pi g$ to be sufficiently small?

Unfortunately, MCMC estimates are often reported with no mention of the corresponding Monte Carlo error. In such cases, it is nearly impossible to objectively evaluate the accuracy of the estimates or the sufficiency of the simulation length. Geometric ergodicity guarantees the existence of rigorous methods for treating these issues. For instance, drift and minorization conditions can be used to derive bounds for sample size n that guarantee

$$\Pr\left(\left|\overline{g}_n - \mathrm{E}_\pi g\right| \leq \omega\right) \geq 1 - \alpha$$

for user-specified ω and α (Latuszynski, 2008). However, these bounds are admittedly conservative and, in our experience, difficult to implement. Therefore, we omit this method from our discussion. Instead, we focus on a more classical approach to answering (Q2) founded in the existence of Markov chain central limit theorems (CLT) and consistent Monte Carlo standard errors (MCSE).

3.2.1 The Markov Chain Central Limit Theorem

Since $\mathrm{E}_\pi g$ is unknown, we cannot directly evaluate the Monte Carlo error $\overline{g}_n - \mathrm{E}_\pi g$. However, deriving the sampling distribution of the error (when possible) provides insight into the accuracy of $\overline{g}_n$. Consider the simple case when $\left\{X^{(0)}, \ldots, X^{(n-1)}\right\}$ is a random sample from π. If $\mathrm{E}_\pi g(X)^2 < \infty$, the usual CLT guarantees

$$\sqrt{n}(\overline{g}_n - \mathrm{E}_\pi g) \xrightarrow{d} N(0, \sigma^2)$$

where $\sigma^2 := \mathrm{Var}_\pi(g(X))$.

In the typical MCMC setting, it is not possible to obtain iid samples from π. Thus the usual CLT is not available. However, central limit theorems for ergodic averages *do* exist under certain conditions, including geometric ergodicity.

Theorem 3.4. *Let Φ be a Harris ergodic Markov chain and $g : \mathcal{X} \mapsto \mathbb{R}$ be a Borel function. Assume Φ is geometrically ergodic and one of the following conditions holds:*

1. *$E_\pi |g(x)|^{2+\delta} < \infty$ for some $\delta > 0$; or*
2. *$E_\pi g^2(x) < \infty$ and for all $x, y \in \mathcal{X}$, Φ satisfies detailed balance*

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx).$$

Then for any initial distribution, the following central limit theorem holds:

$$\sqrt{n}(\overline{g}_n - E_\pi g) \xrightarrow{d} N\left(0, \sigma_g^2\right) \qquad \text{as } n \to \infty$$

where $\sigma_g^2 = Var_\pi\left(g\left(X^{(0)}\right)\right) + 2\sum_{i=1}^{\infty} Cov_\pi\left(g\left(X^{(0)}\right), g\left(X^{(i)}\right)\right)$ and $\sigma_g^2 > 0$.

Remarks:

1. Condition 1 is due to Chan and Geyer (1994) and condition 2 is due to Roberts and Rosenthal (1997).
2. The positivity of σ_g^2 under the theorem assumptions is shown by Flegal and Jones (2008).

A Markov chain need not be geometrically ergodic for a central limit theorem to exist, but it is one of the easier conditions to check. For a nice review of Markov chain central limit theorems, see Jones (2004). In the next section, we discuss the need for and existence of consistent estimators of asymptotic variance σ_g^2.

3.2.2 Consistent Monte Carlo Standard Errors

Suppose a Markov chain CLT holds with corresponding asymptotic variance σ_g^2. If $\hat{\sigma}_g^2$ is consistent for σ_g^2, the user's confidence in $\overline{g}_n$ can be described by the MCSE $\hat{\sigma}_g/\sqrt{n}$. Further, an asymptotically valid $100(1-\alpha)\%$ confidence interval for $\mathrm{E}_\pi g$ is given by

$$\overline{g}_n \pm t_{\alpha/2,n-1}\frac{\hat{\sigma}_g}{\sqrt{n}} \,.$$

Therefore, a sufficient Markov chain sample size can be determined by increasing n until the interval half-width is below some prespecified value. Since Markov chain samples are not iid, the sample variance of the $g\left(X^{(i)}\right)$'s is not consistent for σ_g^2. However, consistent estimators exist under geometric ergodicity. We present two such methods in this chapter, regenerative simulation and consistent batch means.

Regenerative simulation

Resembling methods used in operations research (see, for instance, Glynn and Iglehart (1987)), regenerative simulation (RS) is a theoretically clean approach to obtaining consistent MCSEs. RS requires the simulation of a *split chain*. This is a generalization of the split chain presented in Chapter 3.1.2 and requires the following one-step minorization condition: for some function $s : \mathcal{X} \to [0,1]$ for which $\mathrm{E}_\pi s > 0$ and some probability measure $Q(\cdot)$ on $(\mathcal{X}, \mathcal{B})$,

$$P(x,A) \geq s(x)Q(A) \quad \text{for all } x \in \mathcal{X} \text{ and } A \in \mathcal{B} \,. \tag{3.7}$$

(See Chapter 5.2.3 for an example of constructing (3.7).) It follows that

$$P(x,A) = s(x)Q(A) + (1-s(x))R(x,A) \quad \text{for all } x \in \mathcal{X} \text{ and } A \in \mathcal{B}$$

where $R(x, A) = (1 - s(x))^{-1}[P(x, A) - s(x)Q(A)]$ is a probability measure on $(\mathcal{X}, \mathcal{B})$. Then the following algorithm can be used to simulate split chain $\left\{\left(X^{(n)}, \delta_n\right)\right\}$.

The General Split Chain Algorithm I

1. Select initial value $\left(X^{(0)}, \delta_0\right)$.

2. On the tth iteration, suppose $X^{(t-1)} = x$ and generate $X^{(t)}$ and δ_{t-1} as follows:

   - Draw $\delta_{t-1} \sim \text{Bern}(s(x))$.
   - If $\delta_{t-1} = 0$, draw $X^{(t)} \sim R(x, \cdot)$.
   - Otherwise, if $\delta_{t-1} = 1$, draw $X^{(t)} \sim Q(\cdot)$.

3. Repeat step 2.

Remarks:

1. Notice that when $s(x) = \varepsilon I(x \in C)$, minorization condition (2.4) is just a special case of (3.7). However, the ε values corresponding to (2.4) tend to be too small for the purposes of regenerative simulation (RS).

2. The marginal chain $\{X^{(n)}\}$ and the split chain are co-de-initializing in the terminology of Roberts and Rosenthal (2001). Hence if one is geometrically ergodic, they both are.

Simulating from $R(x, \cdot)$ is often difficult. For such cases, Mykland et al. (1995) present an alternative algorithm for which this is not required.

The General Split Chain Algorithm II

1. Select initial value $\left(X^{(0)}, \delta_0\right)$.

2. On the tth iteration, suppose $X^{(t-1)} = x$ and generate $X^{(t)}$ and δ_{t-1} as follows:

- Draw $X^{(t)} \sim P(x, \cdot)$.
- Draw a Bernoulli δ_{t-1} with the following probability of success:

$$\Pr\left(\delta_{t-1} = 1 \,\middle|\, X^{(t)}, X^{(t-1)}\right) = \frac{s\left(X^{(t-1)}\right) q\left(X^{(t)}\right)}{k\left(X^{(t-1)}, X^{(t)}\right)}$$

where q is the probability density corresponding to Q and k is the transition density corresponding to transition kernel P.

3. Repeat step 2.

When $\delta_{t-1} = 1$, $X^{(t)}$ conditional on $\left(X^{(t-1)}, \delta_{t-1}\right)$ has distribution $Q(\cdot)$. Thus t is a *regeneration time*, ie. a time at which the chain probabilistically restarts itself. Let τ_i denote the ith regeneration time where $0 = \tau_0 < \tau_1 < \ldots$ and

$$\tau_{t+1} = \min\left\{i > \tau_t : \delta_{i-1} = 1\right\}.$$

Then $\{\tau_0, \tau_1, \cdots\}$ partition the chain into a series of independent "tours" where, in general, frequent regenerations are reflective of a quicker convergence rate.

Suppose the simulation is stopped after the Rth time $\delta_i = 1$. Then τ_R is the total simulation length. Further, for $t = 1, \ldots, R$, let $N_t = \tau_t - \tau_{t-1}$ denote the length of the tth tour and define

$$S_t = \sum_{i=\tau_{t-1}}^{\tau_t - 1} g\left(X^{(i)}\right).$$

Since the tours between regeneration times are iid, the (N_t, S_t) pairs are also iid. Appealing to this fact, estimating $\mathrm{E}_\pi g$ is easy. First, let

$$\overline{N} = \frac{1}{R}\sum_{t=1}^{R} N_t \quad \text{and} \quad \overline{S} = \frac{1}{R}\sum_{t=1}^{R} S_t.$$

Then

$$\overline{g}_{\tau_R} = \frac{\overline{S}}{\overline{N}} = \frac{1}{\tau_R} \sum_{i=0}^{\tau_R - 1} g\left(X^{(i)}\right) \to \mathrm{E}_\pi g$$

with probability one as $R \to \infty$. Moreover, under moment conditions, a central limit theorem holds. First, assume $\tau_0 = 0$. This can be accomplished in one of two ways: (i) start from an arbitrary point and discard the simulation up to the time of the first regeneration; or (ii) draw $X^{(0)} \sim Q(\cdot)$, which can often be accomplished with an accept-reject algorithm. Next, assume $\mathrm{E}_Q N_1^2 < \infty$ and $\mathrm{E}_Q S_1^2 < \infty$ where E_Q denotes expectation conditional on $\tau_0 = 0$. Then as $R \to \infty$,

$$\sqrt{R}(\overline{g}_{\tau_R} - \mathrm{E}_\pi g) \xrightarrow{d} N\left(0, \nu_g^2\right) \tag{3.8}$$

where

$$\nu_g^2 = \frac{\mathrm{E}_Q(S_1 - N_1 \mathrm{E}_\pi g)^2}{(\mathrm{E}_Q N_1)^2} \, .$$

Notice that ν_g^2 will typically not equal σ_g^2 from Theorem 3.4. In fact,

$$\frac{\nu_g^2}{\mathrm{E}_\pi s} = \mathrm{Var}_\pi\left(g\left(X^{(0)}\right)\right) + 2\sum_{i=1}^{\infty} \mathrm{Cov}_\pi\left(g\left(X^{(0)}\right), g\left(X^{(i)}\right)\right) = \sigma_g^2$$

for s as defined by (3.7). A consistent estimator of ν_g^2 is given by

$$\hat{\nu}_g^2 = \frac{\sum_{t=1}^{R}(S_t - \overline{g}_{\tau_R} N_t)^2}{R\overline{N}^2} \, . \tag{3.9}$$

The existence of the CLT and the consistency of $\hat{\nu}_g^2$ follow from the assumption that $\mathrm{E}_Q N_1^2 < \infty$ and $\mathrm{E}_Q S_1^2 < \infty$. Unfortunately, these conditions are typically difficult to verify in practice. However, Hobert et al. (2002) show that (3.8) holds and $\hat{\nu}_g^2$ is consistent for ν_g^2 under certain other conditions including geometric ergodicity.

Theorem 3.5. *Hobert et al. (2002)*
Suppose Markov chain Φ *is geometrically ergodic. Also, assume* $E_\pi|g|^{2+\delta} < \infty$ *for some* $\delta > 0$ *and that a general minorization condition (3.7) exists. Then the CLT (3.8) holds and* $\hat{\nu}_g^2$ *given by* (3.9) *is strongly consistent for* ν_g^2*. Further, an asymptotically valid* $100(1-\alpha)\%$ *confidence interval for* $E_\pi g$ *is given by*

$$\bar{g}_{\tau_R} \pm t_{\alpha/2,R-1} \, \frac{\hat{\nu}_g}{\sqrt{R}} \, . \tag{3.10}$$

Remark 3.4. If (3.7) is such that regenerations occur often, then RS is an appealing scheme. It provides both a natural estimator of $\mathrm{E}_\pi g$ and a measure of its accuracy for geometrically ergodic Markov chains.

Batch means

Batch means methods provide an alternative to RS for constructing consistent Monte Carlo standard errors. These methods require more conditions to ensure consistency, yet are typically easier to apply in practice. In fact, the BUGS statistical software calculates standard errors using batch means (though it does so inconsistently).

Whereas RS partitions Markov chain output into independent tours, batch means methods partition output into *batches*. Let a be the number of batches, each of size b, for a Markov chain of length n. That is, $n = ab$. Also, for $k = 1, \ldots, a$ let

$$S_k = \frac{1}{b} \sum_{i=(k-1)b}^{kb-1} g\left(X^{(i)}\right)$$

be the average of the functional of the chain in batch k. The batch means estimate of σ_g^2 from Theorem 3.4 is

$$\hat{\sigma}_g^2 = \frac{b}{a-1} \sum_{k=1}^{a} (S_k - \overline{g}_n)^2 \, . \tag{3.11}$$

Consistency of $\hat{\sigma}_g^2$ requires a balance between the number of batches and the batch size. Jones et al. (2006) formulate these requirements for geometrically ergodic chains.

Theorem 3.6. *Jones et al. (2006)*
Suppose Markov chain Φ is geometrically ergodic. Also, assume $E_\pi|g|^{2+\delta+\varepsilon} < \infty$ for some $\delta > 0$ and $\varepsilon > 0$. Let a_n and b_n denote the number of batches and the batch size for a run of length n. If

1. *$a_n \to \infty$ as $n \to \infty$,*
2. *$b_n \to \infty$ and $b_n/n \to 0$ as $n \to \infty$,*
3. *$b_n^{-1} n^{2\alpha} [\log n]^3 \to 0$ as $n \to \infty$ where $\alpha = 1/(2+\delta)$, and*
4. *there exists a constant $c \geq 1$ such that $\sum_n (b_n/n)^c < \infty$,*

then the batch means estimate $\hat{\sigma}_g^2$ given by (3.11) is consistent for σ_g^2 from Theorem 3.4.

Remark 3.5. Code for implementing the consistent batch means procedure in R is provided by Galin Jones at http://www.stat.umn.edu/∼galin/cbm.R

Chapter 4

Geometric Ergodicity for the Gibbs Sampler

The rate of Markov chain convergence depends on target distribution π as well as on the MCMC algorithm of choice. Therefore, geometric ergodicity is difficult to establish in extremely general settings. However, explorations of convergence in more restricted contexts are abundant in the literature. Though many of these are model-specific, there do exist results for rather broad classes of MCMC algorithms. For instance, Roberts and Tweedie (1996) derive conditions under which the multidimensional Metropolis-Hastings algorithm converges at a geometric rate. Our main goal is to establish conditions for the geometric ergodicity of the Gibbs sampler introduced in the seminal articles by Geman and Geman (1984) and Gelfand and Smith (1990).

Geometric ergodicity for Gibbs samplers has been addressed in the literature by Diaconis et al. (2008a), Diaconis et al. (2008b), Geman and Geman (1984), Hobert and Geyer (1998), Jones and Hobert (2004), Liu et al. (1995), Papaspiliopoulos and Roberts (2007), and Roberts and Rosenthal (1998) (among others). In fact, both Geman and Geman (1984) and Liu et al. (1995) give sufficient conditions for geometric ergodicity in general settings. However, Geman and Geman (1984) only consider Gibbs sampling for finite state spaces and the conditions in Liu et al. (1995) are admittedly difficult to establish in practice (see Chapter 4.2). Furthermore, the other

existing convergence results do not apply to most chains used in realistic MCMC settings. In particular, Gibbs samplers for realistic models have received little attention despite their default use in some software packages.

We establish verifiable conditions under which the Gibbs sampler is geometrically ergodic. Specifically, we derive practical recipes for constructing drift and minorization conditions for the Gibbs sampler under a variety of *scanning strategies*. In addition to guaranteeing geometric ergodicity, these conditions provide us with a foundation on which we can build a rigorous MCMC analysis (see Chapter 3 for details). All proofs are deferred to Appendix A.

4.1 The Gibbs Sampler

Let $\pi(dx_1, dx_2, \cdots, dx_d)$ be a probability distribution having support $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d \subseteq \mathbb{R}^m$ where $\mathcal{X}_i \subseteq \mathbb{R}^{m_i}$ and $\sum_{i=1}^{d} m_i = m$. In a slight abuse of notation, suppose π admits a density $\pi(x_1, \cdots, x_d)$ with respect to some reference measure

$$\mu(dx_1, \cdots, dx_d) = \mu_1(dx_1) \cdots \mu_d(dx_d) \ .$$

For ease of exposition, we will typically assume μ_i is Lebesgue for all i. However, the results herein apply beyond the Lebesgue setting without restriction.

Let x_{-i} denote the vector x excluding the ith component x_i. Also, let $\pi(x_i|x_{-i})$ for $i = 1, \cdots, d$ denote the full conditional densities derived from $\pi(\cdot)$. Therefore, π is used for both the target distribution and distributions derived from it. To distinguish between the two, $\pi(\cdot)$ will always refer to the target distribution. On the other hand, the conditional aspect of the full conditional distributions will always be indicated.

It is often difficult to simulate from $\pi(\cdot)$ but still possible to simulate directly from all $\pi(x_i|x_{-i})$ which we assume throughout. In this case, $\pi(\cdot)$ can be explored using

a d-component Gibbs sampler. We will denote the corresponding Markov chain as $\Phi = \left\{X^{(0)}, X^{(1)}, \cdots\right\}$ where $X^{(t)} = \left(X_1^{(t)}, \cdots, X_d^{(t)}\right)$ denotes the value of the random d-component vector X after the tth iteration.

In general, a Gibbs sampler obtains component-wise updates by sampling from the full conditional distributions. However, several strategies exist for determining the order and frequency of the updates. We consider Gibbs sampling under three different scanning strategies: deterministic scan (DUGS), random permutation scan (RPGS), and random scan (RSGS). Though DUGS is the most widely used, considering RPGS and RSGS allows for more flexibility in our sampling strategy. We will also see that there are some theoretical advantages in employing the latter two strategies. For instance, RSGS is reversible as is RPGS under certain restrictions. Most importantly, this property weakens the conditions for the existence of a CLT (see Theorem 3.4).

4.1.1 DUGS

In every iteration of the DUGS, all components X_i of $X = (X_1, \cdots, X_d)$ are updated in some fixed and predetermined order. As is the case for all Gibbs samplers, each X_i is updated by drawing from the full conditional distribution of X_i given the current values of the other $d-1$ components. Without loss of generality, suppose the DUGS update order is $(1, 2, \cdots, d)$. That is, in each iteration, the d components are updated in consecutive order beginning with X_1 and ending with X_d. The corresponding DUGS algorithm can be summarized as follows:

The DUGS Algorithm

1. Select initial value $X^{(0)}$.

2. On the tth iteration, construct $X^{(t)} = \left(X_1^{(t)}, X_2^{(t)}, \cdots, X_d^{(t)}\right)$:

   Draw $X_1^{(t)} \sim \pi\left(X_1 \middle| X_{-1}^{(t-1)}\right)$.

Draw $X_2^{(t)} \sim \pi\left(X_2 \,\middle|\, X_1^{(t)},\ X_{-(1,2)}^{(t-1)}\right)$.

$\vdots$

Draw $X_d^{(t)} \sim \pi\left(X_d \,\middle|\, X_{-d}^{(t)}\right)$.

3. Repeat step 2.

The corresponding DUGS transition density k_D is the product of full conditional densities. Specifically, for $x = (x_1, \cdots, x_d)$ and $y = (y_1, \cdots, y_d)$,

$$k_D(x, y) = \pi\left(y_1 \,|\, x_{-1}\right) \pi\left(y_2 \,|\, y_1, x_{-(1,2)}\right) \cdots \pi\left(y_d \,|\, y_{-d}\right) \ .$$

It follows that π is invariant for the corresponding Markov chain.

DUGS is not reversible. However, reversible chains can be constructed from DUGS when $d = 2$. In this special case, let $\pi(x_1, x_2)$ denote the invariant density with corresponding marginals $f_1(x_1) := \int \pi(x_1, x_2) dx_2$ and $f_2(x_2) := \int \pi(x_1, x_2) dx_1$. Then the DUGS subchain $\Phi_i = \left\{X_i^{(0)}, X_i^{(1)}, \cdots\right\}$ is reversible with respect to f_i for both $i = 1, 2$. This fact is well-exploited in data augmentation methodology and follows from the structure of the sub-chain transition densities. For instance, let $k_1(x_1, y_1)$ denote the Φ_1 transition density. Also, define full conditional densities $\pi_1(x_1|x_2) = \pi(x_1, x_2)/f_2(x_2)$ and $\pi_2(x_2|x_1) = \pi(x_1, x_2)/f_1(x_1)$. Then $k_1(x_1, y_1) = \int \pi_2(y_2|x_1)\pi_1(y_1|y_2) dy_2$ and satisfies detailed balance with respect to f_1:

$$\begin{aligned}
f_1(x_1)k_1(x_1, y_1) &= f_1(x_1) \int \pi_2(y_2|x_1)\pi_1(y_1|y_2) dy_2 \\
&= f_1(x_1) \int \frac{\pi(x_1, y_2)}{f_1(x_1)} \frac{\pi(y_1, y_2)}{f_2(y_2)} dy_2 \\
&= f_1(y_1) \int \frac{\pi(x_1, y_2)}{f_2(y_2)} \frac{\pi(y_1, y_2)}{f_1(y_1)} dy_2 \\
&= f_1(y_1) \int \pi_1(x_1|y_2)\pi_2(y_2|y_1) dy_2 \\
&= f_1(y_1)k_1(y_1, x_1) \ .
\end{aligned}$$

Reversibility of Φ_1 follows.

4.1.2 RPGS

Like DUGS, RPGS updates all d components in every iteration. However, it does so in a randomly selected order. With d components, there are $d!$ possible update orders. We denote the ith update order as $(i(1), \cdots, i(d))$ where $i(j) \in \{1, \ldots, d\}$ is the index of the jth component to be updated. A set of permutation probabilities $q := \{q_1, q_2, \cdots, q_{d!}\}$ is assigned to the $d!$ possible update orders where

$$\sum_{i=1}^{d!} q_i = 1 \ .$$

Then in each iteration, the update order is randomly selected according to set q.

The RPGS Algorithm

1. Choose permutation probabilities $q = \{q_1, q_2, \cdots, q_{d!}\}$ and initial value $X^{(0)}$.
2. On the tth iteration, construct $X^{(t)} = \left(X_1^{(t)}, X_2^{(t)}, \cdots, X_d^{(t)}\right)$:
   - 2.1 Randomly draw an index $i \in \{1, 2, \cdots, d!\}$ with probability q_i.
   - 2.2 Generate $X^{(t)}$ according to the ith update order $(i(1), \cdots, i(d))$:

     Draw $X_{i(1)}^{(t)} \sim \pi\left(X_{i(1)} \middle| X_{-i(1)}^{(t-1)}\right)$.
     Draw $X_{i(2)}^{(t)} \sim \pi\left(X_{i(2)} \middle| X_{i(1)}^{(t)},\ X_{-(i(1),i(2))}^{(t-1)}\right)$.
     $\vdots$
     Draw $X_{i(d)}^{(t)} \sim \pi\left(X_{i(d)} \middle| X_{-i(d)}^{(t)}\right)$.
3. Repeat step 2.

Remark 4.1. Notice that if $q_i = 1$ for some i, RPGS is equivalent to DUGS.

The RPGS transition density k_P can be written as a mixture of the DUGS transition densities corresponding to the $d!$ possible update orders:

$$k_P(x,y) = \sum_{i=1}^{d!} q_i \pi\left(y_{i(1)} \,\middle|\, x_{-i(1)}\right) \pi\left(y_{i(2)} \,\middle|\, y_{i(1)}, x_{-(i(1),i(2))}\right) \cdots \pi\left(y_{i(d)} \,\middle|\, y_{-i(d)}\right) .$$

For any set of permutation probabilities q, it follows that π is invariant for the RPGS. Further, RPGS is reversible with respect to $\pi(\cdot)$ if $q_i = q_j$ for all pairs (i,j) such that $(i(1), i(2), \cdots, i(d)) = (j(d), j(d-1), \cdots, j(1))$:

$$\begin{aligned}
\pi(x)k_P(x,y) &= \sum_{i=1}^{d!} q_i \pi(x) \pi\left(y_{i(1)} \,\middle|\, x_{-i(1)}\right) \pi\left(y_{i(2)} \,\middle|\, y_{i(1)}, x_{-(i(1),i(2))}\right) \cdots \pi\left(y_{i(d)} \,\middle|\, y_{-i(d)}\right) \\
&= \sum_{i=1}^{d!} q_i \pi\left(x_{i(1)}, x_{-i(1)}\right) \frac{\pi\left(y_{i(1)}, x_{-i(1)}\right)}{\pi\left(x_{-i(1)}\right)} \frac{\pi\left(y_{i(1)}, y_{i(2)}, x_{-(i(1),i(2))}\right)}{\pi\left(y_{i(1)}, x_{-(i(1),i(2))}\right)} \cdots \frac{\pi(y)}{\pi\left(y_{-i(d)}\right)} \\
&= \sum_{i=1}^{d!} q_i \pi(y) \pi\left(x_{i(1)} \,\middle|\, x_{-i(1)}\right) \pi\left(x_{i(2)} \,\middle|\, y_{i(1)}, x_{-(i(1),i(2))}\right) \cdots \pi\left(x_{i(d)} \,\middle|\, y_{-i(d)}\right) \\
&= \sum_{j=1}^{d!} q_j \pi(y) \pi\left(x_{j(1)} \,\middle|\, y_{-j(1)}\right) \pi\left(x_{j(2)} \,\middle|\, x_{j(1)}, y_{-(j(1),j(2))}\right) \cdots \pi\left(x_{j(d)} \,\middle|\, x_{-j(d)}\right) \\
&= \pi(y) k_P(y,x) .
\end{aligned}$$

A special reversible case is the uniform RPGS in which $q_i = 1/d!$ for all i. On the other hand, reversibility does not hold for RPGS sub-chains Φ_i nor do Φ_i even satisfy the Markov property.

4.1.3 RSGS

RSGS differs from DUGS and RPGS in that in each iteration, a single component is updated while the other $d-1$ components remain fixed. Further, the order of the component-wise updates is randomly generated according to some assigned set of selection probabilities $p := \{p_1, p_2, \cdots, p_d\}$ where p_i is the probability of updating the

ith component. The set p must satisfy $p_i > 0$ and $\sum_{i=1}^{d} p_i = 1$ where $p_i > 0$ ensures that the RSGS will visit each component infinitely often.

The RSGS Algorithm

1. Choose selection probabilities $p = \{p_1, p_2, \cdots, p_d\}$ and initial value $X^{(0)}$.
2. On the tth iteration, construct $X^{(t)} = \left(X_1^{(t)}, X_2^{(t)}, \cdots, X_d^{(t)}\right)$:
   - 2.1 Randomly draw an index $i \in \{1, 2, \cdots, d\}$ with probability p_i.
   - 2.2 Generate $X_i^{(t)} \sim \pi\left(X_i \middle| X_{-i}^{(t-1)}\right)$ and set $X^{(t)} = \left(X_1^{(t-1)}, \cdots, X_{i-1}^{(t-1)}, X_i^{(t)}, X_{i+1}^{(t-1)}, \cdots, X_d^{(t-1)}\right)$.
3. Repeat step 2.

The RSGS transition density k_R can be written as

$$k_R(x, y) = \sum_{i=1}^{d} p_i \pi\left(y_i | x_{-i}\right) I\left(x_{-i} = y_{-i}\right)$$

where $I(\cdot)$ denotes the indicator function and $I\left(x_{-i} = y_{-i}\right) = 1$ for one and only one i. It follows that the RSGS satisfies detailed balance with respect to $\pi(\cdot)$:

$$\begin{aligned}
\pi(x) k_R(x, y) &= \sum_{i=1}^{d} p_i \pi(x) \pi\left(y_i \,|\, x_{-i}\right) I\left(x_{-i} = y_{-i}\right) \\
&= \sum_{i=1}^{d} p_i \frac{\pi\left(x_i, x_{-i}\right) \pi\left(y_i, x_{-i}\right)}{\pi\left(x_{-i}\right)} I\left(x_{-i} = y_{-i}\right) \\
&= \sum_{i=1}^{d} p_i \frac{\pi\left(x_i, y_{-i}\right) \pi\left(y_i, y_{-i}\right)}{\pi\left(y_{-i}\right)} I\left(x_{-i} = y_{-i}\right) \\
&= \pi(y) \sum_{i=1}^{d} p_i \pi\left(x_i \,|\, y_{-i}\right) I\left(x_{-i} = y_{-i}\right) \\
&= \pi(y) k_R(y, x) \; .
\end{aligned}$$

Namely, this guarantees that RSGS is reversible with invariant distribution $\pi(\cdot)$. However, as is true for RPGS, the RSGS sub-chains are neither reversible nor Markov.

The most common random scan strategy is to set $p_i = 1/d$ for all i (Amit and Grenander, 1991; Fishman, 1996; Roberts and Sahu, 1997). In this case, the distribution of visits is uniform over the d components. On the other hand, the non-uniform RSGS takes a less balanced tour through the state space. The obvious question that arises in this setting is, what choice of p is "optimal"? Consider the extreme case in which a large proportion of the p_i are close to zero. Under this assignment, the chain will often get stuck exploring a single plane in the state space for multiple iterations in a row. However, when p is chosen carefully, there are certain advantages to implementing a non-uniform RSGS.

A popular non-uniform RSGS strategy is to assign higher selection probabilities to 'more variable' components. Compared to uniform RSGS, visiting such components with higher frequency may hasten convergence and produce more precise Markov chain estimates (Levine et al., 2005; Levine and Casella, 2006; Liu et al., 1995). To this end, Levine et al. (2005) and Levine and Casella (2006) present adaptive random scan algorithms that update p with respect to some decision criterion as the Markov chain evolves. For instance, p can be chosen to minimize the mean squared error or Monte Carlo standard error of a Markov chain estimate. However, the corresponding chains are not Markov. Thus we eliminate these adaptive methods from our discussion.

4.2 Geometric Ergodicity of the Gibbs Sampler

In Lemma 4.1 we provide a simple, yet conservative, set of sufficient conditions for Harris ergodicity of the Gibbs samplers. First, we need the following definition.

Definition 4.1. A measure ν is *absolutely continuous* with respect to measure μ if for all sets $A \in \mathcal{B}$ such that $\mu(A) = 0$, it is also true that $\nu(A) = 0$.

Lemma 4.1. *Let P denote the one-step transition kernel of a d-component Gibbs sampler. Assume $P(x,\cdot)$ is absolutely continuous with respect to invariant distribution π. Also, for DUGS and RPGS, suppose $P(x,A) > 0$ for any $x \in \mathcal{X}$ and $A \in \mathcal{B}$ for which $\pi(A) > 0$. On the other hand, suppose the d-step RSGS transition kernel $P^d(x,A) > 0$ (where d is the number of components). Then the Gibbs sampler is Harris ergodic.*

(See Appendix A for a proof.)

Remark 4.2. The positivity conditions on the transition kernels often hold quite trivially. In addition, most π-irreducible Gibbs samplers satisfy the absolute continuity condition which requires that the chains *not* visit π-null sets. These conditions hold, for example, for the broad class of models in which the general state space $\mathcal{X}$ is the Cartesian product of the marginal state spaces corresponding to the d components.

Assume throughout that the Gibbs sampler is Harris ergodic. Then by Theorem 2.2, the corresponding Markov chain converges to π in total variation distance. Further, Liu et al. (1995) provide conditions under which DUGS and RSGS converge geometrically quickly to π in *Pearson χ^2-distance*. However, there are at least two drawbacks to this result. First, the Pearson χ^2-distance between the chain and the target distribution is defined by

$$d_x^2\left(P^n(x,\cdot),\pi(\cdot)\right) := \int \frac{\left[k^n(x,y)\right]^2}{\pi(y)} dy \ .$$

Notice that d_x^2 is not symmetric, ie. $d_x^2\left(P^n(x,\cdot),\pi(\cdot)\right)$ is typically not equal to $d_x^2\left(\pi(\cdot),P^n(x,\cdot)\right)$. Therefore, d_x^2 measures the discrepancy between $P^n(x,\cdot)$ and $\pi(\cdot)$ but is not a true measure of distance. Second, geometric convergence in Pearson χ^2-distance requires the following condition:

$$\int \left[\frac{k(x,y)}{\pi(y)}\right]^2 \pi(x)\pi(y)dxdy < \infty \ .$$

Unfortunately, this condition is "standard but not easy to check and understand" (Liu et al., 1995).

One of our goals is to derive *verifiable* conditions under which the Gibbs sampler is geometrically ergodic. To this end, we establish practical recipes for constructing drift and minorization conditions for the DUGS, RSGS, and RPGS under a general setting. We consider the 2-component and general d-component cases separately.

4.2.1 Geometric Ergodicity for 2-Component Gibbs Samplers

Consider 2-component Gibbs sampling for $\pi(x)$ where $x = (x_1, x_2)$. This requires sampling from the full conditional distributions of $X_1|X_2$ and $X_2|X_1$ where $(X_1, X_2) \sim \pi(\cdot)$. For simplicity, we will denote the corresponding full conditional densities as $\pi(x_1|x_2) := \pi_{X_1|X_2}(x_1|x_2)$ and $\pi(x_2|x_1) := \pi_{X_2|X_1}(x_2|x_1)$. This setting, though somewhat simple, has many practical applications. For instance, 2-component Gibbs sampling serves as the foundation of data augmentation methods.

Here we establish conditions for the geometric ergodicity of the DUGS, RPGS, and RSGS. A common condition for each sampler (related to the drift conditions) is the existence of functions f_1, f_2 that are positive π-a.e. (ie. π-almost everywhere) and constants a, b, c, and d for which $0 < ac < 1$ and

$$\begin{aligned} \mathrm{E}\left[f_1\left(y_1\right)|x_2\right] &\leq a f_2\left(x_2\right) + b \\ \mathrm{E}\left[f_2\left(y_2\right)|x_1\right] &\leq c f_1\left(x_1\right) + d \end{aligned}. \tag{4.1}$$

Another common requirement (related to the minorization conditions) is the fulfillment of one or both of the following conditions:

i. There exist some positive π-a.e. function g_1 and constant ν_2 such that

$$\inf_{x \in D_2} \pi\left(y_1|x_2\right) \geq g_1\left(y_1\right) \tag{4.2}$$

where $D_2 = \{x : f_2(x_2) \leq \nu_2\}$.

ii. There exist some positive π-a.e. function g_2 and constant ν_1 such that

$$\inf_{x \in D_1} \pi\left(y_2 | x_1\right) \geq g_2\left(y_2\right) \tag{4.3}$$

where $D_1 = \{x : f_1(x_1) \leq \nu_1\}$.

The remaining details of the conditions for geometric ergodicity are given separately for DUGS, RPGS, and RSGS in the following three theorems. Proofs are given in Appendix A. We recommend inspection of these proofs to gain intuition into the given drift and minorization conditions.

Theorem 4.1. *Let Φ denote a 2-component aperiodic and irreducible DUGS under update order (i, j). Further, let π denote the invariant distribution and P_D denote the DUGS transition kernel. Suppose there exist functions f_1 and f_2 for which* (4.1) *holds and define the following constants:*

$$\gamma_D = ac$$

$$L_D = \begin{cases} cb + d & \text{if } (i,j) = (1,2) \\ ad + b & \text{if } (i,j) = (2,1) \end{cases}.$$

Also, if $(i, j) = (1, 2)$ suppose (4.2) *holds for some $\{g_1, \nu_2 > 2L_D/(1-\gamma_D)\}$. Similarly, if $(i, j) = (2, 1)$ suppose* (4.3) *holds for some $\{g_2, \nu_1 > 2L_D/(1-\gamma_D)\}$. Then the following statements are true.*

1. *The DUGS is geometrically ergodic.*

2. *A DUGS drift condition is given by*

$$P_D V_D(x) \leq \gamma_D V_D(x) + L_D$$

where $V_D(x) = f_j(x_j)$.

