CONVERGENCE RATES AND REGENERATION OF THE BLOCK GIBBS SAMPLER FOR BAYESIAN RANDOM EFFECTS MODELS

By

AIXIN TAN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2009

To my parents

ACKNOWLEDGMENTS

First of all, I must express my sincerest appreciation to Dr. James Hobert, who is one of the most wonderful people I have ever met. I feel very lucky to get to know him, let alone having him as my advisor and friend. Among many other things, I thank Dr. Hobert for his high quality mentoring. He always generously shares his ideas and makes great effort to explain them in the clearest way possible — it means so much to me that I can understand and trust what he says. Yet, he is always willing to listen to my random ideas and provide the most informative and encouraging feedback. It has been my greatest pleasure to work with an outstanding researcher and a considerate person like him.

Next, I would like to thank everyone on my supervisory committee. I am grateful to Dr. Hani Doss for his awesome teaching. His hard working attitude very much influenced me. Even his random walk around the building motivates me to work longer hours at my office. I thank Dr. Casella for giving me valuable suggestions and great help on my dissertation. I appreciate all of his advice and encouragement. I thank Dr. Michael Jury and Dr. Scott McCullough from the Department of Mathematics for their help and precious time.

I would also like to thank the faculty at the Department of Statistics who taught me a lot of things, both in and out of the classroom. Special thanks go to Dr. Malay Ghosh, who taught me four different classes (most of which started at 7:25 am). I am in awe of his charm and wisdom and his everlasting enthusiasm and energy in solving mathematical and statistical problems.

I thank Eugenia Buta, Nan Feng, Vikneswaran Gopal, Qin Li, Yao Li, Tezcan Ozrazgat, George Papageorgiou, Vivekananda Roy, Oleksandr Savenkov, Doug Sparks, Jihyun Song, Rebecca Steorts, Chenguang Wang, John Yap, Li Zhang and many others for their friendship. I also thank Ruitao Liu for making every day of the past five years full of excitement and laughter.

Last but not least, I thank my father for teaching me the value of hard working and perseverance. I also thank my mother who is responsible for the cheerful and optimistic side of my personality. I could not have truly enjoyed my studying and my life without these qualities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

CONVERGENCE RATES AND REGENERATION OF THE BLOCK GIBBS SAMPLER
FOR BAYESIAN RANDOM EFFECTS MODELS

By

Aixin Tan

August 2009

Chair: James P. Hobert
Major: Statistics

Markov chain Monte Carlo (MCMC) methods have received considerable attention as powerful computing tools in Bayesian statistical analysis. The idea is to produce Markov chain samples to estimate characteristics of complicated Bayesian posterior distributions. We consider the widely applicable Bayesian one-way random effects model. If the standard diffuse prior is used, there is a simple block Gibbs sampler that can be employed to explore the intractable posterior distribution, π. Indeed, the sampler produces a Markov chain output Φ that has invariant distribution π. Then the sample averages of Φ are used to estimate expectations with respect to π. Consider specifying a different prior, such as the reference prior, for the model. This results in a more complex posterior distribution, π^*. Constructing an MCMC algorithm for π^* is much harder than that for π. So we resort to the importance sampling technique, which uses weighted averages of Φ to estimate expectations with respect to π^*.

Basic Markov chain theory implies that both types of MCMC estimators mentioned above are valid in the following sense. They converge almost surely to their respective estimands as the Markov chain sample size grows to infinity. Nevertheless, it is always important to ask what is an appropriate sample size when running an MCMC algorithm. To answer this question for the above block Gibbs sampler, we develop a regenerative simulation method that yields simple, asymptotically valid Monte Carlo standard errors. These standard errors can be used to construct confidence intervals for the quantities

of interest. Then one method to determine when to stop the Markov chain is to run the simulation until the width of the intervals are below user-specified values. The regenerative method rests on the assumption that the underlying Markov chain converges to its stationary distribution at a geometric rate. We use the drift method to show that, unless the data set is extremely small and unbalanced, the block Gibbs Markov chain is geometrically ergodic. We illustrate the use of the regenerative method with data from a styrene exposure study.

In the final chapter of this dissertation, we discuss a slightly different topic. We study the convergence rate of Markov chains underlying a simple class of two-variable Gibbs samplers. These Markov chains live in a common state space that constitutes points on the diagonal and the subdiagonal of $\mathbb{N} \times \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers. It is shown that a geometric tail decay of the target distribution is almost necessary and sufficient for the corresponding chain to be geometrically ergodic. The argument involves indirect evaluation of the norm of Markov chain operators.

CHAPTER 1
INTRODUCTION

1.1 Bayesian Inference and Intractable Integrals

Bayesian statistical inference has become a well accepted way of analyzing data. Having specified the model and the prior, making inference about the parameters amounts to understanding the posterior distribution, mostly through calculating posterior means of various functions. For posteriors that arise in realistic statistical problems, such calculations involve performing high dimensional integration that can not be done analytically. The Monte Carlo method is a simulation-based method that evaluates these integrals indirectly. The idea is to generate independent and identically distributed (iid) samples from the posterior and use the sample averages to estimate the posterior means. In situations where obtaining iid samples from the posterior is impossible, it may still be easy to run a Markov chain that has the posterior as its invariant distribution (Liu, 2001; Robert and Casella, 2004). Ergodic averages of the Markov chain can then be used to estimate the posterior means. This latter approach is called the Markov chain Monte Carlo (MCMC) method.

Basic Markov chain theory implies that, under simple regularity conditions, MCMC estimators converge almost surely to the quantities of interest as the Markov chain sample size grows to infinity. Nevertheless, any estimator that we use in practice is based on a finite number of samples. Hence, there is always a Monte Carlo error associated with it. We never know the exact value of the error, but we can study its sampling distribution. This knowledge provides us a tool to measure the level of confidence we can have in the estimator. We can also use this information to decide an appropriate Markov chain sample size depending on how confident we need to be of this estimator. The main goal of this dissertation is to explain the above idea rigorously and carry it out for the following Bayesian model.

Consider the classical one-way random effects model given by

$$Y_{ij} = \theta_i + \varepsilon_{ij}\,,\ i = 1, \ldots, q\,,\ j = 1, \ldots, m_i\,,$$

where the random effects $\theta_1, \ldots, \theta_q$ are iid $\mathrm{N}(\mu, \sigma_\theta^2)$, the ε_{ij}s are iid $\mathrm{N}(0, \sigma_e^2)$ and independent of the θ_is, and $(\mu, \sigma_\theta^2, \sigma_e^2)$ is an unknown parameter. There is a long history of Bayesian analysis using this model starting with Hill (1965) and Tiao and Tan (1965). A Bayesian version of the model requires a prior distribution for $(\mu, \sigma_\theta^2, \sigma_e^2)$, call it $p\big(\mu, \sigma_\theta^2, \sigma_e^2\big)$. Actually, since the vector of random effects, $\theta = (\theta_1, \ldots, \theta_q)$, is unobserved, it too is viewed as a parameter with a "built-in" prior density given by

$$\phi(\theta|\mu, \sigma_\theta^2, \sigma_e^2) = \prod_{i=1}^{q} \big(2\pi\sigma_\theta^2\big)^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2\sigma_\theta^2}(\theta_i - \mu)^2 \Big\}\,.$$

Letting $y = \{y_{ij}\}$ denote the vector of observed data, the $(q+3)$-dimensional posterior density is characterized by

$$\pi\big(\theta, \mu, \sigma_\theta^2, \sigma_e^2\big) \propto L(\theta, \mu, \sigma_\theta^2, \sigma_e^2; y)\, \phi(\theta|\mu, \sigma_\theta^2, \sigma_e^2)\, p\big(\mu, \sigma_\theta^2, \sigma_e^2\big)\,, \tag{1–1}$$

where

$$L(\theta, \mu, \sigma_\theta^2, \sigma_e^2; y) = \prod_{i=1}^{q}\prod_{j=1}^{m} \big(2\pi\sigma_e^2\big)^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2\sigma_e^2}(y_{ij} - \theta_i)^2 \Big\}$$

is the likelihood function. Since the observed data are always conditioned upon, we have suppressed the notation of dependence on y in the left-hand side of (1–1).

We consider two types of (improper) priors for $(\mu, \sigma_\theta^2, \sigma_e^2)$. The first type has density

$$p_{a,b}\big(\mu, \sigma_\theta^2, \sigma_e^2\big) = \big(\sigma_\theta^2\big)^{-(a+1)} \big(\sigma_e^2\big)^{-(b+1)}\,, \tag{1–2}$$

where a and b are prespecified hyper-parameters. We use $\pi_{a,b}$ to denote the posterior density under $p_{a,b}$. The members of this are *conditionally conjugate* priors; that is, for each parameter, the prior and the full conditional density have the same form. For example, the prior on σ_θ^2 has an inverse-gamma form, and the corresponding full conditional density, $\pi(\sigma_\theta^2 \,|\, \sigma_e^2, \mu, \theta)$, is an inverse-gamma density. As we will see later, conditionally conjugate

priors are convenient because they allow easy MCMC sampling from the posterior. Within this family of priors, $p_{-\frac{1}{2},0}$ is a popular choice for reasons we now describe.

According to Gelman (2006), the choice of prior for (μ, σ_e^2) is not crucial since the data often contain a good deal of information about these parameters. On the other hand, there is typically relatively little information in the data concerning σ_θ^2, so the choice of prior for this parameter is more important and subtle. A commonly used prior for σ_θ^2 is a proper inverse gamma prior, which is a conditionally conjugate prior. When little or no prior information concerning σ_θ^2 is available, the (shape and scale) hyper-parameters of this prior are often set to very small values in an attempt to be "non-informative". However, in the limit, as the scale-parameter approaches 0 with the shape parameter either approaching 0 or fixed, not only does the prior become improper, but the corresponding posterior also becomes improper. Consequently, the posterior is not robust to small changes in these (somewhat arbitrarily chosen) hyper-parameters. This problem has led several authors, including Daniels (1999) and Gelman (2006), to recommend that the proper inverse gamma prior not be used. In contrast, Gelman (2006) illustrates that the improper prior $\left(\sigma_\theta^2\right)^{-\frac{1}{2}}$ works well unless q is very small (say, below 5). Combining this prior with a uniform prior on $\left(\mu, \log(\sigma_e^2)\right)$ leads to $p_{-\frac{1}{2},0}$. van Dyk and Meng (2001) call $p_{-\frac{1}{2},0}$ the *standard diffuse prior* and we will refer to it as such throughout this dissertation.

However, serious Bayesians disagree about the suitability of the standard diffuse prior. Indeed, Bernardo (1996) states that "...the use of "standard" improper power priors on the variances is a well documented case of careless prior specification ..." Bernardo goes on to recommend the so-called reference prior of Berger and Bernardo (1992), which is the second prior that we consider. When the data set is balanced, i.e., when $m_i = m$ for $i = 1, \ldots, q$, the reference prior takes the form

$$p_r\left(\mu, \sigma_\theta^2, \sigma_e^2\right) \propto \left(\sigma_\theta^2\right)^{-\frac{C_m}{2}} \left(\sigma_e^2\right)^{-1} \left[m - 1 + \left(\frac{\sigma_e^2}{\sigma_e^2 + m\sigma_\theta^2}\right)^2\right]^{\frac{1}{2}}, \tag{1–3}$$

where $C_m = 1 - \sqrt{m-1}/(\sqrt{m}+\sqrt{m-1})^3$. We use $\pi_r\big(\theta, \mu, \sigma_\theta^2, \sigma_e^2\big)$ to denote the posterior density under p_r.

Before we proceed with any of the above priors, a critical step is to check that their corresponding posteriors are genuine probability distributions. Results in Hobert and Casella (1996) show that the posterior $\pi_{a,b}$ is proper if and only if all three of the following conditions hold:

$$a < 0\,, \quad a + \frac{q}{2} > \frac{1}{2}\,, \quad \text{and} \quad a + b > \frac{1-M}{2}\,,$$

where M is the total number of observations; that is, $M = \sum_{i=1}^{q} m_i$. In particular, this implies that the standard diffuse prior yields a proper posterior if and only if $q \geq 3$. We show in Appendix A that $q \geq 3$ is also a necessary and sufficient condition for the propriety of π_r.

Making inference through the posterior distribution often reduces to computing expectations with respect to the posterior density. Unfortunately, even for the simpler prior of the two, $p_{a,b}$, the posterior density is intractable. Indeed, letting $\mathbb{R}_+ = (0, \infty)$, the posterior expectation of $g(\theta, \mu, \sigma_\theta^2, \sigma_e^2)$ is given by

$$\frac{\int_{\mathbb{R}^q}\int_{\mathbb{R}}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+} g(\theta,\mu,\sigma_\theta^2,\sigma_e^2)\, L(\theta,\mu,\sigma_\theta^2,\sigma_e^2;y)\,\phi(\theta|\mu,\sigma_\theta^2,\sigma_e^2)\, p_{a,b}\big(\mu,\sigma_\theta^2,\sigma_e^2\big)\, d\sigma_e^2\, d\sigma_\theta^2\, d\mu\, d\theta}{\int_{\mathbb{R}^q}\int_{\mathbb{R}}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+} L(\theta,\mu,\sigma_\theta^2,\sigma_e^2;y)\,\phi(\theta|\mu,\sigma_\theta^2,\sigma_e^2)\, p_{a,b}\big(\mu,\sigma_\theta^2,\sigma_e^2\big)\, d\sigma_e^2\, d\sigma_\theta^2\, d\mu\, d\theta}\,, \tag{1–4}$$

which is a ratio of two intractable integrals. Results in Section 3.1 show that it is actually possible to integrate θ and μ out of the integral in the denominator in closed form, so the denominator is really a 2-dimensional intractable integral. However, the numerator is usually an intractable integral of dimension $q + 3$.

Intractable integrals like the one in (1–4) are typical in Bayesian posterior analysis. Indeed, posterior expectations that arise from useful statistical models are often high dimensional integrals that have unbounded integration ranges. Therefore methods like analytical approximation or numerical integration are not always applicable. In the next two sections, we review the classical Monte Carlo technique and the MCMC technique

respectively. They are simulation methods that generally better suit the task of evaluating integrals involving probability distributions. Due to their broad applicability beyond the Bayesian one-way random effects model, we discuss the theory behind the Monte Carlo methods in an abstract setting.

1.2 Classical Monte Carlo Methods

Let X denote the parameter space and π the posterior density on X with respect to a measure μ. Let

$$L^1(\pi) = \left\{h : \mathsf{X} \to \mathbb{R} \text{ such that } \int_{\mathsf{X}} \left|h(x)\right|\pi(x)\,\mu(dx) < \infty\right\}$$
$$\text{and} \quad L^2(\pi) = \left\{h : \mathsf{X} \to \mathbb{R} \text{ such that } \int_{\mathsf{X}} h^2(x)\pi(x)\,\mu(dx) < \infty\right\} .$$

Also, let

$$\mathrm{E}_\pi h = \int_{\mathsf{X}} h(x)\pi(x)\,\mu(dx) .$$

Remember our goal is to explore the posterior distribution π by evaluating $\mathrm{E}_\pi g$ for some interesting $g \in L^1(\pi)$. Suppose we could make iid draws $X_0^\#, X_1^\#, \ldots$ from π, then we would estimate $\mathrm{E}_\pi g$ using the classical Monte Carlo estimator

$$\overline{g}_N^\# = \frac{1}{N}\sum_{n=0}^{N-1} g(X_n^\#) .$$

This estimator is unbiased and the strong law of large numbers (SLLN) implies that it converges almost surely to $\mathrm{E}_\pi g$; that is, it is also strongly consistent.

In practice, we need to choose the sample size, N, and this is where the central limit theorem (CLT) comes in. Indeed, if $g \in L^2(\pi)$, then there is a CLT for $\overline{g}_N^\#$; that is, as $N \to \infty$, we have

$$\sqrt{N}\left(\overline{g}_N^\# - \mathrm{E}_\pi g\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \omega^2) ,$$

where $\omega^2 = \mathrm{E}_\pi g^2 - \big(\mathrm{E}_\pi g\big)^2$. Thus, in practice we could choose a preliminary value of N, say N', draw a random sample of size N' from π and compute $\overline{g}^{\#}_{N'}$ and

$$\hat{\omega}^2 = \frac{1}{N'-1} \sum_{n=0}^{N'-1} \big(g(X_n^{\#}) - \overline{g}^{\#}_{N'}\big)^2 \, .$$

Of course, $\hat{\omega}^2$ is a strongly consistent estimator of ω^2. These quantities could then be used to assemble the asymptotic 95% confidence interval (CI) for $\mathrm{E}_\pi g$ given by $\overline{g}^{\#}_{N'} \pm 2\hat{\omega}/\sqrt{N'}$. If we are satisfied with the width of this interval, we stop, whereas if the width is deemed too large, we simply increase the sample size to a level that will ensure an acceptably small standard error. The main message here is that routine use of the CLT allows for straightforward determination of an appropriate Monte Carlo sample size.

Often, we would like to compare the effect of using different priors and/or models on the same dataset. Therefore, we need to obtain expectations of functions with respect to several different posteriors, probably all of which are intractable integrals. Suppose a second posterior that we want to investigate has density π^* with respect to μ on X. Define $S = \big\{x \in \mathsf{X} : \pi(x) > 0\big\}$ and $S^* = \big\{x \in \mathsf{X} : \pi^*(x) > 0\big\}$ and assume that $S^* \subset S$. If we desire an estimate of the intractable expectation $\eta := \mathrm{E}_{\pi^*} g = \int_{\mathsf{X}} g(x)\pi^*(x)\,\mu(dx)$ for some $g \in L^1(\pi^*)$, then instead of obtaining new samples from π^*, we can reuse the iid samples $X_0^{\#}, X_1^{\#}, \ldots$ from π and apply the importance sampling technique. The SLLN implies that

$$\frac{1}{N} \sum_{n=0}^{N-1} g(X_i^{\#}) \frac{\pi^*(X_i^{\#})}{\pi(X_i^{\#})}$$

is a strongly consistent estimator of η. However, if π or π^* is known only up to a normalizing constant, as is typically the case in practice, then this estimator is not computable. Fortunately, there is a simple way to circumvent this difficulty. Assume that $\pi(x) = cf(x)$ and $\pi^*(x) = c^* f^*(x)$, where c and c^* are unknown normalizing constants, and f and f^* are known functions. The standard importance sampling identity that recasts

$\mathrm{E}_{\pi^*} g$ as a ratio of expectations with respect to π is as follows:

$$\begin{aligned}
\mathrm{E}_{\pi^*} g &= \int_{\mathsf{X}} g(x) \frac{\pi^*(x)}{\pi(x)} \pi(x)\, \mu(dx) \Big/ \int_{\mathsf{X}} \frac{\pi^*(x)}{\pi(x)} \pi(x)\, \mu(dx) \\
&= \int_{\mathsf{X}} g(x) \frac{f^*(x)}{f(x)} \pi(x)\, \mu(dx) \Big/ \int_{\mathsf{X}} \frac{f^*(x)}{f(x)} \pi(x)\, \mu(dx) \\
&= \frac{\mathrm{E}_\pi v}{\mathrm{E}_\pi u}\,, \qquad (1\text{–}5)
\end{aligned}$$

where

$$v(x) := g(x) \frac{f^*(x)}{f(x)} \quad \text{and} \quad u(x) := \frac{f^*(x)}{f(x)}\,. \qquad (1\text{–}6)$$

The representation in (1–5) suggests estimating η with

$$\hat{\eta}_N^{\#} = \frac{\overline{v}_N^{\#}}{\overline{u}_N^{\#}}$$

where

$$\overline{v}_N^{\#} := \frac{1}{N} \sum_{n=0}^{N-1} v(X_n^{\#}) \quad \text{and} \quad \overline{u}_N^{\#} := \frac{1}{N} \sum_{n=0}^{N-1} u(X_n^{\#})\,.$$

Again, the SLLN implies that $\hat{\eta}_N^{\#}$ is a strongly consistent estimator of η, and this fact justifies the standard procedure in which we choose some (finite) value of N, simulate a random sample of size N from π, and use the observed value of $\hat{\eta}_N^{\#}$ as an estimate of η.

The procedure of choosing a large enough sample size for $\hat{\eta}_N$ is similar to that for $\overline{g}_N^{\#}$. Again, the key is to obtain the standard error of the estimator. Simple asymptotic arguments show that, if $\mathrm{E}_\pi v^2$ and $\mathrm{E}_\pi u^2$ are both finite, then as $N \to \infty$,

$$\sqrt{N}\big(\hat{\eta}_N^{\#} - \eta\big) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \kappa^2)\,,$$

where

$$\kappa^2 = \frac{\mathrm{E}_\pi\big[(v - u\eta)^2\big]}{\big[\mathrm{E}_\pi u\big]^2}\,.$$

See, for example, Robert and Casella (2004, Lemma 4.3). Moreover, a simple consistent estimator of κ^2 is given by

$$\hat{\kappa}^2 = \frac{1}{N}\sum_{n=0}^{N-1}\left[\left(v(X_n^{\#}) - u(X_n^{\#})\,\hat{\eta}_N^{\#}\right)^2\right] \Bigg/ \left[\frac{1}{N}\sum_{n=0}^{N-1} u(X_n^{\#})\right]^2 .$$

The standard error of the estimate is the observed value of $\hat{\kappa}/\sqrt{N}$.

