Honest Importance Sampling with Multiple Markov Chains

Aixin Tan[1], Hani Doss[2], and James P. Hobert[2]

[1]Department of Statistics, University of Iowa
[2]Department of Statistics, University of Florida

Abstract

Importance sampling is a classical Monte Carlo technique in which a random sample from one probability density, π_1, is used to estimate an expectation with respect to another, π. The importance sampling estimator is strongly consistent and, as long as two simple moment conditions are satisfied, it obeys a central limit theorem (CLT). Moreover, there is a simple consistent estimator for the asymptotic variance in the CLT, which makes for routine computation of standard errors. Importance sampling can also be used in the Markov chain Monte Carlo (MCMC) context. Indeed, if the random sample from π_1 is replaced by a Harris ergodic Markov chain with invariant density π_1, then the resulting estimator remains strongly consistent. There is a price to be paid however, as the computation of standard errors becomes more complicated. First, the two simple moment conditions that guarantee a CLT in the iid case are not enough in the MCMC context. Second, even when a CLT does hold, the asymptotic variance has a complex form and is difficult to estimate consistently. In this paper, we explain how to use regenerative simulation to overcome these problems. Actually, we consider a more general set up, where we assume that Markov chain samples from several probability densities, $\pi_1, \ldots, \pi_k$, are available. We construct multiple-chain importance sampling estimators for which we obtain a CLT based on regeneration. We show that if the Markov chains converge to their respective target distributions at a geometric rate, then under moment conditions similar to those required in the iid case, the MCMC-based importance sampling estimator obeys a CLT. Furthermore, because the CLT is based on a regenerative process, there is a simple consistent estimator of the asymptotic variance. We illustrate the method with two applications in Bayesian sensitivity analysis. The first concerns one-way random effects models under different priors. The second involves Bayesian variable selection in linear regression, and for this application, importance sampling based on multiple chains enables an empirical Bayes approach to variable selection.

Key words and phrases: Geometric ergodicity, importance sampling, Markov chain Monte Carlo, regenerative simulation, standard errors.

1 Introduction

Importance sampling is a classical Monte Carlo technique in which a random sample from one probability density is used to estimate an expectation with respect to another. Let π and π_1 denote two probability densities on the space X with respect to the measure μ, and assume that the support of π is contained in that of π_1. Suppose that $\pi(x) = \nu(x)/m$ and $\pi_1(x) = \nu_1(x)/m_1$ where ν and ν_1 are completely known functions of x, and m and m_1 are the corresponding normalizing constants. Whereas in some applications these normalizing constants are known, in Bayesian analysis they are typically analytically intractable integrals. Suppose that f is a π-integrable function and we want an estimate of the intractable expectation $\eta := E_\pi f$. Note that even without knowledge of the normalizing constants m and m_1, we may express η as a ratio of expectations with respect to π_1 of known functions by writing

$$\eta = \int_{\mathsf{X}} \frac{f(x)\nu(x)/m}{\nu_1(x)/m_1} \pi_1(x)\,\mu(dx) = \int_{\mathsf{X}} \frac{f(x)\nu(x)/m}{\nu_1(x)/m_1} \pi_1(x)\,\mu(dx) \bigg/ \int_{\mathsf{X}} \frac{\nu(x)/m}{\nu_1(x)/m_1} \pi_1(x)\,\mu(dx). \tag{1.1}$$

The ratio m_1/m cancels from the numerator and denominator of the right side of (1.1), and we have $\eta = E_{\pi_1} v / E_{\pi_1} u$, where $u(x) = \nu(x)/\nu_1(x)$ and $v(x) = f(x)u(x)$ are known. Therefore, if we can simulate an iid sequence $X_1, X_2, \ldots$ from π_1, then the strong law of large numbers (SLLN) implies that

$$\widetilde{\eta}_n = \frac{\bar{v}_n}{\bar{u}_n} \tag{1.2}$$

is a strongly consistent estimator of η. Here, for any function h we use $\bar{h}_n$ to denote $n^{-1} \sum_{i=1}^n h(X_i)$.

We will often be interested in estimating $E_\pi f$ where π ranges over a large collection Π, and we will want to do this via importance sampling, so that we do not have to obtain a separate sample from each $\pi \in \Pi$. If we select some $\pi_1 \in \Pi$ and carry out the method described above, we will see poor performance whenever π is not similar to π_1, because then $\nu(X_i)/\nu_1(X_i)$ can take on very large values for some X_i's. The corresponding terms then dominate the sums in $\bar{u}_n$ and $\bar{v}_n$, which causes $\widetilde{\eta}_n$ to be unstable. However, it is typically the case that no $\pi_1 \in \Pi$ is similar to all $\pi \in \Pi$. One way of dealing with this problem is to replace π_1 in (1.1) by a mixture of k densities $\pi_1, \ldots, \pi_k$, in order to "cover more territory." Let $\boldsymbol{a} = (a_1, \ldots, a_k)$ be a vector of positive numbers, let $|\boldsymbol{a}| = \sum_{l=1}^k a_l$, and define $\bar{\pi} = |\boldsymbol{a}|^{-1} \sum_{l=1}^k a_l \pi_l$, which is a probability density. Unfortunately, doing importance sampling with respect to a mixture of densities, each known only up to a normalizing constant, is far more complicated than carrying out the single-density procedure described above (there is no analogue of the simple cancellation of m_1/m that occurred in (1.1)). For $l = 1, \ldots, k$, write $\pi_l(x)$ in the form of $\nu_l(x)/m_l$, let $d_l = m_l/m_1$ and denote $\boldsymbol{d} = (d_2, \ldots, d_k)$. In our initial description, we proceed as if $\boldsymbol{d}$ is known; we will study the case where we need to estimate $\boldsymbol{d}$ in Section 2. As in (1.1), we have

$$\begin{aligned}\eta &= \int_{\mathsf{X}} f(x)\pi(x)\,\mu(dx) = \int_{\mathsf{X}} f(x)\frac{\pi(x)}{\bar{\pi}(x)}\bar{\pi}(x)\,\mu(dx) \bigg/ \int_{\mathsf{X}} \frac{\pi(x)}{\bar{\pi}(x)}\bar{\pi}(x)\,\mu(dx) \\ &= \left(\sum_{l=1}^k a_l \int_{\mathsf{X}} f(x)\frac{\nu(x)}{\sum_{s=1}^k a_s\nu_s(x)/d_s}\pi_l(x)\,\mu(dx)\right) \bigg/ \left(\sum_{l=1}^k a_l \int_{\mathsf{X}} \frac{\nu(x)}{\sum_{s=1}^k a_s\nu_s(x)/d_s}\pi_l(x)\,\mu(dx)\right) \\ &= \frac{\sum_{l=1}^k a_l E_{\pi_l} v(X;\boldsymbol{a},\boldsymbol{d})}{\sum_{l=1}^k a_l E_{\pi_l} u(X;\boldsymbol{a},\boldsymbol{d})}, \end{aligned} \tag{1.3}$$

where

$$u(x;\boldsymbol{a},\boldsymbol{d}) = \frac{\nu(x)}{\sum_{s=1}^k a_s\nu_s(x)/d_s} \quad \text{and} \quad v(x;\boldsymbol{a},\boldsymbol{d}) = f(x)u(x;\boldsymbol{a},\boldsymbol{d}). \tag{1.4}$$

Let $n = (n_1,\dots,n_k)$. Suppose that for each l we can simulate an iid sample of size n_l, $\{X_i^{(l)},\, i = 1,\dots,n_l\}$, from π_l. Then

$$\hat{v}_n := \sum_{l=1}^k \frac{a_l}{n_l}\sum_{i=1}^{n_l} v(X_i^{(l)};\boldsymbol{a},\boldsymbol{d}) \xrightarrow{\text{a.s.}} \sum_{l=1}^k a_l E_{\pi_l} v(X;\boldsymbol{a},\boldsymbol{d}) \qquad \text{as } n_1,\dots,n_k \to \infty,$$

and

$$\hat{u}_n := \sum_{l=1}^k \frac{a_l}{n_l}\sum_{i=1}^{n_l} u(X_i^{(l)};\boldsymbol{a},\boldsymbol{d}) \xrightarrow{\text{a.s.}} \sum_{l=1}^k a_l E_{\pi_l} u(X;\boldsymbol{a},\boldsymbol{d}) \qquad \text{as } n_1,\dots,n_k \to \infty.$$

Therefore,

$$\hat{\eta}_n = \frac{\hat{v}_n}{\hat{u}_n} = \left(\sum_{l=1}^k \frac{a_l}{n_l}\sum_{i=1}^{n_l} \frac{f(X_i^{(l)})\nu(X_i^{(l)})}{\sum_{s=1}^k a_s\nu_s(X_i^{(s)})/d_s}\right) \bigg/ \left(\sum_{l=1}^k \frac{a_l}{n_l}\sum_{i=1}^{n_l} \frac{\nu(X_i^{(l)})}{\sum_{s=1}^k a_s\nu_s(X_i^{(s)})/d_s}\right) \tag{1.5}$$

is a consistent estimator of η by the SLLN, and this fact justifies the standard procedure in which we choose some values $n_1,\dots,n_k$, for each l simulate random samples of size n_l from π_l, and use the observed value of $\hat{\eta}_n$ as an estimate of η. Estimators of the form $\widetilde{\eta}_n$ and $\hat{\eta}_n$ will be called *ratio estimators*.

Most statisticians would agree that a Monte Carlo estimate without an associated standard error is not very useful. We call the importance sampling procedure “honest” if the estimate, $\widetilde{\eta}_n$ or $\hat{\eta}_n$, is accompanied by a valid asymptotic standard error. Of course, the computation of a standard error is based upon the existence of a central limit theorem (CLT). Consider $\widetilde{\eta}_n$ for example. A simple argument involving the delta method shows that, if $E_{\pi_1}[u^2]$ and $E_{\pi_1}[v^2]$ are both finite, then as $n \to \infty$, $n^{1/2}(\widetilde{\eta}_n - \eta) \xrightarrow{d} \mathcal{N}(0,\tau^2)$, where $\tau^2 = E_{\pi_1}[(v-u\eta)^2]/[E_{\pi_1}u]^2$. Moreover, a simple consistent estimator of τ^2 is given by

$$\widetilde{\tau}_n^2 = \frac{1}{n}\sum_{i=1}^n \Big[\big(v(X_i) - u(X_i)\widetilde{\eta}_n\big)^2\Big] \bigg/ \bigg[\frac{1}{n}\sum_{i=1}^n u(X_i)\bigg]^2.$$

Hence the standard error of the estimate is the observed value of $\widetilde{\tau}_n/n^{1/2}$.

Hastings (1970) and others have pointed out that importance sampling can also be used in the Markov chain Monte Carlo (MCMC) context. Consider the version of the importance sampling method that uses only one density π_1, for example. Suppose now that it is not feasible to make exact draws from π_1, but we have at our disposal an MCMC algorithm for exploring π_1. In particular, let the sequence $X_1, X_2, \ldots$ be, instead of a random sample from π_1, a Markov chain with invariant density π_1. Assume that the chain satisfies the basic regularity conditions (irreducibility, aperiodicity and Harris recurrence) that underlie the ergodic theorem (Meyn and Tweedie, 1993, Chapter 17). Then, if h is a π_1-integrable function, the ergodic theorem implies that $\bar{h}_n := n^{-1} \sum_{i=1}^{n} h(X_i)$ is strongly consistent for $E_{\pi_1} h$. This shows that the estimator $\widetilde{\eta}_n = \bar{v}_n / \bar{u}_n$ remains strongly consistent when the random sample is replaced by a well-behaved Markov chain. There is no "free lunch," however, since the computation of standard errors is much more complicated when $\widetilde{\eta}_n$ is based on a Markov chain. First, the two simple moment conditions that guarantee that $\widetilde{\eta}_n$ obeys a CLT in the iid case are no longer enough. Second, even when a CLT does hold, the asymptotic variance has a complex form and is difficult to estimate consistently. These problems have prevented more widespread use of importance sampling in the MCMC context. In this paper, we provide a simple method for computing a valid asymptotic standard error for Markov chain based importance sampling estimators of the form $\widetilde{\eta}_n$ and $\hat{\eta}_n$.

We begin by considering the simpler problem of computing a standard error for $\bar{h}_n$. There are several different approaches to this problem based on time series methods, batching, and regeneration (see, e.g., Glynn and Iglehart, 1987; Geyer, 1992; Mykland et al., 1995; Jones et al., 2006). From both theoretical and practical standpoints, the cleanest of these methods is the one based on regenerative simulation. A *regeneration* is a (random) time at which a stochastic process probabilistically restarts itself. Consider for example a Markov chain on the countable state space $\{0, 1, 2, 3, \ldots\}$, and suppose the chain is started at the point 0. Then the random times at which the chain returns to the point 0 are regeneration times, because at those times the distribution of the process going forward is the same as when it was started. Most of the Markov chains that drive MCMC algorithms have continuous state spaces, and this complicates the identification of regeneration times. However, Mykland et al. (1995) provided a general technique that is based on the construction of a *minorization condition*. The benefit of identifying regeneration times is that the "tours" made by the chain in between these random times are iid, and this fact pays huge dividends in the asymptotic analysis of ergodic averages like $\bar{h}_n$. In particular, Hobert et al. (2002) showed that, if the underlying Markov chain converges to π_1 at a geometric rate and there exists an $\epsilon > 0$ such that $E_{\pi_1} |h|^{2+\epsilon} < \infty$, then $\bar{h}_n$ obeys a CLT whose asymptotic variance is easy to estimate consistently.

Multiple-sample based estimators of the form (1.5) have been discussed by several authors before. Vardi (1985), Gill et al. (1988), Meng and Wong (1996), Kong et al. (2003), and Tan (2004) deal with the iid case. Geyer (1994) and Buta and Doss (2011) work in the setting where the samples are Markov chains. Here, the asymptotic variance is extremely complicated, and these authors largely leave aside the question of how to produce asymptotically valid standard error estimates. In the present paper, we provide an extension of the regenerative methods of Mykland et al. (1995) and Hobert et al. (2002) that can be applied to the multiple-chain importance sampling estimators $\hat{\eta}_n$. A special case of our results pertains to single-chain importance sampling estimators, $\widetilde{\eta}_n$, which was also studied in Bhattacharya (2008). One implication of these results for the analysis of $\widetilde{\eta}_n$ is that if the Markov chain from π_1 is geometrically ergodic, and there exists an $\epsilon > 0$ such that $E_{\pi_1}|u|^{2+\epsilon}$ and $E_{\pi_1}|v|^{2+\epsilon}$ are both finite, then $\widetilde{\eta}_n$ obeys a CLT whose asymptotic variance is easy to estimate consistently. (Note that the moment conditions are only slightly stronger than those required in the iid case.) This result is stated in our Corollary 1. It enables the computation of a valid asymptotic standard error for $\widetilde{\eta}_n$, and hence for honest importance sampling in the MCMC context.

The principal application we have in mind involves analysis of sensitivity to the prior in a Bayesian framework. In an initial instance of this, suppose that p_1 and p are two prior densities and let π_1 and π be the corresponding posteriors. We have $\pi_1(x) = \ell(x)p_1(x)/m_1$ and $\pi(x) = \ell(x)p(x)/m$, where $\ell(x)$ is the likelihood function and the m's are normalizing constants. This is the framework mentioned earlier in that, except for normalizing constants, π_1 and π are known functions and, moreover, the ratio $\ell(x)p(x)/(\ell(x)p_1(x))$ boils down to simply a ratio of priors. If the regularity conditions described earlier are satisfied, then we can use the MCMC algorithm for π_1 to perform honest exploration of π, which obviates the need to develop and study an MCMC algorithm for π.

The main sensitivity analysis problem we have in mind is considerably more complicated and is described as follows. Suppose $\{p_h,\, h \in \mathcal{H}\}$ is a parametric family of priors and let π_h given by $\pi_h(x) = \ell(x)p_h(x)/m_h$ be the corresponding posteriors. Note that m_h is the marginal likelihood of the data under prior p_h. Two problems we wish to consider are: (i) for a given function f, estimate $\eta_h = E_{\pi_h} f$ for all $h \in \mathcal{H}$; this is needed for sensitivity analysis. A closely related problem is: (ii) estimate the family $m_h,\, h \in \mathcal{H}$ and subsequently $h_{\text{opt}} = \text{argmax}_h\, m_h$; this value is by definition the empirical Bayes choice of the hyperparameter h and is needed to implement empirical Bayes methodology. The single chain importance sampling approach does not work well here because for no single $h_1 \in \mathcal{H}$ it is the case that π_{h_1} is similar to π_h for all $h \in \mathcal{H}$, and it is for this reason that multiple-chain importance sampling is necessary. The multiple-chain regenerative method we develop in this paper enables us to obtain valid standard errors for our estimates of η_h and m_h.