3. *A DUGS minorization condition holds on set* $C_D = \{x : V_D(x) \leq \nu_j\} = D_j$*:*

$$P_D(x, A) \geq \varepsilon_D Q_D(A) \quad \text{for all } x \in C_D \text{ and } A \in \mathcal{B}$$

where

$$\varepsilon_D = \int g_i(y_i)\, \pi(y_j|y_i) dy$$

and $Q_D(\cdot)$ *is the probability measure corresponding to density*

$$q_D(y) = \varepsilon_D^{-1} g_i(y_i)\, \pi(y_j|y_i)\ .$$

Remark 4.3. Technically, the DUGS drift condition holds for drift rate $\gamma_D \in [ac, 1)$. However, a smaller drift rate is typically reflective of a quicker convergence rate. Therefore, we simplify the condition by setting $\gamma_D = ac$. We simplify the RPGS and RSGS drift conditions similarly in the next two theorems.

Theorem 4.2. *Let* Φ *denote a 2-component aperiodic and irreducible RPGS with invariant distribution* π *and transition kernel* P_P*. Also, let* q_1 *and* q_2 *be the permutation probabilities corresponding to update orders* $o_1 = (1, 2)$ *and* $o_2 = (2, 1)$*, respectively. Suppose there exist functions* f_1 *and* f_2 *for which* (4.1) *holds and define the following constants:*

$$u = \frac{(q_1 - q_2)ac + \sqrt{ac\,[ac + 4q_1q_2(1 - ac)]}}{2q_2c}$$

$$\gamma_P = q_2c(a + u)$$

$$L_P = q_1\,[b + u(cb + d)] + q_2\,[ud + (ad + b)]\ .$$

Also, suppose that either (4.2) *holds for some* $\{g_1, \nu_2 > (1/u)[2L_P/(1-\gamma_P)]\}$ *or* (4.3) *holds for some* $\{g_2, \nu_1 > 2L_P/(1 - \gamma_P)\}$*. If* (4.2) *does not hold, set* $g_1(x) = 0$ *for all*

$x \in \mathcal{X}$. Similarly, if (4.3) *does not hold, set* $g_2(x) = 0$. *Then if* $u > 0$, *the following statements are true.*

1. *The RPGS is geometrically ergodic.*

2. *An RPGS drift condition is given by*

$$P_P V_P(x) \leq \gamma_P V_P(x) + L_P$$

   where $V_P(x) = f_1(x_1) + u f_2(x_2)$.

3. *An RPGS minorization condition holds on set* $C_P = \{x : V_P(x) \leq \omega\}$ *for any* $2L_P/(1-\gamma_P) < \omega \leq \min\{\nu_1,\ u\nu_2\}$:

$$P_P(x, A) \geq \varepsilon_P Q_P(A) \quad \text{for all } x \in C_P \text{ and } A \in \mathcal{B}$$

   where

$$\varepsilon_P = \int \sum_{j=1}^{2} q_j g_j(y_j) \pi\left(y_{-j} | y_j\right) dy$$

   and $Q_P(\cdot)$ *is the probability measure corresponding to density*

$$q_P(y) = \varepsilon_P^{-1} \sum_{j=1}^{2} q_j g_j(y_j) \pi\left(y_{-j} | y_j\right) \ .$$

Remark 4.4. The requirement that $u > 0$ holds automatically if $q_1 \geq q_2$. This can be guaranteed through our choice of labels for the two components x_1 and x_2.

Theorem 4.3. *Let* Φ *denote a 2-component aperiodic and irreducible RSGS with invariant distribution* π *and transition kernel* P_R. *Also, let* p_1 *and* p_2 *denote the selection probabilities corresponding to components* x_1 *and* x_2, *respectively. Suppose*

there exist functions f_1 and f_2 for which (4.1) *holds and define the following constants:*

$$\begin{aligned} v &= \frac{(p_1 - p_2) + \sqrt{1 - 4p_1p_2(1 - ac)}}{2p_2c} \\ \gamma_R &= p_2(1 + vc) \\ L_R &= p_1b + vp_2d \ . \end{aligned}$$

Also, suppose (4.2) *holds for some* $\{g_1, \nu_2 > (1/v)[2L_R/(1-\gamma_R)]\}$ *and* (4.3) *holds for some* $\{g_2, \nu_1 > 2L_R/(1-\gamma_R)\}$. *Then if* $v > 0$, *the following statements are true.*

1. *The RSGS is geometrically ergodic.*

2. *An RSGS drift condition is given by*

$$P_R V_R(x) \leq \gamma_R V_R(x) + L_R$$

   where $V_R(x) = f_1(x_1) + vf_2(x_2)$.

3. *An RSGS minorization condition holds on set* $C_R = \{x : V_R(x) \leq \omega\}$ *for any* $2L_R/(1-\gamma_R) < \omega \leq \min\{\nu_1, v\nu_2\}$:

$$P_R(x, A) \geq \varepsilon_R Q_R(A) \quad \text{for all } x \in C_R \text{ and } A \in \mathcal{B}$$

   where

$$\varepsilon_R = \int \min_{j \in \{1,2\}} p_j g_j(y_j)\, dy$$

   and $Q_R(\cdot)$ *is the probability measure corresponding to density*

$$q_R(y) = \varepsilon_R^{-1} \min_{j \in \{1,2\}} p_j g_j(y_j) \ .$$

Remark 4.5. The requirement that $v > 0$ holds automatically if $p_1 \geq p_2$. This can be

guaranteed through our choice of labels for the two components x_1 and x_2.

Theorems 4.1, 4.2, and 4.3 provide sets of verifiable conditions for geometric ergodicity of 2-component DUGS, RPGS, and RSGS, respectively. In the following corollary, we combine these results in a way that will be useful for future reference.

Corollary 4.1. *Assume that functions f_1 and f_2 exist for which* (4.1) *holds and let $(\gamma_D, \gamma_P, \gamma_R)$ and (L_D, L_P, L_R) denote the constants defined in Theorems 4.1, 4.2, and 4.3. Also, assume $u > 0$ and $v > 0$ where u and v are defined by Theorems 4.2 and 4.3, respectively. Then if the Gibbs samplers are aperiodic and irreducible, the following statements are true.*

1. *The 2-component DUGS with update order (1,2) is geometrically ergodic if* (4.2) *holds for some $\{g_1, \nu_2 > 2L_D/(1-\gamma_D)\}$. Similarly, the 2-component DUGS with update order (2,1) is geometrically ergodic if* (4.3) *holds for some $\{g_2, \nu_1 > 2L_D/(1-\gamma_D)\}$.*

2. *The 2-component RPGS with permutation probabilities q_1 and q_2 is geometrically ergodic if either* (4.2) *holds for some $\{g_1, \nu_2 > (1/u)[2L_P/(1-\gamma_P)]\}$ or* (4.3) *holds for some $\{g_2, \nu_1 > 2L_P/(1-\gamma_P)\}$.*

3. *The 2-component RSGS with selection probabilities p_1 and p_2 is geometrically ergodic if* (4.2) *holds for some $\{g_1, \nu_2 > (1/v)[2L_R/(1-\gamma_R)]\}$ and* (4.3) *holds for some $\{g_2, \nu_1 > 2L_R/(1-\gamma_R)\}$.*

Remarks:

1. The existence of functions f_1 and f_2 that satisfy (4.1) is a common condition for the geometric ergodicity of DUGS, RPGS, and RSGS. These functions provide the building blocks of the drift conditions given by Theorems 4.1, 4.2, and 4.3. To this end, notice that the drift function for DUGS with update order (i, j)

depends only on $f_j(x_j)$ (hence a single component of x). This follows from the fact that the corresponding DUGS transition density only depends on x through x_j. On the other hand, the RPGS and RSGS transition densities depend on x through both x_1 and x_2. Hence, the RPGS and RSGS drift functions depend on both functions $f_1(x_1)$ and $f_2(x_2)$. See the drift condition proofs for clarification.

2. The conditions for geometric ergodicity are slightly less restrictive for DUGS and RPGS than for RSGS. Specifically, DUGS and RPGS only require that one of (4.2) or (4.3) hold whereas RSGS requires they both hold.

3. Notice that a single (yet conservative) set of conditions can guarantee geometric ergodicity for DUGS, RPGS, and RSGS. Specifically, geometric ergodicity holds for each sampler if there exist functions f_1 and f_2 for which (4.1) holds and if there exist functions g_1 and g_2 and constants

$$\nu_1 > \max\left\{\frac{2L_D}{1-\gamma_D}, \frac{2L_P}{1-\gamma_P}, \frac{2L_R}{1-\gamma_R}\right\}$$
$$\nu_2 > \max\left\{\frac{2L_D}{1-\gamma_D}, \frac{1}{u}\left(\frac{2L_P}{1-\gamma_P}\right), \frac{1}{v}\left(\frac{2L_R}{1-\gamma_R}\right)\right\}$$

for which (4.2) and (4.3) both hold.

The similarities among the conditions for geometric ergodicity suggest there is a systematic connection between the convergence behavior of Gibbs samplers under different scanning strategies. Let γ_D, γ_P, and γ_R denote the DUGS, RPGS, and RSGS drift rates defined by Theorems 4.1, 4.2, and 4.3, respectively. Though it is not as precise as comparing exact convergence rates (which are typically unavailable), comparing γ_D, γ_P, and γ_R provides insight into the convergence relationships. In fact, a more fair comparison might be between γ_D, γ_P, and γ_R^2. Recall that RSGS updates a single component in each iteration (whereas the DUGS and RPGS update both). Therefore, obtaining updates of both components requires at least two RSGS

iterations. Accordingly, we might compare the drift rate of the *two*-step RSGS drift condition to γ_D and γ_P. Under the assumptions of Theorem 4.3, this drift rate equals γ_R^2 and corresponds to the following two-step RSGS drift condition:

$$\begin{aligned}
\mathrm{E}_R\left[V_R\left(X^{(i+2)}\right) \,\middle|\, X^{(i)}=x\right] &= \mathrm{E}_R\left[\mathrm{E}_R\left\{V_R\left(X^{(i+2)}\right) \,\middle|\, X^{(i+1)}\right\} \,\middle|\, X^{(i)}=x\right] \\
&\leq \mathrm{E}_R\left[\gamma_R V_R\left(X^{(i+1)}\right) + L_R \,\middle|\, X^{(i)}=x\right] \\
&\leq \gamma_R^2 V_R(x) + L_R(1+\gamma_R)
\end{aligned}$$

where E_R denotes expectation with respect to the RSGS transition density.

Next, $\gamma_D = ac$ and by the definitions of u and v, γ_P and γ_R can be rewritten as

$$\gamma_P = \frac{1}{2}ac + \frac{1}{2}\sqrt{ac[ac + 4q_1q_2(1-ac)]} \quad \text{and} \quad \gamma_R = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4p_1p_2(1-ac)}\,. \tag{4.4}$$

It follows that

$$\begin{aligned}
ac < \gamma_P &\leq \frac{1}{2}\sqrt{ac}(1+\sqrt{ac}) \\
\frac{1}{2}\sqrt{ac}\left(1+\sqrt{ac}+\frac{1}{2\sqrt{ac}}\left[1-ac\right]\right) &\leq \gamma_R^2 < 1 \\
\frac{1}{2}(1+\sqrt{ac}) &\leq \gamma_R < 1\,.
\end{aligned} \tag{4.5}$$

Remarks:

1. The relationships in (4.5) guarantee $\gamma_D < \gamma_P < \gamma_R^2 < \gamma_R$ where, in general, smaller drift rates are indicative of faster convergence.

2. It is well known that the RSGS convergence rate depends on the choice of selection probabilities. This is illustrated by (4.4) through the dependence of drift rate γ_R on p_1 and p_2. This might guide our choice of the selection probabilities. For instance, we might choose the uniform RSGS ($p_1 = p_2 = 1/2$) as this minimizes γ_R at $\gamma_R = (1+\sqrt{ac})/2$. On the other hand, a uniform strategy

for the RPGS results in *slower* convergence. Specifically, it follows from (4.4) that drift rate γ_P increases as $q_1 \to 1/2$. Further, γ_P converges to its lower bound (DUGS drift rate $\gamma_D = ac$) as q_1 approaches 0 or 1. This indicates that the RPGS drift is quickest when one of the update orders is strongly favored over the other, that is, when the RPGS behaves similarly to the DUGS.

Given the previous results and discussions, there is clearly a connection between the DUGS, RPGS, and RSGS convergence behavior. In fact, uniform ergodicity of DUGS guarantees uniform ergodicity of *uniform* RSGS (Roberts and Rosenthal, 1997). We prove a stronger result that, under very mild conditions, geometric ergodicity of RPGS and RSGS is guaranteed by geometric ergodicity of DUGS. A proof is given in Appendix A.

Theorem 4.4.
Let Φ denote an aperiodic and irreducible DUGS with invariant distribution $\pi(x_1, x_2)$. Further, suppose the DUGS transition kernel is absolutely continuous with respect to $\pi(\cdot)$. Then if DUGS is geometrically ergodic, so are RPGS and RSGS.

Remark 4.6. This general result allows us to move away from the recipes for geometric ergodicity given in Theorems 4.1, 4.2, and 4.3. However, given the results of Chapter 3 and the above discussion of these theorems, constructing drift and minorization conditions still has both practical and theoretical value.

4.2.2 Extension to a 2-Component Mixture Setting

In this section, we extend our results to a more flexible model setting. Specifically, consider a 2-component target density $\pi_m(x_1, x_2)$ derived from a hierarchical model with $\pi_m(x_1|x_2)$ and $\pi_m(x_2)$. That is, $\pi_m(x_1, x_2) = \pi_m(x_1|x_2)\pi_m(x_2)$. Further, for

$s, t \in \{1, 2, \cdots\}$ and *known* $\phi_i, \psi_j \in (0, 1)$, suppose

$$\pi_m(x_1|x_2) = \sum_{i=1}^{s} \phi_i \pi_i(x_1|x_2) \quad \text{and} \quad \pi_m(x_2) = \sum_{j=1}^{t} \psi_j \pi_j(x_2) \tag{4.6}$$

where $\sum_{i=1}^{s} \phi_i = \sum_{j=1}^{t} \psi_j = 1$. In this case, $\pi_m(x_1, x_2)$ and full conditional density $\pi_m(x_2|x_1)$ can also be written as mixture densities. To this end, define $\pi_{ij}(x_1, x_2) = \pi_i(x_1|x_2)\pi_j(x_2)$, $\pi_{ij}(x_1) = \int \pi_{ij}(x_1, x_2)dx_2$, and $\pi_{ij}(x_2|x_1) = \pi_{ij}(x_1, x_2)/\pi_{ij}(x_1)$ for each i, j combination of the mixture components. Then

$$\begin{aligned}
\pi_m(x_1, x_2) &= \pi_m(x_1|x_2)\pi_m(x_2) \\
&= \sum_{i=1}^{s}\sum_{j=1}^{t} \phi_i\psi_j \pi_i(x_1|x_2)\pi_j(x_2) \\
&= \sum_{i=1}^{s}\sum_{j=1}^{t} \phi_i\psi_j \pi_{ij}(x_1, x_2) \ .
\end{aligned}$$

Also, define weights

$$w_{ij}(x_1) = \frac{\phi_i\psi_j\pi_{ij}(x_1)}{\sum_{k=1}^{s}\sum_{l=1}^{t} \phi_k\psi_l\pi_{kl}(x_1)}$$

and notice that $\sum_{i=1}^{s}\sum_{j=1}^{t} w_{ij}(x_1) = 1$ and $0 \leq w_{ij}(x_1) \leq 1$ for all i, j, x_1. Then

$$\begin{aligned}
\pi_m(x_2|x_1) &= \pi_m^{-1}(x_1)\pi_m(x_1, x_2) \\
&= \left[\sum_{k=1}^{s}\sum_{l=1}^{t} \phi_k\psi_l\pi_{kl}(x_1)\right]^{-1} \left[\sum_{i=1}^{s}\sum_{j=1}^{t} \phi_i\psi_j\pi_{ij}(x_1, x_2)\right] \\
&= \sum_{i=1}^{s}\sum_{j=1}^{t} w_{ij}(x_1)\pi_{ij}(x_2|x_1) \ .
\end{aligned}$$

Consider Gibbs sampling for $\pi_m(x_1, x_2)$. In Proposition 4.1 we establish building block functions f_1 and f_2 that satisfy (4.1). In turn, drift conditions for the DUGS, RPGS, and RSGS can be constructed using the recipes given in Theorems 4.1, 4.2,

and 4.3. See Appendix A for a proof.

Proposition 4.1. *Let $\pi_m(x_1, x_2)$ be the joint target density defined by mixture densities $\pi_m(x_1|x_2)$ and $\pi_m(x_2)$ in (4.6). Also, let E_i denote expectation with respect to $\pi_i(x_1|x_2)$ and E_{ij} denote expectation with respect to $\pi_{ij}(x_2|x_1)$. Assume there exist positive π_m-a.e. functions f_1 and f_2 such that for any i, j combination there exist constants a_i, b_i, c_{ij}, d_{ij} for which $0 < a_i c_{ij} < 1$ and*

$$\begin{aligned} E_i[f_1(x_1)|x_2] &\leq a_i f_2(x_2) + b_i \\ E_{ij}[f_2(x_2)|x_1] &\leq c_{ij} f_1(x_1) + d_{ij} \end{aligned} . \tag{4.7}$$

Then with expectation taken with respect to the mixture distribution $\pi_m(x_1, x_2)$, f_1 and f_2 satisfy (4.1) with $a = \max_i a_i$, $b = \sum_{i=1}^{s} \phi_i b_i$, $c = \max_{ij} c_{ij}$, and $d = \max_{ij} d_{ij}$ where the assumptions on a_i and c_{ij} guarantee $ac < 1$.

Remark 4.7. Proposition 4.1 requires that functions f_1 and f_2 are suitable for each mixture combination i, j. This is a restriction of our recipes but, in general, is not required for the existence of a drift condition.

Next, in Proposition 4.2 we establish functions g_1 and g_2 which satisfy (4.2) and (4.3). In conjunction with Theorems 4.1, 4.2 and 4.3, these functions can be used to construct DUGS, RPGS, and RSGS minorization conditions that are associated with the drift conditions derived from (4.7). A proof is given in Appendix A.

Proposition 4.2. *Suppose the assumptions of Proposition 4.1 for mixture distribution $\pi_m(x_1, x_2)$ are met. Further, suppose there exist positive functions g_i and g_{ij} on $\mathcal{X}$ for $i \in \{1, \cdots, s\}$ and $j \in \{1, \cdots, t\}$ such that*

$$\inf_{x \in D_2} \pi_i(y_1|x_2) \geq g_i(y_1) \quad \text{and} \quad \inf_{x \in D_1} \pi_{ij}(y_2|x_1) \geq g_{ij}(y_2)$$

where $D_i = \{x : f_i(x_i) \leq \nu_i\}$ for f_i from Proposition 4.1 and ν_i that satisfy the

conditions of Corollary 4.1. Then

$$g_1(y_1) = \sum_{i=1}^{s} \phi_i g_i(y_1) \quad \textit{and} \quad g_2(y_2) = \min_{i,j} g_{ij}(y_2)$$

satisfy (4.2) *and* (4.3) *on* D_2 *and* D_1. *Specifically,*

$$\inf_{x \in D_2} \pi_m(y_1|x_2) \geq g_1(y_1) \quad \textit{and} \quad \inf_{x \in D_1} \pi_m(y_2|x_1) \geq g_2(y_2) \ .$$

Remark 4.8. Under the assumptions of Propositions 4.1 and 4.2, Corollary 4.1 guarantees geometric ergodicity for DUGS, RPGS, and RSGS for the mixture model $\pi_m(x_1, x_2)$ *as well as* geometric ergodicity for DUGS, RPGS, and RSGS for each of the mixture components $\pi_{ij}(x_1, x_2)$.

4.2.3 Geometric Ergodicity for d-Component Gibbs Samplers

An obvious goal is to extend our results in the 2-component setting to d-component Gibbs sampling where $d \geq 2$. To provide insight into establishing geometric ergodicity when $d > 2$, we begin with a toy example.

Example 4.1.

Let x_1 and x_2 be iid $N(x_3, \sigma^2)$ where x_3 has a $N(0, \sigma^2)$ prior. In this case, the joint distribution is trivariate normal:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \ \sigma^2 \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} \right) \ .$$

Consider three-component Gibbs sampling for this joint distribution. The full condi-

tional distributions required for the Gibbs updates are as follows:

$$\begin{aligned} x_1|x_2,x_3 &\sim N\left(x_3,\sigma^2\right), \\ x_2|x_1,x_3 &\sim N\left(x_3,\sigma^2\right), \quad \text{and} \\ x_3|x_1,x_2 &\sim N\left(\frac{x_1+x_2}{3},\ \frac{\sigma^2}{3}\right). \end{aligned} \tag{4.8}$$

For simplicity, we focus on the following Gibbs samplers:

1. DUGS with update order (1,2,3).
2. RPGS with probabilities $\{1/3, 1/3, 1/3\}$ assigned to update orders $\{(1,2,3),(2,3,1),(3,1,2)\}$.
3. Uniform RSGS.

Remark 4.9. Due to the conditional independence of x_1 and x_2 given x_3, the DUGS with update order (1,2,3) is effectively a 2-component Gibbs sampler with components (x_1, x_2) and x_3. However, we include this sampler for completeness.

Let $\pi(\mu, \tau^2; z)$ denote the $N(\mu, \tau^2)$ density evaluated at some point z. The DUGS, RPGS, and RSGS transition densities k_D, k_P, and k_R, respectively, follow from (4.8):

$$\begin{aligned} k_D(x,y) &= \pi\left(x_3,\sigma^2;y_1\right)\pi\left(x_3,\sigma^2;y_2\right)\pi\left(\frac{y_1+y_2}{3},\frac{\sigma^2}{3};y_3\right) \\ k_P(x,y) &= \frac{1}{3}\pi\left(x_3,\sigma^2;y_1\right)\pi\left(x_3,\sigma^2;y_2\right)\pi\left(\frac{y_1+y_2}{3},\frac{\sigma^2}{3};y_3\right) \\ &\quad + \frac{1}{3}\pi\left(x_3,\sigma^2;y_2\right)\pi\left(\frac{x_1+y_2}{3},\frac{\sigma^2}{3};y_3\right)\pi\left(y_3,\sigma^2;y_1\right) \\ &\quad + \frac{1}{3}\pi\left(\frac{x_1+x_2}{3},\frac{\sigma^2}{3};y_3\right)\pi\left(y_3,\sigma^2;y_1\right)\pi\left(y_3,\sigma^2;y_2\right) \\ k_R(x,y) &= \frac{1}{3}\pi\left(x_3,\sigma^2;y_1\right)I\left(x_{-1}=y_{-1}\right) + \frac{1}{3}\pi\left(x_3,\sigma^2;y_2\right)I\left(x_{-2}=y_{-2}\right) \\ &\quad + \frac{1}{3}\pi\left(\frac{x_1+x_2}{3},\frac{\sigma^2}{3};y_3\right)I\left(x_{-3}=y_{-3}\right). \end{aligned} \tag{4.9}$$

Generalizing (4.1) - (4.2) provides a common set of conditions for geometric ergodicity of the 3-component DUGS, RPGS, and RSGS. We begin by establishing these conditions and utilize them below. First, a generalization of (4.1) provides the basis for the 3-component Gibbs sampler drift conditions. Define functions $f_i(x_i) = x_i^2$ for $i \in \{1, 2, 3\}$ and notice that

$$\mathrm{E}\left[f_1(x_1) \,\middle|\, x_2, x_3\right] = \mathrm{E}\left[f_2(x_2) \,\middle|\, x_1, x_3\right] = f_3(x_3) + \sigma^2$$

and, by Jensen's inequality

$$\begin{aligned}\mathrm{E}\left[f_3(x_3) \,\middle|\, x_1, x_2\right] = \left(\frac{x_1 + x_2}{3}\right)^2 + \frac{\sigma^2}{3} &= \frac{4}{9}\left(\frac{x_1 + x_2}{2}\right)^2 + \frac{\sigma^2}{3} \\ &\leq \frac{4}{9}\left(\frac{x_1^2 + x_2^2}{2}\right) + \frac{\sigma^2}{3} \\ &= \frac{2}{9}\left(f_1(x_1) + f_2(x_2)\right) + \frac{\sigma^2}{3}\,.\end{aligned}$$

Next, a generalization of (4.2) and (4.3) provides the foundation for minorization conditions. To this end, it suffices to find functions g_1, g_2, and g_3 such that for any $y \in \mathbb{R}^3$ and $\nu_1, \nu_2 > 0$

$$\begin{aligned}\inf_{x \in D_2} \pi\left(x_3, \sigma^2; y_i\right) &\geq g_1(y_i) \text{ for } i = 1, 2, \\ \inf_{x \in D_1} \pi\left(\frac{x_1 + x_2}{3}, \frac{\sigma^2}{3}; y_3\right) &\geq g_2(y_3), \text{ and} \\ \inf_{x \in D_1} \pi\left(\frac{x_1 + y_2}{3}, \frac{\sigma^2}{3}; y_3\right) &\geq g_3(y_2, y_3)\end{aligned} \tag{4.10}$$

where

$$\begin{aligned}D_1 &= \{x : f_1(x_1) + f_2(x_2) \leq \nu_1\} = \{x : x_1^2 + x_2^2 \leq \nu_1\}; \text{ and} \\ D_2 &= \{x : f_3(x_3) \leq \nu_2\} = \{x : x_3^2 \leq \nu_2\}\,.\end{aligned} \tag{4.11}$$

This can be achieved by making a general observation about normal densities. Suppose $Z \sim N(\mu, \tau^2)$ where $\mu^2 \leq c$ for some $c > 0$. Then for all $z \in \mathbb{R}$

$$\pi\left(\mu, \tau^2; z\right) \geq g\left(c, \tau^2; z\right)$$

where

$$g\left(c, \tau^2; z\right) = \sqrt{\frac{1}{2\pi\tau^2}} \exp\left\{-\frac{1}{2\tau^2}\left(z + \sqrt{c}\left[I(z \geq 0) - I(z < 0)\right]\right)^2\right\} .$$

Notice that $x_3^2 \leq \nu_2$ for $x \in D_2$ and by Jensen's inequality

$$\begin{aligned}\left(\frac{x_1 + x_2}{3}\right)^2 &\leq \frac{2}{9}\left(x_1^2 + x_2^2\right) \leq \frac{2}{9}\nu_1 \\ \left(\frac{x_1 + y_2}{3}\right)^2 &\leq \frac{2}{9}\left(x_1^2 + y_2^2\right) \leq \frac{2}{9}\left(\nu_1 + y_2^2\right)\end{aligned}$$

for $x \in D_1$. Therefore, the following functions satisfy (4.10) for any $\nu_1, \nu_2 > 0$:

$$\begin{aligned}g_1(y_i) &= g\left(\nu_2, \sigma^2;\ y_i\right), \\ g_2(y_3) &= g\left(\frac{2}{9}\nu_1, \frac{\sigma^2}{3};\ y_3\right), \text{ and} \\ g_3(y_2, y_3) &= g\left(\frac{2}{9}\left(\nu_1 + y_2^2\right), \frac{\sigma^2}{3};\ y_3\right) .\end{aligned} \tag{4.12}$$

We are now ready to establish drift and minorization conditions (and therefore geometric ergodicity) for the 3-component Gibbs samplers. We consider DUGS, RPGS, and RSGS separately. Since the treatments are similar, details are eliminated for RPGS and RSGS.

1. Drift and minorization for DUGS

   In constructing a DUGS drift condition, notice that transition density k_D in (4.9) only depends on the current state x through the third component x_3.

Therefore, we select a drift function that also only depends on x_3. To this end, define $V_D(x) = f_3(x_3) = x_3^2$. Also, let P_D denote the DUGS transition kernel and E_D denote expectation with respect to P_D. Then by the construction of DUGS, for any $x \in \mathbb{R}^3$ we have

$$\begin{aligned} P_D V_D(x) = \mathrm{E}_D\left[f_3(y_3) \,|\, x\right] &= \mathrm{E}\left[\mathrm{E}\left(\mathrm{E}\left[f_3(y_3) \,|\, y_1, y_2\right] \,|\, y_1, x_3\right) \,|\, x_2, x_3\right] \\ &\leq \mathrm{E}\left[\mathrm{E}\left(\frac{4}{9}\left(\frac{f_1(y_1) + f_2(y_2)}{2}\right) + \frac{\sigma^2}{3} \,\middle|\, y_1, x_3\right) \,\middle|\, x_2, x_3\right] \\ &= \mathrm{E}\left[\frac{2}{9}\left(f_1(y_1) + f_3(x_3) + \sigma^2\right) + \frac{\sigma^2}{3} \,\middle|\, x_2, x_3\right] \\ &= \frac{4}{9} f_3(x_3) + \frac{7\sigma^2}{9}\,. \end{aligned}$$

Therefore, a drift condition is satisfied for $\gamma_D = 4/9$ and $L_D = 7\sigma^2/9$:

$$P_D V_D(x) \leq \gamma_D V_D(x) + L_D \quad \text{for all } x \in \mathbb{R}^3\,.$$

Geometric ergodicity will follow from establishing a minorization condition on set $D_2 = \{x : V_D(x) \leq \nu_2\} = \{x : f_3(x_3) \leq \nu_2\}$ from (4.11) where $\nu_2 > 0$. To this end, for any $x \in D_2$ we have

$$\begin{aligned} k_D(x, y) &= \pi\left(x_3, \sigma^2; y_1\right) \pi\left(x_3, \sigma^2; y_2\right) \pi\left(\frac{y_1 + y_2}{3}, \frac{\sigma^2}{3}; y_3\right) \\ &\geq \inf_{x \in D_2}\left[\pi\left(x_3, \sigma^2; y_1\right) \pi\left(x_3, \sigma^2; y_2\right) \pi\left(\frac{y_1 + y_2}{3}, \frac{\sigma^2}{3}; y_3\right)\right] \\ &\geq \inf_{x \in D_2} \pi\left(x_3, \sigma^2; y_1\right) \inf_{x \in D_2} \pi\left(x_3, \sigma^2; y_2\right) \pi\left(\frac{y_1 + y_2}{3}, \frac{\sigma^2}{3}; y_3\right) \\ &\geq g_1\left(y_1\right) g_1\left(y_2\right) \pi\left(\frac{y_1 + y_2}{3}, \frac{\sigma^2}{3}; y_3\right) \\ &:= h_D(y)\,. \end{aligned}$$

Therefore, (2.4) holds for the DUGS with $m = 1$, $\varepsilon = \int h_D(y)dy$, and Q corre-

sponding to probability density $q(y) = \varepsilon^{-1} h_D(y)$.

2. Drift and minorization for uniform RPGS

   Let P_P denote the RPGS transition kernel and E_P denote expectation with respect to P_P. First, the following drift condition holds for drift function $V_P(x) = f_1(x_1) + f_2(x_2) + 3f_3(x_3)$, drift rate $\gamma_P = 2/3$, and $L_P = \sigma^2(110/27)$:

$$P_P V_P(x) \leq \gamma_P V_P(x) + L_P \quad \text{for all } x \in \mathbb{R}^3 .$$

   Next, define

$$\begin{aligned} h_P(y) = \frac{1}{3} g_1(y_1) g_1(y_2) \pi\left(\frac{y_1 + y_2}{3}, \frac{\sigma^2}{3}; y_3\right) + \frac{1}{3} g_1(y_2) g_3(y_2, y_3) \pi\left(y_3, \sigma^2; y_1\right) \\ + \frac{1}{3} g_2(y_3) \pi\left(y_3, \sigma^2; y_1\right) \pi\left(y_3, \sigma^2; y_2\right) \end{aligned}$$

   for g_i defined by (4.12). Then the following minorization condition holds on $C_P = \{x : V_P(x) \leq \omega\}$ for any $\omega > 0$:

$$P_P(x, A) \geq \varepsilon_P Q_P(A) \quad \text{for all } x \in C_P \text{ and } A \in \mathcal{B}$$

   where $\varepsilon_P = \int h_P(y) dy$ and Q_P is the probability measure corresponding to density $q_P(y) = \varepsilon_P^{-1} h_P(y)$.

3. Drift and minorization for uniform RSGS

   Let P_R denote the RSGS transition kernel and E_R denote expectation with respect to P_R. Also, define drift function $V_R(x) = f_1(x_1) + f_2(x_2) + 3f_3(x_3)$ (this is the same as the RPGS drift function). Then an RSGS drift condition holds with $\gamma_R = 8/9$ and $L_R = \sigma^2$:

$$P_R V_R(x) \leq \gamma_R V_R(x) + L_R \quad \text{for all } x \in \mathbb{R}^3 .$$

Further, a minorization condition holds on $C_R = \{x : V_R(x) \leq \omega\}$ for any $\omega > 0$:

$$P_R(x, A) \geq \varepsilon_R Q_R(A) \quad \text{for all } x \in C_R \text{ and } A \in \mathcal{B}$$

where

$$\varepsilon_R = \int \min\left\{g_1\left(y_1\right), g_1\left(y_2\right), g_2\left(y_3\right)\right\} dy$$

for g_i defined by (4.12) and probability measure Q_R corresponds to density $q_R(y) = \varepsilon_R^{-1} \min\left\{g_1\left(y_1\right), g_1\left(y_2\right), g_2\left(y_3\right)\right\}$.

Remark 4.10. The DUGS, RPGS, and RSGS drift rates for the Normal-Normal model are $\gamma_D = 4/9$, $\gamma_P = 2/3$, and $\gamma_R = 8/9$, respectively. As with the 2-component drift rates, these satisfy $\gamma_D < \gamma_P < \gamma_R^3 < \gamma_R$. This suggests that the 3-component DUGS converges to stationarity the quickest and RSGS converges the slowest.

The strategies used in this example can be extended to establish geometric ergodicity for general d-component Gibbs samplers. To this end, we focus on constructing drift conditions for d-component DUGS, RPGS, and RSGS. Recall that in the 2-component setting, drift conditions were constructed from functions f_1, f_2 for which

$$\begin{aligned} \mathrm{E}\left[f_1\left(x_1\right)|x_2\right] &\leq a f_2\left(x_2\right) + b \\ \mathrm{E}\left[f_2\left(x_2\right)|x_1\right] &\leq c f_1\left(x_1\right) + d \end{aligned} \tag{4.13}$$

where $0 < ac < 1$. A natural extension provides a set of functions from which to build drift conditions for the d-component Gibbs samplers.

Assumption $\mathcal{A}$: For all $i \in \{1, \cdots, d\}$ there exist positive π-a.e. functions $f_i\left(x_i\right)$ and constants $0 \leq \alpha_{ij} < 1$ and $\beta_{ij} < \infty$ for which

$$\mathrm{E}\left[f_i\left(x_i\right)|x_{-i}\right] \leq \sum_{k \in \{1,\ldots,d\}, k \neq i} \left[\alpha_{ik} f_k\left(x_k\right) + \beta_{ik}\right] . \tag{4.14}$$

Remark 4.11. Notice that (4.13) satisfies Assumption $\mathcal{A}$ with $\alpha_{12} = a$, $\alpha_{21} = c$, $\beta_{12} = b$, and $\beta_{21} = d$.