1.3 MCMC Methods

In reality it is rare that we can make iid draws from the posterior. For example, we are not aware of any method that produces iid random variables from the $q+3$ dimensional posterior distribution $\pi_{a,b}$ or π_r from Section 1.1. Therefore, to approximate the posterior expectation in (1–4), we use MCMC instead of classical Monte Carlo. The MCMC method applies the strategy outlined in Section 1.2 with a Markov chain in place of the iid sample. And unlike iid sampling, there are always recipes, such as Metropolis-Hastings algorithms and Gibbs samplers, that one can follow to construct a workable Markov chain sample. (By workable, we mean that the Markov chain is Harris ergodic with limiting distribution the posterior distribution. See Chapter 2 for details.) To explore $\pi_{a,b}$, there are at least two Gibbs samplers that we can employ.

The *simple* Gibbs sampler cycles through the $q+3$ components of the vector $(\theta_1, \ldots, \theta_q, \mu, \sigma_\theta^2, \sigma_e^2)$ one at a time and samples each one conditional on the most current values of the other $q+2$ components. This algorithm was first introduced in the seminal paper by Gelfand and Smith (1990) as an application of the Gibbs sampler to a Bayesian version of the one-way model with *proper* conjugate priors.

In this dissertation, we study a *block* Gibbs sampler whose iterations have just two steps. Let $\sigma^2 = \left(\sigma_\theta^2, \sigma_e^2\right)$, $\xi = (\mu, \theta)$ and suppose that the state of the chain at time n is $X_n = \left(\sigma_n^2, \xi_n\right)$. One iteration of our sampler entails drawing σ_{n+1}^2 conditional on ξ_n, and then drawing ξ_{n+1} conditional on σ_{n+1}^2. "Blocking" variables together in this way and doing multivariate updates often leads to improved convergence properties relative to the simple (univariate) version of the Gibbs sampler (see, e.g., Liu et al., 1994). Formally, the

Markov transition density for the update $\left(\sigma_n^2, \xi_n\right) \to \left(\sigma_{n+1}^2, \xi_{n+1}\right)$ is given by

$$k\left(\sigma_{n+1}^2, \xi_{n+1} \mid \sigma_n^2, \xi_n\right) = \pi\left(\sigma_{n+1}^2 \mid \xi_n\right) \pi\left(\xi_{n+1} \mid \sigma_{n+1}^2\right) .$$

(Recall that we are suppressing dependence on the data y.) Straightforward manipulation of (1–1) shows that, given ξ, σ_θ^2 and σ_e^2 are independent random variables each with inverse-gamma distributions, and given σ^2, ξ is multivariate normal. (The specific forms of these distributions are given in Section 3.1). Hence it is easy to program the block Gibbs sampler.

Let $\left\{X_n\right\}_{n=0}^{\infty} = \left\{\left(\sigma_n^2, \xi_n\right)\right\}_{n=0}^{\infty}$ denote the block Gibbs Markov chain and consider applying the strategy outlined in the previous section with the Markov chain in place of the iid sample. First, the analogues of $\overline{g}_N^{\#}$ and $\hat{\eta}_N^{\#}$ are given by

$$\overline{g}_N = \frac{1}{N} \sum_{n=0}^{N-1} g\left(X_n\right) \quad \text{and} \quad \hat{\eta}_N = \frac{\sum_{n=0}^{N-1} v\left(X_n\right)}{\sum_{n=0}^{N-1} u\left(X_n\right)} \tag{1–7}$$

where functions v and u are defined in (1–6). If X_0 is some fixed point, as it would usually be in practice, then $\overline{g}_N$ and $\hat{\eta}_N$ are not unbiased estimators for $\mathrm{E}_\pi g$ and η, but the ergodic theorem (Meyn and Tweedie, 1993, Chapter 17) implies that they are strongly consistent estimators. Thus, at this point in the comparison, all we have lost by using a Markov chain in place of the iid sample is unbiasedness! In other words, the iid sample can be replaced with a Markov chain sample (which is much easier to get) and the strong consistency of the classical Monte Carlo estimator still obtains. Because of this, many view MCMC as a "free lunch" relative to classical Monte Carlo. Unfortunately, as we all know, *there is no free lunch*. Indeed, the routine use of the CLT for choosing an appropriate sample size in the classical Monte Carlo context is far from routine when using MCMC. There are two reasons for this. The first is that, when the iid sequence is replaced by a Markov chain, the second moment conditions are no longer enough to guarantee that CLTs exist. Moreover, the standard method of establishing that there is a CLT involves proving that the underlying Markov chain is *geometrically ergodic* (see

Chapter 2 for the definition), and this often requires difficult theoretical analysis (see, e.g., Jones and Hobert, 2001; Jarner and Hansen, 2000; Mengersen and Tweedie, 1996; Meyn and Tweedie, 1994; Roberts and Rosenthal, 1998, 1999; Roberts and Tweedie, 1996; Roy and Hobert, 2007). The second reason is that, even when a CLT is known to hold, finding a consistent estimator of the the asymptotic variance is a challenging problem because this variance has a fairly complex form and because the dependence among the variables in the Markov chain complicates the asymptotic analysis of estimators based on the chain (see, e.g., Geyer, 1992; Chan and Geyer, 1994; Jones et al., 2006).

In this work, we overcome the problems described above for the block Gibbs sampler through a convergence rate analysis and the development of a regenerative simulation method. In general, regeneration allows one to break a Markov chain up into iid segments (called "tours") so that asymptotic analysis can proceed using standard iid theory. While the theoretical details are fairly involved (Mykland et al., 1995; Hobert et al., 2002), the results of the theory and, more importantly, the application of the results, turns out to be quite straightforward. Indeed, to apply our method, one simply runs the block Gibbs chain as usual, but after each iteration, a single Bernoulli variable is drawn. Specifically, suppose that the value of the chain at time n is X_n. We draw X_{n+1} as usual, and then draw a Bernoulli variable, call it δ_n, whose success probability is a simple function of X_n, X_{n+1} and a few constants (see equation (3–20)). Each time that $\delta_n = 1$ marks the end of a tour. (Note that the tours have random lengths.) Now to estimate $\mathrm{E}_\pi g$, consider running the block Gibbs sampler for R tours; that is, we run the chain until the Rth time that a $\delta_n = 1$. For $t = 1, \ldots, R$, let N_t be the tth tour length and let $S_t = \sum g(X_n)$ where the sum ranges over the values of n that constitute the tth tour. The total number of iterations is $N = \sum_{t=1}^{R} N_t$. The obvious estimate of $\mathrm{E}_\pi g$ based on this simulation is

$$\widetilde{g}_R = \frac{1}{N} \sum_{n=0}^{N-1} g(X_n) = \frac{\sum_{t=1}^{R} S_t}{\sum_{t=1}^{R} N_t} \,, \tag{1–8}$$

which is the same as the usual estimator except that now the length of the simulation is random. The fact that the pairs $(S_1, N_1), \ldots, (S_R, N_R)$ are iid can be used to show that, if the block Gibbs Markov chain is geometrically ergodic and $\mathrm{E}_\pi |g|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$, then there exists a $\gamma^2 \in (0, \infty)$ such that, as $R \to \infty$, we have

$$\sqrt{R}\big(\widetilde{g}_R - \mathrm{E}_\pi g\big) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \gamma^2) \ . \tag{1–9}$$

Furthermore, a simple, consistent estimator of γ^2 is given by

$$\hat{\gamma}^2 = \frac{R \sum_{t=1}^{R} \big(S_t - \widetilde{g}_R N_t\big)^2}{N^2} \ .$$

We conclude that, if the block Gibbs chain is geometrically ergodic and the "$2 + \varepsilon$" moment condition is satisfied, then we can calculate a valid asymptotic standard error for $\widetilde{g}_R$ and only stop the simulation when this standard error is acceptably small. The same strategy can be applied to the MCMC importance sampling estimator $\hat{\eta}_N$ defined in (1–7): geometric ergodicity of the block Gibbs chain plus a pair of moment conditions on the functions u and v will imply a CLT for $\hat{\eta}_N$. In summary, regenerative simulation of the Markov chain allows us to mimic what is done in the classical Monte Carlo context.

Finally, our convergence rate analysis yields conditions under which the block Gibbs chain is geometrically ergodic. Loosely speaking, we are able to say that the chain is geometrically ergodic unless the total sample size M is very small and the within group sample sizes, $m_1, \ldots, m_q$, are highly unbalanced. Our convergence rate result (Proposition 5) applies to all priors $p_{a,b}$ defined in (1–2). Here is a special case.

Corollary 1. *Under the standard diffuse prior, $p_{-\frac{1}{2},0}$, the block Gibbs Markov chain is geometrically ergodic if*

1. $q \geq 4$ *and* $M \geq q + 3$, *or*
2. $q = 3$, $M \geq 6$ *and* $\min\left\{\left(\sum_{i=1}^{3} \frac{m_i}{m_i+1}\right)^{-1}, \frac{m^*}{M}\right\} < 2e^{-\gamma}$, *where* $m^* = \max\{m_1, m_2, m_3\}$ *and* $\gamma \doteq 0.577$ *is Euler's constant.*

Recall that, under the standard diffuse prior, the posterior is proper if and only if $q \geq 3$. When $q \geq 4$, our condition is satisfied for all reasonable data configurations. As for $q = 3$, it turns out that all balanced data sets with $\min\{m_1, m_2, m_3\} \geq 2$ satisfy the conditions, as do most reasonable unbalanced configurations. Table 1-1 displays all unbalanced configurations of (m_1, m_2, m_3) with $m^* \leq 12$ that satisfy the conditions of Corollary 1.

Table 1-1. Unbalanced configurations (m_1, m_2, m_3) with $m^* \leq 12$ that satisfy condition 2 of Corollary 1.

m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3
3	4	4	7	8	10	7	10	11	6	9	12	8	11	12
4	5	5	7	9	10	7	11	11	6	10	12	8	12	12
5	6	6	7	10	10	8	8	11	6	11	12	9	9	12
5	7	7	8	8	10	8	9	11	6	12	12	9	10	12
6	6	7	8	9	10	8	10	11	7	7	12	9	11	12
6	7	7	8	10	10	8	11	11	7	8	12	9	12	12
6	8	8	9	9	10	9	9	11	7	9	12	10	10	12
7	7	8	9	10	10	9	10	11	7	10	12	10	11	12
7	8	8	6	9	11	9	11	11	7	11	12	10	12	12
7	9	9	6	10	11	10	10	11	7	12	12	11	11	12
8	8	9	6	11	11	10	11	11	8	8	12	11	12	12
8	9	9	7	8	11	5	11	12	8	9	12			
6	10	10	7	9	11	5	12	12	8	10	12			

Previous analyses of Gibbs samplers for Bayesian one-way random effects models were performed by Hobert and Geyer (1998) and Jones and Hobert (2001, 2004). However, in each of these studies, the models that were considered have *proper priors* on all parameters. In fact, our Proposition 5 is the first of its kind for random effects models with *improper priors*, which, as we explained above, are the type of priors recommended in the Bayesian literature. It turns out that using improper priors complicates the analysis that is required to study the corresponding Markov chain. Indeed, Proposition 5 is much more than a straightforward extension of the existing results for proper priors.

The Bayesian one-way random effects model that we study is a simplest case of the Bayesian mixed effects model (van Dyk and Meng, 2001, Sec.8). Block Gibbs samplers can be developed to analyze the posterior of these models. However, unified analysis for the convergence rate of the associated Markov chains are difficult. One case that has received

some analysis concerns random effects models with the proper conjugate priors. Under a special restriction about the design matrix and the variance structure, Johnson and Jones (2008) derived a sufficient condition for the geometrically ergodicity of the block Gibbs chain and constructed a regenerative simluation scheme for it. After all, there are many problems still to be resolved concerning complete analysis of Bayesian mixed effects models using MCMC algorithms.

Another related paper is Papaspiliopoulos and Roberts (2008) who studied the convergence rates of Gibbs samplers for Bayesian hierarchical linear models with different symmetric error distributions. What separates our results from theirs is that the variance components in our model are considered unknown parameters, while in their model the variance components are assumed known.

The rest of the dissertation is organized as follows. In Chapter 2, we review some results from general state space Markov chain theory. In particular, we mention that geometric ergodicity plays an important role to ensure the existence of Markov chain CLTs. We also provide an alternative derivation of an existing CLT based on regenerative simulation. In Chapter 3, we carefully study the block Gibbs sampler and provide a sufficient condition for its geometric convergence. Further, we identify regeneration times of the block Gibbs chain by establishing a minorization condition for its transition density. The regenerative method is illustrated using a real data set on styrene exposure where valid estimation for the standard error of MCMC estimators are provided. A slightly different topic is discussed in Chapter 4. We describe an attempt to characterize geometric convergence of Markov chains underlying a specific class of two-variable Gibbs samplers. These Markov chains live on a common state space that constitutes points on the diagonal and the subdiagonal of $\mathbb{N} \times \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers. Our result (Corollary 3) shows that a geometric tail decay of the target distribution is almost necessary and sufficient for the associated chain to be geometrically ergodic.

CHAPTER 2
ON MARKOV CHAIN LIMIT THEORY

2.1 Background Knowledge on General State Space Markov Chains

Let X be a topological space equipped with the Borel σ-algebra $\sigma(\mathsf{X})$ and let $K : \mathsf{X} \times \sigma(\mathsf{X}) \to [0,1]$ be a Markov transition function that defines a discrete time, time homogeneous Markov chain, $\Phi = \{X_n\}_{n=0}^{\infty}$. That is, for $x \in \mathsf{X}$ and $A \in \sigma(\mathsf{X})$,

$$K(x, A) = \Pr(X_1 \in A \mid X_0 = x)\,.$$

Also, let $K^n : \mathsf{X} \times \sigma(\mathsf{X}) \to [0,1]$, $n = 2, 3, \ldots$, denote the n-step Markov transition functions. Suppose that μ is a σ-finite measure on X and that the function $k : \mathsf{X} \times \mathsf{X} \to [0, \infty)$ satisfies $K(x, A) = \int_A k(y|x)\,\mu(dy)$ for any $x \in \mathsf{X}$ and any μ-measurable A. Then k is called a *Markov transition density* of Φ with respect to μ. Suppose that π is an invariant probability measure for the chain; that is,

$$\int_{\mathsf{X}} K(x, dy)\pi(dx) = \pi(dy)\,.$$

We say the chain Φ is *Harris ergodic* if it is ψ-irreducible, aperiodic and Harris recurrent, see Meyn and Tweedie (1993) for definitions. Here ψ represents the maximal irreducibility measure of the chain, see Appendix B for the definition. One method of establishing Harris recurrence is to show that every bounded harmonic function is constant (Nummelin, 1984, Theorem 3.8). Here, a function $h : \mathsf{X} \to \mathbb{R}$ is called harmonic for K if $h = Kh$, where $Kh(x) := \int_{\mathsf{X}} h(x')K(x, dx')$ for all $x \in \mathsf{X}$. Generally speaking, Harris ergodicity is easy to verify in practice. For example, it suffices that Φ has a strictly positive transition density.

Lemma 1. *Suppose Φ is a Markov chain with Markov transition density k (with respect to μ) and invariant probability measure π. If $k(y|x) > 0$ for all $x, y \in \mathsf{X}$, then Φ is Harris ergodic.*

Proof. If $\mu(A) > 0$ then $K(x, A) = \int_A k(x'|x)\mu(dx') > 0$ for all $x \in \mathsf{X}$; i.e., it is possible to get from any point $x \in \mathsf{X}$ to any non measure zero set A in one step. This implies that Φ

is aperiodic and μ-irreducible. Since Φ is μ-irreducible, it is also ψ-irreducible by definition and that μ is absolutely continuous with respect to ψ (denoted $\psi \succ \mu$) .

Since Φ is ψ-irreducible and admits an invariant probability distribution, it is also positive recurrent (Meyn and Tweedie, 1993, Chap. 10). To establish Harris recurrence of Φ, suppose h is a bounded, harmonic function. Since Φ is ψ-irreducible and recurrent, h is constant ψ-a.e. (Nummelin, 1984, Proposition 3.1.3). Thus, there exists a set N with $\psi(N) = 0$ such that $h(x) = c$ for all $x \in \overline{N}$, where $\overline{N}$ denotes the complement set of N. Since $\psi \succ \mu \succ K(x, \cdot)$ for all $x \in \mathsf{X}$, $K(x, N) = 0$. Now, for any $x \in \mathsf{X}$, we have

$$h(x) = \int_{\mathsf{X}} h(x')K(x, dx') = \int_{\overline{N}} h(x')K(x, dx') + \int_{N} h(x')K(x, dx') = c + 0 = c,$$

which implies that $h \equiv c$. It follows that Φ is Harris recurrent. $\square$

Remark 1. Although not necessary for the proof of any of our results, it turns out that, under the conditions of Lemma 1, it is also true that $\mu \succ \psi$, so these two measures are actually equivalent. Indeed, if $\mu(A) = 0$, then it follows that $K^l(x, A) = 0$ for all $x \in \mathsf{X}$ and all $l \in \mathbb{N}$, which implies that $\psi(A) = 0$.

If Φ is Harris ergodic, then for any $x \in \mathsf{X}$,

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

where $\|\cdot\|$ represents the total variation norm defined for a signed measure λ by $\|\lambda\| = 2 \sup_{A \in \sigma(\mathsf{X})} |\lambda(A)|$. Note that this tells us nothing about the *rate* of convergence. A Harris ergodic chain Φ is said to be *geometrically ergodic* if there exists a function $c : \mathsf{X} \to [0, \infty)$ and a constant $0 < r < 1$ such that, for all $x \in \mathsf{X}$ and all $n = 0, 1, \ldots$

$$\|K^n(x, \cdot) - \pi(\cdot)\| \leq c(x)\, r^n \ .$$

One way to show that a chain convergences to its limiting distribution at a geometric rate is to construct a (Lyapunov) drift condition for the Markov transition function.

Proposition 1 below states one version of the drift condition. It is a combination of Lemma 15.2.8 and Theorem 6.0.1 from Meyn and Tweedie (1993).

The Markov chain Φ with transition function K is called a *Feller chain* if $K(\cdot, O)$ is lower semicontinuous for any open set $O \in \sigma(\mathsf{X})$. A function $w : \mathsf{X} \to [0, \infty)$ is said to be *unbounded off compact sets* if the level set $\{x \in \mathsf{X} : w(x) \leq d\}$ is compact for every $d > 0$.

Proposition 1. *Assume that Φ is Harris ergodic and Feller and that the support of the maximal irreducibility measure has nonempty interior. If there exist $\rho < 1$, $L < \infty$ and a function $w : \mathsf{X} \to [0, \infty)$ that is unbounded off compact sets such that*

$$E[w(X_1) \mid X_0 = x] \leq \rho w(x) + L \,, \tag{2–1}$$

then Φ is geometrically ergodic.

The inequality (2–1) is called a *drift condition* and the function w is called a *drift function.*

The concepts of Harris ergodicity and geometric ergodicity concern the distribution of a Markov chain at step n. Hence they are not directly relevant to assessing a single Markov chain output. However, we show in the next section that these properties are desirable for Markov chains used in MCMC algorithoms because they indicate good asymptotic behavior of the estimators defined in (1–7).

2.2 MCMC Estimators, MCMC Importance Sampling Estimators and Their Central Limit Theorems

Assume that the Markov chain Φ is Harris ergodic with invariant distribution π. Recall from Chapter 1 that we construct $\overline{g}_N$ and $\hat{\eta}_N$ based on Φ, and use them to estimate $E_\pi g$ and $\eta = E_{\pi^*} g$ respectively. These estimators are valid because of the ergodic theorem (Meyn and Tweedie, 1993, p.411). The theorem says that, if $g \in L^1(\pi)$, then with probability 1,

$$\overline{g}_N \to E_\pi g \quad \text{as } N \to \infty \,,$$

no matter what the distribution of X_0. In other words, $\overline{g}_N$ is strongly consistent for $E_\pi g$. Similarly, if both v and u defined in equation (1–6) belongs to $L^1(\pi)$, then $\hat{\eta}_N = \overline{v}_N/\overline{u}_N$ is strongly consistent for $\eta = \mathrm{E}_\pi v/\mathrm{E}_\pi u$. However, these assumptions are not enough to guarantee a CLT for these estimators.

Here, we refer to Theorem 2 of Chan and Geyer (1994) for a well known result that relates convergence rate of Markov chains to the existence of a CLT for $\overline{g}_N$. This theorem is a consequence of Theorem 18.5.3 from Ibragimov and Linnik (1971),

Theorem 1. *Suppose that Φ is a Markov chain with invariant measure π. If Φ is geometrically ergodic, and $|g|^{(2+\varepsilon)} \in L^1(\pi)$ for some positive ε, then no matter what the distribution of X_0, $\sqrt{m}\,(\overline{g}_m - E_\pi g)$ converges weakly to a normal distribution with mean 0 and variance*

$$\varsigma^2 = Var\,(g(Y_0)) + 2\sum_{n=1}^{\infty} Cov\,(g(Y_0), g(Y_n)) \ ,$$

where $\{Y_n, n = 0, 1, \dots\}$ is the stationary version of $\{X_n, n = 0, 1, \dots\}$, i.e. $Y_0 \sim \pi$.

Altough a powerful asymptotic result, Theorem 1 does not provide enough guidance for a complete MCMC analysis in practice. The complicated expression for ς^2 does not immediately suggest a consistent estimator. This is not surprising given the dependence in the samples as well as the influence of the starting distribution of the chain.