This paper is organized as follows. In Section 2 we show how regenerative methods can be used to construct CLTs for estimators which are ratios of weighted sums of ergodic averages. The results we obtain apply to estimators of the form $\widetilde{\eta}_n$ and $\hat{\eta}_n$. We also discuss the advantages of using regenerative simulation, and the practical limitations of the approach. In Section 3 we describe the application of these results to the single-chain importance sampling estimators $\widetilde{\eta}_n$, and we illustrate the use of our methods in an example involving a Bayesian one-way random effects model. Specifically, we use a well-studied Gibbs sampler for the posterior associated with a standard diffuse prior to make inferences about an alternative posterior based on the so-called reference prior, which is neither conjugate nor conditionally conjugate. In Section 4 we consider a standard model for variable selection in Bayesian linear regression, in which the prior is indexed by a hyperparameter whose selection plays a critical role in how variable selection is carried out. We show how multiple-chain importance sampling, together with our regenerative methods, can be used to carry out sensitivity analysis and empirical Bayes methodology.

2 A Regeneration-Based Central Limit Theorem for Ratio Estimators

Let $\mathcal{H}$ be an index set, for each $h \in \mathcal{H}$ let π_h be a probability density on the measurable space $(\mathsf{X}, \mathcal{B})$ with respect to the measure μ, and let f be a function defined on X. We consider the situation where for each h the intractable integral $\eta_h = E_{\pi_h} f$ can be represented as a ratio of weighted expectations of the form (1.3). (A sufficient condition for this is that the union of the supports of $\pi_{h_1}, \dots, \pi_{h_k}$ contains that of π_h.) Our goal is to estimate η_h using MCMC methods and to provide a standard error for our estimator. Let $h_1, \dots, h_k \in \mathcal{H}$, and suppose we are able to generate k Markov chains with invariant densities $\pi_{h_1}, \dots, \pi_{h_k}$ (the k chains are generated independently). For $l = 1, \dots, k$, let $\Phi_l = \{X_0^{(l)}, X_1^{(l)}, \dots\}$ denote the l^{th} Markov chain. We consider the estimator of η_h defined by (1.5). Before proceeding we remark on notation. Although we have in mind the situation where $h, h_1, \dots, h_k$ can all vary over $\mathcal{H}$, where $\mathcal{H}$ is a large set, we will write $\pi, \pi_1, \dots, \pi_k$ instead of $\pi_h, \pi_{h_1}, \dots, \pi_{h_k}$ whenever we are not varying $h, h_1, \dots, h_k$, in order to lighten the notation.

For each l, we assume that Φ_l is *Harris ergodic*, which means that Φ_l is ψ-irreducible, aperiodic and Harris recurrent, where ψ represents the maximal irreducibility measure of the chain (see Meyn and Tweedie (1993, Chap. 4 & 9) or Roberts and Rosenthal (2004) for definitions). Harris ergodicity, which is typically easy to check in practice, ensures that the ergodic theorem holds so that ergodic averages are guaranteed to converge (almost surely) to their population counterparts. However, Harris ergodicity is not enough to guarantee that ergodic averages obey CLTs, and it is worth noting that seemingly reasonable MCMC algorithms for which CLTs do not hold are not uncommon (see, e.g., Roberts, 1999). We will further assume that for each l the chain Φ_l converges to π_l at a geometric rate. In what follows, we use the regeneration idea to establish a CLT for $\hat{\eta}_n$. The main benefit of this regeneration-based CLT is that it enables us to obtain consistent estimates of the asymptotic variance in a straightforward manner.

Our first task is to introduce regenerations into the Markov chains. In most statistical applications of MCMC, the state space is continuous, so there are no single points to which a chain returns with positive probability. Mykland et al. (1995) showed that in such cases it may be possible to introduce regenerations through minorization conditions, which we now describe. Let $K_x^{(l)}(A)$ be the Markov transition function for Φ_l, so that for any $A \in \mathcal{B}$ we have $P\bigl(X_{n+1}^{(l)} \in A \,|\, X_n^{(l)} = x\bigr) = K_x^{(l)}(A)$. Suppose that for each l we can identify a function $s_l \colon \mathsf{X} \to [0,1)$ with $E_{\pi_l} s_l > 0$, and a probability measure Q_l on $(\mathsf{X}, \mathcal{B})$, such that

$$K_x^{(l)}(A) \geq s_l(x)\, Q_l(A) \qquad \text{for all } x \in \mathsf{X} \text{ and } A \in \mathcal{B}. \tag{2.1}$$

This so-called minorization condition allows us to express $K^{(l)}$ as a mixture of two probability measures, one of which does not depend on the current state. Indeed, define the Markov transition function $G^{(l)}$ by

$$G_x^{(l)}(A) = \frac{K_x^{(l)}(A) - s_l(x) Q_l(A)}{1 - s_l(x)}.$$

Note that for fixed $x \in \mathsf{X}$, $G_x^{(l)}$ is a probability measure. We may therefore write

$$K_x^{(l)} = s_l(x) Q_l + (1 - s_l(x)) G_x^{(l)}.$$

This mixture representation provides an alternative method for simulating the Markov chain. Given the current state, $X_n^{(l)} = x$, we can draw $X_{n+1}^{(l)}$ by generating $\delta_n \sim \text{Bernoulli}(s_l(x))$, and then drawing $X_{n+1}^{(l)} \sim Q_l$ if $\delta_n = 1$ or $X_{n+1}^{(l)} \sim G_x^{(l)}$ if $\delta_n = 0$. Since Q_l does not depend on the current state, this method of simulation results in a regeneration every time $\delta_n = 1$. In other words, suppose we initiate the chain with

$$X_0^{(l)} \sim Q_l \tag{2.2}$$

and we proceed to simulate using the sequential method just described. Every time that $\delta_n = 1$, $X_{n+1}^{(l)}$ is drawn from Q_l and the process probabilistically restarts itself. The *regeneration times* are $\tau_0^{(l)} = 0$ and $\tau_t^{(l)} = \min\{n > \tau_{t-1}^{(l)} : \delta_{n-1} = 1\}$ for $t = 1, 2, \ldots$. Accordingly, the chain is broken up into "tours" $\bigl\{\bigl(X_{\tau_{t-1}^{(l)}}, \ldots, X_{\tau_t^{(l)}-1}\bigr),\, t = 1, 2, \ldots\bigr\}$ that are independent stochastic replicas of each other. The ability to re-express each chain in this way drastically simplifies the asymptotic analysis of estimators based on the chains. When performing regenerative simulation in practice, it is usually problematic, if not impossible, to draw from $G_x^{(l)}$. Fortunately, there is a simple trick that allows us to circumvent this obstacle (Mykland et al., 1995, following Nummelin 1984, p. 62). It turns out that we can just simulate each Φ_l in the usual way, except that, after the $(n+1)^{\text{th}}$ iteration for $n = 0, 1, \ldots$, we generate a Bernoulli random variable, δ_n, that indicates whether or not a regeneration occurred. Its conditional success probability is given by

$$P\bigl(\delta_n = 1 \,|\, X_n^{(l)} = x, X_{n+1}^{(l)} = \widetilde{x}\bigr) = \left[\frac{d(s_l(x) Q_l)}{dK_x^{(l)}}\right](\tilde{x}), \tag{2.3}$$

where $\bigl[d(s_l(x) Q_l)/dK_x^{(l)}\bigr]$ is the Radon-Nikodym derivative of $s_l(x) Q_l$ with respect to $K_x^{(l)}$, whose existence is implied by (2.1).

Below we explain how regenerative simulation of the k Markov chains helps us analyze the asymptotic behavior of the estimator $\hat{\eta}_n$ defined in (1.5). Note that $\hat{\eta}_n$ involves the vector of ratios $\boldsymbol{d} = (d_2, \ldots, d_l) =$

$(m_2/m_1, \dots, m_k/m_1)$. We discuss two situations. In Section 2.1 we consider the case where $\boldsymbol{d}$ is known, and we do this for two reasons: (i) there are interesting statistical models in which $\boldsymbol{d}$ is known (see Section 4.1), and (ii) understanding the case of known $\boldsymbol{d}$ is necessary in order to understand the case where $\boldsymbol{d}$ is unknown and must be estimated. In Section 2.2 we consider the case where $\boldsymbol{d}$ is unknown. Of course, when $k = 1$, $\hat{\eta}_n$ reduces to the single-chain importance sampling estimator $\widetilde{\eta}_n$, and knowledge of the normalizing constants is not required. Hence the regeneration-based CLT concerning $\widetilde{\eta}_n$ is presented as a special case of the result for the situation where $\boldsymbol{d}$ is known.

2.1 The Case Where $\boldsymbol{d}$ Is Known

Suppose we simulate R_l tours of the l^{th} Markov chain for $l = 1, \dots, k$; that is, we begin the simulation by drawing $X_0^{(l)} \sim Q_l(\cdot)$, and we stop the simulation when we observe the R_l^{th} success among the Bernoulli trials. The length of the l^{th} chain is $n_l = \tau_{R_l}^{(l)}$. Based on these k Markov chains, we construct the importance sampling estimator of η using (1.5). We will write $\hat{u}_n(\boldsymbol{a}, \boldsymbol{d})$, $\hat{v}_n(\boldsymbol{a}, \boldsymbol{d})$, and $\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d})$ instead of $\hat{u}_n$, $\hat{v}_n$, and $\hat{\eta}_n$ respectively, whenever we need to emphasize the dependence of these estimators on a given vector of weights $\boldsymbol{a}$ and the vector of known ratios of normalizing constants $\boldsymbol{d}$. For $t = 1, 2, \dots, R_l$ define

$$V_t^{(l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} v(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d}), \quad U_t^{(l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} u(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d}), \quad \text{and} \quad T_t^{(l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} 1 = \tau_t^{(l)} - \tau_{t-1}^{(l)}, \tag{2.4}$$

where v and u are defined in (1.4) and the sums range over the values of i that constitute the t^{th} tour. The key feature of the above construction is that for each l, $(V_t^{(l)}, U_t^{(l)}, T_t^{(l)})$ are iid triples. Let $\bar{T}^{(l)} = R_l^{-1} \sum_{t=1}^{R_l} T_t^{(l)}$ be the average tour length and, analogously, let $\bar{V}^{(l)} = R_l^{-1} \sum_{t=1}^{R_l} V_t^{(l)}$ and $\bar{U}^{(l)} = R_l^{-1} \sum_{t=1}^{R_l} U_t^{(l)}$. Then the numerator and the denominator of $\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d})$ can be written as

$$\hat{v}_n = \sum_{l=1}^k \frac{a_l}{n_l} \sum_{i=0}^{n_l-1} v(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d}) = \sum_{l=1}^k a_l \frac{\bar{V}^{(l)}}{\bar{T}^{(l)}} \text{ and } \hat{u}_n = \sum_{l=1}^k \frac{a_l}{n_l} \sum_{i=0}^{n_l-1} u(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d}) = \sum_{l=1}^k a_l \frac{\bar{U}^{(l)}}{\bar{T}^{(l)}}, \tag{2.5}$$

respectively. This representation makes it easy to study the asymptotic behavior of $\hat{v}_n(\boldsymbol{a}, \boldsymbol{d})$, $\hat{u}_n(\boldsymbol{a}, \boldsymbol{d})$, and $\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d})$ using iid theory. We have $\bar{V}^{(l)} \xrightarrow{\text{a.s.}} E\big(V_1^{(l)}\big) = E_{\pi_l}(v)/E_{\pi_l}(s_l)$, where the convergence statement follows from the SLLN and the equality follows from Kac's theorem (Meyn and Tweedie, 1993, Thm 10.2.2). Similarly, $\bar{U}^{(l)} \xrightarrow{\text{a.s.}} E_{\pi_l}(u)/E_{\pi_l}(s_l)$, and $\bar{T}^{(l)} \xrightarrow{\text{a.s.}} [E_{\pi_l}(s_l)]^{-1}$. Therefore, from (2.5) we have

$$\hat{\eta}_n = \frac{\hat{v}_n}{\hat{u}_n} \xrightarrow{\text{a.s.}} \frac{\sum_{l=1}^k a_l E_{\pi_l} v}{\sum_{l=1}^k a_l E_{\pi_l} u} = \eta \qquad \text{as } R_l \to \infty \text{ for } l = 1, \dots, k.$$

It is useful to note that the limit of $\hat{u}_n$ itself is a meaningful quantity in Bayesian model comparison settings. We have

$$\hat{u}_n \xrightarrow{\text{a.s.}} \sum_{l=1}^k a_l E_{\pi_l} u = \sum_{l=1}^k a_l E_{\pi_l} u(X; \boldsymbol{a}, \boldsymbol{d}) = \sum_{l=1}^k a_l \int_{\mathsf{X}} \frac{\nu(x)}{\sum_{s=1}^k a_s \nu_s(x)/d_s} \pi_l(x)\, \mu(dx)$$
$$= \int_{\mathsf{X}} \frac{\sum_{l=1}^k a_l \nu_l(x)/m_l}{\sum_{s=1}^k a_s \nu_s(x)/(m_s/m_1)} \nu(x)\, \mu(dx) = \int_{\mathsf{X}} \frac{m\pi(x)}{m_1}\, \mu(dx) = \frac{m}{m_1}.$$

If we are in the Bayesian framework in which $\pi_1(x) = \ell_1(x)p_1(x)/m_1$ and $\pi(x) = \ell(x)p(x)/m$, where the ℓ's are likelihood functions and the m's are normalizing constants, then the ratio m/m_1 is the so-called Bayes factor between the two models. This quantity is often used to carry out model selection.

We now study the asymptotic distributions of $\hat{\eta}_n$ and $\hat{u}_n$. By the same delta method argument that we used earlier, we see that if

$$E\Big[\big(V_1^{(l)}\big)^2\Big] < \infty, \quad E\Big[\big(U_1^{(l)}\big)^2\Big] < \infty \quad \text{and} \quad E\Big[\big(T_1^{(l)}\big)^2\Big] < \infty \qquad \text{for } l = 1, \dots, k, \tag{2.6}$$

then

$$R_l^{1/2}\big(\bar{V}^{(l)}/\bar{T}^{(l)} - E_{\pi_l} v\big) \xrightarrow{d} \mathcal{N}(0, \sigma_l^2) \quad \text{and} \quad R_l^{1/2}\big(\bar{U}^{(l)}/\bar{T}^{(l)} - E_{\pi_l} u\big) \xrightarrow{d} \mathcal{N}(0, \kappa_l^2),$$

where

$$\sigma_l^2 = \frac{E\Big[\big(V_1^{(l)} - T_1^{(l)} E_{\pi_l} v\big)^2\Big]}{\big(ET_1^{(l)}\big)^2} \quad \text{and} \quad \kappa_l^2 = \frac{E\Big[\big(U_1^{(l)} - T_1^{(l)} E_{\pi_l} u\big)^2\Big]}{\big(ET_1^{(l)}\big)^2}.$$

To obtain CLTs for $\hat{u}_n$ and $\hat{\eta}_n$ based on the above results, we assume that $R_l/R_1 \to b_l \in (0, \infty)$ for $l = 2, \dots, k$ as $R_1 \to \infty$, i.e., the relative sizes of the R_l's remain fixed as they grow to infinity. Then the Cramér-Wold Theorem implies that for each l,

$$R_1^{1/2}\left[\begin{pmatrix} \bar{V}^{(l)}/\bar{T}^{(l)} \\ \bar{U}^{(l)}/\bar{T}^{(l)} \end{pmatrix} - \begin{pmatrix} E_{\pi_l} v \\ E_{\pi_l} u \end{pmatrix}\right] = \left(\frac{R_1}{R_l}\right)^{1/2} R_l^{1/2}\left[\begin{pmatrix} \bar{V}^{(l)}/\bar{T}^{(l)} \\ \bar{U}^{(l)}/\bar{T}^{(l)} \end{pmatrix} - \begin{pmatrix} E_{\pi_l} v \\ E_{\pi_l} u \end{pmatrix}\right] \xrightarrow{d} \mathcal{N}_2(0, b_l^{-1}\Sigma_l),$$

where

$$\Sigma_l = \big(ET_1^{(l)}\big)^{-2} \operatorname{Var}\begin{pmatrix} V_1^{(l)} - T_1^{(l)} E_{\pi_l} v \\ U_1^{(l)} - T_1^{(l)} E_{\pi_l} u \end{pmatrix}.$$

Let

$$Z_n = \left(\frac{\bar{V}^{(1)}}{\bar{T}^{(1)}}, \frac{\bar{U}^{(1)}}{\bar{T}^{(1)}}, \dots, \frac{\bar{V}^{(k)}}{\bar{T}^{(k)}}, \frac{\bar{U}^{(k)}}{\bar{T}^{(k)}}\right)^\top \quad \text{and} \quad \xi = \big(E_{\pi_1} v, E_{\pi_1} u, \dots, E_{\pi_k} v, E_{\pi_k} u\big)^\top. \tag{2.7}$$

Since the k chains are independent, we actually have

$$R_1^{1/2}(Z_n - \xi) \xrightarrow{d} \mathcal{N}_{2k}(0, \Sigma), \tag{2.8}$$

where $\Sigma = \operatorname{diag}(\Sigma_1, b_2^{-1}\Sigma_2, \dots, b_k^{-1}\Sigma_k)$ is a block diagonal matrix.

