6-25-2015

Estimating standard errors for importance sampling estimators with multiple Markov chains

Vivekananda Roy
Iowa State University, vroy@iastate.edu

Aixin Tan
University of Iowa

James M. Flegal
University of California, Riverside

Estimating standard errors for importance sampling estimators with multiple Markov chains

Abstract
The naive importance sampling estimator based on the samples from a single importance density can be extremely numerically unstable. We consider multiple distributions importance sampling estimators where samples from more than one probability distributions are combined to consistently estimate means with respect to given target distributions. These generalized importance sampling estimators provide more stable estimators than the naive importance sampling estimators. Importance sampling estimators can also be used in the Markov chain Monte Carlo (MCMC) context, that is, where iid samples are replaced with positive Harris Markov chains with invariant importance distributions. If these Markov chains converge to their respective target distributions at a geometric rate, then under two finite moment conditions a central limit theorem (CLT) holds for the importance sampling estimators. In order to calculate valid asymptotic standard errors, it is required to consistently estimate the asymptotic variance in the CLT. Recently Tan and Doss and Hobert (2015) developed an approach based on regenerative simulation for obtaining consistent estimators of the asymptotic variance. It is well-known that in practice it is often difficult to construct a useful minorization condition that is required in Tan and Doss and Hobert 's (2015) regenerative simulation method. We provide an alternative estimator for these standard errors based on the easy to implement batch means methods. The multi-chain importance sampling estimators depend on Geyer's (1994) reverse logistic estimator (of ratios of normalizing constants) which has wide applications, in its own right, in both frequentist and Bayesian inference. We also provide batch means estimator for calculating asymptotically valid standard errors of Geyer's (1994) reverse logistic estimator. We illustrate the method with an application in Bayesian variable selection in linear regression. In particular, the multi-chain importance sampling estimator is used to perform empirical Bayes variable selection and the batch means estimator is used to obtain standard errors in the large p situation where regenerative method is not applicable.

Keywords
Bayes factors, Geometric ergodicity, importance sampling, Markov chain Monte Carlo, ratios of normalizing constants, standard errors

Disciplines
Statistics and Probability

Estimating standard errors for importance sampling estimators with multiple Markov chains

Vivekananda Roy[1], Aixin Tan[2], and James M. Flegal[3]

[1]Department of Statistics, Iowa State University
[2]Department of Statistics and Actuarial Science, University of Iowa
[3]Department of Statistics, University of California, Riverside

June 25, 2015

Abstract

The naive importance sampling estimator based on the samples from a single importance density can be extremely numerically unstable. We consider multiple distributions importance sampling estimators where samples from more than one probability distributions are combined to consistently estimate means with respect to given target distributions. These generalized importance sampling estimators provide more stable estimators than the naive importance sampling estimators. Importance sampling estimators can also be used in the Markov chain Monte Carlo (MCMC) context, that is, where iid samples are replaced with *positive Harris* Markov chains with invariant importance distributions. If these Markov chains converge to their respective target distributions at a geometric rate, then under two finite moment conditions a central limit theorem (CLT) holds for the importance sampling estimators. In order to calculate valid asymptotic standard errors, it is required to consistently estimate the asymptotic variance in the CLT. Recently Tan and Doss and Hobert (2015) developed an approach based on regenerative simulation for obtaining consistent estimators of the asymptotic variance. It is well-known that in practice it is often difficult to construct a useful minorization condition that is required in Tan and Doss and Hobert 's (2015) regenerative simulation method. We provide an alternative estimator for these standard errors based on the easy to implement batch means methods. The multi-chain importance sampling estimators depend on Geyer's (1994) reverse logistic estimator (of ratios of normalizing constants) which has wide applications, in its own right, in both frequentist and Bayesian inference. We also provide batch means estimator for calculating asymptotically valid standard errors of Geyer's (1994) reverse logistic estimator. We illustrate the method with an application in Bayesian variable selection in linear regression. In particular, the multi-chain importance sampling estimator is used to perform empirical Bayes variable selection and the batch means estimator is used to obtain standard errors in the large p situation where regenerative method is not applicable.

Key words and phrases: Bayes factors, Geometric ergodicity, importance sampling, Markov chain Monte Carlo, ratios of normalizing constants, standard errors.

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 $\{x : \pi(x) = 0\} \subset \{x : \pi_1(x) = 0\}$. The importance sampling estimator of $E_\pi f$ based on iid samples $X_1, \ldots, X_n$ from the importance density π_1 is (see. e.g. Robert and Casella, 2004, chap. 3)

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

The above importance sampling estimator can also be used in the setting where the iid samples $\{X_i\}_{i=1}^n$ are substituted with realizations of a Markov chain, which is suitably irreducible and has π_1 as its stationary density (Hastings, 1970). Note that the estimator (1.1) requires the functions ν, ν_1 to be known. On the other hand, it does not depend on the normalizing constants m, m_1 which are generally unknown in all practical examples.

In this article we consider situations where one wants to estimate $E_\pi f$ for all π belonging to a large collection Π. As mentioned below, this situation arises in different problems, both in 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, (1.1) does not work well unless ν_1 puts appreciable mass under all $\pi \in \Pi$. Since otherwise the ratios $\nu(x)/\nu_1(x)$ can be arbitrarily large for some sample values making the estimator (1.1) unstable. Generally, there is not a single good importance sampling density π_1 which is "close" to all $\pi \in \Pi$ (see e.g. Geyer, 1994). In this case a natural modification to the estimator (1.1) is to replace π_1 in (1.1) with a mixture of k appropriately chosen "scattered" densities, $\overline{\pi} \equiv \sum_{i=1}^k (a_i/|\boldsymbol{a}|)\pi_i$, where $\boldsymbol{a} = (a_1, a_2, \ldots, a_k)$ are k positive constants, $|\boldsymbol{a}| = \sum_{i=1}^k a_i$, and $\pi_i(x) = \nu_i(x)/m_i, i = 1, 2, \ldots, k$ are k densities generally known upto normalizing constants. Suppose $n_1, n_2, \ldots, n_k$ are positive integers, $d_i = m_i/m_1, i = 2, \ldots, k$, with $d_1 \equiv 1$, and the $(k-1)$ dimensional vector

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

Let $\{X_i^{(l)}\}_{i=1}^{n_l}$ be iid sample from π_l or a positive Harris Markov chain with invariant density $\pi_l, l = 1, 2, \ldots, k$. (See Meyn and Tweedie, 1993, chap. 10 for the definition of positive Harris Markov chain.) Then as $n_l \to \infty$, for all $l = 1, 2, \ldots, k$, we have

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

The above multiple samples based estimator has been discussed in the literature before. Vardi (1985), Gill, Vardi and Wellner (1988), Meng and Wong (1996), Kong et al. (2003), and Tan

(2004) consider estimation based on iid samples. The estimator is applicable to a much larger class of problems if Markov chain samples are allowed. Geyer (1994), Buta and Doss (2011) and Tan and Doss and Hobert (2015) study the case when only Markov chain samples are available from the importance densities and this is the setting we work in this paper. The multi-samples importance sampling estimator has many applications including Monte Carlo maximum likelihood estimation and Bayesian sensitivity analysis. In Section 4, we illustrate our methodology with an example of Bayesian sensitivity analysis.

As noted above, the estimator $\hat{\eta}$ is consistent in both settings: iid samples as well as Markov chain samples satisfying the usual regularity conditions. In practice, it is important to provide valid standard errors associated with the Monte Carlo estimate $\hat{\eta}$. Except for the recent work of Tan and Doss and Hobert (2015), other authors largely avoid this very important issue. Tan and Doss and Hobert (2015) provide a way of calculating standard errors of $\hat{\eta}$ using the method of regeneration. The success of this method crucially depends on the construction of an appropriate *minorization condition*. (See Mykland, Tierney and Yu (1995) for definition of minorization condition as well as the description of regeneration method.) As otherwise infrequent regenerations render this method practically useless. Successful applications of the regeneration method for calculating standard errors is problem specific and involves a great deal of trial and error (see e.g. Tan and Hobert (2009) and Roy and Hobert (2007)). In the present paper we provide standard error estimators of $\hat{\eta}$ using easy-to-use batch means method. Our batch means estimator is straightforward to implement and hence can be routinely applied in practice. In the process we also establish central limit theorem (CLT) for $\hat{\eta}$ generalizing some results in Buta and Doss (2011).

The estimator $\hat{\eta}$ in (1.3) depends on the vector $\boldsymbol{d}$ of ratios of normalizing constants which are unknown in all practical applications. We consider the two-stage scheme studied in Buta and Doss (2011) where first an estimate $\hat{\boldsymbol{d}}$ of $\boldsymbol{d}$ is obtained using Geyer's (1994) "reverse logistic regression" method based on samples from $\pi_l, l = 1, 2, \ldots, k$, and then independently of the first stage, *new* samples are used to estimate $E_\pi f, \pi \in \Pi$ using the estimator $\hat{\eta}(\hat{\boldsymbol{d}})$ in (1.3) with $\hat{\boldsymbol{d}}$ substituted for $\boldsymbol{d}$. 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 batch means estimator of the asymptotic variance covariance matrix of $\hat{\boldsymbol{d}}$. Since $\hat{\boldsymbol{d}}$ involves multiple Markov chain samples, here we need to use multivariate batch means estimator. Although, the form of the asymptotic variance covariance matrix of $\hat{\boldsymbol{d}}$ is complicated, our consistent batch means estimator is straightforward to code. Recently Doss and Tan (2014) provide an estimator of the variance covariance matrix of $\hat{\boldsymbol{d}}$ using the method of regenerations, which, as mentioned before, may be difficult to implement in practice.

The problem of estimating $\boldsymbol{d}$, the ratios of normalizing constants of unnormalized densities is important in its own right and has many other applications both in frequentist and Bayesian inference. When the samples are iid sequences, this is the biased sampling problem studied in Vardi (1985). Here are three instances where the problem of estimating ratios of normalizing constants arises naturally– calculation of likelihood ratios in missing data (or latent variable) models, calculation of mixture densities for use in the importance sampling as mentioned before and finally in calculation of Bayes factors which has many applications including the hyperparameter selection problems in Bayesian analysis. We devote an entire section (Section 2) in this paper on this important problem of estimating ratios normalizing constants.

The rest of the paper is organized as follows. In Section 2, we consider the problem of estimating $\boldsymbol{d}$ using Geyer's (1994) reverse logistic regression method. In fact, 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 is appropriate in situations where the multiple Markov chains have different mixing rates. We establish the CLT for the resulting estimators of $\boldsymbol{d}$ and develop the batch means estimators of their asymptotic covariance matrix. Section 3 contains the construction of CLT for $\hat{\eta}$. In this section, we also describe how valid standard errors of $\hat{\eta}$ can be obtained using the batch means method. These easy to compute standard errors of $\hat{\eta}$, developed in Section 3, has been recently used in Roy and Evangelou and Zhu (2014) for choosing the skeleton points in the importance sampling estimator (1.3). Section 4 contains a toy example showing the benefits of different weight functions. In Section 4 we also consider a standard linear regression model with moderately large number of variables and use the batch means estimator developed here for empirical Bayes variable selection. The proofs of the theorems are relegated to the appendix.

2 Estimating ratios of normalizing constants in the Markov chain setting

In this section we consider the problem of estimating ratios of normalizing constants. In particular, we have 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 (1994) proposed a method of estimating $\boldsymbol{d}$, which he called the "reverse logistic regression". We now describe the method. Let $n = \sum n_l$ and set $a_l = n_l/n$ for the moment. Define the vector $\boldsymbol{\zeta}$ by

$$\zeta_l = -\log(m_l) + \log(a_l), \qquad \text{for } l = 1, \ldots, k, \tag{2.1}$$

and let

$$p_l(x, \boldsymbol{\zeta}) = \frac{\nu_l(x)e^{\zeta_l}}{\sum_{s=1}^{k} \nu_s(x)e^{\zeta_s}}, \qquad \text{for } l = 1, \ldots, k. \tag{2.2}$$

Given that 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} sample. 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

$$l_n(\boldsymbol{\zeta}) = \sum_{l=1}^{k} \sum_{i=1}^{n_l} \log\bigl(p_l(X_i^{(l)}, \boldsymbol{\zeta})\bigr) \tag{2.3}$$

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) mentions, 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 $[\boldsymbol{\zeta}_0]_l = [\zeta]_l - \big(\sum_{s=1}^k [\zeta]_s\big)/k$, that is, $\boldsymbol{\zeta}_0$ is 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 optimality properties. Later Geyer (1994) proved the consistency and asymptotic normality of $\hat{\boldsymbol{d}}$ in the more general setting where $\Phi_l, \Phi_2, \ldots, \Phi_k$ 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, although using different computational schemes. None of the papers mentioned above discusses how to consistently estimate the variance covariance matrix of $\hat{\boldsymbol{d}}$ even in the iid setting. Only recently Doss and Tan (2014) address this important issue and, as mentioned in the introduction, obtain a regeneration based estimator of the covariance matrix of $\hat{\boldsymbol{d}}$ in the Markov chain setting. Doss and Tan (2014) mention that the optimality results of Gill et al. (1988) does 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 (2.2) is no more optimal and should instead incorporate the "effective sample size" of different chains as they might have quite different rates of mixing. Doss and Tan (2014) introduce the following more general log quasi-likelihood function

$$\ell_n(\boldsymbol{\zeta}) = \sum_{l=1}^{k} w_l \sum_{i=1}^{n_l} \log\big(p_l(X_i^{(l)}, \boldsymbol{\zeta})\big), \tag{2.4}$$

where the vector $w \in \mathbb{R}^k$ is defined by

$$w_l = a_l \frac{n}{n_l}, \qquad l = 1, \ldots, k, \tag{2.5}$$

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 (2.4) becomes (2.3). In the setting where the regeneration method 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 (2.4). They also obtain a regeneration based estimator of the asymptotic covariance matrix. They describe an empirical method for choosing the optimal $\boldsymbol{a}$ based on minimizing the trace of the estimated covariance matrix of $\hat{\boldsymbol{d}}$. A crucial assumption in Doss and Tan's (2014) results is the minorization condition, that is, in order to implement their regenerative simulation method, it is needed to construct a practically useful minorization condition for each of the k Markov chains Φ_l, $\Phi_2, \ldots \Phi_k$, which is extremely difficult in practice. In the next section, without assuming such minorization condition, we show that $\hat{\boldsymbol{d}}$ is a consistent estimator of $\boldsymbol{d}$ and also satisfies a CLT. We also provide a batch means estimator of the covariance matrix of $\hat{\boldsymbol{d}}$.