To circumvent this difficulty, we resort to an alternative CLT based on *regenerative simulation*. The idea is to identify regeneration times of the Markov chain so that we can partition the dependent samples into iid segments. This allows us to use iid theory to analyze the asymptotic behavior of estimators based on the chain. Mykland et al. (1995) and Hobert et al. (2002) derived CLTs for $\overline{g}_N$ using this idea. Bhattacharya (2008) generalized the proofs in the above papers to get a CLT for $\hat{\eta}_N$. In what follows, we provide an alternative derivation for the CLT of Bhattacharya (2008), which is our Proposition 2. Our proof seems to be more transparent. After that, we show that the CLT of $\overline{g}_N$ described in Hobert et al. (2002) can be viewed as a corollary to Proposition 2.

Comparing to Theorem 1, CLTs based on regenerative simulation clearly require an extra assumption, that is, we can indeed identify regeneration times of the underlying Markov chain. This is a non-issue for discrete state space Markov chains, as the chain starts over every time it returns to a fixed state. See, for example, Asmussen (2003). For general state space Markov chains, the prerequisite is satisfied if one can establish the following *minorization condition* for the Markov transition function:

$$K(x, dy) \geq s(x)\nu(dy) \quad \text{for all } x \in \mathsf{X}, \tag{2–2}$$

where $s : \mathsf{X} \to [0, 1]$ satisfies $\mathrm{E}_\pi s > 0$ and ν is a probability measure on X. The benefit is that K can be rewritten as a mixture

$$K(x, dy) = s(x)\nu(dy) + (1 - s(x))r(x, dy)$$

where $r(x, dy)$ is defined to be $\big(K(x, dy) - s(x)\nu(dy)\big)/\big(1 - s(x)\big)$ for $x \in \{x : s(x) < 1\}$ and any arbitrary probability density for $x \in \{x : s(x) = 1\}$. This mixture provides an alternative method of simulating the Markov chain. Given the current state, $X_n = x$, drawing X_{n+1} from $K(x, \cdot)$ can be done through flipping a coin, $\delta_n \sim \mathrm{Ber}\,(s(x))$, then drawing $X_{n+1} \sim \nu(\cdot)$ if $\delta_n = 1$ and $X_{n+1} \sim r(x, \cdot)$ if $\delta_n = 0$. The point is, every time that $\delta_n = 1$, we have $X_{n+1} \sim \nu(\cdot)$ not dependent of the current value $X_n = x$, which in effect starts the process over at its $(n + 1)$th iteration.

Remark 2. In practice, we can avoid drawing from the potentially problematic $r(\cdot|x)$. Indeed, we can first generate Φ according to k as usual, and then fill in values for $\{\delta_n\}$ by drawing from their respective conditional distributions

$$\delta_n | X_n = x, X_{n+1} = x' \sim \mathrm{Ber}\left(\frac{s(x)\nu(x')}{k(x'|x)}\right). \tag{2–3}$$

Suppose Φ is started with $X_0 \sim \nu(\cdot)$. Denote its regeneration times by $0 = \tau_0 < \tau_1 < \tau_2 < \ldots$; that is, $\tau_{t+1} = \min\{n > \tau_t, \delta_{n-1} = 1\}$ for $t \geq 0$. Let the chain run for R tours where R is fixed. Then the total length of the chain (counting X_0) is a random number,

τ_R. For $t = 1, 2, \ldots, R$, define

$$V_t = \sum_{n=\tau_{t-1}}^{\tau_t - 1} v(X_n) \qquad \text{and} \qquad U_t = \sum_{n=\tau_{t-1}}^{\tau_t - 1} u(X_n) \ ,$$

where the sums range over the values of n that constitute the tth tour. Let $\widetilde{\eta}_R$ denote the estimator of η based on the R tours, that is,

$$\widetilde{\eta}_R = \hat{\eta}_{\tau_R} = \frac{\sum_{n=0}^{\tau_R - 1} v(X_n)}{\sum_{n=0}^{\tau_R - 1} u(X_n)} = \frac{\sum_{t=1}^{R} V_t}{\sum_{t=1}^{R} U_t} \ . \tag{2–4}$$

We now study the asymptotic behavior of $\widetilde{\eta}_R$ as R gets large.

First, by Kac's theorem (Meyn and Tweedie, 1993, Thm 10.2.2), $R/\tau_R \to \mathrm{E}_\pi s$ with probability 1 as $R \to \infty$. In conjunction with the ergodic theorem, this yields

$$\frac{1}{R} \sum_{t=1}^{R} V_t = \frac{\tau_R}{R} \frac{1}{\tau_R} \sum_{n=0}^{\tau_R - 1} v(X_n) \overset{\text{a.s.}}{\to} \big(\mathrm{E}_\pi s\big)^{-1} \mathrm{E}_\pi v \quad \text{as } R \to \infty \ .$$

Now, since each is based on a separate tour, the (V_t, U_t) pairs are iid. Thus, the strong law of large numbers implies that

$$\mathrm{E}_\nu V_1 = \big(\mathrm{E}_\pi s\big)^{-1} \mathrm{E}_\pi v < \infty \ ,$$

where the notation "E_ν" is meant to remind the reader that each tour is started with a draw from ν. An analogous argument shows that, as $R \to \infty$,

$$\frac{1}{R} \sum_{t=1}^{R} U_t \overset{\text{a.s.}}{\to} \big(\mathrm{E}_\pi s\big)^{-1} \mathrm{E}_\pi u = \mathrm{E}_\nu U_1 < \infty \ .$$

Putting all of this together, we see that

$$\mathrm{E}_\nu V_1 = \eta \, \mathrm{E}_\nu U_1 \ .$$

Hence, the random variables $(V_t - \eta U_t)$, $t = 1, 2, \ldots, R$, are iid and have mean zero. Assume for the time being that $\mathrm{E}_\nu V_1^2$ and $\mathrm{E}_\nu U_1^2$ are both finite. Then $\mathrm{E}_\nu \big[(V_1 - \eta \, U_1)^2\big]$ is

also finite and we have the following CLT:

$$\sqrt{R}\,\frac{1}{R}\sum_{t=1}^{R}(V_t-\eta\,U_t)\xrightarrow{\mathrm{d}}\mathrm{N}\Big(0\,,\,\mathrm{E}_\nu\big[(V_1-\eta\,U_1)^2\big]\Big)\ \text{ as }\ R\to\infty\,.$$

This yields a CLT for $\widetilde{\eta}_R$:

$$\sqrt{R}\big(\widetilde{\eta}_R-\eta\big)=\sqrt{R}\bigg(\frac{\sum_{t=1}^{R}V_t}{\sum_{t=1}^{R}U_t}-\eta\bigg)=\frac{\sqrt{R}\frac{1}{R}\sum_{t=1}^{R}(V_t-\eta\,U_t)}{\frac{1}{R}\sum_{t=1}^{R}U_t}\xrightarrow{\mathrm{d}}\mathrm{N}\big(0,\gamma^2\big)\ \text{ as }\ R\to\infty\,,\quad (2\text{–}5)$$

where

$$\gamma^2=\mathrm{E}\big[(V_1-\eta\,U_1)^2\big]\big/(\mathrm{E}_\nu U_1)^2\,.$$

The main benefit of deriving a CLT for $\widetilde{\eta}_R$ using regeneration is the existence of the following simple, strongly consistent estimator of γ^2:

$$\hat{\gamma}^2=\frac{\frac{1}{R}\sum_{t=1}^{R}(V_t-\widetilde{\eta}_R\,U_t)^2}{\Big(\frac{1}{R}\sum_{t=1}^{R}U_t\Big)^2}\,.\qquad (2\text{–}6)$$

To see why $\hat{\gamma}^2$ is indeed a consistent estimator, note that, as R gets large,

$$\tilde{\gamma}^2:=\frac{\frac{1}{R}\sum_{t=1}^{R}(V_t-\eta\,U_t)^2}{\Big(\frac{1}{R}\sum_{t=1}^{R}U_t\Big)^2}\xrightarrow{\text{a.s.}}\gamma^2\,.$$

The strong consistency of $\hat{\gamma}^2$ then follows from the fact that

$$\hat{\gamma}^2-\tilde{\gamma}^2=\frac{\Big(\frac{1}{R}\sum_{t=1}^{R}2U_tV_t\Big)(\eta-\widetilde{\eta}_R)+\Big(\frac{1}{R}\sum_{t=1}^{R}U_t^2\Big)(\widetilde{\eta}_R^2-\eta^2)}{\Big(\frac{1}{R}\sum_{t=1}^{R}U_t\Big)^2}\xrightarrow{\text{a.s.}}0\ \text{ as }R\to\infty\,.$$

Recall that, in order to arrive at the CLT in (2–5), we assumed that $\mathrm{E}_\nu V_1^2$ and $\mathrm{E}_\nu U_1^2$ are both finite. These moment conditions are actually quite difficult to check directly. Indeed, V_1 and U_1 are sums of functions of the states of the Markov chain containing a random number of terms. However, Hobert et al. (2002) show that, if the underlying Markov chain is geometrically ergodic, and there exists an $\varepsilon>0$ such that $\mathrm{E}_\pi|v|^{2+\varepsilon}$ and $\mathrm{E}_\pi|u|^{2+\varepsilon}$ are both finite, then $\mathrm{E}_\nu V_1^2$ and $\mathrm{E}_\nu U_1^2$ are both finite as well. Now, we summarize the above results in Proposition 2. It is the same as the CLT in Bhattacharya (2008).

Proposition 2. *Suppose that Φ is geometrically ergodic. Suppose further that Φ has a Markov transition function, K, satisfying (2–2). Consider estimating the ratio $\eta = E_{\pi^*} g = E_\pi v / E_\pi u$ with the estimator $\widetilde{\eta}_R$ defined at (2–4). If there exists an $\varepsilon > 0$ such that $E_\pi |v|^{2+\varepsilon}$ and $E_\pi |u|^{2+\varepsilon}$ are finite, then for $X_0 \sim \nu(\cdot)$, $R \to \infty$ with probability 1 as $n \to \infty$ and*

$$\sqrt{R}\big(\widetilde{\eta}_R - \eta\big) \xrightarrow{\mathrm{d}} N\big(0, \gamma^2\big) \quad \text{as} \quad R \to \infty \ .$$

Furthermore, $\hat{\gamma}^2$ defined at (2–6) is a strongly consistent estimator of γ^2.

Note that if there exists an $\varepsilon > 0$ such that $\mathrm{E}_\pi \big|g\big|^{2+\varepsilon} < \infty$, then a sufficient condition for $\mathrm{E}_\pi |v|^{2+\varepsilon} < \infty$ and $\mathrm{E}_\pi |u|^{2+\varepsilon} < \infty$ is the existence of a constant $M \in [1, \infty)$ such that

$$\sup_{x \in S^*} \frac{f^*(x)}{f(x)} < M \ . \tag{2–7}$$

This condition basically says that π has heavier tails than π^*. In fact, (2–7) is exactly what is required to use π as the candidate density in an accept/reject algorithm for π^*. Of course, this accept/reject algorithm is not viable if we cannot make exact draws from π. For a thorough review of accept/reject methods, see Robert and Casella (2004, Chapter 2).

Suppose we use the above importance sampling technique to evaluate the effect of a change in the prior distribution or the likelihood function in a Bayesian analysis. Suppose that x and y denote parameters and observed data, respectively. Think of two different posterior densities, π and π^*, with $\pi(x) \propto L(x; y)p(x)$ and $\pi^*(x) \propto L^*(x; y)p^*(x)$, where $L(x; y)$ and $L^*(x; y)$ denote two different likelihood functions, and $p(x)$ and $p^*(x)$ denote two different prior densities. In this setting,

$$\frac{f^*(x)}{f(x)} = \frac{L^*(x; y)p^*(x)}{L(x; y)p(x)} \ . \tag{2–8}$$

If the two Bayesian models differ only in the prior that is used, i.e. if $L = L^*$, then $f^*(x)/f(x) = p^*(x)/p(x)$, which is simply the ratio of the two priors. Hence, the moment conditions in Proposition 2 reduce to $\mathrm{E}_\pi (p^*/p)^{2+\varepsilon} < \infty$ and $\mathrm{E}_\pi |g\, p^*/p|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$.

Proposition 2 is a direct generalization of the result in Hobert et al. (2002), which we state in Corollary 2. This result may be viewed as pertaining to the special case where $u \equiv 1$ and $v = g$. In this case, $U_t = \sum 1 =: N_t$ and $V_t = \sum g(X_n) =: S_t$, where the sum ranges over the values of n that constitute the tth tour. Clearly, U_t is the tth tour length and the moment condition concerning u is satisfied since $\mathrm{E}_\pi |u|^{2+\varepsilon} = 1 < \infty$.

Corollary 2. *If Φ is geometrically ergodic, satisfies the minorization condition (2–2) and $|g|^{2+\varepsilon} \in L^1(\pi)$ for some positive ε, then for $X_0 \sim \nu(\cdot)$, $R \to \infty$ with probability 1 as $n \to \infty$ and*

$$\sqrt{R}\,(\widetilde{g}_R - E_\pi g) \xrightarrow{\mathrm{d}} N(0, \gamma^2) \quad \textit{as } R \to \infty,$$

where

$$\gamma^2 = \frac{E_\nu\left[(S_1 - N_1 E_\pi g)^2\right]}{\left[E_\nu N_1\right]^2}.$$

Remark 3. The accuracy of the estimator $\hat{\gamma}^2$ depends on the variability of the blocks between regeneration times. Let $\overline{N}$ denote the average length of the R tours. According to Mykland et al. (1995, Section 2), small values of the coefficient of variation, $\mathrm{CV}(\overline{N}) = \mathrm{Var}(\overline{N})/\mathrm{E}^2(\overline{N})$ are desirable, and this quantity can be estimated by $\widehat{\mathrm{CV}}(\overline{N}) = \sum_{t=1}^{R}(N_t - \overline{N})^2/(R\overline{N})^2$. Preferably, $\widehat{\mathrm{CV}}(\overline{N}) < 0.01$. In addition to computing $\widehat{\mathrm{CV}}(\overline{N})$, it is useful to examine the pattern of regeneration times graphically, for example, by plotting τ_t/τ_R against t/R, which is called the scaled regeneration quantile (SRQ) plot. If the total run length is long enough, then this plot should be close to a straight line through the origin with unit slope.

One nice feature of Proposition 2 (and Corollary 2) is that the conditions for the existence of a CLT are separated into one concerning the convergence rate of the Markov chain and two others that are simple moment conditions with respect to the target distribution. Hence, if the chain under consideration is known to be geometrically ergodic, then checking the conditions is quite straightforward. In contrast, many results for CLTs for regenerative processes have sufficient conditions that are fairly difficult to check in practice because they involve expectations of complex functions of the underlying process.

For example, Mykland et al. (1995)'s CLT requires $\mathrm{E}_\nu N_1^2 < \infty$ and $\mathrm{E}_\nu S_1^2 < \infty$. Other examples of this can be found in the operations research literature where regenerative simulation is used to assess the variability of MCMC estimators in the analysis of queueing systems (see, e.g., Ripley, 1987; Bratley et al., 1987; Crane and Iglehart, 1975; Lavenberg and Slutz, 1975; Glynn and Iglehart, 1987). The Markov processes that underly these analyses have countable state spaces, which makes the identification of regeneration times trivial. However, unlike Proposition 2, the conditions for CLTs involve unwieldy moment conditions with respect to the underlying Markov process. These conditions are quite difficult to check for all but the simplest queueing systems.

CHAPTER 3
BLOCK GIBBS SAMPLING FOR BAYESIAN ONE-WAY RANDOM EFFECTS MODELS

3.1 Transition Function of the Block Gibbs Sampler

In this chapter, we carefully study the block Gibbs sampler for the Bayesian one-way random effects model described in Chapter 1. Recall that under the conditional conjugate prior

$$p_{a,b}\big(\mu, \sigma_\theta^2, \sigma_e^2\big) = \big(\sigma_\theta^2\big)^{-(a+1)} \big(\sigma_e^2\big)^{-(b+1)} ,$$

the posterior distribution that we would like to draw MCMC samples from has density

$$\begin{aligned}
\pi\big(\theta, \mu, \sigma_\theta^2, \sigma_e^2\big) &\propto L(\theta, \mu, \sigma_\theta^2, \sigma_e^2; y)\, \phi(\theta|\mu, \sigma_\theta^2, \sigma_e^2)\, p_{a,b}\big(\mu, \sigma_\theta^2, \sigma_e^2\big) \\
&= \left[\prod_{i=1}^{q} \prod_{j=1}^{m} \big(2\pi\sigma_e^2\big)^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2\sigma_e^2}(y_{ij} - \theta_i)^2 \Big\} \right] \\
&\quad \times \left[\prod_{i=1}^{q} \big(2\pi\sigma_\theta^2\big)^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2\sigma_\theta^2}(\theta_i - \mu)^2 \Big\} \right] \big(\sigma_\theta^2\big)^{-(a+1)} \big(\sigma_e^2\big)^{-(b+1)} .
\end{aligned} \tag{3–1}$$

Since the observed data are always conditioned upon, we have suppressed the notation of dependence on y in the posterior density π. Of course, whenever an improper prior is used, one must check that the resulting posterior is proper. As we have mentioned in the introduction, results in Hobert and Casella (1996) show that the posterior is proper if and only if all three of the following conditions hold:

$$a < 0 , \quad a + \frac{q}{2} > \frac{1}{2} , \quad \text{and} \quad a + b > \frac{1 - M}{2} , \tag{3–2}$$

where M is the total sample size; that is, $M = \sum_{i=1}^{q} m_i$. Note that (3–2) implies $q > 1 - 2a > 1$ so a necessary condition for propriety is $q \geq 2$. Under the standard diffuse prior, that is, when $(a, b) = (\frac{1}{2}, 0)$, the posterior is proper if and only if $q \geq 3$.

The following block Gibbs sampler provides an efficient way to simulate a Markov chain with invariant distribution π. Let $\sigma^2 = \big(\sigma_\theta^2, \sigma_e^2\big)$, $\xi = (\mu, \theta)$ and suppose that the state of the chain at time n is $\big(\sigma_n^2, \xi_n\big)$. One iteration of our sampler entails drawing σ_{n+1}^2 conditional on ξ_n, and then drawing ξ_{n+1} conditional on σ_{n+1}^2. Formally, the Markov

transition density for the transition $\left(\sigma_n^2, \xi_n\right) \to \left(\sigma_{n+1}^2, \xi_{n+1}\right)$ is given by

$$k\left(\sigma_{n+1}^2, \xi_{n+1} \mid \sigma_n^2, \xi_n\right) = \pi\left(\sigma_{n+1}^2 \mid \xi_n\right) \pi\left(\xi_{n+1} \mid \sigma_{n+1}^2\right) . \tag{3–3}$$

Straightforward manipulation of (3–1) shows that, σ_θ^2 and σ_e^2 are conditionally independent given ξ; that is,

$$\pi(\sigma^2 \mid \xi) = \pi(\sigma_\theta^2 \mid \xi)\, \pi(\sigma_e^2 \mid \xi) ,$$

and that

$$\sigma_\theta^2 \mid \xi \sim \mathrm{IG}\left(\frac{q}{2} + a, \frac{1}{2}\sum_i (\theta_i - \mu)^2)\right) \quad \text{and} \quad \sigma_e^2 \mid \xi \sim \mathrm{IG}\left(\frac{M}{2} + b, \frac{1}{2}\sum_{i,j} (y_{ij} - \theta_i)^2\right) .$$

We say $X \sim \mathrm{IG}(\alpha, \beta)$ if X is a random variable supported on $\mathbb{R}_+$ with density function proportional to $x^{-(\alpha+1)} e^{-\beta/x}$. On the other hand,

$$\xi | \sigma^2 \sim \mathrm{N}\left(\xi_0, \mathbf{V}\right) \tag{3–4}$$

where

$$\mathbf{V}^{-1} = \begin{bmatrix} \mathbf{D}^2 & -\left(\sigma_\theta^2\right)^{-1} \mathbf{1} \\ -\left(\sigma_\theta^2\right)^{-1} \mathbf{1}^T & q\left(\sigma_\theta^2\right)^{-1} \end{bmatrix}$$

and $\mathbf{D}$ is a $q \times q$ diagonal matrix whose ith diagonal element is $d_{ii} = \sqrt{\left(\sigma_\theta^2\right)^{-1} + m_i \left(\sigma_e^2\right)^{-1}}$. The mean, ξ_0, is the solution of

$$\mathbf{V}^{-1} \xi_0 = \left(\sigma_e^2\right)^{-1} \begin{pmatrix} m_1 \overline{y}_1 \\ m_2 \overline{y}_2 \\ \vdots \\ m_q \overline{y}_q \\ 0 \end{pmatrix} \tag{3–5}$$

where $\overline{y}_i = m_i^{-1} \sum_j y_{ij}$, for $i = 1, \ldots, q$.

The Cholesky factorization of the precision matrix is $\mathbf{V}^{-1} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower-triangular matrix. The elements of $\mathbf{L}$ can be calculated "by hand":

$$\mathbf{L} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ -\mathbf{1}^T\left(\sigma_\theta^2\mathbf{D}\right)^{-\mathbf{1}} & \sqrt{t} \end{bmatrix}$$

where

$$t = \sum_{i=1}^{q} \frac{m_i}{\sigma_e^2 + m_i\sigma_\theta^2}.$$

It is straightforward to show that

$$\mathbf{L}^{-1} = \begin{bmatrix} \mathbf{D}^{-1} & \mathbf{0} \\ \frac{\mathbf{c}^T\mathbf{D}^{-1}}{\sqrt{t}} & \frac{1}{\sqrt{t}} \end{bmatrix}$$

where $\mathbf{c}^T$ is a $1 \times q$ row vector whose ith element is $\left(\sigma_\theta^2\, d_{ii}\right)^{-1}$.