From (2.8), it is straightforward to derive CLTs for both $\hat{u}_n$ and $\hat{\eta}_n$. Since $\hat{u}_n$ is simply a linear combination of $\bar{U}^{(1)}/\bar{T}^{(1)}, \dots, \bar{U}^{(k)}/\bar{T}^{(k)}$ (see (2.5)), we have

$$R_1^{1/2}\big(\hat{u}_n(\boldsymbol{a}, \boldsymbol{d}) - m/m_1\big) = \sum_{l=1}^{k} a_l \big[R_1^{1/2}\big(\bar{U}^{(l)}/\bar{T}^{(l)} - E_{\pi_l} u\big)\big] \xrightarrow{d} \mathcal{N}(0, \kappa^2) \qquad \text{as } R_1 \to \infty,$$

where

$$\kappa^2 = \sum_{l=1}^{k} a_l^2 b_l^{-1} \kappa_l^2. \tag{2.9}$$

Letting $g \colon \mathbb{R}^{2k} \to \mathbb{R}$ be defined by

$$g(x) = \frac{a_1 x_1 + a_2 x_3 + \cdots + a_k x_{2k-1}}{a_1 x_2 + a_2 x_4 + \cdots + a_k x_{2k}},$$

we see that $\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d}) = g(Z_n)$ and $\eta = g(\xi)$, where Z_n and ξ are given by (2.7), so by (2.8) and the delta method we have

$$R_1^{1/2}\big(\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d}) - \eta\big) \xrightarrow{d} \mathcal{N}(0, \tau^2) \qquad \text{as } R_1 \to \infty, \tag{2.10}$$

where

$$\tau^2 = \Delta^\top \Sigma \Delta, \tag{2.11}$$

with

$$\Delta = \left(\frac{a_1}{E\hat{u}_n}, -\frac{a_1 E\hat{v}_n}{[E\hat{u}_n]^2}, \dots, \frac{a_k}{E\hat{u}_n}, -\frac{a_k E\hat{v}_n}{[E\hat{u}_n]^2}\right)^\top.$$

The approach to establishing CLTs using regeneration gives rise to the following strongly consistent estimators of the asymptotic variances. Let

$$\hat{\Delta} = \left(\frac{a_1}{\hat{u}_n}, -\frac{a_1 \hat{v}_n}{(\hat{u}_n)^2}, \dots, \frac{a_k}{\hat{u}_n}, -\frac{a_k \hat{v}_n}{(\hat{u}_n)^2}\right)^\top \quad \text{and} \quad \widehat{\Sigma}_l = \frac{1}{R_l\big(\bar{T}^{(l)}\big)^2} \begin{pmatrix} S_l^{(11)} & S_l^{(12)} \\ S_l^{(21)} & S_l^{(22)} \end{pmatrix},$$

where

$$S_l^{(11)} = \sum_{t=1}^{R_l} \Big[V_t^{(l)} - \big(\bar{V}^{(l)}/\bar{T}^{(l)}\big) T_t^{(l)}\Big]^2, \quad S_l^{(22)} = \sum_{t=1}^{R_l} \Big[U_t^{(l)} - \big(\bar{U}^{(l)}/\bar{T}^{(l)}\big) T_t^{(l)}\Big]^2,$$

and

$$S_l^{(12)} = S_l^{(21)} = \sum_{t=1}^{R_l} \Big[V_t^{(l)} - \big(\bar{V}^{(l)}/\bar{T}^{(l)}\big) T_t^{(l)}\Big] \Big[U_t^{(l)} - \big(\bar{U}^{(l)}/\bar{T}^{(l)}\big) T_t^{(l)}\Big],$$

and let $\widehat{\Sigma} = \operatorname{diag}(\widehat{\Sigma}_1, b_2^{-1}\widehat{\Sigma}_2, \dots, b_k^{-1}\widehat{\Sigma}_k)$. Clearly, $\hat{\Delta}$ consistently estimates Δ. And simple calculations show that each component of the difference between $\widehat{\Sigma}_l$ and Σ_l converges almost surely to 0 as $R_l \to \infty$. Finally, let $\hat{\kappa}_l^2$ denote the last entry of $\widehat{\Sigma}_l$. Then

$$\hat{\kappa}^2 = \sum_{l=1}^{k} a_l^2 b_l^{-1} \hat{\kappa}_l^2 \quad \text{and} \quad \hat{\tau}^2 = \hat{\Delta}^\top \widehat{\Sigma} \hat{\Delta} \tag{2.12}$$

are consistent estimators of κ^2 and τ^2 respectively.

Recall that, in order to arrive at the CLT in (2.10), we require the second-moment conditions in (2.6). These conditions are actually quite difficult to check directly. This is because $V_1^{(l)}$ and $U_1^{(l)}$ are sums of functions of the states of the Markov chain containing a random number of terms. However, Hobert et al. (2002) showed that, if the underlying Markov chain is geometrically ergodic and there exists an $\epsilon > 0$ such that $E_{\pi_l}|v(X; \boldsymbol{a}, \boldsymbol{d})|^{2+\epsilon}$ and $E_{\pi_l}|u(X; \boldsymbol{a}, \boldsymbol{d})|^{2+\epsilon}$ are finite, then the second-moment conditions in (2.6) hold. See Section 5 for some discussion concerning geometric ergodicity.

We summarize the above results in the following theorem.

Theorem 1 *Suppose that for each $l = 1, \dots, k$, the following conditions hold.*

1. *The Markov chain $\Phi_l = \{X_0^{(l)}, X_1^{(l)}, \dots\}$ is geometrically ergodic and has π_l as its invariant density.*
2. *The Markov transition function $K^{(l)}$ satisfies the minorization condition* (2.1).
3. *There exists $\epsilon > 0$ such that $E_{\pi_l}|v(X; \boldsymbol{a}, \boldsymbol{d})|^{2+\epsilon}$ and $E_{\pi_l}|u(X; \boldsymbol{a}, \boldsymbol{d})|^{2+\epsilon}$ are finite.*
4. *$R_l/R_1 \to b_l \in (0, \infty)$ as $R_1 \to \infty$.*

Then we have the following CLTs:

$$R_1^{1/2}\big(\hat{u}_n(\boldsymbol{a}, \boldsymbol{d}) - m/m_1\big) \xrightarrow{d} \mathcal{N}(0, \kappa^2) \quad \textit{and} \quad R_1^{1/2}\big(\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d}) - \eta\big) \xrightarrow{d} \mathcal{N}(0, \tau^2) \qquad \textit{as } R_1 \to \infty.$$

Furthermore, $\hat{\kappa}^2$ and $\hat{\tau}^2$ defined in (2.12) *are strongly consistent estimators of κ^2 and τ^2 respectively.*

Now consider the case $k = 1$, for which $\boldsymbol{a} = 1$ and the $(k-1)$-dimensional vector $\boldsymbol{d}$ is irrelevant. Note that $\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d})$ reduces to the single-chain importance sampling estimator $\widetilde{\eta}_n$ defined in (1.2), and $\hat{u}_n(\boldsymbol{a}, \boldsymbol{d})$ reduces to

$$\bar{u}_n = \frac{1}{n}\sum_{i=1}^{n} u(X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{\nu(X_i)}{\nu_1(X_i)}.$$

Hence, even when the ratio of the normalizing constants m and m_1 is unknown, CLTs for $\bar{u}_n$ and $\widetilde{\eta}_n$ follow as special cases of Theorem 1. We summarize the CLT for single-chain importance sampling estimators in the following Corollary. Part of this result also appeared in Bhattacharya (2008).

Corollary 1 *Suppose that $\Phi = \{X_i,\, i = 0, 1, 2, \dots\}$ is a Markov chain which is geometrically ergodic and has π_1 as its invariant density. Suppose further that Φ has a Markov transition function $K^{(l)}$ satisfying* (2.1). *Let $\widetilde{\eta}_n$ defined by* (1.2) *be the estimator of the ratio $\eta = E_\pi v / E_\pi u$ in* (1.1), *where $n = \tau_R$ is the number of iterations required to get R regenerations. If there exists an $\epsilon > 0$ such that $E_{\pi_1}|v|^{2+\epsilon}$ and $E_{\pi_1}|u|^{2+\epsilon}$ are finite, then*

$$R^{1/2}(\bar{u}_n - m/m_1) \xrightarrow{d} \mathcal{N}(0, \underline{\kappa}^2) \quad \textit{and} \quad R^{1/2}(\widetilde{\eta}_n - \eta) \xrightarrow{d} \mathcal{N}(0, \underline{\tau}^2) \qquad \textit{as } R \to \infty.$$

Moreover,

$$\hat{\underline{\kappa}}^2 = \frac{R^{-1}\sum_{t=1}^{R}(U_t - \bar{u}_n T_t)^2}{\big(R^{-1}\sum_{t=1}^{R} T_t\big)^2} \quad \textit{and} \quad \hat{\underline{\tau}}^2 = \frac{R^{-1}\sum_{t=1}^{R}(V_t - \widetilde{\eta}_n U_t)^2}{\big(R^{-1}\sum_{t=1}^{R} U_t\big)^2}$$

are strongly consistent estimators of $\underline{\kappa}^2$ and $\underline{\tau}^2$. Here, U_t is the sum of $u(X_i) = \nu(X_i)/\nu_1(X_i)$ over the t^{th} tour of the Markov chain and V_t is the sum of $v(X_i) = f(X_i)u(X_i)$ over the t^{th} tour.

Remark 1 *Corollary 1 is a direct generalization of the results in Mykland et al. (1995) and Hobert et al. (2002), whose results may be viewed as pertaining to the special case where $u \equiv 1$. In that case, $U_t = \tau_t - \tau_{t-1}$ is the length of the t^{th} tour and, of course, the condition $E_{\pi_1}|u|^{2+\epsilon} < \infty$ holds automatically.*

A nice feature of Theorem 1 and Corollary 1 is that the conditions for the existence of a CLT are separated into ones that concern the convergence rate of the regenerative Markov chain and others that are simple moment conditions with respect to the invariant 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 very difficult to check in practice because they involve expectations of complex functions of the underlying process. An example of this is Mykland et al.'s (1995) main result (which is similar to our Corollary 1), where it is assumed that $EV_1^2 < \infty$. 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; Lavenberg and Slutz, 1975; Glynn and Iglehart, 1987). The Markov processes that underlie these analyses have countable state spaces, which makes the identification of regeneration times trivial. However, unlike Theorem 1 and Corollary 1, 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.

2.2 The Case Where $\boldsymbol{d}$ Is Unknown

Except for Corollary 1, results from the previous section are applicable only if $\boldsymbol{d} = (m_2/m_1, \dots, m_k/m_1)$ is known. In the usual situation where $\boldsymbol{d}$ is unknown, the multiple-chain importance sampling technique can still be applied if in the expressions for $\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d})$ and $\hat{u}_n(\boldsymbol{a}, \boldsymbol{d})$ we replace $\boldsymbol{d}$ by an estimate $\hat{\boldsymbol{d}}$.

Given the k chains $\Phi_1, \dots, \Phi_k$, it is possible to form an estimate $\hat{\boldsymbol{d}}$ of $\boldsymbol{d}$—how to do this is discussed below—and plug in $\hat{\boldsymbol{d}}$ in place of $\boldsymbol{d}$ in (1.5). Call the resulting estimate $\hat{\eta}_n(\boldsymbol{a}, \hat{\boldsymbol{d}})$. It turns out that the variance of $\hat{\eta}_n(\boldsymbol{a}, \hat{\boldsymbol{d}})$ is greater than that of $\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d})$. Thus, the variance decomposes as $\mathrm{Var}\big(\hat{\eta}_n(\boldsymbol{a}, \hat{\boldsymbol{d}})\big) = \mathrm{Var}\big(\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d})\big) + V_d$, where V_d is the increase in variance resulting from using $\hat{\boldsymbol{d}}$ instead of $\boldsymbol{d}$. For the sensitivity analysis problems we have in mind, we wish to compute $E_{\pi_h} f$ for a large number of h's. Because for each h the computational time needed to form $\hat{\eta}_n(\boldsymbol{a}, \hat{\boldsymbol{d}})$ is linear in the total sample size $n_1 + \cdots + n_k$, we are limited in how big the n_l's can be. So if generating the chains is not too computationally demanding, it makes sense to generate preliminary chains $\Phi_1^{\text{prel}}, \dots, \Phi_k^{\text{prel}}$ of very large lengths $N_1, \dots, N_k$, respectively, and use these to form a very accurate estimate $\hat{\boldsymbol{d}}$ of $\boldsymbol{d}$, thus greatly reducing V_d simultaneously for all h. Once $\hat{\boldsymbol{d}}$ is formed, $\Phi_1^{\text{prel}}, \dots, \Phi_k^{\text{prel}}$ are discarded, and the estimate $\hat{\eta}_n(\boldsymbol{a}, \hat{\boldsymbol{d}})$ is based on new and independent chains $\Phi_1, \dots, \Phi_k$. These are shorter, so that the estimate $\hat{\eta}_n(\boldsymbol{a}, \hat{\boldsymbol{d}})$ can be computed for many h's. This two-stage method is proposed in Buta and Doss (2011), who also quantify its benefits relative to the method in which we use a single stage of sampling to form both $\hat{\boldsymbol{d}}$ and $\hat{\eta}_n(\boldsymbol{a}, \hat{\boldsymbol{d}})$. Because these benefits can be quite significant, the two-stage method is the one that we use in the present paper.