2.1 Central limit theorem and covariance estimation using batch means

In this section we discuss central limit theorems of the estimate $\hat{\boldsymbol{\zeta}}$, the maximizer of (2.4). This in turn provides a CLT for $\hat{\boldsymbol{d}}$, the estimate of the ratios of the normalizing constants $\boldsymbol{d}$ defined in

(1.2). We also construct consistent batch means estimator of the variance covariance matrix in the CLT thus leading to asymptotically valid standard errors for $\hat{\boldsymbol{d}}$. Since $\hat{\boldsymbol{\zeta}}$ involves multiple Markov chains, we need to use multivariate batch means estimator to calculate its standard errors. We assume that $n_1, \ldots, n_k \to \infty$ in such a way that $n_l/n \to s_l \in (0,1)$, for $l = 1, \ldots, k$. In order 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

$$g(\boldsymbol{\zeta}) = \begin{pmatrix} e^{\zeta_1 - \zeta_2} a_2/a_1 \\ e^{\zeta_1 - \zeta_3} a_3/a_1 \\ \vdots \\ e^{\zeta_1 - \zeta_k} a_k/a_1 \end{pmatrix}, \tag{2.6}$$

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

$$D = \begin{pmatrix} d_2 & d_3 & \ldots & d_k \\ -d_2 & 0 & \ldots & 0 \\ 0 & -d_3 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & -d_k \end{pmatrix}. \tag{2.7}$$

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 $\hat{\boldsymbol{d}}$. In order to state the CLT of $\hat{\boldsymbol{\zeta}}$, we introduce the following notations.

For $r = 1, 2, \ldots, k$, let

$$Y_i^{(r,l)} = p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big), \qquad i = 1, \ldots, n_l. \tag{2.8}$$

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

$$\begin{aligned} B_{rr} &= \sum_{j=1}^{k} a_j E_{\pi_j}\big(p_r(X, \boldsymbol{\zeta})[1 - p_r(X, \boldsymbol{\zeta})]\big), \qquad r = 1, \ldots, k, \\ B_{rs} &= -\sum_{j=1}^{k} a_j E_{\pi_j}\big(p_r(X, \boldsymbol{\zeta}) p_s(X, \boldsymbol{\zeta})\big), \qquad r, s = 1, \ldots, k,\ r \neq s. \end{aligned} \tag{2.9}$$

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

$$\Omega_{rs} = \sum_{l=1}^{k} \frac{a_l^2}{s_l} \Big[E_{\pi_l}\{Y_1^{(r,l)} Y_1^{(s,l)}\} + \sum_{i=1}^{\infty} E_{\pi_l}\{Y_1^{(r,l)} Y_{1+i}^{(s,l)}\} + \sum_{i=1}^{\infty} E_{\pi_l}\{Y_{1+i}^{(r,l)} Y_1^{(s,l)}\} \Big], \quad r, s = 1, \ldots, k. \tag{2.10}$$

Remark 1. Note that the right hand side of (2.10) 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, which happens if $\Phi_l = \{X_i^{(l)}\}_{i=1}^{n_l}$ is a reversible chain. But in general cases where the chain Φ_l is not reversible for some l, the two expectations are not necessarily the same.

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

$$\begin{aligned}\widehat{B}_{rr} &= \sum_{l=1}^{k} a_l \left(\frac{1}{n_l} \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \big[1 - p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \big] \right), \qquad r = 1, \ldots, k, \\ \widehat{B}_{rs} &= -\sum_{l=1}^{k} a_l \left(\frac{1}{n_l} \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) p_s(X_i^{(l)}, \hat{\boldsymbol{\zeta}}) \right), \qquad r, s = 1, \ldots, k,\ r \neq s. \end{aligned} \tag{2.11}$$

To obtain a batch means 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 $r, l = 1, \ldots, k$

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

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 $l = 1, 2, \ldots, k$, denote $\bar{\bar{Z}}^{(l)} = \left(\bar{\bar{Z}}^{(1,l)}, \bar{\bar{Z}}^{(2,l)}, \ldots, \bar{\bar{Z}}^{(k,l)} \right)^\top$ where $\bar{\bar{Z}}^{(r,l)} = \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \hat{\boldsymbol{\zeta}})/n_l$. Let

$$\widehat{\Sigma}^{(l)} = \frac{b_l}{e_l - 1} \sum_{m=0}^{e_l - 1} \left[\bar{Z}_m^{(l)} - \bar{\bar{Z}}^{(l)} \right] \left[\bar{Z}_m^{(l)} - \bar{\bar{Z}}^{(l)} \right]^T \text{ for } l = 1, 2, \ldots, k. \tag{2.12}$$

Finally, let

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}^{(1)} & & & \\ & \widehat{\Sigma}^{(2)} & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & \widehat{\Sigma}^{(k)} \end{pmatrix} . \tag{2.13}$$

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

$$A_n = \left(-\sqrt{\frac{n}{n_1}} a_1 I_k \quad -\sqrt{\frac{n}{n_2}} a_2 I_k \quad \ldots \quad -\sqrt{\frac{n}{n_k}} a_k I_k \right) , \tag{2.14}$$

where I_k denotes the $k \times k$ identity matrix. Define

$$\widehat{\Omega} = A_n \widehat{\Sigma} A_n^\top. \tag{2.15}$$

We are now ready to state the following theorem which describes the strong consistency, and asymptotic normality of $\hat{\boldsymbol{d}}$. This theorem also provides consistent estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{d}}$ using batch means method. As mentioned in Doss and Tan (2014), the consistency of $\hat{\boldsymbol{d}}$ holds under minimal assumptions on the Markov chains $\Phi_1, \ldots, \Phi_k$. In particular, if $\Phi_1, \ldots, \Phi_k$ are positive Harris chains then $\hat{\boldsymbol{d}}$ is a consistent estimator of $\boldsymbol{d}$. On the other hand, CLTs and consistency of batch means estimator of asymptotic covariance require some mixing conditions on the Markov chains, and the most commonly used condition is that of geometric ergodicity of the 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 (2.4) 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 $\hat{\boldsymbol{d}} \xrightarrow{\text{a.s.}} \boldsymbol{d}$ as $n_1, \ldots, n_k \to \infty$.

(2) If the Markov chains $\Phi_1, \ldots, \Phi_k$ are geometrically ergodic, as $n_1, \ldots, n_k \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{d}} - \boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0, V) \quad \text{where} \quad V = D^\top B^\dagger \Omega B^\dagger D. \tag{2.16}$$

(3) Assume that the Markov chains $\Phi_1, \ldots, \Phi_k$ are geometrically ergodic and for all $l = 1, 2, \ldots, k$, $b_l = \lfloor n_l^\nu \rfloor$ where $1 > \nu > 0$. Let $\widehat{D}$ be the matrix D in (2.7) with $\hat{\boldsymbol{d}}$ in place of $\boldsymbol{d}$, and let $\widehat{B}$ and $\widehat{\Omega}$ be defined by (2.11) and (2.15), respectively. Then, $\widehat{V} := \widehat{D}^\top \widehat{B}^\dagger \widehat{\Omega} \widehat{B}^\dagger \widehat{D}$ is a strongly consistent estimator of V.

The proof of Theorem 1 is given in Appendix A.

Remark 2. Since Theorem 1 does not require the chains $\Phi_l, l = 1 \ldots, k$ to be stationary, there is no need for burnin.

3 Importance sampling with multiple Markov chains

In this section we consider CLT and estimation of standard errors for the multi-chain importance sampling estimator $\hat{\eta}$ given in the Introduction.

From (1.3) we see that $\hat{\eta} \equiv \hat{\eta}^{[f]}(\pi; \boldsymbol{a}, \boldsymbol{d}) = \hat{v}^{[f]}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d}) / \hat{u}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d})$, where

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

with

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

Note that

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

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, the above estimator $\hat{u}$ in (3.1) does not require sample from each density $\pi \in \Pi$. Thus $\hat{u}$ is well suited for the 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 (k) of skeleton densities.

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. The Bayes factors are often used in model selection. Recently, Roy and Evangelou and Zhu (2014) estimate link function and covariance function parameters in spatial generalized linear mixed models by calculating $\hat{u}$ corresponding to a large number (combination) of these parameter values and subsequently choosing that value which maximizes the marginal likelihood function.

The estimators $\hat{u}$ and $\hat{v}$ in (3.1) depend on $\boldsymbol{d}$, which is generally unknown in practice. As mentioned in the Introduction, 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 density 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. In order to avoid introducing more notations and since estimating $\boldsymbol{d}$ is not the primary goal of this section, we use $\tilde{\Phi}_l$'s and N_l's to denote the stage 1 chains and their length 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. This two-stage method is proposed in Buta and Doss (2011) who quantify its benefits over the single stage method where the same MCMC samples are used to estimate both $\boldsymbol{d}$ and $u(\pi, \pi_1)$. In Section 3.1 we present a CLT for $\hat{u}$ and construct batch means estimator of its asymptotic variance. Finally, we discuss the estimation of standard errors of $\hat{\eta}$ in Section 3.2.

3.1 Estimating a large number of ratios of normalizing constants

Before we state a CLT for $\hat{u}(\pi, \pi_1; \boldsymbol{a}, \hat{\boldsymbol{d}})$, we define the following notations:

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

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

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

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

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

and assuming $n_l = e_l m_l$

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

where $\bar{u}_m(\boldsymbol{a}, \boldsymbol{d})$ is the average of the $(m+1)$st block $\{u(X_{mb_l+1}^{(l)}; \boldsymbol{a}, \boldsymbol{d}), \cdots, u(X_{(m+1)b_l}^{(l)}; \boldsymbol{a}, \boldsymbol{d})\}$, and $\bar{\bar{u}}(\boldsymbol{a}, \boldsymbol{d})$ is the overall average of $\{u(X_1^{(l)}; \boldsymbol{a}, \boldsymbol{d}), \cdots, u(X_{n_l}^{(l)}; \boldsymbol{a}, \boldsymbol{d})\}$. 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* $N \equiv \sum_{l=1}^{k} N_l \to \infty$. *Assume that the stage 2 Markov chains* $\Phi_1,\ldots,\Phi_k$ *are geometrically ergodic, and there exists* $\epsilon > 0$ *such that*

$$E_{\pi_l}|u(X;\boldsymbol{a},\boldsymbol{d})|^{2+\epsilon} < \infty$$

for each $l = 1,\cdots,k$. *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* $l = 1,\cdots,k$.

(1) 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 there exist* $\epsilon > 0, \delta > 0$ *such that* $E_{\pi_l}|u(X;\boldsymbol{a},\boldsymbol{d})|^{2+\delta+\epsilon} < \infty$ *for all* $l = 1,2,\ldots,k$, $b_l = \lfloor n_l^\nu \rfloor$ *where* $1 > \nu > 2/(2+\delta)$. *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 strong consistent estimator of the asymptotic variance in* (3.8).

The proof of Theorem 2 is given in Appendix B.

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 1st 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 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 $\boldsymbol{d}$, thus greatly reducing the first term in the variance expression (3.8).

3.2 Estimation of expectations using multiple chain importance sampling

In this section we discuss the estimation of standard errors of the multi chain importance sampling estimator $\hat{\eta}$ given in (1.3). Recall from (3.1) that $\hat{\eta} \equiv \hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d}) = \hat{v}^{[f]}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d})/\hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d})$, where $\hat{v} \equiv \hat{v}^{[f]}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d}) := \sum_{l=1}^{k}(a_l/n_l)\sum_{i=1}^{n_l} v^{[f]}(X_i^{(l)};\boldsymbol{a},\boldsymbol{d})$ and $\hat{u} \equiv \hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d}) := \sum_{l=1}^{k}(a_l/n_l)\sum_{i=1}^{n_l} u(X_i^{(l)};\boldsymbol{a},\boldsymbol{d})$.