This allows us to obtain the following closed form expressions:

$$\begin{aligned} \mathrm{Var}\left(\theta_i \middle| \sigma_\theta^2, \sigma_e^2\right) &= \frac{\sigma_e^2}{\sigma_e^2 + m_i\sigma_\theta^2}\left[\sigma_\theta^2 + \frac{\sigma_e^2}{\left(\sigma_e^2 + m_i\sigma_\theta^2\right)t}\right] \\ \mathrm{Cov}\left(\theta_i, \theta_j \middle| \sigma_\theta^2, \sigma_e^2\right) &= \frac{\left(\sigma_e^2\right)^2}{\left(\sigma_e^2 + m_i\sigma_\theta^2\right)\left(\sigma_e^2 + m_j\sigma_\theta^2\right)t} \\ \mathrm{Cov}\left(\theta_i, \mu \middle| \sigma_\theta^2, \sigma_e^2\right) &= \frac{\sigma_e^2}{\left(\sigma_e^2 + m_i\sigma_\theta^2\right)t} \\ \mathrm{Var}\left(\mu \middle| \sigma_\theta^2, \sigma_e^2\right) &= \frac{1}{t}\ . \end{aligned} \tag{3–6}$$

We can also use (3–5) to write down the closed form of $\boldsymbol{\zeta_0}$, that is, to write down $\mathrm{E}(\mu|\lambda_\theta, \lambda_e)$ and $\mathrm{E}(\theta_k|\lambda_\theta, \lambda_e)$ for $k = 1, \ldots, q$. We have

$$\begin{aligned} \mathrm{E}(\mu|\lambda_\theta, \lambda_e) &= \sum_{i=1}^{q} m_i\lambda_e\overline{y}_i\mathrm{Cov}(\theta_i, \mu|\lambda_\theta, \lambda_e) \\ &= \frac{1}{t}\sum_{i=1}^{q} \frac{m_i\lambda_\theta\lambda_e\overline{y}_i}{\lambda_\theta + m_i\lambda_e}. \end{aligned}$$

Note that $\mathrm{E}(\mu|\lambda_\theta, \lambda_e)$ is a convex combination of the $\overline{y}_i$'s. Thus, it is uniformly bounded by a constant. Furthermore,

$$\begin{aligned}\mathrm{E}(\theta_k|\lambda_\theta, \lambda_e) &= \sum_{i=1}^{q} m_i \lambda_e \overline{y}_i \mathrm{Cov}(\theta_k, \theta_i|\lambda_\theta, \lambda_e) \\ &= \frac{\lambda_\theta}{\lambda_\theta + m_k\lambda_e}\left[\frac{1}{t}\sum_{i=1}^{q}\frac{m_i\lambda_\theta\lambda_e\overline{y}_i}{\lambda_\theta + m_i\lambda_e}\right] + \frac{\lambda_e m_k \overline{y}_k}{\lambda_\theta + m_k\lambda_e}\,.\end{aligned}$$

Note that $\mathrm{E}(\theta_k|\lambda_\theta, \lambda_e)$ is a convex combination of $\mathrm{E}(\mu|\lambda_\theta, \lambda_e)$ and $\overline{y}_k$ and is therefore also uniformly bounded by a constant.

Since $\pi(\xi \mid \sigma^2)$ and $\pi(\sigma^2 \mid \xi)$ are both strictly positive for $(\sigma^2, \xi) \in \mathbb{R}_+^2 \times \mathbb{R}^{q+1}$, it follows that the Markov transition density defined in (3–3) is strictly positive. Thus, Lemma 1 in Section 2.1 implies that the block Gibbs Markov chain, $\{X_n\}_{n=0}^{\infty} = \left\{\left(\sigma_n^2, \xi_n\right)\right\}_{n=0}^{\infty}$, is Harris ergodic. Hence, the MCMC estimators defined in (1–7) are strongly consistent under simple first moment conditions. In the following section, we show that the chain is geometrically ergodic. This will ensure the existence of CLTs for the MCMC estimators.

3.2 Geometric Ergodicity of the Block Gibbs Sampler

As we now explain, an indirect proof is used to show that the block Gibbs chain is geometrically ergodic. It is well known that the two marginal sequences comprising a two-variable Gibbs chain are themselves Markov chains (Liu et al., 1994). Moreover, the Gibbs chain and its two marginal chains all converge at exactly the same rate (Diaconis et al., 2008; Diebolt and Robert, 1994; Roberts and Rosenthal, 2001). In terms of our example above, this means that $\left\{\xi_n\right\}_{n=0}^{\infty}$ is a Markov chain that converges at the same rate as $\left\{(\sigma_n^2, \xi_n)\right\}_{n=0}^{\infty}$. Therefore, we can prove that the block Gibbs chain is geometric by proving that the ξ-chain is geometric. The ξ-chain has a Markov transition density (with respect to Lebesgue measure on $\mathbb{R}^{q+1}$) given by

$$k^*\left(\tilde{\xi} \mid \xi\right) = \int_{\mathbb{R}_+^2} \pi\left(\tilde{\xi} \mid \sigma^2\right) \pi\left(\sigma^2 \mid \xi\right) d\sigma^2\,. \tag{3–7}$$

Clearly, k^* is strictly positive on $\mathbb{R}^{q+1} \times \mathbb{R}^{q+1}$ so another application of Lemma 1 shows that the ξ-chain is Harris ergodic. We also conclude from Lemma 1 that the maximal irreducibility measure of the ξ-chain is equivalent to Lebesgue measure on $\mathbb{R}^{q+1}$ and hence its support has non-empty interior. Finally, it is simple to show that the ξ-chain is Feller using Fatou's Lemma.

Lemma 2. *The Markov chain with transition density k^* is a Feller chain.*

Proof. For any open set $O \subset \mathbb{R}^{q+1}$ and any sequence $\{\xi_l\}_{l=1}^{\infty}$ in $\mathbb{R}^{q+1}$ such that $\xi_l \to \xi$ as $l \to \infty$, we have

$$\begin{aligned}
\liminf_l \int_O k^*(\tilde{\xi} \mid \xi_l)\, d\tilde{\xi} &= \liminf_l \int_O \int_{\mathbb{R}_+^2} \pi(\sigma^2 \mid \xi_l)\, \pi(\tilde{\xi} \mid \sigma^2)\, d\sigma^2\, d\tilde{\xi} \\
&\geq \int_O \int_{\mathbb{R}_+^2} \liminf_l \pi(\sigma^2 \mid \xi_l)\, \pi(\tilde{\xi} \mid \sigma^2)\, d\sigma^2\, d\tilde{\xi} \\
&= \int_O \int_{\mathbb{R}_+^2} \pi(\sigma^2 \mid \xi)\, \pi(\tilde{\xi} \mid \sigma^2)\, d\sigma^2\, d\tilde{\xi} \\
&= \int_O k^*(\tilde{\xi} \mid \xi)\, d\tilde{\xi}\,,
\end{aligned}$$

where the inequality follows from Fatou's Lemma and the second equality is implied by the continuity of $\pi(\sigma^2 \mid \xi)$ in ξ. Hence the chain with transition density k^* is Feller by definition. □

We are now in a position to use Proposition 1 to prove that the ξ-chain is geometric. The real challenge is constructing a valid drift condition. Recall that a function $w : \mathbb{R}^{q+1} \to \mathbb{R}_+$ is said to be *unbounded off compact sets* if the level set $\{\xi : w(\xi) \leq \gamma\}$ is compact for every $\gamma < \infty$. According to Proposition 1, we can prove that the ξ-chain is geometric by finding a w that is unbounded off compact sets and satisfies the drift condition

$$\mathrm{E}\big(w(\tilde{\xi}) \mid \xi\big) \leq \rho w(\xi) + L \quad \text{for all } \xi \in \mathbb{R}^{q+1}, \tag{3–8}$$

where $\rho < 1$ and $L < \infty$. Our drift function takes the form

$$w(\xi) = \epsilon \left[w_1(\xi)\right]^s + \left[w_2(\xi)\right]^s\,,$$

where $w_1(\xi) = \sum_{i=1}^{q}(\theta_i - \mu)^2$, $w_2(\xi) = \sum_{i=1}^{q} m_i(\overline{y}_i - \theta_i)^2$ and $\epsilon > 0$ and $s \in (0, 1]$ are to be determined. It is easy to see that, for fixed $\epsilon > 0$ and $s \in (0, 1]$, the function w is unbounded off compact sets. Indeed, since w is continuous, it is enough to show that, in the level set $\{\xi : w(\xi) \leq \gamma\}$, $|\mu|$ is bounded and $|\theta_i|$ is bounded for each $i \in \{1, 2, \ldots, q\}$. Note that $w_2 \to \infty$ as $|\theta_i| \to \infty$, and hence we have the θ_is contained. Now, given that the θ_is are contained, $w_1 \to \infty$ as $|\mu| \to \infty$, so μ is contained as well.

To keep the notation under control, we use w and $\tilde{w}$ to denote $w(\xi)$ and $w(\tilde{\xi})$, respectively. The left-hand side of (3–8) is

$$\mathrm{E}\big(w(\tilde{\xi}) \mid \xi\big) = \mathrm{E}(\tilde{w} \mid \xi) = \epsilon\, \mathrm{E}\big(\tilde{w}_1^s \mid \xi\big) + \mathrm{E}\big(\tilde{w}_2^s \mid \xi\big) \ .$$

Equation 3–7 shows that we can get the next state, $\tilde{\xi}$, by first drawing $\sigma^2 \sim \pi(\cdot \mid \xi)$, and then drawing $\tilde{\xi} \sim \pi(\cdot \mid \sigma^2)$, so graphically we have $\xi \to \sigma^2 \to \tilde{\xi}$. This allows us to calculate the expectations above by conditioning on σ^2. Indeed, for $k \in \{1, 2\}$, we have

$$\mathrm{E}(\tilde{w}_k^s \mid \xi) = \mathrm{E}\big[\mathrm{E}(\tilde{w}_k^s \mid \sigma^2, \xi) \mid \xi\big] = \mathrm{E}\big[\mathrm{E}(\tilde{w}_k^s \mid \sigma^2) \mid \xi\big] \tag{3–9}$$

where the second equality follows from the fact that $\tilde{\xi}$ is conditionally independent of ξ given σ^2.

Since there are no restrictions on the constant L in (3–8), we do not have to keep track of any constants when calculating $\mathrm{E}(\tilde{w} \mid \xi)$. Hence, we will use the notation "*const*" to refer to a generic constant. The following Lemma provides upper and lower bounds on t, which will come in handy.

Lemma 3. *Let* $m^* = \max\{m_1, \ldots, m_q\}$, *then*

1.

$$t = \sum_{i=1}^{q} \frac{m_i}{m_i\sigma_\theta^2 + \sigma_e^2} \geq \left(\sum_{i=1}^{q} \frac{m_i}{m_i + 1}\right) \frac{1}{\max\left\{\sigma_\theta^2, \sigma_e^2\right\}} \ ; \tag{3–10}$$

2.

$$\frac{m^*\sigma_\theta^2 + \sigma_e^2}{M} \geq \frac{1}{t} \geq \frac{m_i\sigma_\theta^2 + \sigma_e^2}{M(1 + m_i)} \quad \textit{for each } i = 1, ..., q. \tag{3–11}$$

Proof. Equation (3–10) is true because

$$t = \sum_{i=1}^{q} \frac{m_i}{m_i\sigma_\theta^2 + \sigma_e^2} \geq \sum_{i=1}^{q} \frac{m_i}{(1+m_i)\max\{\sigma_e^2, \sigma_\theta^2\}} = \left(\sum_{i=1}^{q} \frac{m_i}{m_i+1}\right) \frac{1}{\max\left\{\sigma_\theta^2, \sigma_e^2\right\}} .$$

As for equation (3–11), the first inequality holds because

$$\frac{1}{t} = \left(\sum_{i=1}^{q} \frac{m_i}{m_i\sigma_\theta^2 + \sigma_e^2}\right)^{-1} \leq \left(\sum_{i=1}^{q} \frac{m_i}{m^*\sigma_\theta^2 + \sigma_e^2}\right)^{-1} = \frac{m^*\sigma_\theta^2 + \sigma_e^2}{M} .$$

To prove the second inequality, note that

$$t = \sum_{i=1}^{q} \frac{m_i}{m_i\sigma_\theta^2 + \sigma_e^2} \leq \sum_{i=1}^{q} \frac{m_i}{m_i\sigma_\theta^2} = \frac{q}{\sigma_\theta^2} \quad \text{and} \quad t = \sum_{i=1}^{q} \frac{m_i}{m_i\sigma_\theta^2 + \sigma_e^2} \leq \sum_{i=1}^{q} \frac{m_i}{\sigma_e^2} = \frac{M}{\sigma_e^2} .$$

If $\sigma_\theta^2 \leq \sigma_e^2$, then

$$\frac{1}{m_i\sigma_\theta^2 + \sigma_e^2}\frac{1}{t} \geq \frac{\sigma_e^2}{M(m_i\sigma_\theta^2 + \sigma_e^2)} \geq \frac{1}{M(1+m_i)} ,$$

else if $\sigma_\theta^2 > \sigma_e^2$, then

$$\frac{1}{m_i\sigma_\theta^2 + \sigma_e^2}\frac{1}{t} \geq \frac{\sigma_\theta^2}{q\left(m_i\sigma_\theta^2 + \sigma_e^2\right)} > \frac{1}{q\left(1+m_i\right)} \geq \frac{1}{M\left(1+m_i\right)} .$$

□

Using Lemma 3 along with the expressions in (3–6), we have

$$\begin{aligned}
\mathrm{E}(\tilde{w}_1|\sigma^2) &= \mathrm{E}\left[\sum_i (\tilde{\theta}_i - \tilde{\mu})^2 \middle| \sigma_\theta^2, \sigma_e^2\right] \\
&= \sum_i \left[\mathrm{Var}\left[(\tilde{\theta}_i - \tilde{\mu})\middle|\sigma_\theta^2, \sigma_e^2\right] + \left(\mathrm{E}\left[(\tilde{\theta}_i - \tilde{\mu})\middle|\sigma_\theta^2, \sigma_e^2\right]\right)^2\right] \\
&= \sum_i \frac{\sigma_\theta^2\sigma_e^2}{\sigma_e^2 + m_i\sigma_\theta^2} + \sum_i \frac{\left(\sigma_e^2\right)^2}{(m_i\sigma_\theta^2 + \sigma_e^2)^2 t} - 2\sum_i \frac{\sigma_e^2}{(m_i\sigma_\theta^2 + \sigma_e^2)t} + \frac{q}{t} + const \\
&\leq \sum_i \frac{\sigma_\theta^2\sigma_e^2}{m_i\sigma_\theta^2 + \sigma_e^2} - \sum_i \frac{\sigma_e^2}{(m_i\sigma_\theta^2 + \sigma_e^2)t} + \frac{q}{t} + const \\
&\leq \sum_i \frac{\sigma_e^2}{m_i} - \sum_i \frac{\sigma_e^2}{M(1+m_i)} + \frac{q}{t} + const
\end{aligned}$$

where the final inequality uses (3–11). We now bound q/t in two different ways. On one hand, by equation (3–10), we get

$$\frac{q}{t} \leq q\left(\sum_{i=1}^{q} \frac{m_i}{m_i+1}\right)^{-1} \max\left\{\sigma_\theta^2, \sigma_e^2\right\} \leq q\left(\sum_{i=1}^{q} \frac{m_i}{m_i+1}\right)^{-1} \left(\sigma_\theta^2 + \sigma_e^2\right) .$$

On the other hand, applying equation (3–11), we get

$$\frac{q}{t} \leq \frac{qm^*\sigma_\theta^2}{M} + \frac{q\sigma_e^2}{M} .$$

Consequently,

$$\begin{aligned}\mathrm{E}(\tilde{w}_1|\sigma^2) \leq\ & q\left(\sum_{i=1}^{q} \frac{m_i}{m_i+1}\right)^{-1} \sigma_\theta^2 \\ & + \left[\sum_i \frac{1}{m_i} - \sum_i \frac{1}{M(1+m_i)} + q\left(\sum_{i=1}^{q} \frac{m_i}{m_i+1}\right)^{-1}\right]\sigma_e^2 + const ,\end{aligned}$$

and

$$\mathrm{E}(\tilde{w}_1|\sigma^2) \leq \frac{qm^*}{M}\sigma_\theta^2 + \left[\sum_i \frac{1}{m_i} - \sum_i \frac{1}{M(1+m_i)} + \frac{q}{M}\right]\sigma_e^2 + const .$$

Now for the function $\tilde{w}_2$,

$$\begin{aligned}\mathrm{E}(\tilde{w}_2|\sigma^2) &= \sum_i m_i \mathrm{E}\left[(\overline{y}_i - \tilde{\theta}_i)^2 \middle| \sigma_\theta^2, \sigma_e^2\right] \\ &= \sum_i m_i \left[\mathrm{Var}(\tilde{\theta}_i \middle| \sigma_\theta^2, \sigma_e^2) + \left(\mathrm{E}\left[(\tilde{\theta}_i - \overline{y}_i) \middle| \sigma_\theta^2, \sigma_e^2\right]\right)^2\right] \\ &= \sum_i \frac{m_i\sigma_\theta^2\sigma_e^2}{m_i\sigma_\theta^2 + \sigma_e^2} + \sum_i \frac{m_i\left(\sigma_e^2\right)^2}{(m_i\sigma_\theta^2 + \sigma_e^2)^2 t} + const \\ &\leq \sum_i \sigma_e^2 + \sum_i \frac{m_i\sigma_e^2}{(m_i\sigma_\theta^2 + \sigma_e^2)t}\sigma_e^2 = (q+1)\sigma_e^2 + const .\end{aligned}$$

In summary, we have shown that

$$\mathrm{E}(\tilde{w}_1 \mid \sigma^2) \leq \Delta_1\sigma_\theta^2 + \Delta_2\sigma_e^2 + const \quad \text{and} \quad \mathrm{E}(\tilde{w}_2 \mid \sigma^2) \leq (q+1)\sigma_e^2 + const$$

where

$$\Delta_1 = \min\left\{q\left(\sum_{i=1}^{q} \frac{m_i}{m_i+1}\right)^{-1}, \frac{qm^*}{M}\right\}$$

and

$$\Delta_2 = \sum_{i=1}^{q} \frac{1}{m_i} - \sum_{i=1}^{q} \frac{1}{M(1+m_i)} + \max\left\{ q\left(\sum_{i=1}^{q} \frac{m_i}{m_i+1} \right)^{-1}, \frac{q}{M} \right\} .$$

To continue the calculation in (3–9), note that for $s \in (0,1]$ and any $A, B > 0$, it follows easily that $(A+B)^s \leq A^s + B^s$. Together with Jensen's inequality, this yields

$$\mathrm{E}(\tilde{w}_1^s \mid \sigma^2) \leq \big[\mathrm{E}(\tilde{w}_1 \mid \sigma^2)\big]^s \leq \big(\Delta_1\sigma_\theta^2 + \Delta_2\sigma_e^2 + const\big)^s \leq \Delta_1^s\big(\sigma_\theta^2\big)^s + \Delta_2^s\big(\sigma_e^2\big)^s + const \ , \quad (3\text{–}12)$$

and

$$\mathrm{E}(\tilde{w}_2^s \mid \sigma^2) \leq \big[\mathrm{E}(\tilde{w}_2 \mid \sigma^2)\big]^s \leq \big((q+1)\sigma_e^2 + const\big)^s \leq (q+1)^s\big(\sigma_e^2\big)^s + const \ . \quad (3\text{–}13)$$

Now to complete the calculation in (3–9), recall that

$$\sigma_\theta^2 \mid \xi \sim \mathrm{IG}\left(\frac{q}{2} + a, \frac{w_1}{2}\right) \quad \text{and} \quad \sigma_e^2 \mid \xi \sim \mathrm{IG}\left(\frac{M}{2} + b, \frac{w_2 + \mathrm{SSE}}{2}\right) ,$$

where $\mathrm{SSE} = \sum_{i,j}(y_{ij} - \overline{y}_i)^2$. This is where we have to make sure that $s \in (0,1]$ is not too large. Define the set

$$S = (0,1] \cap \left(0, \min\left\{\frac{q}{2} + a, \frac{M}{2} + b\right\}\right) .$$

Then, for any $s \in S$, $\mathrm{E}\Big(\big(\sigma_\theta^2\big)^s \mid \xi\Big)$ and $\mathrm{E}\Big(\big(\sigma_e^2\big)^s \mid \xi\Big)$ are both finite. In fact, routine calculations show that

$$\mathrm{E}\Big(\big(\sigma_\theta^2\big)^s \mid \xi\Big) = \frac{\Gamma(\frac{q}{2} + a - s)}{2^s\Gamma(\frac{q}{2} + a)} w_1^s \ , \quad (3\text{–}14)$$

and

$$\mathrm{E}\Big(\big(\sigma_e^2\big)^s \mid \xi\Big) = \frac{\Gamma(\frac{M}{2} + b - s)}{2^s\Gamma(\frac{M}{2} + b)}(w_2 + \mathrm{SSE})^s \leq \frac{\Gamma(\frac{M}{2} + b - s)}{2^s\Gamma(\frac{M}{2} + b)} w_2^s + const \ . \quad (3\text{–}15)$$

Define

$$\delta_1(s) = \frac{(q+1)^s\Gamma(\frac{M}{2} + b - s)}{2^s\Gamma(\frac{M}{2} + b)} \ , \quad \delta_2(s) = \frac{\Delta_2^s\Gamma(\frac{M}{2} + b - s)}{2^s\Gamma(\frac{M}{2} + b)} \quad \text{and} \quad \delta_3(s) = \frac{\Delta_1^s\Gamma(\frac{q}{2} + a - s)}{2^s\Gamma(\frac{q}{2} + a)} \ .$$

Combining (3–9) and (3–12)–(3–15), we have

$$\begin{aligned}
\mathrm{E}(\tilde{w}_1^s \mid \xi) &\le \mathrm{E}\Big(\Delta_1^s\big(\sigma_\theta^2\big)^s + \Delta_2^s\big(\sigma_e^2\big)^s + const \,\big|\, \xi\Big) \\
&\le \Delta_1^s \frac{\Gamma(\frac{q}{2}+a-s)}{2^s\Gamma(\frac{q}{2}+a)} w_1^s + \Delta_2^s \frac{\Gamma(\frac{M}{2}+b-s)}{2^s\Gamma(\frac{M}{2}+b)} w_2^s + const \qquad (3\text{–}16) \\
&= \delta_3(s) w_1^s + \delta_2(s) w_2^s + const\ .
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}(\tilde{w}_2^s \mid \xi) &\le \mathrm{E}\Big((q+1)^s\big(\sigma_e^2\big)^s + const \,\big|\, \xi\Big) \\
&\le (q+1)^s \frac{\Gamma(\frac{M}{2}+b-s)}{2^s\Gamma(\frac{M}{2}+b)} w_2^s + const \qquad (3\text{–}17) \\
&= \delta_1(s) w_2^s + const\ .
\end{aligned}$$

The following result explains how the inequalities (3–16) and (3–17) can be used together to form a valid drift condition.