The problem of estimating $\boldsymbol{d}$ may be stated as follows. We have densities $\pi_1, \dots, \pi_k$ with respect to the measure μ, which are known except for normalizing constants, i.e. we have $\pi_l = \nu_l / m_l$, where the ν_l's are known functions and the m_l's are unknown. We have samples $X_1^{(l)}, \dots, X_{N_l}^{(l)}$ from π_l, and the objective is to estimate all possible ratios m_i/m_j, $i \neq j$ or, equivalently, the vector $\boldsymbol{d} = (m_2/m_1, \dots, m_k/m_1)$. Let $N = N_1 + \cdots + N_k$, let $A_l = N_l/N$, define the vector $\boldsymbol{\zeta}$ by

$$\zeta_l = -\log(m_l) + \log(A_l), \qquad \text{for } l = 1, \dots, k,$$

and form

$$p_l(x, \boldsymbol{\zeta}) = \frac{\nu_l(x) e^{\zeta_l}}{\sum_{s=1}^k \nu_s(x) e^{\zeta_s}}, \qquad \text{for } l = 1, \dots, k. \tag{2.13}$$

Clearly, $\boldsymbol{\zeta}$ determines and is determined by the vector $(m_1, \dots, m_k)$. Geyer (1994) considered the log quasi-likelihood function

$$L_N(\boldsymbol{\zeta}) = \sum_{l=1}^k \sum_{i=1}^{N_l} \log\bigl(p_l(X_i^{(l)}, \boldsymbol{\zeta})\bigr), \tag{2.14}$$

and proposed to estimate $\boldsymbol{\zeta}$ by $\hat{\boldsymbol{\zeta}} = \operatorname{argmax} L_N(\boldsymbol{\zeta})$. Actually, there is a non-identifiability issue regarding L_N: for any constant $a \in \mathbb{R}$, $L_N(\boldsymbol{\zeta})$ and $L_N(\boldsymbol{\zeta} + a 1_k)$ are the same (here, 1_k is the vector of k 1's). So we can estimate $\boldsymbol{\zeta}$ only up to an additive constant (or equivalently, we can estimate $(m_1, \dots, m_k)$ only up to a multiplicative constant, i.e. we can estimate only $\boldsymbol{d} = (m_2/m_1, \dots, m_k/m_1)$). Accordingly, with $\boldsymbol{\zeta}_0$ defined by $[\boldsymbol{\zeta}_0]_l = \zeta_l - \bigl(\sum_{s=1}^k \zeta_s\bigr)/k$, Geyer (1994) proposed to estimate $\boldsymbol{\zeta}_0$ by the maximizer of L_N subject to the linear constraint $\boldsymbol{\zeta}^\top 1_k = 0$, and thus obtain an estimate of $\boldsymbol{d}$. In fact, this estimate of $\boldsymbol{d}$ was originally proposed by Vardi (1985). Gill et al. (1988) showed that it is consistent and asymptotically normal, and established its optimality properties, all under the assumption that for each l, $X_1^{(l)}, \dots, X_{N_l}^{(l)}$ is an iid sequence. Geyer (1994) extended the consistency and asymptotic normality result to the case where the sequences $X_1^{(l)}, \dots, X_{N_l}^{(l)}$ are Markov chains satisfying certain mixing conditions. The estimate was rederived in Meng and Wong (1996), Kong et al. (2003), Tan (2004) from completely different perspectives, all under the iid assumption.

The term $p_l(x, \boldsymbol{\zeta})$ in (2.13) has the appearance of a likelihood ratio, and in the denominator, the probability measure ν_s/m_s is given weight proportional to the length of chain Φ_s. Now Gill et al.'s (1988) optimality result does not apply to the Markov chain case. Doss and Tan (2014) argue that instead of taking $A_s = N_s/N$, we should take the A_s's to reflect the different mixing rates of the chains. They use a modified version of (2.14) and show that the resulting estimator has much better performance. They obtain a regeneration-based CLT for $\hat{\boldsymbol{d}}$ when $(A_1, \ldots, A_k)$ is an arbitrary vector of weights, and give a method for choosing this vector. Here we state their result, since we use it in the version of Theorem 1 that pertains to the case where $\boldsymbol{d}$ is unknown, and we first describe the setup. We assume that in Stage 1, for $l = 1, \ldots, k$, chain l has been run for ρ_l regenerations. So the length of the l^{th} chain, $N_l = T_1^{(l)} + \cdots + T_{\rho_l}^{(l)}$, is random. We assume that $\rho_1, \ldots, \rho_k \to \infty$ in such a way that $\rho_l/\rho_1 \to c_l \in (0, \infty)$, for $l = 1, \ldots, k$. Let $(A_1, \ldots, A_k)$ be an arbitrary (non-random) vector of weights. This vector may depend on $\rho_1, \ldots, \rho_k$, but this dependence is suppressed in the notation. We assume that as $\rho_1, \ldots, \rho_k \to \infty$, $A_l \to \alpha_l$, $l = 1, \ldots, k$ for some probability vector α with $\alpha_l > 0$ for all l. In order to state their CLT, we need to define the quantities that go into the expression for the asymptotic variance. The reader who is not interested in these details can go directly to the statement of their result given in (2.15).

The asymptotic distribution of the vector $\rho_1^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)$ involves the matrices B and Ω defined below. Let B be the $k \times k$ matrix given by

$$\begin{aligned} B_{rr} &= \sum_{l=1}^{k} \alpha_l E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)[1 - p_r(X, \boldsymbol{\zeta}_0)]\big), \qquad r = 1, \ldots, k, \\ B_{rs} &= -\sum_{l=1}^{k} \alpha_l E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0) p_s(X, \boldsymbol{\zeta}_0)\big), \qquad r, s = 1, \ldots, k,\ r \neq s. \end{aligned}$$

Recall that we use $\tau_0^{(l)} < \tau_1^{(l)} < \cdots < \tau_{\rho_l}^{(l)}$ to denote the regeneration times of the l^{th} chain, and that $T_t^{(l)} = \tau_t^{(l)} - \tau_{t-1}^{(l)}$ is the length of the t^{th} tour of the l^{th} chain. Let

$$y_i^{(r,l)} = p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big), \qquad i = 1, \ldots, N_l,$$

for which $E_{\pi_l} y_i^{(r,l)} = 0$, and define

$$Y_t^{(r,l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} y_i^{(r,l)}, \quad \bar{Y}^{(r,l)} = \frac{1}{\rho_l} \sum_{t=1}^{\rho_l} Y_t^{(r,l)}, \quad \text{and} \quad \bar{T}^{(l)} = \frac{1}{\rho_l} \sum_{t=1}^{\rho_l} T_t^{(l)}.$$

Let Ω be the $k \times k$ matrix defined by

$$\Omega_{rs} = \sum_{l=1}^{k} \frac{\alpha_l^2}{c_l} \frac{E\big(Y_1^{(r,l)} Y_1^{(s,l)}\big)}{\big(E T_1^{(l)}\big)^2}, \qquad r, s = 1, \ldots, k,$$

To obtain an estimate $\widehat{\Omega}$, we let

$$Z_t^{(r,l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \quad \text{and} \quad \hat{\mu}_r^{(l)} = \frac{\sum_{i=1}^{N_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}})}{N_l},$$

and define $\widehat{\Omega}$ by

$$\widehat{\Omega}_{rs} = \sum_{l=1}^{k} \frac{A_l^2}{c_l} \frac{1}{\big(\bar{T}^{(l)}\big)^2} \frac{1}{\rho_l} \sum_{t=1}^{\rho_l} \big(Z_t^{(r,l)} - \hat{\mu}_r^{(l)} T_t^{(l)}\big)\big(Z_t^{(s,l)} - \hat{\mu}_r^{(l)} T_t^{(l)}\big), \qquad r, s = 1, \ldots, k.$$

The function $g\colon \mathbb{R}^k \to \mathbb{R}^{k-1}$ that maps $\boldsymbol{\zeta}$ into $\boldsymbol{d}$ and the gradient of this function (in terms of d) are given by

$$g(\boldsymbol{\zeta}) = \begin{pmatrix} e^{\zeta_1-\zeta_2} A_2/A_1 \\ e^{\zeta_1-\zeta_3} A_3/A_1 \\ \vdots \\ e^{\zeta_1-\zeta_k} A_k/A_1 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} d_2 & d_3 & \ldots & d_k \\ -d_2 & 0 & \ldots & 0 \\ 0 & -d_3 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & -d_k \end{pmatrix}.$$

Note that $g(\boldsymbol{\zeta}_0) = \boldsymbol{d}$ and $g(\hat{\boldsymbol{\zeta}}) = \hat{\boldsymbol{d}}$.

For a matrix C, $C^\dagger$ will denote the Moore-Penrose inverse of C. Doss and Tan (2014) show that if for the Stage 1 chains $\Phi_1^{\text{prel}}, \ldots, \Phi_k^{\text{prel}}$ conditions 1 and 2 of Theorem 1 hold, then as $\rho_1 \to \infty$, $\hat{\boldsymbol{d}} \xrightarrow{\text{a.s.}} \boldsymbol{d}$ and

$$\rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0, W) \quad \text{where} \quad W = D^\top B^\dagger \Omega B^\dagger D. \tag{2.15}$$

They show that furthermore, with $\widehat{B}$ and $\widehat{D}$ being the obvious empirical estimates of B and D, respectively,

$$\widehat{W} := \widehat{D}^\top \widehat{B}^\dagger \widehat{\Omega} \widehat{B}^\dagger \widehat{D} \tag{2.16}$$

is a strongly consistent estimate of W.

We now review the big picture (temporarily reverting to the more cumbersome notation): There is a parametric family $\{\pi_h,\ h \in \mathcal{H}\}$, where $\pi_h = \nu_h/m_h$, and we wish to estimate m_h/m_{h_1} for all $h \in \mathcal{H}$. We select "skeleton points" $h_1, \ldots, h_k \in \mathcal{H}$. The Stage 1 chains $\Phi_1^{\text{prel}}, \ldots, \Phi_k^{\text{prel}}$ are used to form an estimate $\hat{\boldsymbol{d}}$ of $\boldsymbol{d} = (m_{h_2}/m_{h_1}, \ldots, m_{h_k}/m_{h_1})$. Stage 2 chains use $\hat{\boldsymbol{d}}$ to form an estimate of m_h/m_{h_1} (and also of $E_{\pi_h} f$) for h not in the skeleton set, and the entire process makes it unnecessary to run a separate Markov chain for each value of $h \in \mathcal{H}$.

In the previous section we proposed a regenerative method for analyzing the variances of $\hat{u}_n(\boldsymbol{a}, \boldsymbol{d})$ and $\hat{\eta}_n(\boldsymbol{a}, \boldsymbol{d})$. When we use $\hat{\boldsymbol{d}}$ instead of $\boldsymbol{d}$ in (1.5), two problems arise. First, the triples $\big(V_t^{(l)}, U_t^{(l)}, T_t^{(l)}\big)$ are no longer independent: because we are using the same $\hat{\boldsymbol{d}}$ throughout, there is dependence across different t's and also across different l's. Second, as mentioned earlier, using $\hat{\boldsymbol{d}}$ instead of $\boldsymbol{d}$ inflates the variance of both estimators.

Choice of the Vector $\boldsymbol{a}$ In the rest of this section we propose a method for dealing with these problems. Recall the estimate $\hat{u}_n = \sum_{l=1}^{k}(a_l/n_l)\sum_{i=1}^{n_l} u(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d})$ where

$$u(x; \boldsymbol{a}, \boldsymbol{d}) = \frac{\nu_h(x)}{\sum_{s=1}^{k} a_s \nu_{h_s}(x)/d_s}. \tag{2.17}$$

This is the estimate of the Bayes factor m_h/m_{h_1} in the Bayesian framework in which $\nu_h(x) = \ell(x)p_h(x)$, and m_h is the marginal likelihood of the data when the prior is p_h. In (2.17) $\boldsymbol{a}$ is a vector with non-negative entries, and in principle we can choose any $\boldsymbol{a}$ that we want. We now discuss the choice of this vector. In forming the log quasi-likelihood function (2.14) which is based on sequences of lengths $N_1, \dots, N_k$ and involves (2.13), all authors mentioned above use $A_s = N_s/N, \ s = 1, \dots, k$, i.e. the weight given ν_{h_s}/d_{h_s} is proportional to the length of the sequence from π_{h_s}, as this choice is generally deemed optimal in some sense (Meng and Wong (1996) establish this optimality in the iid setting for the case $k = 2$). It should be noted that the optimality of the choice $A_s = N_s/N$ pertains to the problem of estimating $\boldsymbol{d}$. In our Stage 2, the problem is different: our goal is to estimate m_h/m_{h_1} for $h \in \mathcal{H}$, and the optimal value of $\boldsymbol{a}$ depends on h, as we now make clear. Intuitively speaking, to minimize the variance of the estimate of m_h/m_{h_1} using $\hat{u}_n$, we should give more weight to probability measures ν_{h_s}/m_{h_s} which are close to ν_h/m_h. In the extreme case where for some j, $\nu_{h_j}/m_{h_j} = \nu_h/m_h$, if we give weight 1 to ν_{h_j}/m_{h_j} and weight 0 to ν_{h_l}/m_{h_l} for $l \neq j$, then $\hat{u}_n$ reduces to

$$\hat{u}_n = \frac{1}{n_j}\sum_{i=1}^{n_j} \frac{\nu_h(X_i^{(j)})}{\nu_{h_j}(X_i^{(j)})/d_j},$$

and this estimate has zero variance, each summand being exactly equal to m_h/m_{h_1}. Now let $h_{\text{opt}} = \text{argmax}_h\, m_h/m_{h_1}$, and note that since h_1 is fixed, h_{opt} is also equal to $\text{argmax}_h\, m_h$, the empirical Bayes choice of h.

Consider the problem of estimating m_h/m_{h_1}, $h \in \mathcal{H}$. After having identified values of h for which m_h is very small, thus eliminating these h's from consideration, we will be especially interested in estimating m_h/m_{h_1} for h near h_{opt}, so that we can accurately identify h_{opt}. To see what choice of $\boldsymbol{a}$ this leads to, consider the measure λ given by $\lambda(dx) = \big(\ell(x)/p_{h_{\text{opt}}}(x)\big)\,\mu(dx)$, and consider the Hilbert space $L^2(\lambda)$ of functions which are square-integrable with respect to λ, with inner product $\langle f_1, f_2\rangle = \int f_1 f_2\, d\lambda$. Note that $p_{h_{\text{opt}}} \in L^2(\lambda)$, with $\|p_{h_{\text{opt}}}\|^2 = \int p_{h_{\text{opt}}}\ell\, d\mu = m_{h_{\text{opt}}}$. Also, for any h such that $p_h \in L^2(\lambda)$, we have $\langle p_h, p_{h_{\text{opt}}}\rangle = m_h$. So the heuristic that we should give more weight to the probability measure ν_{h_j}/m_{h_j} if p_{h_j} is close to $p_{h_{\text{opt}}}$ suggests that we set a_j equal to m_{h_j}, or equivalently, set $(a_1, \dots, a_k) = (1, d_2, \dots, d_k) = (1, \boldsymbol{d})$. (For convenience, we set $d_1 = 1$ and $\hat{d}_1 = 1$.) In some experiments we have done, this choice of $\boldsymbol{a}$ outperforms the more conventional choice $a_j = n_j/n$ (as in, e.g., Tan (2004)) for the problem of estimating m_h/m_{h_1} for h near h_{opt}. This is the choice of $\boldsymbol{a}$ we make when $\boldsymbol{d}$ is known (cf. Section 2.1). When $\boldsymbol{d}$ is unknown, we set $\boldsymbol{a} = (\hat{d}_1, \hat{\boldsymbol{d}}) = (1, \hat{\boldsymbol{d}})$ after noting that there is nothing that requires $\boldsymbol{a}$ to be a constant. With this choice, the expressions for u and v in (1.4) become

$$u\big(x; (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) = \frac{\nu(x)}{\sum_{l=1}^k \nu_l(x)} \quad \text{and} \quad v\big(x; (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) = \frac{f(x)\nu(x)}{\sum_{l=1}^k \nu_l(x)}. \tag{2.18}$$

These do not involve $\hat{\boldsymbol{d}}$, and consequently for each l, the triples $\big(V_t^{(l)}, U_t^{(l)}, T_t^{(l)}\big)$, $t = 0, 1, 2, \ldots$ defined in (2.4) are again iid, and we have independence across l's. The estimator for η reduces to

$$\begin{aligned} \hat{\eta} = \hat{\eta}_{N,n}\big((1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) &= \sum_{l=1}^k \frac{\hat{d}_l}{n_l} \sum_{i=1}^{n_l} \frac{f(X_i^{(l)})\nu(X_i^{(l)})}{\sum_{s=1}^k \nu_s(X_i^{(l)})} \Big/ \sum_{l=1}^k \frac{\hat{d}_l}{n_l} \sum_{i=1}^{n_l} \frac{\nu(X_i^{(l)})}{\sum_{s=1}^k \nu_s(X_i^{(l)})} \\ &= \sum_{l=1}^k \frac{\hat{d}_l}{n_l} \sum_{t=1}^{R_l} V_t^{(l)} \Big/ \sum_{l=1}^k \frac{\hat{d}_l}{n_l} \sum_{t=1}^{R_l} U_t^{(l)} \\ &= \sum_{l=1}^k \hat{d}_l \frac{\bar{V}^{(l)}}{\bar{T}^{(l)}} \Big/ \sum_{l=1}^k \hat{d}_l \frac{\bar{U}^{(l)}}{\bar{T}^{(l)}}, \end{aligned} \tag{2.19}$$

where

$$U_t^{(l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} \frac{\nu(X_i^{(l)})}{\sum_{s=1}^k \nu_s(X_i^{(l)})} \quad \text{and} \quad V_t^{(l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} \frac{f(X_i^{(l)})\nu(X_i^{(l)})}{\sum_{s=1}^k \nu_s(X_i^{(l)})}.$$

Similarly, the estimator for m/m_1 reduces to

$$\hat{u} = \hat{u}_{N,n}\big((1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) = \sum_{l=1}^k \frac{\hat{d}_l}{n_l} \sum_{i=1}^{n_l} \frac{\nu(X_i^{(l)})}{\sum_{s=1}^k \nu_s(X_i^{(l)})} = \sum_{l=1}^k \frac{\hat{d}_l}{n_l} \sum_{t=1}^{R_l} U_t^{(l)} = \sum_{l=1}^k \hat{d}_l \frac{\bar{U}^{(l)}}{\bar{T}^{(l)}}. \tag{2.20}$$

We will often write simply $\hat{\eta}$ and $\hat{u}$ instead of the more involved $\hat{\eta}_{N,n}\big((1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big)$ and $\hat{u}_{N,n}\big((1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big)$ (cf. (2.19) and (2.20)) in order to avoid unnecessarily cumbersome notation, if this will not cause confusion.

