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. Instead, we consider generalized importance sampling esti-
mators 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 poly-
nomial rate, then under two finite moment conditions, we show a central limit theorem holds for
the generalized estimators. Further, we develop an easy to implement method to calculate valid
asymptotic standard errors based on batch means. We also 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 esti-
mator 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 $f : X \to \mathbb{R}$ is a $\pi$ integrable function and we want to estimate $E_\pi f := \int_X f(x)\pi(x)\mu(dx)$. 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 $\pi_1$ is

$$\frac{\sum_{i=1}^n f(X_i)\nu(X_i)/\nu_1(X_i)}{\sum_{i=1}^n \nu(X_i)/\nu_1(X_i)} \overset{a.s.}{\to} \int_X \frac{f(x)\nu(x)/m}{\nu_1(x)/m_1} \pi_1(x)\mu(dx)/\int_X \frac{\nu(x)/m}{\nu_1(x)/m_1} \pi_1(x)\mu(dx) = E_\pi f,$$  \hfill (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 $\pi_1$.
does not depend on normalizing constants \( (\text{Hastings} \ (1970)) \). Note that (1.1) requires the functions \( \nu \) and \( \nu_1 \) 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 \( \pi \) belonging to a large collection, say \( \Pi \). 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 \( \pi_1 \), it does not work well when \( \pi \) differs greatly from \( \pi_1 \). 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 \( \pi_1 \) which is close to all \( \pi \in \Pi \) (see e.g. \text{Geyer} (1994)). Hence a natural modification is to replace \( \pi_1 \) in (1.1) with a mixture of densities where each density in \( \Pi \) is close to a subset of the \( k \) reference densities. To this end, denote \( \pi \equiv \sum_{i=1}^{k} (a_i/|a|) \pi_i \), where \( a = (a_1, \ldots, a_k) \) are \( k \) positive constants, \( |a| = \sum_{i=1}^{k} a_i \), 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

\[
\mathbf{d} = (m_2/m_1, \ldots, m_k/m_1).
\]

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

\[
\tilde{\eta} \equiv \left( \sum_{l=1}^{k} a_l \sum_{i=1}^{n_l} \frac{f(X_i^{(l)}) \nu(X_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_s(X_i^{(l)}/d_s)} \right) / \left( \sum_{l=1}^{k} a_l \sum_{i=1}^{n_l} \frac{\nu(X_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_s(X_i^{(l)}/d_s)} \right)
\]

\[
\Rightarrow \frac{\mathbb{I}_{a.s.}}{m_l} \frac{\sum_{l=1}^{k} a_l \int_X f(x) \nu(x) \pi_i(x) \mu(dx)}{\sum_{l=1}^{k} a_l \int_X \frac{\nu(x)}{\sum_{s=1}^{k} a_s \nu_s(x/d_s)} \pi_i(x) \mu(dx)} = \int_X f(x) \frac{\nu(x)}{\tilde{\pi}(x)} \tilde{\pi}(x) \mu(dx) = E_\pi f.
\]

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. \text{Gill et al.} (1988), \text{Kong et al.} (2003), \text{Meng and Wong} (1996), \text{Tan} (2004), and \text{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. \text{Buta and Doss} (2011), \text{Geyer} (1994), and \text{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 \text{Veach and Guibas} (1995), which is revisited in \text{Owen and Zhou} (2000) as the deterministic mixture. The standard population Monte Carlo algorithm of \text{Cappé et al.} (2004) uses a weighted ratio of the target \( \pi \)
and the proposal $\pi_j$ 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. (2015) have shown 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. 2015, p. 16), 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 given samples of finite size in practice, it is of fundamental importance to provide some measure of uncertainty, such as the standard errors (SEs) associated with this consistent estimator. Take 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 regeneration 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, $d$, which are unknown in practical applications. We consider the two-stage scheme studied in Buta and Doss (2011) where first an estimate $\hat{d}$ is obtained using Geyer’s (1994) “reverse logistic regression” method based on samples from $\pi_1$, and then independently, new samples are used to estimate $E_\pi f$ for $\pi \in \Pi$ using the estimator $\hat{\eta}(\hat{d})$ in (1.3). Buta and Doss (2011) showed that the asymptotic variance of $\hat{\eta}(\hat{d})$ depends on the
asymptotic variance of \( \hat{d} \). Thus we study the CLT of \( \hat{d} \) and provide a BM estimator of the asymptotic covariance matrix of \( \hat{d} \). Since \( \hat{d} \) involves multiple Markov chain samples, we utilize a multivariate BM estimator. Although, the form of the asymptotic covariance matrix of \( \hat{d} \) is complicated, our consistent BM estimator is straightforward to code.