Proposition 3. *Fix $s \in S$. If $\delta_1(s) < 1$ and $\delta_3(s) < 1$, then there exist $\epsilon > 0$, $\rho < 1$ and $L < \infty$ such that*

$$E\big(w(\tilde{\xi}) \mid \xi\big) \le \rho w(\xi) + L \quad \textit{for all } \xi \in \mathbb{R}^{q+1}.$$

Proof. It follows from (3–16) and (3–17) that

$$\begin{aligned}
\mathrm{E}(\epsilon\tilde{w}_1^s + \tilde{w}_2^s \mid \xi) &\le \epsilon\delta_3(s) w_1^s + (\delta_1(s) + \epsilon\delta_2(s)) w_2^s + const \\
&= \rho(\epsilon, s)(\epsilon w_1^s + w_2^s) + \epsilon(\delta_3(s) - \rho(\epsilon, s)) w_1^s + const
\end{aligned}$$

where $\rho(\epsilon, s) = \delta_1(s) + \epsilon\delta_2(s)$. Therefore, we will have a viable drift condition if

$$\rho(\epsilon, s) < 1 \ \text{ and } \ \delta_3(s) - \rho(\epsilon, s) \le 0\ . \qquad (3\text{–}18)$$

Clearly, (3–18) requires that $\delta_1(s) < 1$ and $\delta_3(s) < 1$. We now show that these conditions are also sufficient for the existence of $\epsilon > 0$ such that (3–18) is satisfied.

There are two cases. In the first case, $\delta_1(s) \le \delta_3(s) < 1$. If we take $\epsilon = (\delta_3(s) - \delta_1(s))/\delta_2(s)$, then $\rho(\epsilon, s) = \delta_3(s) < 1$ and $\delta_3(s) - \rho(\epsilon, s) = 0$. In the second case,

$\delta_3(s) < \delta_1(s) < 1$. Now take $\epsilon = \big(1 - \delta_1(s)\big)/\big(2\delta_2(s)\big)$. Then

$$\rho(\epsilon, s) = \delta_1(s) + \frac{1 - \delta_1(s)}{2} = \frac{1 + \delta_1(s)}{2} < 1 \ ,$$

and

$$\delta_3(s) - \rho(\epsilon, s) = \delta_3(s) - \frac{1 + \delta_1(s)}{2} < 0 \ .$$

Hence, if $\delta_1(s) < 1$ and $\delta_3(s) < 1$, then there is a viable drift condition. □

In conjunction with Proposition 1, Proposition 3 shows that the ξ-chain (and hence the block Gibbs Markov chain) is geometrically ergodic as long as there exists an $s \in S$ such that both $\delta_1(s)$ and $\delta_3(s)$ are less than 1. The following result shows when this can happen. See Appendix C for the proof.

Proposition 4. *Let $\Psi(x) = \frac{d}{dx} \log\big(\Gamma(x)\big)$ denote the digamma function. If $M + 2b \geq q + 3$ and $\Delta_1 < 2\exp\big(\Psi\big(\frac{q}{2} + a\big)\big)$, then there exists $s \in S$ such that $\delta_1(s) < 1$ and $\delta_3(s) < 1$.*

Finally, we can state our main convergence rate result.

Proposition 5. *The Markov chain underlying the block Gibbs sampler is geometrically ergodic if*

1. $\Delta_1 < 2\exp\Big(\Psi\big(\frac{q}{2} + a\big)\Big)$, and
2. $M + 2b \geq q + 3$,

where $\Delta_1 = q \min\left\{ \left(\sum_{i=1}^{q} \frac{m_i}{m_i+1} \right)^{-1}, \frac{m^*}{M} \right\}$.

Loosely speaking, Proposition 5 shows that geometric ergodicity holds unless the data set is small and unbalanced at the same time. Indeed, consider the first condition. The left-hand side will be large (and the condition will fail) only if *both* $\big(\sum_{i=1}^{q} \frac{m_i}{m_i+1}\big)^{-1}$ and $\frac{m^*}{M}$ are large. The first term increases as the m_is get smaller and the second term increases as m^* gets larger relative to M; that is, as the data become more unbalanced. The second condition is a weak condition on the sample size.

Finally, we show that Corollary 1 of Chapter 1, which deals with the special case of $(a, b) = (-\frac{1}{2}, 0)$, follows straightforwardly from Proposition 5. It follows from (3–2) that,

when $a = -\frac{1}{2}$, the posterior is improper if $q \leq 2$. Hence, we only care about $q \geq 3$. When $q \geq 4$, we have

$$2\exp\left(\Psi\left(\frac{q}{2}+a\right)\right) \geq 2\exp\left(\Psi\left(\frac{3}{2}\right)\right) = 2\exp(-\gamma - 2(\log 2 - 1)) \doteq 2.074 \ .$$

Since $\frac{m_i}{1+m_i} \geq \frac{1}{2}$, we have

$$\Delta_1 \leq \frac{q}{\sum_i \frac{m_i}{m_i+1}} \leq \frac{q}{\sum_i \frac{1}{2}} = 2 < 2\exp\left(\Psi\left(\frac{q}{2}+a\right)\right) \ ,$$

so, when $a = -\frac{1}{2}$ and $q \geq 4$, the condition $\Delta_1 < 2\exp\left(\Psi\left(\frac{q}{2}+a\right)\right)$ is *always satisfied*.

Now, when $q = 3$, we have

$$2\exp\left(\Psi\left(\frac{q}{2}+a\right)\right) = 2\exp\left(\Psi(1)\right) = 2\exp(-\gamma) \doteq 1.123 \ .$$

For balanced data,

$$\Delta_1 \leq \frac{qm^*}{M} = 1 < 2\exp\left(\Psi(1)\right) \ .$$

Hence, when $a = -\frac{1}{2}$ and $q = 3$ and the data are balanced, the condition $\Delta_1 < 2\exp\left(\Psi\left(\frac{q}{2}+a\right)\right)$ is satisfied. Finally, if $q = 3$ and the data are unbalanced, then $\Delta_1 < 2\exp(-\gamma)$ if and only if

$$\sum_i \frac{m_i}{m_i+1} > \frac{3}{2\exp(-\gamma)} \doteq 2.67 \quad \text{or} \quad m^* < \frac{2\exp(-\gamma)}{3} M \doteq 0.374M \ . \tag{3–19}$$

Table 1-1 in Chapter 1 provides a list of all unbalanced data configurations with $q = 3$ and $m^* \leq 12$ that satisfy (3–19).

3.3 Minorization for the Block Gibbs Sampler

Recall from Chapter 2 that regenerative simulation of a Markov chain allow us to compute asymptotically valid standard errors for the MCMC estimators. In this section, we construct a minorization condition that drives the regenerative simulation of our block Gibbs sampler. The idea that we use to find this minorization condition is outlined in Mykland et al. (1995). Recall that the transition density of the block Gibbs chain is given

by

$$k\big(\tilde{\sigma}^2, \tilde{\xi} \mid \sigma^2, \xi\big) = \pi\big(\tilde{\sigma}^2 \mid \xi\big)\, \pi\big(\tilde{\xi} \mid \tilde{\sigma}^2\big)\ .$$

Fix $0 < d_1 < d_2 < \infty$ and $0 < d_3 < d_4 < \infty$ and let D denote the closed rectangle $[d_1, d_2] \times [d_3, d_4] \subset \mathbb{R}^2_+$. Also, fix a *distinguished point* $\xi^* \in \mathbb{R}^{q+1}$. Then

$$\begin{aligned}
k(\tilde{\sigma}^2, \tilde{\xi} \mid \sigma^2, \xi) &= \frac{\pi(\tilde{\sigma}^2 \mid \xi)}{\pi(\tilde{\sigma}^2 \mid \xi^*)} \pi(\tilde{\xi} \mid \tilde{\sigma}^2) \pi(\tilde{\sigma}^2 \mid \xi^*) \\
&\geq \left[\inf_{\sigma^2 \in D} \frac{\pi(\sigma^2 \mid \xi)}{\pi(\sigma^2 \mid \xi^*)} \right] \pi(\tilde{\xi} \mid \tilde{\sigma}^2) \pi(\tilde{\sigma}^2 \mid \xi^*) I_D(\tilde{\sigma}^2) \\
&= \left\{ c \frac{\pi(\underline{\sigma}^2 \mid \xi)}{\pi(\underline{\sigma}^2 \mid \xi^*)} \right\} \left\{ \frac{1}{c} \pi(\tilde{\xi} \mid \tilde{\sigma}^2) \pi(\tilde{\sigma}^2 \mid \xi^*) I_D(\tilde{\sigma}^2) \right\} \\
&=: s(\xi)\, \nu(\tilde{\sigma}^2, \tilde{\xi})
\end{aligned}$$

where $\underline{\sigma}^2 = (\underline{\sigma}^2_\theta, \underline{\sigma}^2_e)$ denotes the minimizer of $\pi(\sigma^2 \mid \xi)/\pi(\sigma^2 \mid \xi^*)$ as σ^2 ranges over the set D, and c is the normalizing constant; that is,

$$c = \int_{\mathbb{R}^2_+} \int_{\mathbb{R}^{q+1}} \pi(\xi \mid \sigma^2) \pi(\sigma^2 \mid \xi^*) I_D(\sigma^2)\, d\xi\, d\sigma^2 = \left[\int_{d_1}^{d_2} \pi(\sigma^2_\theta \mid \xi^*)\, d\sigma^2_\theta \right] \left[\int_{d_3}^{d_4} \pi(\sigma^2_e \mid \xi^*)\, d\sigma^2_e \right] .$$

The value of c is actually not required in practice. Indeed, the probability (2–3), which must be calculated after each iteration of the Markov chain, involves s and ν only through their product. Thus, c cancels out.

We now develop a closed form expression for s. Since $\pi(\sigma^2 \mid \xi)$ factors into $\pi(\sigma^2_\theta \mid \xi)$ and $\pi(\sigma^2_e \mid \xi)$, the bivariate minimization problem becomes two separate univariate minimization problems. Let w^*_k stand for w_k evaluated at ξ^* for $k = 1, 2$. Then

$$\begin{aligned}
\frac{\pi(\sigma^2_\theta \mid \xi)}{\pi(\sigma^2_\theta \mid \xi^*)} &= \frac{(\frac{1}{2} w_1)^{\frac{q}{2}+a} / \Gamma(\frac{q}{2} + a)}{(\frac{1}{2} w^*_1)^{\frac{q}{2}+a} / \Gamma(\frac{q}{2} + a)} \frac{\big(\sigma^2_\theta\big)^{-(\frac{q}{2}+a+1)} \exp[-\frac{1}{2} w_1 / \sigma^2_\theta]}{\big(\sigma^2_\theta\big)^{-(\frac{q}{2}+a+1)} \exp[-\frac{1}{2} w^*_1 / \sigma^2_\theta]} \\
&= \left(\frac{w_1}{w^*_1} \right)^{\frac{q}{2}+a} \exp\left[-\frac{1}{2} (w_1 - w^*_1) / \sigma^2_\theta \right] ,
\end{aligned}$$

and

$$\frac{\pi(\sigma_e^2 \mid \xi)}{\pi(\sigma_e^2 \mid \xi^*)} = \frac{\left[\frac{1}{2}(w_2 + \mathrm{SSE})\right]^{\frac{M}{2}+b} / \Gamma(\frac{M}{2}+b)}{\left[\frac{1}{2}(w_2^* + \mathrm{SSE})\right]^{\frac{M}{2}+b} / \Gamma(\frac{M}{2}+b)} \frac{\left(\sigma_e^2\right)^{-(\frac{M}{2}+b+1)} \exp\left[-\frac{1}{2}(w_2 + \mathrm{SSE})/\sigma_e^2\right]}{\left(\sigma_e^2\right)^{-(\frac{M}{2}+b+1)} \exp\left[-\frac{1}{2}(w_2^* + \mathrm{SSE})/\sigma_e^2\right]}$$
$$= \left(\frac{w_2 + \mathrm{SSE}}{w_2^* + \mathrm{SSE}}\right)^{\frac{M}{2}+b} \exp\left[-\frac{1}{2}(w_2 - w_2^*)/\sigma_e^2\right].$$

Hence, $\underline{\sigma}_\theta^2 = d_1$ when $w_1 > w_1^*$ and $\underline{\sigma}_\theta^2 = d_2$ when $w_1 \leq w_1^*$. Similarly, $\underline{\sigma}_e^2 = d_3$ when $w_2 > w_2^*$ and $\underline{\sigma}_e^2 = d_4$ when $w_2 \leq w_2^*$. Finally,

$$\begin{aligned}
&\Pr\Big(\delta_n = 1 \;\Big|\; (\sigma_n^2, \xi_n) = (\sigma^2, \xi), (\sigma_{n+1}^2, \xi_{n+1}) = (\tilde{\sigma}^2, \tilde{\xi})\Big) \\
&= \frac{s(\xi)\,\nu(\tilde{\sigma}^2, \tilde{\xi})}{k\big(\tilde{\sigma}^2, \tilde{\xi} \mid \sigma^2, \xi\big)} = \frac{\pi(\underline{\sigma}^2 \mid \xi)\pi(\tilde{\sigma}^2 \mid \xi^*)}{\pi(\underline{\sigma}^2 \mid \xi^*)\pi(\tilde{\sigma}^2 \mid \xi)} I_D(\tilde{\sigma}^2) \\
&= \left(\frac{w_1}{w_1^*}\right)^{\frac{q}{2}+a} \exp\left[-\frac{1}{2}(w_1 - w_1^*)/\underline{\sigma}_\theta^2\right] \left(\frac{w_2 + \mathrm{SSE}}{w_2^* + \mathrm{SSE}}\right)^{\frac{M}{2}+b} \exp\left[-\frac{1}{2}(w_2 - w_2^*)/\underline{\sigma}_e^2\right] \\
&\times \left(\frac{w_1^*}{w_1}\right)^{\frac{q}{2}+a} \exp\left[-\frac{1}{2}(w_1^* - w_1)/\tilde{\sigma}_\theta^2\right] \left(\frac{w_2^* + \mathrm{SSE}}{w_2 + \mathrm{SSE}}\right)^{\frac{M}{2}+b} \exp\left[-\frac{1}{2}(w_2^* - w_2)/\tilde{\sigma}_e^2\right] I_D(\tilde{\sigma}^2) \\
&= \exp\left\{\frac{1}{2}\left[(w_1 - w_1^*)\left(\frac{1}{\tilde{\sigma}_\theta^2} - \frac{1}{\underline{\sigma}_\theta^2}\right) + (w_2 - w_2^*)\left(\frac{1}{\tilde{\sigma}_e^2} - \frac{1}{\underline{\sigma}_e^2}\right)\right]\right\} I_D(\tilde{\sigma}^2)\,.
\end{aligned} \tag{3–20}$$

Theoretically, we could use any set $D = [d_1, d_2] \times [d_3, d_4]$ and any distinguished point ξ^* to run the regenerative simulation. However, the asymptotics for $\hat{\gamma}^2$ involve $R \to \infty$, so we would like for the chain to regenerate fairly often. Thus, we should choose D and ξ^* so that the probability in (3–20) is frequently close to one. Not surprisingly, there is trade-off between the size of the set D and the magnitude of the exponential term in (3–20) (when the indicator is unity). Our strategy for choosing D and ξ^* is as follows. We run the block Gibbs sampler for an initial n_0 iterations (using starting value $\xi_0 = (\overline{y}_1, \ldots, \overline{y}_q, \overline{y})$ for example). We take $[d_1, d_2]$ to be the shortest interval that contains 60% of the n_0 values of σ_θ^2, and we calculate $[d_3, d_4]$ similarly using the n_0 values of σ_e^2. The regeneration probability (3–20) involves ξ^* only through $w_1(\xi^*)$ and $w_2(\xi^*)$. Hence, instead of setting ξ^* equal to the median, say, of the n_0 values of ξ in the initial run of the chain, we calculate w_1 and w_2 for each of the n_0 values of ξ and we set w_1^* to be the median of the w_1 values

and w_2^* to be the median of the w_2 values. There is one small caveat. This approach makes sense only if there happens to exist a $\hat{\xi} \in \mathbb{R}^{q+1}$ such that $\big(w_1(\hat{\xi}), w_2(\hat{\xi})\big) = (w_1^*, w_2^*)$. For balanced data, such a $\hat{\xi}$ exists if and only if $\sqrt{mw_1^*} + \sqrt{w_2^*} \geq \sqrt{\text{SST}}$, where m is the number of observations in each group and $\text{SST} = m\sum(\overline{y}_i - \overline{y})^2$ is the treatment sum of squares. See Appendix D for a proof of this result as well as guidelines for the unbalanced case.

3.4 An Example: Styrene Exposure Data

In this section, we analyze a real data set from Lyles et al. (1997) using the Bayesian random effects models. Thirteen workers were randomly selected from a group within a boat manufacturing plant and each one's styrene exposure was measured on three separate occasions. So we have $q = 13$, $m_i \equiv m = 3$ and $M = m \times q = 39$. The data are summarized in Tables 3-1 and 3-2.

Table 3-1. Average styrene exposure level for each of the 13 workers.

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

Table 3-2. Summary statistics for the styrene exposure data

$\overline{y} = M^{-1}\sum_{i=1}^{13}\sum_{j=1}^{3} y_{ij} = 4.809$
$\text{SST} = 3\sum_{i=1}^{13}(\overline{y}_i - \overline{y})^2 = 11.430$
$\text{SSE} = \sum_{i=1}^{13}\sum_{j=1}^{3}(y_{ij} - \overline{y}_i)^2 = 14.711$

We would like to investigate the effect of using two different priors in the model. They are the standard diffuse prior, $p = p_{-\frac{1}{2},0}$, from equation (1–2), and the reference prior, $p^* = p_r$, defined in (1–3). Denote their associated posterior distributions by $\pi(\theta, \mu, \sigma_\theta^2, \sigma_e^2)$ and $\pi^*(\theta, \mu, \sigma_\theta^2, \sigma_e^2)$, respectively. We know that both posteriors are well-defined (proper) since the data set is balanced and $q = 13 \geq 3$. We will focus on the posterior expectations of three functions: $g_1(\mu, \sigma_\theta^2, \sigma_e^2) = \sigma_\theta^2$, $g_2(\mu, \sigma_\theta^2, \sigma_e^2) = \sigma_e^2$, and $g_3(\mu, \sigma_\theta^2, \sigma_e^2) = \sigma_\theta^2/(\sigma_\theta^2 + \sigma_e^2)$. Here, g_3 is the correlation between observations on the same worker. All three functions

have finite expectations with respect to π. Actually, they all satisfy the stronger "$2+\varepsilon$" moment condition, $\mathrm{E}_\pi|g_i|^{2+\varepsilon} < \infty$. This is clearly true for g_3 since it is bounded above by 1. And a straightforward application of Proposition 6 from Appendix A implies that $\mathrm{E}_\pi|g_1|^3 < \infty$ and $\mathrm{E}_\pi|g_2|^3 < \infty$. Therefore, all of the assumptions underlying the regenerative simulation of the Block Gibbs chain are satisfied. This simulation can be used to produce estimates and standard errors for $\mathrm{E}_\pi g_i(\mu, \sigma_\theta^2, \sigma_e^2)$, $i = 1, 2, 3$.