Theorem 2 below gives the asymptotic distributions of $\hat{\eta}$ and $\hat{u}$. To state it, we first need to define the expressions that go into the asymptotic variance. Let M and H be the vectors of length $k - 1$ for which the $(j-1)^{\text{th}}$ coordinates are

$$M_{j-1} = E_{\pi_j} u \quad \text{and} \quad H_{j-1} = \frac{E_{\pi_j} v}{\sum_{l=1}^k d_l E_{\pi_l} u} - \frac{\big(\sum_{l=1}^k d_l E_{\pi_l} v\big)\big(E_{\pi_j} u\big)}{\big(\sum_{l=1}^k d_l E_{\pi_l} u\big)^2}, \qquad j = 2, \ldots, k. \tag{2.21}$$

Recall that N is the total sample size used in Stage 1 sampling and R_1 is the number of regenerations for chain 1 in Stage 2 sampling.

Theorem 2 *Suppose that for the Stage* 1 *chains, conditions* 1 *and* 2 *of Theorem 1 hold, and that for the Stage* 2 *chains, conditions* 1–4 *of Theorem 1 hold, where condition* 3 *refers to the functions* u *and* v *given in* (2.18)*. If* $\rho_1 \to \infty$ *and* $R_1 \to \infty$ *in such a way that* $R_1/\rho_1 \to q \in [0, \infty)$*, then*

$$R_1^{1/2}\big(\hat{u} - m/m_1\big) \xrightarrow{d} \mathcal{N}\big(0, qM^\top W M + \kappa^2\big)$$

and

$$R_1^{1/2}\big(\hat{\eta}-\eta\big) \xrightarrow{d} \mathcal{N}\big(0, qH^\top W H + \tau^2\big),$$

where M, H, W, κ^2 and τ^2 are given in equations (2.21), (2.15), (2.9), *and* (2.11), *respectively. In* (2.9) *and* (2.11), $\boldsymbol{a}$ *is taken to be* $\boldsymbol{a} = (1, \boldsymbol{d})$. *Furthermore, we can form strongly consistent estimates of the asymptotic variances if we use* $\widehat{W}$, $\hat{\kappa}^2$, *and* $\hat{\tau}^2$ *defined in* (2.16) *and* (2.12), *respectively, and use the obvious empirical estimates of M and H.*

The proof of the theorem is given in the Appendix.

2.3 Advantages and Limitations of the Regeneration-Based Approach

There are several important benefits to using regeneration, the most obvious one being the existence of the strongly consistent estimators of the asymptotic variances in Theorems 1 and 2. Another advantage is that because (2.2) in effect states that we start each chain at a regeneration point, we completely bypass all problems with burn-in.

There are also significant potential computational advantages in using regeneration to estimate standard errors. Recall that we have in mind a parametric family $\{\pi_h,\ h \in \mathcal{H}\}$, and we are interested in a sensitivity analysis problem in which we wish to estimate $\eta_h = E_{\pi_h} f$ for all $h \in \mathcal{H}$. To keep this discussion as simple as possible, suppose we are in the case where $k = 1$, and we are using the estimate (1.2), based on a Markov chain $X_1, X_2, \ldots$ with stationary distribution π_{h_1}. Here, $u_h(x) = \nu_h/\nu_{h_1}$, $v_h(x) = f(x)u_h(x)$, and $\widetilde{\eta}_{h,n} = \bar{v}_{h,n}/\bar{u}_{h,n}$. If w_h is equal to either v_h or u_h, under our regularity conditions, the asymptotic variance of $\bar{w}_{h,n}$ is

$$\operatorname{Var}\big(w_h(X_1)\big) + 2\textstyle\sum_{j=1}^{\infty} \operatorname{Cov}\big(w_h(X_1), w_h(X_{1+j})\big), \tag{2.22}$$

where the variance and covariances are calculated under the assumption that $X_1 \sim \pi_{h_1}$. Now whereas the calculation of $\bar{w}_{h,n}$ requires $O(n)$ operations, estimates of the asymptotic variance of $\bar{w}_{h,n}$ based on conventional spectral methods involve estimating the first M_n terms of the series (2.22), where M_n is of the order of n^α, and $\alpha > 0$ depends on the method used. Since the estimate of each term requires $O(n)$ operations, for each h, $O(n^{1+\alpha})$ operations are required, and the additional computational burden is problematic because we are dealing with a large number of h's. In contrast, when we use regeneration, once we have identified the sequence of regeneration times, the estimates (2.12) use *only* quantities that have already been computed in the process of forming $\widetilde{\eta}_{h,n}$, so estimation of standard errors comes at a trivial additional computational cost.

Additionally, the regeneration method gives a potential approach for obtaining very difficult distributional results. Consider the parameter m_h/m_{h_1} and its estimator $\hat{u}_n$, which we will denote temporarily by

$B(h)$ and $\hat{B}_n(h)$, respectively, in order to emphasize the dependence of these quantities on h. It is of interest to provide a confidence band (region, if h is multidimensional) for $B(h)$ that is valid simultaneously for all $h \in \mathcal{H}$. A closely related problem is to produce a confidence interval for $\operatorname{argmax}_{h \in \mathcal{H}} B(h)$. The traditional way of forming confidence bands that are valid globally is to proceed as follows:

1 Establish a functional central limit theorem that says that $n^{1/2}\big(\hat{B}_n(h) - B(h)\big)$ converges in distribution to a Gaussian process $W(h);\ h \in \mathcal{H}$.

2 Find the distribution of $\sup_{h \in \mathcal{H}} |W(h)|$.

If s_α is the $(1-\alpha)$-quantile of the distribution of this supremum, then the band $\hat{B}_n(h) \pm s_\alpha / n^{1/2}$ has asymptotic coverage probability equal to $1-\alpha$. The value s_α is typically too difficult to compute analytically, but can be obtained by simulation [see, e.g. Burr and Doss (1993) among many others]. Establishing functional central limit theorems at this level of generality can be done only using empirical process theory, which requires an iid structure. For this reason we believe that the regeneration method offers the best hope for establishing such theorems.

We now discuss the scope of the problems to which regenerative simulation can be applied. Mykland et al. (1995, Section 4.1) showed that for independence Metropolis-Hastings chains, there is always a scheme for producing regeneration sequences, and Gilks et al. (1998) show that this is also true for random walk Metropolis-Hastings chains. A general approach for producing a minorization condition is the so-called "distinguished point technique" of Mykland et al. (1995, Section 4.1); a short description is given near the end of the Appendix to the present paper, where we implement it for the Markov chain used in the Bayesian variable selection problem of Section 4.1. The technique requires that we find a point $x \in \mathsf{X}$ and a set $D \subset \mathsf{X}$; these are in effect tuning parameters, whose choice inevitably requires preliminary trial and error experimentation. It is widely acknowledged that the distinguished point technique is very difficult to implement successfully in high dimensions. Gilks et al. (1998, Theorem 4.1) show that in a certain class of problems, the regeneration rates for random walk Metropolis-Hastings chains go down exponentially with the dimension. To conclude, there is no all-purpose method for devising minorization conditions that will give regeneration rates that are large enough to produce useful results, and to a large extent, successful implementation of regenerative simulation is a matter of art.

3 Single-Chain Importance Sampling in Bayesian Analysis

Importance sampling can be very useful in Bayesian analysis when we wish to see the effect of a change in the prior distribution or the likelihood function on a posterior expectation. Here we focus on the application of single-chain importance sampling estimators in such a situation.

Let x and y denote parameters and observed data, respectively. We think of π and π_1 as two different posterior densities with $\pi(x) = \ell(x)p(x)/m$, and $\pi_1(x) = \ell_1(x)p_1(x)/m_1$, where $\ell(x)$ and $\ell_1(x)$ denote two different likelihood functions, $p(x)$ and $p_1(x)$ denote two different prior densities, and m and m_1 are unknown normalizing constants. (When the dependence of the likelihood function on the data needs to be noted we will write $\ell(x; y)$ instead of $\ell(x)$.) We imagine a situation where an MCMC algorithm for π is yet to be developed, but we have available a Markov chain with invariant distribution π_1, say $\Phi = \{X_i,\, i = 0, 1, \ldots\}$, which satisfies the assumptions of Corollary 1. Suppose we are interested in approximating the expectation $E_\pi f = \int_{\mathsf{X}} f(x)\pi(x)\,\mu(dx)$, where $f\colon \mathsf{X} \to \mathbb{R}$ is a π-integrable function. According to equation (1.2), the ratio estimator of η based on the Markov chain Φ from π_1 is given by

$$\widetilde{\eta}_n = \sum_{i=1}^{n} \frac{f(X_i)\ell(X_i)p(X_i)}{\ell_1(X_i)p_1(X_i)} \Big/ \sum_{i=1}^{n} \frac{\ell(X_i)p(X_i)}{\ell_1(X_i)p_1(X_i)}.$$

If the two Bayesian models differ only in the prior, i.e., if $\ell = \ell_1$, then

$$\widetilde{\eta}_n = \sum_{i=1}^{n} \frac{f(X_i)p(X_i)}{p_1(X_i)} \Big/ \sum_{i=1}^{n} \frac{p(X_i)}{p_1(X_i)},$$

and the cancellation of the potentially very complicated likelihood gives a convenient simplification. According to Corollary 1, if there exists $\epsilon > 0$ such that

$$E_{\pi_1}|\ell p/\ell_1 p_1|^{2+\epsilon} < \infty \quad \text{and} \quad E_{\pi_1}|f\ell p/\ell_1 p_1|^{2+\epsilon} < \infty, \tag{3.1}$$

then we can estimate $E_\pi f$ with $\widetilde{\eta}_n$, and a valid asymptotic standard error for this estimate is given by $\hat{\underline{\tau}}/R^{1/2}$.

Remark 2 *If there exists an $\epsilon > 0$ such that $E_\pi|f|^{2+\epsilon} < \infty$, then a sufficient condition for* (3.1) *is the existence of a constant $M \in [1, \infty)$ such that*

$$\sup_x \frac{\ell(x)p(x)}{\ell_1(x)p_1(x)} < M. \tag{3.2}$$

This condition basically says that the tails of π_1 are heavier than those of π. In fact, (3.2) *is exactly what is required to use π_1 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_1. For a thorough review of accept/reject methods, see Robert and Casella (2004, Chap. 2).*

On the other hand, there are some situations where we may want to change the likelihood function. An example arises in binary regression, where one can use either the probit or the logistic model. Under the probit model there exists a very convenient data augmentation algorithm (Albert and Chib, 1993) that gives rise to a chain for which geometric ergodicity and minorization conditions have been established (Roy and

Hobert, 2007). But in certain biostatistical applications one strongly prefers the logistic model because the parameters are then equal to log odds ratios, and these have a nice interpretation when dealing with case-control studies. Unfortunately, chains that implement logistic regression models are very difficult to analyze, and to the best of our knowledge there do not exist any results regarding rates of convergence for chains that implement these models. Informal calculations show that when ℓ_1 is the probit likelihood and p_1 is a flat prior (the setup analyzed by Roy and Hobert (2007)), and when ℓ is the logistic likelihood and p is a normal prior, then the key condition $E_{\pi_1}|\ell p/\ell_1 p_1|^{2+\epsilon} < \infty$ is satisfied, so that a geometrically ergodic chain run to implement the probit model can be used to do an honest analysis of the logistic model. The next section gives a concrete example of the comparison of posteriors corresponding to two different priors using the importance sampling idea described above.

3.1 Bayesian Analysis of the One-Way Random Effects Model

Consider the classical balanced one-way random effects model given by

$$Y_{ij} = \theta_i + \varepsilon_{ij} \qquad i = 1, \dots, q, \; j = 1, \dots, m,$$

where the random effects $\theta_1, \dots, \theta_q$ are iid $\mathcal{N}(\mu, \sigma_\theta^2)$, the ε_{ij}'s are iid $\mathcal{N}(0, \sigma_e^2)$ and independent of the θ_i's. To avoid trivial special cases, assume that $q \geq 2$ and $m \geq 2$. A Bayesian version of the model requires a prior distribution for $(\mu, \sigma_\theta^2, \sigma_e^2)$, call it $p(\mu, \sigma_\theta^2, \sigma_e^2)$. Actually, since the vector of random effects, $\theta = (\theta_1, \dots, \theta_q)$, is unobserved, it too is viewed as a parameter with a "built-in" prior density given by

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

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

$$\pi(\theta, \mu, \sigma_\theta^2, \sigma_e^2) \propto \ell(\theta, \mu, \sigma_\theta^2, \sigma_e^2; y) \, \bigl[\phi(\theta \,|\, \mu, \sigma_\theta^2, \sigma_e^2) \, p(\mu, \sigma_\theta^2, \sigma_e^2)\bigr], \tag{3.3}$$

where

$$\ell(\theta, \mu, \sigma_\theta^2, \sigma_e^2; y) = \prod_{i=1}^{q} \prod_{j=1}^{m} (2\pi\sigma_e^2)^{-1/2} \exp\Bigl\{-\frac{1}{2\sigma_e^2}(y_{ij} - \theta_i)^2\Bigr\}.$$

We consider two (improper) priors for $(\mu, \sigma_\theta^2, \sigma_e^2)$. The first is the widely-used "standard diffuse prior" given by

$$p_s(\mu, \sigma_\theta^2, \sigma_e^2) \propto \bigl(\sigma_e^2 (\sigma_\theta^2)^{1/2}\bigr)^{-1}.$$

Serious Bayesians disagree about the suitability of this prior. Indeed, Gelman (2006) endorses it, but 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. This prior takes the form

$$p_r(\mu, \sigma_\theta^2, \sigma_e^2) \propto (\sigma_\theta^2)^{-C_m/2} (\sigma_e^2)^{-1} \Big[m - 1 + \big(\sigma_e^2 / (\sigma_e^2 + m\sigma_\theta^2) \big)^2 \Big]^{1/2},$$

where $C_m = 1 - \sqrt{m-1} \big/ (\sqrt{m} + \sqrt{m-1}\,)^3$. Let $\pi_r(\theta, \mu, \sigma_\theta^2, \sigma_e^2)$ denote the posterior density under p_r. A necessary and sufficient condition for propriety of the posterior under either p_s or p_r is $q \geq 3$ (see Proposition 1 below). In the remainder of this subsection, we explain how to use a well studied Gibbs sampler for $\pi_s(\theta, \mu, \sigma_\theta^2, \sigma_e^2)$ to approximate intractable posterior expectations under $\pi_r(\theta, \mu, \sigma_\theta^2, \sigma_e^2)$, which is the more complex of the two posterior densities.

The standard diffuse prior, $p_s(\mu, \sigma_\theta^2, \sigma_e^2)$, is a conditionally conjugate prior; 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, $\pi_s(\sigma_\theta^2 \,|\, \sigma_e^2, \mu, \theta)$, is an inverse-gamma density. In fact, straightforward manipulation of (3.3) shows that $\pi_s(\mu, \theta \,|\, \sigma_\theta^2, \sigma_e^2)$ is a multivariate normal density, and that $\pi_s(\sigma_\theta^2, \sigma_e^2 \,|\, \mu, \theta)$ factors into a product of two inverse-gamma densities. Thus, it is easy to simulate a *block* Gibbs sampler for the posterior π_s that alternates between drawing $(\sigma_\theta^2, \sigma_e^2)$ and (μ, θ), conditioning on the most recent value of the other. The precise forms of $\pi_s(\mu, \theta \,|\, \sigma_\theta^2, \sigma_e^2)$ and $\pi_s(\sigma_\theta^2, \sigma_e^2 \,|\, \mu, \theta)$ are provided by Tan and Hobert (2009), who proved that the Markov chain underlying the block Gibbs algorithm is geometrically ergodic. These authors also developed a minorization condition of the form (2.1). In other words, Tan and Hobert (2009) showed that the block Gibbs sampler for π_s satisfies the conditions of Corollary 1.