The problem of estimating \( d \), 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.

Our work considers the problem of estimating \( d \) 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 which are appropriate in situations where the multiple Markov chains have different mixing rates. We establish the CLT for the resulting estimators of \( d \) and develop the BM estimators of their asymptotic covariance matrix.

Thus we consider two related problems in this paper–firstly, estimating (ratios of) normalizing constants given samples from \( k \) densities, and secondly, 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 substantially in that we only require the chains to be polynomially ergodic. To this end, let \( K_t(x, \cdot) \) be the Markov transition function for the Markov chain \( \Phi_t = \{X_t^{(l)}\}_{t \geq 1} \), so that for any measurable set \( A \), and \( s, t \in \{1, 2, \ldots\} \) we have \( P(X_{s+t}^{(l)} \in A \mid X_s^{(l)} = x) = K_t^s(x, A) \). Let \( \| \cdot \| \) denote the total variation norm and \( \Pi_t \) be the probability measure corresponding to the density \( \pi_t \). The Markov chain \( \Phi_t \) is polynomially ergodic of order \( m \) where \( m > 0 \) if there exists \( W : X \rightarrow \mathbb{R}^+ \) with \( E_{\pi_t} W < \infty \) such that

\[
\| K_t^s(x, \cdot) - \Pi_t(\cdot) \| \leq W(x)t^{-m}.
\]

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 important problem of estimating the ratios of normalizing constants of unnormalized densities, that is estimating $d$. 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. All 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 $\mu$, where the $\nu_l$’s are known functions and the $m_l$’s are unknown constants. For each $l$ we have a positive Harris Markov chain $\Phi_l = \{X_l^{(i)}, i = 1, \ldots, n_l\}$ with invariant density $\pi_l$. Our objective is to estimate all possible ratios $m_i/m_j, i \neq j$ or, equivalently, the vector $d$ 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 $\zeta_l$ by

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

and let

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

Given the value $x$ belongs to the pooled sample $\{X_l^{(i)}, i = 1, \ldots, n_l, l = 1, \ldots, k\}$, $p_l(x, \zeta_l)$ 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 $\zeta$ by maximizing the log quasi-likelihood function

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

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

As Geyer (1994) mentioned, there is a non-identifiability issue regarding $l_n(\zeta)$: for any constant $c \in \mathbb{R}$, $l_n(\zeta)$ is same as $l_n(\zeta + c1_k)$ where $1_k$ is the vector of $k$ 1’s. So we can estimate the true $\zeta$ only
up to an additive constant. Thus, we can estimate $m$ only up to an overall multiplicative constant, that is, we can estimate only $d$. Let $\zeta_0 \in \mathbb{R}^k$ be defined by $[\zeta_0]_l = [\zeta_l] - \left(\sum_{s=1}^{k} [\zeta_s]_l\right)/k$, the true $\zeta$ normalized to add to zero. Geyer (1994) proposed to estimate $\zeta_0$ by $\hat{\zeta}$, the maximizer of $l_n$ subject to the linear constraint $\zeta^\top 1_k = 0$, and thus obtain an estimate of $d$. The estimator $\hat{d}$ (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, $\hat{d}$ is consistent and asymptotically normal, and established its efficiency. Geyer (1994) proved the consistency and asymptotic normality of $\hat{d}$ when $\Phi_1, \ldots, \Phi_k$ 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.

However, none of these articles discuss how to consistently estimate the covariance matrix of $\hat{d}$, even in the iid setting. Recently Doss and Tan (2014) address this important issue and obtain a RS estimator of the covariance matrix of $\hat{d}$ in the Markov chain setting. Doss and Tan (2014) also mention 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 $\nu_j/m_j$ 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 following more general log quasi-likelihood function

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

where the vector $w \in \mathbb{R}^k$ is defined by $w_l = a_l n/l$ for $l = 1, \ldots, k$ for an arbitrary probability vector $a$. (Note the change of notation from $l$ to $\ell$.) 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 $\zeta_0$) and asymptotic normality of the constrained maximizer $\hat{\zeta}$ (subject to the constraint $\zeta^\top 1_k = 0$) 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 $a$ based on minimizing the trace of the estimated covariance matrix of $\hat{d}$. 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 $\hat{d}$ is a consistent estimator of $d$, show $\hat{d}$ satisfies a CLT under significantly weaker mixing conditions, and provide a strongly consistent BM estimator of the covariance matrix of $\hat{d}$.