In order to state a CLT for $\hat{\eta}$ we define the following notations:

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

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

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

(Note that, γ_l^{22} is 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

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

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

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

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

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

where $[c(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1}$ is defined in (3.6). We use the same two-stage procedure, as in Section 3.1, for evaluating $\hat{\eta}$. Again assuming $n_l = e_l m_l$, let

$$\widehat{\Gamma}_l(\pi;\boldsymbol{a},\boldsymbol{d}) \equiv \frac{b_l}{e_l-1}\sum_{m=0}^{e_l-1}\left[\begin{pmatrix} \bar{v}_m^{[f]} \\ \bar{u}_m \end{pmatrix} - \begin{pmatrix} \bar{\bar{v}}^{[f]} \\ \bar{\bar{u}} \end{pmatrix}\right]\left[\begin{pmatrix} \bar{v}_m \\ \bar{u}_m \end{pmatrix} - \begin{pmatrix} \bar{\bar{v}}^{[f]} \\ \bar{\bar{u}} \end{pmatrix}\right]^\top \tag{3.13}$$

$$= \frac{b_l}{e_l-1}\begin{pmatrix} \sum_{m=0}^{e_l-1}\left[\bar{v}_m^{[f]} - \bar{\bar{v}}^{[f]}\right]^2 & \sum_{m=0}^{e_l-1}\left[\bar{v}_m^{[f]} - \bar{\bar{v}}^{[f]}\right][\bar{u}_m - \bar{\bar{u}}] \\ \sum_{m=0}^{e_l-1}\left[\bar{v}_m^{[f]} - \bar{\bar{v}}^{[f]}\right][\bar{u}_m - \bar{\bar{u}}] & \sum_{m=0}^{e_l-1}[\bar{u}_m - \bar{\bar{u}}]^2 \end{pmatrix} \tag{3.14}$$

$$= \begin{pmatrix} \hat{\gamma}^{11}(\pi;\boldsymbol{a},\boldsymbol{d}) & \hat{\gamma}^{12}(\pi;\boldsymbol{a},\boldsymbol{d}) \\ \hat{\gamma}^{21}(\pi;\boldsymbol{a},\boldsymbol{d}) & \hat{\gamma}^{22}(\pi;\boldsymbol{a},\boldsymbol{d}) \end{pmatrix}, \text{say,} \tag{3.15}$$

where $\bar{v}_m^{[f]}$ is the average of the $(m+1)$st block $\{v^{[f]}(X_{mb_l+1}^{(l)};\boldsymbol{a},\boldsymbol{d}),\cdots,v^{[f]}(X_{(m+1)b_l}^{(l)};\boldsymbol{a},\boldsymbol{d})\}$, $\bar{\bar{v}}^{[f]}$ 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 $\bar{u}_m \equiv \bar{u}_m(\pi,\boldsymbol{a},\boldsymbol{d}), \bar{\bar{u}} \equiv \bar{\bar{u}}(\pi,\boldsymbol{a},\boldsymbol{d})$ 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 holds such that* $N^{1/2}(\hat{\boldsymbol{d}}-\boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0,V)$ *as* $N \equiv \sum_{l=1}^k N_l \to \infty$. *Assume that the stage 2 Markov chains* $\Phi_1,\ldots,\Phi_k$ *are geometrically ergodic, and there exists* $\epsilon > 0$ *such that*

$$E_{\pi_l}|u(X;\boldsymbol{a},\boldsymbol{d})|^{2+\epsilon} < \infty \;\text{ and }\; E_{\pi_l}|v^{[f]}(X;\boldsymbol{a},\boldsymbol{d})|^{2+\epsilon} < \infty \tag{3.16}$$

for each $l = 1,\cdots,k$. *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* $l = 1,\cdots,k$.

(1) Then as $n_1, \ldots, n_k \to \infty$,

$$\sqrt{n}(\hat{\eta}^{[f]}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - E_\pi f) \xrightarrow{d} N(0, qe(\pi; \boldsymbol{a}, \boldsymbol{d})^\top V e(\pi; \boldsymbol{a}, \boldsymbol{d}) + \rho(\pi; \boldsymbol{a}, \boldsymbol{d})). \tag{3.17}$$

(2) Let $\widehat{V}$ *be the consistent estimator of* V *given in Theorem 1 (3). Assume that there exists* $\epsilon > 0$ *such that* (3.16) *holds for all* $l = 1, 2, \ldots, k$, $b_l = \lfloor n_l^\nu \rfloor$ *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 strong consistent estimator of the asymptotic variance in* (3.17).

The proof of Theorem 3 is given in Appendix C.

Remark 3. Theorem 2 (1) and Theorem 3 (1) extend Buta and Doss's (2011) Theorem 1 and Theorem 3 respectively who consider the special case when $a_l = n_l/n$. Tan and Doss and Hobert (2015) mention that $a_l = n_l/n$ is not an optimal choice for $\boldsymbol{a}$. This is why here we consider a general arbitrary vector $\boldsymbol{a}$.

Remark 4. In the case when $\boldsymbol{d}$ is unknown, Tan and Doss and Hobert (2015) provide a regeneration based consistent estimator 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. As can be seen in the proofs of Theorem 2 in Appendix B and Theorem 3 in Appendix C, proving consistency of the batch means estimators of the variances of $\hat{u}$ and $\hat{\eta}$ in the general case requires careful, deep calculations.

Remark 5. As mentioned in Remark 4, Tan and Doss and Hobert (2015) provide regeneration estimators for calculating standard errors of $\hat{u}$ and $\hat{\eta}$ in the special case when $\boldsymbol{a} = (1, \hat{\boldsymbol{d}})$. We now describe a trick that will essentially allow any choice of $\boldsymbol{a}$ for Tan and Doss and Hobert 's (2015) regeneration estimator. In particular, set $\boldsymbol{a} = w * (1, \hat{\boldsymbol{d}})$ for any user specified fixed vector w. This general choice of $\boldsymbol{a}$ still allow the expressions in (2.18) of Tan and Doss and Hobert (2015) to be free of $\hat{\boldsymbol{d}}$, thus leading to the independence of certain quantities required for their estimator to work (details are given in the supplementary materials). This method was not mentioned in Tan and Doss and Hobert (2015).

4 Illustrations

In Section 4.1 through a toy example, we first discuss the different choices of weight functions and their effect on the estimates of expectations and ratios of normalizing constants. Next in Section 4.2, we use our batch means method for empirical Bayes variable selection in the context of standard linear regression models with moderately large number of variables where regeneration based simulation is impractical.

4.1 Toy example

Let $t_{r,\mu}$ denote the t-distribution with degree of freedom r and central parameter μ. We consider a toy example where $\pi_1(\cdot)$ and $\pi_2(\cdot)$ are the density functions for $t_{5,\mu_1=1}$ and $t_{5,\mu_2=0}$ respectively. For simplicity, let $\nu_i(\cdot) = \pi_i(\cdot)$ for $i = 1, 2$. Our plan is (1) to estimate the ratio between the two normalizing constants, $d = m_2/m_1$, and (2) to study a sea of t-distributions, say, $\Pi = \{t_{r,\mu} :$

$\mu \in M\}$ where M is a fine grid over $[0,1]$, say $M = \{0, .01, \cdots, .99, 1\}$. For each $\mu \in M$, we assume that $\nu_\mu(\cdot) = \pi_\mu(\cdot)$, and we want to learn the ratio between the normalizing constants through $d_\mu := \frac{m_\mu}{m_1}$, as well as the expectation of each distribution in Π, $E_{t_{5,\mu}}X$, written as $E_\mu X$ for short. Clearly, we know the exact answer to all the questions above: $d = d_\mu = 1$ and $E_{t_{5,\mu}}X = \mu$ for any $\mu \in M$. Nevertheless, we follow the two-stage procedure from section 3 to generate Markov chains from π_1 and π_2 respectively, and build MCMC estimators as described in Theorem 1, 2, and 3. The main goal is to check the performance of the batch means (BM) and the regeneration based simulation (RS) estimators for the asymptotic variance of these estimators.

We will draw iid samples from π_1, and draw Markov chain samples from π_2 using the so called "independent Metropolis Hastings (MH) algorithm" with proposal density $t_{5,1}$. For $i = 1, 2$, we will draw N_i observations from π_i in stage 1, and n_i observations from π_i in stage 2. We set $N_1 \approx N_2$ and $n_1 \approx n_2$, and further ask stage 2 sample sizes to be smaller than that of stage 1, specifically, $n_1 = N_1/10$, due to reasons about computing cost explained right after Theorem 2. We consider an increasing sequence of sample sizes, from $n_1 = 10^3$ to 10^5, in order to examine trace plots for the BM and the RS estimates of the asymptotic variances. Such two stage procedures are repeated 1000 times, so that empirical estimates of the asymptotic variances can be calculated and used to evaluate our estimators.

Note that, for estimators based on stage 1 samples, Theorem 1 allows any choice of weight, $\boldsymbol{a}^1$. And for estimators based on stage 2 samples, Theorem 2 and 3 allow any choice of weight, $\boldsymbol{a}^2$, in constructing BM estimators. As for RS estimators, similar Theorems hold and are described in Doss and Tan (2014); Tan and Doss and Hobert (2015). Recall from Remark 5 that we made an important generalization to existing theorems concerning RS estimators. That is, essentially any non-negative numerical vector can now be used as weight. We will discuss the choice of weights and their impact on the estimators in a separate section below. For now, we use the toy example to check the aforementioned Theorems, to see if both the BM estimator and the RS estimator are consistent, regardless of the weight chosen.

Figure 1 displays the BM and the RS estimates of $\hat{d}$ in stage 1, obtained from each of the replications. They are evaluated at a sequence of sample sizes, from $n_1 = 10^3$ to 10^5, and we can see that both the BM and the RS estimates approach the empirical asymptotic variance as the sample size increases, suggesting their consistency. Also, as expected, the BM estimator are more variable than the RS estimates. Similarly, Figures 2 to 4 show convergence of the BM and the RS estimators to the empirical asymptotic variance of $\hat{d}_\mu$ and $\hat{E}_\mu(X)$ respectively, in stage 2. (Small deviations between the limit of the estimators and the empirical asymptotic variance are probably due to the fact that we do not know the true asymptotic variance, but are using an empirical estimate of it based on a finite sample size over a finite number of repetitions. Also, only plots at selected values of $\mu \in M$ are shown due to space limit, but convergence of the estimators are indeed observed in all the case we have inspected.)

Overall, the simulations study suggests that both the BM and the RS methods provide consistent estimators for the true asymptotic variance. Also, the RS estimators enjoy smaller mse in most cases. Nevertheless, when the number of regenerations is not great, BM estimators could be the more stable estimator. For example, in the top left plot of Figure 2, at sample size $n_2 = 1000$, in more than 5 out of the 1000 replications, RS estimators wildly over estimated the target. Further, in the cases where regeneration is "not viable", i.e., where the number of regenerations is extremely small for any affordable sample size, BM would be the more stable estimator, or the

only reliable estimator between the two.

Choice of weights in stage 1

In practice, for stage 1, we recommend obtaining a close-to-optimal weight, $\boldsymbol{a}^{1\text{opt}}$, using a short pilot study. See Doss and Tan (2014) for details on how the pilot study can be done. The right panel of Figure 1 displays empirical asymptotic variance of $\hat{d}$ over 1000 replications with $n_1 = 10^5$. For comparison, the left panel of Figure 1 presents such results based on the naive choice of $\boldsymbol{a}^1 = (.5, .5)$. The naive choice is determined by the relative sample sizes of the reference chains, which is an asymptotically optimal choice had both chains been independent. But in our example where chain 2 involves dependent samples, using $\boldsymbol{a}^{1\text{opt}}$ results in $\hat{d}$ with smaller standard errors.

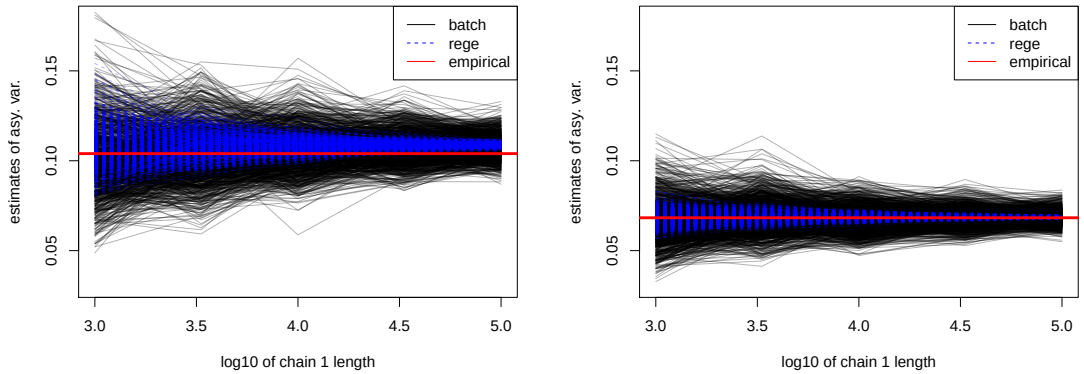

Figure 1: Estimates of the asymptotic variance of $\hat{d}$ in stage 1.

Choice of weights in stage 2

As for stage 2, where we have a sea of parameter values of interest, for each $\mu \in M$, there would be a weight $a^2(\mu)$ that is optimal for estimating d_μ, yet another weight $a^2(\mu)$ that is optimal for estimating $E_{t_{5,\mu}} X$. Again, a pilot study can be used to find approximations for the $2|M|$ sets of optimal weights, or a selected subset of them scattered inside M. A less costly alternative is to set $\boldsymbol{a}^2(\mu)$ to be inversely proportional to the deviation of π_μ from the k reference points. Different ways of defining deviation may be appropriate in different problems, here we simply measure the deviation between $t_{5,\mu}$ and $t_{5,\mu'}$ by $|\mu - \mu'|$. Further adjustment of $\boldsymbol{a}^2(\mu)$ are possible that assigns the more efficient chains heavier weight. After all, we experiment with the following strategies for stage 2 estimates.

