Ann. Inst. Statist. Math.
Vol. 54, No. 2, 338–354 (2002)

PARTITION-WEIGHTED MONTE CARLO ESTIMATION

MING-HUI CHEN[1] AND QI-MAN SHAO[2]

[1]*Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, CT 06269-4120, U.S.A.*, e-mail: mhchen@merlot.stat.uconn.edu
[2]*Department of Mathematics, University of Oregon, Eugene, OR 97403-1222, U.S.A.*, e-mail: qmshao@darkwing.uoregon.edu

(Received October 1, 1999; revised July 27, 2000)

Abstract. Although various efficient and sophisticated Markov chain Monte Carlo sampling methods have been developed during the last decade, the *sample mean* is still a dominant in computing Bayesian posterior quantities. The sample mean is simple, but may not be efficient. The weighted sample mean is a natural generalization of the sample mean. In this paper, a new weighted sample mean is proposed by partitioning the support of posterior distribution, so that the same weight is assigned to observations that belong to the same subset in the partition. A novel application of this new weighted sample mean in computing ratios of normalizing constants and necessary theory are provided. Illustrative examples are given to demonstrate the methodology.

Key words and phrases: Bayesian computation, importance sampling, Markov chain Monte Carlo, posterior distribution, simulation.

1. Introduction

In Bayesian inference, Monte Carlo (MC) methods are often used to compute posterior expectation

$$E(h(\theta) \mid D) = \int_{R^d} h(\theta)\pi(\theta \mid D)d\theta, \tag{1.1}$$

where θ is a d-dimensional vector of model parameters, $h(\theta)$ is a real-valued function, D denotes the data, and $\pi(\theta \mid D)$ is the posterior distribution, because the analytical evaluation of $E(h(\theta) \mid D)$ is typically not available. The use of Monte Carlo methods for computing high dimensional integrations has a long history. In the MC literature, one of the excellent early references is Hammersley and Handscomb (1964), and many early MC methods such as *importance sampling* and *conditional Monte Carlo*, which are still useful now, can be found therein. Trotter and Tukey (1956) proposed a general MC scheme based on the weighted average. More specifically, instead of sampling θ alone, they suggested generating a pair (θ, w), where w is a real-valued weight, from some joint distribution $\pi(\theta, w)$ so that for all reasonable real-valued functions $h(\theta)$,

$$\int_{R^{d+1}} wh(\theta)\pi(\theta, w)d\theta dw = E(h(\theta) \mid D). \tag{1.2}$$

The authorship of this article is based on alphabetical order.

We note that by taking $h(\theta) = 1$, (1.2) reduces to

$$\int_{R^{d+1}} w\pi(\theta, w)d\theta dw = 1.$$

In (1.2), $h(\cdot)$ is said to be reasonable if

$$\int_{R^{d+1}} |wh(\theta)|\pi(\theta, w)d\theta dw < \infty \quad \text{and} \quad E(|h(\theta)| \mid D) < \infty.$$

Assuming that $\{(\theta_i, w_i), i = 1, 2, \ldots, n\}$ is a (dependent or independent) sample from $\pi(\theta, w)$, the weighted sample mean of $E(h(\theta) \mid D)$ is given by

$$\hat{E}_w(h) = \frac{1}{n}\sum_{i=1}^{n} w_i h(\theta_i). \tag{1.3}$$

The weighted sample mean is very general, and many important estimators, such as the sample mean and importance sampling, are a special case, in which the weight w_i is fixed, and not random, conditionally on the θ_i's. By taking $w_i = 1$, the weighted sample mean (1.3) reduces to the usual sample mean of the $h(\theta_i)$'s:

$$\hat{E}_{avg}(h) = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i). \tag{1.4}$$