Assumption $\mathcal{A}$ was satisfied in Example 4.1 by $f_i(x_i) = x_i^2$ for $i \in \{1, 2, 3\}$. However, establishing this assumption might be difficult for some d-component Gibbs samplers. For instance, one restriction is that each f_i is only allowed to depend on a single component x_i. Fortunately, Assumption $\mathcal{A}$ can be relaxed. First, define index set $I = \{1, \cdots, d\}$. Also, for any subset of indices $D \subset I$, let x_D denote the vector consisting of the components x_i for which $i \in D$. For instance, suppose $d = 5$ and $I = \{1, \cdots, 5\}$. Then if $D = \{3, 4\}$, $x_D = (x_3, x_4)$. Our drift condition recipes require the following assumption.

> *Assumption $\mathcal{B}$:* There exists some set $J = \{J_1, J_2, \cdots, J_m\}$ where each J_i is a set of indices ($J_i \subseteq I$) and for all $J_i \in J$ there exists some positive π-a.e. function $f_i(x_{J_i})$ such that for any $j \in I$
>
> $$\mathrm{E}\left[f_i\left(x_{J_i}\right)|x_{-j}\right] \leq \sum_{k=1}^{m} \left[\alpha_{ijk} f_k\left(x_{J_k}\right) + \beta_{ijk}\right] \tag{4.15}$$
>
> for some constants α_{ijk} and β_{ijk}.

Remarks:

1. Assumption $\mathcal{A}$ is a special case of Assumption $\mathcal{B}$ with $J = \{J_1, \ldots, J_d\}$ and $J_i = \{i\}$ for all i.

2. Each f_i in (4.15) can depend on any subset of the d components. However, there are some key relationships to keep in mind when searching for these functions. First, notice that when $j \notin J_i$, the function $f_i(x_{J_i})$ does not depend on x_j. Therefore, $\mathrm{E}\left[f_i\left(x_{J_i}\right)|x_{-j}\right] = f_i\left(x_{J_i}\right)$ and the inequality in (4.15) can be replaced by an equality where $\alpha_{iji} = 1$, $\beta_{iji} = 0$, and the remaining α, β terms are zero.

In a similar spirit, when $j \in J_i$, $f_i(x_{J_i})$ depends on x_j and the sum on the right-hand side of (4.15) can be taken over k for which $j \notin J_k$. In other words, if $j \in J_i$, it must be true that $\alpha_{ijk} = \beta_{ijk} = 0$ for any k such that $j \in J_k$.

In Proposition 4.3, we provide drift condition recipes for the d-component DUGS, RPGS, and RSGS under Assumption $\mathcal{B}$. A proof is given in Appendix A.

Proposition 4.3. *Let Φ denote a d-component Gibbs sampler with invariant distribution π. Suppose Assumption $\mathcal{B}$ holds and define $\beta_{ij\cdot} = \sum_{k=1}^{m} \beta_{ijk}$.*

1. *Drift condition for DUGS*

   Without loss of generality, suppose the update order for the DUGS is $(1, 2, \cdots, d)$. Also, assume there exist constants w_k for $k \in \{1, \cdots, m\}$ for which

$$\gamma_{Dk} := \frac{1}{w_k} \sum_{i_1, \cdots, i_d = 1}^{m} w_{i_1} \alpha_{i_1 d i_2} \alpha_{i_2 (d-1) i_3} \cdots \alpha_{i_{d-1} 2 i_d} \alpha_{i_d 1 k} < 1 \qquad \text{when } w_k \neq 0$$

   and $\gamma_{Dk} := 0$ when $w_k = 0$. Then a drift condition holds for drift function $V_D(x) = \sum_{i=1}^{m} w_i f_i(x_{J_i})$, drift rate $\gamma_D = \max_k \gamma_{Dk}$, and constant $L_D = \sum_{i_1=1}^{m} w_{i_1} L_{D i_1}$ where

$$L_{D i_1} = \beta_{i_1 d \cdot} + \sum_{i_2=1}^{m} \alpha_{i_1 d i_2} \beta_{i_2 (d-1) \cdot} + \cdots + \sum_{i_2, \cdots, i_d = 1}^{m} \alpha_{i_1 d i_2} \alpha_{i_2 (d-1) i_3} \cdots \alpha_{i_{d-1} 2 i_d} \beta_{i_d 1 \cdot} \, .$$

2. *Drift condition for RPGS*

   Let $q = \{q_1, q_2, \cdots, q_{d!}\}$ denote the RPGS permutation probabilities where q_j corresponds to $(j(1), j(2), \cdots, j(d))$, the jth permutation of the update order. Assume there exist constants u_k for $k \in \{1, \cdots, m\}$ for which

$$\gamma_{Pk} := \frac{1}{u_k} \sum_{j=1}^{d!} \sum_{i_1, \cdots, i_d = 1}^{m} q_j u_{i_1} \alpha_{i_1 j(d) i_2} \alpha_{i_2 j(d-1) i_3} \cdots \alpha_{i_{d-1} j(2) i_d} \alpha_{i_d j(1) k} < 1$$

when $u_k \neq 0$ *and* $\gamma_{Pk} := 0$ *when* $u_k = 0$. *Then a drift condition holds for drift function* $V_P(x) = \sum_{i=1}^{m} u_i f_i(x_{J_i})$, *drift rate* $\gamma_P = \max_k \gamma_{Pk}$, *and* $L_P = \sum_{j=1}^{d!} \sum_{i_1=1}^{m} q_j u_{i_1} L_{Pji_1}$ *where*

$$L_{Pji_1} = \beta_{i_1 j(d)\cdot} + \sum_{i_2=1}^{m} \alpha_{i_1 j(d) i_2} \beta_{i_2 j(d-1)\cdot} + \cdots + \sum_{i_2,\cdots,i_d=1}^{m} \alpha_{i_1 j(d) i_2} \alpha_{i_2 j(d-1) i_3} \cdots \alpha_{i_{d-1} j(2) i_d} \beta_{i_d j(1)\cdot} \ .$$

3. *Drift condition for RSGS*

   Let $p = \{p_1, \cdots, p_d\}$ *denote the RSGS selection probabilities and assume there exist constants* v_k *for* $k \in \{1, \cdots, m\}$ *for which*

$$\gamma_{Rk} := \frac{1}{v_k} \sum_{i=1}^{m} \sum_{j=1}^{d} v_i p_j \alpha_{ijk} < 1 \ .$$

   Then a drift condition holds for drift function $V_R(x) = \sum_{i=1}^{m} v_i f_i(x_{J_i})$, *drift rate* $\gamma_R = \max_k \gamma_{Rk}$, *and* $L_R = \sum_{i=1}^{m} \sum_{j=1}^{d} v_i p_j \beta_{ij\cdot}$.

Remarks:

1. Drift condition recipes can be somewhat simplified under Assumption $\mathcal{A}$ and follow directly from Proposition 4.3.

2. The drift conditions for the 2-component Gibbs samplers given by Theorems 4.1, 4.2, and 4.3 follow from Proposition 4.3. First, in the notation of Assumption $\mathcal{B}$, let $J = \{J_1, J_2\} := \{\{1\}, \{2\}\}$. Then (4.1) satisfies (4.15) by defining $\alpha_{112} = a$, $\alpha_{221} = c$, $\alpha_{121} = \alpha_{212} = 1$, $\beta_{112} = b$, $\beta_{221} = d$, and setting the remaining α and β terms to 0.

3. As in the 2-component case, a set of functions f_i serve as the building blocks of the drift conditions. However, in the d-component setting we cannot explicitly formulate coefficients w_i, u_i, and v_i for the general DUGS, RPGS, and RSGS

recipes, respectively. Therefore, directly comparing the d-component drift conditions is more meaningful in the context of a specific application.

4. The RSGS drift rate γ_R depends on selection probabilities, p_i. So long as $\inf \gamma_R > 0$, it should be possible (at least numerically) to find a set of p_i that minimize γ_R while maintaining $\gamma_R < 1$. Similarly, the drift rate γ_P of the RPGS drift condition depends on permutation probabilities q_i.

Establishing geometric ergodicity requires the existence of both a drift condition and an associated minorization condition. In fact, this requirement should help guide the selection of building block functions f_i. It is possible to write down recipes for the minorization conditions. However, due to the increased complexity of the drift conditions and transition densities of the d-component Gibbs samplers, such conditions are best constructed on a case-by-case basis. See Example 4.1 for an illustration. Finally, in the 2-component case we showed that geometric ergodicity of DUGS guaranteed geometric ergodicity of RPGS and RSGS. It is not clear to us if and how this result can be extended to the d-component setting. This will be a topic of future research.

Chapter 5

Examples and Applications

In this chapter we apply our results to two broad classes of models. First, in Chapter 5.1 we illustrate the applications of our recipes using a toy example. Specifically, we construct drift and minorization conditions for Gibbs samplers for hierarchical exponential family models. In Chapter 5.2 we establish geometric ergodicity of Gibbs samplers for a Bayesian hierarchical general linear model. This example is practically relevant in the sense that it is not straightforward to sample directly from the distributions of interest. To conclude, we apply these results in the analysis of US government health maintenance organization (HMO) data.

5.1 The Exponential Family

Let λ be a nonzero Borel measure on $\mathcal{X} \subseteq \mathbb{R}^d$. The Laplace transform of λ is

$$c(\theta) = \int_{\mathcal{X}} e^{\langle x,\theta \rangle} \lambda(dx)$$

where $\langle \cdot, \cdot \rangle$ is the usual Euclidean inner product. Consider the exponential family $\{\pi(\cdot|\theta) : \theta \in \Theta\}$ with respect to measure λ where

$$\Theta = \left\{\theta \in \mathbb{R}^d : c(\theta) < \infty\right\} \quad \text{and} \quad \pi(x|\theta) = \frac{1}{c(\theta)} e^{\langle x,\theta \rangle} .$$

Suppose $X_1, \cdots, X_n$ are iid $\pi(x|\theta)$. Further, let γ be a nonzero Borel measure on Θ and let the prior on Θ be

$$p(\theta) = \frac{1}{d(a,k)} e^{\langle a,\theta\rangle - k\log c(\theta)}$$

where

$$d(a,k) = \int_\Theta e^{\langle a,\theta\rangle - k\log c(\theta)} \gamma(d\theta)$$

and (a,k) are hyperparameters in the domain of d. The corresponding posterior distribution is

$$\pi(\theta|x) = \frac{1}{d\left(a + \sum_{i=1}^n x_i, k+n\right)} e^{\left\langle a + \sum_{i=1}^n x_i, \theta\right\rangle - (k+n)\log c(\theta)} .$$

Therefore, $p(\theta)$ is a conjugate prior.

We restrict our attention to exponential families with quadratic variance functions. Diaconis et al. (2008b) consider a similar setting for a slightly less general formulation of the exponential family. Under the above set-up, the quadratic variance property guarantees the existence of scalars $\tilde{a}, \tilde{c}, \tilde{e}, \tilde{f}, \tilde{h}, \tilde{k}$ and d-vectors $\tilde{b}, \tilde{d}, \tilde{g}, \tilde{j}$ such that

$$\begin{aligned}
\mathrm{E}(X_i|\theta) &= \tilde{a}\theta + \tilde{b} \\
\mathrm{E}(\langle X_i, X_i\rangle |\theta) &= \tilde{c}\langle\theta,\theta\rangle + \left\langle \tilde{d}, \theta \right\rangle + \tilde{e} \\
\mathrm{E}(\theta|X) &= \tilde{f}\sum_{i=1}^n X_i + \tilde{g} \\
\mathrm{E}(\langle\theta,\theta\rangle |X) &= \tilde{h}\left\langle \sum_{i=1}^n X_i, \sum_{i=1}^n X_i \right\rangle + \left\langle \tilde{j}, \sum_{i=1}^n X_i \right\rangle + \tilde{k}
\end{aligned} \tag{5.1}$$

where $X := \left(X_1^T, X_2^T, \cdots, X_n^T\right)^T$. Diaconis et al. (2008b) discuss six examples of this special class of models for $d = 1$. One of these is the Normal-Normal model.

Example 5.1. The Normal-Normal Model

Let N_k denote a k-variate Normal distribution and I_k denote the k-dimensional identity matrix. Then the Normal-Normal model for $d \in \{1, 2, \cdots\}$ is

$$\begin{aligned} X_i|\theta &\sim N_d\left(\theta, \sigma^2 I_d\right) \\ \theta &\sim N_d\left(\mu, \tau^2 I_d\right) \end{aligned}$$

where $0 < \sigma^2, \tau^2 < \infty$. In this case, the posterior is

$$\theta|X \sim N_d\left(\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i + \frac{1}{\tau^2}\mu\right),\ \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2} I_d\right)$$

and (5.1) holds since

$$\begin{aligned} \mathrm{E}(X_i|\theta) &= \theta \\ \mathrm{E}(\langle X_i, X_i\rangle|\theta) &= \langle\theta, \theta\rangle + d\sigma^2 \\ \mathrm{E}(\theta|X) &= \frac{\tau^2}{\sigma^2 + n\tau^2}\sum_{i=1}^{n} X_i + \frac{\sigma^2}{\sigma^2 + n\tau^2}\mu \\ \mathrm{E}(\langle\theta, \theta\rangle|X) &= \frac{\tau^4}{(\sigma^2 + n\tau^2)^2}\left\langle\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i\right\rangle + \left\langle\frac{2\sigma^2\tau^2}{(\sigma^2 + n\tau^2)^2}\mu, \sum_{i=1}^{n} X_i\right\rangle \\ &\quad + \left[\frac{d\sigma^2\tau^2}{\sigma^2 + n\tau^2} + \frac{\sigma^4}{(\sigma^2 + n\tau^2)^2}\langle\mu, \mu\rangle\right]. \end{aligned}$$

5.1.1 The 2-component Gibbs Sampler

Consider 2-component Gibbs sampling for the joint distribution $\pi(X, \theta)$. Though it is straightforward to sample directly from $\pi(X, \theta)$, this provides a clean illustration of our geometric ergodicity results among and within the DUGS, RPGS, and RSGS

settings.

The transition densities corresponding to the DUGS (under update order (X, θ)), RPGS, and RSGS for this model are as follows:

$$\begin{aligned}
k_D\left((x', \theta'), (x, \theta)\right) &= \pi\left(x|\theta'\right)\pi(\theta|x) \\
k_P\left((x', \theta'), (x, \theta)\right) &= q_1\pi\left(x|\theta'\right)\pi(\theta|x) + q_2\pi\left(\theta|x'\right)\pi(x|\theta) \\
k_R\left((x', \theta'), (x, \theta)\right) &= p_1\pi\left(x|\theta'\right) I\left(\theta = \theta'\right) + p_2\pi\left(\theta|x'\right) I\left(x = x'\right) \ .
\end{aligned}$$

Therefore, $k_D\left((x', \theta'), (x, \theta)\right)$, $k_P\left((x', \theta'), (x, \theta)\right)$, and $k_R^2\left((x', \theta'), (x, \theta)\right)$ are strictly positive for all (x', θ') and (x, θ) in $\mathbb{R}^{nd} \times \Theta$. It follows that the corresponding transition kernels are absolutely continuous with respect to $\pi(x, \theta)$. Therefore, the resulting Markov chains are Harris ergodic by Lemma 4.1. We can also establish geometric ergodicity for the Gibbs samplers by constructing drift and minorization conditions. To this end, we construct functions f_1 and f_2 which satisfy (4.1) by exploiting the quadratic variance structure of the exponential families. In addition, we construct functions g_1 and g_2 which satisfy (4.2) and (4.3). Proofs are given in Appendix B.

Lemma 5.1. *Define functions f_1 and f_2*

$$f_1(x) = \left\langle \sum_{i=1}^{n} x_i - \tilde{u}, \sum_{i=1}^{n} x_i - \tilde{u} \right\rangle \quad \text{and} \quad f_2(\theta) = \langle \theta - \tilde{v}, \theta - \tilde{v} \rangle \tag{5.2}$$

for

$$\tilde{u} = \frac{\tilde{f}\tilde{d} + \tilde{c}\tilde{j} + (n-1)\tilde{a}\left(\tilde{a}\tilde{j} + 2\tilde{f}\tilde{b}\right)}{2\left[\tilde{a}\tilde{f} - \tilde{h}(\tilde{c} + (n-1)\tilde{a}^2)\right]} \quad \text{and} \quad \tilde{v} = \frac{\tilde{h}\tilde{d} + \tilde{a}\tilde{j} + 2(n-1)\tilde{a}\tilde{h}\tilde{b}}{2\left[\tilde{a}\tilde{f} - \tilde{h}(\tilde{c} + (n-1)\tilde{a}^2)\right]} \ .$$

Then f_1 and f_2 satisfy (4.1) *with*

$$\begin{aligned}
a &= n\tilde{c} + n(n-1)\tilde{a}^2, \\
b &= n\tilde{e} + n(n-1)\left\langle \tilde{b}, \tilde{b} \right\rangle - 2n\left\langle \tilde{u}, \tilde{b} \right\rangle + \langle \tilde{u}, \tilde{u} \rangle - a\langle \tilde{v}, \tilde{v} \rangle, \\
c &= \tilde{h}, \text{ and} \\
d &= \langle \tilde{v}, \tilde{v} \rangle - 2\langle \tilde{v}, \tilde{g} \rangle + \tilde{k} - \tilde{h}\langle \tilde{u}, \tilde{u} \rangle
\end{aligned}$$

where we assume $ac < 1$.

Remark 5.1. The assumption that $ac < 1$ may require restrictions on some of the exponential family hyperparameters.

Lemma 5.2. *Define $D_1 = \{(x, \theta) : f_1(x) \leq \nu_1\}$ and $D_2 = \{(x, \theta) : f_2(\theta) \leq \nu_2\}$ for f_i defined by Lemma 5.1 where ν_1, ν_2 are any positive numbers for which the D_i are non-empty and $D_i \subset \mathbb{R}^{nd} \times \Theta$. Also, define positive functions g_1 and g_2*

$$\begin{aligned}
g_1(x) = \exp &\left\{ \left\langle \sum_{i=1}^n x_i, \tilde{v} - \sqrt{\nu_2} sgn\left(\sum_{i=1}^n x_i \right) \right\rangle \right\} \\
&\cdot \left[\int_{\mathcal{X}} \exp\left\{ \langle x, \tilde{v} + \sqrt{\nu_2} sgn(x) \rangle \right\} \lambda(dx) \right]^{-n} \\
g_2(\theta) = \exp &\left\{ \langle \theta, a + \tilde{u} - \sqrt{\nu_1} sgn(\theta) \rangle - (k+n) \log c(\theta) \right\} \\
&\cdot \left[\int_{\Theta} \exp\left\{ \langle \theta, a + \tilde{u} + \sqrt{\nu_1} sgn(\theta) \rangle - (k+n) \log c(\theta) \right\} \gamma(d\theta) \right]^{-1}
\end{aligned}$$

where $sgn(y) = (sgn(y_1), \cdots, sgn(y_d))^T$ for any $y \in \mathbb{R}^d$ and $sgn(y_i)$ denotes the usual sign function. Then g_1 and g_2 satisfy (4.2) *and* (4.3)*:*

$$\inf_{(x', \theta') \in D_2} \pi(x|\theta') \geq g_1(x) \quad \text{and} \quad \inf_{(x', \theta') \in D_1} \pi(\theta|x') \geq g_2(\theta)$$

where $\pi(x|\theta')$ is the full conditional density of x with respect to $\lambda(dx)$ and $\pi(\theta|x')$ is the full conditional density of θ with respect to $\gamma(d\theta)$.

Theorems 4.1, 4.2, and 4.3 can be used in conjuction with Lemmas 5.1 and 5.2 to construct drift and minorization conditions for the DUGS, RPGS, and RSGS for $\pi(x, \theta)$. Geometric ergodicity for these samplers follows directly.

Proposition 5.1. *Under the assumptions of Lemma 5.1, suppose Lemma 5.2 holds for ν_1 and ν_2 that satisfy the conditions of Corollary 4.1 for the appropriate drift condition parameters. Then the DUGS, RPGS, and RSGS for the joint distribution $\pi(x, \theta)$ arising from the exponential family with quadratic variance functions are geometrically ergodic.*

Remarks:

1. We make no claim that the above drift and minorization conditions are optimal for all (or even any) of the exponential family models of interest. In fact, it is easy to cook up many more functions V for which a drift condition is satisfied, especially if each family is considered individually. We prefer drift functions that lead to larger values of ε in a minorization condition (2.4) for $C = \{x : V(x) \leq \omega\}$.

2. Diaconis et al. (2008b) use alternative methods to establish geometric ergodicity of the DUGS and RSGS for the univariate case of the exponential family models considered here. However, their methods do not extend to the multivariate case and do not produce drift and minorization conditions.

5.1.2 Example: The Normal-Normal Model

As an illustration of the 2-component Gibbs sampler for $\pi(X, \theta)$, we consider the Normal-Normal model where $X_1, X_2, \ldots, X_n$ are iid $X_i|\theta \sim N(\theta, 1/4)$ and $\theta \sim N(0, 1/4)$. In this case, the joint distribution is multivariate normal. Further, it

follows from Example 5.1 that

$$\theta|X \sim N\left(\frac{1}{n+1}\sum_{i=1}^{n} X_i,\ \frac{1}{4(n+1)}\right)$$

and

$$\begin{aligned} \mathrm{E}(X_i|\theta) &= \theta & \mathrm{E}(\theta|X) &= \frac{1}{n+1}\sum_{i=1}^{n} X_i \\ \mathrm{E}\left(X_i^2|\theta\right) &= \theta^2 + \frac{1}{4} & \mathrm{E}\left(\theta^2|X\right) &= \frac{1}{(n+1)^2}\left(\sum_{i=1}^{n} X_i\right)^2 + \frac{1}{4(n+1)} \end{aligned} . \tag{5.3}$$

Therefore, (5.1) holds for $\tilde{a} = \tilde{c} = 1$, $\tilde{e} = 1/4$, $\tilde{f} = 1/(n+1)$, $\tilde{h} = 1/(n+1)^2$, $\tilde{k} = 1/[4(n+1)]$, and $\tilde{b} = \tilde{d} = \tilde{g} = \tilde{j} = 0$.

Since the Normal-Normal model follows the exponential family framework with quadratic variance functions, we can apply Lemma 5.1 to $\tilde{a}, \tilde{b}$, etc to obtain functions f_1 and f_2 for the DUGS, RPGS, and RSGS drift conditions.

Lemma 5.3. *Let*

$$f_1(x) = \left(\sum_{i=1}^{n} x_i\right)^2 \quad \text{and} \quad f_2(\theta) = \theta^2 \,. \tag{5.4}$$

Then f_1 and f_2 satisfy (4.1) with $a = n^2$, $b = n/4$, $c = 1/(n+1)^2$, and $d = 1/[4(n+1)]$ where $ac < 1$.

We can also obtain functions g_1 and g_2 for the Gibbs sampler minorization conditions. A proof is similar to that for Lemma 5.2, hence is left to the reader.

Lemma 5.4. *Define functions*

$$\begin{aligned} g_1(x) &= \left(\frac{2}{\pi}\right)^{n/2} \exp\left\{-2\sum_{i=1}^{n}\left(x_i + sgn(x_i)\sqrt{\nu_2}\right)^2\right\} \quad \text{and} \\ g_2(\theta) &= \left(\frac{2(n+1)}{\pi}\right)^{1/2} \exp\left\{-2(n+1)\left(\theta + sgn(\theta)\frac{\sqrt{\nu_1}}{n+1}\right)^2\right\} \end{aligned}$$

for any $\nu_1, \nu_2 > 0$. *Then* g_1 *and* g_2 *satisfy* (4.2) *and* (4.3). *Specifically,*

$$\inf_{(x',\theta')\in D_2} \pi(x|\theta') \geq g_1(x) \quad \textit{and} \quad \inf_{(x',\theta')\in D_1} \pi(\theta|x') \geq g_2(\theta)$$

where $D_1 = \{(x,\theta) : f_1(x) \leq \nu_1\}$ *and* $D_2 = \{(x,\theta) : f_2(\theta) \leq \nu_2\}$ *for* f_1 *and* f_2 *defined by* (5.4).

Remark 5.2. Theorems 4.1, 4.2, and 4.3 in conjunction with Lemmas 5.3 and 5.4 establish geometric ergodicity for the DUGS, RPGS, and RSGS for the Normal-Normal joint distribution. Further, explicit drift and minorization conditions can be constructed by applying Theorems 4.1, 4.2, and 4.3 to Lemmas 5.3 and 5.4.

Our goal is to explore the convergence relationships among the Gibbs samplers as well as the potential implications of these relationships for MCMC estimation. To this end, we consider estimating $\mathrm{E}(\theta) = 0$ and measure the quality of MCMC estimators using mean squared error (MSE). Let $\overline{\theta}$ denote a Monte Carlo average based on m iterations. Then the MSE of $\overline{\theta}$ for estimating $\mathrm{E}(\theta) = 0$ is

$$\mathrm{MSE}(\overline{\theta}) = \mathrm{E}\left(\overline{\theta} - \mathrm{E}(\theta)\right)^2 = \mathrm{E}\left(\overline{\theta}\right)^2 \ .$$

In the typical MCMC setting, it is not possible to directly evaluate MSE. However, we can estimate this quantity by

$$\widehat{\mathrm{MSE}}(\overline{\theta}) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\overline{\theta}^{\,(i)}\right)^2$$

where $\left\{\overline{\theta}^{\,(1)}, \cdots, \overline{\theta}^{\,(1000)}\right\}$ is an independent sample of Monte Carlo averages. Further, standard errors of $\widehat{\mathrm{MSE}}(\overline{\theta})$ can be estimated by $\mathrm{se}(\widehat{\mathrm{MSE}}) = s/\sqrt{1000}$ where s is the standard deviation of the $\left\{\left(\overline{\theta}^{\,(i)}\right)^2\right\}$ sample.

To compare the quality of estimates $\overline{\theta}_1$ and $\overline{\theta}_2$ resulting from two different Gibbs

samplers, we can estimate

$$\frac{\mathrm{MSE}(\overline{\theta}_1)}{\mathrm{MSE}(\overline{\theta}_2)} \quad \text{by} \quad \frac{\widehat{\mathrm{MSE}}(\overline{\theta}_1)}{\widehat{\mathrm{MSE}}(\overline{\theta}_2)} := \frac{\widehat{\mathrm{MSE}}_1}{\widehat{\mathrm{MSE}}_2}$$

where ratios greater than one favor $\overline{\theta}_2$. A consistent standard error of this ratio estimate can be derived using a first-order Taylor expansion:

$$\mathrm{se}\left(\frac{\widehat{\mathrm{MSE}}_1}{\widehat{\mathrm{MSE}}_2}\right) = \frac{\widehat{\mathrm{MSE}}_1}{\widehat{\mathrm{MSE}}_2}\sqrt{\left(\frac{\mathrm{se}(\widehat{\mathrm{MSE}}_1)}{\widehat{\mathrm{MSE}}_1}\right)^2 + \left(\frac{\mathrm{se}(\widehat{\mathrm{MSE}}_2)}{\widehat{\mathrm{MSE}}_2}\right)^2}\,.$$

See Example 5.5.27 of Casella and Berger (2002) for details.

We begin by comparing the relationships between DUGS, uniform RPGS ($q_1 = q_2 = 0.50$), and uniform RSGS ($p_1 = p_2 = 0.50$). In this case, Theorems 4.1, 4.2, and 4.3 can be applied to Lemma 5.3 to obtain corresponding drift rates

$$\gamma_D = \left(\frac{n}{n+1}\right)^2, \quad \gamma_P = \left(\frac{n}{n+1}\right)\left(\frac{2n+1}{2n+2}\right), \quad \text{and} \quad \gamma_R = \frac{2n+1}{2n+2}\,. \tag{5.5}$$

Notice that these drift rates satisfy $\gamma_D < \gamma_P < \gamma_R^2 < \gamma_R$, a relationship observed in the general setting (see (4.4)). Furthermore, each drift rate converges to one as n increases. This indicates that, theoretically, the Gibbs sampler becomes less efficient as the model sample size increases. In addition, the gain in efficiency observed for DUGS diminishes as n increases. These relationships within and among the drift rates are illustrated in Figure 5.1.

Using the MSE procedure described above, we can explore the potential implications of these convergence relationships for the quality of Markov chain estimates. To this end, for every combination of $n \in \{10, 100\}$, Markov chain simulation length $m \in \{1000, 10000\}$, and scanning strategy (DUGS, uniform RPGS, uniform RSGS), we ran 1000 independent Markov chains, each of length m and each started at

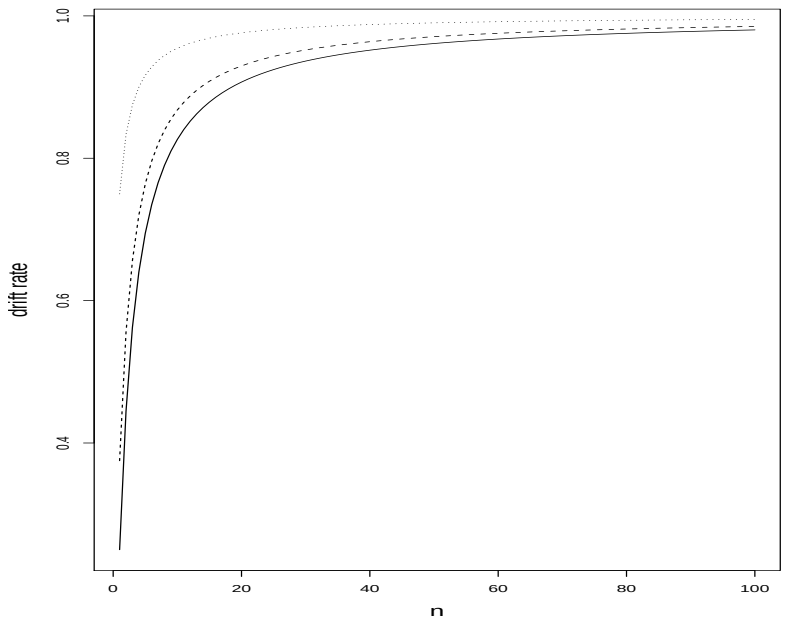


Figure 5.1: The 2-component DUGS, uniform RPGS, and uniform RSGS drift rates for the Normal-Normal model are given by (5.5). In this plot, the drift rates γ_D (solid line), γ_P (dashed line), and γ_R (dotted line) are plotted versus sample size n.

$\left(x^{(0)}, \theta^{(0)}\right) = 0_{n+1}$ (where 0_m denotes an $m \times 1$ vector of ones). From each chain, we obtained a Monte Carlo average

$$\overline{\theta} = \frac{1}{m} \sum_{i=1}^{m} \theta^{(i)} \ .$$

In other words, for each combination of n, m, and Gibbs scanning strategy, we obtained 1000 independent estimates of $\mathrm{E}(\theta)$. In each setting, these estimates were used to estimate $\mathrm{MSE}(\overline{\theta})$. The MSE estimates are reported in Table 5.1. It is clear that, as expected, MSE's shrink as the simulation length increases and increase as the model sample size n increases.

Next, to better compare MSE's among scanning strategies, we also estimated MSE

Table 5.1: MSE's (and standard errors) for the Markov chain estimates of $\mathrm{E}(\theta)$.

	m	$n = 10$	$n = 100$
DUGS	1000	0.0048 (0.0002)	0.0413 (0.0018)
	10000	0.0005 (0.00002)	0.0047 (0.0002)
Uniform RPGS	1000	0.0062 (0.0003)	0.0537 (0.0025)
	10000	0.0007 (0.00003)	0.0066 (0.0003)
Uniform RSGS	1000	0.0203 (0.0009)	0.0935 (0.0041)
	10000	0.0020 (0.00008)	0.0177 (0.0008)

ratios relative to DUGS using

$$\frac{\widehat{\mathrm{MSE}}(\overline{\theta}_{\mathrm{RPGS}}, m)}{\widehat{\mathrm{MSE}}(\overline{\theta}_{\mathrm{DUGS}}, m)} \quad \text{and} \quad \frac{\widehat{\mathrm{MSE}}(\overline{\theta}_{\mathrm{RSGS}}, m)}{\widehat{\mathrm{MSE}}(\overline{\theta}_{\mathrm{DUGS}}, m)}$$

where m denotes the Gibbs sampler simulation length. These estimates are reported in Table 5.2. Conclusions from this table are similar to those drawn from the comparison of drift rates. First, ratios greater than one favor the DUGS. Therefore, in terms of MSE, DUGS is more efficient than the uniform RPGS and uniform RSGS in each model and Monte Carlo sample size setting. In addition, the gain in efficiency is larger when compared to the uniform RSGS than the uniform RPGS. Also, though the MSE's decrease as simulation length increases, the MSE ratios remain relatively constant when $n = 10$. In other words, it appears that the MSE's for the different samplers decrease with simulation length at similar rates. On the other hand, when $n = 100$, the DUGS MSE's appear to decrease at a slightly quicker rate as m increases in comparison to RPGS and RSGS.

Recall that RSGS needs to be run for twice as many iterations as DUGS to have the potential to obtain the same number of updates for both components. Therefore, it might be more fair to compare DUGS with m iterations to RSGS with $2m$ iterations. To this end, for every combination of $n \in \{10, 100\}$ and $m \in \{1000, 10000\}$, we also ran 1000 independent uniform RSGS chains, each of length $2m$. The corresponding

Table 5.2: MSE ratios relative to DUGS (and standard errors) for the Markov chain estimates of $\mathrm{E}(\theta)$.

	m	$n = 10$	$n = 100$
Uniform RPGS	1000	1.304 (0.081)	1.301 (0.083)
	10000	1.276 (0.080)	1.396 (0.089)
Uniform RSGS	1000	4.268 (0.266)	2.265 (0.141)
	10000	3.912 (0.241)	3.772 (0.236)

estimates $\overline{\theta}$ were used to estimate $\mathrm{MSE}(\overline{\theta}_{\mathrm{RSGS}}, 2m)$. In turn, MSE ratios relative to DUGS were estimated by

$$\frac{\widehat{\mathrm{MSE}}(\overline{\theta}_{\mathrm{RSGS}}, 2m)}{\widehat{\mathrm{MSE}}(\overline{\theta}_{\mathrm{DUGS}}, m)} .$$

These are reported in Table 5.3.

Table 5.3: MSE ratios of RSGS ($2m$ iterations) relative to DUGS (m iterations) for the Markov chain estimates of $\mathrm{E}(\theta)$.

m	$n = 10$	$n = 100$
1000	1.991 (0.124)	1.806 (0.108)
10000	1.991 (0.126)	1.990 (0.123)

Even when run for half as many iterations, DUGS outperforms RSGS in terms of MSE. Similarly, RSGS is still less efficient than RPGS (see Table 5.2).

We now consider the impact of permutation probabilities q_i and selection probabilities p_i on the RPGS and RSGS. In the general non-uniform case, an application of Theorems 4.1, 4.2, and 4.3 to Lemma 5.3 produces drift rates

$$\begin{aligned} \gamma_P &= \frac{n}{2(n+1)^2}\left[n + \sqrt{n^2 + 4q_1q_2(2n+1)}\right]; \text{ and} \\ \gamma_R &= \frac{1}{2(n+1)}\left[(n+1) + \sqrt{(n+1)^2 - 4p_1p_2(2n+1)}\right] . \end{aligned} \tag{5.6}$$

The RPGS drift rate γ_P is maximized for $q_1 = 1/2$ and decreases as q_1 approaches either 0 or 1. On the other hand, the RSGS drift rate γ_R is minimized for $p_1 = 1/2$

and increases as p_1 approaches either 0 or 1. Also, for fixed q_1 and p_1, both γ_P and γ_R converge to 1 as n increases. These relationships are depicted in Figure 5.2.

