Statistica Sinica

ESTIMATING STANDARD ERRORS FOR IMPORTANCE SAMPLING ESTIMATORS WITH MULTIPLE MARKOV CHAINS

Vivekananda Roy, Aixin Tan, and James M. Flegal

Iowa State University, University of Iowa, and University of California, Riverside

Abstract: The naive importance sampling estimator, based on samples from a single importance density, can be numerically unstable. We consider generalized importance sampling estimators where samples from more than one probability distribution are combined. We study this problem in the Markov chain Monte Carlo context, where independent samples are replaced with Markov chain samples. If the chains converge to their respective target distributions at a polynomial rate, then under two finite moment conditions, we show a central limit theorem holds for the generalized estimators. We develop an easy-to-implement method to calculate valid asymptotic standard errors based on batch means. We provide a batch means estimator for calculating asymptotically valid standard errors of Geyer's (1994) reverse logistic estimator. We illustrate the method via three examples. In particular, the generalized importance sampling estimator is used for Bayesian spatial modeling of binary data and to perform empirical Bayes variable selection where the batch means estimator enables standard error calculations in high-dimensional settings.

Key words and phrases: Bayes factors, Markov chain Monte Carlo, polynomial ergodicity, ratios of normalizing constants, reverse logistic estimator.

1 Introduction

Let $\pi(x) = \nu(x)/m$ be a probability density function (pdf) on X with respect to a measure $\mu(\cdot)$. Suppose $f : \mathsf{X} \to \mathbb{R}$ is a π integrable function and we want to estimate $E_\pi f := \int_{\mathsf{X}} f(x)\pi(x)\mu(dx)$. Let $\pi_1(x) = \nu_1(x)/m_1$ be another pdf on X such that $\{x : \pi_1(x) = 0\} \subset \{x : \pi(x) = 0\}$. The importance sampling (IS) estimator of $E_\pi f$ based on independent and identically distributed (iid) samples $X_1, \ldots, X_n$ from the importance density π_1 is

$$\frac{\sum_{i=1}^n f(X_i)\nu(X_i)/\nu_1(X_i)}{\sum_{i=1}^n \nu(X_i)/\nu_1(X_i)} \xrightarrow{\text{a.s.}} \int_{\mathsf{X}} \frac{f(x)\nu(x)/m}{\nu_1(x)/m_1}\pi_1(x)\,\mu(dx) \Big/ \int_{\mathsf{X}} \frac{\nu(x)/m}{\nu_1(x)/m_1}\pi_1(x)\,\mu(dx) = E_\pi f, \quad (1.1)$$

as $n \to \infty$. This estimator can also be used in the Markov chain Monte Carlo (MCMC) context when $X_1, \ldots, X_n$ are realizations from a suitably irreducible Markov chain with stationary density π_1 (Hastings (1970)). Note that (1.1) requires the functions ν and ν_1 to be known. On the other hand, it does not depend on normalizing constants m and m_1, which are generally unknown.

In this article, we consider situations where one wants to estimate $E_\pi f$ for all π belonging to a large collection, say Π. This situation arises in both frequentist and Bayesian statistics. Although (1.1) provides consistent estimators of $E_\pi f$ for all $\pi \in \Pi$ based on a single Markov chain $\{X_n\}_{n\geq 0}$ with stationary density π_1, it does not work well when π differs greatly from π_1. In that case the ratios $\nu(x)/\nu_1(x)$ can be arbitrarily large for some sample values making the estimator at (1.1) unstable. In general, there is not a single good importance density π_1 which is close to all $\pi \in \Pi$ (see e.g. Geyer (1994)). Hence a natural modification is to replace π_1 in (1.1) with a mixture of densities where each density in Π is close to a subset of the k reference densities. To this end, let $\overline{\pi} \equiv \sum_{i=1}^k (a_i/|\boldsymbol{a}|)\pi_i$,

where $\boldsymbol{a} = (a_1, \ldots, a_k)$ are k positive constants, $|\boldsymbol{a}| = \sum_{i=1}^k a_i$, and $\pi_i(x) = \nu_i(x)/m_i$ for $i = 1, \ldots, k$ are k densities known up to their normalizing constants. Suppose further that $n_1, \ldots, n_k$ are positive integers and $d_i := m_i/m_1$ for $i = 2, \ldots, k$, with $d_1 \equiv 1$. Then define the $(k-1)$ dimensional vector

$$\boldsymbol{d} = (m_2/m_1, \ldots, m_k/m_1). \tag{1.2}$$

Finally for $l = 1, \ldots, k$, let $\{X_i^{(l)}\}_{i=1}^{n_l}$ be an iid sample from π_l or realizations from a positive Harris Markov chain with invariant density π_l (for definitions see Meyn and Tweedie (1993)). Then as $n_l \to \infty$, for all $l = 1, \ldots, k$, we have

$$\begin{aligned}
\hat{\eta} &\equiv \left(\sum_{l=1}^k \frac{a_l}{n_l} \sum_{i=1}^{n_l} \frac{f(X_i^{(l)})\nu(X_i^{(l)})}{\sum_{s=1}^k a_s \nu_s(X_i^{(l)})/d_s}\right) \Bigg/ \left(\sum_{l=1}^k \frac{a_l}{n_l} \sum_{i=1}^{n_l} \frac{\nu(X_i^{(l)})}{\sum_{s=1}^k a_s \nu_s(X_i^{(l)})/d_s}\right) \\
&\xrightarrow{\text{a.s.}} \left(\sum_{l=1}^k a_l \int_{\mathsf{X}} f(x) \frac{\nu(x)}{\sum_{s=1}^k a_s \nu_s(x)/d_s} \pi_l(x)\, \mu(dx)\right) \Bigg/ \left(\sum_{l=1}^k a_l \int_{\mathsf{X}} \frac{\nu(x)}{\sum_{s=1}^k a_s \nu_s(x)/d_s} \pi_l(x)\, \mu(dx)\right) \\
&= \int_{\mathsf{X}} f(x) \frac{\nu(x)}{\bar{\pi}(x)} \bar{\pi}(x)\, \mu(dx) \Bigg/ \int_{\mathsf{X}} \frac{\nu(x)}{\bar{\pi}(x)} \bar{\pi}(x)\, \mu(dx) = E_\pi f.
\end{aligned} \tag{1.3}$$

The generalized IS estimator (1.3) has been discussed widely in the literature, e.g. applications include Monte Carlo maximum likelihood estimation and Bayesian sensitivity analysis. Gill et al. (1988), Kong et al. (2003), Meng and Wong (1996), Tan (2004), and Vardi (1985) consider estimation using (1.3) based on iid samples. The estimator is applicable to a much larger class of problems if Markov chain samples are allowed, see e.g. Buta and Doss (2011), Geyer (1994), and Tan et al. (2015), which is the setting of this paper.

Alternative importance weights have also been proposed. In the case when the normalizing constants m_i's are known, the estimator (1.3) resembles the *balance heuristic* estimator of Veach

and Guibas (1995), which is revisited in Owen and Zhou (2000) as the *deterministic mixture*. The standard population Monte Carlo algorithm of Cappé et al. (2004) uses a weighted ratio of the target π and the proposal π_j it was drawn from (evaluated at the sample itself). However, when iid samples are available from $\pi_j, j = 1, 2, \ldots, k$, Elvira et al. (2017) shows that the normalized estimator (m_i's known) version of (1.3) always has a smaller variance than that of the population Monte Carlo algorithm. Further, it may be difficult in practice to find fully known importance densities that approximate the target densities. Indeed, applications such as in empirical Bayes analysis and Bayesian sensitivity analysis routinely select representatives from the large number of target posterior densities to serve as proposal densities, and they are known only up to normalizing constants. See Buta and Doss (2011), Doss (2010), as well as Section 5 for examples. Although there is no known proof for the self normalized estimator (Elvira et al. (2017), p. 18), it is reasonable to assume the superiority of (1.3) over estimators corresponding to other weighting schemes.

As noted in (1.3), the estimator $\hat{\eta}$ converges to $E_\pi f$ as the sample sizes increase to infinity, for iid samples as well as Markov chain samples satisfying the usual regularity conditions. Now for samples of finite size, it is of fundamental importance to provide some measure of uncertainty, such as the standard errors (SEs) associated with this consistent estimator. For estimators that are sample averages based on iid Monte Carlo samples, for example, it is a basic requirement to report their SEs. But the very same issue is often overlooked in practice when the estimators have more complicated structure, and when they are based on MCMC samples, largely due to the difficulty of doing so. See, for e.g. Flegal et al. (2008) on the issue concerning MCMC experiments and Koehler et al. (2009)

for more general simulation studies. For calculating SEs of $\hat{\eta}$ based on MCMC samples, Tan et al. (2015) provide a solution using the method of regenerative simulation (RS). However, this method crucially depends on the construction of a practical minorization condition, i.e. one where sufficient regenerations are observed in finite simulations (for definitions and a description of RS see Mykland et al. (1995)). Further, the usual method of identifying regeneration times by splitting becomes impractical for high-dimensional problems (Gilks et al. (1998)). Hence, successful applications of RS involve significant trial and error and are usually limited to low-dimensional Gibbs samplers (see e.g. Tan and Hobert (2009); Roy and Hobert (2007)). In this paper we avoid RS and provide SE estimators of $\hat{\eta}$ using the batch means (BM) method, which is straightforward to implement and can be routinely applied in practice. In obtaining this estimator, we also establish a central limit theorem (CLT) for $\hat{\eta}$ that generalizes some results in Buta and Doss (2011).