### 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 $n_l/n \to s_l \in (0, 1)$. In order to obtain the CLT result for $\hat{d}$, we first establish a CLT for $\hat{\zeta}$. Note that the function $g: \mathbb{R}^k \to \mathbb{R}^{k-1}$
2.1 Central limit theorem and asymptotic covariance estimator

that maps $\zeta_0$ into $d$ is given by

$$g(\zeta) = \begin{pmatrix} e^{\zeta_1-\zeta_2/a_1} \\ e^{\zeta_1-\zeta_3/a_1} \\ \vdots \\ e^{\zeta_1-\zeta_k/a_1} \end{pmatrix},$$

and its gradient at $\zeta_0$ (in terms of $d$) is

$$D = \begin{pmatrix} d_2 & d_3 & \ldots & d_k \\ -d_2 & 0 & \ldots & 0 \\ 0 & -d_3 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & -d_k \end{pmatrix}.$$ (2.4)

Since $d = g(\zeta_0)$, and by definition $\hat{d} = g(\hat{\zeta})$, we can use the CLT result of $\hat{\zeta}$ to get a CLT for $\hat{d}$.

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

$$Y_i^{(r,l)} = p_r(X_i^{(l)}, \zeta_0) - E_{\pi_l}(p_r(X, \zeta_0)).$$ (2.6)

The asymptotic covariance matrix in the CLT of $\hat{\zeta}$, involves two $k \times k$ matrices $B$ and $\Omega$, which we now define. The matrix $B$ is given by

$$B_{rr} = \sum_{j=1}^{k} a_j E_{\pi_j}(p_r(X, \zeta) [1 - p_r(X, \zeta)])$$ and

$$B_{rs} = -\sum_{j=1}^{k} a_j E_{\pi_j}(p_r(X, \zeta)p_s(X, \zeta))$$ for $r \neq s$. (2.7)

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

$$\Omega_{rs} = \sum_{i=1}^{k} \frac{a_i^2}{s_i} \left[ E_{\pi_i} \{ Y_1^{(r,l)} Y_1^{(s,l)} \} + \sum_{i=1}^{\infty} E_{\pi_i} \{ Y_1^{(r,l)} Y_{1+i}^{(s,l)} \} + \sum_{i=1}^{\infty} E_{\pi_i} \{ Y_{1+i}^{(r,l)} Y_1^{(s,l)} \} \right].$$ (2.8)

**Remark 1.** The right hand side of (2.8) involves terms of the form $E_{\pi_i} \{ Y_1^{(r,l)} Y_{1+i}^{(s,l)} \}$ and $E_{\pi_i} \{ Y_{1+i}^{(r,l)} Y_1^{(s,l)} \}$.

For any fixed $l, r, s$ and $i$, the two expectations are the same if $X_i^{(l)}$ and $X_{1+i}^{(l)}$ are exchangeable, e.g. if the chain $\Phi_i$ is reversible. In general, the two expectations are not equal.

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

$$\hat{B}_{rr} = \sum_{i=1}^{k} a_i \left( \frac{1}{n_i} \sum_{i=1}^{n_i} p_r(X_i^{(l)}, \zeta_0) [1 - p_r(X_i^{(l)}, \zeta_0)] \right)$$ and

$$\hat{B}_{rs} = -\sum_{i=1}^{k} a_i \left( \frac{1}{n_i} \sum_{i=1}^{n_i} p_r(X_i^{(l)}, \zeta_0)p_s(X_i^{(l)}, \zeta_0) \right)$$ for $r \neq s$. (2.9)
To obtain a BM estimate $\hat{\Omega}$, suppose we simulate the Markov chain $\Phi_l$ 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 $r, l = 1, \ldots, k$
\[ Z^{(r,l)}_m := \frac{1}{b_l} \sum_{j=m b_l+1}^{(m+1)b_l} p_r(X^{(l)}_j, \hat{\zeta}) \quad \text{for } m = 0, \ldots, e_l - 1. \]

Now set $\tilde{Z}^{(l)}_m = \left( \tilde{Z}^{(1,l)}_m, \ldots, \tilde{Z}^{(k,l)}_m \right)^T$ for $m = 0, \ldots, e_l - 1$. For $l = 1, \ldots, k$, denote $\tilde{Z}^{(l)} = \left( \tilde{Z}^{(1,l)}, \ldots, \tilde{Z}^{(k,l)} \right)^T$ where $\tilde{Z}^{(r,l)} = \sum_{i=1}^{n_l} p_r(X^{(l)}_i, \hat{\zeta})/n_l$. Let
\[ \tilde{\Sigma}^{(l)} = \frac{b_l}{e_l - 1} \sum_{m=0}^{e_l - 1} \left[ \tilde{Z}^{(l)}_m - \overline{Z}^{(l)} \right] \left[ \tilde{Z}^{(l)}_m - \overline{Z}^{(l)} \right]^T \quad \text{for } l = 1, \ldots, k. \] (2.10)