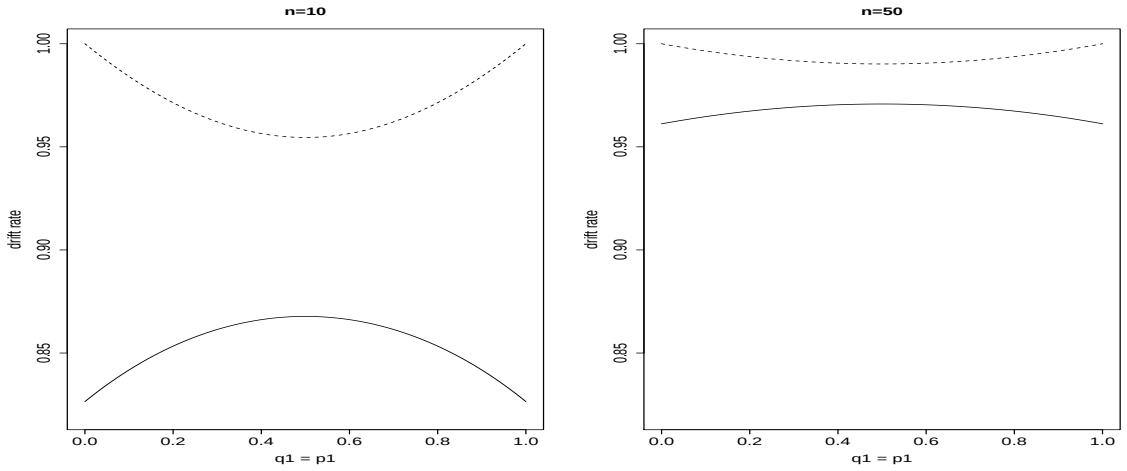


Figure 5.2: Drift rate γ_P is plotted versus permutation probability q_1 (solid line) and drift rate γ_R is plotted versus selection probability p_1 (dashed line). Drift rates are calculated using $n = 10$ (left) and $n = 50$ (right).

The impact of q_i and p_i is also reflected in a comparison of Markov chain estimates. For every combination of $n \in \{10, 100\}$, simulation length $m \in \{1000, 10000\}$, and permutation probability $q_1 \in \{0.01, 0.25, 0.50, 0.75, 0.99\}$ we ran 1000 independent Markov chains of length m under the RPGS strategy. Similarly, for every combination of $n \in \{10, 100\}$, simulation length $m \in \{1000, 10000\}$, and selection probability $p_1 \in \{0.01, 0.25, 0.50, 0.75, 0.99\}$ we ran 1000 independent Markov chains of length m under the RSGS strategy. All chains were started in state $(x^{(0)}, \theta^{(0)}) = 0_{n+1}$. In each setting, this produced 1000 independent averages $\overline{\theta}$ which were used to estimate $\mathrm{MSE}(\overline{\theta})$. In all but two of these settings, the MSE's decreased as m increased. The exceptions are the RSGS for $n = 100$ and $p_1 \in \{0.01, 0.99\}$.

Estimates of the MSE ratios relative to the uniform scanning strategies,

$$\frac{\widehat{\mathrm{MSE}}(\overline{\theta}_{\mathrm{RPGS}}, m)}{\widehat{\mathrm{MSE}}(\overline{\theta}_{\mathrm{RPGS},q_1=0.50}, m)} \quad \text{and} \quad \frac{\widehat{\mathrm{MSE}}(\overline{\theta}_{\mathrm{RSGS}}, m)}{\widehat{\mathrm{MSE}}(\overline{\theta}_{\mathrm{RSGS},p_1=0.50}, m)},$$

are reported in Table 5.4. Ratios greater than one favor the uniform settings. As suspected, the efficiency of RPGS as measured by MSE is greatest for small or large values of q_1 ($q_1 = 0.01$ or $q_1 = 0.99$). Also, for fixed q_1, the improvement over the uniform strategy remains somewhat constant across changes in n and m. On the other hand, RSGS is most efficient in the uniform setting. The contrast between MSE's is even sharper when $m = 10000$. Again, the only exception is the RSGS setting with $n = 100$, $m = 1000$, and $p_1 \in \{0.01, 0.99\}$.

Table 5.4: MSE ratios relative to the uniform settings ($q_1 = 0.5$ for RPGS and $p_1 = 0.5$ for RSGS) for the Markov chain estimates of $\mathrm{E}(\theta)$.

	n	m	$q_1 = p_1 = 0.01$	$q_1 = p_1 = 0.25$	$q_1 = p_1 = 0.75$	$q_1 = p_1 = 0.99$
RPGS	10	1000	0.794 (0.049)	0.966 (0.062)	0.899 (0.060)	0.835 (0.052)
	10	10000	0.808 (0.050)	0.952 (0.060)	0.867 (0.054)	0.697 (0.044)
	100	1000	0.760 (0.048)	0.885 (0.056)	0.885 (0.054)	0.784 (0.050)
	100	10000	0.738 (0.045)	0.882 (0.056)	0.856 (0.053)	0.736 (0.047)
RSGS	10	1000	3.809 (0.243)	1.255 (0.078)	1.390 (0.086)	4.074 (0.255)
	10	10000	18.914 (1.226)	1.213 (0.076)	1.271 (0.084)	21.261 (1.321)
	100	1000	0.157 (0.011)	0.994 (0.065)	1.030 (0.068)	0.155 (0.011)
	100	10000	3.884 (0.253)	1.241 (0.080)	1.250 (0.080)	3.975 (0.261)

5.2 A Bayesian Hierarchical General Linear Model

Consider the following version of the normal theory general linear model: Let Y denote an $N \times 1$ response vector and suppose β is a $p \times 1$ vector of regression coefficients, u is a $k \times 1$ vector of parameters, X is a known $N \times p$ design matrix having full column

rank, and Z is a known $N \times k$ matrix. Then for $r, s, t \in \{1, 2, \cdots\}$, the hierarchy is

$$\begin{aligned} Y|\beta, u, \lambda_R, \lambda_D &\sim \mathrm{N}_N\left(X\beta + Zu, \lambda_R^{-1} I_N\right) \\ \beta|u, \lambda_R, \lambda_D &\sim \sum_{i=1}^{r} \eta_i \mathrm{N}_p\left(\beta_i, B^{-1}\right) \\ u|\lambda_R, \lambda_D &\sim \mathrm{N}_k\left(0, \lambda_D^{-1} I_k\right) \\ \lambda_R &\sim \sum_{j=1}^{s} \phi_j \mathrm{Gamma}\left(r_{j1}, r_{j2}\right) \\ \lambda_D &\sim \sum_{l=1}^{t} \psi_l \mathrm{Gamma}\left(d_{l1}, d_{l2}\right) \ . \end{aligned} \tag{5.7}$$

where β and u are assumed to be conditionally independent given λ_R and λ_D and we say $W \sim \mathrm{Gamma}(a, b)$ if it has density proportional to $w^{a-1}e^{-bw}$ for $w > 0$. Also, B positive definite is assumed known and hyperparameters r_1, r_2, d_1, and d_2 are all assumed to be positive. The general linear model is one of the most powerful tools statisticians have at their disposal. Hence, this is one of the most popular Bayesian hierarchical models (see, for instance, Gelman et al. (2004) and Spiegelhalter et al. (2005)). The incorporation of the mixture priors increases this model's flexibility.

Let $\xi = \left(u^T, \beta^T\right)^T$ and $\lambda = (\lambda_R, \lambda_D)^T$. Then for some observed data y, the posterior is characterized by

$$\pi(\xi, \lambda|y) \propto \pi(y|\xi, \lambda)\pi(\xi|\lambda)\pi(\lambda)$$

and is proper under the positivity of hyperparameters r_1, r_2, d_1, and d_2. Since the corresponding posterior distribution is complicated, finding closed form solutions to integrals involving the posterior can be prohibitively difficult as is direct simulation from the posterior. Therefore, inference may require MCMC methods.

We consider 2-component DUGS, RPGS, and RSGS for $\pi(\xi, \lambda|y)$. Geometric er-

godicity for these samplers will be established by applying the results of Chapter 4. We know of three other papers that have addressed geometric ergodicity of Gibbs samplers in the context of the normal theory linear model with proper priors. These are Hobert and Geyer (1998), Jones and Hobert (2004), and Papaspiliopoulos and Roberts (2007). The linear model we consider substantively differs from those in Papaspiliopoulos and Roberts (2007) in that we do not assume the variance components are known. Our model is also much more general than the one-way random effects model in Hobert and Geyer (1998) and Jones and Hobert (2004). Gibbs sampling for the balanced one-way random effects model is also considered in Rosenthal (1995b) where coupling techniques were used to establish upper bounds on the total variation distance to stationarity. However, these results fall short of establishing geometric ergodicity of the associated Markov chain. Thus our results are the first that give practitioners verifiable conditions which guarantee geometric ergodicity of MCMC algorithms for such a general and widely applicable Bayesian hierarchical model. Proofs of our results are deferred to Appendix B.

5.2.1 Gibbs Sampling for $\pi(\xi, \lambda|y)$

Let $\Phi = \{(\xi_0, \lambda_0), (\xi_1, \lambda_1), \cdots\}$ denote the Markov chain produced by a 2-component Gibbs sampler for $\pi(\xi, \lambda|y)$. The full conditional distributions required for the Gibbs updates are as follows:

$$\begin{aligned}
\xi|\lambda, y &\sim \sum_{i=1}^{r} \eta_i \mathrm{N}_{k+p}(\xi_{i0}, \Sigma^{-1}) \\
\lambda|\xi, y &\sim \sum_{j=1}^{s}\sum_{l=1}^{t} \phi_j\psi_l \mathrm{Gamma}\left(r_{j1} + \frac{N}{2}, r_{j2} + \frac{1}{2}v_1(\xi)\right) \mathrm{Gamma}\left(d_{l1} + \frac{k}{2}, d_{l2} + \frac{1}{2}v_2(\xi)\right)
\end{aligned} \tag{5.8}$$

where

$$v_1(\xi) := (y - X\beta - Zu)^T(y - X\beta - Zu), \qquad v_2(\xi) := u^T u, \tag{5.9}$$

and

$$\begin{aligned} \Sigma^{-1} &= \begin{pmatrix} \left(\lambda_R Z^T Z + \lambda_D I_k\right)^{-1} & 0 \\ 0 & \left(\lambda_R X^T X + B\right)^{-1} \end{pmatrix} \\ \xi_{i0} &= \begin{pmatrix} \lambda_R \left(\lambda_R Z^T Z + \lambda_D I_k\right)^{-1} Z^T y \\ \left(\lambda_R X^T X + B\right)^{-1} \left(\lambda_R X^T y + B\beta_i\right) \end{pmatrix}. \end{aligned} \tag{5.10}$$

See Appendix B for a proof. For ease of exposition, we will often drop the dependence of the full conditionals on the data y.

The structures of the full conditional distributions guarantee that the DUGS, RPGS, and RSGS transition densities

$$k_D((\xi', \lambda'), (\xi, \lambda)), \quad k_P((\xi', \lambda'), (\xi, \lambda)), \quad \text{and} \quad k_R^2((\xi', \lambda'), (\xi, \lambda))$$

are strictly positive for all (ξ', λ'), (ξ, λ) in $\mathbb{R}^{k+p} \times \mathbb{R}_+^2$. The corresponding transition kernels are also absolutely continuous with respect to posterior $\pi(\xi, \lambda|y)$. Therefore, the DUGS, RPGS, and RSGS are Harris ergodic by Lemma 4.1.

5.2.2 Geometric Ergodicity

In discussing geometric ergodicity, we will need the following notation. For all $j \in \{1, \cdots, s\}$ and $l \in \{1, \cdots, t\}$, define constants

$$\begin{aligned} \delta_{j1} &= \frac{\sum_{i=1}^N z_i \left(Z^T Z\right)^{-1} z_i^T}{2r_{j1} + N - 2}; \quad \delta_{l2} = \frac{k}{2d_{l1} + k - 2}; \\ \delta_{j3} &= \frac{\sum_{i=1}^N x_i \left(X^T X\right)^{-1} x_i^T}{4(2r_{j1} + N - 2)}; \text{ and } \quad \delta_{l4} = \frac{k + \sum_{i=1}^N z_i z_i^T}{2d_{l1} + k - 2}. \end{aligned} \tag{5.11}$$

and let x_i, z_i denote the ith rows of matrices X and Z, respectively. Also, let y_i and u_i denote the ith elements of vectors y and u, respectively.

Establishing geometric ergodicity for the Gibbs samplers first requires the construction of functions f_1 and f_2 that satisfy (4.1). We consider this construction for the setting in which Z^TZ is nonsingular, as well as that in which no restrictions are placed on Z. The restricted form of (5.7) for which Z^TZ is nonsingular accommodates, for instance, Bayesian random effects and random intercept models. The following lemma provides constructions of f_1 and f_2 for both settings. A proof can be found in Appendix B.

Lemma 5.5. *Fix $c' \in (0, \min_{j,l}\{d_{l2}, r_{j2}\})$ and let constants δ be defined by (5.11). Also, assume there exists some $\Delta^2 < \infty$ such that for all λ*

$$\sum_{i=1}^{N} \left[E\left(y_i - x_i\beta - z_iu|\lambda, y\right)\right]^2 + \sum_{i=1}^{k} \left[E\left(u_i|\lambda, y\right)\right]^2 \leq \Delta^2 \,. \tag{5.12}$$

Then the following two conditions hold.

1. *Suppose Z^TZ is nonsingular and define functions*

$$\begin{aligned} f_1(\xi) &= v_1(\xi) + v_2(\xi) \\ f_2(\lambda) &= \lambda_R^{-1} \sum_{i=1}^{N} z_i(Z^TZ)^{-1}z_i^T + \lambda_D^{-1}k + \exp\{c'\lambda_D\} + \exp\{c'\lambda_R\} \end{aligned} \tag{5.13}$$

for $v_1(\cdot)$ and $v_2(\cdot)$ as defined by (5.9). Then, if

$$r_{j1} > 0 \vee 0.5\left(\sum_{i=1}^{N} z_i(Z^TZ)^{-1}z_i^T - N + 2\right) \quad \text{for all } j \in \{1, \cdots, s\}$$

and $d_{l1} > 1$ for all $l \in \{1, \cdots, t\}$, f_1 and f_2 satisfy (4.1) with $a = 1$, $c =$

$\max_{j,l}\{\delta_{j1}, \delta_{l2}\}$,

$$b = \sum_{i=1}^{N} x_i B^{-1} x_i^T + \Delta^2, \quad \textit{and}$$
$$d = \max_{j,l}\left\{2r_{j2}\delta_{j1} + 2d_{l2}\delta_{l2} + \left(\frac{d_{l2}}{d_{l2}-c'}\right)^{k/2+d_{l1}} + \left(\frac{r_{j2}}{r_{j2}-c'}\right)^{N/2+r_{j1}}\right\}.$$

2. *Suppose Z^TZ is possibly singular and define functions*

$$\begin{aligned} f_1(\xi) &= v_1(\xi) + v_2(\xi) \\ f_2(\lambda) &= \lambda_R^{-1}\frac{1}{4}\sum_{i=1}^{N} x_i(X^TX)^{-1}x_i^T + \lambda_D^{-1}\left[k + \sum_{i=1}^{N} z_i z_i^T\right] \\ &\quad + \exp\{c'\lambda_D\} + \exp\{c'\lambda_R\} \end{aligned} \tag{5.14}$$

for $v_1(\cdot)$ and $v_2(\cdot)$ as defined by (5.9). *Then, if*

$$d_{l1} > 0.5\left(2 + \sum_{i=1}^{N} z_i z_i^T\right) \quad \textit{for all } l \in \{1, \cdots, t\}$$
$$r_{j1} > 0 \vee 0.5\left(0.25\sum_{i=1}^{N} x_i(X^TX)^{-1}x_i^T - N + 2\right) \quad \textit{for all } j \in \{1, \cdots, s\},$$

f_1 and f_2 satisfy (4.1) *with $a = 1$, $c = \max_{j,l}\{\delta_{j3}, \delta_{l4}\}$,*

$$b = \frac{1}{4}\sum_{i=1}^{N} x_i B^{-1} x_i^T + \Delta^2, \quad \textit{and}$$
$$d = \max_{j,l}\left\{2r_{j2}\delta_{j3} + 2d_{l2}\delta_{l4} + \left(\frac{d_{l2}}{d_{l2}-c'}\right)^{k/2+d_{l1}} + \left(\frac{r_{j2}}{r_{j2}-c'}\right)^{N/2+r_{j1}}\right\}.$$

Remarks:

1. The restrictions on hyperparameters r_{j1} and d_{l1} guarantee $0 < ac < 1$.

2. When Z^TZ is nonsingular, Conditions 1 and 2 of Lemma 5.5 both hold. However, Condition 1 requires weaker conditions for d_{l1}.

3. Lemma 5.5 requires (5.12) to hold for some Δ. We have been unable to show this except in some special cases. For example, (5.12) holds when $\beta_i = 0$ for all $i \in \{1, \cdots, r\}$ and Z^TZ nonsingular. See Appendix B for a proof.

   Proposition 5.2. *Assume Z^TZ is nonsingular and choose $\beta_i = 0$ for all $i \in \{1, \cdots, r\}$. Then* (5.12) *holds with*

   $$\Delta^2 = y^Ty + y^TZ(Z^TZ)^{-2}Z^Ty \, .$$

   When the direct calculation of a bound is not possible, it may be possible to find one numerically.

Drift conditions for the DUGS, RPGS, and RSGS can be constructed by applying Theorems 4.1, 4.2, and 4.3 to functions f_1 and f_2 given by either (5.13) or (5.14). Recall that the corresponding drift rates are $\gamma_D = ac$,

$$\gamma_P = \frac{1}{2}ac + \frac{1}{2}\sqrt{ac[ac + 4q_1q_2(1-ac)]} \quad \text{and} \quad \gamma_R = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4p_1p_2(1-ac)} \, .$$

Whether ac is given by Condition 1 or Condition 2 of Lemma 5.5, it is clear that each drift rate converges to 1 as $ac \to 1$. To better understand the implication of this relationship, consider the following examples.

Example 5.2. A Balanced Random Intercept Model

Consider the balanced random intercept model derived from (5.7) for k subjects with m observations each. In this case, $Z = I_k \otimes 1_m$ where $\otimes$ denotes the Kronecker product and 1_m represents a vector of ones of length m. Hence $Z^TZ = mI_k$ is nonsingular

and, if $d_{l1} > 1$ for all l, Condition 1 of Lemma 5.5 establishes

$$ac = \max_{j,l}\{\delta_{j1}, \delta_{l2}\} = \max_{j,l}\left\{\frac{k}{2r_{j1} + N - 2}, \frac{k}{2d_{l1} + k - 2}\right\} .$$

In this special case, ac (hence γ_D, γ_P, and γ_R) converges to 1 as $k \to \infty$. This observation supports our intuition that the Gibbs sampler, or any Markov chain, should converge more slowly as the number of model parameters increases.

On the other hand, if k is fixed but m increases so that $N = km \to \infty$, then

$$ac = \max_{l}\left\{\frac{k}{2d_{l1} + k - 2}\right\}$$

for sufficiently large N and drift rates γ_D, γ_P, and γ_R remain fixed. Thus increasing the number of observations per subject does not have the same negative, qualitative impact as increasing the number of subjects. Finally, the drift rates converge to one when k and N are held constant and $d_{l1} \to 1$ for any l.

Example 5.3. A Balanced Two-Way Model

Consider the special case of (5.7) corresponding to a balanced two-way layout with m levels for the first factor, n levels for the second factor, and r observations per treatment level combination. In this case $k = m + n$, $N = mnr$, and $X = 1_N$. Also, $Z = (I_m \otimes 1_{nr} \quad 1_m \otimes I_n \otimes 1_r)$ and $Z^T Z$ is singular. By Lemma 5.5, the block Gibbs samplers are geometrically ergodic if $d_{l1} > N + 1$ for all $l \in \{1, \cdots, t\}$. Therefore, an increase in k results in a stronger restriction on the choice of hyperparameters d_{l1}. The same is true when k is fixed but the number of observations per treatment level combination, r, increases. From Condition 2 of Lemma 5.5 we also have

$$ac = \max_{j,l}\left\{\frac{1}{4(2r_{j1} + N - 2)}, \frac{k + 2N}{2d_{l1} + k - 2}\right\} .$$

However, the dependence of ac on k and N is more complex than in Example 5.2.

We now provide the building blocks necessary for constructing minorization conditions for the DUGS, RPGS, and RSGS. This requires some new notation. First, define constants $\tilde{a}$ and $\tilde{b}$ for the setting in which Z^TZ is nonsingular and the setting in which Z^TZ is possibly singular.

Case 1: Z^TZ nonsingular

$$\tilde{a} = \sum_{i=1}^{N} z_i (Z^TZ)^{-1} z_i^T \quad \text{and} \quad \tilde{b} = k \tag{5.15}$$

Case 2: Z^TZ possibly singular

$$\tilde{a} = (1/4) \sum_{i=1}^{N} x_i (X^TX)^{-1} x_i^T \quad \text{and} \quad \tilde{b} = k + \sum_{i=1}^{N} z_i z_i^T \tag{5.16}$$

Next, define functions g_{i1} on $\mathbb{R}^{k+p}$ for $i \in \{1, \cdots, r\}$ such that

$$g_{i1}(\xi) = h_1(u) h_{i2}(\beta)$$

where

$$h_1(u) = \left(\frac{\tilde{b}}{2\pi\nu_2} \right)^{k/2} \exp\left\{ -\frac{1}{2} f(u) \right\}$$
$$h_{i2}(\beta) = (2\pi)^{-p/2} |B|^{1/2} \exp\left\{ -\frac{1}{2} g_i(\beta) \right\}$$

for

$$
\begin{aligned}
f(u) &= \frac{\log \nu_2}{c'} u^T (Z^T Z + I_k) u + \frac{\nu_2}{\tilde{b}} \left(\frac{\log \nu_2}{c'} \right)^2 y^T Z Z^T y - 2\nu(u) u^T Z^T y \\
\nu(u) &= \begin{cases} \frac{\log \nu_2}{c'} & u^T Z^T y < 0 \\ \frac{\tilde{a}}{\nu_2} & u^T Z^T y \geq 0. \end{cases} \\
g_i(\beta) &= \left(\frac{\log \nu_2}{2c'} \right)^2 y^T X B^{-1} X^T y + \frac{\nu_2}{4\tilde{a}} \beta_i^T B \left(X^T X \right)^{-1} B \beta_i \\
&\quad + \frac{\log \nu_2}{c'} \left[\beta^T X^T X \beta + \frac{1}{4} y^T X \left(X^T X \right)^{-1} X^T y \right] - \omega_i(\beta) \left[2\beta - \frac{1}{2}\beta_i \right]^T X^T y \\
&\quad + \left[\beta^T B \beta - 2\beta^T B \beta_i + \frac{1}{4} \beta_i^T B \beta_i + \frac{1}{2} y^T X \left(X^T X \right)^{-1} B \beta_i \right] \\
\omega_i(\beta) &= \begin{cases} \frac{\log \nu_2}{c'} & \left[2\beta - \frac{1}{2}\beta_i \right]^T X^T y < 0 \\ \frac{\tilde{a}}{\nu_2} & \left[2\beta - \frac{1}{2}\beta_i \right]^T X^T y \geq 0 \end{cases}.
\end{aligned}
$$

Finally, define functions g_{jl2} on $\mathbb{R}_+^2$ for $j \in \{1, \cdots, s\}$ and $l \in \{1, \cdots, t\}$ such that

$$
g_{jl2}(\lambda) = h_{j1}(\lambda_R) h_{l2}(\lambda_D)
$$

where

$$
h_{j1}(\lambda_R) = \begin{cases} \text{Gamma}\left(r_{j1} + \frac{N}{2}, r_{j2}; \lambda_R \right) & \text{if } \lambda_R < \lambda_{jR}^* \\ \text{Gamma}\left(r_{j1} + \frac{N}{2}, r_{j2} + \frac{\nu_1}{2}; \lambda_R \right) & \text{if } \lambda_R \geq \lambda_{jR}^* \end{cases}
$$

for

$$
\lambda_{jR}^* = \frac{2r_{j1} + N}{\nu_1} \log \left(1 + \frac{\nu_1}{2r_{j2}} \right)
$$

and

$$
h_{l2}(\lambda_D) = \begin{cases} \text{Gamma}\left(d_{l1} + \frac{k}{2}, d_{l2}; \lambda_D \right) & \text{if } \lambda_D < \lambda_{lD}^* \\ \text{Gamma}\left(d_{l1} + \frac{k}{2}, d_{l2} + \frac{\nu_1}{2}; \lambda_D \right) & \text{if } \lambda_D \geq \lambda_{lD}^* \end{cases}
$$

for

$$\lambda^*_{lD} = \frac{2d_{l1} + k}{\nu_1} \log\left(1 + \frac{\nu_1}{2d_{l2}}\right) .$$

In the following lemma, we construct functions g_1 and g_2 that satisfy (4.2) and (4.3) for the Bayesian hierarchical model.

Lemma 5.6. *Let Φ denote the 2-component Gibbs sampler for $\pi(\xi, \lambda|y)$. Define constants ν_1 and ν_2 where $\nu_1 > 0$, $\nu_2 > 1$, and $\nu_2 \log \nu_2 \geq c' \max\left\{\tilde{a}, \tilde{b}\right\}$. Also, let*

$$g_1(\xi) = \sum_{i=1}^{r} \eta_i g_{i1}(\xi) \quad \text{and} \quad g_2(\lambda) = \min_{j,l} g_{jl2}(\lambda) .$$

Then g_1 and g_2 satisfy (4.2) *and* (4.3)*:*

$$\inf_{(\lambda',\xi')\in D_2} \pi(\xi|\lambda', y) \geq g_1(\xi) \quad \text{and} \quad \inf_{(\lambda',\xi')\in D_1} \pi(\lambda|\xi', y) \geq g_2(\lambda)$$

for $D_1 = \{(\xi, \lambda) : \ f_1(\xi) \leq \nu_1\}$, $D_2 = \{(\xi, \lambda) : \ f_2(\lambda) \leq \nu_2\}$, and f_1 and f_2 as defined by either (5.13) *or* (5.14)*.*

Remark 5.3. We can apply Theorems 4.1, 4.2, and 4.3 to construct minorization conditions since Lemma 5.6 holds for any $\nu_1 > 0$, $\nu_2 > 1$, and $\nu_2 \log(\nu_2) \geq c' \max\{\tilde{a}, \tilde{b}\}$.

Geometric ergodicity of the 2-component Gibbs sampler under any of the three scanning strategies follows directly from Corollary 4.1 and Lemmas 5.5 and 5.6.

Theorem 5.1. *Suppose there exists Δ for which* (5.12) *holds and that at least one of the following conditions are satisfied:*

1. *$Z^T Z$ is nonsingular and for all $j \in \{1, \cdots, s\}$ and $l \in \{1, \cdots, t\}$, $d_{l1} > 1$ and*

$$r_{j1} > 0 \vee 0.5 \left[\sum_{i=1}^{N} z_i \left(Z^T Z\right)^{-1} z_i^T - N + 2\right] \quad \text{or}$$

2. $Z^T Z$ is possibly singular and for all $j \in \{1, \cdots, s\}$ and $l \in \{1, \cdots, t\}$,

$$d_{l1} > 0.5 \left[2 + \sum_{i=1}^{N} z_i z_i^T \right] ,$$

$$r_{j1} > 0 \vee 0.5 \left[\frac{1}{4} \sum_{i=1}^{N} x_i \left(X^T X \right)^{-1} x_i^T - N + 2 \right] .$$

Then DUGS, RPGS, and RSGS for $\pi(\xi, \lambda|y)$ are geometrically ergodic.

5.2.3 Regenerative Simulation for DUGS

Let $g : \mathbb{R}^{k+p} \times \mathbb{R}_+^2 \to \mathbb{R}$ and suppose our goal is to estimate

$$E_\pi g(\xi, \lambda) = \int g(\xi, \lambda) \pi(\xi, \lambda|y) \, d\xi \, d\lambda \, .$$

We will assume throughout that there exists $\rho > 0$ such that $E_\pi |g|^{2+\rho} < \infty$. In this case, we can use regenerative simulation (RS) to estimate $E_\pi g$ and obtain a valid Monte Carlo standard error. The details of this procedure are laid out in Chapter 3.2.2. In this section, we derive the tools required for implementing RS for $\pi(\xi, \lambda|y)$. For ease of exposition, we only consider RS for DUGS in the non-mixture setting where $\pi(\xi, \lambda|y)$ follows (5.7) with $r = s = t = 1$:

$$\begin{aligned} Y|\beta, u, \lambda_R, \lambda_D &\sim \mathrm{N}_N \left(X\beta + Zu, \lambda_R^{-1} I_N \right) \\ \beta|u, \lambda_R, \lambda_D &\sim \mathrm{N}_p \left(\beta_0, B^{-1} \right) \\ u|\lambda_R, \lambda_D &\sim \mathrm{N}_k \left(0, \lambda_D^{-1} I_k \right) \\ \lambda_R &\sim \mathrm{Gamma} \left(r_1, r_2 \right) \\ \lambda_D &\sim \mathrm{Gamma} \left(d_1, d_2 \right) \, . \end{aligned} \tag{5.17}$$

Results for the RPGS and RSGS should be similar.

Let k_D denote the DUGS transition density having support $\mathbb{R}^{k+p} \times \mathbb{R}^2_+$. Then RS requires the following general minorization condition: for some $s : \mathbb{R}^{k+p} \times \mathbb{R}^2_+ \to [0,1]$ for which $\int s(\xi,\lambda)\pi(\xi,\lambda|y)d\xi d\lambda > 0$ and some density $q(\cdot)$

$$k_D((\xi',\lambda'),\ (\xi,\lambda)) \geq s(\xi',\lambda')q(\xi,\lambda)\ . \tag{5.18}$$

When (5.18) holds, recall that the *split chain* $\{((\xi^{(n)},\lambda^{(n)}),\delta_n)\}_{n=0}^{\infty}$ may be obtained by the following recipe: Draw $\left(\xi^{(n+1)},\lambda^{(n+1)}\right) \mid \left(\xi^{(n)},\lambda^{(n)}\right)$ using transition density k_D, then draw a Bernoulli δ_n with success probability

$$\Pr\left(\delta_n \mid \xi^{(n)},\lambda^{(n)},\xi^{(n+1)},\lambda^{(n+1)}\right) = \frac{s(\xi^{(n)},\lambda^{(n)})q\left(\xi^{(n+1)},\lambda^{(n+1)}\right)}{k_D\left((\xi^{(n)},\lambda^{(n)}),\ (\xi^{(n+1)},\lambda^{(n+1)})\right)}\ .$$

The DUGS transition density (hence (5.18)) depends on the chosen update order; i.e., (ξ,λ) or (λ,ξ). We construct (5.18) for update order (ξ,λ) and recall a result from Hobert et al. (2006) for update order (λ,ξ). First, we define some notation. Let

$$\begin{aligned}
\Sigma_\beta(\lambda) &= \lambda_R X^T X + B \\
\mu_\beta(\lambda) &= \Sigma_\beta(\lambda)^{-1}(\lambda_R X^T y + B\beta_0) \\
\Sigma_u(\lambda) &= \lambda_R Z^T Z + \lambda_D I_k \\
\mu_u(\lambda) &= \lambda_R \Sigma_u(\lambda)^{-1} Z^T y
\end{aligned}$$

and recall from (5.8) that

$$\beta|\lambda,y \sim \mathrm{N}_p\left(\mu_\beta(\lambda),\Sigma_\beta(\lambda)^{-1}\right) \quad \text{and} \quad u|\lambda,y \sim \mathrm{N}_k\left(\mu_u(\lambda),\Sigma_u(\lambda)^{-1}\right)\ .$$

Also, define functions $g_1(\cdot)$ and $g_2(\cdot)$ as follows. First,

$$g_1(\lambda',\tilde{\lambda}) = g(\lambda',\tilde{\lambda})\exp\left\{-\frac{(\lambda'_R-\tilde{\lambda}_R)}{2}(y-X\breve{\beta})^T(y-X\breve{\beta})\right\} \tag{5.19}$$

for

$$g(\lambda', \tilde{\lambda}) = \frac{|\Sigma_\beta(\lambda')|^{1/2}}{|\Sigma_\beta(\tilde{\lambda})|^{1/2}} \cdot \frac{\exp\left\{-\frac{1}{2}\mu_\beta(\lambda')^T \Sigma_\beta(\lambda')\mu_\beta(\lambda')\right\}}{\exp\left\{-\frac{1}{2}\mu_\beta(\tilde{\lambda})^T \Sigma_\beta(\tilde{\lambda})\mu_\beta(\tilde{\lambda})\right\}} \exp\left\{\frac{\lambda'_R - \tilde{\lambda}_R}{2} y^T y\right\}$$

$$\breve{\beta} = \begin{cases} (X^T X)^{-1} X^T y & \text{if } \lambda'_R \leq \tilde{\lambda}_R \\ (X^T X)^{-1} X^T y + v_\beta & \text{if } \lambda'_R > \tilde{\lambda}_R \end{cases}.$$

Next,

$$\begin{aligned} g_2(\lambda', \tilde{\lambda}) = h(\lambda', \tilde{\lambda}) \exp\left\{-\frac{(\lambda'_D - \tilde{\lambda}_D)}{2} \sum_{i=1}^{k} \hat{u}_i^2\right\} \\ \cdot \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2}(y - Z\breve{u})^T (y - Z\breve{u})\right\} \end{aligned} \tag{5.20}$$

for

$$h(\lambda', \tilde{\lambda}) = \frac{|\Sigma_u(\lambda')|^{1/2}}{|\Sigma_u(\tilde{\lambda})|^{1/2}} \cdot \frac{\exp\left\{-\frac{1}{2}\mu_u(\lambda')^T \Sigma_u(\lambda')\mu_u(\lambda')\right\}}{\exp\left\{-\frac{1}{2}\mu_u(\tilde{\lambda})^T \Sigma_u(\tilde{\lambda})\mu_u(\tilde{\lambda})\right\}} \exp\left\{\frac{\lambda'_R - \tilde{\lambda}_R}{2} y^T y\right\}$$

$$\breve{u} = \begin{cases} (Z^T Z)^{-1} Z^T y & \text{if } \lambda'_R \leq \tilde{\lambda}_R \\ (Z^T Z)^{-1} Z^T y + v_u & \text{if } \lambda'_R > \tilde{\lambda}_R \end{cases}$$

and where $\hat{u}_i$ is defined as follows. Let $z(i)$ denote the ith element of vector z and $\text{abs}(\cdot)$ denote the absolute value operator. Then, if $\lambda'_D \leq \tilde{\lambda}_D$

$$\hat{u}_i = \begin{cases} 0 & u_1(i) \leq 0 \leq u_2(i) \\ \min\left\{\text{abs}\left(u_1(i)\right), \ \text{abs}\left(u_2(i)\right)\right\} & \text{otherwise} \end{cases}.$$

Otherwise, if $\lambda'_D > \tilde{\lambda}_D$

$$\hat{u}_i = \max\left\{\text{abs}\left(u_1(i)\right),\ \ \text{abs}\left(u_2(i)\right)\right\}\ .$$

Proposition 5.3. *Assume Z^TZ is nonsingular. Fix $\tilde{\lambda} \in \mathbb{R}^2_+$ and define sets $\mathbb{M}_\beta = \{\beta : \beta \in (X^TX)^{-1}X^Ty \pm v_\beta\}$ for some $v_\beta \in \mathbb{R}^p_+$ and $\mathbb{M}_u = \{u : u_1 \leq u \leq u_2\}$ where $u_1 = (Z^TZ)^{-1}Z^Ty - v_u$ and $u_2 = (Z^TZ)^{-1}Z^Ty + v_u$ for some $v_u \in \mathbb{R}^k_+$.*

Let $q(\xi, \lambda)$ be a density on $\mathbb{R}^{k+p} \times \mathbb{R}^2_+$ such that

$$q(\xi, \lambda) = c_q^{-1}\pi(\beta|\tilde{\lambda})\pi(u|\tilde{\lambda})\pi(\lambda|\xi)I(\beta \in \mathbb{M}_\beta)I(u \in \mathbb{M}_u)$$

where

$$c_q = \left[\int \pi(u|\tilde{\lambda})I(u \in \mathbb{M}_u)du\right]\left[\int \pi(\beta|\tilde{\lambda})I(\beta \in \mathbb{M}_\beta)d\beta\right]\ .$$

Also, let

$$s(\lambda', \tilde{\lambda}) = c_q\ g_1(\lambda', \tilde{\lambda})g_2(\lambda', \tilde{\lambda})$$

for g_1 and g_2 as defined by (5.19) *and* (5.20)*, respectively. Then the following minorization condition is satisfied for the Markov transition density of the block Gibbs sampler corresponding to update order (ξ, λ):*

$$k_2((\xi', \lambda'), (\xi, \lambda)) \geq s(\lambda', \tilde{\lambda})q(\xi, \lambda)$$

where $k_2((\xi', \lambda'), (\xi, \lambda)) = \pi(\xi|\lambda', y)\pi(\lambda|\xi, y)$.