The estimator $\hat{\eta}$ in (1.3) depends on the ratios of normalizing constants, $\boldsymbol{d}$, that are unknown in applications. We consider the two-stage scheme studied in Buta and Doss (2011) where first an estimate $\hat{\boldsymbol{d}}$ is obtained using Geyer's (1994) "reverse logistic regression" method based on samples from π_l, and then, independently, new samples are used to estimate $E_\pi f$ for $\pi \in \Pi$ using the estimator $\hat{\eta}(\hat{\boldsymbol{d}})$ in (1.3). Buta and Doss (2011) showed that the asymptotic variance of $\hat{\eta}(\hat{\boldsymbol{d}})$ depends on the asymptotic variance of $\hat{\boldsymbol{d}}$. Thus we study the CLT of $\hat{\boldsymbol{d}}$ and provide a BM estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{d}}$. Since $\hat{\boldsymbol{d}}$ involves multiple Markov chain samples, we utilize a multivariate BM estimator. Although, the form of the asymptotic covariance matrix of $\hat{\boldsymbol{d}}$ is complicated, our consistent BM estimator is straightforward to code.

The problem of estimating $\boldsymbol{d}$, the ratios of normalizing constants of unnormalized densities is important in its own right and has many applications in frequentist and Bayesian inference. For example, when the samples are iid sequences this is the biased sampling problem studied in Vardi (1985). In addition, the problem arises naturally in the calculations of likelihood ratios in missing data (or latent variable) models, mixture densities for use in IS, and Bayes factors.

We consider the problem of estimating $\boldsymbol{d}$ using Geyer's (1994) reverse logistic regression method. Specifically, we study the general quasi-likelihood function proposed in Doss and Tan (2014). Unlike Geyer's (1994) method, this extended quasi-likelihood function has the advantage of using user defined weights that are appropriate to situations where the multiple Markov chains have different mixing rates. We establish the CLT for the resulting estimators of $\boldsymbol{d}$ and develop the BM estimators of their asymptotic covariance matrix.

Thus we consider two related problems in this paper: estimating (ratios of) normalizing constants given samples from k densities; estimating expectations with respect to a large number of (other) target distributions using these samples. In both cases, we establish CLTs for our estimators and provide easy-to-calculate SEs using BM methods.

Prior results of Buta and Doss (2011), Doss and Tan (2014), Geyer (1994), and Tan et al. (2015) all assume that the underlying Markov chains are geometrically ergodic. We weaken this condition in that we only require the chains to be *polynomial ergodic*. To this end, let $K_l(x, \cdot)$ be the Markov transition function for the Markov chain $\Phi_l = \{X_t^{(l)}\}_{t \geq 1}$, so that for any measurable set A, and $s, t \in \{1, 2, \ldots\}$ we have $P\big(X_{s+t}^{(l)} \in A \,|\, X_s^{(l)} = x\big) = K_l^t(x, A)$. Let $\|\cdot\|$ denote the total variation norm

and Π_l the probability measure corresponding to the density π_l. The Markov chain Φ_l is *polynomially ergodic of order* m where $m > 0$ if there exists $W : \mathsf{X} \to \mathbb{R}^+$ with $E_{\pi_l} W < \infty$ such that

$$\|K_l^t(x, \cdot) - \Pi_l(\cdot)\| \leq W(x) t^{-m}.$$

There is substantial MCMC literature establishing that Markov chains are at least polynomially ergodic (see Vats et al. (2016+) and the references therein).

We illustrate the generalized IS method and importance of obtaining SEs through three examples. First, we consider a toy example to demonstrate that BM and RS estimators are consistent and investigate the benefit of allowing general weights to be used in generalized IS. Second, we consider a Bayesian spatial model for a root rot disease dataset where we illustrate the importance of calculating SEs by considering different designs and performing samples size calculations. Finally, we consider a standard linear regression model with a large number of variables and use the BM estimator developed here for empirical Bayes variable selection.

The rest of the paper is organized as follows. Section 2 is devoted to the problem of estimating the ratios of normalizing constants of unnormalized densities, that is estimating $\boldsymbol{d}$. Section 3 contains the construction of a CLT for $\hat{\eta}$ and describes how valid SEs of $\hat{\eta}$ can be obtained using BM. Section 4 contains a toy example illustrating the benefits of different weight functions. Section 5 considers a Bayesian spatial models for binary responses. The empirical Bayes variable selection example is contained in the supplement. We conclude with a discussion in Section 6. Proofs are relegated to the online supplementary material.

2 Estimating ratios of normalizing constants

Consider k densities $\pi_l = \nu_l/m_l, l = 1, \ldots, k$ with respect to the measure μ, where the ν_l's are known functions and the m_l's are unknown constants. For each l we have a positive Harris Markov chain $\Phi_l = \{X_1^{(l)}, \ldots, X_{n_l}^{(l)}\}$ with invariant density π_l. Our objective is to estimate all possible ratios m_i/m_j, $i \neq j$ or, equivalently, the vector $\boldsymbol{d}$ defined in (1.2).

Geyer's (1994) reverse logistic regression is described as follows. Let $n = \sum n_l$ and set $a_l = n_l/n$ for now. For $l = 1, \ldots, k$ define the vector $\boldsymbol{\zeta}$ by

$$\zeta_l = -\log(m_l) + \log(a_l)$$

and let

$$p_l(x, \boldsymbol{\zeta}) = \frac{\nu_l(x)e^{\zeta_l}}{\sum_{s=1}^{k} \nu_s(x)e^{\zeta_s}}. \tag{2.1}$$

Given the value x belongs to the pooled sample $\left\{X_i^{(l)},\ i = 1, \ldots, n_l,\ l = 1, \ldots, k\right\}$, $p_l(x, \boldsymbol{\zeta})$ is the probability that x came from the l^{th} distribution. Of course, we know which distribution the sample x came from, but here we pretend that the only thing we know about x is its value and estimate $\boldsymbol{\zeta}$ by maximizing the log quasi-likelihood function

$$l_n(\boldsymbol{\zeta}) = \sum_{l=1}^{k} \sum_{i=1}^{n_l} \log\big(p_l(X_i^{(l)}, \boldsymbol{\zeta})\big) \tag{2.2}$$

with respect to $\boldsymbol{\zeta}$. Since $\boldsymbol{\zeta}$ has a one-to-one correspondence with $\boldsymbol{m} = (m_1, \ldots, m_k)$, by estimating $\boldsymbol{\zeta}$ we can estimate $\boldsymbol{m}$.

As Geyer (1994) mentioned, there is a non-identifiability issue regarding $l_n(\boldsymbol{\zeta})$: for any constant $c \in \mathbb{R}$, $l_n(\boldsymbol{\zeta})$ is same as $l_n(\boldsymbol{\zeta} + c1_k)$ where 1_k is the vector of k 1's. So we can estimate the true $\boldsymbol{\zeta}$ only

up to an additive constant. Thus, we can estimate $\boldsymbol{m}$ only up to an overall multiplicative constant, that is, we can estimate only $\boldsymbol{d}$. Let $\boldsymbol{\zeta}_0 \in \mathbb{R}^k$ be defined by $[\boldsymbol{\zeta}_0]_l = [\zeta]_l - \big(\sum_{s=1}^k [\zeta]_s\big)/k$, the true $\boldsymbol{\zeta}$ normalized to add to zero. Geyer (1994) proposed to estimate $\boldsymbol{\zeta}_0$ by $\hat{\boldsymbol{\zeta}}$, the maximizer of l_n subject to the linear constraint $\boldsymbol{\zeta}^\top 1_k = 0$, and thus obtain an estimate of $\boldsymbol{d}$. The estimator $\hat{\boldsymbol{d}}$ (written explicitly in Section 2.1), was introduced by Vardi (1985), and studied further by Gill et al. (1988), who proved that in the iid setting, $\hat{\boldsymbol{d}}$ is consistent and asymptotically normal, and established its efficiency. Geyer (1994) proved the consistency and asymptotic normality of $\hat{\boldsymbol{d}}$ when $\Phi_l, \ldots, \Phi_k$ are k Markov chains satisfying certain mixing conditions. In the iid setting, Meng and Wong (1996), Kong et al. (2003), and Tan (2004) rederived the estimate under different computational schemes.

None of these articles discuss how to consistently estimate the covariance matrix of $\hat{\boldsymbol{d}}$, even in the iid setting. Recently, Doss and Tan (2014) address this important issue and obtain a RS estimator of the covariance matrix of $\hat{\boldsymbol{d}}$ in the Markov chain setting. They also mention that the optimality results of Gill et al. (1988) do not hold in the Markov chain case. In particular, when using Markov chain samples, the choice of the weights $a_j = n_j/n$ to the probability density ν_j/m_j in the denominator of (2.1) is no more optimal and should instead incorporate the effective sample size of different chains as they might have quite different rates of mixing. They introduce the more general log quasi-likelihood function

$$\ell_n(\boldsymbol{\zeta}) = \sum_{l=1}^{k} w_l \sum_{i=1}^{n_l} \log\big(p_l(X_i^{(l)}, \boldsymbol{\zeta})\big), \tag{2.3}$$

where the vector $w \in \mathbb{R}^k$ is defined by $w_l = a_l n/n_l$ for $l = 1, \ldots, k$ for an arbitrary probability vector $\boldsymbol{a}$. (Note the change of notation from l to ℓ.) Clearly if $a_l = n_l/n$, then $w_l = 1$ and (2.3) becomes

(2.2).