Assume that the posterior density $\pi(\theta \mid D)$ has the form $\pi(\theta \mid D) = \frac{L(\theta|D)\pi(\theta)}{c(D)}$, where $L(\theta \mid D)$ is the likelihood function, $\pi(\theta)$ is the prior, and $c(D)$ is an unknown normalizing constant. Also let $g(\theta)$ be an importance sampling density, which is known up to a normalizing constant. Suppose $\{\theta_i, i = 1, 2; \ldots, n\}$ is a sample from $g(\theta)$. Write the importance sampling weight as

$$w_i = \frac{L(\theta_i \mid D)\pi(\theta_i)/g(\theta_i)}{\frac{1}{n}\sum_{l=1}^{n} L(\theta_l \mid D)\pi(\theta_l)/g(\theta_l)}. \tag{1.5}$$

Then, the weighted sample mean (1.3) with w_i given by (1.5) reduces to an importance sampling estimator:

$$\hat{E}_I(h) = \frac{\sum_{i=1}^{n} h(\theta_i)L(\theta_i \mid D)\pi(\theta_i)/g(\theta_i)}{\sum_{l=1}^{n} L(\theta_l \mid D)\pi(\theta_l)/g(\theta_l)}. \tag{1.6}$$

In (1.6), $g(\theta)$ needs not to be completely known, since the unknown normalizing constant in $g(\theta)$ cancels out in the ratio. The weight given in (1.5) is mainly used to adjust the importance sampling estimator so that it is a consistent estimator of $E(h(\theta) \mid D)$ with respect to $\pi(\theta \mid D)$. In general, it has no guarantee that the importance sampling estimator is better than the sample mean given by (1.4) if a sample directly from $\pi(\theta \mid D)$ is available.

Asymptotic or small sample properties of the weighted sample mean depend on the choice of the joint distribution $\pi(\theta, w)$ and the algorithm used to generate the weighted sample. Under certain regularity conditions such as *ergodicity*, the weighted sample mean

$\hat{E}_w(h)$ is consistent. Ideally, the pair (θ, w) should be sampled jointly from some distribution $\pi(\theta, w)$. In general, it is difficult to construct $\pi(\theta, w)$. The dynamic weighting method of Wong and Liang (1997) is an attempt in this regard. The dynamic weighting method extends the basic Markov chain equilibrium concept of Metropolis *et al.* (1953) to a more general weighted equilibrium of a Markov chain. The basic idea of dynamic weighting is to augment the original sample space by a positive scalar w, called a weight function, which can automatically adjust its own value to help the sampler move more freely. However, many of the weighted transition rules proposed by Liu *et al.* (1998) lead to a marginal weight distribution that is long-tailed. This long-tailed weight distribution makes the resulting weighted sample mean $\hat{E}_w(h)$ converge very slowly.

Instead of sampling (θ, w) jointly from certain distribution $\pi(\theta, w)$, Casella and Robert (1996) proposed a post-simulation improvement for two common Monte Carlo methods, the acceptance-rejection and Metropolis algorithms. The improvement is based on a Rao-Balckwellization method that integrates over the uniform random variables involved in the algorithm. They showed how the Rao-Blackwellized versions of these algorithms can be implemented and how the weights w_i's can be constructed. In the same spirit, Casella and Robert (1998) proposed alternative methods for constructing estimators from accept-reject samples by incorporating the variables rejected by the algorithm. They showed that these estimators are superior asymptotically to the classical accept-reject estimator, which ignores the rejected variables. The Rao-Balckwellization method and the recycling method are intuitively appealing. The weighted sample mean $\hat{E}_w(h)$ can be better than the sample mean $\hat{E}_{avg}(h)$ *only if* the additional information can be used to construct the weight w_i. The additional information can be obtained either from the sampling scheme or from the variables involved in the algorithm. This is very much alike a typical Bayesian analysis. When the informative prior is available, the more accurate posterior estimates can be resulted in.