Let
\[ \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\Sigma}^{(k)} \end{pmatrix} \] (2.11)
and define the following $k \times k^2$ matrix
\[ A_n = \left( -\sqrt{\frac{n}{n_1}} a_1 I_k \quad -\sqrt{\frac{n}{n_2}} a_2 I_k \quad \ldots \quad -\sqrt{\frac{n}{n_k}} a_k I_k \right), \] (2.12)
where $I_k$ denotes the $k \times k$ identity matrix. Finally, define
\[ \hat{\Omega} = A_n \hat{\Sigma} A_n^T. \] (2.13)

We are now ready to describe conditions that ensure strong consistency and asymptotic normality of $\hat{d}$. The following theorem also provides consistent estimate of the asymptotic covariance matrix of $\hat{d}$ using BM method. Consistency of $\hat{d}$ 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 $\{X^{(l)}_1, X^{(l)}_2, \ldots\}$ has invariant distribution $\pi_l$.

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 $\zeta^\top 1_k = 0$. Let $\hat{\zeta}$ denote this maximizer, and let $\hat{d} = g(\hat{\zeta})$. Then $\hat{d} \xrightarrow{a.s.} d$ 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 $n_1, \ldots, n_k \to \infty$, $\sqrt{n}(\hat{d} - d) \xrightarrow{d} N(0, V)$ 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 $l = 1, \ldots, k$, $b_l = \lfloor n_l^r \rfloor$ where $1 > \nu > 0$. Let $\tilde{D}$ be the matrix $D$ in (2.5) with $\tilde{d}$ in place of $d$, and let $\tilde{B}$ and $\tilde{\Omega}$ be defined by (2.9) and (2.13), respectively. Then, $\tilde{V} := \tilde{D}^\top \tilde{B}^\top \tilde{\Omega} \tilde{B}^\top \tilde{D}$ 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) we see that

$$\hat{\eta} \equiv \hat{\eta}^{[f]}(\pi; a, d) = \hat{v}^{[f]}(\pi, \pi_1; a, d)/\hat{u}(\pi, \pi_1; a, d),$$

where

$$\hat{v} \equiv \hat{v}^{[f]}(\pi, \pi_1; a, d) := \sum_{l=1}^k \frac{a_l}{n_l} \sum_{i=1}^{n_l} v^{[f]}(X_i^{(l)}; a, d)$$

and

$$\hat{u} \equiv \hat{u}(\pi, \pi_1; a, d) := \sum_{l=1}^k \frac{a_l}{n_l} \sum_{i=1}^{n_l} u(X_i^{(l)}; a, d)$$

with

$$v^{[f]}(x; a, d) := f(x)u(x; a, d) \quad \text{and} \quad u(x; a, d) := \frac{\nu(x)}{\sum_{s=1}^k a_s \nu_s(x)/d_s}.$$  \hspace{1cm} (3.1)

Note that $\hat{u}$ converges almost surely to

$$\sum_{l=1}^k a_l E_{\pi_1} u(X; a, d) = \int x \frac{\sum_{l=1}^k a_l \nu_l(x)/m_l}{\sum_{s=1}^k a_s \nu_s(x)/(m_s/m_1)} \nu(x) \mu(dx) = \frac{m}{m_1},$$

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 $\{u(\pi, \pi_1) = m/m_1 | \pi \in \Pi\}$. Unlike the estimator $\tilde{d}$ in Section 2, $\hat{d}$ 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 $\pi$’s based on samples from a small number of skeleton densities, say $k$. Thus this method is particularly efficient when obtaining samples from the target distributions is computationally demanding and the distributions within $\Pi$ are similar.