When RS can be used, Doss and Tan (2014) proved the consistency (to the true value $\boldsymbol{\zeta}_0$) and asymptotic normality of the constrained maximizer $\hat{\boldsymbol{\zeta}}$ (subject to the constraint $\boldsymbol{\zeta}^\top 1_k = 0$) of (2.3) under geometric ergodicity. They also obtain a RS estimator of the asymptotic covariance matrix and describe an empirical method for choosing the optimal $\boldsymbol{a}$ based on minimizing the trace of the estimated covariance matrix of $\hat{\boldsymbol{d}}$. However, their procedure requires a practical minorization condition for each of the k Markov chains, which can be extremely difficult. Without a minorization condition, we show $\hat{\boldsymbol{d}}$ is a consistent estimator of $\boldsymbol{d}$, show $\hat{\boldsymbol{d}}$ satisfies a CLT under significantly weaker mixing conditions, and provide a strongly consistent BM estimator of the covariance matrix of $\hat{\boldsymbol{d}}$.

2.1 Central limit theorem and asymptotic covariance estimator

Within each Markov chain $l = 1, \ldots, k$, assume $n_l \to \infty$ in such a way that $n_l/n \to s_l \in (0, 1)$. To obtain the CLT result for $\hat{\boldsymbol{d}}$, we first establish a CLT for $\hat{\boldsymbol{\zeta}}$. Note that the function $g\colon \mathbb{R}^k \to \mathbb{R}^{k-1}$ that maps $\boldsymbol{\zeta}_0$ into $\boldsymbol{d}$ is given by

$$g(\boldsymbol{\zeta}) = \begin{pmatrix} e^{\zeta_1 - \zeta_2} a_2/a_1 \\ e^{\zeta_1 - \zeta_3} a_3/a_1 \\ \vdots \\ e^{\zeta_1 - \zeta_k} a_k/a_1 \end{pmatrix}, \tag{2.4}$$

and its gradient at $\boldsymbol{\zeta}_0$ (in terms of $\boldsymbol{d}$) is

$$D = \begin{pmatrix} d_2 & d_3 & \dots & d_k \\ -d_2 & 0 & \dots & 0 \\ 0 & -d_3 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -d_k \end{pmatrix}. \tag{2.5}$$

Since $\boldsymbol{d} = g(\boldsymbol{\zeta}_0)$, and by definition $\hat{\boldsymbol{d}} = g(\hat{\boldsymbol{\zeta}})$, we can use the CLT result of $\hat{\boldsymbol{\zeta}}$ to get a CLT for $\hat{\boldsymbol{d}}$.

First, we introduce the following notations. For $r = 1, \dots, k$, let

$$Y_i^{(r,l)} = p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big), \qquad i = 1, \dots, n_l. \tag{2.6}$$

The asymptotic covariance matrix in the CLT of $\hat{\boldsymbol{\zeta}}$, involves two $k \times k$ matrices B and Ω. The matrix B is given by

$$\begin{aligned} B_{rr} &= \sum_{j=1}^{k} a_j E_{\pi_j}\big(p_r(X, \boldsymbol{\zeta})[1 - p_r(X, \boldsymbol{\zeta})]\big) \text{ and} \\ B_{rs} &= -\sum_{j=1}^{k} a_j E_{\pi_j}\big(p_r(X, \boldsymbol{\zeta}) p_s(X, \boldsymbol{\zeta})\big) \text{ for } r \neq s. \end{aligned} \tag{2.7}$$

Let Ω be the $k \times k$ matrix defined (for $r, s = 1, \dots, k$) by

$$\Omega_{rs} = \sum_{l=1}^{k} \frac{a_l^2}{s_l} \Big[E_{\pi_l}\{Y_1^{(r,l)} Y_1^{(s,l)}\} + \sum_{i=1}^{\infty} E_{\pi_l}\{Y_1^{(r,l)} Y_{1+i}^{(s,l)}\} + \sum_{i=1}^{\infty} E_{\pi_l}\{Y_{1+i}^{(r,l)} Y_1^{(s,l)}\} \Big]. \tag{2.8}$$

Remark 1. The right hand side of (2.8) involves terms of the form $E_{\pi_l}\{Y_1^{(r,l)} Y_{1+i}^{(s,l)}\}$ and $E_{\pi_l}\{Y_{1+i}^{(r,l)} Y_1^{(s,l)}\}$. For any fixed l, r, s and i, the two expectations are the same if $X_1^{(l)}$ and $X_{1+i}^{(l)}$ are exchangeable, e.g. if the chain Φ_l is reversible. In general, the two expectations are not equal.

The matrix B will be estimated by its natural estimate $\widehat{B}$ defined by

$$\begin{aligned}\widehat{B}_{rr} &= \sum_{l=1}^{k} a_l \left(\frac{1}{n_l} \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \left[1 - p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \right] \right) \text{ and} \\ \widehat{B}_{rs} &= -\sum_{l=1}^{k} a_l \left(\frac{1}{n_l} \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) p_s(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \right) \text{ for } r \neq s.\end{aligned} \tag{2.9}$$

To obtain a BM estimate $\widehat{\Omega}$, suppose we simulate the Markov chain Φ_l for $n_l = e_l b_l$ iterations (hence $e_l = e_{n_l}$ and $b_l = b_{n_l}$ are functions of n_l) and define for $r, l = 1, \ldots, k$,

$$\bar{Z}_m^{(r,l)} := \frac{1}{b_l} \sum_{j=mb_l+1}^{(m+1)b_l} p_r(X_j^{(l)}, \hat{\boldsymbol{\zeta}}) \qquad \text{for } m = 0, \ldots, e_l - 1 \, .$$

Now set $\bar{Z}_m^{(l)} = \left(\bar{Z}_m^{(1,l)}, \ldots, \bar{Z}_m^{(k,l)} \right)^\top$ for $m = 0, \ldots, e_l - 1$. For $l = 1, \ldots, k$, denote $\bar{\bar{Z}}^{(l)} = \left(\bar{\bar{Z}}^{(1,l)}, \ldots, \bar{\bar{Z}}^{(k,l)} \right)^\top$ where $\bar{\bar{Z}}^{(r,l)} = \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}})/n_l$. Let

$$\widehat{\Sigma}^{(l)} = \frac{b_l}{e_l - 1} \sum_{m=0}^{e_l - 1} \left[\bar{Z}_m^{(l)} - \bar{\bar{Z}}^{(l)} \right] \left[\bar{Z}_m^{(l)} - \bar{\bar{Z}}^{(l)} \right]^T \text{ for } l = 1, \ldots, k, \tag{2.10}$$

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}^{(1)} & & 0 \\ & \ddots & \\ & & \\ 0 & & \widehat{\Sigma}^{(k)} \end{pmatrix} \tag{2.11}$$

and define the $k \times k^2$ matrix

$$A_n = \begin{pmatrix} -\sqrt{\frac{n}{n_1}} a_1 I_k & -\sqrt{\frac{n}{n_2}} a_2 I_k & \ldots & -\sqrt{\frac{n}{n_k}} a_k I_k \end{pmatrix} , \tag{2.12}$$

where I_k denotes the $k \times k$ identity matrix. Finally, define

$$\widehat{\Omega} = A_n \widehat{\Sigma} A_n^\top . \tag{2.13}$$

We are now ready to describe conditions that ensure strong consistency and asymptotic normality of $\hat{\boldsymbol{d}}$. The following theorem also provides consistent estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{d}}$ using BM method. Consistency of $\hat{\boldsymbol{d}}$ holds under minimal assumptions, i.e. if $\Phi_1, \ldots, \Phi_k$ are positive Harris chains. On the other hand, CLTs and consistency of BM estimator of asymptotic covariance require some mixing conditions on the Markov chains. For a square matrix C, let $C^\dagger$ denote the Moore-Penrose inverse of C.

Theorem 1. *Suppose that for each $l = 1, \ldots, k$, the Markov chain $\{X_1^{(l)}, X_2^{(l)}, \ldots\}$ has invariant distribution π_l.*

1. *If the Markov chains $\Phi_1, \ldots, \Phi_k$ are positive Harris, the log quasi-likelihood function (2.3) has a unique maximizer subject to the constraint $\boldsymbol{\zeta}^\top 1_k = 0$. Let $\hat{\boldsymbol{\zeta}}$ denote this maximizer, and let $\hat{\boldsymbol{d}} = g(\hat{\boldsymbol{\zeta}})$. Then $\hat{\boldsymbol{d}} \xrightarrow{a.s.} \boldsymbol{d}$ as $n_1, \ldots, n_k \to \infty$.*

2. *If the Markov chains $\Phi_1, \ldots, \Phi_k$ are polynomially ergodic of order $m > 1$, as $n_1, \ldots, n_k \to \infty$, $\sqrt{n}(\hat{\boldsymbol{d}} - \boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0, V)$ where $V = D^\top B^\dagger \Omega B^\dagger D$.*