Unlike the aforementioned methods, we propose a new weighted sample mean by partitioning the support of posterior distribution $\pi(\theta \mid D)$ so that the same weight is assigned to observations that belong to the same subset in the partition. Our approach resembles the stratified sampling method (Thompson (1992)) in the sense that we use the weighted method to partition the sample $\{\theta_i, i = 1, 2, \ldots, n\}$ into several subsets so that within each subset, the $h(\theta_i)$'s are more homogeneous. The technical detail is given in Section 2. In Section 3, we present a novel application of this new weighted sample mean in computing ratios of normalizing constants. Examples are given in Section 4, and we conclude the article with a brief discussion.

2. The partition-weighted estimation

Assume that $\{\theta_1, \theta_2, \ldots, \theta_n\}$ is an independent or dependent stationary sample from $\pi(\theta \mid D)$. Suppose a weighted sample mean is of the form

$$\hat{E}_w(h) = \frac{1}{n}\sum_{i=1}^{n} w_i h(\theta_i), \tag{2.1}$$

where w_i's are the fixed weights subject to

$$\frac{1}{n}\sum_{i=1}^{n} w_i = 1. \tag{2.2}$$

Then, we are led to the following proposition.

PROPOSITION 2.1. *Let Σ denote the covariance matrix of $h(\theta_1), h(\theta_2), \ldots, h(\theta_n)$. Then, the value of $w = (w_1, w_2, \ldots, w_n)'$ that minimizes the variance of $\hat{E}_w(h)$ in* (2.1) *is*

$$w_{opt} = \frac{n\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}, \tag{2.3}$$

where $1 = (1, 1, \ldots, 1)'$, and the optimal weighted sample mean is given by

$$\hat{E}_{opt}(h) = \frac{1}{n}(h(\theta_1), h(\theta_2), \ldots, h(\theta_n))w_{opt} \tag{2.4}$$

with variance

$$\mathrm{Var}(\hat{E}_{opt}(h)) = \frac{\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}. \tag{2.5}$$

The proof of this proposition simply follows from the Lagrange multiplier method, and thus it is omitted for brevity. We notice that this result is also given in Peng (1998).

Remark 2.1. If $\Sigma = \sigma^2 I_n$, then $w_{opt} = \mathbf{1}$, and the optimal weighted sample mean $\hat{E}_{opt}(h)$ reduces to $\hat{E}_{avg}(h) = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i)$ with variance $\frac{\sigma^2}{n}$. Thus, for an i.i.d. sample, the sample mean of the $h(\theta_i)$'s is the best estimator of $E(h(\theta) \mid D)$.

Remark 2.2. Assume that $\{\theta_1, \theta_2, \ldots, \theta_n\}$ is a dependent sample from an AR(1) process with marginal variance σ^2 and lag-one autocorrelation ρ. Consider $h(\theta) = \theta$. Then, the variance of the sample mean of the $h(\theta_i)$'s is given by

$$\mathrm{Var}(\hat{E}_{avg}(h)) = \frac{\sigma^2}{n}\left[\frac{1+\rho}{1-\rho} - \frac{2\rho(1-\rho^n)}{n(1-\rho)^2}\right].$$

Using (2.3) and (2.4), it can be shown that the optimal fixed weighted sample mean is given by

$$\hat{E}_{opt}(h) = \frac{\theta_1 + (1-\rho)\sum_{i=2}^{n-1}\theta_i + \theta_n}{n - (n-2)\rho}$$

with variance

$$\mathrm{Var}(\hat{E}_{opt}(h)) = \frac{\sigma^2(1+\rho)}{n - (n-2)\rho}.$$

Thus, $\lim_{n\to\infty}[\mathrm{Var}(\hat{E}_{opt}(h))/\mathrm{Var}(\hat{E}_{avg}(h))] = 1$, which implies that the sample mean is as efficient as the optimal weighted sample mean asymptotically.

Remarks 2.1 and 2.2 indicate that the weighted sample mean may not substantially improve the simulation efficiency over the sample mean if the weight w_i is fixed (not random). Thus, in order to obtain a better weighted sample mean, the weight w_i should be random or depend on the θ_i's in a particular functional form.