1. basic: $\boldsymbol{a}^2 = (.5, .5)$
2. inv-dist: $\boldsymbol{a}^2(\mu) \propto (\frac{1}{|\mu-\mu_1|}, \frac{1}{|\mu-\mu_2|})$

3. ess by inv-dist: $\boldsymbol{a}^2(\mu) \propto (\text{ess}_1, \text{ess}_2) * (\frac{1}{|\mu-\mu_1|}, \frac{1}{|\mu-\mu_2|})$, where $*$ denotes point-wise product.

Not surprisingly, none of the three strategies is a clear winner, and their performances vary greatly depending on the quantity being estimated (d_μ or $E_\mu(X)$, or the expectation of other functions), as well as on whether we are only interested in μ falling in between the reference values $\mu_1 = 1$ and $\mu_2 = 0$, or beyond. Still, we visualize the simulation results in Figure 5 and summarize the situation briefly.

For estimating d_μ

1. For $\mu \in (0, 1)$, strategy 2 works the best.

2. For $\mu = 0$, strategy 2 and 3 work better than strategy 1. Indeed, both of them simply set their stage 2 estimates $\hat{d}_0$ to be the stage 1 estimate, $\hat{d}$. This would be a better choice than strategy 1 because in a two-step procedure, stage 1 chains are often much longer than stage 2 chains, and hence $\hat{d}$ is already a very accurate estimate for $d_0 = d$.

3. Though not recommended, in case it is of interest to explore $\mu \notin [0, 1]$, strategy 2 and 3 generally leads to more stable estimates of d_μ. However, all strategies lead to much larger asymptotic variances than desired. Indeed, in case $\mu \notin [0, 1]$, it's better to reconsider the choice of reference chains to be drawn at μ in its vicinity.

For estimating $E_\mu(X)$

1. For $\mu \in (0, 1)$, strategy 2 works the best in general. Strategy 3 is very unstable.

2. For either $\mu = 0$ or 1, strategy 2 and 3 are the same, and they only utilize the reference chain from μ. This was a wise choice for estimating d_μ as explained above, but not so for any other quantities of interest.

After all, strategies 2 and 3 have an advantage over using the naive weight when the estimands are ratios between normalizing constants. However, when estimating $E_\mu(X)$, the situation is more complicated. Our impression is that assigning any extreme weight will lead to high variability in the estimator. So it is reasonable to simply use the naive weight, or use other strategies that bound the weights from 0 and 1.

5 Discussion

In this paper we consider two separate but related problems. The first problem is to estimate the ratios of unknown normalizing constants given Markov chain samples from each of the k probability densities for some integer k larger than 1. The second problem is to estimate expectations of a function with respect to a large number of probability distributions. The two problems are related in the sense that the multiple chains importance sampling estimators used for the latter uses the estimates derived for solving the first problem. The first situation also arises in a variety of contexts other than the multiple chain importance sampling estimators. In both situations, we consider estimators derived by methods involving flexible choice of weights and thus these estimators are appropriate for Markov chains with different mixing behaviors. We establish CLTs

for these estimators and develop batch means methods for consistently estimating their standard errors. Although we compare batch means and regeneration based methods in this paper, spectral methods can also be used for variance estimation. Generally estimation by spectral methods is computationally more expensive (Doss and Tan, 2014, p. 703). Flegal and Jones (2010) compare the performance of confidence intervals produced by batch means, regeneration 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.

Appendices

A Proof of Theorem 1

Proof. The proof of the consistency of $\hat{\boldsymbol{d}}$ follows from Doss and Tan (2014) (section A.1) and is omitted. Now onward, we use D&T to denote Doss and Tan (2014). Establishing a CLT for $\hat{\boldsymbol{d}}$ is although analogous to section A.2 of D&T, there are some significant differences. Below we establish the CLT for $\hat{\boldsymbol{d}}$ and finally, we show that $\widehat{V}$ is a consistent estimator of V.

We begin by considering $n^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)$. As before, let ∇ represents the gradient operator. As in the classical proof of asymptotic normality of maximum likelihood estimators, we expand $\nabla \ell_n$ at $\hat{\boldsymbol{\zeta}}$ around $\boldsymbol{\zeta}_0$, and using the appropriate scaling factor, we get

$$-n^{-1/2}\big(\nabla \ell_n(\hat{\boldsymbol{\zeta}}) - \nabla \ell_n(\boldsymbol{\zeta}_0)\big) = -n^{-1}\nabla^2 \ell_n(\boldsymbol{\zeta}_*) n^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0), \tag{A.1}$$

where $\boldsymbol{\zeta}_*$ is between $\hat{\boldsymbol{\zeta}}$ and $\boldsymbol{\zeta}_0$. Consider the left side of (A.1), which is just $n^{1/2}n^{-1}\nabla \ell_n(\boldsymbol{\zeta}_0)$, since $\nabla \ell_n(\hat{\boldsymbol{\zeta}}) = 0$. There are several nontrivial components to the proof, so we first give an outline.

1. Following D&T we show that each element of the vector $n^{-1}\nabla \ell_n(\boldsymbol{\zeta}_0)$ can be represented as a linear combination of mean 0 averages of functions of the k chains.
2. Based on Step 1 above, applying CLT for each of the k Markov chain averages, we obtain a CLT for the scaled score vector. In particular, we show that $n^{1/2}n^{-1}\nabla \ell_n(\boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0, \Omega)$, where Ω defined in (2.10) involves infinite sums of auto-covariances of each chain.
3. Following Geyer (1994) it can be shown that $-n^{-1}\nabla^2 \ell_n(\boldsymbol{\zeta}_*) \xrightarrow{\text{a.s.}} B$ and that $\big(-n^{-1}\nabla^2 \ell_n(\boldsymbol{\zeta}_*)\big)^{\dagger} \xrightarrow{\text{a.s.}} B^{\dagger}$, where B is defined in (2.9).
4. We conclude that $n^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0, B^{\dagger}\Omega B^{\dagger})$.
5. Since $\boldsymbol{d} = g(\boldsymbol{\zeta}_0)$ and $\hat{\boldsymbol{d}} = g(\hat{\boldsymbol{\zeta}})$, where g is defined in (2.6), by the delta method it follows that $n^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0, V)$ where $V = D^{\top}B^{\dagger}\Omega B^{\dagger}D$.

We now provide the details.

1. Following D&T we start by considering $n^{-1}\nabla\ell_n(\boldsymbol{\zeta}_0)$. For $r = 1, \ldots, k$, from D&T we have

$$\begin{aligned}\frac{\partial \ell_n(\boldsymbol{\zeta}_0)}{\partial \zeta_r} &= w_r \sum_{i=1}^{n_r} \big(1 - p_r(X_i^{(r)}, \boldsymbol{\zeta}_0)\big) - \sum_{\substack{l=1 \\ l\neq r}}^{k} w_l \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) \\ \text{(can be shown to)} \quad &= w_r \sum_{i=1}^{n_r} \Big(1 - p_r(X_i^{(r)}, \boldsymbol{\zeta}_0) - \big[1 - E_{\pi_r}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big]\Big) \qquad \text{(A.2)} \\ &\quad - \sum_{\substack{l=1 \\ l\neq r}}^{k} w_l \sum_{i=1}^{n_l} \big[p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big].\end{aligned}$$

   That is, (A.2) can be used to view $n^{-1}\partial\ell_n(\boldsymbol{\zeta}_0)/\partial\zeta_r$ as a linear combination of mean 0 averages of functions of the k chains.

2. Next, we need to write a CLT for the vector $\nabla\ell_n(\boldsymbol{\zeta}_0) = (\partial\ell_n(\boldsymbol{\zeta}_0)/\partial\zeta_1, \cdots, \partial\ell_n(\boldsymbol{\zeta}_0)/\partial\zeta_k)^T$, that is, to show that

$$n^{-1/2}\nabla\ell_n(\boldsymbol{\zeta}_0) \xrightarrow{d} N(0, \Omega) \qquad \text{as } n \to \infty.$$

   Note that,

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\zeta}_0)}{\partial \zeta_r} &= -\frac{1}{\sqrt{n}} \sum_{l=1}^{k} w_l \sum_{i=1}^{n_l} \big[p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big] \\ &= -\sum_{l=1}^{k} \sqrt{\frac{n}{n_l}} a_l \frac{1}{\sqrt{n_l}} \sum_{i=1}^{n_l} \big[p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big] \\ &= -\sum_{l=1}^{k} \sqrt{n} a_l \bar{\bar{Y}}^{(r,l)}, \qquad \text{(A.3)}\end{aligned}$$

   where $\bar{\bar{Y}}^{(r,l)} := \frac{1}{n_l}\sum_{i=1}^{n_l} Y_i^{(r,l)}$ and $Y_i^{(r,l)}$ is as defined in (2.8). Since $p_r(x, \boldsymbol{\zeta}) \in (0, 1)$ for all x, r and $\boldsymbol{\zeta}$, we have $E_{\pi_l}\big(|p_r(X, \boldsymbol{\zeta}_0) - E_{\pi_l}(p_r(X, \boldsymbol{\zeta}_0))|^{2+\epsilon}\big) < \infty$ for any $\epsilon > 0$. Then since Φ_l is geometrically ergodic, we have asymptotic normality for the univariate quantities $\sqrt{n_l}\bar{\bar{Y}}^{(r,l)}$. Since $n_l/n \to s_l$ for $l = 1, 2, \ldots, k$ and a_l's are known, by independence of the k chains, we conclude that

$$\frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\zeta}_0)}{\partial \zeta_r} \xrightarrow{d} \mathcal{N}(0, \Omega_{rr}) \qquad \text{as } n \to \infty,$$

   where Ω is defined in (2.10). Next, we extend the component-wise CLT to a joint CLT. Consider any $\boldsymbol{t} \in (t_1, \cdots, t_k) \in \mathbb{R}^k$, we have

$$\begin{aligned}&t_1\frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\zeta}_0)}{\partial \zeta_1} + \cdots + t_k\frac{1}{\sqrt{n}}\frac{\partial \ell_n(\boldsymbol{\zeta}_0)}{\partial \zeta_k} \\ &= -\sum_{l=1}^{k}\left(t_1\sqrt{n}a_l\frac{\sum_{i=1}^{n_l} Y_i^{(1,l)}}{n_l} + \cdots + t_k\sqrt{n}a_l\frac{\sum_{i=1}^{n_l} Y_i^{(k,l)}}{n_l}\right) \\ &= -\sum_{l=1}^{k}\sqrt{\frac{n}{n_l}}a_l\frac{\sum_{i=1}^{n_l}\left(t_1 Y_i^{(1,l)} + \cdots + t_k Y_i^{(k,l)}\right)}{\sqrt{n_l}} \xrightarrow{d} \mathcal{N}(0, \boldsymbol{t}^T\Omega\boldsymbol{t}) \qquad \text{as } n \to \infty.\end{aligned}$$

Hence, the Cramér-Wold device implies the joint CLT,

$$n^{-1/2}\nabla\ell_n(\boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0,\Omega) \qquad \text{as } n\to\infty. \tag{A.4}$$

3. Items 3-5 are omitted here as the derivations are basically the same as in D&T.

Next we provide a proof of the consistency of the estimate of the asymptotic variance covariance matrix V, that is, we show that $\widehat{V} \equiv \widehat{D}^\top \widehat{B}^\dagger \widehat{\Omega} \widehat{B}^\dagger \widehat{D} \xrightarrow{\text{a.s.}} V \equiv D^\top B^\dagger \Omega B^\dagger D$ as $n\to\infty$. Since $\hat{\boldsymbol{\zeta}} \xrightarrow{\text{a.s.}} \boldsymbol{\zeta}_0$ and $\hat{\boldsymbol{d}} \xrightarrow{\text{a.s.}} \boldsymbol{d}$, it implies that $\widehat{D} \xrightarrow{\text{a.s.}} D$. From D&T, we know that $\widehat{B} \xrightarrow{\text{a.s.}} B$ and using the spectral representation of $\widehat{B}$ and of B, it follows that $\widehat{B}^\dagger \xrightarrow{\text{a.s.}} B^\dagger$.

To complete the proof, we now show that $\widehat{\Omega} \xrightarrow{\text{a.s.}} \Omega$ where the batch means estimator $\widehat{\Omega}$ is defined in (2.15). This will be proved in couple of steps. First, we consider a single chain Φ_l used to calculate k quantities and establish a multivariate CLT. We use the results in Flegal et al. (2014) who obtain conditions for the nonoverlapping batch means estimator to be strongly consistent in multivariate settings. Second, we combine results from the k independent chains. Finally, we show that $\widehat{\Omega}$ is a strongly consistent estimator of Ω.