3. *Assume that the Markov chains $\Phi_1, \ldots, \Phi_k$ are polynomially ergodic of order $m > 1$ and for all $l = 1, \ldots, k$, $b_l = \lfloor n_l^\nu \rfloor$ where $1 > \nu > 0$. Let $\widehat{D}$ be the matrix D in (2.5) with $\hat{\boldsymbol{d}}$ in place of $\boldsymbol{d}$, and $\widehat{B}$ and $\widehat{\Omega}$ are given by (2.9) and (2.13), respectively. Then, $\widehat{V} := \widehat{D}^\top \widehat{B}^\dagger \widehat{\Omega} \widehat{B}^\dagger \widehat{D}$ is a strongly consistent estimator of V.*

3 IS with multiple Markov chains

This section considers a CLT and SEs for the generalized IS estimator $\hat{\eta}$. From (1.3), $\hat{\eta} \equiv \hat{\eta}^{[f]}(\pi; \boldsymbol{a}, \boldsymbol{d}) = \hat{v}^{[f]}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d})/\hat{u}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d})$, where

$$\begin{aligned} \hat{v} \equiv \hat{v}^{[f]}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d}) &:= \sum_{l=1}^{k} \frac{a_l}{n_l} \sum_{i=1}^{n_l} v^{[f]}(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d}) \text{ and} \\ \hat{u} \equiv \hat{u}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d}) &:= \sum_{l=1}^{k} \frac{a_l}{n_l} \sum_{i=1}^{n_l} u(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d}) \end{aligned} \tag{3.1}$$

with

$$v^{[f]}(x; \boldsymbol{a}, \boldsymbol{d}) := f(x) u(x; \boldsymbol{a}, \boldsymbol{d}) \quad \text{and} \quad u(x; \boldsymbol{a}, \boldsymbol{d}) := \frac{\nu(x)}{\sum_{s=1}^{k} a_s \nu_s(x)/d_s}. \tag{3.2}$$

Here, $\hat{u}$ converges almost surely to

$$\sum_{l=1}^{k} a_l E_{\pi_l} u(X; \boldsymbol{a}, \boldsymbol{d}) = \int_{\mathsf{X}} \frac{\sum_{l=1}^{k} a_l \nu_l(x)/m_l}{\sum_{s=1}^{k} a_s \nu_s(x)/(m_s/m_1)} \nu(x)\, \mu(dx) = \frac{m}{m_1}, \tag{3.3}$$

as $n_1, \ldots, n_k \to \infty$. Thus $\hat{u}$ itself is a useful quantity as it consistently estimates the ratios of normalizing constants $\{u(\pi, \pi_1) \equiv m/m_1 | \pi \in \Pi\}$. Unlike the estimator $\hat{\boldsymbol{d}}$ in Section 2, $\hat{u}$ does not require a sample from each density $\pi \in \Pi$. Thus $\hat{u}$ is well suited for situations where one wants to estimate the ratios $u(\pi, \pi_1)$ for a very large number of π's based on samples from a small number of skeleton densities, say k. This method is particularly efficient when obtaining samples from the target distributions is computationally demanding and the distributions within Π are similar.

In the context of Bayesian analysis, let $\pi(x) = \mathrm{lik}(x)p(x)/m$ be the posterior density corresponding to the likelihood function $\mathrm{lik}(x)$ and prior $p(x)$ with normalizing constant m. In this case, $u(\pi, \pi_1)$ is the so-called Bayes factor between the two models, which is commonly used in model selection.

The estimators $\hat{u}$ and $\hat{v}$ in (3.1) depend on $\boldsymbol{d}$, which is generally unknown in practice. Here we consider a two-stage procedure for evaluating $\hat{u}$. In the 1st stage, $\boldsymbol{d}$ is estimated by its reverse logistic regression estimator $\hat{\boldsymbol{d}}$, described in Section 2, using Markov chains $\tilde{\Phi}_l \equiv \{\tilde{X}_i^l\}_{i=1}^{N_l}$ with stationary densities π_l, for $l = 1, \ldots, k$. Note the change of notation from Section 2 where we used n_l's to denote the length of the Markov chains. We use $\tilde{\Phi}_l$'s and N_l's to denote the stage 1 chains and their lengths, respectively. Once $\hat{\boldsymbol{d}}$ is formed, new MCMC samples $\Phi_l \equiv \{X_i^l\}_{i=1}^{n_l}, l = 1 \ldots, k$ are obtained and $u(\pi, \pi_1)(E_\pi f)$ is estimated using $\hat{u}(\pi, \pi_1; \boldsymbol{a}, \hat{\boldsymbol{d}})$ ($\hat{\eta}^{[f]}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}})$) based on these 2nd stage samples. Buta and Doss (2011) propose this two-stage method and quantify its benefits over the method where the same MCMC samples are used to estimate both $\boldsymbol{d}$ and $u(\pi, \pi_1)$.

3.1 Estimating ratios of normalizing constants

Before we state a CLT for $\hat{u}(\pi, \pi_1; \boldsymbol{a}, \hat{\boldsymbol{d}})$, we need some notation. Let

$$\tau_l^2(\pi; \boldsymbol{a}, \boldsymbol{d}) = \mathrm{Var}_{\pi_l}(u(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d})) + 2\sum_{g=1}^{\infty} \mathrm{Cov}_{\pi_l}(u(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d}), u(X_{1+g}^{(l)}; \boldsymbol{a}, \boldsymbol{d})) \tag{3.4}$$

and $\tau^2(\pi; \boldsymbol{a}, \boldsymbol{d}) = \sum_{l=1}^{k} (a_l^2/s_l)\tau_l^2(\pi; \boldsymbol{a}, \boldsymbol{d})$. Define $c(\pi; \boldsymbol{a}, \boldsymbol{d})$ as a vector of length $k-1$ with $(j-1)$th coordinate

$$[c(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1} = \frac{u(\pi, \pi_1)}{d_j^2} \int_{\mathsf{X}} \frac{a_j \nu_j(x)}{\sum_{s=1}^{k} a_s \nu_s(x)/d_s} \pi(x) dx \text{ for } j = 2, \ldots, k, \tag{3.5}$$

and $\hat{c}(\pi; \boldsymbol{a}, \boldsymbol{d})$ as a vector of length $k-1$ with $(j-1)$th coordinate

$$[\hat{c}(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1} = \sum_{l=1}^{k} \frac{1}{n_l} \sum_{i=1}^{n_l} \frac{a_j a_l \nu(X_i^{(l)}) \nu_j(X_i^{(l)})}{(\sum_{s=1}^{k} a_s \nu_s(X_i^{(l)})/d_s)^2 d_j^2} \text{ for } j = 2, \ldots, k. \tag{3.6}$$

Assuming $n_l = e_l b_l$, let

$$\hat{\tau}_l^2(\pi; \boldsymbol{a}, \boldsymbol{d}) = \frac{b_l}{e_l - 1} \sum_{m=0}^{e_l - 1} \left[\bar{u}_m(\boldsymbol{a}, \boldsymbol{d}) - \bar{\bar{u}}(\boldsymbol{a}, \boldsymbol{d})\right]^2, \tag{3.7}$$

where $\bar{u}_m(\boldsymbol{a}, \boldsymbol{d})$ is the average of the $(m+1)$st block $\{u(X^{(l)}_{mb_l+1}; \boldsymbol{a}, \boldsymbol{d}), \cdots, u(X^{(l)}_{(m+1)b_l}; \boldsymbol{a}, \boldsymbol{d})\}$, and $\bar{\bar{u}}(\boldsymbol{a}, \boldsymbol{d})$ is the overall average of $\{u(X^{(l)}_1; \boldsymbol{a}, \boldsymbol{d}), \cdots, u(X^{(l)}_{n_l}; \boldsymbol{a}, \boldsymbol{d})\}$. Here, b_l and e_l are the block sizes and the number of blocks, respectively. Finally let $\hat{\tau}^2(\pi; \boldsymbol{a}, \boldsymbol{d}) = \sum_{l=1}^k (a_l^2/s_l)\hat{\tau}_l^2(\pi; \boldsymbol{a}, \boldsymbol{d})$.

Theorem 2. *Suppose that for the stage 1 chains, conditions of Theorem 1 holds such that* $N^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0, V)$ *as* $N \equiv \sum_{l=1}^k N_l \to \infty$. *Suppose there exists* $q \in [0, \infty)$ *such that* $n/N \to q$ *where* $n = \sum_{l=1}^k n_l$ *is the total sample size for stage 2, and let* $n_l/n \to s_l$ *for* $l = 1, \cdots, k$.

1. *Assume that the stage 2 Markov chains* $\Phi_1, \ldots, \Phi_k$ *are polynomially ergodic of order* m, *and for some* $\delta > 0$ $E_{\pi_l}|u(X; \boldsymbol{a}, \boldsymbol{d})|^{2+\delta} < \infty$ *for each* $l = 1, \cdots, k$ *where* $m > 1 + 2/\delta$. *Then as* $n_1, \ldots, n_k \to \infty$,

$$\sqrt{n}(\hat{u}(\pi, \pi_1; \boldsymbol{a}, \hat{\boldsymbol{d}}) - u(\pi, \pi_1)) \xrightarrow{d} N(0, qc(\pi; \boldsymbol{a}, \boldsymbol{d})^\top V c(\pi; \boldsymbol{a}, \boldsymbol{d}) + \tau^2(\pi; \boldsymbol{a}, \boldsymbol{d})). \tag{3.8}$$

2. *Let* $\widehat{V}$ *be the consistent estimator of* V *given in Theorem 1 (3). Assume that the Markov chains* $\Phi_1, \ldots, \Phi_k$ *are polynomially ergodic of order* $m \geq (1+\epsilon)(1+2/\delta)$ *for some* $\epsilon, \delta > 0$ *such that* $E_{\pi_l}|u(X; \boldsymbol{a}, \boldsymbol{d})|^{4+\delta} < \infty$, *and for all* $l = 1, \ldots, k$, $b_l = \lfloor n_l^\nu \rfloor$ *where* $1 > \nu > 0$. *Then* $q\hat{c}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}})^\top \widehat{V} \hat{c}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) + \hat{\tau}^2(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}))$ *is a strongly consistent estimator of the asymptotic variance in* (3.8).