We now propose a new weighted sample mean by partitioning the support of the posterior distribution, and show that this new weighted sample mean can always be better than the sample mean for an i.i.d. sample. Our approach is somewhat related to the stratified sampling method (see, for example, Thompson (1992)), in the sense that

we use the stratified sampling idea to construct the weight. Let $\theta_1, \theta_2, \ldots, \theta_n$ denote n i.i.d. random variables from $\pi(\theta \mid D)$ and let h be a real-valued function. Assume that

$$\mu = E[h(\theta) \mid D] \neq 0 \quad \text{and} \quad \sigma^2 = \text{Var}[h(\theta) \mid D] < \infty,$$

where the expectation and variance are taken with respect to the posterior distribution $\pi(\theta \mid D)$. Let $\Omega \subset R^d$ denote the support of $\pi(\theta \mid D)$, and let $A_1, A_2, \ldots, A_k$ be a partition of Ω such that (i) $\cup_{l=1}^{k} A_1 = \Omega$, (ii) $A_l \cap A_{l^*} = \emptyset$ for $l \neq l^*$, and (iii) $\int_{A_l} \pi(\theta \mid D) d\theta > 0$ for $l = 1, 2, \ldots, k$. Also, let

$$\mu_l = E[h(\theta) 1\{\theta \in A_l\} \mid D] \quad \text{and} \quad b_l = E[h^2(\theta) 1\{\theta \in A_l\} \mid D]. \tag{2.6}$$

Then, a partition-weighted sample mean of $\mu = E[h(\theta) \mid D]$ is given by

$$\hat{E}_a(h) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{k} a_l h(\theta_i) 1\{\theta_i \in A_l\}, \tag{2.7}$$

where $a = (a_1, a_2, \ldots, a_k)'$ is a vector of fixed weights subject to

$$\sum_{l=1}^{k} a_l \mu_l = \mu. \tag{2.8}$$

The constraint given in (2.8) guarantees the unbiasedness of the partition-weighted sample mean $\hat{E}_a(h)$. It follows from straightforward algebra that the variance of $\hat{E}_a(h)$ is given by

$$\text{Var}(\hat{E}_a(h)) = \frac{1}{n} \left(\sum_{l=1}^{k} a_l^2 b_l - \mu^2 \right). \tag{2.9}$$

We note that the partition-weighted sample mean $\hat{E}_a(h)$ is strictly speaking not an estimator, since the weights, a_l's, are unknown in general. We also note that although the a_l's are fixed, the partition-weighted sample mean $\hat{E}_a(h)$ indeed uses the random weights. To see this, we let

$$w_i = \sum_{l=1}^{k} a_l 1\{\theta_i \in A_l\}. \tag{2.10}$$

Then, we can rewrite (2.7) as

$$\hat{E}_a(h) = \frac{1}{n} \sum_{i=1}^{n} w_i h(\theta_i). \tag{2.11}$$

Therefore, w_i is random, and in fact, it is a function of θ_i. This property also distinguishes the partition-weighted sample mean from a usual stratified weighted estimator such as the Horvitz-Thompson estimator (see, Thompson (1992), p. 49), in which a fixed weight is assigned to each $h(\theta_i)$.

Since θ_i's are i.i.d. observations, the sample mean of the $h(\theta_i)$'s has variance σ^2/n. The following theorem states that the partition-weighted sample mean given by (2.7) or (2.11) can be always better than the sample mean.

THEOREM 2.1. *The value of* $a = (a_1, a_2, \ldots, a_k)'$ *that minimizes the variance of* $\hat{E}_a(h)$ *in* (2.9) *is given by*

$$a_{opt,l} = \frac{\mu_l}{b_l} \frac{\mu}{\sum_{j=1}^k \mu_j^2/b_j}, \quad \text{for} \quad l = 1, 2, \ldots, k. \tag{2.12}$$

Let $a_{opt} = (a_{opt,1}, a_{opt,2}, \ldots, a_{opt,k})'$. *Then, the optimal partition-weighted sample mean* $\hat{E}_{a_{opt}}(h)$ *has variance*

$$\mathrm{Var}(\hat{E}_{a_{opt}}(h)) = \frac{1}{n}\left(\frac{\mu^2}{\sum_{l=1}^k \mu_l^2/b_l} - \mu^2\right), \tag{2.13}$$

and

$$\mathrm{Var}(\hat{E}_{a_{opt}}(h)) \le \mathrm{Var}(\hat{E}_{avg}(h)) = \frac{\sigma^2}{n}, \tag{2.14}$$

where μ_l *and* b_l *are defined as in* (2.6).