For a proof of Proposition 5.3 see Appendix B. The following result was established in Hobert et al. (2006) but will be used in the sequel, hence we report it for completeness.

Proposition 5.4. *Fix $\tilde{\xi} \in \mathbb{R}^{k+p}$ and define sets $\mathbb{M}_D = [a_1, a_2]$ and $\mathbb{M}_R = [b_1, b_2]$. Let $q(\lambda, \xi)$ be a density on $\mathbb{R}_+^2 \times \mathbb{R}^{k+p}$ such that*

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

where

$$c_q = \left[\int \pi(\lambda_D|\tilde{\xi}) I(\lambda_D \in \mathbb{M}_D) d\lambda_D \right] \left[\int \pi(\lambda_R|\tilde{\xi}) I(\lambda_R \in \mathbb{M}_R) d\lambda_R \right] .$$

Also, define

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

where

$$\inf_{\lambda_D \in \mathbb{M}_D} \frac{\pi(\lambda_D|\xi')}{\pi(\lambda_D|\tilde{\xi})} = \left(\frac{d_2 + \frac{1}{2} v_2(\xi')}{d_2 + \frac{1}{2} v_2(\tilde{\xi})} \right)^{d_1 + k/2} \exp \left\{ -\frac{g(\xi', \tilde{\xi})}{2} \left(v_2(\xi') - v_2(\tilde{\xi}) \right) \right\}$$

for

$$g(\xi', \tilde{\xi}) = \begin{cases} a_1 & \text{if } v_2(\xi') - v_2(\tilde{\xi}) \leq 0 \\ a_2 & \text{if } v_2(\xi') - v_2(\tilde{\xi}) > 0 \end{cases}$$

and

$$\inf_{\lambda_R \in \mathbb{M}_R} \frac{\pi(\lambda_R|\xi')}{\pi(\lambda_R|\tilde{\xi})} = \left(\frac{r_2 + \frac{1}{2} v_1(\xi')}{r_2 + \frac{1}{2} v_1(\tilde{\xi})} \right)^{r_1 + N/2} \exp \left\{ -\frac{h(\xi', \tilde{\xi})}{2} \left(v_1(\xi') - v_1(\tilde{\xi}) \right) \right\}$$

for

$$h(\xi', \tilde{\xi}) = \begin{cases} b_1 & \text{if } v_1(\xi') - v_1(\tilde{\xi}) \leq 0 \\ b_2 & \text{if } v_1(\xi') - v_1(\tilde{\xi}) > 0 \end{cases} .$$

The following minorization condition is satisfied for the Markov transition density

of the block Gibbs sampler corresponding to update order (λ, ξ)*:*

$$k_1((\lambda', \xi'), (\lambda, \xi)) \geq s(\xi', \tilde{\xi}) q(\lambda, \xi)$$

where $k_1((\lambda', \xi'), (\lambda, \xi)) = \pi(\lambda|\xi', y)\pi(\xi|\lambda, y)$.

5.3 A Simulation Study: The Random Intercept Model

In this section we consider Gibbs sampling for the balanced random intercept model of Example 5.2 with k subjects and $m \geq 2$ observations on each subject. The frequentist version of this model writes the jth observation on subject i as

$$y_{ij} = x_{ij}\beta + u_i + \varepsilon_{ij}$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, m$ where x_{ij} is the $1 \times p$ row vector containing the jth set of observations on the ith subject, β is a $p \times 1$ vector of regression parameters, u_i is the random subject term, and ε_{ij} is the residual error. Let $Y = (y_1^T, \ldots, y_k^T)^T$ represent the overall $N \times 1$ response vector where $N = km$ and $y_i = (y_{i1}, \ldots, y_{im})^T$ is the response vector for subject i. Then the model can be written in matrix form as

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

where $u = (u_1, \ldots, u_k)^T$, ε is the vector of residual errors,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix} \quad \text{and} \quad Z = \begin{pmatrix} 1_m & & 0 \\ & \ddots & \\ 0 & & 1_m \end{pmatrix} = I_k \otimes 1_m$$

where X_i is the $m \times p$ design matrix for the ith subject.

We consider the non-mixture Bayesian hierarchical version of this model given by (5.17). Further, we focus on the special case with $m = 5$ and $p = 1$. In this setting, $Z^T Z = mI_k = 5I_k$ and is therefore nonsingular. Theorem 5.1 in conjunction with Proposition 5.2 guarantees the DUGS, RPGS, and RSGS for $\pi(\xi, \lambda|y)$ are geometrically ergodic if the hyperparameter setting satisfies $\beta_0 = 0$, $d_1 > 1$, and $r_1 > 0$. There are no extra restrictions on r_1 since Theorem 5.1 only requires

$$r_1 > 0 \vee 0.5 \left(\sum_{i=1}^{N} z_i (Z^T Z)^{-1} z_i^T - N + 2 \right)$$

where $m = 5$ guarantees

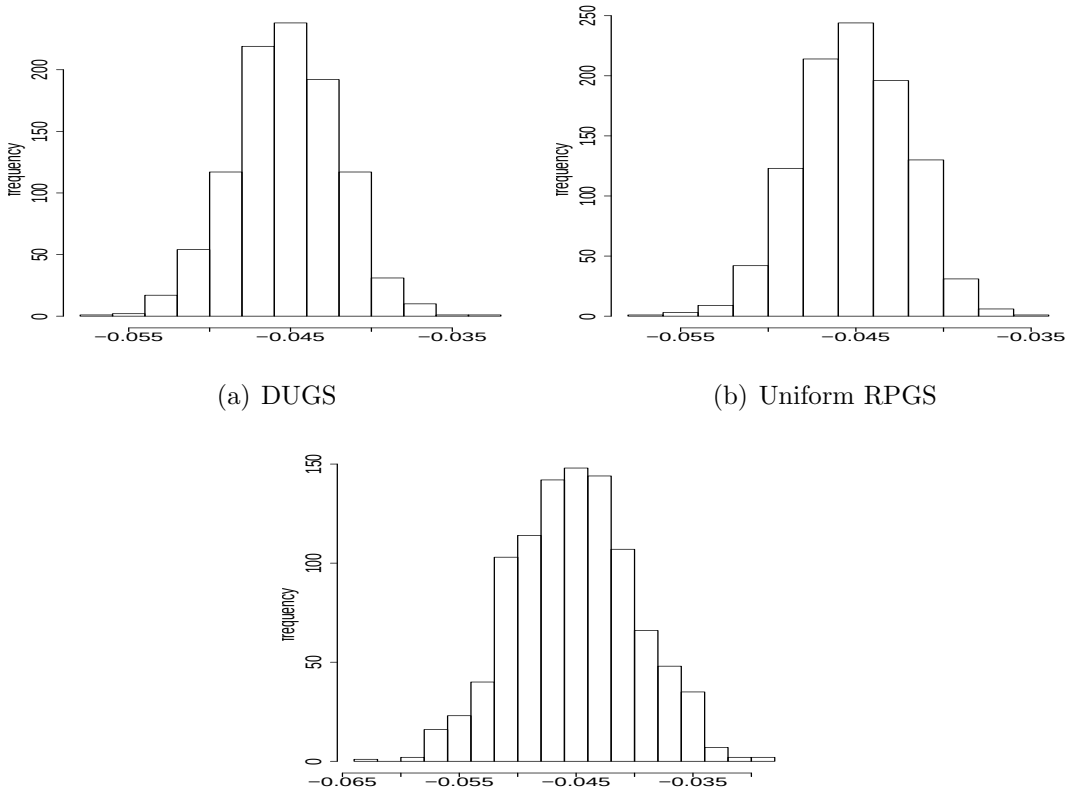
$$\sum_{i=1}^{N} z_i \left(Z^T Z \right)^{-1} z_i - N + 2 = k - 5k + 2 \leq 0 \, .$$

Unfortunately, we have not been able to eliminate the restriction on d_1. This does *not* imply that a Gibbs sampler with $d_1 < 1$ is sub-geometric. It only implies that we cannot use our results to *guarantee* geometric ergodicity. In turn, we cannot guarantee that Markov chain CLTs exist. However, a short simulation suggests CLTs hold even when $d_1 < 1$. First, we generated data for $k = 2$ subjects using the procedure outlined in the next section. Then setting $\beta_0 = 0$, $B^{-1} = 0.1$, and $r_1 = r_2 = d_1 = d_2 = 0.5$, we ran 1000 independent DUGS for 10000 iterations each. We repeated this for the

uniform RPGS and uniform RSGS. In each repetition we calculated the average

$$\overline{\beta} = \frac{1}{10000} \sum_{i=1}^{10000} \beta^{(i)} \ .$$

Histograms of the 1000 independent ergodic averages in each sampling setting are included in Figure 5.3. These suggest CLTs exist even when $d_1 < 1$.



(a) DUGS

(b) Uniform RPGS

(c) Uniform RSGS

Figure 5.3: Histograms of the Gibbs sampler ergodic averages, $\overline{\beta}$

This argument aside, we now restrict our attention to Gibbs sampling with $d_1 > 1$. Though this guarantees geometric ergodicity, there are many open questions regarding the behavior of the Gibbs sampler. For instance, how does scanning strategy affect the efficiency of Markov chain estimation? What effect do model dimension

and hyperparameter setting have on the behavior of Gibbs samplers? To explore these issues, we conducted a series of simulation studies. We begin by outlining our procedure for generating the data required by each of these simulations.

5.3.1 Simulated Data

Throughout we simulate data (ie. values of Y) from (5.17) using the following settings. Our balanced random intercept model has a single regression parameter β ($p = 1$) with k subjects and $m = 5$ observations per subject. In this case, $X = (x_1^T \cdots x_k^T)^T$ where x_i is the design vector for the ith subject. Also, for all i, $x_i = (-0.50, -0.25, 0, 0.25, 0.50)$ so that $x_i^T 1_5 = 0$. Therefore, $X^T Z = 0$ which guarantees conditional independence of β and u given λ. Finally, we set $\beta_0 = 0$, $B^{-1} = 0.1$, and $r_1 = r_2 = d_1 = d_2 = 2$ and simulate *one* set of data for each $k \in \{2, 5, 10, 20, 25, 35, 100\}$.

5.3.2 An MSE Comparison of DUGS, RPGS, and RSGS

Suppose we are interested in estimating the posterior mean $\mathrm{E}(\beta|y)$. To compare the quality of estimates between and within the DUGS, RPGS, and RSGS, we perform an analysis of the corresponding MSE's. To this end, we follow a procedure similar to that outlined in Chapter 5.1.2 for the Normal-Normal model. First, for each value of $k \in \{2, 10\}$ and the associated simulated data, we simulated the Gibbs sampler under a variety of scanning strategies. Assuming we do not know the true nature of the data, in each case we set $\beta_0 = 0$, $B^{-1} = 0.1$, and $r_1 = r_2 = d_1 = d_2 = 3$ and started the chain from $(\xi_0, \lambda_0) = (0_{k+1}, 1)$. Further, for each combination of k and scanning strategy we ran 1000 independent Gibbs samplers for 10000 iterations each.

In each setting this produced 1000 independent estimates of $\mathrm{E}(\beta|y)$,

$$\overline{\beta} = \frac{1}{10000} \sum_{i=1}^{10000} \beta^{(i)} \ .$$

Estimating $\mathrm{MSE}(\overline{\beta})$ requires the true value of $\mathrm{E}[\beta|y]$, but this is unknown. Our solution is to estimate this quantity with an independent run of the DUGS. For $k = 2$ and $k = 10$ we used runs of length $n = 10^6$ and $n = 5 * 10^6$, respectively. This produced estimates $\tilde{\beta}$ which we treat as the "true" value of $\mathrm{E}(\beta|y)$. Then in each Gibbs sampling setting, we estimated

$$\mathrm{MSE}(\overline{\beta}) = \mathrm{E}\left(\overline{\beta} - \mathrm{E}(\beta|y)\right)^2,$$

using

$$\widehat{\mathrm{MSE}}(\overline{\beta}) = \frac{1}{1000} \sum_{i=1}^{1000} \left(\overline{\beta}^{(i)} - \tilde{\beta}\right)^2$$

where $\overline{\beta}^{(i)}$ denotes the ith Monte Carlo estimate.

We begin by comparing the quality of $\overline{\beta}$ among the DUGS, uniform RPGS, and uniform RSGS. Estimates of the MSE ratios relative to the DUGS sampler,

$$\frac{\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{RPGS}})}{\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{DUGS}})} \quad \text{and} \quad \frac{\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{RSGS}})}{\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{DUGS}})} \ ,$$

are reported in Table 5.5. First, with MSE ratios close to one, there does not appear to be an obvious choice between the DUGS and uniform RPGS strategies. On the other hand, DUGS is more efficient than RSGS for both $k = 2$ and $k = 10$.

As in Chapter 5.1.2, we also want to compare DUGS to RSGS with twice as many iterations. To this end, for each $k \in \{2, 10\}$ we ran 1000 independent uniform RSGS for 20000 iterations each. The corresponding estimates $\overline{\beta}$ were used to obtain

Table 5.5: MSE ratios relative to DUGS (and standard errors) for the Markov chain estimates of $\mathrm{E}(\beta|y)$.

	$k=2$	$k=10$
Uniform RPGS	1.110 (0.066)	0.990 (0.063)
Uniform RSGS	3.240 (0.202)	3.093 (0.190)

$\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{RSGS}}, 20000)$ and MSE ratios relative to DUGS were estimated by

$$\frac{\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{RSGS}}, 20000)}{\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{DUGS}}, 10000)} \,. \tag{5.21}$$

For $k = 2$, the MSE ratio estimate was 1.488 with a standard error of 0.094. For $k = 10$, the estimate was 1.653 with a standard error of 0.106. Therefore, the DUGS (and RPGS) with half as many iterations is still more efficient than RSGS in terms of MSE.

Next, consider the impact of RPGS permutation probabilities q_i and RSGS selection probabilities p_i on the estimation of $\mathrm{E}(\beta|y)$. Let $\{q_1, q_2\}$ be the RPGS permutation probabilities corresponding to update orders $\{(\xi, \lambda), (\lambda, \xi)\}$, respectively, and $\{p_1, p_2\}$ be the RSGS selection probabilities corresponding to components $\{\xi, \lambda\}$, respectively. For every combination of $k \in \{2, 10\}$ and $q_1 \in \{0.01, 0.25, 0.50, 0.75, 0.99\}$ we ran 1000 independent RPGS for 10000 iterations each. Similarly, for every combination of $k \in \{2, 10\}$ and $p_1 \in \{0.01, 0.25, 0.50, 0.75, 0.99\}$ we ran 1000 independent RSGS for 10000 iterations each. The resulting estimates of the MSE ratios relative to the uniform settings,

$$\frac{\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{RPGS}})}{\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{RPGS},q_1=0.50})} \quad \text{and} \quad \frac{\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{RSGS}})}{\widehat{\mathrm{MSE}}(\overline{\beta}_{\mathrm{RSGS},p_1=0.50})} \,,$$

are reported in Table 5.6.

It does not appear that the choice of q_1 has a significant impact on the quality of

RPGS MSE's. The opposite is true for the choice of p_1. Specifically, RSGS MSE's appear to decrease as p_1 increases. Exploring the full conditional distributions $\xi|\lambda, y$ and $\lambda|\xi, y$ provides insight into this observation. First,

$$\mathrm{Var}(\xi \mid \lambda, y) = \begin{pmatrix} (5\lambda_R + \lambda_D)I_k & 0 \\ 0 & \lambda_R X^T X + 10 \end{pmatrix}^{-1}$$

whereas

$$\mathrm{Var}(\lambda_R \mid \xi, y) = \frac{12 + 10k}{(6 + v_1(\xi))^2} \quad \text{and} \quad \mathrm{Var}(\lambda_D \mid \xi, y) = \frac{12 + 2k}{(6 + v_2(\xi))^2}\,.$$

Therefore, the variability in $\xi|\lambda, y$ can be huge when λ is small. On the other hand, for any $\xi \in \mathbb{R}^{k+1}$, $\mathrm{Var}(\lambda_R|\xi, y) \leq (6+5k)/18$ and $\mathrm{Var}(\lambda_D|\xi, y) \leq (6+k)/18$. By choosing large p_1 we are forcing the Markov chain to visit the more variable component, ξ, with higher frequency. Hence Table 5.6 supports our intuition and the literature which suggests that the accuracy of Markov chain estimates may be improved by visiting more variable components with higher frequency (Levine and Casella, 2006).

Table 5.6: MSE ratios relative to the uniform settings ($q_1 = 0.50$ for RPGS and $p_1 = 0.50$ for RSGS) for the estimation of $\mathrm{E}(\beta|y)$. Standard errors are given in parentheses.

	k	$q_1 = p_1 = 0.01$	$q_1 = p_1 = 0.25$	$q_1 = p_1 = 0.75$	$q_1 = p_1 = 0.99$
RPGS	2	0.933 (0.058)	1.044 (0.065)	0.991 (0.063)	0.990 (0.060)
	10	1.004 (0.063)	1.003 (0.063)	1.055 (0.066)	1.002 (0.063)
RSGS	2	65.801 (4.080)	2.433 (0.152)	0.535 (0.032)	0.715 (0.044)
	10	69.471 (4.493)	2.678 (0.169)	0.555 (0.035)	0.342 (0.021)

5.3.3 Regenerative Simulation

In our experience and in the published literature (see e.g. Gilks et al. (1998)) RS is often seen as being particularly difficult to implement as the dimension of the Markov

chain increases. This is apparently related to our discussion of Example 5.2 where we found that the drift rate approaches 1 as k increases. We also found that drift rate is affected by the hyperparameter setting. In this section, we address the effect of model dimension and hyperparameter setting on the performance of RS. In particular, we consider DUGS for the random intercept model and evaluate RS with respect to two measures: (1) average time between regenerations (ie. average tour length); and (2) empirical performance of the resulting confidence intervals defined in (3.10).

Implementing the Split Chain

Exploring the regenerative behavior of DUGS requires simulation of the appropriate split chains. The tools required for this simulation are given in Chapter 5.2.3. Let Φ_1' denote the split chain on $\mathbb{R}_+^2 \times \mathbb{R}^{k+1} \times \{0,1\}$ corresponding to update order (λ, ξ) and let Φ_2' denote the split chain on $\mathbb{R}^{k+1} \times \mathbb{R}_+^2 \times \{0,1\}$ corresponding to update order (ξ, λ). The split chains are simulated with respect to the general minorization conditions given in Propositions 5.3 and 5.4. This requires the definition of sets $\mathbb{M}_D$, $\mathbb{M}_R$, $\mathbb{M}_\beta$, and $\mathbb{M}_u$ and fixed points $\tilde{\lambda}$ and $\tilde{\xi}$. For each value of k and hyperparameter setting of interest, we ran the DUGS corresponding to update order (ξ, λ) for 1×10^4 iterations starting from $\lambda_D = \lambda_R = 1$. Letting $\tilde{\lambda}_D$, $\tilde{\lambda}_R$, $\tilde{u}$, and $\tilde{\beta}$ denote the resulting estimates of the posterior expectations, we set $\tilde{\lambda} = (\tilde{\lambda}_D, \tilde{\lambda}_R)$ and $\tilde{\xi} = (\tilde{u}, \tilde{\beta})$. We also define $\mathbb{M}_D = \tilde{\lambda}_D \pm w s_{\lambda_D}$ and $\mathbb{M}_R = \tilde{\lambda}_R \pm w s_{\lambda_R}$ where s_{λ_D}, s_{λ_R} denote the sample standard deviations of the Markov chain samples for λ_D and λ_R, respectively, and $w > 0$. Finally, set $\mathbb{M}_\beta = (X^T X)^{-1} X^T y \pm w(\tilde{\lambda}_R X^T X)^{-1}$ and $\mathbb{M}_u = (Z^T Z)^{-1} Z^T y \pm w(\tilde{\lambda}_D X^T X)^{-1}$. Notice that as w increases, Φ_1' visits the set $\mathbb{M}_D \times \mathbb{M}_R \times \mathbb{R}^{k+1} \times \{0,1\}$ and Φ_2' visits the set $\mathbb{M}_u \times \mathbb{M}_\beta \times \mathbb{R}_+^2 \times \{0,1\}$ with increased frequency. However, the probability of regeneration decreases. Throughout this section, a simple grid search is used to choose w that minimizes the average tour lengths. These are reported in Table 5.8.

Finally, split chain simulation requires $(\lambda_0, \xi_0) \sim q(\cdot)$ where q is defined by Propo-

sitions 5.4 and 5.3 for split chains Φ_1' and Φ_2', respectively. Sampling from $q(\cdot)$ was achieved by applying an accept-reject algorithm using the appropriate full transition density as the candidate distribution. For instance, for split chain Φ_1', candidate values (λ, ξ) were drawn from $\pi(\lambda|\tilde{\xi})\pi(\xi|\lambda)$ until $\lambda_D \in \mathbb{M}_D$ and $\lambda_R \in \mathbb{M}_R$, in which case we set $(\lambda_0, \xi_0) = (\lambda, \xi)$.

Regeneration Rates

Here we address the practicality of RS, ie. the frequency with which regenerations occur, as a function of k (the dimension of u) and the hyperparameter setting. We begin by studying the effect k has on regeneration rate. Set $\beta_0 = 0$, $B^{-1} = 0.1$ and $r_1 = r_2 = d_1 = d_2 = 3$. Under this hyperparameter setting and for both DUGS update orders, we simulated 5×10^3 regenerations for each $k \in \{2, 5, 10, 20, 25, 35\}$. The results are summarized in Figure 5.4 and Table 5.8. For both update orders, the average tour length increases with k. In addition, the average tour lengths for update order (ξ, λ) are significantly larger than for order (λ, ξ). In the former case, regeneration can only occur when the $\xi = (u, \beta)$ component visits set $\mathbb{M}_u \times \mathbb{M}_\beta \subset \mathbb{R}^{k+1}$. On the other hand, regeneration in the latter case only requires the $\lambda = (\lambda_D, \lambda_R)$ component to visit set $\mathbb{M}_D \times \mathbb{M}_R \subset \mathbb{R}^2_+$.

Remark 5.4. We also considered this simulation using $k = 100$. However, we could not find a value of w that would produce a regeneration in fewer than 10^5 iterations. We had similar difficulties when $k = 50$. Thus the range of applicability of RS based on our minorization condition is limited to the situation where k is not too large.

We now turn our attention to the effect of hyperparameter setting on the regeneration rate. Recall that $\lambda_D \sim \text{Gamma}(d_1, d_2)$ and $\lambda_R \sim \text{Gamma}(r_1, r_2)$. We consider hyperparameter settings that satisfy $d_1 = d_2$ and $r_1 = r_2$. In this case, the prior means equal one and the prior variances are $1/d_2$ and $1/r_2$. Therefore, the priors for

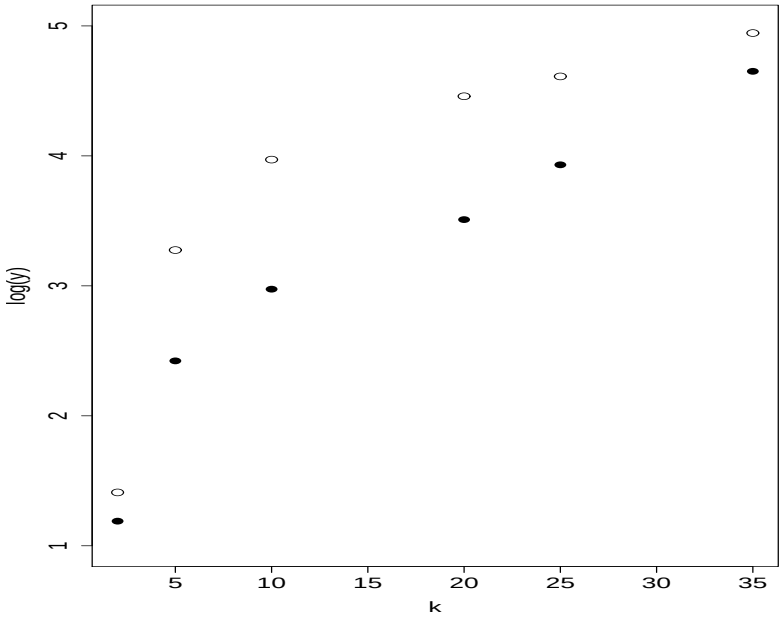


Figure 5.4: The log transformation of the average RS tour lengths, $\log(y)$, is plotted against k for DUGS with order (λ, ξ) (solid dots) and order (ξ, λ) (open circles).

λ_D and λ_R become more "vague" as d_2 and r_2 decrease, respectively.

Set $k = 2$ and $m = 5$. Under a variety of hyperparameter settings satisfying $d_1 = d_2$ and $r_1 = r_2$ and for both update orders, we simulated 5×10^3 regenerations for the DUGS. The results are displayed in Figure 5.5. First, the average tour length decreases slightly as $d_1 = d_2 \searrow 1$ (left plot). There is also a noticeable spike in the average tour lengths when $d_1 = d_2 < 1$. This suggests RS may not be practical when λ_D has a large prior variance. (Recall that our results do not guarantee geometric ergodicity when $d_1 < 1$.) The impact of a large prior variance for λ_R is not as dramatic (right plot). However, the average tour length increases as $r_1 = r_2$ decreases. Overall, our results are consistent with empirical results in Jones and Hobert (2004) and Natarajan and McCulloch (1998) regarding the convergence of block Gibbs samplers in some hierarchical model settings.

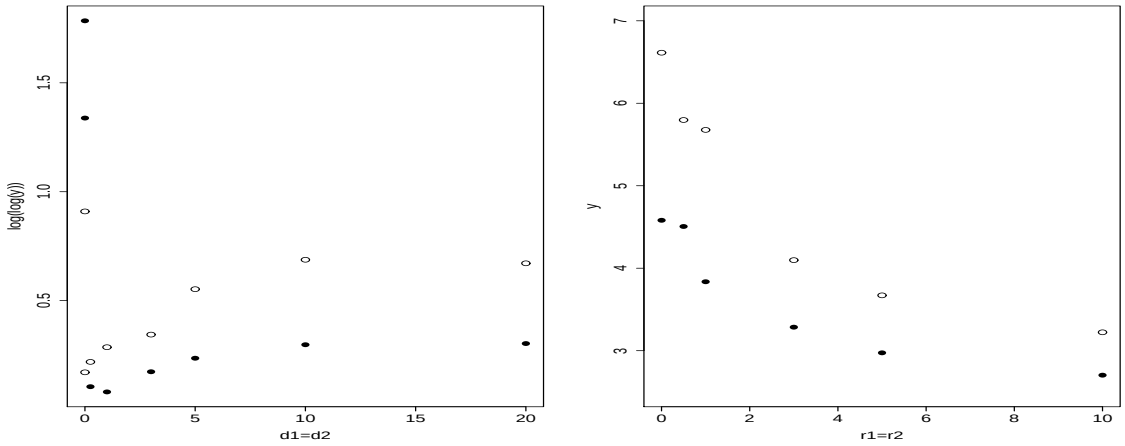


Figure 5.5: Let y denote the average tour length of the regenerations. The left plot graphs $\log(\log(y))$ against $d_1 = d_2$ while fixing $r_1 = r_2 = 3$. Similarly, the right plot graphs y against $r_1 = r_2$ while fixing $d_1 = d_2 = 3$. In both plots, solid dots represent DUGS with order (λ, ξ) and open circles represent DUGS with order (ξ, λ).

Performance of RS

Recall from (3.10) that an asymptotically valid $100(1-\alpha)\%$ confidence interval for $E(\beta|y)$ is given by

$$\bar{\beta}_{\tau_R} \pm t_{\alpha/2,R-1} \frac{\hat{\nu}}{\sqrt{R}} \tag{5.22}$$

where τ_R is the total simulation length for R regenerations and $\hat{\nu}$ is given by (3.9). In light of our previous discussions, we now investigate the impact of increasing k (ie. the dimension of u) on the performance of these confidence intervals. In particular, we will evaluate the empirical coverage probabilities and simulation effort.

Consider estimating $E[\beta|y]$ using DUGS with update order (λ, ξ) and the following hyperparameter setting: $\beta_0 = 0$, $B^{-1} = 0.1$ and $r_1 = r_2 = d_1 = d_2 = 3$. For every combination of $k \in \{2, 10, 25\}$ and $R \in \{10, 25, 50, 100\}$, we simulated 500 independent DUGS for R regenerations each. Then, for each individual chain, we calculated 95% confidence intervals for $E[\beta|y]$ using (5.22). To estimate the true coverage probability of these intervals we need $E[\beta|y]$, but this is unknown. Our

solution is to estimate this quantity with an independent run of the Gibbs sampler. For $k = 2$ and $k = 10$ we used runs of 6×10^6 regenerations ($\approx 2 \times 10^7$ and $\approx 1.1 \times 10^8$ iterations, respectively). For $k = 25$ this was impractical so we settled for 6×10^4 regenerations ($\approx 3.2 \times 10^6$ iterations). The estimated coverage probabilities based on these estimates of $\mathrm{E}[\beta|y]$ are reported in Table 5.7 along with associated 95% confidence intervals. Also reported are the average simulation lengths for each combination of k and R.

Table 5.7: Estimated coverage probabilities with associated 95% confidence intervals. Also reported are the average simulation lengths, $\overline{n}$, of the 500 independent samplers in each setting.

		$R = 10$	$R = 25$	$R = 50$	$R = 100$
$k = 2$	estimate	0.888	0.934	0.954	0.946
	95% CI	(0.860,0.916)	(0.912,0.956)	(0.936,0.972)	(0.926,0.966)
	$\overline{n}$	32.95	82.02	164.60	329.96
$k = 10$	estimate	0.882	0.912	0.948	0.932
	95% CI	(0.854,0.910)	(0.887,0.937)	(0.929,0.967)	(0.910,0.954)
	$\overline{n}$	201.86	498.65	1002.17	2002.40
$k = 25$	estimate	0.854	0.910	0.934	0.952
	95% CI	(0.823,0.885)	(0.885,0.935)	(0.912,0.956)	(0.933,0.971)
	$\overline{n}$	525.39	1320.81	2623.83	5249.55

Clearly, using too few regenerations results in significant undercoverage. However, using enormous numbers of regenerations does not appear to have any obvious benefits. On the other hand, larger values of k require additional simulation time (though a similar number of regenerations) to attain the same level of coverage as smaller k values. Nonetheless, even for moderate values of k, RS performs well with a relatively small amount of simulation effort.

Table 5.8: Summary statistics of the regenerative simulation tour lengths (N): average tour length ($\text{avg}(N)$), standard deviation of the sample ($\text{sd}(N)$), maximum observed tour length ($\max(N)$). Also reported is the value of w used for the regenerative simulation

		Update Order (λ, ξ)	Update Order (ξ, λ)
$k = 2$	$\text{avg}(N)$	3.3	4.1
	$\text{sd}(N)$	2.8	3.7
	$\max(N)$	24	27
	w	1.5	1.5
$k = 5$	$\text{avg}(N)$	11.3	26.4
	$\text{sd}(N)$	11.2	26.0
	$\max(N)$	81	244
	w	0.75	6
$k = 10$	$\text{avg}(N)$	19.6	53.1
	$\text{sd}(N)$	20.4	52.4
	$\max(N)$	199	466
	w	0.8	375
$k = 20$	$\text{avg}(N)$	33.4	86.4
	$\text{sd}(N)$	34.3	86.0
	$\max(N)$	320	1020
	w	0.75	500
$k = 25$	$\text{avg}(N)$	51.0	100.6
	$\text{sd}(N)$	54.9	101.1
	$\max(N)$	484	865
	w	1.5	1100
$k = 35$	$\text{avg}(N)$	104.6	140.5
	$\text{sd}(N)$	112.4	142.4
	$\max(N)$	1033	1380
	w	2	500

5.4 A Numerical Example: The HMO Data

To study the cost-effectiveness of transferring military retirees from a Defense Department health plan to health plans for government employees, information was gathered from 341 state-based health maintenance organizations (HMOs). These plans represent 42 states, the District of Columbia, Puerto Rico, and Guam. An HMO plan's cost is measured by its monthly premium for individual subscribers. Two possible factors in this cost are (1) the typical hospital expenses in the state in which the HMO operates; and (2) the region in which the HMO operates. In Figure 5.6, the individual monthly premiums for the 341 HMOs is plotted against the average expenses per admission in the state of operation (both in US dollars).

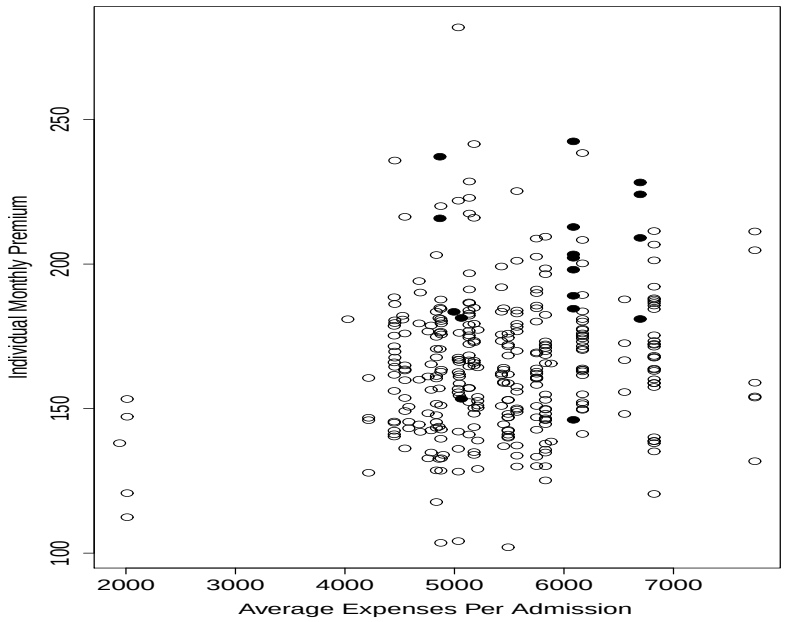


Figure 5.6: Individual monthly HMO premiums are plotted against the average expenses per admission in the state in which the HMO operates. Solid circles represent states in New England.