Note that the asymptotic variance in (3.8) has two components. The second term is the variance of $\hat{u}$ when $\boldsymbol{d}$ is known. The first term is the increase in the variance of $\hat{u}$ resulting from using $\hat{\boldsymbol{d}}$

instead of $\boldsymbol{d}$. Since we are interested in estimating $u(\pi, \pi_1)$ for a large number of π's and for every π, the computational time needed to calculate $\hat{u}$ in (3.1) is linear in the total sample size n, this cannot be very large. If generating MCMC samples is not computationally demanding, then long chains can be used in the 1st stage to obtain a precise estimate of $\boldsymbol{d}$, and thus greatly reduce the first term in the variance expression (3.8).

3.2 Estimation of expectations using generalized IS

This section discusses estimating SEs of the generalized IS estimator $\hat{\eta}$ given in (1.3). We use the following notation

$$\gamma_l^{11} \equiv \gamma_l^{11}(\pi; \boldsymbol{a}, \boldsymbol{d}) = \mathrm{Var}_{\pi_l}(v^{[f]}(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d})) + 2\sum_{g=1}^{\infty} \mathrm{Cov}_{\pi_l}(v^{[f]}(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d}), v^{[f]}(X_{1+g}^{(l)}; \boldsymbol{a}, \boldsymbol{d})),$$

$$\gamma_l^{12} \equiv \gamma_l^{12}(\pi; \boldsymbol{a}, \boldsymbol{d}) = \gamma_l^{21} \equiv \gamma_l^{21}(\pi; \boldsymbol{a}, \boldsymbol{d}) = \mathrm{Cov}_{\pi_l}(v^{[f]}(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d}), u(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d}))$$
$$+ \sum_{g=1}^{\infty} [\mathrm{Cov}_{\pi_l}(v^{[f]}(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d}), u(X_{1+g}^{(l)}; \boldsymbol{a}, \boldsymbol{d})) + \mathrm{Cov}_{\pi_l}(v^{[f]}(X_{1+g}^{(l)}; \boldsymbol{a}, \boldsymbol{d}), u(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d}))],$$

$$\gamma_l^{22} \equiv \gamma_l^{22}(\pi; \boldsymbol{a}, \boldsymbol{d}) = \mathrm{Var}_{\pi_l}(u(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d})) + 2\sum_{g=1}^{\infty} \mathrm{Cov}_{\pi_l}(u(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d}), u(X_{1+g}^{(l)}; \boldsymbol{a}, \boldsymbol{d})),$$

(note γ_l^{22} is the same as $\tau_l^2(\pi; \boldsymbol{a}, \boldsymbol{d})$ defined in (3.4)) and

$$\Gamma_l(\pi; \boldsymbol{a}, \boldsymbol{d}) = \begin{pmatrix} \gamma^{11} & \gamma^{12} \\ \gamma^{21} & \gamma^{22} \end{pmatrix}; \Gamma(\pi; \boldsymbol{a}, \boldsymbol{d}) = \sum_{l=1}^{k} \frac{a_l^2}{s_l} \Gamma_l(\pi; \boldsymbol{a}, \boldsymbol{d}). \tag{3.9}$$

Since $\hat{\eta}$ has the form of a ratio, to establish a CLT for it, we apply the Delta method on the function $h(x, y) = x/y,$ with $\nabla h(x, y) = (1/y, -x/y^2)'$. Let

$$\rho(\pi; \boldsymbol{a}, \boldsymbol{d}) = \nabla h(E_\pi f u(\pi, \pi_1), u(\pi, \pi_1))' \Gamma(\pi; \boldsymbol{a}, \boldsymbol{d}) \nabla h(E_\pi f u(\pi, \pi_1), u(\pi, \pi_1)), \tag{3.10}$$

$e(\pi;\boldsymbol{a},\boldsymbol{d})$ be a vector of length $k-1$ with $(j-1)$th coordinate

$$[e(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1} = \frac{a_j}{d_j^2}\int_{\mathsf{X}} \frac{[f(x)-E_\pi f]\nu_j(x)}{\sum_{s=1}^k a_s\nu_s(x)/d_s}\pi(x)dx, \quad j=2,\ldots,k, \tag{3.11}$$

and $\hat{e}(\pi;\boldsymbol{a},\boldsymbol{d})$ be a vector of length $k-1$ with $(j-1)$th coordinate

$$[\hat{e}(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1} \equiv \frac{\sum_{l=1}^k \frac{a_l}{n_l}\sum_{i=1}^{n_l}\frac{a_j f(X_i^{(l)})\nu(X_i^{(l)})\nu_j(X_i^{(l)})}{d_j^2(\sum_{s=1}^k a_s\nu_s(X_i^{(l)})/d_s)^2}}{\hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d})} - \frac{[c(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1}\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d})}{\hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d})}, \tag{3.12}$$

where $[c(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1}$ is defined in (3.6). Assuming $n_l = e_l b_l$, let

$$\begin{aligned}
\widehat{\Gamma}_l(\pi;\boldsymbol{a},\boldsymbol{d}) &= \frac{b_l}{e_l-1}\sum_{m=0}^{e_l-1}\left[\begin{pmatrix} \bar{v}_m^{[f]} \\ \bar{u}_m \end{pmatrix} - \begin{pmatrix} \bar{\bar{v}}^{[f]} \\ \bar{\bar{u}} \end{pmatrix}\right]\left[\begin{pmatrix} \bar{v}_m \\ \bar{u}_m \end{pmatrix} - \begin{pmatrix} \bar{\bar{v}}^{[f]} \\ \bar{\bar{u}} \end{pmatrix}\right]^\top \\
&= \frac{b_l}{e_l-1}\begin{pmatrix} \sum_{m=0}^{e_l-1}\left[\bar{v}_m^{[f]}-\bar{\bar{v}}^{[f]}\right]^2 & \sum_{m=0}^{e_l-1}\left[\bar{v}_m^{[f]}-\bar{\bar{v}}^{[f]}\right][\bar{u}_m-\bar{\bar{u}}] \\ \sum_{m=0}^{e_l-1}\left[\bar{v}_m^{[f]}-\bar{\bar{v}}^{[f]}\right][\bar{u}_m-\bar{\bar{u}}] & \sum_{m=0}^{e_l-1}[\bar{u}_m-\bar{\bar{u}}]^2 \end{pmatrix} \\
&= \begin{pmatrix} \hat{\gamma}^{11}(\pi;\boldsymbol{a},\boldsymbol{d}) & \hat{\gamma}^{12}(\pi;\boldsymbol{a},\boldsymbol{d}) \\ \hat{\gamma}^{21}(\pi;\boldsymbol{a},\boldsymbol{d}) & \hat{\gamma}^{22}(\pi;\boldsymbol{a},\boldsymbol{d}) \end{pmatrix},
\end{aligned}$$

where $\bar{v}_m^{[f]}$ is the average of the $(m+1)$st block $\{v^{[f]}(X^{(l)}_{mb_l+1};\boldsymbol{a},\boldsymbol{d}),\cdots,v^{[f]}(X^{(l)}_{(m+1)b_l};\boldsymbol{a},\boldsymbol{d})\}$, $\bar{\bar{v}}^{[f]}$ is the overall average of $\{v^{[f]}(X_1^{(l)};\boldsymbol{a},\boldsymbol{d}),\cdots,v^{[f]}(X_{n_l}^{(l)};\boldsymbol{a},\boldsymbol{d})\}$, and $\bar{u}_m \equiv \bar{u}_m(\pi,\boldsymbol{a},\boldsymbol{d}), \bar{\bar{u}} \equiv \bar{\bar{u}}(\pi,\boldsymbol{a},\boldsymbol{d})$ as defined in Section 3.1. Finally let $\widehat{\Gamma}(\pi;\boldsymbol{a},\boldsymbol{d}) = \sum_{l=1}^k (a_l^2/s_l)\widehat{\Gamma}_l(\pi;\boldsymbol{a},\boldsymbol{d})$, and

$$\hat{\rho}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}}) = \nabla h(\hat{v}^{[f]}(\hat{\boldsymbol{d}}),\hat{u}(\hat{\boldsymbol{d}}))'\widehat{\Gamma}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}})\nabla h(\hat{v}^{[f]}(\hat{\boldsymbol{d}}),\hat{u}(\hat{\boldsymbol{d}})).$$

Theorem 3. *Suppose that for the stage 1 chains, conditions of Theorem 1 hold such that* $N^{1/2}(\hat{\boldsymbol{d}}-\boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0,V)$ *as* $N \equiv \sum_{l=1}^k N_l \to \infty$. *Suppose there exists* $q \in [0,\infty)$ *such that* $n/N \to q$ *where* $n = \sum_{l=1}^k n_l$ *is the total sample size for stage 2, and let* $n_l/n \to s_l$ *for* $l = 1,\cdots,k$.