Next consider reusing the block Gibbs output to approximate $\mathrm{E}_{\pi^*} g_i(\mu, \sigma_\theta^2, \sigma_e^2)$ for $i = 1, 2, 3$. According to the discussion following Proposition 2, our importance sampling results are applicable if we can find an $\varepsilon > 0$ such that $\mathrm{E}_\pi|g_i\, p^*/p|^{2+\varepsilon} < \infty$ for $i = 0, 1, 2, 3$, where $g_0 \equiv 1$. First, note that $|g_i p^*/p| = g_i\, p^*/p$ since all the terms are positive. Now

$$\frac{p^*}{p} = \frac{\left(\sigma_\theta^2\right)^{-\frac{C_3}{2}}\left(\sigma_e^2\right)^{-1}\left[2+\left(\frac{\sigma_e^2}{\sigma_e^2+3\sigma_\theta^2}\right)^2\right]^{\frac{1}{2}}}{\left(\sigma_\theta^2\right)^{-\frac{1}{2}}\left(\sigma_e^2\right)^{-1}} = \left(\sigma_\theta^2\right)^{\frac{1-C_3}{2}}\left[2+\left(\frac{\sigma_e^2}{\sigma_e^2+3\sigma_\theta^2}\right)^2\right]^{\frac{1}{2}} \leq \sqrt{3}\left(\sigma_\theta^2\right)^{\frac{1-C_3}{2}},$$

where $C_3 \doteq 0.96$. This inequality together with Proposition 6 shows that $\mathrm{E}_\pi|g_i\, p^*/p|^3 < \infty$ for $i = 0, 1, 2, 3$. Hence, our importance sampling technique is applicable.

Implementation of the regenerative simulation requires us to specify R, the total number of regenerations; that is, the number of iid tours in the chain. The procedure to determine R has two steps. In the first step, we run the chain for a preliminary R regenerations that is believed to lead to a reasonable estimator of the asymptotic variance, γ^2. Here, we used $R = 5,000$ which took 87,169 iterations and consumed 20 seconds (coded in R). The simulation results are summarized in Table 3-3. For each of the three functions of interest, the table provides the estimator, $\tilde{g}_R$ (or the importance sampling estimator, $\tilde{\eta}_R$), the estimated asymptotic variance, $\hat{\gamma}^2$, the estimated standard error $\sqrt{\hat{\gamma}^2/R}$, and an approximate 95% CI, $\overline{g}_R \pm 2\sqrt{\hat{\gamma}^2/R}$ (or $\tilde{\eta}_R \pm 2\sqrt{\hat{\gamma}^2/R}$) under both priors. Recall Remark 3 from Chapter 2, it is not recommended to use $\hat{\gamma}^2$ to estimate γ^2 unless the average tour length, $\overline{N} = R^{-1}\sum_{t=1}^{R} N_t$ has a coefficient of variation, $\mathrm{CV}(\bar{N})$ smaller than 1%. For our simulation above, $\widehat{\mathrm{CV}}(\overline{N}) = 0.034\%$ clearly meets the criteria. We also

examined trace plots of $\hat{\gamma}^2$ for the parameters of interest, σ_θ^2, σ_e^2 and $\sigma_\theta^2/(\sigma_\theta^2 + \sigma_e^2)$, and all suggest that the variance estimators have stabilized by the 5,000th regeneration. Hence they are reasonable approximations of their respective estimands.

In the second step of the procedure, we decide how large R needs to be for the resulting CI to be shorter than a user-specified width based on the preliminary analysis above. Take the 95% CI of $\mathrm{E}_\pi \sigma_\theta^2$ for example. Suppose that we desire its margin of error to be around 1% of the magnitude of $\mathrm{E}_\pi \sigma_\theta^2$. Since $\tilde{g}_{5000} = 0.19003$, the desired width of the 95% CI, l, is approximately $2 \times 0.19003 \times 1\% \doteq 0.0038$ and will require about $16\hat{\gamma}^2/l^2 = 16 \times 0.03074/0.0038^2 \doteq 38,371$ regenerations. To take into account the possibility that the asymptotic variance can be slightly underestimated by the $\hat{\gamma}^2$ obtained in the preliminary analysis, we run the chain a little longer than what the calculation suggests. Here, we decide to produce a chain with $R = 40,000$ regenerations. Actually, only 35,000 more iid tours need to be simulated, and then combined with the 5,000 iid tours we already have. The final chain with 40,000 regenerations accounted for 697,869 iterations and took 3 minutes to generate.

The simulation results under both priors are summarized in Table 3-4 and a crude diagnostic was provided using the SRQ plot in Figure 3-1. Further, the trace plots of the estimated asymptotic 95% CIs for $E_\pi \sigma_\theta^2$ and $\hat{\gamma}^2$ are provided in Figure 3-2. These plots suggest that things have stabilized quite nicely by the 40,000th regeneration. Finally, we see from Table 3-4 that the estimated posterior means under the two priors are very close, so the issues regarding choice of the prior raised by Bernardo (1996) are not a concern for this data set.

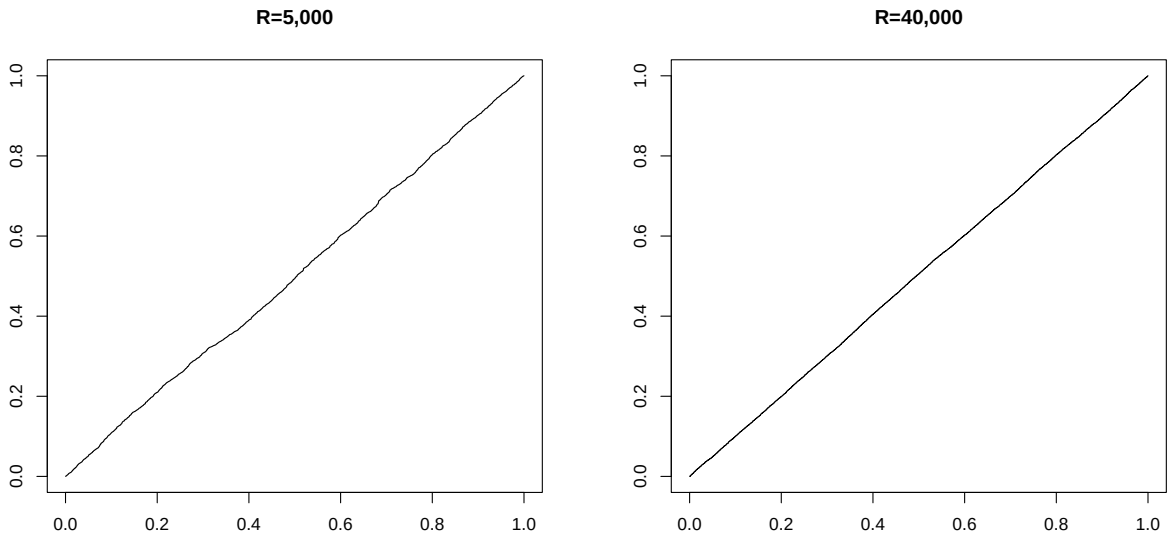

Figure 3-1. SRQ plots of τ_t/τ_R (vertical axes) against t/R (horizontal axes) of the block Gibbs sampler for the styrene exposure data based on runs of length 5,000 and 40,000; the associated $\widehat{CV}(\overline{M})$ are 0.034% and 0.004% respectively.

Table 3-3. In the styrene exposure data analysis, there are three quantities of interest, σ_θ^2, σ_e^2, and $\sigma_\theta^2/(\sigma_\theta^2+\sigma_e^2)$, and two different priors. For each combination of these, the table provides estimates of the posterior expectation and the corresponding asymptotic variance, as well as the standard error and a 95% asymptotic CI. These results are based on $R = 5{,}000$ regenerations.

	Prior	Estimate	$\hat{\gamma}^2$	$\sqrt{\hat{\gamma}^2/R}$	Estimate $\pm 2\sqrt{\hat{\gamma}^2/R}$
σ_θ^2	Diffuse	0.19003	0.03463	0.00263	(0.18477, 0.19529)
	Reference	0.18688	0.25732	0.00717	(0.17253, 0.20123)
σ_e^2	Diffuse	0.61777	0.00883	0.00133	(0.61511, 0.62043)
	Reference	0.61964	0.06613	0.00364	(0.61237, 0.62691)
$\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_e^2}$	Diffuse	0.21288	0.03532	0.00266	(0.20757, 0.21820)
	Reference	0.21005	0.26278	0.00725	(0.19555, 0.22455)

Table 3-4. Results based on $R = 40,000$ regenerations

	Prior	Estimate	$\hat{\gamma}^2$	$\sqrt{\hat{\gamma}^2/R}$	Estimate $\pm 2\sqrt{\hat{\gamma}^2/R}$
σ_θ^2	Diffuse	0.19023	0.03523	0.00094	(0.18835, 0.19210)
	Reference	0.18720	0.03252	0.00090	(0.18540, 0.18900)
σ_e^2	Diffuse	0.61849	0.00966	0.00049	(0.61751, 0.61947)
	Reference	0.62031	0.00897	0.00047	(0.61936, 0.62126)
$\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_e^2}$	Diffuse	0.21304	0.03687	0.00096	(0.21112, 0.21496)
	Reference	0.21033	0.03404	0.00092	(0.20849, 0.21218)

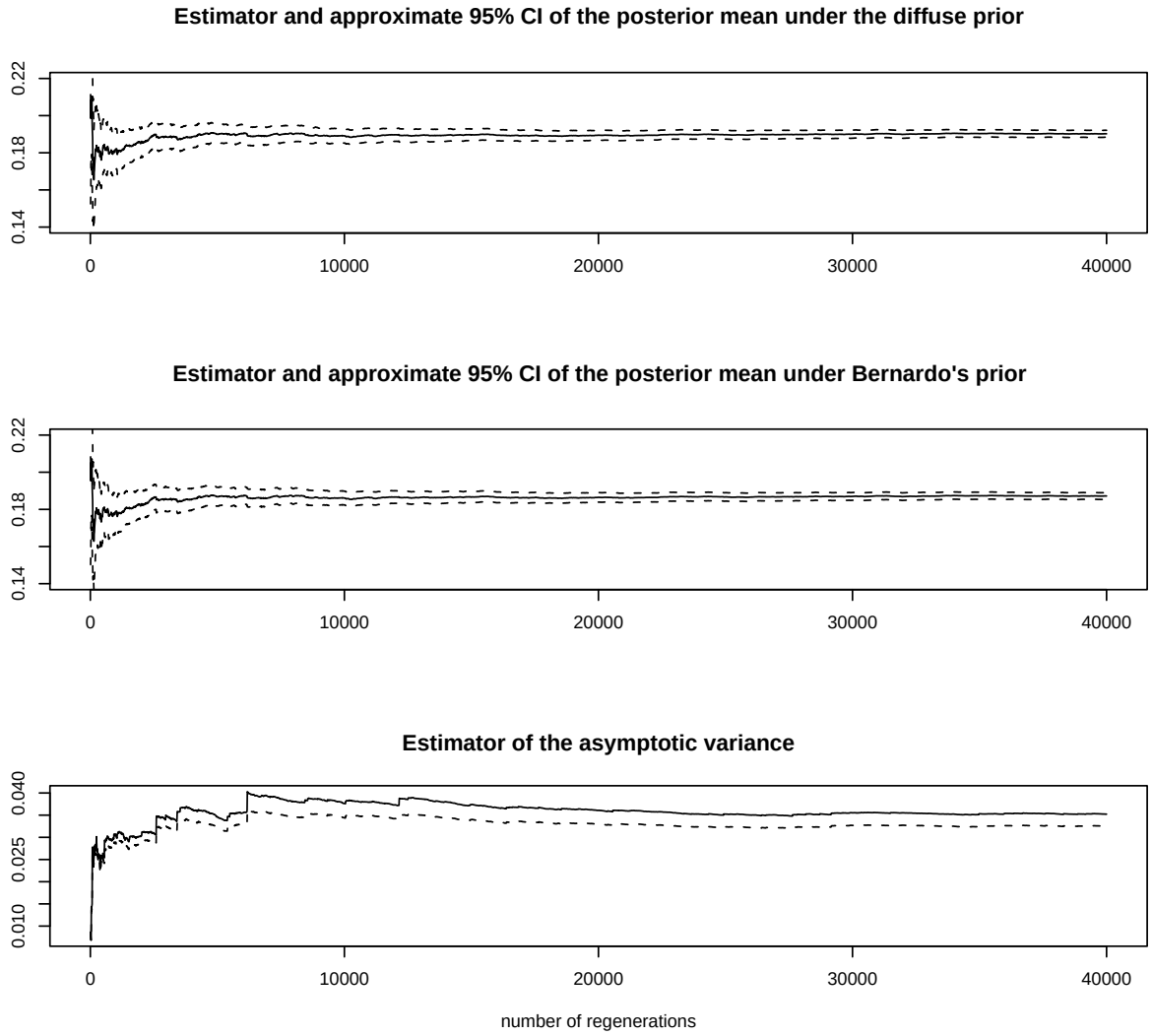

Figure 3-2. The top plot shows the evolution of the estimator of $\mathrm{E}_{\pi}\sigma_{\theta}^{2}$ and the corresponding 95% asymptotic CI as the number of regenerations grows. The solid line represents the estimator and the dashed lines denote the upper and lower endpoints of the CI. The middle plot does the same for $\mathrm{E}_{\pi^*}\sigma_{\theta}^{2}$. The bottom plot displays the evolution of the estimate of the asymptotic variance, $\hat{\gamma}^2$, of the estimators of $\mathrm{E}_{\pi}\sigma_{\theta}^{2}$ (solid line) and $\mathrm{E}_{\pi^*}\sigma_{\theta}^{2}$ (dashed line). Since trace plots of the estimators under the two different priors are calculated using the same run of a Markov chain with the same regeneration times, it should not be surprising that they track each other closely.

CHAPTER 4
ON THE GEOMETRIC ERGODICITY OF TWO-VARIABLE GIBS SAMPLERS

4.1 Introduction

In this chapter, we discuss a slightly different topic. Moreover, the notation we use here is not necessarily consistent with that of the earlier chapters. Suppose we are given a joint density $\pi(x, y)$ on $\mathsf{X} \times \mathsf{Y}$ with respect to some measure $\mu = \mu_X \times \mu_Y$. Suppose we need to analyze the X-marginal density

$$\pi_X(x) = \int_{\mathsf{Y}} \pi(x, y)\, \mu_Y(dy)$$

by evaluating quantities like $\mathrm{E}_{\pi_X} g = \int_{\mathsf{X}} g(x)\, \pi_X(dx)$, where g is some function of interest. As we mentioned in Chapter 1, it is usually impossible to perform the integration directly. Instead, we can approximate the integrals using Monte Carlo methods. Classical Monte Carlo methods are not applicable when it is hard to simulate from π or π_X directly. In such cases, we resort to the MCMC methods. Below, we describe a class of useful MCMC algorithms the transition densities of which have a common structure.

Let $\pi_Y(y) = \int_{\mathsf{X}} \pi(x, y)\, \mu_X(dx)$ denote the Y-marginal density of π. Further, let $\pi_{X|Y}(x|y) = \pi(x, y)/\pi_Y(y)$ and $\pi_{Y|X}(y|x) = \pi(x, y)/\pi_X(x)$ denote the conditional densities. Then the following Markov transition density defines a Markov chain that has invariant distribution π:

$$k(x', y' \mid x, y) = \pi_{X|Y}(x'|y')\, \pi_{Y|X}(y'|x)\ . \tag{4–1}$$

It is possible to simulate the chain if we can sample from the conditionals, $\pi_{X|Y}$ and $\pi_{Y|X}$. Suppose that the state of the chain at time n is $(X_n, Y_n) = (x, y)$. Then the next state, (X_{n+1}, Y_{n+1}), is drawn as follows. First, simulate $X_{n+1} \sim \pi_{X|Y}(\cdot|y)$ and call the result x'; secondly, simulate $Y_{n+1} \sim \pi_{Y|X}(\cdot|x')$. Since simulating each iteration of the Markov chain entails two steps, we refer to this MCMC algorithm as the *two-variable Gibbs sampler* (TGS).

Despite the simple structure of TGS, it is applicable in the posterior analysis of many complex Bayesian models. For instance, the block Gibbs sampler that we studied in Chapter 3 is a TGS. Actually, there is a large family of very useful MCMC algorithms, called the *data augmentation* algorithms (Tanner and Wong, 1987), that fall in this category. The wide application of TGSs as well as their relatively simple structure tempt us to study their convergence rates in a unified manner. As a first small step towards this big goal, we study the simplest possible TGS. Recall that a TGS is determined by the joint density π on $\mathsf{X} \times \mathsf{Y}$. It is well known that if either X or Y is finite, then any irreducible chain on $\mathsf{X} \times \mathsf{Y}$ is also geometrically ergodic. So the situation where both X and Y are countable provides the simplest case for which geometric chains can not be easily characterized. Therefore, we decide to study TGSs that are associated with a special class of densities on $\mathsf{X} \times \mathsf{Y} = \mathbb{N} \times \mathbb{N}$. First, we use the drift method to find a sufficient condition for geometric ergodicity of the Markov chain underlying the TGS. Secondly, and more interestingly, we resort to some operator theory to derive a necessary condition for geometric convergence of the chain. Since the invariant distribution completely determines the corresponding TGS, it also determines the convergence rate of the algorithm. As expected, our result (Corollary 3) shows that a geometric tail decay of the target distribution is almost necessary and sufficient for the associated chain to be geometrically ergodic.

We are not aware of much results on characterizing geometric ergodicity of Gibbs samplers in the literature. But there have been several attempts at this for Metropolis-Hastings algorithms. For example, Mengersen and Tweedie (1996); Roberts and Tweedie (1996); Jarner and Hansen (2000) studied random-walk Metropolis algorithms with invariant density π on $\mathbb{R}^k$. Their results provide necessary and sufficient condition for geometric ergodicity of the algorithm in terms of the tail behavior of π.

4.2 A Special Kind of Discrete Two-variable Gibbs sampler

Let $\mathbb{N}$ denote the set of natural numbers. Consider a discrete bivariate distribution, π, that has density $\pi(x, y)$ with respect to the counting measure, μ, on $\mathsf{X} \times \mathsf{Y} = \mathbb{N} \times \mathbb{N}$. Suppose that $\pi(x, y)$ has the following form:

$$\pi(x,y) = \begin{cases} a_x & \text{if } x = y,\ y = 1, 2, \dots ; \\ b_y & \text{if } x = y+1,\ y = 1, 2, \dots ; \\ 0 & \text{otherwise}\,; \end{cases} \tag{4–2}$$

where $\{a_k\}_{k=1}^{\infty}$ and $\{b_k\}_{k=1}^{\infty}$ are strictly positive sequences that satisfy $\sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} b_k = 1$. Also, let $b_0 = 0$. Define the class $\mathcal{P}$ to consist of the densities π satisfying (4–2). The marginal densities associated with π are given by

$$\pi_X(x) = a_x + b_{x-1} \ \text{ for } x \in \mathbb{N}\,,$$

and

$$\pi_Y(y) = a_y + b_y \ \text{ for } y \in \mathbb{N}.$$

And the conditional densities are given by

$$\pi_{X|Y}(x|y) = \begin{cases} \frac{a_y}{a_y+b_y} =: 1 - \beta_y & \text{if } x = y,\ y \in \mathbb{N}; \\ \frac{b_y}{a_y+b_y} =: \beta_y & \text{if } x = y+1,\ y \in \mathbb{N}; \\ 0 & \text{otherwise}; \end{cases}$$

and

$$\pi_{Y|X}(y|x) = \begin{cases} \frac{b_{x-1}}{a_x+b_{x-1}} & \text{if } y = x-1,\ x \in \mathbb{N}; \\ \frac{a_x}{a_x+b_{x-1}} & \text{if } y = x,\ x \in \mathbb{N}; \\ 0 & \text{otherwise}. \end{cases}$$

The conditional probability distributions now determine a Markov chain through transition density (4–1). Let $\Phi = \{(X_n, Y_n)\}_{n=0}^{\infty}$ denote the Markov chain with an arbitrary starting point (X_0, Y_0). For any strictly positive sequences $\{a_k\}$ and $\{b_k\}$, the associated Markov chain, Φ, is clearly irreducible, aperiodic and positive recurrent with

invariant density π. Since recurrence and Harris recurrence are identical for countable space Markov chains (Meyn and Tweedie, 1993, Prop.8.1.3), Φ is also Harris ergodic. The question that we are interested in is, when is Φ geometrically ergodic? In other words, for what positive sequences, $\{a_k\}$ and $\{b_k\}$, does the associated Markov chain converge at a geometric rate?