Now consider what the development and analysis of an MCMC algorithm for π_r would entail. Since the reference prior, $p_r(\mu, \sigma_\theta^2, \sigma_e^2)$, is *not* conditionally conjugate, the Gibbs sampler for π_r would not be as straightforward to simulate as that for π_s. In particular, the full conditional densities for the variance components have non-standard forms, so sampling from these densities would require some type of rejection sampling. Another possibility would be to replace the Gibbs updates of the variance components with Metropolis-Hastings moves. No matter how we choose to deal with the variance components, one thing is clear: the MCMC algorithm for π_r will be more difficult to implement and more challenging to analyze than would be the block Gibbs sampler for π_s.

For researchers who simply want reliable estimators of posterior expectations with respect to π_r, and are not interested in conquering the theoretical difficulties associated with the reference prior, this is an ideal situation in which to apply the importance sampling methods described in the previous section. In particular, the block Gibbs sampler for π_s can be used to construct estimators (and valid asymptotic standard errors) for intractable posterior expectations of the form $E_{\pi_r} f(\mu, \sigma_\theta^2, \sigma_e^2)$, as long as the moment conditions in (3.1) are satisfied. The following result is useful for checking the moment conditions and will be applied in the next subsection.

Proposition 1 *A necessary and sufficient condition for propriety of the posterior under either p_s or p_r is $q \geq 3$. Moreover, if $q \geq 3$ and $s, t \in \mathbb{R}$, then $E_{\pi_s}\left[(\sigma_\theta^2)^s(\sigma_e^2)^t\right] < \infty$ if and only if the following two conditions are satisfied:*

1. $-1/2 < s < q/2 - 1$,

2. $s + t < mq/2 - 1$.

The proof of the proposition is in the Appendix.

3.2 A Numerical Example: Styrene Exposure Data

Here we use the Bayesian model from the previous subsection to analyze a real data set from Lyles et al. (1997). Thirteen workers were randomly selected from a group within a boat manufacturing plant and each worker's styrene exposure was measured on three separate occasions. The data set was previously analyzed in Tan and Hobert (2009), which reproduces the data and also gives some summary statistics. The two posterior distributions for $(\theta, \mu, \sigma_\theta^2, \sigma_e^2)$ under consideration are π_s, which is based on the standard diffuse prior, and π_r, which is based on the reference prior. Both posteriors are proper since $q = 13$. We will focus on the posterior expectations of three functions: $f_1(\mu, \sigma_\theta^2, \sigma_e^2) = \sigma_\theta^2$, $f_2(\mu, \sigma_\theta^2, \sigma_e^2) = \sigma_e^2$, and $f_3(\mu, \sigma_\theta^2, \sigma_e^2) = \sigma_\theta^2/(\sigma_\theta^2 + \sigma_e^2)$. Note that f_3 is the correlation between observations on the same worker.

Tan and Hobert (2009) used their minorization condition to simulate $R = 40{,}000$ regenerations of the block Gibbs sampler for π_s. This simulation was used to produce estimates and standard errors for $E_{\pi_s} f_i(\mu, \sigma_\theta^2, \sigma_e^2)$, $i = 1, 2, 3$, and their results are summarized in Table 1. Figure 1 gives trace plots that justify the choice $R = 40{,}000$. Now consider re-using Tan and Hobert's (2009) block Gibbs output to produce estimates and standard errors for $E_{\pi_r} f_i(\mu, \sigma_\theta^2, \sigma_e^2)$, $i = 1, 2, 3$. According to Corollary 1, our importance sampling results are applicable if we can find $\epsilon > 0$ such that $E_{\pi_s}|f_i\, p_r/p_s|^{2+\epsilon} < \infty$ for $i = 0, 1, 2, 3$, where $f_0 \equiv 1$. First, note that $|f_i p_r/p_s| = f_i\, p_r/p_s$, since all the terms are positive. Now

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

where $C_3 \doteq 0.96$. This inequality together with Proposition 1 shows that $E_{\pi_s}|f_i\, p_r/p_s|^3 < \infty$ for $i = 0, 1, 2, 3$. Hence, our importance sampling technique is applicable, and the results are given in Table 1. Again, Figure 1 justifies the use of $R = 40{,}000$. From Table 1 we see that the estimates 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.

	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	0.18625	0.03411	0.00092	$(0.18440, 0.18813)$
σ_e^2	Diffuse	0.61849	0.00966	0.00049	$(0.61751, 0.61947)$
	Reference	0.62134	0.00937	0.00048	$(0.62037, 0.62230)$
$\sigma_\theta^2/(\sigma_\theta^2 + \sigma_e^2)$	Diffuse	0.21304	0.03687	0.00096	$(0.21112, 0.21496)$
	Reference	0.20881	0.03571	0.00094	$(0.20692, 0.21070)$

Table 1: 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. Results are based on $R = 40{,}000$ regenerations.

4 Multiple-Chain Importance Sampling in Bayesian Posterior Analysis

Here we consider a standard Bayesian setup in which we have a parametric family of prior densities $\{p_h,\ h \in \mathcal{H}\}$ on the parameter θ, we observe a data vector y, whose distribution is given by the likelihood function $\ell(\theta; y)$, and we have posterior densities given by $\pi_h(\theta) = \ell(\theta; y)p_h(\theta)/m_h$. We are interested in estimating certain features of π_h for all $h \in \mathcal{H}$. As mentioned earlier, the single-chain importance sampling method does not work well here, and we use multiple-chain importance sampling instead. This section illustrates our methodology on a model for variable selection in Bayesian linear regression. Section 4.1 presents the model, reviews an MCMC algorithm for making inference using the model and describes its regenerative features. Section 4.2 gives an illustration on a data set.

In Section 1 we stated that our goals are to estimate $\eta_h = E_{\pi_h} f$ and m_h for all $h \in \mathcal{H}$, and before proceeding we discuss the problem of estimating m_h and the objective in doing so. The quantity $m_h := m_h(y) = \int \ell(\theta; y)p_h(\theta)\, d\theta$ is the marginal likelihood of the data y when the prior is p_h, and may be viewed as a measure of compatibility of the prior with the data y: priors for which m_h is very small are deemed implausible, and $h_{\text{opt}} = \text{argmax}_h\, m_h$ is the empirical Bayes choice of h. It turns out that estimating m_h itself is computationally very difficult—for example the harmonic mean estimator (Newton and Raftery, 1994) almost always has infinite variance—but for fixed $h_1 \in \mathcal{H}$, estimating the ratio m_h/m_{h_1} is computationally far easier, and statistically equivalent: the information regarding h in the function m_h is the same as the information regarding h in the function m_h/m_{h_1} (these two functions have the same shape, and in particular, $\text{argmax}_h\, m_h = \text{argmax}_h(m_h/m_{h_1})$). For this reason, we consider only estimation of m_h/m_{h_1}.

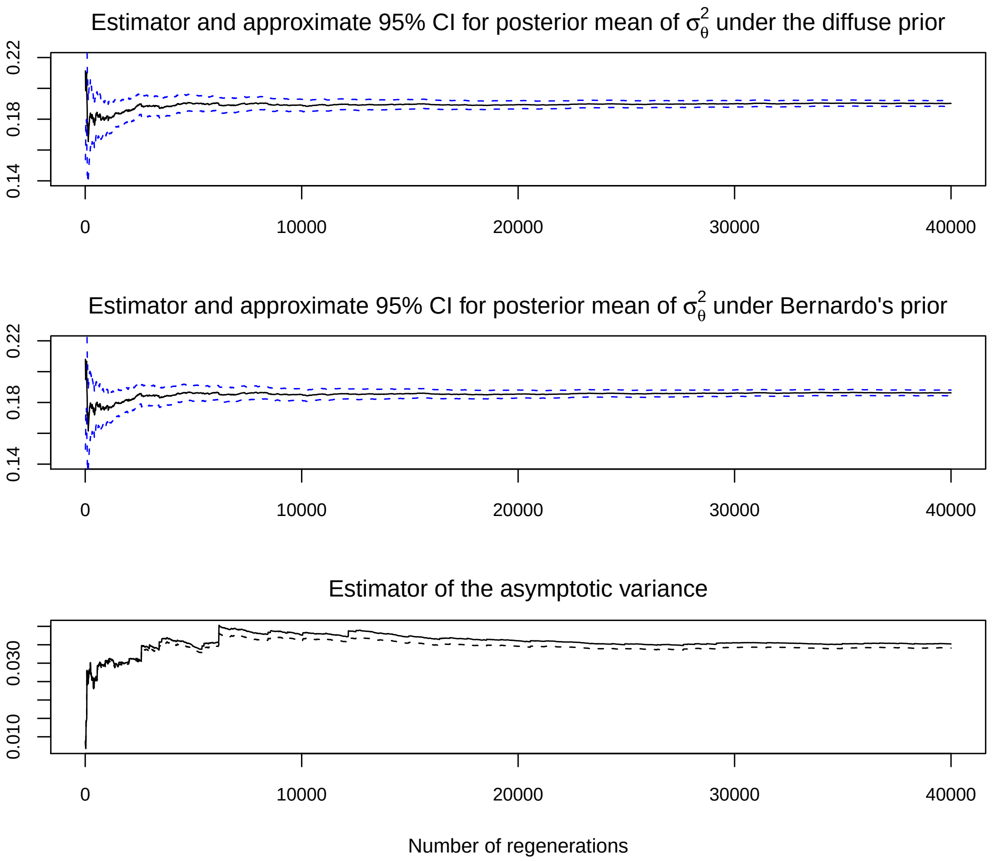

Figure 1: Top plot shows evolution of the estimator of $E_{\pi_s}\sigma_\theta^2$ and the corresponding 95% asymptotic CI as the number of regenerations grows. Solid line represents the estimator and dashed lines denote the upper and lower endpoints of the CI. Middle plot does the same for $E_{\pi_r}\sigma_\theta^2$. Bottom plot displays evolution of the estimate of the asymptotic variance, $\hat{\tau}^2$, of the estimators of $E_{\pi_s}\sigma_\theta^2$ (solid line) and $E_{\pi_r}\sigma_\theta^2$ (dashed line).

4.1 Illustration on a Model for Variable Selection in Bayesian Linear Regression

The most commonly used setup for variable selection in Bayesian linear regression is described as follows. We have a response vector $Y = (Y_1, \dots, Y_m)^\top$ and a set of potential predictors $X_1, \dots, X_q$, each a vector of length m. Every subset of predictors is identified with a binary vector $\gamma = (\gamma_1, \dots, \gamma_q)^\top \in \{0,1\}^q$, where $\gamma_j = 1$ if X_j is included in the model and $\gamma_j = 0$ otherwise. For every γ, we have a model given by

$$Y = 1_m\beta_0 + X_\gamma\beta_\gamma + \epsilon,$$

where 1_m is the vector of m 1's, X_γ is the design matrix whose columns consist of the predictor vectors corresponding to γ, β_γ is the vector of coefficients for that subset, and $\epsilon \sim \mathcal{N}_m(0, \sigma^2 I)$. For this setup, the unknown parameter is $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$, which includes the indicator of the subset of variables that go into the regression model. The prior on θ is a hierarchy in which we first select the variables that go into the regression model, then a "non-informative prior" is given to (σ^2, β_0), and given γ and σ, we choose β_γ from some proper distribution. The specific instance of this model that we will consider is indexed by two hyperparameters, $w \in (0, 1)$ and $g > 0$, and is given in detail as follows:

$$\text{given } \gamma, \sigma, \beta_0, \beta_\gamma, \qquad Y \sim \mathcal{N}_m(1_m\beta_0 + X_\gamma\beta_\gamma, \sigma^2 I), \tag{4.1a}$$

$$\text{given } \gamma, \sigma, \qquad \beta_\gamma \sim \mathcal{N}_{q_\gamma}\big(0, g\sigma^2(X_\gamma^\top X_\gamma)^{-1}\big), \tag{4.1b}$$

$$(\sigma^2, \beta_0) \sim p(\beta_0, \sigma^2) \propto 1/\sigma^2, \tag{4.1c}$$

$$\gamma \sim p(\gamma) = w^{q_\gamma}(1 - w)^{q - q_\gamma}. \tag{4.1d}$$

The prior on γ given by (4.1d) is the so-called independence Bernoulli prior, in which every variable goes into the model with probability w, independently of all the other variables. In (4.1b), $q_\gamma = \sum_{j=1}^q \gamma_j$ is the number of predictors that go in the regression, and the prior on β_γ is Zellner's g-prior (Zellner, 1986). Because (σ^2, β_0) is given an improper prior (line (4.1c)), the prior on θ is improper; however, it turns out that the posterior distribution of θ is proper. Models of the type (4.1) were introduced by Mitchell and Beauchamp (1988) and have been studied in dozens of papers; see Liang et al. (2008) for a review and recent developments.

The hyperparameter $h = (w, g)$ plays a critical role: if w is small and g is large, the prior p_h concentrates its mass on models with few variables and large coefficients, while if w is large and g is small, p_h concentrates its mass on models with many variables and small coefficients. (To appreciate the importance of the role played by h, note that George and Foster (2000) have shown that for the slightly different version of (4.1) in which σ^2 is assumed known, h can be chosen so that the highest posterior probability model is exactly the best model under the AIC/C_p, BIC, or RIC criteria.) Therefore, h plays a central role, and it is important to choose it properly. It is in order to do this that we need to estimate m_h, $h \in \mathcal{H}$, or rather, as discussed earlier estimate cm_h, $h \in \mathcal{H}$, where c is a constant. For our first goal, namely to estimate the family of posterior expectations $E_{\pi_h} f$, $h \in \mathcal{H}$, an example of a function of interest is $f(\theta) = I(\gamma_1 = 1)$, in which case $E_{\pi_h} f$ is the posterior probability that variable 1 is included in the model.

The marginal likelihood of the data, m_h, is in general the sum of 2^q integrals (George and Foster, 2000), and is computable when q is relatively small (q less than 20 or 25). The constant $\boldsymbol{d} = (m_{h_2}/m_{h_1}, \ldots, m_{h_k}/m_{h_1})$ is then available, and this case provides an example of a situation where the methods of Section 2.1 apply.

A Regenerative Markov Chain for Estimating the Posterior Distribution MCMC methodology for estimating posterior distributions for model (4.1) fall into two categories. Smith and Kohn (1996) developed a Markov chain algorithm which runs only on γ, the other variables being integrated out. Their chain is a simple Gibbs sampler which runs on the vector $(\gamma_1, \dots, \gamma_q)^\top$, updating one component at a time. Many variants of their scheme have been proposed; see Clyde and George (2004) for a review. It is also possible to devise a Markov chain that runs over $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$. The algorithm in Buta and Doss (2011) is of this sort. Accordingly, there are two ways of considering model (4.1).

1. We can take γ as the sole parameter. In this case, (4.1d) is the prior, and the likelihood, given by (4.1a)–(4.1c), is now an integral. Fortunately, this integral is available in closed form. We have

$$\ell_h(\gamma; y) = c_m \left\| y - (1/m)1_m 1_m^\top y \right\|^{-(m-1)} (1+g)^{(m-q_\gamma-1)/2} \left[1 + g(1 - R_\gamma^2)\right]^{-(m-1)/2}, \qquad (4.2)$$

   where R_γ^2 is the usual coefficient of determination of model γ, and c_m is a constant that depends only on m. Note that the likelihood now depends on the hyperparameter $h = (w, g)$ (through its second component) and our notation $\ell_h(\gamma; y)$ emphasizes this fact. Markov chains of the kind that were developed by Smith and Kohn (1996) are compatible with this point of view.

2. We can take the parameter to be $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$. In this case, (4.1b)–(4.1d) specify the prior on θ, and the likelihood is (4.1a). The Markov chain on θ developed by Buta and Doss (2011) is compatible with this view.