1. *Assume that the stage 2 Markov chains* $\Phi_1, \ldots, \Phi_k$ *are polynomially ergodic of order* m*, and for some* $\delta > 0$*,* $E_{\pi_l}|u(X; \boldsymbol{a}, \boldsymbol{d})|^{2+\delta} < \infty$ *and* $E_{\pi_l}|v^{[f]}(X; \boldsymbol{a}, \boldsymbol{d})|^{2+\delta} < \infty$*, for each* $l = 1, \cdots, k$ *where* $m > 1 + 2/\delta$*. Then as* $n_1, \ldots, n_k \to \infty$*,*

$$\sqrt{n}(\hat{\eta}^{[f]}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - E_\pi f) \xrightarrow{d} N(0, qe(\pi; \boldsymbol{a}, \boldsymbol{d})^\top V e(\pi; \boldsymbol{a}, \boldsymbol{d}) + \rho(\pi; \boldsymbol{a}, \boldsymbol{d})). \tag{3.13}$$

2. *Let* $\widehat{V}$ *be the consistent estimator of* V *given in Theorem 1 (3). Assume that the Markov chains* $\Phi_1, \ldots, \Phi_k$ *are polynomially ergodic of order* $m \geq (1+\epsilon)(1+2/\delta)$ *for some* $\epsilon, \delta > 0$ *such that* $E_{\pi_l}|u(X; \boldsymbol{a}, \boldsymbol{d})|^{4+\delta} < \infty, E_{\pi_l}|v^{[f]}(X; \boldsymbol{a}, \boldsymbol{d})|^{4+\delta} < \infty$*, and for each* $l = 1, \cdots, k$*,* $b_l = \lfloor n_l^\nu \rfloor$ *where* $1 > \nu > 0$*. Then* $q\hat{e}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}})^\top \widehat{V} \hat{e}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) + \hat{\rho}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}))$ *is a strongly consistent estimator of the asymptotic variance in* (3.13)*.*

Remark 2. Part (1) of Theorems 2 and 3 extend Buta and Doss's (2011) Theorems 1 and 3, respectively. Specifically, they require that $a_l = n_l/n$, which is a non-optimal choice for $\boldsymbol{a}$ (Tan et al. (2015)). Our results also substantially weaken the Markov chain mixing conditions.

Remark 3. Theorems 2 and 3 prove consistency of the BM estimators of the variances of $\hat{u}$ and $\hat{\eta}$ for a general $\boldsymbol{a}$. This extends results in Tan et al. (2015), which provides RS based estimators of the asymptotic variance of $\hat{u}$ and $\hat{\eta}$ in the special case when $\boldsymbol{a} = (1, \hat{\boldsymbol{d}})$. With this particular choice, $u(x; \boldsymbol{a}, \hat{\boldsymbol{d}})$ and $v^{[f]}(x; \boldsymbol{a}, \hat{\boldsymbol{d}})$ in (3.2) become free of $\hat{\boldsymbol{d}}$, leading to independence among certain quantities. However, one can set $\boldsymbol{a} = w * (1, \hat{\boldsymbol{d}})$ for any user-specified fixed vector w, which allows the expression in (2.18) of Tan et al. (2015) to be free of $\hat{\boldsymbol{d}}$ and thus the necessary independence. Hence, their RS estimator can also be applied to an arbitrary vector $\boldsymbol{a}$ (details are given in the supplement).

Remark 4. A sufficient condition for the moment assumptions for u in Theorems 2 and 3 is that, for any $\pi \in \Pi$, $\sup_x \left\{\pi(x)/\sum_{s=1}^k a_s \pi_s(x)\right\} < \infty$. That is, in any given direction, the tail of at least one of $\{\pi_s, s = 1, \ldots, k\}$ is heavier than that of π. This is not hard to achieve in practice by properly choosing $\{\pi_s\}$ with regard to Π (see e.g. Roy , 2014). Further, if $E_\pi |f|^{4+\delta} < \infty$, then the moment assumptions for $v^{[f]}$ are satisfied.

4 Toy example

Here, we confirm that both the BM and the RS estimators are consistent, as well as demonstrate the benefit of allowing general weights to be used in the generalized IS estimator. Let $t_{r,\mu}$ denote the t-distribution with degree of freedom r and central parameter μ. We set $\pi_1(\cdot) = \nu_1(\cdot)$ and $\pi_2(\cdot) = \nu_2(\cdot)$, the density functions for a $t_{5,\mu_1=1}$ and $t_{5,\mu_2=0}$, respectively. Pretending that we do not know the value of the ratio between the two normalizing constants, $d = m_2/m_1 = 1/1$, we estimate it by the stage 1 estimator $\hat{d}$ from Section 2, and compare the BM and the RS method in estimating the asymptotic variance. As for the stage 2 estimators from Section 3, the choice of weight and performance of the BM and the RS methods in assessing estimators' uncertainty are studied in the supplement.

We drew iid samples from π_1 and Markov chain samples from π_2 using the independent Metropolis Hastings algorithm with proposal density $t_{5,1}$. It is simple to show $\inf_x \frac{t_{5,\mu}(x)}{t_{5,0}(x)} > 0$, which implies the algorithm is uniformly ergodic (Mengersen and Tweedie (1996), Theorem 2.1) and hence polynomially ergodic and geometrically ergodic. For RS, our carefully tuned minorization condition enables the Markov chain for π_2 to regenerate about every 3 iterations. In contrast, the BM method proposed

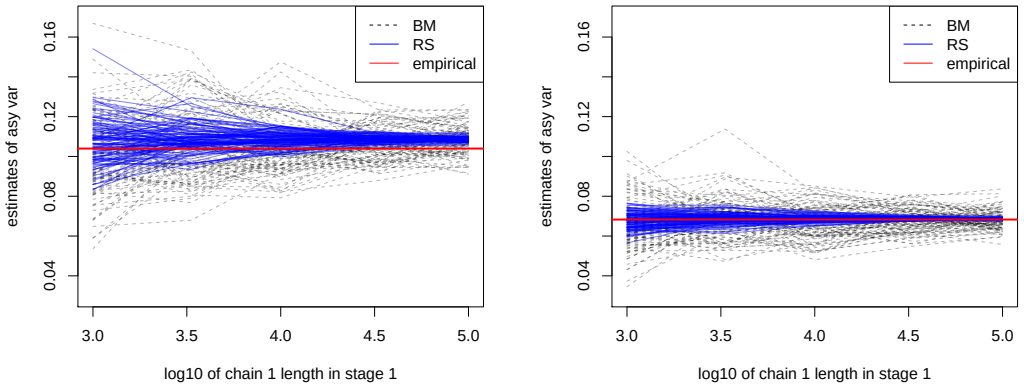

Figure 1: Plots of BM and RS estimates of the asymptotic variance of $\hat{d}$ in stage 1 for 100 randomly chosen replications. The left panel is based on the naive weight, $\boldsymbol{a}^{[1]} = (0.5, 0.5)$ and the right panel is based on a close-to-optimal weight, $\boldsymbol{a}^{[1]} = (0.82, 0.18)$. Horizontal lines represent the empirical asymptotic variance of $\hat{d}$ obtained over all replications.

here requires no such theoretical development.

We evaluated the variance estimators at various sample sizes with different choices of weight. Figure 1 displays traces of the BM and the RS estimates of the asymptotic variance of $\hat{d}$, in dashed and solid lines, respectively. Overall, the BM and the RS estimates approach the empirical asymptotic variance as the sample size increases, suggesting their consistency. Due to the frequency of regenerations, BM estimates are generally more variable than RS estimates. The left panel of Figure 1 is for estimators based on the naive weight, $\boldsymbol{a} = (0.5, 0.5)$, that is proportional to the sample sizes; the

right panel is for estimators based on $\boldsymbol{a} = (0.82, 0.18)$, that emphasizes the iid sample more than the Markov chain sample. Indeed, the latter weight is a close-to-optimal weight obtained with a small pilot study (see the supplement for details). Using such a method to choose weights can lead to big improvement in the efficiency of $\hat{d}$ when the mixing rate of the multiple samples differ a lot.

5 Bayesian spatial models for binary responses

In this section, we analyze a root rot disease dataset collected from a 90-acre farm in the state of Washington (Zhang (2002)). All computations were done in R, using the package geoBayes (Evangelou and Roy (2015)). Recorded at $M = 100$ chosen sites are the longitude and the latitude s_i, the root counts $\ell(s_i)$, and the number of infected roots $y(s_i), i = 1, \ldots, M$. Of interest is a map of the disease rate over the entire area for precision farming. We consider a spatial generalized linear mixed model (SGLMM), similar to that used by Zhang (2002) and Roy et al. (2016). Taking $\ell(s_i)$ and s_i as fixed, let

$$y(s_i)|z(s_i) \overset{\text{ind}}{\sim} \text{Binomial}\left(\ell(s_i), \Phi\left(z(s_i)\right)\right), i = 1, \ldots, M.$$

Here $\boldsymbol{z} = (z(s_1), \ldots, z(s_M))$ is a vector of latent variables, assumed to be a subvector of a Gaussian random field (GRF) $\{z_s, s \in S\}$, that has a constant mean μ, and a covariance function

$$\text{Cov}\left(z(s), z(s')\right) = \sigma^2 \rho_\phi(\|s - s'\|) + \omega\, \sigma^2 I_s(s')\,.$$

Here, σ^2 is the partial sill, $\|\cdot\|$ denotes the Euclidean distance, and ρ_ϕ is a correlation function from the spherical family with range parameter ϕ. That is, $\rho_\phi(u) = 1 - \frac{3}{2}\frac{u}{\phi} + \frac{1}{2}\left(\frac{u}{\phi}\right)^3$ for $u \in (0, \phi)$.