Let y_i denote the individual monthly premium of the ith HMO plan. To analyze

these data, Hodges (1998) considered a Bayesian version of the following frequentist model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \tag{5.23}$$

where the ε_i are iid $\mathrm{N}\left(0, \lambda_R^{-1}\right)$, x_{i1} denotes the centered and scaled average expenses per admission in the state in which the ith HMO operates, and x_{i2} is an indicator for New England. The x_{i1} values were centered and scaled to avoid collinearity. Specifically, if $\tilde{x}_{i1}$ is the raw average expense per admission and $\overline{x}_1$ is the overall average expense per admission, $x_{i1} = (\tilde{x}_{i1} - \overline{x}_1)/1000$.

We perform a Bayesian regression analysis based on the following hierarchical version of (5.23):

$$\begin{aligned} y|\beta, \lambda_R &\sim \mathrm{N}_N\left(X\beta, \lambda_R^{-1} I_N\right) \\ \beta|\lambda_R &\sim \mathrm{N}_3\left(\tilde{\beta}, B^{-1}\right) \\ \lambda_R &\sim \mathrm{Gamma}(r_1, r_2) \end{aligned} \tag{5.24}$$

where $N = 341$, y is the $N \times 1$ vector of individual premiums, $\beta = (\beta_0, \beta_1, \beta_2)$ is the vector of regression parameters, and X is the $N \times 3$ data matrix

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} \end{pmatrix}.$$

With no state random effect, this model follows (5.17) by setting $Z = 0$. (For an alternative treatment of the data, see Hodges (1998) who included a random effect for the state of operation and did not place a prior on the β_i.)

Complete specification of the model requires values for hyperparameters $(\tilde{\beta}, B, r_1, r_2)$.

Though there are many hyperparameter settings we might consider, we take an empirical Bayes point of view and choose one that is reflective of the data. To this end, we fit (5.23) using least squares regression. The results are summarized in Table 5.9.

Table 5.9: Least squares regression results for (5.23).

Parameter	Estimate	Standard Error
β_0	164.989	1.322
β_1	3.910	1.508
β_2	32.799	5.961
$N = 341$		
degrees of freedom = 338		
MSE = SSE/338 = $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2/338 = 23.79^2$		

Accordingly, we chose the following prior mean and covariance matrix for β:

$$\tilde{\beta} = \begin{pmatrix} 164.989 \\ 3.910 \\ 32.799 \end{pmatrix} \quad \text{and} \quad B^{-1} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 36 \end{pmatrix}$$

where $\tilde{\beta}$ is the vector of least squares estimates and the diagonal elements of B^{-1} are reflective of the corresponding squared standard errors in Table 5.9. Next, we set the prior mean and variance for λ_R to

$$\begin{aligned} \mathrm{E}(\lambda_R) &= \frac{r_1}{r_2} = \frac{1}{\mathrm{MSE}} = 0.00177; \quad \text{and} \\ \mathrm{Var}(\lambda_R) &= \frac{r_1}{r_2^2} = 1 \end{aligned}$$

where MSE is the least squares estimate of λ_R^{-1} given in Table 5.9. Solving for r_1 and r_2 gives

$$r_1 = 3.122 * 10^{-6} \quad \text{and} \quad r_2 = 0.00177\,.$$

To perform the Bayesian regression analysis, consider Gibbs sampling for $\pi(\beta, \lambda_R|y)$. Since (5.24) does not contain any random effects, it follows from Theorem 5.1 that the Gibbs sampler (DUGS, RPGS, or RSGS) will be geometrically ergodic if

$$r_1 > 0 \vee 0.5\,[2 - N] = 0$$

and if there exists Δ for which (5.12) holds. The first assumption is clearly satisfied. The second assumption requires Δ such that for any $\lambda_R \in \mathbb{R}_+$

$$\sum_{i=1}^{N} [\mathrm{E}(y_i - x_i\beta|\lambda_R, y)]^2 = (y - X\mathrm{E}(\beta|\lambda_R, y))^T (y - X\mathrm{E}(\beta|\lambda_R, y)) \leq \Delta^2$$

where $\mathrm{E}(\beta|\lambda_R, y) = (\lambda_R X^T X + B)^{-1}(\lambda_R X^T y + B\tilde{\beta})$. This can be established using numerical techniques. Specifically, using the R "optimize" function which optimizes a function using *golden section search* and *successive parabolic interpolation*, it can be shown that this holds for $\Delta^2 = 191241$. Therefore, the Gibbs sampler is geometrically ergodic. Most importantly, this guarantees the existence of central limit theorems and consistent standard errors for the Markov chain ergodic averages (under moment conditions). Throughout this section, the standard errors are computed using the consistent batch means technique. See Chapter 3.2 for details.

We first consider the Bayesian regression analysis using DUGS. To this end, we ran independent DUGS for $\pi(\beta, \lambda_R|y)$ from a variety of starting values $\left(\beta^{(0)}, \lambda_R^{(0)}\right)$ and using update order (λ_R, β). For each chain, simulation continued until the Monte Carlo standard errors for the estimates of the posterior means of β_0, β_1, β_2, and λ_R were below 0.05, 0.01, 0.05, and 0.0001, respectively. The results were consistent across starting values. That is, DUGS with different starting values produced similar estimates and required similar simulation effort to meet the above specifications. Here, we present the results for DUGS started from the prior means $(\beta^{(0)}, \lambda_R^{(0)}) = (\tilde{\beta}, r_1/r_2)$.

Under this setting, the Monte Carlo standard error thresholds were met after 15000 iterations. The corresponding estimates of the posterior means are reported in Table 5.10 with corresponding standard errors. Also included are 95% credible intervals computed from the Monte Carlo sample 2.5th and 97.5th quantiles.

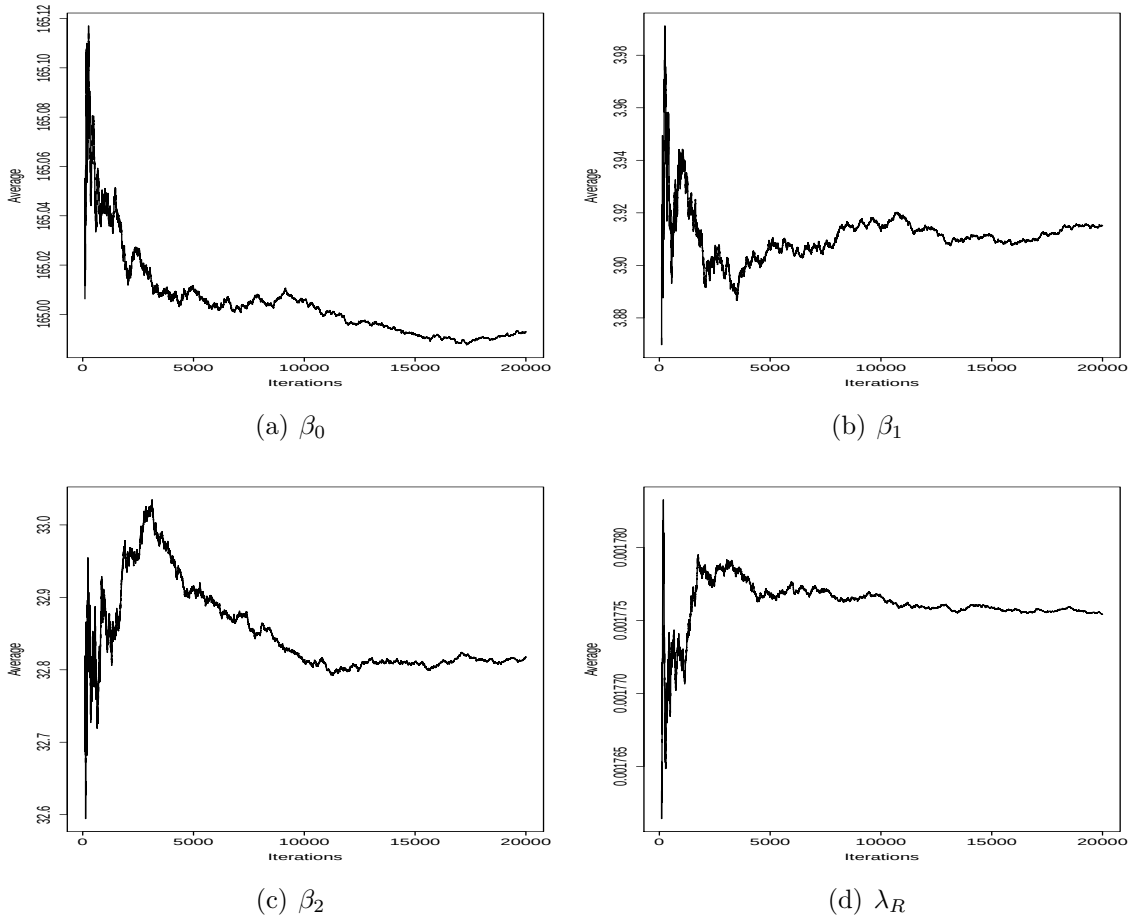
Table 5.10: DUGS estimates of posterior means with corresponding standard errors.

Parameter	Estimate	Standard Error	95% Credible Interval
β_0	164.992	0.0080	(163.1,166.9)
β_1	3.912	0.0099	(1.72,6.13)
β_2	32.810	0.0329	(25.5,41.0)
λ_R	0.00178	$10 * 10^{-7}$	(0.0015,0.0020)

In Figure 5.7, the running means from this DUGS run with 5000 added iterations are plotted versus iteration number (starting from the one hundredth iteration). The estimates for each parameter appear to *sufficiently* stabilize by the 15000th iteration, thus providing further evidence that 15000 iterations is "long enough".

To bolster confidence in our analysis, we also investigated the raw DUGS output. For each parameter we produced a time series plot of the Markov chain values as well as a corresponding autocorrelation plot. The plots for β_1 (the primary parameter of interest) are given in Figure 5.8. The time series plot of the β_1 output (left) indicates this sub-chain is mixing quickly. This is supported by the corresponding autocorrelation plot (right) where estimates of the autocorrelation function (ACF), $\gamma(r) = \text{Cor}\left(\beta_1^{(i)}, \beta_1^{(i+r)}\right)$, are plotted versus lag r for $r \in \{0, 1, \cdots, 20\}$. The time series and autocorrelation plots for the other parameters are similar.

Similar analyses can be performed using RPGS or RSGS with comparable results. However, the simulation effort required to attain the same level of accuracy in the estimation procedure differs. For simplicity, consider estimating the posterior mean of β_1. After 15000 iterations, DUGS produced an estimate of this quantity ($\overline{\beta}_1$) with a Monte Carlo standard error less than 0.01. Using this as a target accuracy level,



(a) β_0 (b) β_1

(c) β_2 (d) λ_R

Figure 5.7: Running mean plots for the DUGS estimates of the posterior expectations of β_0, β_1, β_2, and λ_R.

we ran RPGS and RSGS under a variety of specifications for the permutation and selection probabilities, respectively. In each case, the Markov chain was started from $(\tilde{\beta}, r_1/r_2)$ and stopped when the Monte Carlo standard error for $\overline{\beta}_1$ was below 0.01. The results are summarized in Table 5.11.

In this table, q_1 is the probability of choosing update order (β, λ_R) in any RPGS iteration and p_1 is the probability of selecting β for update in any RSGS iteration (while fixing λ_R). The estimates $\overline{\beta}_1$ are consistent across scanning strategies. Further, the required RPGS simulation effort is similar to that for DUGS. This is also true

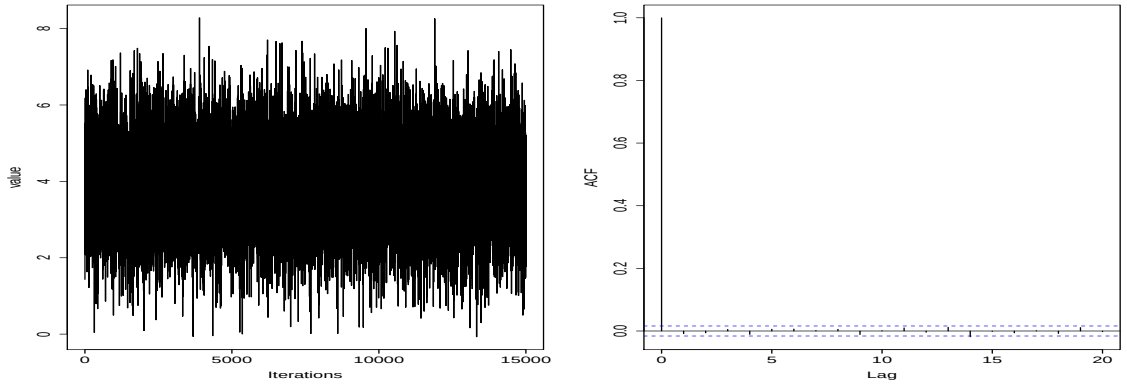


Figure 5.8: A time series plot (left) and autocorrelation plot (right) for the DUGS β_1 iterations.

for the RSGS with $p_1 = 0.99$. However, the smaller p_1 is, the longer the RSGS chain must run in order to attain the prespecified level of accuracy. This coincides with our observations for the Bayesian random intercept model in Chapter 5.3 where the MSE of the RSGS ergodic average increased as p_1 decreased.

Table 5.11: Estimates of the posterior mean of β_1 using DUGS, RPGS, and RSGS.

Scanning Strategy	Estimate	Standard Error	Iterations
DUGS	3.912	0.00987	15000
RPGS with $q_1 = 0.01$	3.906	0.00982	15000
RPGS with $q_1 = 0.50$	3.908	0.00945	16000
RPGS with $q_1 = 0.99$	3.920	0.00942	17000
RSGS with $p_1 = 0.01$	3.917	0.00963	2500000
RSGS with $p_1 = 0.50$	3.918	0.00994	45000
RSGS with $p_1 = 0.99$	3.914	0.00928	15000

References

Amit, Y. and Grenander, U. (1991). Comparing sweep strategies for stochastic relaxation. *Journal of Multivariate Analysis*, **37** 197–222.

Casella, G. and Berger, R. (2002). *Statistical Inference, Second edition*. Duxbury.

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

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, **91** 883–904.

Diaconis, P., Khare, K. and Saloff-Coste, L. (2008a). Gibbs sampling, conjugate priors and coupling. Tech. rep., Stanford University.

Diaconis, P., Khare, K. and Saloff-Coste, L. (2008b). Gibbs sampling, exponential families and orthogonal polynomials. *Statistical Science*, **23** 151–178.

Fishman, G. S. (1996). Coordinate selection rules for Gibbs sampling. *Annals of Applied Probability*, **6** 444–465.

Flegal, J. M., Haran, M. and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science,* to appear.

Flegal, J. M. and Jones, G. L. (2008). Batch means and spectral variance estimators in Markov chain Monte Carlo. Tech. rep., University of Minnesota, School of Statistics.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85** 398–409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis, Second edition*. Chapman & Hall/CRC.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6** 721–741.

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

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

Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, **23** 53–60.

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

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

Hobert, J. P., Jones, G. L. and Robert, C. P. (2006). Using a Markov chain to construct a tractable approximation of an intractable probability distribution. *Scandinavian Journal of Statistics*, **33** 37–51.

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*, **14** 1295–1305.

HODGES, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society, Series B*, **60** 497–536.

JAIN, N. and JAMISON, B. (1967). Contributions to Doeblin's theory of Markov processes. *Z. Wahrsch. Verw. Geb.*, **8** 19–40.

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

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, **101** 1537–1547.

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*, **32** 784–817.

KENDALL, W. S. (2004). Geometric ergodicity and perfect simulation. *Electronic Communications in Probability*, **9** 140–151.

LATUSZYNSKI, K. (2008). MCMC (ε-α)-approximation under drift condition with application to Gibbs samplers for a hierarchical random effects model. *Preprint.*

LEVINE, R. A. and CASELLA, G. (2006). Optimizing random scan Gibbs samplers. *Journal of Multivariate Analysis*, **97** 2071–2100.

LEVINE, R. A., YU, Z., HANLEY, W. G. and NITAO, J. J. (2005). Implementing random scan Gibbs samplers. *Computational Statistics*, **20** 177–196.

LINDVALL, T. (1992). *Lectures on the coupling method.* Wiley, New York.

LIU, J. S., WONG, W. H. and KONG, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society, Series B*, **57** 157–169.

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, London.

MYKLAND, P., TIERNEY, L. and YU, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, **90** 233–241.

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

PAPASPILIOPOULOS, O. and ROBERTS, G. (2007). Stability of the Gibbs sampler for Bayesian hierarchical models. *The Annals of Statistics,* to appear.

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

ROBERTS, G. O. and ROSENTHAL, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, **2** 13–25.

ROBERTS, G. O. and ROSENTHAL, J. S. (1998). On convergence rates of Gibbs samplers for uniform distributions. *Annals of Applied Probability*, **8** 1291–1302.

ROBERTS, G. O. and ROSENTHAL, J. S. (2001). Markov chains and de-initializing processes. *Scandinavian Journal of Statistics*, **28** 489–504.

ROBERTS, G. O. and ROSENTHAL, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, **1** 20–71.

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

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

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

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.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2005). Winbugs version 2.10. Tech. rep., MRC Biostatistics Unit, Cambridge: UK.

Tan, A. and Hobert, J. P. (2008). Block Gibbs sampling for Bayesian random effects models with improper priors: convergence and regeneration. *Preprint*.

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

Watkins, W. (1974). Convex matrix functions. *Proceedings of the American Mathematical Society*, **44** 31–34.

Appendix A

Proofs of Chapter 4 Results

Proof of Lemma 4.1

Harris ergodicity for the DUGS and RPGS follows directly from Lemma 1 of Tan and Hobert (2008). Therefore, it only remains to establish Harris ergodicity for the RSGS. That is, we must establish that the RSGS has invariant distribution π and is irreducible, aperiodic, and Harris recurrent. First, π is invariant by the reversibility of the RSGS with respect to π. Also, the RSGS is π-irreducible since for any $x \in \mathcal{X}$ and any set $A \in \mathcal{B}$ with $\pi(A) > 0$, $P^d(x, A) > 0$ by assumption. By the construction of the RSGS, this also guarantees $P^n(x, A) > 0$ for all $n \geq d$. Hence the RSGS is aperiodic. Finally, Harris recurrence follows from Corollary 1 of Tierney (1994).

Corollary A.1. *Let Φ be a π-irreducible Markov chain with invariant distribution π and transition kernel $P(x, \cdot)$. Then Φ is Harris recurrent so long as $P(x, \cdot)$ is absolutely continuous with respect to π for all $x \in \mathcal{X}$.*

Proof of Theorem 4.1

Without loss of generality, consider the 2-component DUGS with update order (1,2). We will construct the drift and minorization conditions for this sampler. Geometric ergodicity then follows directly from Proposition 2.1.

1. Drift condition for DUGS

Let E_D denote expectation with respect to the DUGS transition density and notice that

$$\mathrm{E}_D\left[V_D(y) \,|\, x\right] = \mathrm{E}\left[f_2(y_2) \,|\, x_2\right] = \mathrm{E}\left[\mathrm{E}\left(f_2(y_2) \,|\, y_1\right) \,|\, x_2\right].$$

Therefore, an application of (4.1) shows that

$$\begin{aligned}
\mathrm{E}_D\left[V_D(y) \,|\, x\right] = \mathrm{E}\left[\mathrm{E}\left(f_2(y_2) \,|\, y_1\right) \,|\, x_2\right] &\leq \mathrm{E}\left[cf_1(y_1) + d \,|\, x_2\right] \\
&\leq c\left[af_2(x_2) + b\right] + d \\
&= acV_D(x) + (cb + d) \\
&= \gamma_D V_D(x) + L_D \,.
\end{aligned}$$

Since ac is positive and strictly less than one by assumption, the result holds.

2. Minorization condition for DUGS

   To establish the minorization condition on C_D, notice that $C_D = D_2$. Therefore, for any $x \in C_D$, the DUGS transition density k_D for update order $(1, 2)$ satisfies

$$\begin{aligned}
k_D(x, y) &= \pi\left(y_1 \,|\, x_2\right) \pi\left(y_2 \,|\, y_1\right) \\
&\geq \inf_{x \in D_2} \pi\left(y_1 \,|\, x_2\right) \pi\left(y_2 \,|\, y_1\right) \\
&\geq g_1\left(y_1\right) \pi\left(y_2 \,|\, y_1\right) \\
&= \varepsilon_D q_D(y)
\end{aligned}$$

   and the minorization condition holds.

Proof of Theorem 4.2

We will construct the RPGS drift and minorization conditions. Geometric ergodicity then follows directly from Proposition 2.1.

1. Drift condition for RPGS

   First, recall that q_1 and q_2 are the permutation probabilities corresponding

to update orders $o_1 = (1, 2)$ and $o_2 = (2, 1)$, respectively. Therefore, the 2-component RSGS transition density is

$$k_P(x, y) = \sum_{i=1}^{2} q_i \pi\left(y_i | x_{-i}\right) \pi(y_{-i} | y_i).$$

It follows that

$$\mathrm{E}_P\left[V_P(y) \,|\, x\right] = \sum_{i=1}^{2} q_i \mathrm{E}\left[V_P(y) \,|\, x, o_i\right] = \sum_{i=1}^{2} q_i \mathrm{E}\left[f_1(y_1) + u f_2(y_2) \,|\, x, o_i\right] \quad \text{(A.1)}$$

where E_P denotes expectation with respect to the RPGS transition density. Also, notice that by (4.1)

$$\begin{aligned} \mathrm{E}\left[f_1(y_1) + u f_2(y_2) \,|\, x, o_1\right] &= \mathrm{E}\left[f_1(y_1) + u f_2(y_2) \,|\, x_2\right] \\ &\leq a f_2(x_2) + b + u \mathrm{E}\left[\mathrm{E}\left(f_2(y_2) \,|\, y_1\right) \,|\, x_2\right] \\ &\leq a f_2(x_2) + b + u \mathrm{E}\left[c f_1(y_1) + d \,|\, x_2\right] \\ &\leq a(1 + uc) f_2(x_2) + [b + u(cb + d)] \ . \end{aligned} \quad \text{(A.2)}$$

Similarly, we can show that

$$\mathrm{E}\left[f_1(y_1) + u f_2(y_2) \,|\, x, o_2\right] \leq c(a + u) f_1(x_1) + [ud + (ad + b)] \ . \quad \text{(A.3)}$$

Combining (A.1) – (A.3) gives

$$\begin{aligned} \mathrm{E}_P\left[V_P(y) \,|\, x\right] &\leq q_1 \left\{a(1 + uc) f_2(x_2) + [b + u(cb + d)]\right\} \\ &\quad + q_2 \left\{c(a + u) f_1(x_1) + [ud + (ad + b)]\right\} \\ &= q_2 c(a + u) f_1(x_1) + q_1 a \left(\frac{1}{u} + c\right) u f_2(x_2) + L_P \\ &= \gamma_P V_P(x) + L_P \end{aligned}$$

since u is a solution to $q_2c(a+u) = q_1a\,(1/u + c)$. To see this, notice that

$$\begin{aligned} u &= \frac{(q_1 - q_2)ac + \sqrt{ac\,[ac + 4q_1q_2(1-ac)]}}{2q_2c} \\ &= \frac{-(q_2 - q_1)ac + \sqrt{[(q_2 - q_1)ac]^2 + 4q_1q_2ac}}{2q_2c} \end{aligned}$$

so that by the quadratic formula, u is a solution to the equation

$$q_2cu^2 + (q_2 - q_1)acu - q_1a = 0\ .$$

This gives

$$q_2c(a+u) - q_1a\left(\frac{1}{u} + c\right) = \frac{1}{u}\left[q_2cu^2 + (q_2 - q_1)acu - q_1a\right] = 0\ .$$

Finally, the result holds since $0 < \gamma_P < 1$ by (4.5) and the assumption that $ac < 1$.

2. Minorization condition for RPGS

Notice that $\omega \le \nu_1$ and $\omega/u \le \nu_2$ with $u > 0$ guarantee $C_P \subset D_1 \cap D_2$ since

$$\begin{aligned} C_P = \{x : V_P \le \omega\} &= \{x : f_1(x_1) + uf_2(x_2) \le \omega\} \\ &\subset \{x : f_1(x_1) \le \omega\} \cap \{x : f_2(x_2) \le \omega/u\} \\ &\subset \{x : f_1(x_1) \le \nu_1\} \cap \{x : f_2(x_2) \le \nu_2\} \\ &= D_1 \cap D_2\ . \end{aligned}$$

Therefore, it suffices to establish the RPGS minorization condition on $D_1 \cap D_2$.

To this end, for any $x \in D_1 \cap D_2$, the RPGS transition density k_P is such that

$$\begin{aligned} k_P(x,y) &= \sum_{j=1}^{2} q_j \pi\left(y_j|x_{-j}\right) \pi(y_{-j}|y_j) \\ &\geq \sum_{j=1}^{2} q_j \inf_{x \in D_1 \cap D_2} \pi\left(y_j|x_{-j}\right) \pi(y_{-j}|y_j) \\ &\geq \sum_{j=1}^{2} q_j g_j\left(y_j\right) \pi(y_{-j}|y_j) \\ &= \varepsilon_P q_P(y) \end{aligned}$$

and the minorization condition holds.

Proof of Theorem 4.3

We will construct the RSGS drift and minorization conditions. Geometric ergodicity then follows directly from Proposition 2.1.

1. Drift condition for RSGS

   First, in the 2-component setting, the RSGS transition density can be written as

   $$k_R(x,y) = p_1 \pi\left(y_1|x_2\right) I\left(x_2 = y_2\right) + p_2 \pi\left(y_2|x_1\right) I\left(x_1 = y_1\right) \ .$$

   Also, let E_R denote expectation with respect to the RSGS transition density. Then by the construction of the RSGS

   $$\begin{aligned} \mathrm{E}_R\left[V_R(y) \,\middle|\, x\right] &= \mathrm{E}_R\left[f_1(y_1) + v f_2(y_2) \,\middle|\, x\right] \\ &= \sum_{i=1}^{2} p_i \mathrm{E}\left[f_1(y_1) + v f_2(y_2) \,\middle|\, x_{-i}\right] I(y_{-i} = x_{-i}) \\ &\leq p_1 \mathrm{E}\left[f_1(y_1) + v f_2(x_2) \,\middle|\, x_2\right] + p_2 \mathrm{E}\left[f_1(x_1) + v f_2(y_2) \,\middle|\, x_1\right] \\ &= p_1\left[\mathrm{E}(f_1(y_1)|x_2) + v f_2(x_2)\right] + p_2\left[f_1(x_1) + v\mathrm{E}(f_2(y_2)|x_1)\right] \ . \end{aligned}$$

Therefore, by (4.1)

$$\begin{aligned}\mathrm{E}_R\left[V_R(y)\,\middle|\,x\right] &\leq p_1\left[(a+v)f_2(x_2)+b\right]+p_2\left[(1+vc)f_1(x_1)+vd\right]\\ &= p_2(1+vc)f_1(x_1)+p_1\left(\frac{a}{v}+1\right)vf_2(x_2)+\left[p_1b+p_2vd\right]\\ &= \gamma_R V_R(x)+L_R\end{aligned}$$

since v is a solution to $p_2(1+vc)=p_1(a/v+1)$. This follows from the fact that

$$v=\frac{(p_1-p_2)+\sqrt{1-4p_1p_2(1-ac)}}{2p_2c}=\frac{-(p_2-p_1)+\sqrt{(p_2-p_1)^2+4p_1p_2ac}}{2p_2c}\,.$$

Therefore, by the quadratic formula, v is a solution to the equation

$$p_2cv^2+(p_2-p_1)v-p_1a=0$$

so that

$$p_2(1+vc)-p_1\left(\frac{a}{v}+1\right)=\frac{1}{v}\left[p_2cv^2+(p_2-p_1)v-p_1a\right]=0\,.$$

Finally, $0<\gamma_R<1$ by (4.5) and the result holds.

2. Minorization condition for RSGS

Similar to the proof of the RPGS minorization condition, $\omega\leq\nu_1$ and $\omega/v\leq\nu_2$ with $v>0$ guarantee $C_R\subset D_1\cap D_2$. Therefore, it suffices to establish a minorization condition on $D_1\cap D_2$. To this end, for any $x\in D_1\cap D_2$, the RSGS transition density k_R satisfies

$$\begin{aligned}k_R(x,y) &= \sum_{j=1}^{2}p_j\pi\left(y_j|x_{-j}\right)I\left(x_{-j}=y_{-j}\right)\\ &\geq \sum_{j=1}^{2}p_j\inf_{x\in D_1\cap D_2}\pi\left(y_j|x_{-j}\right)I\left(x_{-j}=y_{-j}\right)\\ &\geq \sum_{j=1}^{2}p_jg_j\left(y_j\right)I\left(x_{-j}=y_{-j}\right)\quad.\end{aligned}$$

Since it is true by the construction of the RSGS that $I\left(x_{-j} = y_{-j}\right) = 1$ for one and only one j, this gives

$$k_R(x, y) \geq \min_{j \in \{1,2\}} \left\{p_j g_j\left(y_j\right)\right\} = \varepsilon_R q_R(y)$$

for $x \in D_1 \cap D_2$ and the minorization condition holds.

Proof of Theorem 4.4

Let $\pi(dx_1, dx_2)$ denote the target distribution on $(\mathcal{X}, \mathcal{B})$ where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. In a slight abuse of notation, suppose π admits a density $\pi(x_1, x_2)$ with respect to some reference measure $\mu(dx_1, dx_2) = \mu_1(dx_1)\mu_2(dx_2)$. For simplicity, we will assume μ_i is Lebesgue for both $i = 1, 2$. Finally, let $\pi_1(x_1) = \int_{\mathcal{X}_2} \pi(x_1, x_2)dx_2$ and $\pi_2(x_2) = \int_{\mathcal{X}_1} \pi(x_1, x_2)dx_1$ denote the associated marginal densities on $(\mathcal{X}_1, \mathcal{B}_1)$ and $(\mathcal{X}_2, \mathcal{B}_2)$, respectively.

Consider DUGS for π and let $\Phi = \left\{X^{(0)}, X^{(1)}, X^{(2)}, \ldots\right\}$ denote the corresponding Markov chain where $X^{(t)} = \left(X_1^{(t)}, X_2^{(t)}\right)$. Also, for $i = 1, 2$, denote the DUGS subchains as $\Phi_i = \left\{X_i^{(0)}, X_i^{(1)}, X_i^{(2)}, \ldots\right\}$. The transition densities k_1 and k_2 for Φ_1 and Φ_2, respectively, can be written as follows:

$$k_1(x_1, y_1) = \int \pi(y_2|x_1)\pi(y_1|y_2)dy_2 \quad \text{and} \quad k_2(x_2, y_2) = \int \pi(y_1|x_2)\pi(y_2|y_1)dy_1 \, .$$

It follows that Φ_i is an aperiodic and irreducible Markov chain with invariant density π_i. Further, geometric ergodicity of Φ_i follows from geometric ergodicity of Φ (Roberts and Rosenthal, 2001). Therefore, the following conditions hold for $i = 1, 2$.

(C1) Let P_i denote the transition kernel corresponding to Φ_i. Then by Theorem 2.3, there exists drift function $V_i \geq 1$, small set C_i, and constants $b_i < \infty$ and $\beta_i < 1$ for which

$$P_i V_i(x_i) \leq \beta_i V_i(x_i) + b_i I_{C_i}(x_i) \quad \text{for all } x_i \in \mathcal{X}_i \, . \tag{A.4}$$

(C2) By Lemma 15.2.2 and Theorem 5.5.7 of Meyn and Tweedie (1993), V_i is unbounded off small sets. That is, for any $w > 0$, $\{x_i : V_i(x_i) \leq w\}$ is small with respect to P_i.

Geometric ergodicity of RPGS and RSGS will be established by appealing to Theorem 2.3 in conjunction with (C1) and (C2). We begin by constructing RPGS and RSGS drift conditions.

Drift Conditions

Define $g_1(x_1) = \int V_2(z_2)\pi(z_2|x_1)dz_2$ and $g_2(x_2) = \int V_1(z_1)\pi(z_1|x_2)dz_1$ where V_1 and V_2 are the Φ_1 and Φ_2 drift functions, respectively, guaranteed by (C1). For both RPGS and RSGS, we establish a drift condition of the form

$$PV(x) \leq \rho V(x) + bI_C(x) \quad \text{for all } x \in \mathcal{X} \tag{A.5}$$

where

$$V(x) = V_1(x_1) + V_2(x_2) + ug_2(x_2) + vg_1(x_1) \ ,$$

$C = \{x : V(x) \leq b/(1-\rho)\}$, and (P, ρ, b, u, v) depend on scanning strategy. We consider RPGS and RSGS separately.

1. RPGS drift condition

   Let P denote the RPGS transition kernel with associated transition density

   $$k_P(x, y) = q_1\pi(y_1|x_2)\pi(y_2|y_1) + q_2\pi(y_2|x_1)\pi(y_1|y_2) \tag{A.6}$$

   and in the notation of (A.4), let u and v be any constants for which

   $$\frac{q_1}{1-q_1\beta_1} < u < \frac{1-q_2\beta_1}{q_2\beta_1} \quad \text{and} \quad \frac{q_2}{1-q_2\beta_2} < v < \frac{1-q_1\beta_2}{q_1\beta_2} \ .$$

   Also, define

   $$\gamma = \max\left\{q_2\beta_1(1+u),\ q_1\beta_2(1+v),\ \frac{q_1(1+\beta_1 u)}{u},\ \frac{q_2(1+\beta_2 v)}{v}\right\}$$
   $$b = b_1(q_2+u) + b_2(q_1+v)$$
   $$\rho = \frac{\gamma+1}{2} \ .$$

Then $V(x) \geq 1$ and $\gamma < 1$. To establish (A.5), first notice that

$$\begin{aligned}
(PV_1)(x) &= \int\int V_1(y_1)\left[q_1\pi(y_1|x_2)\pi(y_2|y_1) + q_2\pi(y_2|x_1)\pi(y_1|y_2)\right] dy_1 dy_2 \\
&= q_1 \int V_1(y_1)\pi(y_1|x_2)dy_1 + q_2 \int V_1(y_1) \int \pi(y_2|x_1)\pi(y_1|y_2)dy_2 dy_1 \\
&= q_1 g_2(x_2) + q_2 \int V_1(y_1)k_1(x_1, y_1)dy_1 \\
&= q_1 g_2(x_2) + q_2 P_1 V_1(x_1) \\
&\leq q_1 g_2(x_2) + q_2\beta_1 V_1(x_1) + q_2 b_1
\end{aligned} \tag{A.7}$$

where the inequality follows from (C1). Similarly,

$$(PV_2)(x) \leq q_2 g_1(x_1) + q_1\beta_2 V_2(x_2) + q_1 b_2 \ . \tag{A.8}$$

Next,

$$\begin{aligned}
(Pg_2)(x) &= \int\int g_2(y_2)\left[q_1\pi(y_1|x_2)\pi(y_2|y_1) + q_2\pi(y_2|x_1)\pi(y_1|y_2)\right] dy_1 dy_2 \\
&= \int\int\int V_1(z_1)\pi(z_1|y_2)\left[q_1\pi(y_1|x_2)\pi(y_2|y_1) + q_2\pi(y_2|x_1)\pi(y_1|y_2)\right] dz_1 dy_1 dy_2 \\
&= q_1 \int \pi(y_1|x_2) \int V_1(z_1) \int \pi(z_1|y_2)\pi(y_2|y_1)dy_2 dz_1 dy_1 \\
&\quad + q_2 \int V_1(z_1) \int \pi(z_1|y_2)\pi(y_2|x_1)dy_2 dz_1 \\
&= q_1 \int \pi(y_1|x_2) \int V_1(z_1)k_1(y_1, z_1)dz_1 dy_1 + q_2 \int V_1(z_1)k_1(x_1, z_1)dz_1 \\
&= q_1 \int \pi(y_1|x_2)P_1 V_1(y_1)dy_1 + q_2 P_1 V_1(x_1) \\
&\leq q_1 \int \pi(y_1|x_2)\left[\beta_1 V_1(y_1) + b_1 I_{C_1}(y_1)\right] dy_1 + q_2\left[\beta_1 V_1(x_1) + b_1 I_{C_1}(x_1)\right] \\
&\leq q_1\beta_1 g_2(x_2) + q_2\beta_1 V_1(x_1) + q_1 b_1 + q_2 b_1 I_{C_1}(x_1) \\
&\leq q_1\beta_1 g_2(x_2) + q_2\beta_1 V_1(x_1) + b_1 \ .
\end{aligned} \tag{A.9}$$

Similarly,

$$(Pg_1)(x) \leq q_2\beta_2 g_1(x_1) + q_1\beta_2 V_2(x_2) + b_2 \; . \tag{A.10}$$

Combining (A.7) – (A.10) gives

$$\begin{aligned} PV(x) &= (PV_1)(x) + (PV_2)(x) + u(Pg_2)(x) + v(Pg_1)(x) \\ &\leq q_2\beta_1(1+u)V_1(x_1) + q_1\beta_2(1+v)V_2(x_2) \\ &\quad + \left(\frac{q_1(1+\beta_1 u)}{u}\right) u g_2(x_2) + \left(\frac{q_2(1+\beta_2 v)}{v}\right) v g_1(x_1) + b \\ &\leq \gamma\left[V_1(x_1) + V_2(x_2) + u g_2(x_2) + v g_1(x_1)\right] + b \\ &= \gamma V(x) + b \; . \end{aligned}$$

Finally, recall $\rho = (\gamma + 1)/2$. Therefore, (A.5) holds since

$$\begin{aligned} PV(x) &\leq \gamma V(x) + b \\ &= (2\rho - 1)V(x) + b \\ &= \rho V(x) - (1-\rho)V(x) + b \\ &\leq \rho V(x) + b I_C(x) \end{aligned} \tag{A.11}$$

where $C = \{x : V(x) \leq w\}$ for $w = b/(1-\rho)$.