Denote $\bar{\bar{Y}}^{(l)} = \left(\bar{\bar{Y}}^{(1,l)}, \bar{\bar{Y}}^{(2,l)}, \ldots, \bar{\bar{Y}}^{(k,l)}\right)^\top$. By a proof similar to what we used to derive (A.4) using the Cramér-Wold device, we have the following joint CLT for W_l:

$$\sqrt{n_l}\bar{\bar{Y}}^{(l)} \xrightarrow{d} \mathcal{N}(0,\Sigma^{(l)}) \qquad \text{as } n_l\to\infty,$$

where $\Sigma^{(l)}$ is a $k\times k$ variance covariance matrix with

$$\Sigma^{(l)}_{rs} = E_{\pi_l}\{Y_1^{(r,l)}Y_1^{(s,l)}\} + \sum_{i=1}^{\infty} E_{\pi_l}\{Y_1^{(r,l)}Y_{1+i}^{(s,l)}\} + \sum_{i=1}^{\infty} E_{\pi_l}\{Y_{1+i}^{(r,l)}Y_1^{(s,l)}\}, \quad r,s=1,\ldots,k. \tag{A.5}$$

The nonoverlapping batch means estimator of $\Sigma^{(l)}$ is given in (2.12). We now prove the strong consistency of $\widehat{\Sigma}^{(l)}$. Note that $\widehat{\Sigma}^{(l)}$ is defined using the terms $\bar{Z}_m^{(r,l)}$'s which involve the random quantity $\hat{\boldsymbol{\zeta}}$. We define $\widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0)$ to be $\widehat{\Sigma}^{(l)}$ with $\boldsymbol{\zeta}_0$ substituted for $\hat{\boldsymbol{\zeta}}$, that is,

$$\widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0) = \frac{b_l}{e_l-1}\sum_{m=0}^{e_l-1}\left[\bar{Y}_m^{(l)} - \bar{\bar{Y}}^{(l)}\right]\left[\bar{Y}_m^{(l)} - \bar{\bar{Y}}^{(l)}\right]^\top \quad \text{for } \; l=1,2,\ldots,k,$$

where $\bar{Y}_m^{(l)} = \left(\bar{Y}_m^{(1,l)},\ldots,\bar{Y}_m^{(k,l)}\right)^\top$ with $\bar{Y}_m^{(r,l)} := \sum_{j=mb_l+1}^{(m+1)b_l} Y_j^{(r,l)}/b_l$. We prove $\widehat{\Sigma}^{(l)} \xrightarrow{\text{a.s.}} \Sigma^{(l)}$ in two steps:

1. $\widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0) \xrightarrow{\text{a.s.}} \Sigma^{(l)}$ and
2. $\widehat{\Sigma}^{(l)} - \widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0) \xrightarrow{\text{a.s.}} 0$.

Strong consistency of the multivariate batch means estimator $\widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0)$ requires both $e_l\to\infty$ and $b_l\to\infty$. Since for all r, $E_{\pi_l}\left(|p_r(X,\boldsymbol{\zeta}_0) - E_{\pi_l}(p_r(X,\boldsymbol{\zeta}_0))|^{2+\epsilon}\right) < \infty$ for any $\epsilon>0$, Φ_l is geometrically ergodic, and $b_l = \lfloor n_l^\nu \rfloor$ where $1>\nu>0$, it follows from a more general result in Flegal et al. (2014) that $\widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0) \xrightarrow{\text{a.s.}} \Sigma^{(l)}$ as $n_l\to\infty$. We show $\widehat{\Sigma}^{(l)}_{rs} - \widehat{\Sigma}^{(l)}_{rs}(\boldsymbol{\zeta}_0) \xrightarrow{\text{a.s.}} 0$ where $\widehat{\Sigma}^{(l)}_{rs}$ and $\widehat{\Sigma}^{(l)}_{rs}(\boldsymbol{\zeta}_0)$ are the (r,s)th elements of the $k\times k$ matrices $\widehat{\Sigma}^{(l)}_{rs}$ and $\widehat{\Sigma}^{(l)}_{rs}(\boldsymbol{\zeta}_0)$ respectively. By the

mean value theorem (in multiple variables), there exists $\boldsymbol{\zeta}^* = t\hat{\boldsymbol{\zeta}} + (1-t)\boldsymbol{\zeta}_0$ for some $t \in (0,1)$, such that

$$\widehat{\Sigma}_{rs}^{(l)} - \widehat{\Sigma}_{rs}^{(l)}(\boldsymbol{\zeta}_0) = \nabla \widehat{\Sigma}_{rs}^{(l)}(\boldsymbol{\zeta}^*) \cdot (\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0), \tag{A.6}$$

where $\cdot$ represents the dot product. Note that

$$\widehat{\Sigma}_{rs}^{(l)}(\boldsymbol{\zeta}) = \frac{b_l}{e_l - 1} \sum_{m=0}^{e_l-1} [\bar{Z}_m^{(r,l)}(\boldsymbol{\zeta}) - \bar{\bar{Z}}^{(r,l)}(\boldsymbol{\zeta})][\bar{Z}_m^{(s,l)}(\boldsymbol{\zeta}) - \bar{\bar{Z}}^{(s,l)}(\boldsymbol{\zeta})],$$

where $\bar{Z}_m^{(r,l)}(\boldsymbol{\zeta}) := \sum_{j=mb_l+1}^{(m+1)b_l} p_r(X_j^{(l)}, \boldsymbol{\zeta})/b_l$ and $\bar{\bar{Z}}^{(r,l)}(\boldsymbol{\zeta}) := \sum_{j=1}^{n_l} p_r(X_j^{(l)}, \boldsymbol{\zeta})/n_l$. Some calculations show that for $t \neq r$

$$\frac{\partial \bar{Z}_m^{(r,l)}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_t} = -\frac{1}{b_l} \sum_{j=mb_l+1}^{(m+1)b_l} p_r(X_j^{(l)}, \boldsymbol{\zeta}) p_t(X_j^{(l)}, \boldsymbol{\zeta})$$

and

$$\frac{\partial \bar{Z}_m^{(r,l)}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_r} = \frac{1}{b_l} \sum_{j=mb_l+1}^{(m+1)b_l} p_r(X_j^{(l)}, \boldsymbol{\zeta})(1 - p_r(X_j^{(l)}, \boldsymbol{\zeta})).$$

We denote $\bar{U}_m^r := \bar{Z}_m^{(r,l)}(\boldsymbol{\zeta}) - E_{\pi_l}[p_r(X, \boldsymbol{\zeta})]$, $\bar{\bar{U}}^r := \bar{\bar{Z}}^{(r,l)}(\boldsymbol{\zeta}) - E_{\pi_l}[p_r(X, \boldsymbol{\zeta})]$, and similarly the centered versions of $\partial \bar{Z}_m^{(r,l)}(\boldsymbol{\zeta})/\partial \boldsymbol{\zeta}_t$ and $\partial \bar{\bar{Z}}^{(r,l)}(\boldsymbol{\zeta})/\partial \boldsymbol{\zeta}_t$ by $\bar{V}_m^{(r,t)}$ and $\bar{\bar{V}}^{(r,t)}$ respectively. Since $p_r(X, \boldsymbol{\zeta})$ is uniformly bounded by 1 and Φ_l is geometrically ergodic, there exist $\sigma_r^2, \tau_{r,t}^2 < \infty$ such that $\sqrt{b_l}\bar{U}_m^r \xrightarrow{d} N(0, \sigma_r^2)$, $\sqrt{n_l}\bar{\bar{U}}^r \xrightarrow{d} N(0, \sigma_r^2)$, $\sqrt{b_l}\bar{V}_m^{(r,t)} \xrightarrow{d} N(0, \tau_{r,t}^2)$, and $\sqrt{n_l}\bar{\bar{V}}^{(r,t)} \xrightarrow{d} N(0, \tau_{r,t}^2)$. We have

$$\begin{aligned}
\frac{\partial \widehat{\Sigma}_{rs}^{(l)}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_t} &= \frac{1}{e_l - 1} \sum_{m=0}^{e_l-1} [\sqrt{b_l}(\bar{U}_m^r - \bar{\bar{U}}^r)\sqrt{b_l}(\bar{V}_m^{(s,t)} - \bar{\bar{V}}^{(s,t)}) + \sqrt{b_l}(\bar{V}_m^{(r,t)} - \bar{\bar{V}}^{(r,t)})\sqrt{b_l}(\bar{U}_m^s - \bar{\bar{U}}^s)] \\
&= \frac{1}{e_l - 1} \sum_{m=0}^{e_l-1} \left[\sqrt{b_l}\bar{U}_m^r \sqrt{b_l}\bar{V}_m^{(s,t)} + \sqrt{b_l}\bar{V}_m^{(r,t)} \sqrt{b_l}\bar{\bar{U}}_m^s \right] - \frac{1}{e_l - 1} \left[\sqrt{n_l}\bar{\bar{U}}^r \sqrt{n_l}\bar{\bar{V}}^{(s,t)} + \sqrt{n_l}\bar{\bar{V}}^{(r,t)} \sqrt{n_l}\bar{\bar{U}}^s \right].
\end{aligned}$$

It is easy to see that the negative term in the above expression goes to zero as $e_l \to \infty$. Further, since

$$\left| \sqrt{b_l}\bar{U}_m^r \sqrt{b_l}\bar{V}_m^{(s,t)} \right| \leq \frac{1}{2} \left[b_l (\bar{U}_m^r)^2 \right] + \frac{1}{2} \left[b_l (\bar{V}_m^{(s,t)})^2 \right],$$

we have

$$\left| \frac{1}{e_l - 1} \sum_{m=0}^{e_l-1} \sqrt{b_l}\bar{U}_m^r \sqrt{b_l}\bar{V}_m^{(s,t)} \right| \leq \frac{1}{2} \frac{1}{e_l - 1} \sum_{m=0}^{e_l-1} \left[b_l (\bar{U}_m^r)^2 \right] + \frac{1}{2} \frac{1}{e_l - 1} \sum_{m=0}^{e_l-1} \left[b_l (\bar{V}_m^{(s,t)})^2 \right] \xrightarrow{\text{a.s.}} \frac{1}{2}\sigma_r^2 + \frac{1}{2}\tau_{s,t}^2,$$

where the last step above is due to strong consistency of the batch means estimators for the asymptotic variances of the sequences $\{p_r(X_j^{(l)}, \boldsymbol{\zeta}), j = 1, \cdots, n_l\}$ and $\{\partial p_s(X_j^{(l)}, \boldsymbol{\zeta})/\partial \boldsymbol{\zeta}_t, j = 1, \cdots, n_l\}$ respectively. Similarly, we have

$$\left| \frac{1}{e_l - 1} \sum_{m=0}^{e_l-1} \sqrt{b_l}\bar{V}_m^{(r,t)} \sqrt{b_l}\bar{U}_m^s \right| \leq \frac{1}{2} \frac{1}{e_l - 1} \sum_{m=0}^{e_l-1} \left[b_l (\bar{V}_m^{(r,t)})^2 + \frac{1}{2} \frac{1}{e_l - 1} \sum_{m=0}^{e_l-1} \left[b_l (\bar{U}_m^s)^2 \right] \right] \xrightarrow{\text{a.s.}} \frac{1}{2}\tau_{r,t}^2 + \frac{1}{2}\sigma_s^2.$$

Note that the terms $U_m^r V_m^{(r,t)}, \sigma_r^2, \tau_{r,t}^2$, etc, above actually depends on $\boldsymbol{\zeta}$, and we are indeed concerned with the case where $\boldsymbol{\zeta}$ takes on the value $\boldsymbol{\zeta}^*$, lying between $\hat{\boldsymbol{\zeta}}$ and $\boldsymbol{\zeta}_0$. Since, $\hat{\boldsymbol{\zeta}} \xrightarrow{\text{a.s.}} \boldsymbol{\zeta}_0$, $\boldsymbol{\zeta}^* \xrightarrow{\text{a.s.}} \boldsymbol{\zeta}_0$ as $n_l \to \infty$. Let $\|u\|$ denotes the L_1 norm of a vector $u \in \mathbb{R}^k$. So from (A.6), and the fact that $\partial \widehat{\Sigma}_{rs}^{(l)}(\boldsymbol{\zeta})/\partial \boldsymbol{\zeta}_t$ is bounded with probability one, we have

$$|\widehat{\Sigma}_{rs}^{(l)} - \widehat{\Sigma}_{rs}^{(l)}(\boldsymbol{\zeta}_0)| \leq \max_{1 \leq t \leq k} \left\{ \left| \frac{\partial \widehat{\Sigma}_{rs}^{(l)}(\boldsymbol{\zeta}^*)}{\partial \boldsymbol{\zeta}_t} \right| \right\} \|\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\| \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \to \infty.$$

Since $\widehat{\Sigma}^{(l)} \xrightarrow{\text{a.s.}} \Sigma^{(l)}$, for $l = 1, \ldots, k$, it follows that $\widehat{\Sigma} \xrightarrow{\text{a.s.}} \Sigma$ where $\widehat{\Sigma}$ is defined in (2.13) and Σ is the corresponding $k^2 \times k^2$ variance covariance matrix, that is, Σ is a block diagonal matrix as $\widehat{\Sigma}$ with $\Sigma^{(l)}$ substituted for $\widehat{\Sigma}^{(l)}, l = 1, \ldots, k$. Since $n_l/n \to s_l$ for $l = 1, 2, \ldots, k$, we have $A_n \to A_c$ as $n \to \infty$ where A_n is defined in (2.14) and

$$A_c = \left(-\sqrt{\frac{1}{c_1}} a_1 I_k \quad -\sqrt{\frac{1}{c_2}} a_2 I_k \quad \ldots \quad -\sqrt{\frac{1}{c_k}} a_k I_k \right).$$

Finally from (2.10) and (A.5) we see that $\Omega = A_c \Sigma A_c^T$. So from (2.15) we have $\widehat{\Omega} \equiv A_n \widehat{\Sigma} A_n^T \xrightarrow{\text{a.s.}} A_c \Sigma A_c^T = \Omega$ as $n \to \infty$. □

B Proof of Theorem 2

Proof. As in Buta and Doss (2011) we write

$$\sqrt{n}(\hat{u}(\pi, \pi_1; \boldsymbol{a}, \hat{\boldsymbol{d}}) - u(\pi, \pi_1)) = \sqrt{n}(\hat{u}(\pi, \pi_1; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \hat{u}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d})) + \sqrt{n}(\hat{u}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d}) - u(\pi, \pi_1)). \tag{B.1}$$

First, consider the 2nd term, which involves randomness only from the 2nd stage. From (3.3) note that $\sum_{l=1}^k a_l E_{\pi_l} u(X; \boldsymbol{a}, \boldsymbol{d}) = u(\pi, \pi_1)$. Then from (3.1) we have

$$\sqrt{n}(\hat{u}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d}) - u(\pi, \pi_1)) = \sum_{l=1}^k a_l \sqrt{\frac{n}{n_l}} \frac{\sum_{i=1}^{n_l} (u(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d}) - E_{\pi_l} u(X; \boldsymbol{a}, \boldsymbol{d}))}{\sqrt{n_l}}.$$

Since Φ_l is geometrically ergodic and $E_{\pi_l}|u(X; \boldsymbol{a}, \boldsymbol{d})|^{2+\epsilon}$ is finite, it follows that $\sum_{i=1}^{n_l} (u(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d}) - E_{\pi_l} u(X; \boldsymbol{a}, \boldsymbol{d}))/\sqrt{n_l} \xrightarrow{d} N(0, \tau_l^2(\pi; \boldsymbol{a}, \boldsymbol{d}))$ where $\tau_l^2(\pi; \boldsymbol{a}, \boldsymbol{d})$ is defined in (3.4). As $n_l/n \to s_l$ and the Markov chains Φ_l's are independent, it follows that $\sqrt{n}(\hat{u}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d}) - u(\pi, \pi_1)) \xrightarrow{d} N(0, \tau^2(\pi; \boldsymbol{a}, \boldsymbol{d}))$.