Recall from Chapter 3 that the two marginal sequences $\Phi_X = \{X_n\}$ and $\Phi_Y = \{Y_n\}$ of a two-variable Gibbs chain are themselves Markov chains (Liu et al., 1994). Moreover, the joint chain and the marginal chains all converge at exactly the same rate (Diaconis et al., 2008; Roberts and Rosenthal, 2001). Therefore, we can learn whether Φ is geometric or not by analyzing Φ_X. It is easy to see that Φ_X is a birth and death (BD) chain on $\mathbb{N}$ with the following transition probabilities:

$$k_X(x' \,|\, x) = \begin{cases} = \frac{b_{x-1}}{a_x+b_{x-1}} \frac{a_{x-1}}{a_{x-1}+b_{x-1}} =: q_x & \text{if } x' = x-1,\ x \geq 2\,; \\ = \frac{a_x}{a_x+b_{x-1}} \frac{a_x}{a_x+b_x} + \frac{b_{x-1}}{a_x+b_{x-1}} \frac{b_{x-1}}{a_{x-1}+b_{x-1}} =: 1 - p_x - q_x & \text{if } x' = x,\ x \geq 2\,; \\ = \frac{a_x}{a_x+b_{x-1}} \frac{b_x}{a_x+b_x} =: p_x & \text{if } x' = x+1,\ x \geq 1\,; \\ = 1 - p_1 & \text{if } x' = x = 1\,; \\ 0 & \text{otherwise.} \end{cases} \tag{4–3}$$

4.3 A Sufficient Condition for Φ_X to be Geometric

Lemma 4. *A sufficient condition for Φ_X to be geometrically ergodic is*

$$\limsup_{x\to\infty} \frac{p_x}{q_x} =: r < 1 \quad \textit{and} \quad \liminf_{x\to\infty} q_x := q > 0\,. \tag{4–4}$$

Proof. According to Proposition 1, Φ_X is geometric if we can find a function $V : \mathbb{N} \to [0, \infty)$ that is unbounded off compact sets and satisfies the drift condition

$$\mathrm{E}\big(V(X_1) \mid X_0 = x\big) \leq \rho V(x) + L \quad \text{for some } \rho < 1,\ L < \infty, \text{ and all } x \in \mathbb{N}\,. \tag{4–5}$$

Consider $V(x) = z^x$, where $z > 1$ is to be determined later. Then V is nonnegative and clearly unbounded off compact sets. For $x \geq 2$,

$$\begin{aligned}
\mathrm{E}(V(X_1)|X_0 = x) &= p_x z^{x+1} + (1 - p_x - q_x)z^x + q_x z^{x-1} \\
&= p_x(z^{x+1} - z^x) + q_x(z^{x-1} - z^x) + z^x \\
&= z^x\big(p_x(z-1) + q_x(z^{-1} - 1) + 1\big) \\
&= \big(p_x(z-1) + q_x(z^{-1} - 1) + 1\big)V(x)\,.
\end{aligned}$$

Since $\limsup_{x\to\infty} \frac{p_x}{q_x} = r < 1$, there exists $x_0 \geq 2$ such that $p_x/q_x < (r+1)/2$ for all $x > x_0$. Then for any fixed $z \in (1, \frac{2}{r+1})$,

$$\begin{aligned}
&p_x(z-1) + q_x(z^{-1} - 1) + 1 < \frac{r+1}{2} q_x(z-1) + q_x \frac{1}{z}(1-z) + 1 \\
=&q_x(z-1)\Big(\frac{r+1}{2} - \frac{1}{z}\Big) + 1 \leq q(z-1)\Big(\frac{r+1}{2} - \frac{1}{z}\Big) + 1 =: \rho\,.
\end{aligned}$$

Note that

$$(z-1) \in \Big(0\,, \frac{1-r}{1+r}\Big) \subset (0,1)\,, \quad \Big(\frac{r+1}{2} - \frac{1}{z}\Big) \in \Big(\frac{r-1}{2}\,, 0\Big) \subset \big(-\frac{1}{2}, 0\big) \quad \text{and} \quad q \in (0,1)\,.$$

It follows that $q(z-1)\big(\frac{r+1}{2} - \frac{1}{z}\big) \in (-1, 0)$, hence $\rho \in (0,1)$. Therefore V satisfies (4–5) with $L := \max_{x \leq x_0}\{\mathrm{E}(V(X_1)|X_0 = x)\}$ and $\rho < 1$ defined above. Therefore, (4–4) implies geometric ergodicity of Φ_X. □

4.4 A Sufficient Condition for Φ_X to be Subgeometric

Operator theory has always been a useful tool in the analysis of Markov chains. The reason is that every Markov transition function defines an operator on the space of functions that are square integrable with respect to the invariant density. To analyze the Markov chain Φ_X, denote the square integrable space by $L^2(\pi_X)$ and define

$$L_0^2(\pi_X) = \left\{ h \in L^2(\pi_X) : \int_{\mathsf{X}} h(x)\pi_X(x)\mu_X(dx) = \sum_x h(x)\,\pi_X(x) = 0 \right\}.$$

Note that $L_0^2(\pi_X)$ is a Hilbert space with inner product defined by

$$\langle h_1, h_2 \rangle = \int_{\mathsf{X}} h_1(x)\, h_2(x)\, \pi_X(x)\, \mu_X(dx) = \sum_x h_1(x)\, h_2(x)\, \pi_X(x) \text{ for all } h_1, h_2 \in L_0^2(\pi_X),$$

which induces a norm on $L_0^2(\pi_X)$ defined by $\|h\| = \langle h, h \rangle^{\frac{1}{2}}$. Then $k_X(x'|x)$, the transition function of Φ_X, defines an operator, P_X, on $L_0^2(\pi_X)$. That is, for any $h \in L_0^2(\pi_X)$,

$$P_X\, h(x) = \int_{\mathsf{X}} k_X(x'|x) h(x') \mu_X(dx') = \sum_{x'} k_X(x'|x)\, h(x') \text{ for all } x\,.$$

One nice feature of the Markov chain Φ_X is that it is *reversible* with respect to its invariant density π_X. That is,

$$k_X(x' \,|\, x)\, \pi_X(x) = k_X(x \,|\, x')\, \pi_X(x') \text{ for any } x, x' \in \mathsf{X}$$

This implies that P_X is a self-adjoint operator:

$$\begin{aligned}
\langle P_X h_1, h_2 \rangle &= \int_{\mathsf{X}} \int_{\mathsf{X}} k_X(x'|x) h_1(x') h_2(x) \pi(x) \mu_X(dx') \mu_X(dx) \\
&= \int_{\mathsf{X}} \int_{\mathsf{X}} k_X(x|x') h_1(x') h_2(x) \pi(x') \mu_X(dx) \mu_X(dx') \\
&= \langle h_1, P_X h_2 \rangle\,.
\end{aligned}$$

Finally, the norm of P_X is defined as

$$\|P_X\| = \sup_h \|P_X h\|$$

where the supremum is taken over $h \in L_{0,1}^2(\pi_X) = \left\{ h \in L_0^2(\pi_X) : \int_{\mathsf{X}} h^2(x) \pi_X(x) \mu_X(dx) = 1 \right\}$.

According to the results in Roberts and Rosenthal (1997); Roberts and Tweedie (2001) that apply to reversible Markov chains, Φ_X is not geometrically ergodic if and only if the norm of P_X, $\|P_X\|$, is equal to 1. Let X_0, X_1 represent the first two random variables of Φ_X with $X_0 \sim \pi_X$. Also let (X, Y) be a pair of random variables with joint distribution

π. We have

$$\begin{aligned}\|P_X\| &= \sup_h |\langle P_X h, h\rangle| \\ &= \sup_h \Big|\mathrm{E}\big(h(X_1)h(X_0)\big)\Big| \\ &= \sup_h \mathrm{Var}\big(\mathrm{E}(h(X)|Y)\big) \\ &= 1 - \inf_h \mathrm{E}\big(\mathrm{Var}(h(X)|Y)\big)\end{aligned} \tag{4–6}$$

where the supremums and infimum range over $h \in L^2_{0,1}(\pi_X)$. The first equality holds for all self-adjoint operators (Rudin, 1991, Thm 12.25), the second equality is the definition of inner product, the third equality follows from Liu et al. (1994, Lemma 3.2) and the last equality is implied by the formula

$$\mathrm{Var}(h(X)) = \mathrm{E}\big[\mathrm{Var}(h(X)|Y)\big] + \mathrm{Var}\big[\mathrm{E}(h(X)|Y)\big].$$

Consequently, if we can find a sequence $\{h_i \in L^2_{0,1}(\pi_X)\}$ such that

$$\liminf_{i\to\infty} \mathrm{E}\big[\mathrm{Var}(h_i(X)|Y)\big] = 0\,, \tag{4–7}$$

then $\|P_X\| = 1$ according to (4–6) and Φ_X is not geometric. We use this idea to derive the following lemma:

Lemma 5. *If any one of the following conditions hold, then Φ_X is not geometrically ergodic.*

1. $$\limsup_{i\to\infty} \frac{\sum_{x=i}^{\infty}(a_x + b_x)}{a_{i-1}} = \infty\,,$$

2. $$\limsup_{i\to\infty} \frac{\sum_{x=i}^{\infty}(a_x + b_x)}{b_{i-1}} = \infty\,,$$

3. $$\limsup_{i\to\infty} \frac{b_i}{a_i} = \infty\,.$$

Proof. Let (X, Y) be a pair of random variables with joint distribution π. For $i = 1, 2, \ldots,$ let

$$H_i(x) = \begin{cases} 0 & x < i, \\ 1 & x \geq i. \end{cases}$$

The means and variances of $\{H_i\}$ with respect to π_X are

$$\mu_i := \mathrm{E}\big(H_i(X)\big) = \mathrm{E}\big(H_i^2(X)\big) = \sum_{x=1}^{\infty} \pi_X(x) H_i(x) = \sum_{x=i}^{\infty} (a_x + b_{x-1}) < \infty$$

and

$$V_i := \mathrm{Var}\big(H_i(X)\big) = \mathrm{E}\big(H_i^2(X)\big) - \Big(\mathrm{E}\big(H_i(X)\big)\Big)^2 = \mu_i(1 - \mu_i) < \infty.$$

Define $h_i(x) = \big(H_i(x) - \mu_i\big)/\sqrt{V_i}$, then $h_i(x) \in L^2_{0,1}(\pi_X)$. Next, we evaluate $\mathrm{E}\big[\mathrm{Var}(h_i(X)|Y)\big]$, which is part of the left hand side of equation (4–7). Recall that we defined β_y to be $b_y/(a_y + b_y) = \pi_{X|Y}(y + 1 \,|\, y)$ in Section 4.2. Then

$$\begin{aligned} &\mathrm{E}\big(H_i(X)|Y = y\big) = \mathrm{E}\big(H_i^2(X)|Y = y\big) \\ &= \pi_{X|Y}(y \,|\, y)\, H_i(y) + \pi_{X|Y}(y + 1 \,|\, y)\, H_i(y + 1) = \begin{cases} 0 & y \leq i - 2\,, \\ \beta_{i-1} & y = i - 1\,, \\ 1 & y \geq i\,. \end{cases} \end{aligned}$$

Furthermore,

$$\mathrm{Var}\big(H_i(X)|Y = y\big) = \mathrm{E}\big(H_i^2(X)|Y = y\big) - \Big(\mathrm{E}\big(H_i(X)|Y = y\big)\Big)^2 = \begin{cases} \beta_{i-1}(1 - \beta_{i-1}) & y = i - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} \mathrm{E}\Big[\mathrm{Var}\big(H_i(X)|Y\big)\Big] &= \sum_{y=1}^{\infty} \pi_Y(y)\, \mathrm{Var}\big(H_i(X)|Y = y\big) \\ &= \pi_Y(i - 1)\, \mathrm{Var}\big(H_i(X)|Y = i - 1\big) \\ &= (a_{i-1} + b_{i-1})\, \beta_{i-1}\, (1 - \beta_{i-1}) \\ &= \frac{a_{i-1} b_{i-1}}{a_{i-1} + b_{i-1}}\,. \end{aligned}$$

Finally,

$$\mathrm{E}\Big[\mathrm{Var}\big(h_i(X)|Y\big)\Big] = V_i^{-1}\mathrm{E}\Big[\mathrm{Var}\big(H_i(X)|Y\big)\Big] = (\mu_i(1-\mu_i))^{-1}\frac{a_{i-1}b_{i-1}}{a_{i-1}+b_{i-1}}.$$

Note that

$$\begin{aligned}\left(\mathrm{E}\Big[\mathrm{Var}\big(h_i(X)|Y\big)\Big]\right)^{-1} &= \mu_i(1-\mu_i)\left(\frac{a_{i-1}+b_{i-1}}{a_{i-1}b_{i-1}}\right)\\ &= (1-\mu_i)\left[\sum_{x=i}^{\infty}(a_x+b_x)+b_{i-1}\right]\left(\frac{1}{a_{i-1}}+\frac{1}{b_{i-1}}\right)\\ &= (1-\mu_i)\left[\frac{\sum_{x=i}^{\infty}(a_x+b_x)}{a_{i-1}}+\frac{\sum_{x=i}^{\infty}(a_x+b_x)}{b_{i-1}}+\frac{b_{i-1}}{a_{i-1}}+1\right]\end{aligned} \tag{4–8}$$

where $\lim_{i\to\infty}(1-\mu_i) = \lim_{i\to\infty}\sum_{x=1}^{i}(a_x+b_{x-1}) = 1$. Hence, equation (4–7) holds if and only if

$$\begin{aligned}&\limsup_{i\to\infty}\frac{\sum_{x=i}^{\infty}(a_x+b_x)}{a_{i-1}} = \infty\,,\\ \textit{or}\quad &\limsup_{i\to\infty}\frac{\sum_{x=i}^{\infty}(a_x+b_x)}{b_{i-1}} = \infty\,,\\ \textit{or}\quad &\limsup_{i\to\infty}\frac{b_i}{a_i} = \infty\,.\end{aligned}$$

□

Basically, Lemma 5 says that the chain is subgeometric if the tail mass associated with a_i or b_i converges to zero at a slower rate than a_i or b_i themselves as i goes to infinity. We are able to understand this result better in the next section, where Lemma 4 and 5 together provide characterization of geometric ergodicity of a class of Markov chains.

4.5 An Attempt to Characterize Geometric Ergodicity of Φ_X

So far, we have developed a sufficient condition for Φ_X to be geometrically ergodic and a sufficient condition for it to be subgeometric. Using these tools, we examine all $\pi \in \mathcal{P}$ (defined in Section 4.2 at (4–2)). For most π, we are able to tell whether its corresponding Gibbs chain is geometric or subgeometric. However, there are a few πs that we can not categorize.

Case 1. If $\limsup_{i\to\infty} \frac{a_i}{b_{i-1}} = \infty$, then by condition 2 of Lemma 5, the corresponding chain is not geometric.

Case 2. If $\limsup_{i\to\infty} \frac{b_i}{a_i} = \infty$, then by condition 3 of Lemma 5, the corresponding chain is not geometric.

Case 3. Otherwise, suppose that $\limsup_{i\to\infty} \frac{a_i}{b_{i-1}} < \infty$ and $\limsup_{i\to\infty} \frac{b_i}{a_i} < \infty$. Note that

$$q_i = \frac{b_{i-1}}{a_i + b_{i-1}} \frac{a_{i-1}}{a_{i-1} + b_{i-1}} = \frac{1}{1 + \frac{a_i}{b_{i-1}}} \frac{1}{1 + \frac{b_{i-1}}{a_{i-1}}} . \tag{4–9}$$

Hence

$$\liminf q_i \geq \frac{1}{1 + \limsup \frac{a_i}{b_{i-1}}} \frac{1}{1 + \limsup \frac{b_{i-1}}{a_{i-1}}} > 0 .$$

Note that

$$\frac{p_i}{q_i} = \frac{a_i}{a_{i-1}} \frac{b_i}{b_{i-1}} \frac{a_{i-1} + b_{i-1}}{a_i + b_i} = \frac{a_i}{a_{i-1}} \frac{1 + \frac{a_{i-1}}{b_{i-1}}}{1 + \frac{a_i}{b_i}} . \tag{4–10}$$

Hence

$$\limsup \frac{p_i}{q_i} \leq \left[\limsup \frac{a_i}{a_{i-1}}\right] \left[\frac{1 + \limsup \frac{a_{i-1}}{b_{i-1}}}{1 + \liminf \frac{a_i}{b_i}}\right] = A \frac{1 + M}{1 + m} ,$$

where $A := \limsup \frac{a_i}{a_{i-1}}$, $m := \liminf \frac{a_i}{b_i}$ and $M := \limsup \frac{a_i}{b_i}$. Therefore, when $A\frac{1+M}{1+m} < 1$, the chain is geometric by Lemma 4. Otherwise we'll have to check the three conditions in Lemma 5. If any of them are satisfied, the chain is not geometric, and if none of them are satisfied, it remains unknown what the convergence rate is.

Nevertheless, there is one collection of Markov chains for which geometric ergodicity *can* be fully characterized.

Corollary 3. *Assume that both* $A = \lim \frac{a_i}{a_{i-1}}$ *and* $\lim \frac{a_i}{b_i}$ *exist, then all the limits below are well defined and the following statements are equivalent:*

(i)

$$\lim_{i\to\infty} \frac{a_i}{b_{i-1}} < \infty \quad \text{and} \quad \lim_{i\to\infty} \frac{b_i}{a_i} < \infty \quad \text{and} \quad A < 1 ;$$

(ii)

$$r = \lim_{i\to\infty} \frac{p_i}{q_i} =< 1 \quad \text{and} \quad q = \lim_{i\to\infty} q_i > 0 ;$$

(iii) *The chain* Φ_X *is geometric.*

Proof. (i)→(ii): This is true because under (i), $r = \lim_{i\to\infty} \frac{p_i}{q_i} = A < 1$ by (4–10) and $q > 0$ by (4–9).

(ii)→(iii): Immediate by Lemma 4.

(iii)→(i): If the chain is geometric, then $\lim_{i\to\infty} \frac{a_i}{b_{i-1}} < \infty$ and $\lim_{i\to\infty} \frac{b_i}{a_i} < \infty$ by Condition 2 and 3 of Lemma 5. Next, if $A = 1$, then for any fixed positive integer K, we have

$$\lim_{i\to\infty} \frac{a_{i+1}}{a_i} = 1,\ \lim_{i\to\infty} \frac{a_{i+2}}{a_i} = 1, \cdots, \lim_{i\to\infty} \frac{a_{i+K}}{a_i} = 1\,.$$

Then there exists i_0 such that for any $i \geq i_0$,

$$\frac{a_{i+1}}{a_i} > \frac{1}{2},\ \frac{a_{i+2}}{a_i} > \frac{1}{2}, \cdots, \frac{a_{i+K}}{a_i} > \frac{1}{2}\,.$$

Hence, given any K, there exists i_0 such that for any $i > i_0$,

$$\frac{\sum_{x=i}^{\infty}(a_x + b_x)}{a_{i-1}} \geq \frac{\sum_{x=i}^{(i+K-1)} a_x}{a_{i-1}} > \frac{K}{2}\,.$$

This implies that $\limsup_{i\to\infty} \frac{\sum_{x=i}^{\infty}(a_x+b_x)}{a_{i-1}} = \infty$. By condition 1 of Lemma 5 the chain is not geometric. This contradicts statement (iii). So $A \neq 1$. But, A can not be greater than 1 either. Otherwise, $\sum_x a_x = \infty$, which contradicts the fact that $\sum a_x + \sum b_x = 1$. Therefore, $A < 1$. □

In the above result, Condition 1 shows the types of distribution π for which good or bad behavior of the associated chain may be expected. The key part of this condition is $A < 1$, which means that the probabilities in the diagonal die down geometrically in the tail. On the other hand, Condition 2 shows us how to characterize geometric convergence of the chain in terms of transition probabilities of the marginal BD chain. One needs to be careful when interpreting this condition, because an underlying assumption on the sequences $\{p_i\}$ and $\{q_i\}$ is that there do exist sequences $\{a_i\}$ and $\{b_i\}$ such that, together, they satisfy the relationship in (4–3). Finally, we end this chapter by applying the above results to four concrete examples.

Example 1. Let $a_x = c_1 x^{-d}$ and $b_x = c_2 x^{-d}$, where d is some fixed integer that is greater than 1 and $(c_1 + c_2)\sum_{x=1}^{\infty} x^{-d} = 1$. Then both $\lim_{i\to\infty} \frac{a_i}{a_{i-1}}$ and $\lim_{i\to\infty} \frac{a_i}{b_i}$ exist, with $A = \lim_{i\to\infty} \frac{a_i}{a_{i-1}} = 1$. Therefore Φ_X is not geometrically ergodic by Corollary 3.

Example 2. Let $a_x = ce^{-x}$ and $b_x = e^{-x}$, where c is the solution to $(1+c)\,e^{-1}/(1-e^{-1}) = 1$. Then both $A = \lim_{i\to\infty} \frac{a_i}{a_{i-1}}$ and $\lim_{i\to\infty} \frac{a_i}{b_i}$ exist. Furthermore, $A = e^{-1} < 1$, $\limsup_{i\to\infty} \frac{a_i}{b_{i-1}} = \lim_{i\to\infty} ce^{-1} < \infty$ and $\limsup_{i\to\infty} \frac{b_i}{a_i} = c^{-1} < \infty$. Therefore Φ_X is not geometrically ergodic by Corollary 3.

Example 3. Let $a_x = ce^{-x}$ and $b_x = e^{-2x}$, where c is the solution to $c\,e^{-1}/(1-e^{-1}) + e^{-2}/(1-e^{-2}) = 1$. Then both $A = \lim_{i\to\infty} \frac{a_i}{a_{i-1}}$ and $\lim_{i\to\infty} \frac{a_i}{b_i}$ exist, and $\limsup_{i\to\infty} \frac{a_i}{b_{i-1}} = \lim_{i\to\infty} ce^{i-2} = \infty$. Therefore Φ_X is geometrically ergodic by Corollary 3.

Example 4. Let

$$a_x = \begin{cases} = ce^{-x} & \text{if } x \text{ is even,} \\ = e^{-2x} & \text{if } x \text{ is odd,} \end{cases} \qquad \text{and} \qquad b_x = \begin{cases} = e^{-2x} & \text{if } x \text{ is even,} \\ = ce^{-x} & \text{if } x \text{ is odd,} \end{cases}$$

where c is the solution to $ce^{-1}/(1-e^{-1}) + e^{-2}/(1-e^{-2}) = 1$. Then $\lim_{i\to\infty} \frac{a_i}{b_i}$ does not exist. Hence Corollary 3 is not applicable. However, we can resort to the more general Lemma 5. Note that

$$\limsup_{i\to\infty} \frac{b_i}{a_i} \geq \lim_{i\to\infty} \frac{a_{2i+1}}{b_{2i+1}} = \lim_{i\to\infty} \frac{ce^{-(2i+1)}}{e^{-2(2i+1)}} = \lim_{i\to\infty} ce^{2i+1} = \infty.$$

Therefore Φ_X is not geometrically ergodic.