In the context of Bayesian analysis, let $\pi(x) = \text{lik}(x)p(x)/m$ be the posterior density corresponding to the likelihood function $\text{lik}(x)$ 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 $d$, which is generally unknown in practice. Here we consider a two-stage procedure for evaluating $\hat{u}$. In the 1st stage, $d$ is estimated by its reverse logistic regression estimator $\tilde{d}$ described in Section 2 using Markov chains $\tilde{\Phi}_l \equiv \{X_i^{(l)}\}_{i=1}^{N_l}$ with stationary densities $\pi_l$, 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. In order to avoid more notations, we use $\tilde{\Phi}_l$’s and $N_l$’s to denote the stage 1 chains and their length respectively. Once $d$ is formed, new MCMC samples $\Phi_l \equiv \{X_i^{(l)}\}_{i=1}^{n_l}$, $l = 1, \ldots, k$
are obtained and \(u(\pi, \pi_1)(E_x f)\) is estimated using \(\hat{u}(\pi, \pi_1; a, \hat{d})\) \((\hat{\eta}^{[f]}(\pi; a, \hat{d}))\) based on these 2nd stage samples. \[\text{Buta and Doss} (2011)\] propose this two-stage method and quantify its benefits over the method where the same MCMC samples are used to estimate both \(d\) and \(u(\pi, \pi_1)\).

### 3.1 Estimating ratios of normalizing constants

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

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

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

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

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

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

Assuming \(n_l = e_l b_l\), then let

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

where \(\tilde{u}_m(a, d)\) is the average of the \((m + 1)\)st block \(\{u(X_{m_b+1}^{(l)}; a, d), \ldots, u(X_{m+1}^{(l)}; a, d)\}\), and \(\tilde{u}(a, d)\) is the overall average of \(\{u(X_1^{(l)}; a, d), \ldots, u(X_{N_l}^{(l)}; a, d)\}\). Here, \(b_l\) and \(e_l\) are the block sizes and the number of blocks respectively. Finally let \(\tilde{\tau}^2(\pi; a, d) = \sum_{l=1}^{k} (a_l^2/s_l) \tilde{\tau}_l^2(\pi; a, d)\).

**Theorem 2.** Suppose that for the stage 1 chains, conditions of Theorem 1 holds such that \(N^{1/2}(\hat{d} - d) \rightarrow N(0, V)\) as \(N \equiv \sum_{l=1}^{k} n_l \rightarrow \infty\). Suppose there exists \(q \in (0, \infty)\) such that \(n/N \rightarrow q\) where \(n = \sum_{l=1}^{k} n_l\) is the total sample size for stage 2. In addition, let \(n_l/n \rightarrow s_l\) for \(l = 1, \ldots, k\).

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

\[
\sqrt{n}(\hat{u}(\pi, \pi_1; a, \hat{d}) - u(\pi, \pi_1)) \rightarrow N(0, q c(\pi; a, d) + \tau^2(\pi; a, d)). \tag{3.8}
\]

2. Let \(\hat{V}\) be the consistent estimator of \(V\) given in Theorem 1. 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 \( E_{\pi}|u(X; a, d)|^{1+\delta} < \infty \), and for all \( l = 1, \ldots, k \), \( b_l = |n^*_l| \) where \( 1 > \nu > 0 \). Then \( \hat{q}c(\pi; a, \hat{d})^\top \hat{V}c(\pi; a, \hat{d}) + \hat{\tau}^2(\pi; a, \hat{d}) \) 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 \( d \) is known. The first term is the increase in the variance of \( \hat{u} \) resulting from using \( \hat{d} \) instead of \( d \). Since we are interested in estimating \( u(\pi, \pi_1) \) for a large number of \( \pi \)’s and for every \( \pi \) the computational time needed to calculate \( \hat{u} \) in (3.1) is linear in the total sample size \( n \), this can not be very large. If generating MCMC samples is not computationally demanding, then long chains can be used in the 1st stage (that is, large \( N_l \)’s can be used) to obtain a precise estimate of \( d \), and thus greatly reducing 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 (3.3). In order to state a CLT for \( \hat{\eta} \) we define the following notations:

\[
\gamma_{l}^{11} \equiv \gamma_{l}^{11}(\pi; a, d) = \text{Var}_{\pi}(v^{[f]}(X_1^{(l)}; a, d)) + 2 \sum_{g=1}^{\infty} \text{Cov}_{\pi}(v^{[f]}(X_1^{(l)}; a, d), v^{[f]}(X_1^{(l)}_{1+g}; a, d)),
\]

\[
\gamma_{l}^{12} \equiv \gamma_{l}^{12}(\pi; a, d) = \text{Cov}_{\pi}(v^{[f]}(X_1^{(l)}; a, d), u(X_1^{(l)}; a, d))
\]

\[
+ \sum_{g=1}^{\infty} \text{[Cov}_{\pi}(v^{[f]}(X_1^{(l)}; a, d), u(X_1^{(l)}; a, d)) + \text{Cov}_{\pi}(v^{[f]}(X_1^{(l)}_{1+g}; a, d), u(X_1^{(l)}; a, d))],
\]

\[
\gamma_{l}^{22} \equiv \gamma_{l}^{22}(\pi; a, d) = \text{Var}_{\pi}(u(X_1^{(l)}; a, d)) + 2 \sum_{g=1}^{\infty} \text{Cov}_{\pi}(u(X_1^{(l)}; a, d), u(X_1^{(l)}_{1+g}; a, d)),
\]