2. RSGS drift condition

Let P denote the RSGS transition kernel with associated transition density

$$k_R(x, y) = p_1\pi(y_1|x_2)I(y_2 = x_2) + p_2\pi(y_2|x_1)I(y_1 = x_1) \; . \tag{A.12}$$

In the notation of (A.4), let u and v be any values for which

$$\frac{p_1}{p_2} < u < \frac{p_1}{p_2\beta_1} \quad \text{and} \quad \frac{p_2}{p_1} < v < \frac{p_2}{p_1\beta_2} \; .$$

Also, define

$$\begin{aligned}\gamma &= \max\left\{p_2(1+u\beta_1),\ p_1(1+v\beta_2),\ \frac{p_1(1+u)}{u},\ \frac{p_2(1+v)}{v}\right\}\\ b &= up_2b_1 + vp_1b_2\\ \rho &= \frac{\gamma+1}{2}\ .\end{aligned}$$

Then $V(x) \geq 1$, $\gamma < 1$, and (A.5) holds. To establish this result, notice that

$$\begin{aligned}(PV_1)(x) &= \sum_{i=1}^{2} p_i \mathrm{E}\left[V_1(y_1)|x_{-i}\right] I(y_{-i} = x_{-i})\\ &\leq p_1 \mathrm{E}\left[V_1(y_1)|x_2\right] + p_2 \mathrm{E}\left[V_1(x_1)|x_1\right]\\ &= p_1 g_2(x_2) + p_2 V_1(x_1)\ .\end{aligned} \tag{A.13}$$

Similarly,

$$(PV_2)(x) \leq p_2 g_1(x_1) + p_1 V_2(x_2)\ . \tag{A.14}$$

Next,

$$\begin{aligned}(Pg_2)(x) &= \sum_{i=1}^{2} p_i \mathrm{E}\left[g_2(y_2)|x_{-i}\right] I(y_{-i} = x_{-i})\\ &\leq p_1 \mathrm{E}\left[g_2(x_2)|x_2\right] + p_2 \mathrm{E}\left[g_2(y_2)|x_1\right]\\ &= p_1 g_2(x_2) + p_2 \int\int V_1(z_1)\pi(z_1|y_2)\pi(y_2|x_1)dz_1dy_2\\ &= p_1 g_2(x_2) + p_2 \int V_1(z_1)k_1(x_1, z_1)dz_1\\ &= p_1 g_2(x_2) + p_2 P_1 V_1(x_1)\\ &\leq p_1 g_2(x_2) + p_2\beta_1 V_1(x_1) + p_2 b_1\ .\end{aligned} \tag{A.15}$$

Similarly,

$$(Pg_1)(x) \leq p_2 g_1(x_1) + p_1\beta_2 V_2(x_2) + p_1 b_2\ . \tag{A.16}$$

The result holds by combining (A.13) – (A.16):

$$\begin{aligned} PV(x) &= (PV_1)(x) + (PV_2)(x) + u(Pg_2)(x) + v(Pg_1)(x) \\ &\leq p_2(1+u\beta_1)V_1(x_1) + p_1(1+v\beta_2)V_2(x_2) \\ &\quad + \left(\frac{p_1(1+u)}{u}\right) ug_2(x_2) + \left(\frac{p_2(1+v)}{v}\right) vg_1(x_1) + b \\ &\leq \gamma V(x) + b \\ &\leq \rho V(x) + bI_C(x) \end{aligned}$$

where $C = \{x : V(x) \leq w\}$ for $w = b/(1-\rho)$ and the last inequality holds by (A.11) .

Minorization Conditions

We have now established (A.5) for both RPGS and RSGS. Then by Theorem 2.3, RPGS and RSGS are geometrically ergodic if the corresponding sets $C = \{x : V(x) \leq w\}$ are small where $w = b/(1-\rho)$. We first make some observations that will be useful in establishing this result.

Let P_2 and k_2 denote the transition kernel and density, respectively, of sub-chain Φ_2. Also, let $D_2 = \{x_2 : V_2(x_2) \leq w\}$. It will suffice to establish $\mathcal{X}_1 \times D_2$ is small since $C \subset \mathcal{X}_1 \times D_2$. To this end, (C2) guarantees D_2 is small with respect to P_2. Hence there exist m, $0 < \varepsilon < 1$, and probability measure $Q(\cdot)$ on $(\mathcal{X}_2, \mathcal{B}_2)$ for which

$$P_2^m(x_2, A) \geq \varepsilon Q(A) \quad \text{for all } x_2 \in D_2 \text{ and } A \in \mathcal{B}_2 \,. \tag{A.17}$$

Next, we make the following claim, proved below.

Claim 1.

Probability measure $Q(\cdot)$ has an associated density $q(\cdot)$. Further, for all $x_2 \in D_2$ the set $D = \{y_2 \in \mathcal{X}_2 : k_2^m(x_2, y_2) \geq \varepsilon q(y_2)\}$ has positive Lebesgue measure.

Proof. Under the assumptions of Theorem 4.4, the DUGS transition kernel is absolutely continuous with respect to $\pi(\cdot)$. It follows that $P_2^m(x_2, \cdot)$ is absolutely continuous with respect to marginal $\pi_2(\cdot)$ for all $x_2 \in \mathcal{X}_2$. Then for any $A \in \mathcal{B}_2$ with

$\pi_2(A) = 0$ and any $x_2 \in D_2$,

$$0 = P_2^m(x_2, A) \geq \varepsilon Q(A) \ .$$

Therefore, $Q(\cdot)$ is also absolutely continuous with respect to $\pi_2(\cdot)$. By the Radon-Nikodym theorem, $Q(\cdot)$ has an associated density $\tilde{q}(\cdot)$ with respect to π_2. That is, for any $A \in \mathcal{B}_2$

$$Q(A) = \int_A \tilde{q}(y_2)\pi_2(dy_2) = \int_A q(y_2)dy_2$$

where $q(y_2) = \tilde{q}(y_2)\pi_2(y_2)$.

It remains to show that D has positive Lebesgue measure for all $x_2 \in D_2$. For a proof by contradiction, suppose D has Lebesgue measure zero for some $x_2 \in D_2$. Also, for any $A \in \mathcal{B}_2$ let $\tilde{A} = A \cap D$. Since $\tilde{A} \subseteq D$, $\tilde{A}$ also has Lebesgue measure zero. Therefore,

$$\begin{aligned} P_2^m(x_2, A) &= \int_{\tilde{A}} k_2^m(x_2, y_2)dy_2 + \int_{A/\tilde{A}} k_2^m(x_2, y_2)dy_2 \\ &= \int_{A/\tilde{A}} k_2^m(x_2, y_2)dy_2 \\ &< \int_{A/\tilde{A}} \varepsilon q(y_2)dy_2 \\ &< \varepsilon Q(A) \ . \end{aligned}$$

This contradicts (A.17) so the result holds. □

Let $q(\cdot)$ be the density and D be the set in Claim 1. Also, define $\tilde{g}(y) = \tilde{\varepsilon}^{-1} \int_D q(z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2$ for

$$\tilde{\varepsilon} = \int_{\mathcal{X}} \int_D q(z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2 d_y \ .$$

We establish $\mathcal{X}_1 \times D_2$ is small for RPGS and RSGS separately.

1. RPGS minorization condition

   We will establish an $(m+1)$-step minorization condition for the RPGS. There

are many possible paths the RPGS can take in $m+1$ iterations. For instance, with probability q_1^{m+1}, the RPGS follows update order (1,2) in each iteration. Further, the probability of moving from some state x to state y on this path is equivalent to

$$\int_{\mathcal{X}_2} k_2^m(x_2, z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2$$

where k_2 is the Φ_2 transition density. Therefore,

$$k_P^{m+1}(x, y) \geq q_1^{m+1} \int_{\mathcal{X}_2} k_2^m(x_2, z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2$$

where k_P is the RPGS transition density defined by (A.6). A minorization condition follows since for any $x \in \mathcal{X}_1 \times D_2$ and $y \in \mathcal{X}$,

$$\begin{aligned} k_P^{m+1}(x, y) &\geq q_1^{m+1} \int_{\mathcal{X}_2} k_2^m(x_2, z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2 \\ &\geq q_1^{m+1} \int_{D} k_2^m(x_2, z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2 \\ &\geq \varepsilon q_1^{m+1} \int_{D} q(z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2 \\ &= \tilde{\varepsilon}\tilde{g}(y) \, . \end{aligned}$$

2. RSGS minorization condition

We will establish a $(2m+2)$-step minorization condition for RSGS. Again, there are many possible paths the RSGS might take in $2m+2$ iterations. For instance, with probability $(p_1p_2)^{m+1}$ the RSGS alternately updates the first and second components $m+1$ times. Since RSGS only updates a single component in each iteration, the probability of moving from some state x to state y on this path is equivalent to

$$\int_{\mathcal{X}_2} k_2^m(x_2, z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2$$

where k_2 is the Φ_2 transition density. Therefore,

$$k_R^{2m+2}(x, y) \geq (p_1p_2)^{m+1} \int_{\mathcal{X}_2} k_2^m(x_2, z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2$$

where k_R is the RSGS transition density defined by (A.12). A minorization condition follows since for any $x \in \mathcal{X}_1 \times D_2$ and $y \in \mathcal{X}$,

$$\begin{aligned} k_R^{2m+2}(x,y) &\geq (p_1 p_2)^{m+1} \int_{\mathcal{X}_2} k_2^m(x_2, z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2 \\ &\geq (p_1 p_2)^{m+1} \int_{D} k_2^m(x_2, z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2 \\ &\geq \varepsilon(p_1 p_2)^{m+1} \int_{D} q(z_2)\pi(y_1|z_2)\pi(y_2|y_1)dz_2 \\ &= \tilde{\varepsilon}\tilde{g}(y) \ . \end{aligned}$$

Remark A.1. Notice that the $(m+1)$-step RPGS and $(2m+2)$-step RSGS minorization conditions have the same lower bounds:

$$k_P^{m+1}(x,y) \geq \tilde{\varepsilon}\tilde{q}(y) \quad \text{and} \quad k_R^{2m+2}(x,y) \geq \tilde{\varepsilon}\tilde{q}(y) \ .$$

Proof of Proposition 4.1

Let E denote expectation with respect to the mixture model $\pi_m(x_1, x_2)$. First, notice that

$$\begin{aligned} \mathrm{E}\left[f_1(x_1)|x_2\right] = \sum_{i=1}^{s} \phi_i \mathrm{E}_i\left[f_1(x_1)|x_2\right] &\leq \sum_{i=1}^{s} \phi_i \left[a_i f_2(x_2) + b_i\right] \\ &\leq \max_i\{a_i\} f_2(x_2) + \sum_{i=1}^{s} \phi_i b_i \ . \end{aligned}$$

Also,

$$\begin{aligned} \mathrm{E}\left[f_2(x_2)|x_1\right] = \sum_{i=1}^{s}\sum_{j=1}^{t} w_{ij}(x_1)\mathrm{E}_{ij}\left[f_2(x_2)|x_1\right] &\leq \sum_{i=1}^{s}\sum_{j=1}^{t} w_{ij}(x_1)[c_{ij} f_1(x_1) + d_{ij}] \\ &\leq \max_{ij}\{c_{ij}\} f_1(x_1) + \max_{ij}\{d_{ij}\} \end{aligned}$$

since for all x_1, i, j,

$$\sum_{i=1}^{s}\sum_{j=1}^{t} w_{ij}(x_1) = \sum_{i=1}^{s}\sum_{j=1}^{t}\left[\frac{\phi_i\psi_j\pi_{ij}(x_1)}{\sum_{k=1}^{s}\sum_{l=1}^{t}\phi_k\psi_l\pi_{kl}(x_1)}\right] = 1\ .$$

Proof of Proposition 4.2

First, notice that

$$\inf_{x\in D_2}\pi_m(y_1|x_2) = \inf_{x\in D_2}\sum_{i=1}^{s}\phi_i\pi_i(y_1|x_2) \geq \sum_{i=1}^{s}\phi_i\inf_{x\in D_2}\pi_i(y_1|x_2) = g_1(y_1)\ .$$

Also,

$$\begin{aligned}
\inf_{x\in D_1}\pi_m(y_2|x_1) &= \inf_{x\in D_1}\sum_{i=1}^{s}\sum_{j=1}^{t} w_{ij}(x_1)\pi_{ij}(y_2|x_1) \\
&\geq \sum_{i=1}^{s}\sum_{j=1}^{t} w_{ij}(x_1)\inf_{x\in D_1}\pi_{ij}(y_2|x_1) \\
&\geq \sum_{i=1}^{s}\sum_{j=1}^{t} w_{ij}(x_1)g_{ij}(y_2) \\
&\geq g_2(y_2)
\end{aligned}$$

and the result holds.

Proof of Proposition 4.3

Let E_D, E_P, and E_R denote expectation with respect to the DUGS, RPGS, and RSGS transition densities, respectively. We consider the construction of the DUGS, RPGS, and RSGS drift conditions separately.

1. Drift condition for DUGS

   For ease of exposition, we construct the drift condition for the 3-component

DUGS with update order (1,2,3). By the construction of this DUGS,

$$\mathrm{E}_D\left[V_D(y)\,|\,x\right] = \mathrm{E}\left[\mathrm{E}\left(\mathrm{E}\left[V_D(y)\,|\,y_1,y_2\right]\,|\,y_1,x_3\right)\,|\,x_2,x_3\right]\ .$$

Therefore, the drift condition holds since by (4.15) we have

$$\begin{aligned}
\mathrm{E}_D\left[V_D(y)\,|\,x\right] &= \sum_{i=1}^{m} w_i \mathrm{E}\left[\mathrm{E}\left(\mathrm{E}\left[f_i\left(y_{J_i}\right)\,|\,y_1,y_2\right]\,|\,y_1,x_3\right)\,|\,x_2,x_3\right] \\
&\leq \sum_{i=1}^{m}\sum_{j=1}^{m} w_i\alpha_{i3j}\mathrm{E}\left[\mathrm{E}\left(f_j\left(y_{J_j}\right)\,|\,y_1,x_3\right)\,|\,x_2,x_3\right] + \sum_{i=1}^{m} w_i\beta_{i3\cdot} \\
&\leq \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} w_i\alpha_{i3j}\alpha_{j2k}\mathrm{E}\left[f_k\left(y_{J_k}\right) I\left(y_3 = x_3\right)\,|\,x_2,x_3\right] \\
&\quad + \sum_{i=1}^{m} w_i\left(\beta_{i3\cdot} + \sum_{j=1}^{m}\alpha_{i3j}\beta_{j2\cdot}\right) \\
&\leq \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} w_i\alpha_{i3j}\alpha_{j2k}\alpha_{k1l}f_l\left(x_{J_l}\right) \\
&\quad + \sum_{i=1}^{m} w_i\left(\beta_{i3\cdot} + \sum_{j=1}^{m}\alpha_{i3j}\beta_{j2\cdot} + \sum_{j=1}^{m}\sum_{k=1}^{m}\alpha_{i3j}\alpha_{j2k}\beta_{k1\cdot}\right) \\
&= \sum_{l=1}^{m}\gamma_{Dl}w_l f_l\left(x_{J_l}\right) + L_D \\
&\leq \gamma_D V_D(x) + L_D
\end{aligned}$$

where $\gamma_D < 1$ by assumption.

2. Drift condition for RPGS

Again, for ease of exposition consider RPGS for a 3-component model. Let $o_t = (t(1), t(2), t(3))$ denote the tth permutation of the block Gibbs update order where $t \in \{1, 2, \cdots, 6\}$. Then by the construction of the RPGS,

$$\mathrm{E}_P\left[V_P(y)\,|\,x\right] = \sum_{t=1}^{6} q_t\mathrm{E}\left[V_P(y)\,|\,x,o_t\right] = \sum_{t=1}^{6}\sum_{i=1}^{m} q_t u_i\mathrm{E}\left[f_i\left(y_{J_i}\right)\,|\,x,o_t\right]\ .$$

Also, by the proof of the drift construction for DUGS,

$$\mathrm{E}\left[f_i\left(y_{J_i}\right) \,\middle|\, x, o_t\right] \leq \sum_{j,k,l=1}^{m} \alpha_{it(3)j}\alpha_{jt(2)k}\alpha_{kt(1)l} f_l\left(x_{J_l}\right) + L_{Pti}$$

for $i \in \{1, \cdots, m\}$ and $t \in \{1, \cdots, 6\}$. Therefore, the RPGS drift follows from combining these results:

$$\begin{aligned}
\mathrm{E}_P\left[V_P(y) \,|\, x\right] &\leq \sum_{t=1}^{6} \sum_{i,j,k,l=1}^{m} q_t u_i \alpha_{it(3)j}\alpha_{jt(2)k}\alpha_{kt(1)l} f_l\left(x_{J_l}\right) + \sum_{t=1}^{6}\sum_{i=1}^{m} q_t u_i L_{Pti} \\
&= \sum_{l=1}^{m} f_l\left(x_{J_l}\right) \sum_{t=1}^{6} \sum_{i,j,k=1}^{m} q_t u_i \alpha_{it(3)j}\alpha_{jt(2)k}\alpha_{kt(1)l} + L_P \\
&= \sum_{l=1}^{m} \gamma_{Pl} u_l f_l\left(x_{J_l}\right) + L_P \\
&\leq \gamma_P V_P(x) + L_P \ .
\end{aligned}$$

3. Drift condition for RSGS

To establish the drift condition, notice that by the construction of the RSGS

$$\mathrm{E}_R\left[V_R(y) \,|\, x\right] = \mathrm{E}\left[\sum_{i=1}^{m} v_i f_i\left(y_{J_i}\right) \,\middle|\, x\right] = \sum_{i=1}^{m}\sum_{j=1}^{d} v_i p_j \mathrm{E}\left[f_i\left(y_{J_i}\right) \,\middle|\, x_{-j}\right] I\left(y_{-j} = x_{-j}\right) \ .$$

Also, the construction of the drift condition is simplified by the fact that $d-1$ components remain fixed in each iteration of the RSGS. Specifically, from (4.15)

we have

$$\begin{aligned}
\mathrm{E}_R\left[V_R(y) \,\middle|\, x\right] &\leq \sum_{i=1}^{m}\sum_{j=1}^{d}\sum_{k=1}^{m} v_i p_j \left[\alpha_{ijk} f_k\left(x_{J_k}\right) + \beta_{ijk}\right] \\
&= \sum_{k=1}^{m} f_k\left(x_{J_k}\right) \sum_{i=1}^{m}\sum_{j=1}^{d} v_i p_j \alpha_{ijk} + \sum_{i=1}^{m}\sum_{j=1}^{d}\sum_{k=1}^{m} v_i p_j \beta_{ijk} \\
&= \sum_{k=1}^{m} \gamma_{Rk} v_k f_k\left(x_{J_k}\right) + L_R \\
&\leq \gamma_R V_R(x) + L_R
\end{aligned}$$

and a drift condition holds since $\gamma_R < 1$ by assumption.

Appendix B

Proofs of Chapter 5 Results

Proof of Lemma 5.1

First, notice that

$$\begin{aligned}
\mathrm{E}\left[\left\langle \sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i \right\rangle \middle| \theta\right] &= \mathrm{E}\left[\sum_{i=1}^{n} \langle X_i, X_i\rangle + \sum_{i\neq j} \langle X_i, X_j\rangle \middle| \theta\right] \\
&= \sum_{i=1}^{n} \mathrm{E}\left[\langle X_i, X_i\rangle \mid \theta\right] + \sum_{i\neq j} \left\langle \mathrm{E}\left[X_i \,|\, \theta\right], \mathrm{E}\left[X_j \,|\, \theta\right]\right\rangle \\
&= n\left[\tilde{c}\,\langle\theta,\theta\rangle + \left\langle \tilde{d}, \theta\right\rangle + \tilde{e}\right] + n(n-1)\left\langle \tilde{a}\theta + \tilde{b}, \tilde{a}\theta + \tilde{b}\right\rangle \\
&= a\,\langle\theta,\theta\rangle + \left\langle n\tilde{d} + 2n(n-1)\tilde{a}\tilde{b}, \theta\right\rangle + \left[n\tilde{e} + n(n-1)\left\langle \tilde{b}, \tilde{b}\right\rangle\right] .
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathrm{E}[f_1(X)|\theta] &= \mathrm{E}\left[\left\langle \sum_{i=1}^n X_i - \tilde{u}, \sum_{i=1}^n X_i - \tilde{u} \right\rangle \middle| \theta\right] \\
&= \mathrm{E}\left[\left\langle \sum_{i=1}^n X_i, \sum_{i=1}^n X_i \right\rangle - 2\left\langle \tilde{u}, \sum_{i=1}^n X_i \right\rangle \middle| \theta\right] + \langle \tilde{u}, \tilde{u}\rangle \\
&= a\langle \theta, \theta\rangle + \left\langle n\tilde{d} + 2n(n-1)\tilde{a}\tilde{b}, \theta \right\rangle + \left[n\tilde{e} + n(n-1)\left\langle \tilde{b}, \tilde{b}\right\rangle\right] \\
&\quad - 2n\left\langle \tilde{u}, \tilde{a}\theta + \tilde{b}\right\rangle + \langle \tilde{u}, \tilde{u}\rangle \\
&= a\left[\langle \theta, \theta\rangle - 2\left\langle \frac{n(2\tilde{u}\tilde{a} - \tilde{d} - 2(n-1)\tilde{a}\tilde{b})}{2a}, \theta \right\rangle\right] \\
&\quad + \left[n\tilde{e} + n(n-1)\left\langle \tilde{b}, \tilde{b}\right\rangle - 2n\left\langle \tilde{u}, \tilde{b}\right\rangle + \langle \tilde{u}, \tilde{u}\rangle\right] \\
&= a\left[\langle \theta, \theta\rangle - 2\langle \tilde{v}, \theta\rangle\right] + \left[n\tilde{e} + n(n-1)\left\langle \tilde{b}, \tilde{b}\right\rangle - 2n\left\langle \tilde{u}, \tilde{b}\right\rangle + \langle \tilde{u}, \tilde{u}\rangle\right] \\
&= af_2(\theta) + b\,.
\end{aligned}$$

Next, we have

$$\begin{aligned}
\mathrm{E}[f_2(\theta)|x] &= \mathrm{E}[\langle \theta - \tilde{v}, \theta - \tilde{v}\rangle\,|x] \\
&= \mathrm{E}[\langle \theta, \theta\rangle - 2\langle \tilde{v}, \theta\rangle\,|x] + \langle \tilde{v}, \tilde{v}\rangle \\
&= \tilde{h}\left\langle \sum_{i=1}^n x_i, \sum_{i=1}^n x_i \right\rangle - 2\left\langle \frac{2\tilde{f}\tilde{v} - \tilde{j}}{2}, \sum_{i=1}^n x_i \right\rangle + \left[\langle \tilde{v}, \tilde{v}\rangle - 2\langle \tilde{v}, \tilde{g}\rangle + \tilde{k}\right] \\
&= \tilde{h}\left\langle \sum_{i=1}^n x_i, \sum_{i=1}^n x_i \right\rangle - 2\tilde{h}\left\langle \tilde{u}, \sum_{i=1}^n x_i \right\rangle + \left[\langle \tilde{v}, \tilde{v}\rangle - 2\langle \tilde{v}, \tilde{g}\rangle + \tilde{k}\right] \\
&= cf_1(x) + d\,.
\end{aligned}$$

Therefore, f_1 and f_2 satisfy (4.1) since $ac < 1$ is assumed.

Proof of Lemma 5.2

We will show that $\inf_{(x',\theta')\in D_2} \pi(x|\theta') \geq g_1(x)$. The proof that $\inf_{(x',\theta')\in D_1} \pi(\theta|x') \geq g_2(\theta)$ is very similar and is left to the reader. First, recall that $x = (x_1, \cdots, x_n)$ and

let x_{ij} denote the jth component of the d-dimensional vector x_i. Also, let θ_j denote the jth component of θ. Then for the multivariate exponential family of interest, we have

$$\pi(x|\theta) = \frac{1}{[c(\theta)]^n} \exp\left\{ \left\langle \sum_{i=1}^{n} x_i, \theta \right\rangle \right\} .$$

Also, notice that we can rewrite set D_2 as

$$D_2 = \left\{ (x,\theta) : \sum_{j=1}^{d} (\theta_j - \tilde{v}_j)^2 \leq \nu_2 \right\} .$$

Therefore, $D_2 \subset C_2$ where

$$\begin{aligned} C_2 &= \left\{ (x,\theta) : (\theta_j - \tilde{v}_j)^2 \leq \nu_2 \ \text{ for all } j \in \{1,\cdots,d\} \right\} \\ &= \left\{ (x,\theta) : \tilde{v}_j - \sqrt{\nu_2} \leq \theta_j \leq \tilde{v}_j + \sqrt{\nu_2} \ \text{ for all } j \in \{1,\cdots,d\} \right\} . \end{aligned}$$

It follows that

$$\begin{aligned} \inf_{(x',\theta')\in D_2} \pi(x|\theta') &\geq \inf_{(x',\theta')\in C_2} \pi(x|\theta') \\ &\geq \inf_{(x',\theta')\in C_2} \frac{1}{[c(\theta')]^n} \cdot \inf_{(x',\theta')\in C_2} \exp\left\{ \left\langle \sum_{i=1}^{n} x_i, \theta' \right\rangle \right\} \end{aligned} \tag{B.1}$$

where

$$\inf_{(x',\theta')\in C_2} \exp\left\{ \left\langle \sum_{i=1}^{n} x_i, \theta' \right\rangle \right\} \geq \exp\left\{ \left\langle \sum_{i=1}^{n} x_i, \tilde{v} - \sqrt{\nu_2}\text{sgn}\left(\sum_{i=1}^{n} x_i \right) \right\rangle \right\} ; \text{ and}$$

$$\begin{aligned} \inf_{(x',\theta')\in C_2} [c(\theta')]^{-n} &= \left[\sup_{(x',\theta')\in C_2} c(\theta') \right]^{-n} \\ &= \left[\sup_{(x',\theta')\in C_2} \int_{\mathcal{X}} \exp\left\{ \langle x, \theta' \rangle \right\} \lambda(dx) \right]^{-n} \\ &\geq \left[\int_{\mathcal{X}} \sup_{(x',\theta')\in C_2} \exp\left\{ \langle x, \theta' \rangle \right\} \lambda(dx) \right]^{-n} \\ &\geq \left[\int_{\mathcal{X}} \exp\left\{ \langle x, \tilde{v} + \sqrt{\nu_2}\text{sgn}(x) \rangle \right\} \lambda(dx) \right]^{-n} . \end{aligned} \tag{B.2}$$

Together, (B.1) and (B.2) establish that $\inf_{(x',\theta')\in D_2} \pi(x|\theta') \geq g_1(x)$.

Preliminaries for Chapter 5.2 Results

We will require the following Lemmas throughout the remainder of this chapter. A proof of Lemma B.1 is given in Henderson and Searle (1981) and Lemma B.2 follows from the convexity of the inverse function (see, for example, Watkins (1974)).

Lemma B.1. *Let A be a nonsingular $n \times n$ matrix, B be a nonsingular $s \times s$ matrix, U be an $n \times s$ matrix, and V be an $s \times n$ matrix. Then*

$$(A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

When $U = V$ this implies

$$x^T(A + UBV)^{-1}x \leq x^T A^{-1}x$$

for any $n \times 1$ vector x.

Lemma B.2. *Let x be an $m \times 1$ vector. Also, let A and B be nonsingular, $m \times m$ matrices. Then*

$$x^T(A + B)^{-1}x \leq \frac{1}{4}x^T\left(A^{-1} + B^{-1}\right)x\,.$$

Proof of Full Conditionals (5.8)

For simplicity, we consider the proof in the non-mixture setting ($r = s = 1$). In this case the model is represented by (5.17). First, for the precision parameters with hyperparameters r_1, r_2, d_1, and d_2 we have

$$\begin{aligned}\pi(\lambda_R|\xi, y) &\propto \pi(y|\lambda_R, \xi)f(\lambda_R) \\ &\propto \lambda_R^{N/2}\exp\{-0.5\lambda_R v_1(\xi)\} \cdot \lambda_R^{r_1-1}\exp\{-r_2\lambda_R\} \\ &\propto \lambda_R^{(r_1+N/2)-1}\exp\{-\lambda_R\left(r_2 + 0.5v_1(\xi)\right)\}\end{aligned}$$

and

$$\begin{aligned}\pi(\lambda_D|\xi,y) &\propto \pi(u|\lambda_D,\xi)f(\lambda_D)\\ &\propto \lambda_D^{k/2}\exp\{-0.5\lambda_D v_2(\xi)\}\cdot\lambda_D^{d_1-1}\exp\{-d_2\lambda_D\}\\ &\propto \lambda_D^{(d_1+k/2)-1}\exp\{-\lambda_D(d_2+0.5v_2(\xi))\}.\end{aligned}$$

Therefore, $\lambda_R|\xi,y \sim \text{Gamma}(r_1+N/2, r_2+0.5v_1(\xi))$ and $\lambda_D|\xi,y \sim \text{Gamma}(d_1+k/2, d_2+0.5v_2(\xi))$. Next, notice that for any $W \sim N_m\left(\mu_W, \Sigma_W^{-1}\right)$, the density of W is

$$\pi(w) \propto \exp\left\{-0.5(w-\mu_W)^T\Sigma_W(w-\mu_W)\right\} \propto \exp\left\{-0.5w^T\Sigma_W w + w^T\Sigma_W\mu_W\right\}\ .$$

Then for the full conditional of ξ, notice that

$$\begin{aligned}\pi(\xi|\lambda,y) &\propto \pi(y|\beta,u,\lambda)\pi(\beta|u,\lambda)\pi(u|\lambda)\\ &\propto \exp\{-0.5\lambda_R(y-X\beta-Zu)^T(y-X\beta-Zu)\}\\ &\qquad\cdot\exp\{-0.5\left[(\beta-\beta_0)^TB(\beta-\beta_0)+\lambda_D u^Tu\right]\}\\ &\propto \exp\{-0.5\left[\beta^T(\lambda_R X^TX+B)\beta + u^T(\lambda_R Z^TZ+\lambda_D I_k)u\right]\}\\ &\qquad\cdot\exp\{-0.5\left[-2\lambda_R u^TZ^Ty - 2\beta^T(\lambda_R X^Ty + B\beta_0) + 2\lambda_R\beta^TX^TZu\right]\}.\end{aligned}$$

Therefore,

$$\begin{aligned}\pi(\xi|\lambda,y) \propto \exp&\left\{-0.5\xi^T\begin{pmatrix}\lambda_R Z^TZ+\lambda_D I_k & \lambda_R Z^TX\\ \lambda_R X^TZ & \lambda_R X^TX+B\end{pmatrix}\xi\right\}\\ &\cdot\exp\left\{\xi^T\begin{pmatrix}\lambda_R Z^Ty\\ \lambda_R X^Ty + B\beta_0\end{pmatrix}\right\}\end{aligned}$$

and setting $X^TZ=0$ gives the unnormalized $N_{k+p}(\xi_0,\Sigma^{-1})$ density for ξ_0 and Σ^{-1} as defined by (5.10).

Proof of Lemma 5.5

We begin by defining some notation. Let $\pi_i(\xi|\lambda, y)$ denote the probability density corresponding to the $N_{k+p}(\xi_{i0}, \Sigma^{-1})$ distribution and define $\pi_{jl}(\lambda|\xi, y) = \pi_j(\lambda_R|\xi, y)\pi_l(\lambda_D|\xi, y)$ where $\pi_j(\lambda_R|\xi, y)$ and $\pi_l(\lambda_D|\xi, y)$ denote the probability densities corresponding to

$$\text{Gamma}\left(r_{j1} + \frac{N}{2}, r_{j2} + \frac{1}{2}v_1(\xi)\right) \quad \text{and} \quad \text{Gamma}\left(d_{l1} + \frac{k}{2}, d_{l2} + \frac{1}{2}v_2(\xi)\right) ,$$

respectively. From (5.8) it follows that

$$\pi(\xi|\lambda, y) = \sum_{i=1}^{r} \eta_i \pi_i(\xi|\lambda, y) \quad \text{and} \quad \pi(\lambda|\xi, y) = \sum_{j=1}^{s}\sum_{l=1}^{t} \phi_j \psi_l \pi_{jl}(\lambda|\xi, y) .$$

Finally, let E_i and Var_i denote expectation and variance with respect to $\pi_i(\xi|\lambda, y)$. Similarly, let E_{jl} and Var_{jl} denote expectation and variance with respect to $\pi_{jl}(\lambda|\xi, y) = \pi_j(\lambda_R|\xi, y)\pi_l(\lambda_D|\xi, y)$.