Next, $I_s(s')$ is an indicator that is 1 if $s = s'$, and 0 otherwise. Finally, $\omega\sigma^2$ is the nugget effect, accounting for any remaining variability at site s such as measurement error, while $\omega \in \mathcal{R}^+$ is the relative size of the nugget to the partial sill. Following Roy et al. (2016) we assign a non-informative Normal-inverse-Gamma prior to (μ, σ^2) which is (conditionally) conjugate for the model,

$$\mu|\sigma^2 \sim \mathrm{N}(0, 100\sigma^2), \text{ and } f(\sigma^2) \propto \left(\sigma^2\right)^{-\frac{1}{2}-1} \exp\left(-\frac{1}{2\sigma^2}\right).$$

Assigning priors for $h = (\phi, \omega)$ in the correlation function of the Gaussian random field is usually difficult, and the choice of prior influences the inference (Christensen (2004)). Hence we perform a sensitivity analysis focused on obtaining the Bayes factor (BF) of the model at h relative to a baseline h_0 for a range of values $h \in \mathcal{H}$. For a fixed $h = (\phi, \omega)$, this Bayesian model has parameters $\psi = (\mu, \sigma^2)$. Conditioning on the observed data $\boldsymbol{y} = (y(s_1), \ldots, y(s_M))$, inference is based on the posterior density

$$\pi_h(\psi|\boldsymbol{y}) = \frac{L_h(\psi|\boldsymbol{y})\pi(\psi)}{m_h(\boldsymbol{y})}, \tag{5.1}$$

where $L_h(\psi|\boldsymbol{y}) = \int_{\mathcal{R}^M} f(\boldsymbol{y}|\boldsymbol{z}) f_h(\boldsymbol{z}|\psi) d\boldsymbol{z}$ is the likelihood, $\pi(\psi)$ is the prior on ψ, and $m_h(\boldsymbol{y}) = \int_{\mathcal{R}\times\mathcal{R}_+} L_h(\psi|\boldsymbol{y})\pi(\psi)d\psi$ is the normalizing constant, also called the marginal likelihood. The BF between any two models indexed by h and h_0 is $m_h(\boldsymbol{y})/m_{h_0}(\boldsymbol{y})$, and the empirical Bayes choice of h is $\arg\max_{h\in\mathcal{H}} m_h(\boldsymbol{y}) = \arg\max_{h\in\mathcal{H}} [m_h(\boldsymbol{y})/m_{h_0}(\boldsymbol{y})]$. Our plan is to get MCMC samples for a small reference set of h, to estimate the BF among them using the reverse logistic estimator, and then get new samples to estimate $\{m_h(\boldsymbol{y})/m_{h'}(\boldsymbol{y}), h \in \mathcal{H}\}$ using the generalized IS method. Below, we describe the MCMC algorithms and the practical concern of how long to run them, which illustrates the importance of calculating a SE.

While the two high-dimensional integrals leave the posterior density in (5.1) intractable, there are MCMC algorithms to sample from the augmented posterior distribution,

$$\pi_h(\psi, \boldsymbol{z}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{z}) f_h(\boldsymbol{z}|\psi)\pi(\psi)}{m_h(\boldsymbol{y})}. \tag{5.2}$$

Note that $\int_{\mathcal{R}} \pi_h(\psi, \boldsymbol{z}|\boldsymbol{y}) d\boldsymbol{z} = \pi_h(\psi|\boldsymbol{y})$. Hence, a two-component Gibbs sampler that updates ψ and $\boldsymbol{z}$ in turn from their respective conditional distributions based on (5.2) yields a Markov chain $\{\psi^{(i)}, \boldsymbol{z}^{(i)}\}_{i\geq 1}$ with stationary distribution $\pi_h(\psi, \boldsymbol{z}|\boldsymbol{y})$. As a result, the marginal $\{\psi^{(i)}\}_{i\geq 1}$ is also a Markov chain with stationary distribution $\pi_h(\psi|\boldsymbol{y})$ (Tanner and Wong (1987)).

As a starting point, we used a small pilot study to identify a range for $h = (\phi, \omega)$ that corresponds to reasonably large BF values. This step was carried out by obtaining the reverse logistic estimator of BF at a coarse grid of h values over a wide area, based on short runs of Markov chains. Specifically, $(\phi, \omega) \in [80, 200] \times [0.2, 2]$ and, within this range, the minimum BF was about 1% the size of the maximum. To more carefully estimate BF over this range, we examined a fine grid $\mathcal{H}$ that consisted of 130 different h values, with increments of size 10 for the ϕ component, and that of .2 for the w component.

A natural choice for the set of skeleton points was $S = \{80, 140, 200\} \times \{0.5, 1, 2\}$, with an arbitrarily chosen baseline at $(200, 2)$. We first experimented with samples of sizes $n_1 = \cdots = n_9 = 500$ at the skeleton points (after a burn-in period of 500 iterations and a thinning procedure that kept one sample every 10 iterations), of which the first 80% were used in stage 1, and the remaining in stage 2 of the generalized IS procedure. BF estimates at all $h \in \mathcal{H}$ were obtained, though not shown. Given the current Monte Carlo sample sizes, it was natural to consider how trustworthy these BF

estimates were. The point-wise SEs at all $h \in \mathcal{H}$ were obtained via the BM method (not shown). In this setting for some h, the magnitude of the SE was about 6.6% of the corresponding BF estimate. This pilot stage took less than 8 seconds to compute on a 3.4GHz Intel Core i7 running linux.

Suppose it is desirable to reduce the relative SE to 1% or less for all $h \in \mathcal{H}$, we increased the sample sizes to $n_1 = \cdots = n_9 = 22,000$, approximately $(6.64\%/1\%)^2$ times the common pilot sample size. This new process took 8 minutes to run. The resulting BF estimates are shown in Figure 2, with maximum relative SE reduced to 0.96%. For the sake of comparison, we tried a few other designs that used different sets of skeleton points, including $S_4 = \{80, 200\} \times \{0.5, 2\}$, $S_{6a} = \{80, 200\} \times \{0.5, 1, 2\}$, $S_{6b} = \{80, 140, 200\} \times \{0.5, 2\}$, and $S_{12} = \{80, 140, 200\} \times \{0.5, 1, 1.5, 2\}$, all while keeping the baseline unchanged at $(200, 0.2)$. To achieve SEs at all $h \in \mathcal{H}$ below 1% of the corresponding BF estimates, it took sample sizes $55,000$, $35,000$, $32,000$, and $18,000$, for each simulated chain in these designs, respectively. The computing time for each turned out to be similar to that of design S and ranged from 8 to 10 minutes. In short, easily obtainable SE estimates allow us to experiment with different designs and perform samples size calculations in the pilot step, as well as providing reliable SE calculations for the final estimates.

The simplicity of the method matters when it comes to estimating SEs in practice. Using the BM method to obtain SE requires no extra input beyond what is needed for obtaining the generalized IS estimates. Indeed, as long as one can run existing software to obtain the Markov chain samples, there is no need to know the Markov transition kernels utilized in the background. Unlike the BM method, the RS method depends on identifying regeneration times, typically through constructing

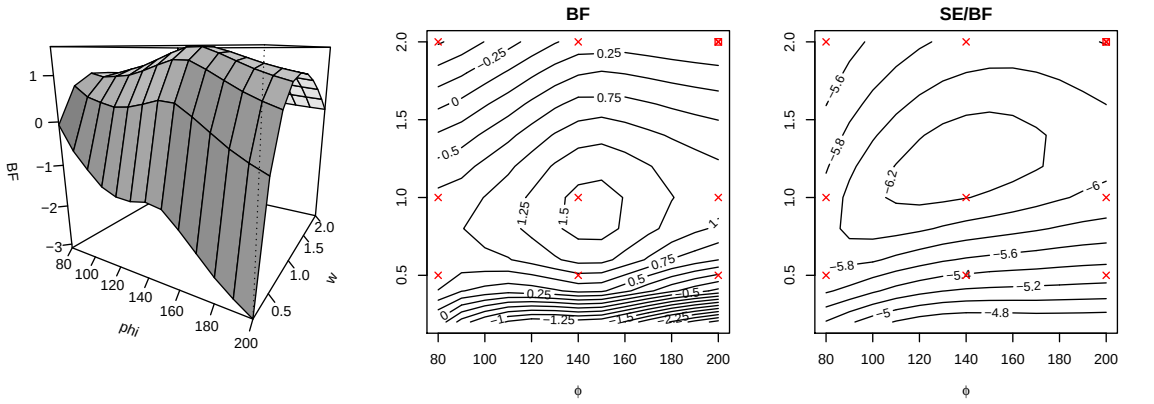

Figure 2: The left and middle panels display surface and contour plots for BF estimates in log scale (based on nine Markov chains with 22,000 iterations each). The right panel shows the ratio of the SE to the BF estimates in log scale, where SEs were evaluated using the BM method.