(note \( \gamma_{l}^{22} \) is the same as \( \tau_{l}^2(\pi; a, d) \) defined in (3.4)) and

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

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

\[
\rho(\pi; a, d) = \nabla h(E_{\pi}f u(\pi, \pi_1), u(\pi, \pi_1))' \Gamma_l(\pi; a, d) \nabla h(E_{\pi}f u(\pi, \pi_1), u(\pi, \pi_1)), \tag{3.10}
\]

\( e(\pi; a, d) \) is a vector of length \( k - 1 \) with \((j - 1)\)th coordinate as

\[
[e(\pi; a, d)]_{j-1} = \frac{a_j}{a_j^2} \int x \frac{[f(x) - E_{\pi}f]v_j(x)}{\sum_{s=1}^{k} a_s \nu_s(x)/d_s} \pi(x)dx, \quad j = 2, \ldots, k, \tag{3.11}
\]
and $\hat{e}(\pi; a, d)$ is a vector of length $k - 1$ with $(j - 1)$th coordinate as
\[
[\hat{e}(\pi; a, d)]_{j-1} = \frac{\sum_{l=1}^d \sum_{i=1}^n a_{i,j}(X^{(i)}(l))\nu(X^{(i)}(l))}{\bar{u}(\pi, \pi_1; a, d)} - \frac{[c(\pi; a, d)]_{j-1}\hat{\gamma}^f(\pi; a, d)}{\bar{u}(\pi, \pi_1; a, d)},
\]
(3.12)
where $[c(\pi; a, d)]_{j-1}$ is defined in (3.6). Assuming $n_l = c_l b_l$ let
\[
\hat{\Gamma}_l(\pi; a, d) = \frac{b_l}{c_l - 1} \sum_{m=0}^{e_l-1} \left[ \begin{array}{c}
\bar{v}_m^f \\
\bar{u}_m
\end{array} \right] - \left[ \begin{array}{c}
\bar{v}_l^f \\
\bar{u}_l
\end{array} \right]
\]
\[
= \frac{b_l}{c_l - 1} \left( \begin{array}{cc}
\sum_{m=0}^{\epsilon_l-1} [v_m^f - \bar{v}_l^f]^2 & \sum_{m=0}^{\epsilon_l-1} [v_m^f - \bar{v}_l^f] [u_m - \bar{u}_l] \\
\sum_{m=0}^{\epsilon_l-1} [v_m^f - \bar{v}_l^f] [u_m - \bar{u}_l] & \sum_{m=0}^{\epsilon_l-1} [u_m - \bar{u}_l]^2
\end{array} \right)
\]
\[
= \left( \begin{array}{cc}
\bar{\gamma}^{11}(\pi; a, d) & \bar{\gamma}^{12}(\pi; a, d) \\
\bar{\gamma}^{21}(\pi; a, d) & \bar{\gamma}^{22}(\pi; a, d)
\end{array} \right),
\]
where $\bar{v}_m^f$ is the average of the $(m+1)$st block $\{v_l^f(X_{mb+1}^{(l)}; a, d), \ldots, v_l^f(X_{(m+1)b}^{(l)}; a, d)\}$, $\bar{v}_l^f$ is the overall average of $\{v_l^f(X_1^{(l)}; a, d), \ldots, v_l^f(X_{mb}^{(l)}; a, d)\}$ and $\bar{u}_m = \bar{u}(\pi, a, d), \bar{u} = \bar{u}(\pi, a, d)$ defined in Section 3.1 Finally let $\hat{\Gamma}(\pi; a, d) = \sum_{l=1}^N (a_l^2/s_l)\hat{\Gamma}_l(\pi; a, d)$, and
\[
\hat{\rho}(\pi; a, d) = \nabla h(\hat{v}_l^f(\hat{d}), \hat{u}^l(\hat{d}))\hat{\Gamma}(\pi; a, d)\nabla h(\hat{v}_l^f(\hat{d}), \hat{u}^l(\hat{d})).
\]

**Theorem 3.** Suppose that for the stage 1 chains, conditions of Theorem 1 hold such that $N^{1/2}(\hat{d} - d) \xrightarrow{d} \mathcal{N}(0, V)$ as $N = \sum_{l=1}^k N_l \to \infty$. 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. In addition, let $n_l/n \to s_l$ for $l = 1, \ldots, k$.