1. Case 1: $Z^T Z$ nonsingular

   For all $j \in \{1, \cdots, s\}$ and $l \in \{1, \cdots, t\}$, define

   $$c_{jl} = \max\{\delta_{j1}, \delta_{l2}\}; \text{ and}$$
   $$d_{jl} = 2r_{j2}\delta_{j1} + 2d_{l2}\delta_{l2} + \left(\frac{d_{l2}}{d_{l2} - c'}\right)^{k/2+d_{l1}} + \left(\frac{r_{j2}}{r_{j2} - c'}\right)^{N/2+r_{j1}} .$$

   Our goal is to show that

   $$\text{E}[f_1(\xi)|\lambda] \le a f_2(\lambda) + b \quad \text{and} \quad \text{E}[f_2(\lambda)|\xi] \le c f_1(\xi) + d$$

   where $a = 1$, $c = \max_{j,l} c_{j,l}$, $d = \max_{j,l} d_{j,l}$, and

   $$b = \sum_{i=1}^{N} x_i B^{-1} x_i^T + \Delta^2 .$$

Appealing to Proposition 4.1, it will suffice to establish

$$\begin{aligned} \mathrm{E}_i[f_1(\xi)|\lambda] &\leq a f_2(\lambda) + b \\ \mathrm{E}_{jl}[f_2(\lambda)|\xi] &\leq c_{jl} f_1(\xi) + d_{jl} \end{aligned} .$$

To this end, we begin by considering $\mathrm{E}_i[f_1(\xi)|\lambda] = \mathrm{E}_i[v_1(\xi) + v_2(\xi)|\lambda]$. First, for $v_1(\xi)$ and $v_2(\xi)$ as defined by (5.9), note that

$$\begin{aligned} \mathrm{E}_i[v_1(\xi)|\lambda] &= \sum_{i=1}^{N} \mathrm{E}_i\left[(y_i - x_i\beta - z_i u)^2|\lambda\right] \\ &= \sum_{i=1}^{N} \mathrm{Var}_i(y_i - x_i\beta - z_i u|\lambda) + \sum_{i=1}^{N} \left[\mathrm{E}_i(y_i - x_i\beta - z_i u|\lambda)\right]^2 \end{aligned} \tag{B.3}$$

where by Lemma B.1

$$\begin{aligned} \mathrm{Var}_i(y_i - x_i\beta - z_i u|\lambda) &= x_i \left(\lambda_R X^T X + B\right)^{-1} x_i^T + z_i \left(\lambda_R Z^T Z + \lambda_D I_k\right)^{-1} z_i^T \\ &\leq x_i B^{-1} x_i^T + \lambda_R^{-1} z_i \left(Z^T Z\right)^{-1} z_i^T \ . \end{aligned} \tag{B.4}$$

Also, let e_i denote the $k \times 1$ vector with the ith element being 1 and the rest of the elements being 0. Then

$$\mathrm{E}_i[v_2(\xi)|\lambda] = \sum_{i=1}^{k} \mathrm{E}_i(u_i^2|\lambda) = \sum_{i=1}^{k} \mathrm{Var}_i(u_i|\lambda) + \sum_{i=1}^{k} [\mathrm{E}_i(u_i|\lambda)]^2 \tag{B.5}$$

where by Lemma B.1,

$$\mathrm{Var}_i(u_i|\lambda) = e_i^T \left(\lambda_R Z^T Z + \lambda_D I_k\right)^{-1} e_i \ \leq \ \frac{1}{\lambda_D} e_i^T e_i \ = \ \frac{1}{\lambda_D} \ . \tag{B.6}$$

Combining (B.3) – (B.6) gives

$$\begin{aligned}
\mathrm{E}_i\left[f_1(\xi)\,|\,\lambda\right] &\leq \lambda_R^{-1}\sum_{i=1}^{N} z_i\left(Z^TZ\right)^{-1}z_i^T + \lambda_D^{-1}k + \sum_{i=1}^{N} x_iB^{-1}x_i^T + \tilde{\Delta}^2 \\
&\leq f_2(\lambda) + \sum_{i=1}^{N} x_iB^{-1}x_i^T + \Delta^2 \\
&= af_2(\lambda) + b\ .
\end{aligned}$$

We next consider $\mathrm{E}_{jl}[f_2(\lambda)|\xi]$. First, from (5.8) we have

$$\mathrm{E}_{jl}\left[\lambda_R^{-1}\,\middle|\,\xi\right] = \frac{2r_{j2}+v_1(\xi)}{2r_{j1}+N-2} \quad \text{and} \quad \mathrm{E}_{jl}\left[\lambda_D^{-1}\,\middle|\,\xi\right] = \frac{2d_{l2}+v_2(\xi)}{2d_{l1}+k-2}\ . \tag{B.7}$$

Also, it is straightforward to show that

$$\mathrm{E}_{jl}\left[e^{c'\lambda_D}\,\middle|\,\xi\right] \leq \left(\frac{d_{l2}}{d_{l2}-c'}\right)^{k/2+d_{l1}} \quad \text{and} \quad \mathrm{E}_{jl}\left[e^{c'\lambda_R}\,\middle|\,\xi\right] \leq \left(\frac{r_{j2}}{r_{j2}-c'}\right)^{N/2+r_{j1}}\ . \tag{B.8}$$

Combining (B.7) and (B.8) gives

$$\begin{aligned}
\mathrm{E}_{jl}\left[f_2(\lambda)\,|\,\xi\right] &= \mathrm{E}_{jl}\left[\lambda_R^{-1}\sum_{i=1}^{N} z_i(Z^TZ)^{-1}z_i^T + \lambda_D^{-1}k + \exp\{c'\lambda_D\} + \exp\{c'\lambda_R\}\,\middle|\,\xi\right] \\
&\leq \delta_{j1}(2r_{j2}+v_1(\xi)) + \delta_{l2}(2d_{l2}+v_2(\xi)) + \left(\frac{d_{l2}}{d_{l2}-c'}\right)^{k/2+d_{l1}} \\
&\quad + \left(\frac{r_{j2}}{r_{j2}-c'}\right)^{N/2+r_{j1}} \\
&\leq c_{jl}f_1(\xi) + d_{jl}\ .
\end{aligned}$$

2. Case 2: Z^TZ is possibly singular

For all $j \in \{1, \cdots, s\}$ and $l \in \{1, \cdots, t\}$, define

$$\begin{aligned} c_{jl} &= \max\{\delta_{j3}, \delta_{l4}\}; \text{ and} \\ d_{jl} &= 2r_{j2}\delta_{j3} + 2d_{l2}\delta_{l4} + \left(\frac{d_{l2}}{d_{l2} - c'}\right)^{k/2+d_{l1}} + \left(\frac{r_{j2}}{r_{j2} - c'}\right)^{N/2+r_{j1}} . \end{aligned}$$

Our goal is to show that

$$\mathrm{E}[f_1(\xi)|\lambda] \leq a f_2(\lambda) + b \quad \text{and} \quad \mathrm{E}[f_2(\lambda)|\xi] \leq c f_1(\xi) + d$$

where $a = 1$, $c = \max_{j,l} c_{j,l}$, $d = \max_{j,l} d_{j,l}$, and

$$b_i = \frac{1}{4}\sum_{i=1}^{N} x_i B^{-1} x_i^T + \Delta^2 .$$

Again, appealing to Proposition 4.1, it will suffice to establish

$$\begin{aligned} \mathrm{E}_i[f_1(\xi)|\lambda] &\leq a f_2(\lambda) + b \\ \mathrm{E}_{jl}[f_2(\lambda)|\xi] &\leq c_{jl} f_1(\xi) + d_{jl} \end{aligned} .$$

First, consider $\mathrm{E}_i[f_1(\xi)|\lambda] = \mathrm{E}_i[v_1(\xi) + v_2(\xi)|\lambda]$. From (B.3),

$$\mathrm{E}_i[v_1(\xi)|\lambda] = \sum_{i=1}^{N} \mathrm{Var}_i(y_i - x_i\beta - z_i u|\lambda) + \sum_{i=1}^{N} [\mathrm{E}_i(y_i - x_i\beta - z_i u|\lambda)]^2$$

where by Lemmas B.1 and B.2

$$\begin{aligned} \mathrm{Var}_i(y_i - x_i\beta - z_i u|\lambda) &= x_i \left(\lambda_R X^T X + B\right)^{-1} x_i^T + z_i \left(\lambda_R Z^T Z + \lambda_D I_k\right)^{-1} z_i^T \\ &\leq \frac{1}{4} x_i \left(\frac{1}{\lambda_R}\left(X^T X\right)^{-1} + B^{-1}\right) x_i^T + \frac{1}{\lambda_D} z_i z_i^T . \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{E}_i[v_1(\xi)|\lambda] \leq \lambda_R^{-1}\frac{1}{4}\sum_{i=1}^{N} x_i \left(X^T X\right)^{-1} x_i^T + \lambda_D^{-1}\sum_{i=1}^{N} z_i z_i^T + \frac{1}{4}\sum_{i=1}^{N} x_i B^{-1} x_i^T \\ + \sum_{i=1}^{N}\left[\mathrm{E}(y_i - x_i\beta - z_i u|\lambda)\right]^2 \ . \end{aligned} \tag{B.9}$$

and combining (B.5), (B.6), and (B.9) gives

$$\begin{aligned} \mathrm{E}_i\left[f_1(\xi)\,\middle|\,\lambda\right] &\leq \lambda_R^{-1}\frac{1}{4}\sum_{i=1}^{N} x_i \left(X^T X\right)^{-1} x_i^T + \lambda_D^{-1}\left[k + \sum_{i=1}^{N} z_i z_i^T\right] \\ &\quad + \frac{1}{4}\sum_{i=1}^{N} x_i B^{-1} x_i^T + \Delta^2 \\ &\leq f_2(\lambda) + \frac{1}{4}\sum_{i=1}^{N} x_i B^{-1} x_i^T + \Delta^2 \\ &= a f_2(\lambda) + b \ . \end{aligned}$$

Finally, consider $\mathrm{E}_{jl}[f_2(\lambda)|\xi]$. By (B.7) and (B.8),

$$\begin{aligned} \mathrm{E}_{jl}\left[f_2(\lambda)\,\middle|\,\xi\right] &= \mathrm{E}_{jl}\left[\lambda_R^{-1}\frac{1}{4}\sum_{i=1}^{N} x_i (X^T X)^{-1} x_i^T + \lambda_D^{-1}\left(k + \sum_{i=1}^{n} z_i z_i^T\right)\,\middle|\,\xi\right] \\ &\quad + \mathrm{E}_{jl}\left[\exp\{c'\lambda_D\} + \exp\{c'\lambda_R\}\,\middle|\,\xi\right] \\ &\leq \delta_{j3}(2r_{j2} + v_1(\xi)) + \delta_{l4}(2d_{l2} + v_2(\xi)) + \left(\frac{d_{l2}}{d_{l2} - c'}\right)^{k/2+d_{l1}} \\ &\quad + \left(\frac{r_{j2}}{r_{j2} - c'}\right)^{N/2+r_{j1}} \\ &\leq c_{jl} f_1(\xi) + d_{jl} \end{aligned}$$

and Condition 2 holds.

Proof of Proposition 5.2

Let

$$f(\lambda) = (y - X\mathrm{E}(\beta|\lambda) - Z\mathrm{E}(u|\lambda))^T(y - X\mathrm{E}(\beta|\lambda) - Z\mathrm{E}(u|\lambda)) + \mathrm{E}(u|\lambda)^T\mathrm{E}(u|\lambda)$$

and note that the claim will be proven if we can show that $f(\lambda) \leq \Delta^2$ for all λ. To this end, define functions g, and h as

$$\begin{aligned} g(\lambda) &= (y - X\mathrm{E}(\beta|\lambda))^T(y - X\mathrm{E}(\beta|\lambda)) \\ h(\lambda) &= \mathrm{E}(u|\lambda)^T Z^T Z\mathrm{E}(u|\lambda) + \mathrm{E}(u|\lambda)^T\mathrm{E}(u|\lambda) - 2y^T Z\mathrm{E}(u|\lambda)\ . \end{aligned}$$

Since the conditional independence of β and u given λ implies $X^TZ = 0$, a little algebra shows that $f(\lambda) = g(\lambda) + h(\lambda)$. Thus, it suffices to find Δ_g and Δ_h such that for all λ, $g(\lambda) \leq \Delta_g^2$ and $h(\lambda) \leq \Delta_h^2$.

We begin by bounding $g(\lambda)$. Define $A_g := \lambda_R X^T X + B$. From (5.10), setting $\beta_i = 0$ for all i gives $\mathrm{E}(\beta|\lambda) = \lambda_R A_g^{-1} X^T y$ so that

$$\begin{aligned} g(\lambda) &= y^T y + \mathrm{E}(\beta|\lambda)^T X^T X\mathrm{E}(\beta|\lambda) - 2y^T X\mathrm{E}(\beta|\lambda) \\ &= y^T y + \lambda_R^2 y^T X A_g^{-1} X^T X A_g^{-1} X^T y - 2\lambda_R y^T X A_g^{-1} X^T y \\ &= y^T y - \lambda_R y^T X A_g^{-1} B A_g^{-1} X^T y + \lambda_R y^T X A_g^{-1} A_g A_g^{-1} X^T y \\ &\quad - 2\lambda_R y^T X A_g^{-1} X^T y \\ &= y^T y - \lambda_R y^T X A_g^{-1} B A_g^{-1} X^T y - \lambda_R y^T X A_g^{-1} X^T y \\ &\leq y^T y \\ &:= \Delta_g^2 \end{aligned}$$

by the positive definiteness of B and A_g^{-1}.

Now consider $h(\lambda)$ and define $A_h := \lambda_R Z^T Z + \lambda_D I_k$. Then $\mathrm{E}(u|\lambda) = \lambda_R A_h^{-1} Z^T y$

and

$$\begin{aligned} h(\lambda) &= \lambda_R^2 y^T Z A_h^{-1} Z^T Z A_h^{-1} Z^T y + \lambda_R^2 y^T Z A_h^{-2} Z^T y - 2\lambda_R y^T Z A_h^{-1} Z^T y \\ &= \lambda_R y^T Z A_h^{-1} A_h A_h^{-1} Z^T y - \lambda_R \lambda_D y^T Z A_h^{-2} Z^T y + \lambda_R^2 y^T Z A_h^{-2} Z^T y \\ &\quad - 2\lambda_R y^T Z A_h^{-1} Z^T y \\ &= (\lambda_R^2 - \lambda_R \lambda_D) y^T Z A_h^{-2} Z^T y - \lambda_R y^T Z A_h^{-1} Z^T y. \end{aligned}$$

Since A_h^{-1} and A_h^{-2} are positive semidefinite we have

$$\begin{aligned} h(\lambda) &\leq \lambda_R^2 y^T Z A_h^{-2} Z^T y \\ &= \lambda_R^2 y^T Z \left(\left(\lambda_R Z^T Z\right)^2 + \lambda_D \left(2\lambda_R Z^T Z + \lambda_D I_k\right) \right)^{-1} Z^T y \\ &\leq \lambda_R^2 y^T Z (\lambda_R Z^T Z)^{-2} Z^T y \\ &= y^T Z (Z^T Z)^{-2} Z^T y \\ &:= \Delta_h^2 \end{aligned}$$

where the last inequality holds by Lemma B.1. The result now follows by setting $\Delta^2 = \Delta_g^2 + \Delta_h^2$.

Proof of Lemma 5.6

We begin by clarifying some notation. If $Z^T Z$ is nonsingular, define $\tilde{a}$, $\tilde{b}$ by (5.15). If $Z^T Z$ is possibly singular, define $\tilde{a}$, $\tilde{b}$ by (5.16). Then making no assumptions about $Z^T Z$, f_1,f_2 from (5.13) and (5.14) can both be written as

$$\begin{aligned} f_1(\xi) &= v_1(\xi) + v_2(\xi) \\ f_2(\lambda) &= \lambda_R^{-1} \tilde{a} + \lambda_D^{-1} \tilde{b} + \exp\{c'\lambda_D\} + \exp\{c'\lambda_R\} \, . \end{aligned}$$

Our goal is to establish that

$$\inf_{(\lambda',\xi') \in D_2} \pi(\xi|\lambda', y) \geq g_1(\xi) \quad \text{and} \quad \inf_{(\lambda',\xi') \in D_1} \pi(\lambda|\xi', y) \geq g_2(\lambda)$$

for $D_1 = \{(\xi, \lambda) : \ f_1(\xi) \leq \nu_1\}$, $D_2 = \{(\xi, \lambda) : \ f_2(\lambda) \leq \nu_2\}$. From Proposition 4.2, it will suffice to establish that for all $i \in \{1, \cdots, r\}$, $j \in \{1, \cdots, s\}$, and $l \in \{1, \cdots, t\}$

$$\inf_{(\lambda', \xi') \in D_2} \pi_i(\xi | \lambda', y) \geq g_{i1}(\xi) \quad \text{and} \quad \inf_{(\lambda', \xi') \in D_1} \pi_{jl}(\lambda | \xi', y) \geq g_{jl2}(\lambda)$$

where π_i and π_{jl} are defined in the proof of Lemma 5.5.

1. Proof of $\inf_{(\lambda', \xi') \in D_2} \pi_i(\xi | \lambda', y) \geq g_{i1}(\xi)$

   Define sets

   $$C_1 = \left\{(\xi, \lambda) : \frac{\tilde{a}}{\nu_2} \leq \lambda_R \leq \frac{\log \nu_2}{c'}\right\} \quad \text{and} \quad C_2 = \left\{(\xi, \lambda) : \frac{\tilde{b}}{\nu_2} \leq \lambda_D \leq \frac{\log \nu_2}{c'}\right\}$$

   and notice that $D_2 \subset C_1 \cap C_2$ since

   $$\begin{aligned} D_2 &= \{(\xi, \lambda) : f_2(\lambda) \leq \nu_2\} \\ &= \left\{(\xi, \lambda) : \lambda_R^{-1}\tilde{a} + \lambda_D^{-1}\tilde{b} + \exp\{c'\lambda_D\} + \exp\{c'\lambda_R\} \leq \nu_2\right\} . \end{aligned}$$

   Also, the restrictions placed on ν_2 guarantee $C_1 \cap C_2$ is nonempty. Therefore,

   $$\begin{aligned} \inf_{(\lambda', \xi') \in D_2} \pi_i(\xi | \lambda', y) &\geq \inf_{(\lambda', \xi') \in C_1 \cap C_2} \pi_i(\xi | \lambda', y) \\ &= \inf_{(\lambda', \xi') \in C_1 \cap C_2} \pi_i(\beta | \lambda', y) \pi_i(u | \lambda', y) \\ &\geq \left[\inf_{(\lambda', \xi') \in C_1 \cap C_2} \pi_i(\beta | \lambda', y)\right] \left[\inf_{(\lambda', \xi') \in C_1 \cap C_2} \pi_i(u | \lambda', y)\right] . \end{aligned} \tag{B.10}$$

   First, consider $\inf_{(\lambda', \xi') \in C_1 \cap C_2} \pi_i(u | \lambda', y)$. From (5.8),

   $$u | \lambda', y \sim \mathrm{N}_k\left(\mu_u, \Sigma_u^{-1}\right)$$

   where

   $$\Sigma_u = \lambda_R' Z^T Z + \lambda_D' I_k \quad \text{and} \quad \mu_u = \lambda_R' \Sigma_u^{-1} Z^T y .$$

Then for $(\lambda', \xi') \in C_1 \bigcap C_2$,

$$\begin{aligned}
(u-\mu_u)^T \Sigma_u (u-\mu_u) &= u^T\Sigma_u u + \mu_u^T\Sigma_u\mu_u - 2u^T\Sigma_u\mu_u \\
&= u^T\Sigma_u u + \lambda_R'^2 y^T Z\Sigma_u^{-1}Z^T y - 2\lambda_R' u^T Z^T y \\
&= u^T(\lambda_R' Z^T Z + \lambda_D' I_k)u \\
&\quad + \lambda_R'^2 y^T Z\left(\lambda_R' Z^T Z + \lambda_D' I_k\right)^{-1} Z^T y - 2\lambda_R' u^T Z^T y \\
&\leq u^T(\lambda_R' Z^T Z + \lambda_D' I_k)u + \frac{\lambda_R'^2}{\lambda_D'} y^T Z Z^T y \\
&\quad - 2\lambda_R' u^T Z^T y \\
&\leq f(u)
\end{aligned} \tag{B.11}$$

where the first inequality holds by Lemma B.1. Also, it follows from the fact that Z^TZ and I_k are positive definite, symmetric matrices that

$$|\Sigma_u| = |\lambda_R' Z^T Z + \lambda_D' I_k| \geq |\lambda_R' Z^T Z| + |\lambda_D' I_k| \geq |\lambda_D' I_k| \geq \left(\frac{\tilde{b}}{\nu_2}\right)^k \tag{B.12}$$

for $(\lambda', \xi') \in C_1 \bigcap C_2$. Since

$$\pi_i(u|\lambda', y) = (2\pi)^{-k/2}|\Sigma_u|^{1/2}\exp\left\{-\frac{1}{2}(u-\mu_u)^T\Sigma_u(u-\mu_u)\right\}$$

we have from (B.11) and (B.12) that

$$\inf_{(\lambda',\xi')\in C_1\bigcap C_2} \pi_i(u|\lambda', y) \geq h_1(u)\ . \tag{B.13}$$

Next, consider $\inf_{(\lambda',\xi')\in C_1\cap C_2}\pi_i(\beta|\lambda', y)$. The argument here is similar to the one above. From (5.8),

$$\beta|\lambda', y \sim \mathrm{N}_p\left(\mu_{i\beta}, \Sigma_\beta^{-1}\right)$$

where

$$\Sigma_\beta = \lambda_R' X^T X + B \quad \text{and} \quad \mu_{i\beta} = \Sigma_\beta^{-1}\left(\lambda_R' X^T y + B\beta_i\right)\ .$$

Notice that for $(\lambda', \xi') \in C_1 \bigcap C_2$,

$$
\begin{aligned}
&(\beta - \mu_{i\beta})^T \Sigma_\beta (\beta - \mu_{i\beta}) \\
&= \beta^T \Sigma_\beta \beta + \mu_{i\beta}^T \Sigma_\beta \mu_{i\beta} - 2\beta^T \Sigma_\beta \mu_{i\beta} \\
&= \beta^T \Sigma_\beta \beta + \left(\lambda_R' X^T y + B\beta_i\right)^T \Sigma_\beta^{-1} \left(\lambda_R' X^T y + B\beta_i\right) - 2\beta^T \left(\lambda_R' X^T y + B\beta_i\right) \\
&\leq \beta^T \Sigma_\beta \beta + \frac{1}{4} \left(\lambda_R' X^T y + B\beta_i\right)^T \left(\frac{1}{\lambda_R'} \left(X^T X\right)^{-1} + B^{-1}\right) \left(\lambda_R' X^T y + B\beta_i\right) \\
&\quad - 2\beta^T \left(\lambda_R' X^T y + B\beta_i\right) \\
&= \lambda_R'^2 \left[\frac{1}{4} y^T X B^{-1} X^T y\right] + \frac{1}{\lambda_R'} \left[\frac{1}{4} \beta_i^T B \left(X^T X\right)^{-1} B\beta_i\right] \\
&\quad + \lambda_R' \left[\beta^T X^T X \beta + \frac{1}{4} y^T X \left(X^T X\right)^{-1} X^T y + \frac{1}{2} \beta_i^T X^T y - 2\beta^T X^T y\right] \\
&\quad + \left[\beta^T B \beta - 2\beta^T B \beta_i + \frac{1}{4} \beta_i^T B \beta_i + \frac{1}{2} y^T X \left(X^T X\right)^{-1} B\beta_i\right] \\
&\leq g_i(\beta)
\end{aligned}
\tag{B.14}
$$

where the first inequality holds by Lemma B.2. Also, $X^T X$ and B are positive definite, symmetric matrices so that

$$
|\Sigma_\beta| = |\lambda_R' X^T X + B| \geq |\lambda_R' X^T X| + |B| \geq |B| \,. \tag{B.15}
$$

Since

$$
\pi_i(\beta|\lambda', y) = (2\pi)^{-p/2} |\Sigma_\beta|^{1/2} \exp\left\{-\frac{1}{2} (\beta - \mu_{i\beta})^T \Sigma_\beta (\beta - \mu_{i\beta})\right\}
$$

we have from (B.14) and (B.15) that

$$
\inf_{(\lambda', \xi') \in C_1 \bigcap C_2} \pi_i(\beta|\lambda', y) \geq h_{i2}(\beta) \,. \tag{B.16}
$$

Combining (B.10), (B.13), and (B.16) establishes

$$\inf_{(\lambda',\xi')\in D_2} \pi_i\left(\xi|\lambda', y\right) \geq h_1(u)h_{i2}(\beta) = g_{i1}(\xi) \ .$$

2. Proof of $\inf_{(\lambda',\xi')\in D_1} \pi_{jl}(\lambda|\xi', y) \geq g_{jl2}(\lambda)$

Define set $C_i = \{(\lambda, \xi) : v_i(\xi) \leq \nu_1\}$ for $i = 1, 2$ and notice that $D_1 \subset C_1 \cap C_2$ since

$$D_1 = \{(\lambda, \xi) : f_1(\xi) \leq \nu_1\} = \{(\lambda, \xi) : v_1(\xi) + v_2(\xi) \leq \nu_1\} \ .$$

Therefore it suffices to establish the result on $C_1 \cap C_2$ since $C_1 \cap C_2$ is nonempty. First, consider the following result due to Jones and Hobert (2004) that will be useful in establishing the result.

Lemma B.3. *Jones and Hobert (2004)*

Let $Gamma\,(\alpha, \beta; x)$ *denote the value of the* $Gamma(\alpha, \beta)$ *density at the point* $x > 0$. *If* $\alpha > 1$, $b > 0$, *and* $c > 0$ *are fixed, then, as a function of* x

$$\inf_{0<\beta<c} Gamma\,(\alpha, b + \beta/2; x) = \begin{cases} Gamma\,(\alpha, b; x) & \text{if } x < x^* \\ Gamma\,(\alpha, b + c/2; x) & \text{if } x > x^* \end{cases}$$

where

$$x^* = \frac{2\alpha}{c} \log\left(1 + \frac{c}{2b}\right) \ .$$

Recall that

$$\pi_{jl}(\lambda_R|\xi', y) = \text{Gamma}\left(r_{j1} + \frac{N}{2}, r_{j2} + \frac{v_1(\xi')}{2}; \lambda_R\right)$$
$$\pi_{jl}(\lambda_D|\xi', y) = \text{Gamma}\left(d_{l1} + \frac{k}{2}, d_{l2} + \frac{v_2(\xi')}{2}; \lambda_D\right) \ .$$

Then by Lemma B.3

$$\inf_{(\lambda',\xi')\in C_1} \pi_{jl}(\lambda_R|\xi', y) \geq \inf_{\xi':v_1(\xi')\leq\nu_1} \pi_{jl}(\lambda_R|\xi', y) = h_{j1}(\lambda_R)$$

and

$$\inf_{(\lambda',\xi')\in C_2} \pi_{jl}(\lambda_D|\xi',y) \geq \inf_{\xi':v_2(\xi')\leq\nu_1} \pi_{jl}(\lambda_D|\xi',y) = h_{l2}(\lambda_D) \; .$$

It follows that

$$\begin{aligned}
\inf_{(\lambda',\xi')\in D_1} \pi_{jl}(\lambda|\xi',y) &\geq \inf_{(\lambda',\xi')\in C_1\cap C_2} \pi_{jl}(\lambda|\xi',y) \\
&= \inf_{(\lambda',\xi')\in C_1\cap C_2} \pi_{jl}(\lambda_R|\xi',y)\pi_{jl}(\lambda_D|\xi',y) \\
&\geq \left[\inf_{(\lambda',\xi')\in C_1\cap C_2} \pi_{jl}(\lambda_R|\xi',y)\right]\left[\inf_{(\lambda',\xi')\in C_1\cap C_2} \pi_{jl}(\lambda_D|\xi',y)\right] \\
&\geq \left[\inf_{(\lambda',\xi')\in C_1} \pi_{jl}(\lambda_R|\xi',y)\right]\left[\inf_{(\lambda',\xi')\in C_2} \pi_{jl}(\lambda_D|\xi',y)\right] \\
&\geq h_{j1}(\lambda_R)h_{l2}(\lambda_D) \\
&= g_{jl2}(\lambda) \; .
\end{aligned}$$

Proof of Proposition 5.3

The block Gibbs sampler corresponding to (ξ,λ) has transition density

$$k_2((\xi',\lambda'),(\xi,\lambda)) = \pi(\xi|\lambda',y)\pi(\lambda|\xi,y) = \pi(\beta|\lambda',y)\pi(u|\lambda',y)\pi(\lambda|\xi,y)$$

where for $\beta\in\mathbb{M}_\beta$ and $u\in\mathbb{M}_u$

$$\begin{aligned}
k_2((\xi',\lambda'),(\xi,\lambda)) &= \frac{\pi(\beta|\lambda',y)\pi(u|\lambda',y)}{\pi(\beta|\tilde{\lambda},y)\pi(u|\tilde{\lambda},y)}\pi(\beta|\tilde{\lambda},y)\pi(u|\tilde{\lambda},y)\pi(\lambda|\xi,y) \\
&\geq \left[\inf_{\beta\in\mathbb{M}_\beta}\frac{\pi(\beta|\lambda')}{\pi(\beta|\tilde{\lambda})}\right]\left[\inf_{u\in\mathbb{M}_u}\frac{\pi(u|\lambda')}{\pi(u|\tilde{\lambda})}\right]\pi(\beta|\tilde{\lambda},y)\pi(u|\tilde{\lambda},y)\pi(\lambda|\xi,y) \; .
\end{aligned}$$

Therefore, the minorization condition will follow directly from establishing

$$\inf_{\beta\in\mathbb{M}_\beta}\frac{\pi(\beta|\lambda')}{\pi(\beta|\tilde{\lambda})} \geq g_1(\lambda',\tilde{\lambda}) \quad \text{and} \quad \inf_{u\in\mathbb{M}_u}\frac{\pi(u|\lambda')}{\pi(u|\tilde{\lambda})} \geq g_2(\lambda',\tilde{\lambda}) \; .$$

First,

$$\beta|\lambda,y \sim \mathrm{N}_k\left(\mu_\beta(\lambda),\Sigma_\beta(\lambda)^{-1}\right)$$

gives

$$\begin{aligned}
\frac{\pi(\beta|\lambda')}{\pi(\beta|\tilde{\lambda})} &= \frac{|\Sigma_\beta(\lambda')|^{1/2}}{|\Sigma_\beta(\tilde{\lambda})|^{1/2}} \cdot \frac{\exp\left\{-\frac{1}{2}(\beta-\mu_\beta(\lambda'))^T\Sigma_\beta(\lambda')(\beta-\mu_\beta(\lambda'))\right\}}{\exp\left\{-\frac{1}{2}(\beta-\mu_\beta(\tilde{\lambda}))^T\Sigma_\beta(\tilde{\lambda})(\beta-\mu_\beta(\tilde{\lambda}))\right\}} \\
&= \frac{|\Sigma_\beta(\lambda')|^{1/2}}{|\Sigma_\beta(\tilde{\lambda})|^{1/2}} \cdot \frac{\exp\left\{-\frac{1}{2}\mu_\beta(\lambda')^T\Sigma_\beta(\lambda')\mu_\beta(\lambda')\right\}}{\exp\left\{-\frac{1}{2}\mu_\beta(\tilde{\lambda})^T\Sigma_\beta(\tilde{\lambda})\mu_\beta(\tilde{\lambda})\right\}} \\
&\quad \cdot \frac{\exp\left\{-\frac{1}{2}\left[\beta^T(\lambda'_R X^T X + B)\beta - 2\beta^T(\lambda'_R X^T y + B\beta_0)\right]\right\}}{\exp\left\{-\frac{1}{2}\left[\beta^T(\tilde{\lambda}_R X^T X + B)\beta - 2\beta^T(\tilde{\lambda}_R X^T y + B\beta_0)\right]\right\}} \\
&= g(\lambda', \tilde{\lambda}) \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2} y^T y\right\} \\
&\quad \cdot \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2}\left(\beta^T X^T X \beta - 2\beta^T X^T y\right)\right\} \\
&= g(\lambda', \tilde{\lambda}) \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2}(y - X\beta)^T(y - X\beta)\right\} .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\inf_{\beta\in\mathbb{M}_\beta} \frac{\pi(\beta|\lambda')}{\pi(\beta|\tilde{\lambda})} &= g(\lambda', \tilde{\lambda}) \inf_{\beta\in\mathbb{M}_\beta} \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2}(y - X\beta)^T(y - X\beta)\right\} \\
&= g(\lambda', \tilde{\lambda}) \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2}(y - X\breve{\beta})^T(y - X\breve{\beta})\right\} \\
&= g_1(\lambda', \tilde{\lambda}) .
\end{aligned}$$

Similarly, since

$$u|\lambda, y \sim \mathrm{N}_k\left(\mu_u(\lambda), \Sigma_u(\lambda)^{-1}\right)$$

we have

$$\begin{aligned}
\frac{\pi(u|\lambda')}{\pi(u|\tilde{\lambda})} &= \frac{|\Sigma_u(\lambda')|^{1/2}}{|\Sigma_u(\tilde{\lambda})|^{1/2}} \cdot \frac{\exp\left\{-\frac{1}{2}(u-\mu_u(\lambda'))^T\Sigma_u(\lambda')(u-\mu_u(\lambda'))\right\}}{\exp\left\{-\frac{1}{2}(u-\mu_u(\tilde{\lambda}))^T\Sigma_u(\tilde{\lambda})(u-\mu_u(\tilde{\lambda}))\right\}} \\
&= \frac{|\Sigma_u(\lambda')|^{1/2}}{|\Sigma_u(\tilde{\lambda})|^{1/2}} \cdot \frac{\exp\left\{-\frac{1}{2}\mu_u(\lambda')^T\Sigma_u(\lambda')\mu_u(\lambda')\right\}}{\exp\left\{-\frac{1}{2}\mu_u(\tilde{\lambda})^T\Sigma_u(\tilde{\lambda})\mu_u(\tilde{\lambda})\right\}} \\
&\quad \cdot \frac{\exp\left\{-\frac{1}{2}\left[u^T(\lambda'_R Z^T Z + \lambda'_D I_k)u - 2\lambda'_R u^T Z^T y\right]\right\}}{\exp\left\{-\frac{1}{2}\left[u^T(\tilde{\lambda}_R Z^T Z + \tilde{\lambda}_D I_k)u - 2\tilde{\lambda}_R u^T Z^T y\right]\right\}} \\
&= h(\lambda', \tilde{\lambda}) \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2} y^T y\right\} \exp\left\{-\frac{(\lambda'_D - \tilde{\lambda}_D)}{2} u^T u\right\} \\
&\quad \cdot \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2}\left(u^T Z^T Z u - 2u^T Z^T y\right)\right\} \\
&= h(\lambda', \tilde{\lambda}) \exp\left\{-\frac{(\lambda'_D - \tilde{\lambda}_D)}{2} u^T u\right\} \\
&\quad \cdot \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2}(y - Zu)^T(y - Zu)\right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\inf_{u\in\mathbb{M}_u} \frac{\pi(u|\lambda')}{\pi(u|\tilde{\lambda})} &\geq h(\lambda', \tilde{\lambda}) \inf_{u\in\mathbb{M}_u} \exp\left\{-\frac{(\lambda'_D - \tilde{\lambda}_D)}{2} u^T u\right\} \\
&\qquad \cdot \inf_{u\in\mathbb{M}_u} \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2}(y - Zu)^T(y - Zu)\right\}. \\
&= h(\lambda', \tilde{\lambda}) \exp\left\{-\frac{(\lambda'_D - \tilde{\lambda}_D)}{2} \sum_{i=1}^{k} \hat{u}_i^2\right\} \\
&\qquad \cdot \exp\left\{-\frac{(\lambda'_R - \tilde{\lambda}_R)}{2}(y - Z\breve{u})^T(y - Z\breve{u})\right\} \\
&= g_2(\lambda', \tilde{\lambda})
\end{aligned}$$