Now we consider the 1st term in the right hand side of (B.1). Letting $F(\boldsymbol{z}) = \hat{u}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{z})$, by Taylor series expansion of F about $\boldsymbol{d}$ we have

$$\sqrt{n}(F(\hat{\boldsymbol{d}}) - F(\boldsymbol{d})) = \sqrt{n} \nabla F(\boldsymbol{d})^\top (\hat{\boldsymbol{d}} - \boldsymbol{d}) + \frac{\sqrt{n}}{2} (\hat{\boldsymbol{d}} - \boldsymbol{d})^\top \nabla^2 F(\boldsymbol{d}^*) (\hat{\boldsymbol{d}} - \boldsymbol{d}), \tag{B.2}$$

where $\boldsymbol{d}^*$ is between $\boldsymbol{d}$ and $\hat{\boldsymbol{d}}$.

Simple calculations show that

$$[\nabla F(\boldsymbol{d})]_{j-1} = \sum_{l=1}^{k} \frac{a_l}{n_l} \sum_{i=1}^{n_l} \frac{a_j \nu_j(X_i^{(l)}) \nu(X_i^{(l)})}{(\sum_{s=1}^{k} a_s \nu_s(X_i^{(l)})/d_s)^2 d_j^2} \xrightarrow{\text{a.s.}} [c(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1} \tag{B.3}$$

where $[c(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1}$ is defined in (3.5). We know that $n/N \to q$. Using similar arguments as in Buta and Doss (2011), it follows that $\nabla^2 F(\boldsymbol{d}^*)$ is bounded in probability. Thus from (B.2) we have

$$\sqrt{n}(F(\hat{\boldsymbol{d}}) - F(\boldsymbol{d})) = \sqrt{q} c(\pi; \boldsymbol{a}, \boldsymbol{d})^\top \sqrt{N}(\hat{\boldsymbol{d}} - \boldsymbol{d}) + o_p(1).$$

Then Theorem 2 (1) follow from (B.1) and the independence of the two stages of Markov chain sampling.

Next to prove Theorem 2 (2), note that, we already have a consistent batch means estimator $\widehat{V}$ of V. From (B.3), we have $[\hat{c}(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1} = [\nabla F(\boldsymbol{d})]_{j-1} \xrightarrow{\text{a.s.}} [c(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1}$. Applying mean value theorem on $[\nabla F(\boldsymbol{d})]_{j-1}$ and the fact that $\nabla^2 F(\boldsymbol{d}^*)$ is bounded in probability, it follows that $[\hat{c}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}})]_{j-1} - [\hat{c}(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1} \xrightarrow{\text{a.s.}} 0$. Writing $c(\pi; \boldsymbol{a}, \boldsymbol{d})^\top V c(\pi; \boldsymbol{a}, \boldsymbol{d})$ as $\sum_{i=1}^{k-1} \sum_{j=1}^{k-1} c_i V_{ij} c_j$, it then follows that $\hat{c}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}})^\top \widehat{V} \hat{c}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) \xrightarrow{\text{a.s.}} c(\pi; \boldsymbol{a}, \boldsymbol{d})^\top V c(\pi; \boldsymbol{a}, \boldsymbol{d})$.

We will now show that $\hat{\tau}_l^2(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}})$ is a consistent estimator of $\tau_l^2(\pi; \boldsymbol{a}, \boldsymbol{d})$ where τ_l^2 and $\hat{\tau}_l^2$ are defined in (3.4) and (3.7) respectively. Since the Markov chains $\{X_i^{(l)}\}_{i=1}^{n_l}$, $l = 1, 2, \ldots, k$ are independent, it then follows that $\tau^2(\pi; \boldsymbol{a}, \boldsymbol{d})$ is consistently estimated by $\hat{\tau}^2(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}})$ completing the proof of Theorem 2 (2).

If $\boldsymbol{d}$ is known from the assumptions of Theorem 2 (2) and the results in Jones et al. (2006) (also Bednorz and Latuszynski (2007)), we know that $\tau_l^2(\pi; \boldsymbol{a}, \boldsymbol{d})$ is consistently estimated by its batch means estimator $\hat{\tau}_l^2(\pi; \boldsymbol{a}, \boldsymbol{d})$. Note that, $\hat{\tau}_l^2(\pi; \boldsymbol{a}, \boldsymbol{d})$ is defined in terms of the quantities $u(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d})$'s. We now show that $\hat{\tau}_l^2(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \hat{\tau}_l^2(\pi; \boldsymbol{a}, \boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$.

Denoting $\hat{\tau}_l^2(\pi; \boldsymbol{a}, \boldsymbol{z})$ by $G(\boldsymbol{z})$, by the mean value theorem (in multiple variables), there exists $\boldsymbol{d}^* = t\hat{\boldsymbol{d}} + (1-t)\boldsymbol{d}$ for some $t \in (0, 1)$, such that

$$G(\hat{\boldsymbol{d}}) - G(\boldsymbol{d}) = \nabla G(\boldsymbol{d}^*) \cdot (\hat{\boldsymbol{d}} - \boldsymbol{d}).$$

For any $j \in \{2, \cdots, k\}$, and $\boldsymbol{z} \in R^{+k-1}$,

$$\frac{\partial G(\boldsymbol{z})}{\partial z_j} = \frac{b_l}{e_l - 1} \left[\sum_{m=0}^{e_l - 1} 2(\bar{u}_m(\boldsymbol{a}, \boldsymbol{z}) - \bar{\bar{u}}(\boldsymbol{a}, \boldsymbol{z})) \left(\frac{\partial \bar{u}_m(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} - \frac{\partial \bar{\bar{u}}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} \right) \right] \tag{B.4}$$

Let $\bar{W}_m := \bar{u}_m(\boldsymbol{a}, \boldsymbol{z}) - E_{\pi_l}(u(X; \boldsymbol{a}, \boldsymbol{z}))$ and $\bar{\bar{W}} := \bar{\bar{u}}(\boldsymbol{a}, \boldsymbol{z}) - E_{\pi_l}(u(X; \boldsymbol{a}, \boldsymbol{z}))$. Note that, there exists, $\sigma^2 < \infty$ such that $\sqrt{b_l}\bar{W}_m \xrightarrow{d} \mathrm{N}(0, \sigma^2)$, and $\sqrt{n_l}\bar{\bar{W}} \xrightarrow{d} \mathrm{N}(0, \sigma^2)$. Simple calculations show that

$$\frac{\partial \bar{u}_m(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} = \frac{a_j}{z_j^2} \frac{1}{b_l} \sum_{i=mb_l+1}^{(m+1)b_l} \left[\frac{\nu(X_i^{(l)}) \nu_j(X_i^{(l)})}{\left(\sum_s a_s \nu_s(X_i^{(l)})/z_s \right)^2}. \right]$$

Hence, letting $\alpha_j = E_{\pi_l}[\nu(X)\nu_j(X)/\left(\sum_s a_s \nu_s(X)/z_s\right)^2]$, we write

$$\frac{\partial \bar{u}_m(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} - \frac{\partial \bar{\bar{u}}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} \equiv \frac{a_j}{z_j^2} \left\{ \bar{Z}_{m,j} \right\} - \frac{a_j}{z_j^2} \left\{ \bar{\bar{Z}}_j \right\},$$

where $\bar{Z}_{1,j} = (1/b_l)\sum_{i=1}^{b_l}[\nu(X_i^{(l)})\nu_j(X_i^{(l)})/\{\sum_s a_s\nu_s(X_i^{(l)})/z_s\}^2] - \alpha_j$ and $\bar{\bar{Z}}_j$ is similarly defined. Note that, there exists $\tau_j^2 < \infty$, such that $\sqrt{b_l}\bar{Z}_{m,j} \xrightarrow{d} \mathrm{N}(0,\tau_j^2)$, and $\sqrt{n_l}\bar{\bar{Z}}_j \xrightarrow{d} \mathrm{N}(0,\tau_j^2)$. From (C.6) we have

$$\begin{aligned}
\frac{\partial G(\boldsymbol{z})}{\partial z_j} =& \frac{a_j}{z_j^2}\frac{2}{e_l-1}\sum_{m=0}^{e_l-1}\left[\sqrt{b_l}(\bar{W}_m-\bar{\bar{W}})\sqrt{b_l}\left(\bar{Z}_{m,j}-\bar{\bar{Z}}_j\right)\right] \\
=& \frac{a_j}{z_j^2}\frac{2}{e-1}\sum_{m=0}^{e-1}\left[\sqrt{b}\bar{W}_m\sqrt{b}\bar{Z}_{m,j}\right] - \frac{a_j}{z_j^2}2b\left[\bar{\bar{Z}}_j\frac{1}{e-1}\sum_{m=0}^{e-1}\bar{W}_m + \bar{\bar{W}}\frac{1}{e-1}\sum_{m=0}^{e-1}\bar{Z}_{m,j} - \frac{e}{e-1}\bar{\bar{W}}\bar{\bar{Z}}_j\right] \\
=& \frac{a_j}{z_j^2}\frac{2}{e-1}\sum_{m=0}^{e-1}\left[\sqrt{b}\bar{W}_m\sqrt{b}\bar{Z}_{m,j}\right] - \frac{a_j}{z_j^2}\frac{2}{e-1}\left[\sqrt{n}\bar{\bar{W}}\sqrt{n}\bar{\bar{Z}}_j\right].
\end{aligned}$$

Then using similar arguments as in the proof of Theorem 1, it can be shown that $\partial G(\boldsymbol{z})/\partial z_j$ is bounded with probability one. Then it follows that

$$|G(\hat{\boldsymbol{d}}) - G(\boldsymbol{d})| \le \max_{1\le j\le k-1}\left\{\left|\frac{\partial G(\boldsymbol{d}^*)}{\partial z_j}\right|\right\}\|\hat{\boldsymbol{d}}-\boldsymbol{d}| \xrightarrow{\text{a.s.}} 0. \tag{B.5}$$

□

C Proof of Theorem 3

Proof. As in the proof of Theorem 2 we write

$$\sqrt{n}(\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}}) - E_\pi f) = \sqrt{n}(\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}}) - \hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d})) + \sqrt{n}(\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d}) - E_\pi f). \tag{C.1}$$

First, consider the 2nd term, which involves randomness only from the 2nd stage. Since

$$\hat{v} \xrightarrow{\text{a.s.}} \sum_{l=1}^{k} a_l E_{\pi_l} v^{[f]}(X;\boldsymbol{a},\boldsymbol{d}) = \int_{\mathsf{X}} \frac{f(x)\sum_{l=1}^k a_l\nu_l(x)/m_l}{\sum_{s=1}^k a_s\nu_s(x)/(m_s/m_1)}\nu(x)\,\mu(dx) = \frac{m}{m_1}E_\pi f, \tag{C.2}$$

we have $\sum_{l=1}^k a_l E_{\pi_l} v^{[f]}(X;\boldsymbol{a},\boldsymbol{d}) = E_\pi f u(\pi,\pi_1)$. Then from (3.1) we have

$$\sqrt{n}\begin{pmatrix} \hat{v}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d}) - E_\pi f u(\pi,\pi_1) \\ \hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d}) - u(\pi,\pi_1)\end{pmatrix} = \sum_{l=1}^{k} a_l\sqrt{\frac{n}{n_l}}\frac{1}{\sqrt{n_l}}\sum_{i=1}^{n_l}\begin{pmatrix} v^{[f]}(X_i^{(l)};\boldsymbol{a},\boldsymbol{d}) - E_{\pi_l}v^{[f]}(X;\boldsymbol{a},\boldsymbol{d}) \\ u(X_i^{(l)};\boldsymbol{a},\boldsymbol{d}) - E_{\pi_l}u(X;\boldsymbol{a},\boldsymbol{d})\end{pmatrix}. \tag{C.3}$$

From the conditions of Theorem 3 and the fact that the Markov chains $\Phi_l, l = 1,\ldots,k$ are independent, it follows that the above vector (C.3) converges in distribution to the bivariate normal distribution with mean 0 and covariance matrix $\Gamma(\pi;\boldsymbol{a},\boldsymbol{d})$ defined in (3.9). Then applying the Delta method to the function $g(x,y) = x/y$ we have a CLT for the ratio estimator $\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d})$, that is, we have $\sqrt{n}(\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d}) - E_\pi f) \xrightarrow{d} N(0,\rho(\pi;\boldsymbol{a},\boldsymbol{d}))$ where $\rho(\pi;\boldsymbol{a},\boldsymbol{d}))$ is defined in (3.10).