1. Assume that the stage 2 Markov chains $\Phi_1, \ldots, \Phi_k$ are polynomially ergodic of order $m$, and for some $\delta > 0$ $E_\pi [u(X; a, d)]^{2+\delta} < \infty$ and $E_\pi [v_l^f(X; a, d)]^{2+\delta} < \infty$, for each $l = 1, \ldots, k$ where $m > 1/2/\delta$. Then as $n_1, \ldots, n_k \to \infty$,
\[
\sqrt{n}[\hat{\gamma}^f(\pi; a, d) - E_\pi f] \xrightarrow{d} \mathcal{N}(0, qe(\pi; a, d)^T V e(\pi; a, d) + \rho(\pi; a, d)).
\]
(3.13)

2. Let $\tilde{V}$ be the consistent estimator of $V$ given in Theorem 2 (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 $E_\pi [u(X; a, d)]^{4+\delta} < \infty, E_\pi [v_l^f(X; a, d)]^{4+\delta} < \infty$, and for each $l = 1, \ldots, k$, $b_l = \lfloor n_l^\alpha \rfloor$ where $1 > \nu > 0$. Then $q\bar{e}(\pi; a, d)^T \tilde{V} \bar{e}(\pi; a, d) + \bar{\rho}(\pi; a, d))$ 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 $a_l = n_l/n$ which is a non-optimal choice for $a$ (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 $a$. 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 $a = (1, \hat{d})$. With this particular choice, $u(x; a, \hat{d})$ and $v^{[f]}(x; a, \hat{d})$ in (3.2) become free of $\hat{d}$ leading to independence among certain quantities. However, one can set $a = w^* (1, \hat{d})$ for any user specified fixed vector $w$, which allows the expression in (2.18) of Tan et al. (2015) to be free of $\hat{d}$ and thus the necessary independence. Hence, their RS estimator can also be applied to an arbitrary vector $a$ (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 $\pi$. This is not hard to achieve in practice by properly choosing $\{\pi_s\}$ with regard to $\Pi$ (see e.g. Roy, 2014). Further, if $E_x |f|^{4+\delta} < \infty$ then the moment assumptions for $v^{[f]}$ are satisfied.

4 Toy example

In this section, we employ a toy example to 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 $\mu$. We set $\pi_1(\cdot) = \nu_1(\cdot)$ and $\pi_2(\cdot) = \nu_2(\cdot)$, which are 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 draw iid samples from $\pi_1$ and Markov chain samples from $\pi_2$ using the independent Metropolis Hastings algorithm with proposal density $t_{5,1}$. It is simple to show $\inf_x t_{5,\mu}(x) > 0$, 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 $\pi_2$ to regenerate about every 3 iterations. In contrast, the BM method proposed here requires no such theoretical development.

We evaluated the variance estimators at various sample sizes with different choices of weight. Figure 4 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. Further, the left panel of Figure 1 is for estimators based on the naive weight, \( a = (0.5, 0.5) \), that is proportional to the sample sizes; and the right panel is for estimators based on \( a = (0.82, 0.18) \), 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 weight can lead to big improvement in the efficiency of \( \hat{d} \) if the mixing rate of the multiple samples differ a lot.

![Figure 1](image-url)  

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, \( a^{[1]} = (0.5, 0.5) \) and the right panel is based on a close-to-optimal weight, \( a^{[1]} = (0.82, 0.18) \). Horizontal lines represent the empirical asymptotic variance of \( \hat{d} \) obtained over all replications.

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 are 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 the following 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

\[
y(s_i)|z(s_i) \sim \text{Binomial} (\ell(s_i), \Phi (z(s_i))), i = 1, \ldots, M.
\]
Here \( z = (z(s_1), \ldots, z(s_M)) \) is a vector of latent variables, which is assumed to be a subvector of a Gaussian random field (GRF) \( \{z_s, s \in S\} \), that has a constant mean \( \mu \), and a covariance function

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

Here, \( \sigma^2 \) is the partial sill, \( \| \cdot \| \) denotes the Euclidean distance, and \( \rho_\phi \) is a correlation function from the spherical family with range parameter \( \phi \). That is, \( \rho_\phi(u) = 1 - \frac{3u^2}{2\phi^2} + \frac{1}{2} \left( \frac{u}{\phi} \right)^3 \) for \( u \in (0, \phi) \). Next, \( I_s(s') \) is an indicator that equals 1 if \( s = s' \), and equals 0 otherwise. Finally, \( \omega \sigma^2 \) is the nugget effect, accounting for any remaining variability at site \( s \) such as measurement error, while \( \omega \in \mathbb{R}^+ \) 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 \((\mu, \sigma^2)\) which is (conditionally) conjugate for the model. Specifically,