We now summarize. Suppose we are considering two hyperparameters $h_1 = (w_1, g_1)$ and $h_2 = (w_2, g_2)$, we have available a (single) Markov chain with invariant density π_{h_1}, and we wish to estimate $E_{\pi_{h_2}} f$ for some function f. As explained in the beginning of Section 3, we need to take averages which involve ratios of the sort

$$\frac{\ell_{h_2}(\vartheta) p_{h_2}(\vartheta)}{\ell_{h_1}(\vartheta) p_{h_1}(\vartheta)}, \qquad (4.3)$$

where ϑ is the parameter (either γ or θ).

In the first view of model (4.1), we generate a chain on only γ, and this chain can be a simple Gibbs sampler. Also, the ratio of the prior densities is trivial: from (4.1d) we have

$$p_{h_2}(\gamma)/p_{h_1}(\gamma) = (w_2/w_1)^{q_\gamma} \big((1-w_2)/(1-w_1)\big)^{q-q_\gamma}. \qquad (4.4)$$

But the price to pay is that we need to calculate $\ell_{h_2}(\gamma)/\ell_{h_1}(\gamma)$ which, in view of (4.2), involves the coefficient of determination R_γ^2. Calculation of R_γ^2 requires $O(q_\gamma^2)$ operations.

In the second way of viewing model (4.1), the likelihood does not involve the hyperparameter, and therefore cancels in (4.3). On the other hand, the price to be paid is that (i) we need to generate a chain that runs on $(\gamma, \sigma, \beta_0, \beta_\gamma)$, and (ii) the prior distributions are not absolutely continuous with respect to the product of counting measure on $\{0, 1\}^q$ and Lebesgue measure on $(0, \infty) \times \mathbb{R}_+ \times \mathbb{R}^{q+1}$ (the dimension of β_γ is not fixed). The "ratio of densities" $p_{h_2}(\theta)/p_{h_1}(\theta)$ then needs to be taken as a Radon-Nikodym derivative, and while a formula exists (Doss, 2007, eq. (7)), it is not nearly as simple as (4.4). Although it is possible to proceed with either view, in the present paper we take the first view.

Let $\pi_h(\gamma)$ denote the posterior distribution of γ in model (4.1). All existing MCMC algorithms that run on γ (be they Gibbs samplers or Metropolis-Hastings algorithms) require, in one way or another, the calculation of $\pi_h(\widetilde{\gamma})/\pi_h(\gamma)$ for $\widetilde{\gamma}, \gamma \in \{0,1\}^q$. This ratio is available from the relation

$$\pi_h(\gamma) \propto (1+g)^{(m-q_\gamma-1)/2} \left[1 + g(1-R_\gamma^2)\right]^{-(m-1)/2} w^{q_\gamma}(1-w)^{q-q_\gamma}, \tag{4.5}$$

which follows directly from (4.2). Here, we use the simple Gibbs sampler devised by Smith and Kohn (1996): the explicit nature of the Markov transition density for this chain makes it easy to develop the needed minorization condition, and this is done in the Appendix.

4.2 Analysis of the Ozone Data

Here we illustrate our methods on the ozone data set originally presented in Breiman and Friedman (1985). The data were re-analyzed by many authors and were recently analyzed in a Bayesian framework by Casella and Moreno (2006) and Liang et al. (2008). This data set seems ideal because it has been studied in several papers already, so we can compare our results with previous analyses. The dataset consists of daily measurements of the maximum ozone concentration near Los Angeles and eight meteorological variables for 330 days. Following Casella and Moreno (2006) and Liang et al. (2008), we consider the eight meteorological variables, plus two-way interactions and squares, leading to 44 possible predictors. This results in $2^{44} \approx 1.7 \times 10^{13}$ potential regression models. A description of the variables is given in Appendix D of Liang et al. (2008).

We have two goals. First, we wish to make a plot of the Bayes factor surface m_h/m_{h_1} as h varies, and also provide pointwise error margins. The literature has several (conflicting) suggestions for the value of g to use, and these are reviewed in Liang et al. (2008). The Bayes factor plot will enable us to rule out the h's which are obviously inappropriate (and in particular to rule out certain values of g), and will also provide us with an estimate of h_{opt}.

Liang et al. (2008) give the models (i.e. list of variables to include) for ten variable selection algorithms, including several Bayesian ones (all of which use $w = .5$). Interestingly, certain variables, notably hum and the interaction $\mathsf{hum}\times\mathsf{ibt}$, are included by several of the algorithms but not by others. Consider the quantities $E_{\pi_{(.5,g)}} I(\gamma_{\mathsf{hum}} = 1)$ and $E_{\pi_{(.5,g)}} I(\gamma_{\mathsf{hum}\times\mathsf{ibt}} = 1)$, which are the posterior probabilities that these two variables are included. Our second objective is to plot estimates of these as functions of g, and also to provide pointwise confidence bands.

The left panel of Figure 2 gives a plot of the estimates of the Bayes factors m_h/m_{h_1}. To form the plot we took as skeleton set the grid consisting of the 16 values

$$(w, g) \in \{.05, .15, .25, .35\} \times \{75, 125, 225, 375\}.$$

In Stage 1 we ran 16 chains, corresponding to the same skeleton values, for 500 regenerations each (a regeneration required about 60 iterations on average). We used these to obtain $\hat{\boldsymbol{d}}$ via the method described in Section 2.2. In Stage 2, we ran the same 16 regenerative chains, and used these, together with $\hat{\boldsymbol{d}}$, to form the estimates $\hat{u}_n$ for the 1311 values of h which result when w ranges from .01 to .45 in increments of .02 and g ranges from 20 to 300 in increments of 5. (We carried out a small pilot experiment to identify the set of h's having relatively high marginal likelihoods, and from this experiment we determined the skeleton grid and the range of h's for which the plot is made. Also, we selected $h_1 = (.15, 125)$ because this was a point of fairly high marginal likelihood.) Each chain was run for 200 regenerations. (Our theory does not require that we use the same number of regenerations for all the chains, and chains that mix more slowly could be run for a greater number of regenerations; however, we did not see a need to do this in the present example.) From the left panel of Figure 2 we see that the value of h at which the maximum is attained is $h = (.15, 150)$. The right panel gives a plot of the estimated standard errors for the estimates in the left panel. These are obtained using the estimate of variance given in Theorem 2. From the right panel we see that the estimated standard error is less than .039 over the entire range of the plot.

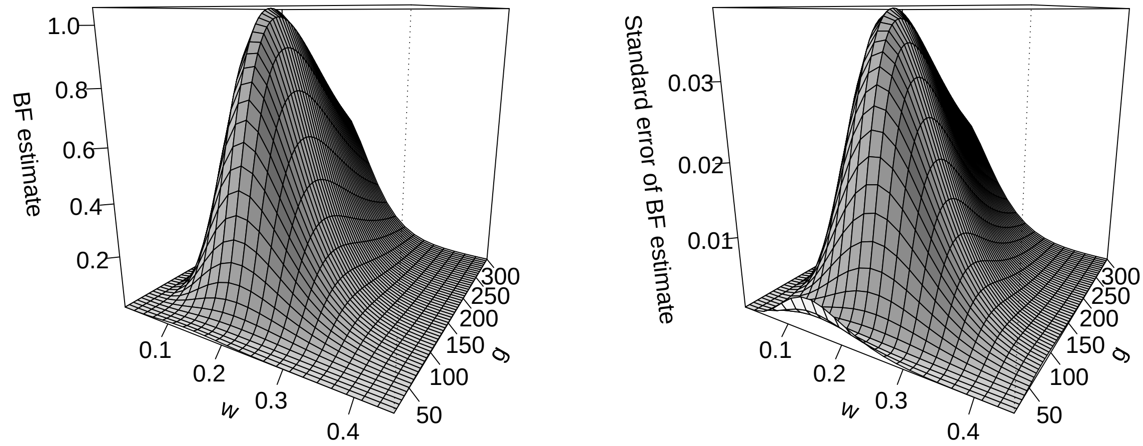

Figure 2: Model assessment for the ozone data. Left panel gives plot of the Bayes factor estimate as a function of hyperparameter values (w, g) when the baseline hyperparameter value is given by $w = .15$ and $g = 125$. Right panel gives the corresponding standard error estimates.

Figure 3 gives, for each of the variables hum and hum×ibt, the estimate of the posterior inclusion probability as a function of g, together with 95% confidence bands (the bands are formed using the estimate given in Theorem 2). From the figure, we see clearly that for any g that is deemed a reasonable choice by Figure 2, the variable hum should be included in the model, while there is considerable uncertainty regarding whether or not to include hum×ibt.

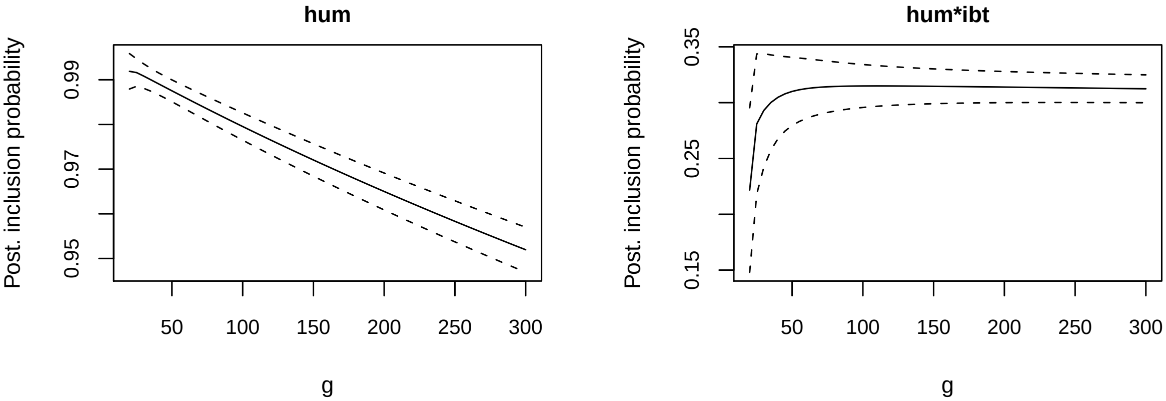

Figure 3: Uncertainty in variable inclusion for the ozone data. Left panel gives a plot of the Monte Carlo estimate of the posterior inclusion probability for the variable hum, as a function of g, together with 95% confidence bands, valid pointwise. Right panel is for the variable hum×ibt.

Figure 4 shows the evolution of the estimate of variance of the Bayes factor estimate for three values of h, as the number of tours increases (the ratio of number of Stage 1 tours to number of Stage 2 tours is kept fixed at 2.5). The plots suggest that the asymptotic variance estimates have stabilized when we use 200 tours for each chain in stage 2, and justifies our choice of $R_j = 200$ for $j = 1, \ldots, 16$.

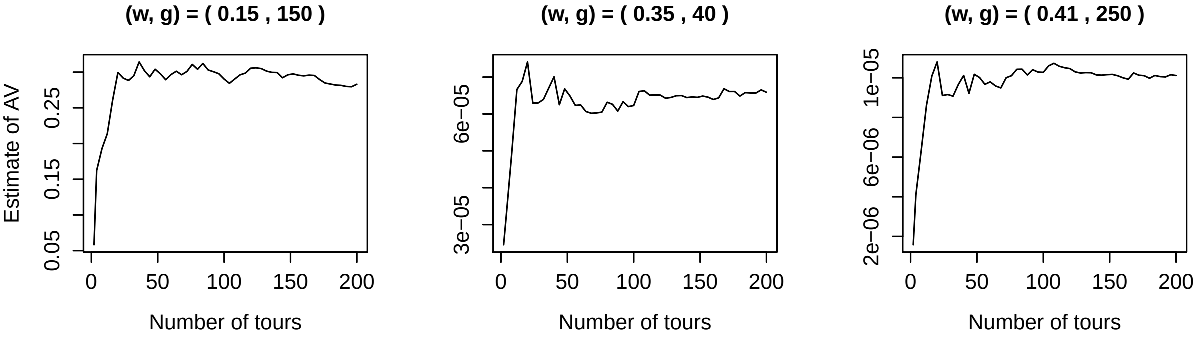

Figure 4: Plot of evolution of estimate of asymptotic variance of the Bayes factor estimate as the number of tours increases.

5 Discussion

The only hypothesis in Theorems 1 and 2 that may be difficult to check is the geometric ergodicity of the Markov chains. We note, however, that geometric ergodicity is a standard assumption that underlies almost every method for calculating standard errors of MCMC-based estimators (see, e.g., Roberts and Rosenthal, 1998; Jones and Hobert, 2001; Flegal et al., 2008). Currently, the main technique available for proving that a Markov chain is geometrically ergodic is the construction of a *geometric drift condition* as described in Chapter 15 of Meyn and Tweedie (1993). Successful applications of this technique include Roberts and Tweedie's (1996) analysis of Metropolis-Hastings algorithms, Roberts and Rosenthal's (1999) examination of slice samplers, and the studies of Jones and Hobert (2004) and Roy and Hobert (2007), who looked at Gibbs samplers. Despite these successes, it is generally understood that establishing geometric ergodicity of Monte Carlo Markov chains via drift conditions is quite challenging. Indeed, Fill et al. (2000) describe it as "difficult theoretical analysis," while Diaconis et al. (2008) lament that the required drift functions are hard to identify and refer to the method as "a matter of art."

In this paper we have shown that if $\Phi_1, \dots, \Phi_k$ are k geometrically ergodic Markov chains with invariant densities $\pi_{h_1}, \dots, \pi_{h_k}$, respectively, and if the density π_h is similar to at least one of $\pi_{h_1}, \dots, \pi_{h_k}$, then the scope of the k MCMC algorithms can be extended to the honest exploration of π_h. The methodology can be used in two ways. On occasion, π_h has a structure that is fundamentally different from that of $\pi_{h_1}, \dots, \pi_{h_k}$, and our approach enables us to circumvent the construction and analysis of a new MCMC algorithm; this is the case for the application in Section 3.1, in which $k = 1$. In other situations, all the π_h's come from the same parametric family, and the issue is not the need to develop new MCMC algorithms, but rather how to handle many π_h's simultaneously; this is the case for the application in Section 4.1.

Appendix

Proof of Theorem 2 We first prove the statement regarding $\hat{\eta}$. Note that

$$R_1^{1/2}\big[\hat{\eta}\big((1,\hat{\boldsymbol{d}}),\hat{\boldsymbol{d}}\big) - \eta\big] = R_1^{1/2}\big[\hat{\eta}\big((1,\hat{\boldsymbol{d}}),\hat{\boldsymbol{d}}\big) - \hat{\eta}\big((1,\boldsymbol{d}),\boldsymbol{d}\big)\big] + R_1^{1/2}\big[\hat{\eta}\big((1,\boldsymbol{d}),\boldsymbol{d}\big) - \eta\big]. \tag{A.1}$$

The second term on the right side of (A.1) involves randomness coming only from Stage 2 sampling, and its distribution is given by Theorem 1: it is asymptotically normal with mean 0 and an easy-to-estimate variance τ^2. The first term involves randomness from both Stage 1 and Stage 2 sampling. However, we will show that for this term, the randomness from Stage 2 is asymptotically negligible, so that only Stage 1 sampling contributes to its asymptotic distribution. This will enable us to obtain its asymptotic distribution, and the asymptotic normality of the left side of (A.1) will follow immediately, since the two stages of sampling are independent.