APPENDIX A
PROPRIETY OF THE POSTERIORS AND FINITE MOMENT CONDITIONS

Proposition 6. *A necessary and sufficient condition for propriety of the posterior under either the standard diffuse prior, $p_{-\frac{1}{2},0}$ (defined in (1–2)), or the reference prior, p_r (defined in (1–3)), is $q \geq 3$. Moreover, if $q \geq 3$ and $s, t \in \mathbb{R}$, then*

$$E_{\pi_{-\frac{1}{2},0}}\left[\left(\sigma_\theta^2\right)^s\left(\sigma_e^2\right)^t\right] < \infty$$

if and only if the following two conditions are satisfied:

(i) $\quad -\frac{1}{2} < s < \frac{q}{2} - 1$

(ii) $\quad s + t < \frac{mq}{2} - 1$

Proof. Hobert and Casella (1996) show that the posterior density under $p_{a,b}$ is proper if and only if

$$\frac{1-q}{2} < a < 0 \qquad \text{and} \qquad a + b > \frac{1-mq}{2}\,. \tag{A–1}$$

In other words, the integral

$$m_{a,b}(y) := \int_{\mathbb{R}_+}\int_{\mathbb{R}_+}\int_{\mathbb{R}}\int_{\mathbb{R}^q}\left(\sigma_\theta^2\right)^{-(a+1)}\left(\sigma_e^2\right)^{-(b+1)}\phi(\theta|\mu,\sigma_\theta^2,\sigma_e^2)\,L\big(\theta,\mu,\sigma_\theta^2,\sigma_e^2;y\big)\,d\theta\,d\mu\,d\sigma_\theta^2\,d\sigma_e^2$$

is finite if and only if (A–1) is satisfied. The standard diffuse prior corresponds to $p_{a,b}$ with $a = -\frac{1}{2}$ and $b = 0$, and in this special case, (A–1) is satisfied if and only if $q \geq 3$. Now consider the reference prior, and note that

$$\left[m - 1 + \left(\frac{\sigma_e^2}{\sigma_e^2 + m\sigma_\theta^2}\right)^2\right]^{\frac{1}{2}} \in \left[\sqrt{m-1}, \sqrt{m}\right]$$

for all $(\sigma_e^2, \sigma_\theta^2) \in \mathbb{R}_+ \times \mathbb{R}_+$. Consequently, as far as propriety of the posterior is concerned, the reference prior behaves like $p_{a,b}$ with $a = -1 + C_m/2$ and $b = 0$. Combining Hobert & Casella's propriety result with the fact that $C_m \in (0.92, 1)$ for all $m \geq 2$ shows that the reference prior yields a proper posterior if and only if $q \geq 3$.

Now assume that $q \geq 3$ so that $\pi_{-\frac{1}{2},0}$ is well-defined, and note that

$$
\begin{aligned}
& \mathrm{E}_{\pi_{-\frac{1}{2},0}}\left[\left(\sigma_\theta^2\right)^s\left(\sigma_e^2\right)^t\right] \\
= \; & \frac{1}{m_{-\frac{1}{2},0}(y)} \int_{\mathbb{R}_+}\int_{\mathbb{R}_+}\int_{\mathbb{R}}\int_{\mathbb{R}^q} \left(\sigma_\theta^2\right)^{s-\frac{1}{2}}\left(\sigma_e^2\right)^{t-1} \phi(\theta|\mu,\sigma_\theta^2,\sigma_e^2)\, L\big(\theta,\mu,\sigma_\theta^2,\sigma_e^2;y\big)\, d\theta\, d\mu\, d\sigma_\theta^2\, d\sigma_e^2 \\
= \; & m_{-s-\frac{1}{2},-t}(y) \Big/ m_{-\frac{1}{2},0}(y) \; .
\end{aligned}
$$

The result then follows directly from another application of Hobert & Casella's propriety result. $\square$

APPENDIX B
ψ-IRREDUCIBILITY

Recall from Section 2.1 that Φ denotes a Markov chain with Markov transition function K. Suppose that μ is a non-trivial, σ-finite measure on X. Then Φ is called *μ-irreducible* if for each $x \in \mathsf{X}$ and each set A with $\mu(A) > 0$, there exists an $n \in \{1, 2, \dots\}$, which may depend on x and A, such that $K^n(x, A) > 0$. In words, the chain is μ-irreducible if every set with positive μ-measure is accessible from every point in the state space. According to Meyn and Tweedie (1993, Prop. 4.2.2), if Φ is μ-irreducible for some μ, then there exists a probability measure ψ, called the *maximal irreducibility measure*, such that

1. Φ is ψ-irreducible, and
2. for any other measure μ', the chain Φ is μ'-irreducible if and only if μ' is absolutely continuous with respect to ψ (denoted $\psi \succ \mu'$).

It is clear from above that a maximal irreducibility measure is unique up to equivalence; i.e., if ψ_1 and ψ_2 are both maximal irreducibility measures, then $\psi_1 \succ \psi_2$ and $\psi_2 \succ \psi_1$ (denoted $\psi_1 \equiv \psi_2$). Hence, the terminology "Φ is ψ-irreducible" is well defined, if by this we mean that Φ is μ-irreducible for some μ and that ψ is a maximal irreducibility measure for Φ. We can see that ψ represents the "natural" irreducibility measure of a chain.

APPENDIX C
BOUNDING δ_1 AND δ_3 BELOW 1

We now prove Proposition 4 by establishing that

- $\delta_1(s) < 1$ for any $s \in (0,1)$ if $M + 2b \geq q + 3$, and
- $\delta_3(s_0) < 1$ for some small positive s_0 if $\Delta_1 < 2\exp\big(\Psi\big(\frac{q}{2} + a\big)\big)$.

It is well known that $\Psi'(x) > 0$ for all $x > 0$ and that $\Psi(x+1) = \Psi(x) + \frac{1}{x}$. A couple of common values of the digamma function that we need below are $\Psi(1) = -\gamma$ and $\Psi(\frac{1}{2}) = -\gamma - 2\log(2)$, where $\gamma := \lim_{p\to\infty}\big(1 + \frac{1}{2} + \cdots + \frac{1}{p} - \log(p)\big) \doteq 0.577$ is Euler's constant. Also, the recurrence formula yields: $\Psi(\frac{3}{2}) = -\gamma - 2\log(2) + 2$.

C.1 Bounding $\delta_1(s)$

Recall that $\delta_1(s)$ is actually a function of s, q and $\frac{M}{2} + b$. Indeed,

$$\delta_1\left(s, q, \frac{M}{2} + b\right) = \frac{\Gamma\big(\frac{M}{2} + b - s\big)}{\Gamma\big(\frac{M}{2} + b\big)}\left(\frac{q+1}{2}\right)^s .$$

Now, for any fixed $s > 0$, $\Gamma(x-s)/\Gamma(x)$ is decreasing in x for $x > s$, because

$$\frac{d}{dx}\Big[\log(\Gamma(x-s)) - \log(\Gamma(x))\Big] = \Psi(x-s) - \Psi(x) < 0 \quad \text{for all } x > s > 0 .$$

Therefore, with (s,q) fixed, $\delta_1\big(s, q, \frac{M}{2} + b\big)$ is decreasing in $\big(\frac{M}{2} + b\big)$ as long as $\frac{M}{2} + b > s$. Consequently, to show that $\delta_1\big(s, q, \frac{M}{2} + b\big) < 1$ for all $s \in (0,1)$ if $M + 2b \geq q + 3$, we need only prove the following.

Lemma 6. *$\delta_1\big(s, q, \frac{q+3}{2}\big) < 1$ for all $s \in (0,1)$ and $q \geq 2$.*

Proof. Fix $s \in (0,1)$ and define

$$T(x) = \frac{\Gamma(x+s)}{\Gamma(x)}(x+s-1)^{-s} \quad \text{for } x > 1 - s .$$

We claim that

1. $T(x)$ is strictly decreasing in x, and
2. $\lim_{x\to\infty} T(x) = 1$.

To prove claim 1, we will show that $Q(x) = \log(T(x))$ is decreasing in x. First, for $x > 0$

$$\Psi(x) = -\gamma + \sum_{p=1}^{\infty} \left(\frac{1}{p} - \frac{1}{x+p-1} \right)$$

(Abramowitz and Stegun, 1964, p.259). Note that $\left(\frac{1}{p} - \frac{1}{x+p-1}\right)$ is nonnegative for all p when $x \geq 1$ and negative for all p when $x < 1$. Hence, the above series is absolutely convergent for all $x > 0$. Clearly, $Q(x) = -s\log(x+s-1) + \log(\Gamma(x+s)) - \log(\Gamma(x))$ and its derivative can be expressed as follows

$$Q'(x) = -s\frac{1}{x+s-1} + \sum_{p=1}^{\infty} \left(\frac{1}{p} - \frac{1}{x+s+p-1} \right) - \sum_{p=1}^{\infty} \left(\frac{1}{p} - \frac{1}{x+p-1} \right) . \tag{C–1}$$

The fraction in the first term of (C–1) can be written as the following absolutely convergent telescoping series

$$\frac{1}{x+s-1} = \sum_{p=1}^{\infty} \left(\frac{1}{x+s+p-2} - \frac{1}{x+s+p-1} \right) .$$

Therefore,

$$\begin{aligned} Q'(x) &= s\sum_{p=1}^{\infty} \left(\frac{1}{x+s+p-1} - \frac{1}{x+s+p-2} \right) + \sum_{p=1}^{\infty} \left(\frac{1}{x+p-1} - \frac{1}{x+s+p-1} \right) \\ &= \sum_{p=1}^{\infty} \left[-(1-s)\frac{1}{x+s+p-1} - s\frac{1}{x+s+p-2} + \frac{1}{x+p-1} \right] . \end{aligned}$$

The convexity of the function $h(z) = \frac{1}{z}$ on $\mathbb{R}_+$ combined with the fact that $(1-s)(x+s+p-1) + s(x+s+p-2) = x+p-1$ can be used to show that every term in the series above is negative. It follows that $Q(x)$ and $T(x)$ are both decreasing in x for $x > 1-s$.

We now prove claim 2. Fix $s \in (0,1)$ and define $S(x) = x^{-s}\Gamma(x+s)/\Gamma(x)$. As $x \to \infty$, $S(x) \to 1$ (Abramowitz and Stegun, 1964, p.257). As a consequence,

$$\lim_{x\to\infty} T(x) = \lim_{x\to\infty} S(x) \left(\frac{x}{x+s-1} \right)^s = 1 .$$

Finally, for fixed $s \in (0,1)$, note that $\frac{q+3}{2} - s > 1 - s$ and

$$\delta_1\big(s, q, \frac{q+3}{2}\big) = \big(\frac{q+1}{2}\big)^s \frac{\Gamma\big(\frac{q+3}{2} - s\big)}{\Gamma\big(\frac{q+3}{2}\big)} = \Big(T\big(\frac{q+3}{2} - s\big)\Big)^{-1} .$$

It follows from claims 1 and 2 that $T\big(\frac{q+3}{2} - s\big) > 1$ and hence $\delta_1\big(s, q, \frac{q+3}{2}\big) < 1$. □

C.2 Bounding $\delta_3(s)$

Recall that $\delta_3(s)$ is actually a function of s, m and a. If we define

$$A(s, q, a) = \frac{\Gamma(\frac{q}{2} + a - s)}{2^s \Gamma(\frac{q}{2} + a)} ,$$

then we have $\delta_3(s, m, a) = A(s, q, a)\Delta_1^s(m)$. Note that there exists an $s_0 \in S$ such that $\delta_3(s_0, m, a) < 1$ if and only if $\Delta_1(m) < A^*(q, a)$, where

$$A^*(q, a) := \sup_{s \in S} A^{-\frac{1}{s}}(s, q, a) = 2 \sup_{s \in S} \left(\frac{\Gamma(\frac{q}{2} + a)}{\Gamma(\frac{q}{2} + a - s)} \right)^{\frac{1}{s}} .$$

We now establish a lower bound for $A^*(q, a)$. Define

$$g(s, q, a) = \log\left(\frac{1}{2} A^{-\frac{1}{s}}(s, q, a)\right) = \frac{1}{s}\left[\log\left(\Gamma\left(\frac{q}{2} + a\right)\right) - \log\left(\Gamma\left(\frac{q}{2} + a - s\right)\right)\right] .$$

Then

$$\lim_{s \to 0} g(s, q, a) = \left.\frac{d \log(\Gamma(x))}{dx}\right|_{x = \frac{q}{2} + a} = \Psi\left(\frac{q}{2} + a\right) .$$

Hence,

$$A^*(q, a) \geq \lim_{s \to 0} 2 \exp\left[g(s, q, a)\right] = 2 \exp\left(\Psi\left(\frac{q}{2} + a\right)\right) . \tag{C–2}$$

We conclude that there exists $s \in S$ such that $\delta_3(s, m, a) < 1$ if $\Delta_1(m) < 2 \exp\left(\Psi\left(\frac{q}{2} + a\right)\right)$.

Remark 4. It is easy to show that $\left.\frac{\partial g(s,q,a)}{\partial s}\right|_{s=0} < 0$. In other words, for fixed q and a, $g(s, q, a)$ is decreasing in s in a neighborhood of $s = 0$. Furthermore, numerical calculations suggest that $g(s, q, a)$ is decreasing on the entire set S for any fixed q and a. Hence, we believe the lower bound on $A^*(q, a)$ in (C–2) is sharp. Nevertheless, our proof for Proposition 4 does not rely on this conjecture.

APPENDIX D
VALID CHOICES OF THE DISTINGUISHED POINT

Note that w_1 and w_2 are both functions of ξ. Hence for any pair, (w_1^*, w_2^*), that we use as a distinguished point in our minorization condition, it is necessary to check that there exists a $\xi \in \mathbb{R}^{q+1}$ such that $(w_1^*, w_2^*) = (w_1(\xi), w_2(\xi))$. Let $\tilde{y} = (\overline{y}_1, \cdots, \overline{y}_q)$ and $\tilde{m} = (m_1, \cdots, m_q)$. Also, let $\overline{y} = \frac{1}{M}\sum m_i \overline{y}_i$, $\text{SST} = \sum m_i (\overline{y}_i - \overline{y})^2$ and $m_* = \min\{m_1, \ldots, m_q\}$.

Lemma 7. *For a given pair (w_1^*, w_2^*), a solution to*

$$w_1(\xi) = w_1^* \quad \textit{and} \quad w_2(\xi) = w_2^* \tag{D–1}$$

exists if

$$\sqrt{m_* w_1^*} + \sqrt{w_2^*} \geq \sqrt{SST}. \tag{D–2}$$

When $m_1 = \cdots = m_q$, condition (D–2) is not only sufficient but also necessary.

Proof. If $\xi = (\theta, \mu)$ is a solution to (D–1), then

$$\sqrt{w_2^*} = \sqrt{\sum_i m_i(\overline{y}_i - \theta_i)^2} = \sqrt{\sum_i (\sqrt{m_i}\overline{y}_i - \sqrt{m_i}\theta_i)^2} = \|\sqrt{\tilde{m}} \cdot \tilde{y} - \sqrt{\tilde{m}} \cdot \theta\| \tag{D–3}$$

and

$$\sqrt{m_* w_1^*} = \sqrt{m_* \sum_i (\theta_i - \mu)^2} \leq \sqrt{\sum_i (\sqrt{m_i}\theta_i - \sqrt{m_i}\mu)^2} = \|\sqrt{\tilde{m}} \cdot \theta - \mu\sqrt{\tilde{m}}\| \tag{D–4}$$

with equality holding in (D–4) when $m_1 = \cdots = m_q = m_*$. In p-dimensional space, let l denote the line that passes through the origin and through the point $\sqrt{\tilde{m}} = (\sqrt{m_1}, \cdots, \sqrt{m_q})$. The point $\sqrt{\tilde{m}} \cdot \tilde{y}$ is determined by the data, so it is fixed. According to (D–3), the point $\sqrt{\tilde{m}} \cdot \theta$ should fall on the sphere C_1 that is centered at $\sqrt{\tilde{m}} \cdot \tilde{y}$ with radius $\sqrt{w_2^*}$. But by (D–4), $\sqrt{\tilde{m}} \cdot \theta$ should also fall on the sphere $C_2(\xi)$ that is centered at $\mu\sqrt{\tilde{m}}$ with radius $\|\sqrt{\tilde{m}} \cdot \theta - \mu\sqrt{\tilde{m}}\| \geq \sqrt{m_* w_1^*}$, where $\mu\sqrt{\tilde{m}}$ is a point on l. Therefore, a solution to (D–1) exists if and only if there exists ξ such that C_1 and $C_2(\xi)$ intersect.

Note that the (shortest) distance between the point $\sqrt{\tilde{m}} \cdot \tilde{y}$ and the line l is

$$d(\sqrt{\tilde{m}} \cdot \tilde{y}, l) = \|\sqrt{\tilde{m}} \cdot \tilde{y} - \overline{y}\sqrt{\tilde{m}}\| = \sqrt{\sum m_i(\overline{y}_i - \overline{y})^2} = \sqrt{\mathrm{SST}}.$$

Some straightforward analysis then shows that C_1 and $C_2(\xi)$ intersect for some ξ if and only if

$$\|\sqrt{\tilde{m}} \cdot \theta - \mu\sqrt{\tilde{m}}\| + \sqrt{w_2^*} \geq \sqrt{\mathrm{SST}}\,. \tag{D–5}$$

Finally, because of (D–4), (D–2) implies (D–5). Thus (D–2) is sufficient for the existence of a solution. Now, if $m_1 = \cdots = m_q$, then (D–2) and (D–5) are exactly the same. So, in this case, (D–2) is also necessary.

□

REFERENCES

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Tables*. Dover.

Asmussen, S. (2003). *Applied Probability and Queues*. Springer Verlag.

Berger, J. O. and Bernardo, J. M. (1992). Reference priors in a variance components problem. *Bayesian Analysis in Statistics and Econometrics (P.K. Goel and N.S. Iyengar, eds.)* 323–340.

Bernardo, J. M. (1996). Discussion of "statistical inference and Monte Carlo algorithms" by G. Casella. *Test*, **5** 289–291.

Bhattacharya, S. (2008). Consistent estimation of the accuracy of importance sampling using regenerative simulation. *Statistics and Probability Letters*, **78** 2522–2527.

Bratley, P., Fox, B. L. and Schrage, L. E. (1987). *A Guide to Simulation (2nd ed.)*. Springer-Verlag New York, Inc., New York, NY, USA.

Chan, K. S. and Geyer, C. J. (1994). Discussion of "Markov chains for exploring posterior distributions". *The Annals of Statistics*, **22** 1747–1757.

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

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, **27** 567–578.

Diaconis, P., Khare, K. and Saloff-Coste, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials (with discussion). *Statistical Science*, **23** 151–200.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society,* Series B 363–375.

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. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1** 515–533.

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

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

Hill, B. M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Association*, **60** 806–825.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, **91** 1461–1473.

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.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Groningen: Wolters-Noordhoff.

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

Johnson, A. A. and Jones, G. L. (2008). Gibbs sampling for a Bayesian hierarchical version of the general linear mixed model. Tech. rep., University of Minnesota.

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** 312–34.

Lavenberg, S. S. and Slutz, D. R. (1975). Introduction to regenerative simulation. *IBM Journal of Research and Development*, **19** 458–462.

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

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

Lyles, R. H., Kupper, L. L. and Rappaport, S. M. (1997). Assessing regulatory compliance of occupational exposures via the balanced one-way random effects ANOVA model. *Journal of Agricultural, Biological, and Environmental Statistics*, **2** 417–439.

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

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer Verlag, London.

Meyn, S. P. and Tweedie, R. L. (1994). Computable bounds for geometric convergence rates of markov chains. *The Annals of Applied Probability*, **4** 981–1011.

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, **90** 233–41.

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

Papaspiliopoulos, O. and Roberts, G. (2008). Stability of the Gibbs sampler for Bayesian hierarchical models. *The Annals of Statistics*, **36** 95–117.

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

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods.* Springer.

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

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

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

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. (2001). Markov chains and de-initializing processes. *Scandinavian Journal of Statistics*, **28** 489–504.

Roberts, G. O. and Tweedie, R. L. (2001). Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains. *Journal of Applied Probability*, **38A** 37–41.

Roy, V. and Hobert, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society,* Series B, **69** 607–623.

Rudin, W. (1991). *Funtional Analysis.* 2nd ed. McGraw-Hill.

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

Tiao, G. C. and Tan, W. Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance components. *Biometrika*, **52** 37–53.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics*, **10** 1–50.

BIOGRAPHICAL SKETCH

Aixin Tan was born in 1984 in Beijing, China. She was recruited to the gifted children's program of Beijing No. 8 middle school in 1995. In 1999, she was admitted to the School of Mathematical Science of Peking University. As a sophomore, she made a life-changing decision to learn to play tennis and joined the university team soon after. She was increasingly attracted to the sport as well as a teammate of hers who eventually became her beloved boyfriend. In 2003, she earned her bachelor's degree in probability and statistics. After that, she joined the Department of Statistics at the University of Florida. She received her master's degree in 2005 and her doctorate degree in 2009. After graduation, Aixin will join the Department of Statistics and Actuarial Science at the University of Iowa, as an assistant professor of the Hawkeyes. But of course, she always bears the heart of a Gator.