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

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

\[
\pi_h(\psi|y) = \frac{L_h(\psi|y)\pi(\psi)}{m_h(y)}, \quad (5.1)
\]

where \( L_h(\psi|y) = \int_{\mathbb{R}^M} f(y|z)f_h(z|\psi)dz \) is the likelihood, \( \pi(\psi) \) is the prior on \( \psi \), and \( m_h(y) = \int_{\mathbb{R} \times \mathbb{R}^+} L_h(\psi|y)\pi(\psi)d\psi \) is the normalizing constant, also called the marginal likelihood. The BF between any two models indexed by \( h \) and \( h_0 \) is \( m_h(y)/m_{h_0}(y) \), and the empirical Bayes choice of \( h \) is \( \arg \max_{h \in \mathcal{H}} m_h(y) = \arg \max_{h \in \mathcal{H}} [m_h(y)/m_{h_0}(y)] \). 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 \( \{m_h(y)/m_{h_0}(y), h \in \mathcal{H}\} \) 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.

The two high-dimensional integrals lend the posterior density in (5.1) intractable. But there exist MCMC algorithms to sample from the augmented posterior distribution, that is

\[
\pi_h(\psi, z|y) = \frac{f(y|z)f_h(z|\psi)\pi(\psi)}{m_h(y)}. \quad (5.2)
\]

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

As a starting point, we use a small pilot study to identify a range for \( h = (\phi, \omega) \) that corresponds to reasonably large BF values. This step is 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 is about 1% the size of the maximum.

To more carefully estimate BF over this range, we examine a fine grid \( \mathcal{H} \) that consists of 130 different \( h \) values, with increments of size 10 for the \( \phi \) component, and that of 0.2 for the \( \omega \) component.

A natural choice for the set of skeleton points is \( S = \{80, 140, 200\} \times \{0.5, 1, 2\} \), with an arbitrarily chosen baseline at \((200, 2)\). We first experiment 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 keeps one sample every 10 iterations), of which the first 80% are used in stage 1, and the remaining in stage 2 of the generalized IS procedure. BF estimates at all \( h \in \mathcal{H} \) are obtained, though not shown. Given the current Monte Carlo sample sizes, it is natural to consider how trustworthy these BF estimates are. The point-wise SEs at all \( h \in \mathcal{H} \) are obtained via the BM method (not shown). In this setting for some \( h \), the magnitude of the SE is about 6.6% of the corresponding BF estimate. This pilot stage takes 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 \( h \in \mathcal{H} \), then we increase the sample sizes to \( n_1 = \cdots = n_9 = 22,000 \), which are approximately \((6.64\%/1\%)^2\) times the common pilot sample size. This new process takes 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 use 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 turns out to be similar to that of design \( S \) and ranges 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 minorization
Figure 2: The left and middle panels display surface and contour plots for BF estimates in log scale (based on 9 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 are evaluated using the BM method.

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 for the current example the amount of effort needed to obtain a regeneration can be prohibitively high. Recall the MCMC scheme involves sampling from $\pi_h(\psi|z,y)$ and $\pi_h(z|\psi,y)$ 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 $y$, the transition kernel of the Markov chain can be represented as

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

According to a common 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,

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

Further, the above condition can be established if

$$f_1(z_1'|z_2, \cdots, z_M, \psi)f_2(z_2'|z_1, z_3, \cdots, z_M, \psi) \cdots f_M(z_M'|z_1', \cdots, z_{M-1}', \psi)\pi_h(\psi'|z') \geq I_D(z, \psi) \epsilon_1 k_1(z_1') \epsilon_2 k_2(z_2') \cdots \epsilon_M k_M(z_1', \cdots, z_M')\pi_h(\psi'|z') \text{ for all } (z', \psi') \in \mathbb{R}^M \times \mathbb{R} \times \mathbb{R}^+. $$
where the common term $\pi_\theta$ on both sides of the inequality will 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 $\epsilon = \prod_{i=1}^{M}\epsilon_j$ can possibly be, where $\epsilon$ can be interpreted as the conditional regeneration rate given $D$ is visited. Suppose we take $D$ to be small enough such that $\epsilon_j$ 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 billion 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 $\psi$ and prediction of the spatial random field can be done in the standard method using MCMC samples from $\pi_\theta(\psi|y)$ (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 $a$ 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 expensive computationally. Further, 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 these three methods 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 to Theorems 1 to 3, as well as a proof to 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 the 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


REFERENCES


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