minorization conditions for the Markov transition kernels (see Mykland et al. (1995) for details). Despite the fact that minorization conditions can be established for any Markov transition kernel, we demonstrate that for the current example the amount of effort needed to obtain a regeneration can be prohibitively high. Recall the MCMC scheme involves sampling from $\pi_h(\psi|\boldsymbol{z},\boldsymbol{y})$ and $\pi_h(\boldsymbol{z}|\psi,\boldsymbol{y})$ in turn. The former is a standard distribution hence easy to sample from. The latter is not, and we followed Diggle et al. (1998) that updates $z_j, j = 1,\cdots,M$ in turn, each using a one-dimensional Metropolis-Hastings step that keeps invariant the conditional posterior distribution of z_j given all other components. Denote the transition density of these MH steps as $f_1,\cdots,f_M$, and suppressing

the notations of their dependence on $\boldsymbol{y}$, the transition kernel of the Markov chain can be represented as

$$p(\boldsymbol{z}', \psi' | \boldsymbol{z}, \psi) = f_1(z_1' | z_2, \cdots, z_M, \psi) f_2(z_2' | z_1', z_3, \cdots, z_M, \psi) \cdots f_n(z_M' | z_1', \cdots, z_{M-1}', \psi) \pi_h(\psi' | \boldsymbol{z}') \,.$$

According to a method described in Jones and Hobert (2004), one can build a minorization condition by finding $D \subset \mathbb{R}^M \times \mathbb{R} \times \mathbb{R}^+$, $\epsilon > 0$, and $k(\cdot)$ such that,

$$p(\boldsymbol{z}', \psi' | \boldsymbol{z}, \psi) \geq \epsilon I_D(\boldsymbol{z}, \psi)\, k(\boldsymbol{z}', \psi') \;\text{ for all } (\boldsymbol{z}', \psi') \in \mathbb{R}^M \times \mathbb{R} \times \mathbb{R}^+ \,.$$

Further, this condition can be established if

$$\begin{aligned} & f_1(z_1' | z_2, \cdots, z_M, \psi) f_2(z_2' | z_1', z_3, \cdots, z_M, \psi) \cdots f_M(z_M' | z_1', \cdots, z_{M-1}', \psi) \pi_h(\psi' | \boldsymbol{z}') \\ \geq & I_D(\boldsymbol{z}, \psi)\, \epsilon_1 k_1(z_1')\, \epsilon_2 k_2(z_1', z_2') \cdots \epsilon_M k_M(z_1', \cdots, z_M') \pi_h(\psi' | \boldsymbol{z}') \;\text{ for all } (\boldsymbol{z}', \psi') \in \mathbb{R}^M \times \mathbb{R} \times \mathbb{R}^+, \end{aligned}$$

where the common term π_h on both sides of the inequality cancel, and hence the work is in finding $\epsilon_1, \cdots, \epsilon_M$, and $k_1(\cdot), \cdots, k_M(\cdot)$. It's easy to see that the smaller the set D, the larger $\epsilon = \Pi_{i=1}^M \epsilon_j$ can possibly be, where ϵ can be interpreted as the conditional regeneration rate given D is visited. If we take D to be small enough such that ϵ_j takes on a very large value of 0.8 for each j, then the probability of getting a regeneration given a visit to D is $\epsilon = (0.8)^{100} \approx 2 \times 10^{-10}$. Being overoptimistic that the Markov chain visits D with probability close to 1, it would still take 100 billon iterations for the chain to regenerate about twenty times, barely enough for the RS method to be effective.

Using the EB estimate $\hat{h}$ of h, estimation of the remaining parameters ψ and prediction of the spatial random field can be done in the standard method using MCMC samples from $\pi_{\hat{h}}(\psi | \boldsymbol{y})$ (see

e.g. Roy et al. (2016), section 3.2).

Thus we can produce the root rot disease prediction map similar to that in Roy et al. (2016, Web Fig. 10).

6 Discussion

In this paper we consider two separate but related problems. The first problem is estimating the ratios of unknown normalizing constants given Markov chain samples from each of the $k > 1$ probability densities. The second problem is estimating expectations of a function with respect to a large number of probability distributions. These problems are related in the sense that generalized IS estimators used for the latter utilize estimates derived when solving the first problem. The first situation also arises in a variety of contexts other than the generalized IS estimators.

For both problems, we derive estimators with flexible weights and thus these estimators are appropriate for Markov chains with different mixing behaviors. We establish CLTs for these estimators and develop BM methods for consistently estimating their SEs. These easy-to-calculate SEs are important for at least three reasons. First, SEs are needed to assess the quality of the estimates. Second, our ability to calculate SEs allows us to search for optimal weights $\boldsymbol{a}$ for both stage 1 and 2. And last but not least, SEs form the basis for comparison of generalized IS with other available methods for estimating large number of (ratios of) normalizing constants.

Although we compare BM and RS in this paper, spectral estimators can also be derived for variance estimation using the results in Vats et al. (2016+). However, estimation by spectral methods

is generally more computationally expensive. Flegal and Jones (2010) compare the performance of confidence intervals produced by BM, RS, and spectral methods for the time average estimator, and they conclude that if tuning parameters are chosen appropriately, all three perform equally well. Control variates can be used to further improve the accuracy of our generalized IS estimators (Owen and Zhou (2000); Doss (2010)). A direction of future research would be to establish a BM estimator of the SEs for control variate based methods.

Supplementary Materials

The supplement to this paper contains proofs of Theorems 1 to 3, as well as a proof of the extension of the CLT based on regenerative simulation mentioned in Remark 3. Also included is a simulation study that demonstrates consistency of the BM and the RS estimators in stage 2 of the generalized IS estimators, as well as a comparison among three different weighting strategies. Finally, we study a linear regression model and use the BM estimator to aid the process of empirical Bayes variable selection.

Acknowledgements

The authors thank Hani Doss, James Hobert, and Galin Jones for helpful discussions, and the anonymous referee and associate editor for constructive comments which resulted in many improvements. The third author's work was partially supported by NSF grant DMS-13-08270.

References

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

Cappé, O. and Guillin, A. and Marin, J. M. and Robert, C. P.(2004). Population Monte Carlo. *J. Comput. Graph. Statist.* **13**, 907–929.

Christensen, O. F. (2004). Monte Carlo maximum likelihood in model based geostatistics. *J. Comput. Graph. Statist.* **13**, 702–718.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics. *Applied Statistics* **47**, 299–350.

Doss, H. (2010). Estimation of large families of Bayes factors from Markov chain output. *Statist. Sinica*, **20**, 537–560.

Doss, H. and Tan, A. (2014). Estimates and standard errors for ratios of normalizing constants from multiple Markov chains via regeneration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76**, 683–712.

Elvira, V. and Martino, L. and Luengo, D. and Bugallo, M. F. (2017). Generalized multiple importance sampling. *ArXiv e-prints*.

Evangelou, E. and Roy, V. (2015). *geoBayes*. http://cran.r-project.org/web/packages/geoBayes. R package version 0.3-3.

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

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.*, **38**, 1034–1070.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Tech. Rep. 568r, Department of Statistics, University of Minnesota.

Gilks, W. R., Roberts, G. O. and Sahu, S. K. (1998) Adaptive Markov chain Monte Carlo through regeneration. *J. Amer. Statist. Assoc.* **93**, 1045–1054.

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

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

Johns, G. and Hobert, J. (2004). Sufficient Burn-in for Gibbs Samplers for a Hierarchical Random Effects Model. *Ann. Statist.* **32**, 784–817.

Koehler, E. and Brown, E. and Haneuse, S. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *Amer. Statist.* **63**, 155–162.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D. and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65**, 585–618.

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

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

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

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.* **90**, 233–41.

Owen A. and Zhou, Y. (2000). Safe and effective importance sampling. *J. Amer. Statist. Assoc.* **95**, 135-143.

Roy, V. (2014) Efficient estimation of the link function parameter in a robust Bayesian binary regression model. *Comput. Stat. Data. Anal.*, **73**, 87–102.

Roy, V. and Evangelou, E. and Zhu, Z. (2016). Efficient estimation and prediction for the Bayesian binary spatial model with flexible link functions. *Biometrics* **72**, 289–298.

Roy, V. and Hobert, J. P. (2007) Convergence rates and asymptotic standard errors for MCMC algorithms for Bayesian probit regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69**, 607–623.

Tan, A. and Doss, H. and Hobert, J. P. (2015). Honest importance sampling with multiple Markov chains. *J. Comput. Graph. Statist.*, **24**, 792–826.

Tan, A. and Hobert, J. P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: convergence and regeneration. *J. Comput. Graph. Statist.* **18**, 861-878.

Tan, Z. (2004). On a likelihood approach for Monte Carlo integration. *J. Amer. Statist. Assoc.* **99**, 1027–1036.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation(with discussion). *J. Amer. Statist. Assoc.* **82**, 528–550.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178–203.

Vats, D., Flegal, J. M., and Jones, G. L.. (2016+). Strong consistency of the multivariate spectral variance estimator in Markov chain Monte Carlo. *Bernoulli, to appear.*

Veach, E. and Guibas, L. (1995). Optimally combining sampling techniques for Monte Carlo rendering. *SIGGRAPH 95 Conference Proceedings, Reading MA. Addison-Wesley*, 419-428.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics* **58**, 129–136.

Department of Statistics, Iowa State University

E-mail: vroy@iastate.edu

Department of Statistics and Actuarial Science, University of Iowa

E-mail: aixin-tan@uiowa.edu

Department of Statistics, University of California, Riverside

E-mail: jflegal@ucr.edu