Next letting $L(\boldsymbol{z}) = \hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{z})$, by Taylor series expansion of L about $\boldsymbol{d}$ we have

$$\sqrt{n}(L(\hat{\boldsymbol{d}}) - L(\boldsymbol{d})) = \sqrt{n}\nabla L(\boldsymbol{d})^\top(\hat{\boldsymbol{d}}-\boldsymbol{d}) + \frac{\sqrt{n}}{2}(\hat{\boldsymbol{d}}-\boldsymbol{d})^\top\nabla^2 L(\boldsymbol{d}^*)(\hat{\boldsymbol{d}}-\boldsymbol{d}), \tag{C.4}$$

where $\boldsymbol{d}^*$ is between $\boldsymbol{d}$ and $\hat{\boldsymbol{d}}$.

Simple calculations show that

$$[\nabla L(\boldsymbol{d})]_{j-1} = [\hat{e}(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1} \xrightarrow{\text{a.s.}} [e(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1} \tag{C.5}$$

where $[e(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1}$ and $[\hat{e}(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1}$ are defined in (3.11) and (3.12) respectively. It can be shown that $\nabla^2 L(\boldsymbol{d}^*)$ is bounded in probability. Thus from (C.4) we have

$$\sqrt{n}(L(\hat{\boldsymbol{d}}) - L(\boldsymbol{d})) = \sqrt{q} e(\pi; \boldsymbol{a}, \boldsymbol{d})^\top \sqrt{N}(\hat{\boldsymbol{d}} - \boldsymbol{d}) + o_p(1).$$

Then Theorem 3 (1) follow from (C.1) and the independence of the two stages of Markov chain sampling.

Next to prove Theorem 3 (2), note that, we already know that $\widehat{V}$ is a consistent batch means estimator of V. From (C.5), we have $[\hat{e}(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1} \xrightarrow{\text{a.s.}} [e(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1}$. Applying mean value theorem on $[\nabla L(\boldsymbol{d})]_{j-1}$ and the fact that $\nabla^2 L(\boldsymbol{d}^*)$ is bounded in probability, it follows that $[\hat{e}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}})]_{j-1} - [\hat{e}(\pi; \boldsymbol{a}, \boldsymbol{d})]_{j-1} \xrightarrow{\text{a.s.}} 0$.

From (3.8) we know that $\hat{u}(\pi, \pi_1; \boldsymbol{a}, \hat{\boldsymbol{d}}) \xrightarrow{\text{a.s.}} u(\pi, \pi_1)$. From (3.17) we know $\hat{\eta}^{[f]}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) \xrightarrow{\text{a.s.}} E_\pi f$. Since $\hat{v}^{[f]}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d}) = \hat{\eta}^{[f]}(\pi; \boldsymbol{a}, \boldsymbol{d})\hat{u}(\pi, \pi_1; \boldsymbol{a}, \boldsymbol{d})$, it follows that $\hat{v}^{[f]}(\pi, \pi_1; \boldsymbol{a}, \hat{\boldsymbol{d}}) \xrightarrow{\text{a.s.}} E_\pi f u(\pi, \pi_1)$. Thus $\nabla h(\hat{v}^{[f]}(\pi, \pi_1; \boldsymbol{a}, \hat{\boldsymbol{d}}), \hat{u}(\pi, \pi_1; \boldsymbol{a}, \hat{\boldsymbol{d}})) \xrightarrow{\text{a.s.}} \nabla h(E_\pi f u(\pi, \pi_1), u(\pi, \pi_1))$. Thus to prove Theorem 3 (2), we only need to show that $\widehat{\Gamma}_l(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) \xrightarrow{\text{a.s.}} \Gamma_l(\pi; \boldsymbol{a}, \boldsymbol{d})$.

If $\boldsymbol{d}$ is known from the assumptions of Theorem 3 (2) and the results in Flegal et al. (2014), we know that $\Gamma_l(\pi; \boldsymbol{a}, \boldsymbol{d})$ is consistently estimated by its batch means estimator $\widehat{\Gamma}_l(\pi; \boldsymbol{a}, \boldsymbol{d})$. We now show that $\widehat{\Gamma}_l(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \widehat{\Gamma}_l(\pi; \boldsymbol{a}, \boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$.

From Theorem 2 (2), we know that $\hat{\gamma}_l^{22}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \hat{\gamma}_l^{22}(\pi; \boldsymbol{a}, \boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$. We now show $\hat{\gamma}_l^{11}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \hat{\gamma}_l^{11}(\pi; \boldsymbol{a}, \boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$.

Letting $\hat{\gamma}_l^{11}(\pi; \boldsymbol{a}, \boldsymbol{z})$ by $H(\boldsymbol{z})$, by the mean value theorem, there exists $\boldsymbol{d}^* = t\hat{\boldsymbol{d}} + (1-t)\boldsymbol{d}$ for some $t \in (0, 1)$, such that

$$H(\hat{\boldsymbol{d}}) - H(\boldsymbol{d}) = \nabla H(\boldsymbol{d}^*) \cdot (\hat{\boldsymbol{d}} - \boldsymbol{d}).$$

For any $j \in \{2, \cdots, k\}$, and $\boldsymbol{z} \in R^{+k-1}$,

$$\frac{\partial H(\boldsymbol{z})}{\partial z_j} = \frac{b_l}{e_l - 1} \left[\sum_{m=0}^{e_l - 1} 2(\bar{v}_m^{[f]}(\boldsymbol{a}, \boldsymbol{z}) - \bar{\bar{v}}^{[f]}(\boldsymbol{a}, \boldsymbol{z})) \left(\frac{\partial \bar{v}_m^{[f]}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} - \frac{\partial \bar{\bar{v}}^{[f]}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} \right) \right]. \tag{C.6}$$

Let $\bar{W}_m^{[f]} := \bar{v}_m^{[f]}(\boldsymbol{a}, \boldsymbol{z}) - E_{\pi_l}(v^{[f]}(X; \boldsymbol{a}, \boldsymbol{z}))$ and $\bar{\bar{W}}^{[f]} := \bar{\bar{v}}^{[f]}(\boldsymbol{a}, \boldsymbol{z}) - E_{\pi_l}(v^{[f]}(X; \boldsymbol{a}, \boldsymbol{z}))$. Note that, there exists, $\sigma_f^2 < \infty$ such that $\sqrt{b_l}\bar{W}_m^{[f]} \xrightarrow{d} \text{N}(0, \sigma_f^2)$, and $\sqrt{n_l}\bar{\bar{W}}^{[f]} \xrightarrow{d} \text{N}(0, \sigma_f^2)$. Simple calculations show that

$$\frac{\partial \bar{v}_m^{[f]}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} = \frac{a_j}{z_j^2} \frac{1}{b_l} \sum_{i=mb_l+1}^{(m+1)b_l} \left[\frac{f(X_i^{(l)}) \nu(X_i^{(l)}) \nu_j(X_i^{(l)})}{\left(\sum_s a_s \nu_s(X_i^{(l)})/z_s \right)^2} . \right]$$

Hence, letting $\alpha_j^{[f]} = E_{\pi_l}[f(X)\nu(X)\nu_j(X)/\left(\sum_s a_s \nu_s(X)/z_s\right)^2]$, we write

$$\frac{\partial \bar{v}_m^{[f]}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} - \frac{\partial \bar{\bar{u}}^{[f]}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} \equiv \frac{a_j}{z_j^2} \left\{ \bar{Z}_{m,j}^{[f]} \right\} - \frac{a_j}{z_j^2} \left\{ \bar{\bar{Z}}_j^{[f]} \right\},$$

where $\bar{Z}_{1,j}^{[f]} = (1/b_l)\sum_{i=1}^{b_l}[f(X_i^{(l)})\nu(X_i^{(l)})\nu_j(X_i^{(l)})/\{\sum_s a_s\nu_s(X_i^{(l)})/z_s\}^2] - \alpha_j^{[f]}$ and $\bar{\bar{Z}}_j^{[f]}$ is similarly defined. Note that, there exists $\tau_{j,f}^2 < \infty$, such that $\sqrt{b_l}\bar{Z}_{m,j} \xrightarrow{d} \mathrm{N}(0, \tau_{j,f}^2)$, and $\sqrt{n_l}\bar{\bar{Z}}_j \xrightarrow{d} \mathrm{N}(0, \tau_{j,f}^2)$. Doing similar calculations as in the proof of Theorem 2 we have

$$\frac{\partial H(\boldsymbol{z})}{\partial z_j} = \frac{a_j}{z_j^2}\frac{2}{e-1}\sum_{m=0}^{e-1}\left[\sqrt{b}\bar{W}_m^{[f]}\sqrt{b}\bar{Z}_{m,j}^{[f]}\right] - \frac{a_j}{z_j^2}\frac{2}{e-1}\left[\sqrt{n}\bar{\bar{W}}^{[f]}\sqrt{n}\bar{\bar{Z}}_j^{[f]}\right].$$

Then using similar arguments as in the proof of Theorem 2, it can be shown that $\hat{\gamma}_l^{11}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \hat{\gamma}_l^{11}(\pi; \boldsymbol{a}, \boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$. Similarly we can show $\hat{\gamma}_l^{12}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \hat{\gamma}_l^{12}(\pi; \boldsymbol{a}, \boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$. □

References

Bednorz, W. and Latuszynski K. (2007). A few remarks on Fixed-width output analysis for Markov chain Monte Carlo by Jones et al., *Journal of the American Statistical Association* **102** 1485–1486.

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

Doss, H. (2010). Estimation of large families of Bayes factors from Markov chain output. *Statistica 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. *Journal of the Royal Statistical Society,* Series B **76** 683–712.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38:1034–1070.

Flegal, J. M., Jones, G. L., and Vats, D. (2014). Covariance matrix estimation in high-dimensional Markov chain Monte Carlo. *In preparation*.

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

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

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

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101:1537–1547.

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

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

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

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

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

Roy, V. and Evangelou, E. and Zhu, Z. (2014). Efficient estimation and prediction for the Bayesian binary spatial model with flexible link functions Tech. Rep. 568r, Department of Statistics, Iowa State University.

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

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

Tan, A. and Doss, H. and Hobert, J. P. (2015). Honest importance sampling with multiple Markov chains. *Journal of Computational and Graphical Statistics,* (to appear) .

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

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

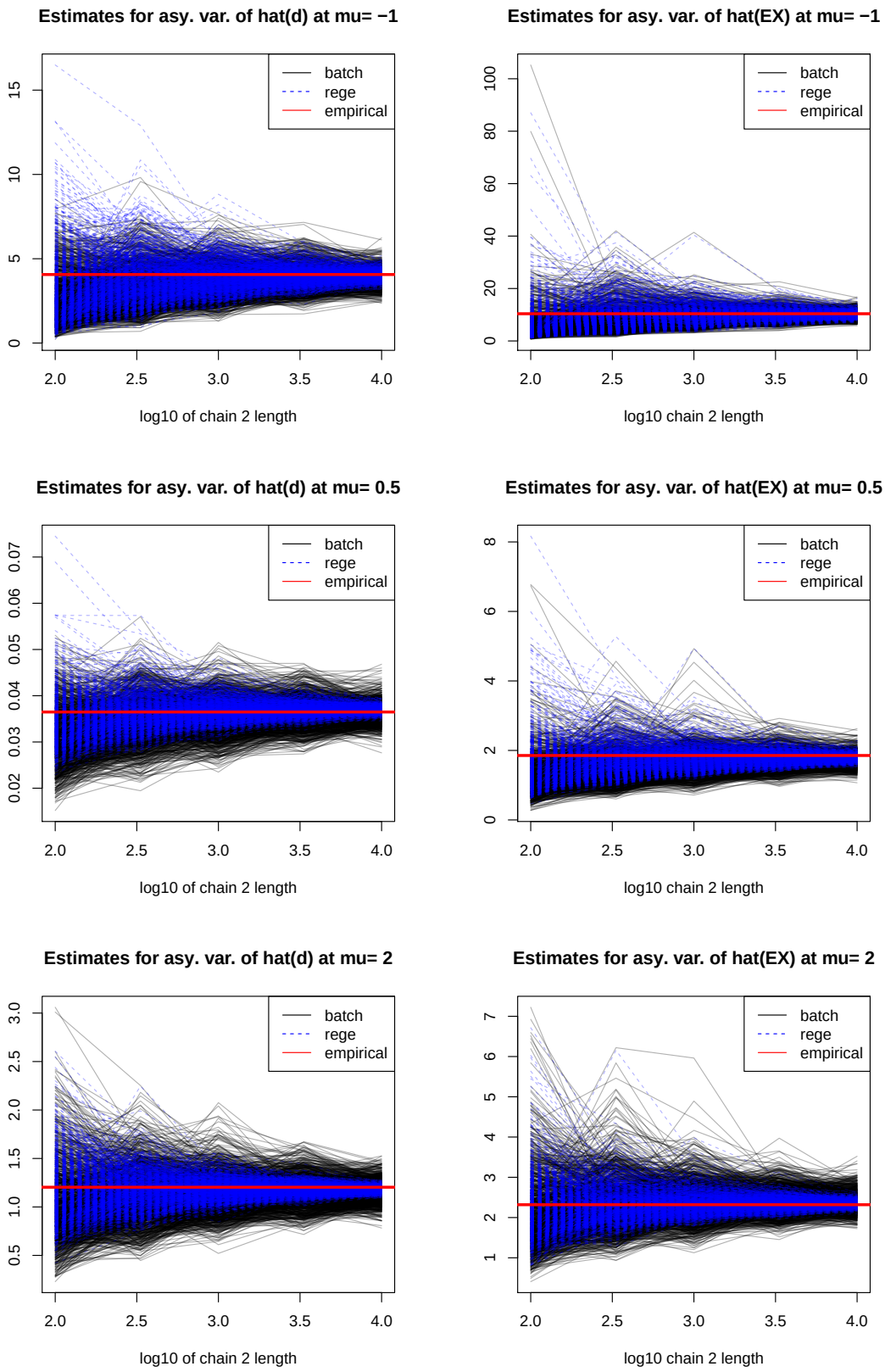

Figure 2: Estimates of the asy. var. of $\hat{d}_\mu$ and $\hat{E}_\mu(X)$ in stage 2, with basic weight $\boldsymbol{a}^2 = (.5, .5)$