Now consider the first term on the right side of (A.1). Recall that if $\boldsymbol{a} = (1, \boldsymbol{d})$, then

$$v(x) := v(x; \boldsymbol{a}, \boldsymbol{d}) = \frac{f(x)\nu(x)}{\sum_{l=1}^{k} \nu_l(x)} \quad \text{and} \quad u(x) := u(x; \boldsymbol{a}, \boldsymbol{d}) = \frac{\nu(x)}{\sum_{l=1}^{k} \nu_l(x)}.$$

With (2.19) and (2.20) in mind, define the function

$$A(\boldsymbol{z}) = \hat{\eta}\big((1, \boldsymbol{z}), \boldsymbol{z}\big) = \sum_{l=1}^{k} \frac{z_l}{n_l} \sum_{i=1}^{n_l} v(X_i^{(l)}) \Big/ \sum_{l=1}^{k} \frac{z_l}{n_l} \sum_{i=1}^{n_l} u(X_i^{(l)})$$

for $\boldsymbol{z} = (z_2, \dots, z_k)^\top$, with $z_l > 0$ for $l = 2, \dots, k$, and $z_1 = 1$. Note that setting $\boldsymbol{z} = \boldsymbol{d}$ gives $A(\boldsymbol{d}) = \hat{\eta}((1, \boldsymbol{d}), \boldsymbol{d})$, and setting $\boldsymbol{z} = \hat{\boldsymbol{d}}$ gives $A(\hat{\boldsymbol{d}}) = \hat{\eta}((1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}})$. By a Taylor series expansion of A about $\boldsymbol{d}$ we get

$$\begin{aligned} R_1^{1/2}\big[\hat{\eta}\big((1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) - \hat{\eta}\big((1, \boldsymbol{d}), \boldsymbol{d}\big)\big] &= R_1^{1/2} \nabla A(\boldsymbol{d})^\top (\hat{\boldsymbol{d}} - \boldsymbol{d}) + \frac{R_1^{1/2}}{2} (\hat{\boldsymbol{d}} - \boldsymbol{d})^\top \nabla^2 A(\boldsymbol{d}^*)(\hat{\boldsymbol{d}} - \boldsymbol{d}) \\ &= R_1^{1/2} \nabla A(\boldsymbol{d})^\top (\hat{\boldsymbol{d}} - \boldsymbol{d}) + \frac{R_1^{1/2}}{2\rho_1} \big(\rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d})\big)^\top \nabla^2 A(\boldsymbol{d}^*) \big(\rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d})\big), \end{aligned}$$

where $\boldsymbol{d}^*$ is between $\boldsymbol{d}$ and $\hat{\boldsymbol{d}}$. As $R_1 \to \infty$, $n_l \to \infty$ for each l. We first show that the gradient $\nabla A(\boldsymbol{d})$ converges almost surely to a finite constant vector by proving that each one of its components, $[A(\boldsymbol{d})]_{j-1}$, $j = 2, \dots, k$, converges almost surely as $R_1 \to \infty$. As $n_l \to \infty$ for $l = 1, \dots, k$, for $j = 2, \dots, k$, we have

$$\begin{aligned} [\nabla A(\boldsymbol{d})]_{j-1} &= \frac{(1/n_j)\sum_{i=1}^{n_j} v(X_i^{(j)})}{\sum_{l=1}^{k}(d_l/n_l)\sum_{i=1}^{n_l} u(X_i^{(l)})} - \frac{\big(\sum_{l=1}^{k}(d_l/n_l)\sum_{i=1}^{n_l} v(X_i^{(l)})\big)\big((1/n_j)\sum_{i=1}^{n_j} u(X_i^{(j)})\big)}{\big(\sum_{l=1}^{k}(d_l/n_l)\sum_{i=1}^{n_l} u(X_i^{(l)})\big)^2} \\ &\xrightarrow{\text{a.s.}} \frac{E_{\pi_j} v}{\sum_{l=1}^{k} d_l E_{\pi_l} u} - \frac{\big(\sum_{l=1}^{k} d_l E_{\pi_l} v\big)\big(E_{\pi_j} u\big)}{\big(\sum_{l=1}^{k} d_l E_{\pi_l} u\big)^2}. \end{aligned} \tag{A.2}$$

The expression in (A.2) corresponds to H_{j-1}, which was defined in (2.21), and it is finite by Assumption 3 of Theorem 1. Next, we show that the random Hessian matrix $\nabla^2 A(\boldsymbol{d}^*)$ is bounded in probability, i.e., each element of this matrix is $O_p(1)$. As $n_l \to \infty$ for $l = 1, \dots, k$, for any $j, t \in \{2, \dots, k\}$, we have

$$\begin{aligned} [\nabla^2 F(\boldsymbol{d}^*)]_{t-1,j-1} &= -\frac{\big(\frac{1}{n_j}\sum_{i=1}^{n_j} v(X_i^{(j)})\big)\big(\frac{1}{n_t}\sum_{i=1}^{n_t} u(X_i^{(t)})\big)}{\big(\sum_{l=1}^{k} \frac{d_l^*}{n_l}\sum_{i=1}^{n_l} u(X_i^{(l)})\big)^2} \\ &\quad - \left(\frac{1}{n_j}\sum_{i=1}^{n_j} u(X_i^{(j)})\right)\left[\frac{\frac{1}{n_t}\sum_{i=1}^{n_t} v(X_i^{(t)})}{\big(\sum_{l=1}^{k} \frac{d_l^*}{n_l}\sum_{i=1}^{n_l} u(X_i^{(l)})\big)^2} - 2\frac{\big(\sum_{l=1}^{k} \frac{d_l^*}{n_l}\sum_{i=1}^{n_l} v(X_i^{(l)})\big)\big(\frac{1}{n_t}\sum_{i=1}^{n_t} u(X_i^{(t)})\big)}{\big(\sum_{l=1}^{k} \frac{d_l^*}{n_l}\sum_{i=1}^{n_l} u(X_i^{(l)})\big)^3}\right] \\ &\xrightarrow{\text{a.s.}} -\frac{(E_{\pi_j} v)(E_{\pi_t} u)}{\big(\sum_{l=1}^{k} d_l E_{\pi_l} u\big)^2} - (E_{\pi_j} u)\left[\frac{E_{\pi_t} v}{\big(\sum_{l=1}^{k} d_l E_{\pi_l} u\big)^2} - 2\frac{\big(\sum_{l=1}^{k} d_l E_{\pi_l} v\big)(E_{\pi_t} u)}{\big(\sum_{l=1}^{k} d_l E_{\pi_l} u\big)^3}\right], \end{aligned}$$

where the limits are also finite.

Now, we can rewrite (A.1) as

$$\begin{aligned}
R_1^{1/2}\big[\hat\eta\big((1,\hat{\boldsymbol d}),\hat{\boldsymbol d}\big)-\eta\big] &= (R_1/\rho_1)^{1/2}\nabla A(\boldsymbol d)^\top \rho_1^{1/2}(\hat{\boldsymbol d}-\boldsymbol d) + R_1^{1/2}\big[\hat\eta\big((1,\boldsymbol d),\boldsymbol d\big)-\eta\big] \\
&\quad + \frac{1}{2\rho_1^{1/2}}(R_1/\rho_1)^{1/2}\big[\rho_1^{1/2}(\hat{\boldsymbol d}-\boldsymbol d)\big]^\top \nabla^2 A(\boldsymbol d^*)\big[\rho_1^{1/2}(\hat{\boldsymbol d}-\boldsymbol d)\big] \\
&= q^{1/2}[\nabla A(\boldsymbol d)]^\top \rho_1^{1/2}(\hat{\boldsymbol d}-\boldsymbol d) + R_1^{1/2}\big[\hat\eta\big((1,\boldsymbol d),\boldsymbol d\big)-\eta\big] + o_p(1).
\end{aligned}$$

Since the two sampling stages are assumed to be independent, we conclude that

$$R_1^{1/2}\big[\hat\eta\big((1,\hat{\boldsymbol d}),\hat{\boldsymbol d}\big)-\eta\big] \xrightarrow{d} \mathcal N\big(0, q[\nabla A(\boldsymbol d)]^\top W[\nabla A(\boldsymbol d)]+\tau^2\big).$$

The proof of the CLT for $\hat u$ is similar but easier. As in (A.1), we have

$$R_1^{1/2}\big[\hat u\big((1,\hat{\boldsymbol d}),\hat{\boldsymbol d}\big)-m/m_1\big] = R_1^{1/2}\big[\hat u\big((1,\hat{\boldsymbol d}),\hat{\boldsymbol d}\big)-\hat u\big((1,\boldsymbol d),\boldsymbol d\big)\big]+R_1^{1/2}\big[\hat u\big((1,\boldsymbol d),\boldsymbol d\big)-m/m_1\big]. \quad \text{(A.3)}$$

The asymptotic distribution of the second term in (A.3) is given in Theorem 1. The first term is linear in $\hat{\boldsymbol d}-\boldsymbol d$:

$$\hat u\big((1,\hat{\boldsymbol d}),\hat{\boldsymbol d}\big)-\hat u\big((1,\boldsymbol d),\boldsymbol d\big) = \sum_{j=2}^k \bigg(\frac{1}{n_j}\sum_{i=1}^{n_j} u(X_i^{(j)})\bigg)(\hat d_j - d_j). \quad \text{(A.4)}$$

For $j=2,\dots,k$, the coefficient of $(\hat d_j - d_j)$ in (A.4) converges almost surely to $E_{\pi_j}u$, which is the term M_{j-1} defined in (2.21). Finally, from the independence of the two terms in (A.3) we conclude that

$$R_1^{1/2}\big[\hat u\big((1,\hat{\boldsymbol d}),\hat{\boldsymbol d}\big)-m/m_1\big] \xrightarrow{d} \mathcal N\big(0, qM^\top WM+\kappa^2\big). \qquad \square$$

Proof of Proposition 1 Consider a family of priors for $(\mu,\sigma_\theta^2,\sigma_e^2)$ given by

$$r_{a,b}(\mu,\sigma_\theta^2,\sigma_e^2) \propto (\sigma_\theta^2)^{-(a+1)}(\sigma_e^2)^{-(b+1)},$$

where a and b are known hyperparameters. Hobert and Casella (1996) showed that the posterior density under $r_{a,b}$ is proper if and only if

$$\frac{1-q}{2} < a < 0 \quad \text{and} \quad a+b > \frac{1-mq}{2}. \quad \text{(A.5)}$$

In other words, the integral

$$m_{a,b}(y) := \int_0^\infty\int_0^\infty\int_{\mathbb R}\int_{\mathbb R^q} (\sigma_\theta^2)^{-(a+1)}(\sigma_e^2)^{-(b+1)}\,\phi(\theta\,|\,\mu,\sigma_\theta^2,\sigma_e^2)\ell(\theta,\mu,\sigma_\theta^2,\sigma_e^2;y)\,d\theta\,d\mu\,d\sigma_\theta^2\,d\sigma_e^2$$

is finite if and only if (A.5) is satisfied. The standard diffuse prior corresponds to $r_{a,b}$ with $a=-1/2$ and $b=0$, and in this special case, (A.5) is satisfied if and only if $q\ge 3$. Now consider the reference prior, and note that

$$\Big[m-1+\big(\sigma_e^2/(\sigma_e^2+m\sigma_\theta^2)\big)^2\Big]^{1/2} \in \big[\sqrt{m-1},\sqrt{m}\,\big]$$

for all $(\sigma_e^2, \sigma_\theta^2) \in (0,\infty) \times (0,\infty)$. Consequently, as far as propriety of the posterior is concerned, the reference prior behaves like $r_{a,b}$ with $a = -1 + C_m/2$ and $b = 0$. Combining Hobert and Casella's (1996) 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 π_r is well defined. Noting that

$$
\begin{aligned}
&E_{\pi_r}\left[(\sigma_\theta^2)^s(\sigma_e^2)^t\right] \\
&\quad = \frac{1}{m_{-1/2,0}(y)} \int_0^\infty \int_0^\infty \int_{\mathbb{R}} \int_{\mathbb{R}^q} (\sigma_\theta^2)^{s-1/2}(\sigma_e^2)^{t-1}\, \phi(\theta \mid \mu, \sigma_\theta^2, \sigma_e^2)\, \ell(\theta, \mu, \sigma_\theta^2, \sigma_e^2; y)\, d\theta\, d\mu\, d\sigma_\theta^2\, d\sigma_e^2,
\end{aligned}
$$

the result follows directly from another application of Hobert and Casella's (1996) propriety result. □

Minorization Condition for the Gibbs Sampler of Smith and Kohn (1996) The transition density of the Gibbs sampler of Smith and Kohn (1996) is given by

$$
k(\widetilde{\gamma} \mid \gamma) = \pi_h(\widetilde{\gamma}_1 \mid \gamma_2, \ldots, \gamma_q)\pi_h(\widetilde{\gamma}_2 \mid \widetilde{\gamma}_1, \gamma_3 \ldots, \gamma_q) \ldots \pi_h(\widetilde{\gamma}_q \mid \widetilde{\gamma}_1, \ldots, \widetilde{\gamma}_{q-1}), \qquad \gamma, \widetilde{\gamma} \in \{0,1\}^q,
$$

where π_h is given by (4.5). It is easy to see that the underlying Markov chain is irreducible and aperiodic. And because the state space is finite, the chain is actually uniformly ergodic. We now give a minorization condition that can be used as the basis for regeneration in the Gibbs sampler. The method we use involves the "distinguished point" technique discussed in Mykland et al. (1995). Let γ^* denote a fixed model, which we will refer to as a distinguished point, and let $D \subset \{0,1\}^q$ be a set of models. (Both γ^* and D are arbitrary, but below we give guidelines for making a practical choice.) We have

$$
\begin{aligned}
k(\widetilde{\gamma} \mid \gamma) &= \frac{k(\widetilde{\gamma} \mid \gamma)}{k(\widetilde{\gamma} \mid \gamma^*)} k(\widetilde{\gamma} \mid \gamma^*) \\
&\geq \left[\min_{\gamma' \in D} \frac{k(\gamma' \mid \gamma)}{k(\gamma' \mid \gamma^*)}\right] k(\widetilde{\gamma} \mid \gamma^*) I_D(\widetilde{\gamma}) \\
&= \left\{c \min_{\gamma' \in D} \frac{k(\gamma' \mid \gamma)}{k(\gamma' \mid \gamma^*)}\right\} \left\{\frac{1}{c} k(\widetilde{\gamma} \mid \gamma^*) I_D(\widetilde{\gamma})\right\} \\
&=: s(\gamma) q(\widetilde{\gamma}),
\end{aligned}
$$

where $c = \sum_{\gamma' \in D} k(\gamma' \mid \gamma^*)$ is a normalizing constant. Note that the value of c is not required to carry out regenerative simulation because the regeneration probability (2.3) involves the product of s and q in which c cancels out.

The choice of γ^* and D affects the regeneration rate. Referring to equation (2.3), ideally we would like the regeneration probability to be as big as possible. Notice that regeneration can occur only if $\widetilde{\gamma} \in D$. This suggests making D large. However, increasing the size of D makes $s(\gamma)$ smaller. We have found that a reasonable tradeoff consists of taking D to be the smallest set of models that encompasses 10% of the posterior probability. Also, the obvious choice for γ^* is the highest probability model. The distinguished point γ^* and the set D are selected from the output of a pilot (relatively short) chain.

Acknowledgements

We thank two referees and the associate editor for suggestions which improved the paper. Both Doss and Hobert's research were supported by NSF grants DMS-08-05860 and DMS-11-06395. Doss's research was also supported by NIH grant P30 AG028740.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88** 669–679.

Berger, J. O. and Bernardo, J. M. (1992). Reference priors in a variance components problem. In *Bayesian Analysis in Statistics and Econometrics* (P. K. Goel and N. S. Iyengar, eds.). Lecture Notes in Statistics, Springer, New York, 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.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80** 580–598.

Burr, D. and Doss, H. (1993). Confidence bands for the median survival time as a function of the covariates in the Cox model. *Journal of the American Statistical Association* **88** 1330–1340.

Buta, E. and Doss, H. (2011). Computational approaches for empirical Bayes methods and Bayesian sensitivity analysis. *The Annals of Statistics* **39** 2658–2685.

Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association* **101** 157–167.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* **19** 81–94.

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

Doss, H. (2007). Bayesian model selection: Some thoughts on future directions. *Statistica Sinica* **17** 413–421.

Doss, H. and Tan, A. (2014). Estimates and standard errors for ratios of normalizing constants from multiple Markov chains via regeneration. *Journal of the Royal Statistical Society,* Series B (to appear) .

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

Flegal, J. M., Haran, M. and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science* **23** 250–260.

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

George, E. and Foster, D. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731.

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

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. *Technical Report, School of Statistics, University of Minnesota* .

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

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics* **16** 1069–1112.

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.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

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., 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.

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.

Jones, G. L. and Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *The Annals of Statistics* **32** 784–817.

Kong, A., McCullagh, P., Meng, X., Nicolae, D. and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *Journal of the Royal Statistical Society,* Series B **65** 585–618.

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

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association* **103** 410–423.

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.

Meng, X. and Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6** 831–860.

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

Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83** 1023–1036.

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

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society,* Series B **56** 3–48.

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

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

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. 2nd ed. Springer, New York.

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

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

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

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys* **1** 20–71.

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

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.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75** 317–343.

Tan, A. and Hobert, J. P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: Convergence and regeneration. *Journal of Computational and Graphical Statistics* **18** 861–878.

Tan, Z. (2004). On a likelihood approach for Monte Carlo integration. *Journal of the American Statistical Association* **99** 1027–1036.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics* **13** 178–203.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* **6** 233–243.