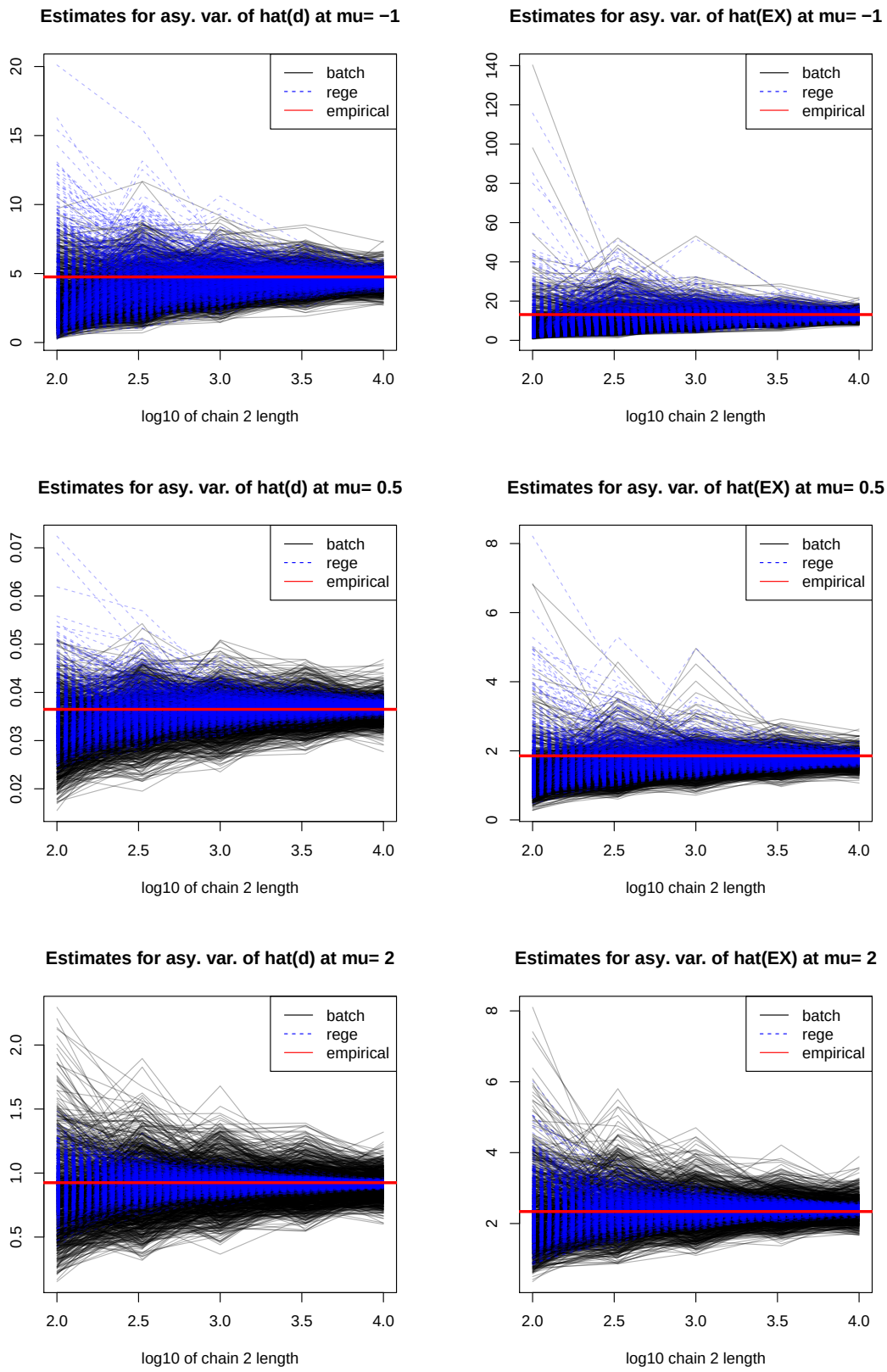

Figure 3: Estimates of the asy. var. of $\hat{d}_\mu$ and $\hat{E}_\mu(X)$ in stage 2, with weight $\boldsymbol{a}^2(\mu)$ chosen by strategy 2.

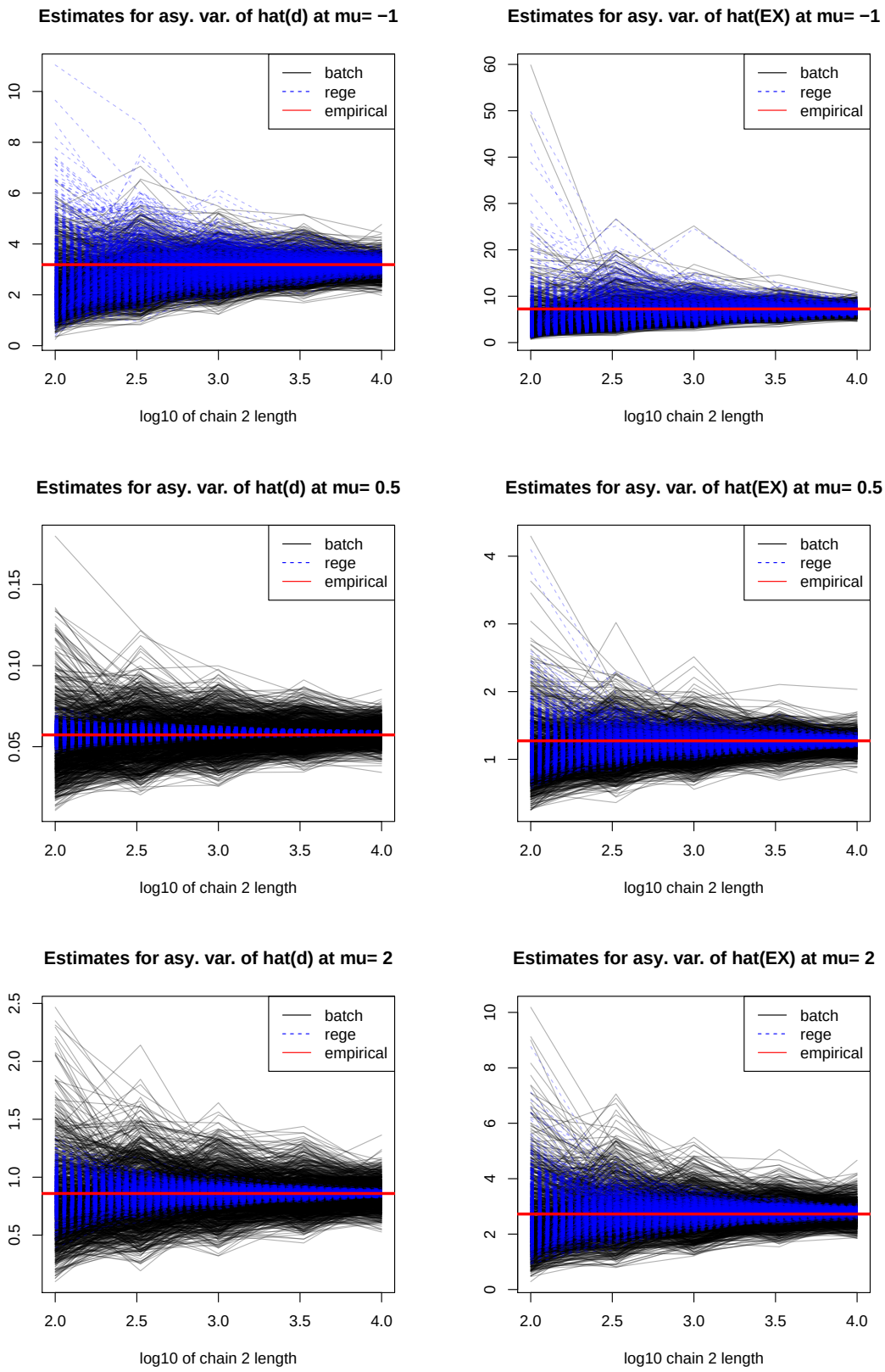

Figure 4: Estimates of the asy. var. of $\hat{d}_\mu$ and $\hat{E}_\mu(X)$ in stage 2, with weight $\boldsymbol{a}^2(\mu)$ chosen by strategy 3.

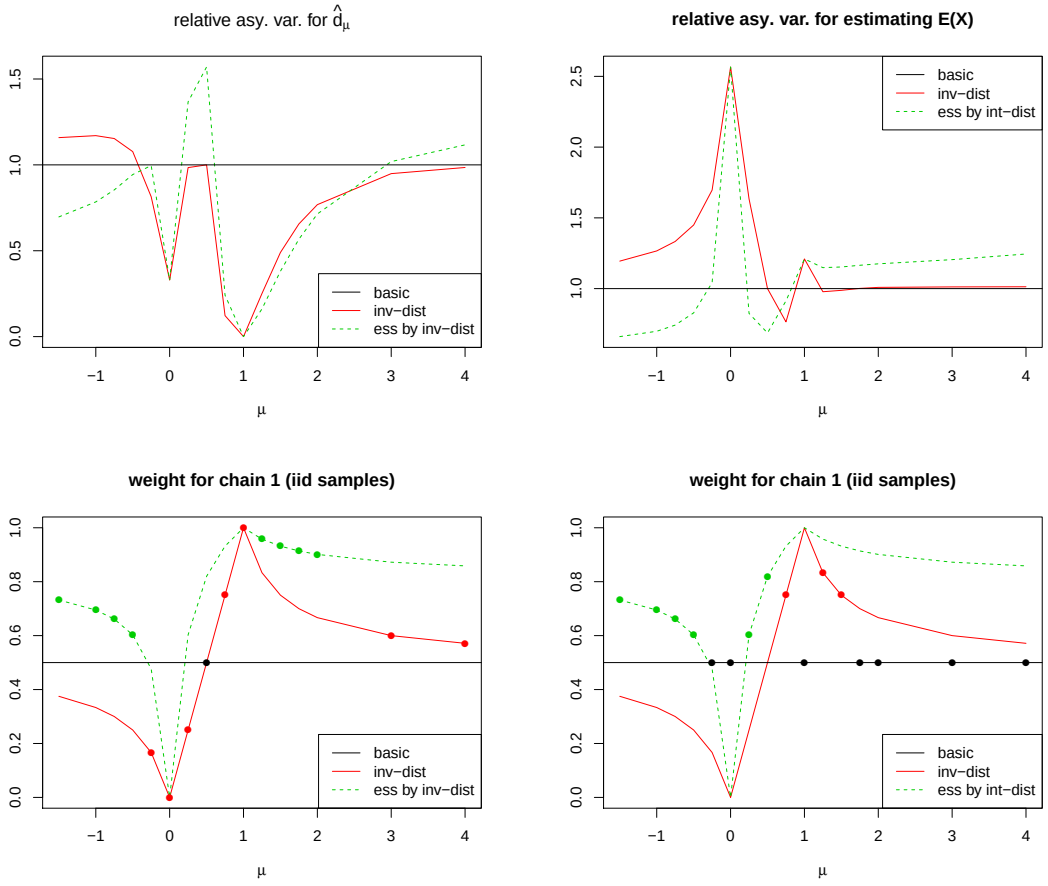

Figure 5: Three strategies of choosing $\boldsymbol{a}^2(\mu)$ in stage 2 are compared, in terms of the asymptotic variance of the corresponding estimators $\hat{d}_\mu$ and $\hat{E}_\mu(X)$. In the bottom two graphs, the color of the solid dots shows which strategy achieves the smallest asymptotic variance among the three at any given μ. In case of ties, the color of the more basic strategy is shown.