PROOF. The derivation of the optimal a_{opt} given in (2.12) directly follows from the Lagrange multiplier method. We obtain (2.13) by plugging a_{opt} in (2.9). Noting that $\sum_{l=1}^k b_l = E[h^2(\theta) \mid D]$, the Cauchy-Schwarz inequality yields

$$\begin{aligned}\frac{\mu^2}{\sum_{l=1}^k \mu_l^2/b_l} &= \frac{\mu^2}{(\sum_{l=1}^k |\mu_l|)^2} \frac{(\sum_{l=1}^k b_l^{1/2}|\mu_l|/b_l^{1/2})^2}{\sum_{l=1}^k \mu_l^2/b_l} \\ &\le \frac{\mu^2}{(\sum_{l=1}^k |\mu_l|)^2} \sum_{l=1}^k b_l \le E[h^2(\theta) \mid D].\end{aligned}$$

This shows that the variance of $\hat{E}_{a_{opt}}(h)$ is always less than or equal to σ^2/n. □

We note that the equality in (2.14) holds if and only if

$$b_l = c_0|\mu_l|, \quad \text{for} \quad l = 1, 2, \ldots, k, \tag{2.15}$$

and $|\mu| = \sum_{l=1}^k |\mu_l|$, where $c_0 = E(h^2(\theta) \mid D)/|\mu|$ is a constant.

From (2.13), it can be observed that if $h(\theta_i)$ is close to constant for $\theta_i \in A_l$,

$$\frac{\mu_l^2}{b_l} = \frac{[\int_{A_l} h(\theta)\pi(\theta \mid D)d\theta]^2}{\int_{A_l}[h(\theta)]^2\pi(\theta \mid D)d\theta} \approx \frac{[\int_{A_l} \pi(\theta \mid D)d\theta]^2}{\int_{A_l} \pi(\theta \mid D)d\theta} = \pi(A_l \mid D),$$

where $\pi(A_l \mid D)$ denotes the posterior probability of A_l. Then,

$$\frac{\mu^2}{\sum_{l=1}^k \mu_l^2/b_l} - \mu^2 \approx \frac{\mu^2}{\sum_{l=1}^k \pi(A_l \mid D)} - \mu^2 = \mu^2 - \mu^2 = 0,$$

since $\sum_{l=1}^k \pi(A_l \mid D) = 1$. Thus, the resulting optimal variance is approximately 0. This is a useful feature, which indicates that if the values of $h(\theta_i)$ for $\theta_i \in A_l$ are similar, a smaller variance of $\hat{E}_{a_{opt}}(h)$ can be obtained. In other words, if we partition the sample $\{\theta_i, i = 1, 2, \ldots, n\}$ into several subsets so that within each subset, the $h(\theta_i)$'s are roughly constant, a better partition-weighted sample mean can be resulted in. This is in fact

closely related to stratified sampling, in which homogeneous units are grouped into the same stratum. Moreover, this property can also be used as a guideline to construct the partition $\{A_l, l = 1, 2, \ldots, k\}$. Since $h(\theta)$ is one-dimensional, we can choose a finite partition of the real line R^1, $-\infty = h_0 < h_1 < \cdots < h_{k-1} < h_k = \infty$, and then compute

$$A_l = \{\theta : h_{l-1} < h(\theta) \leq h_l\}, \tag{2.16}$$

for $l = 1, 2, \ldots, k$. We note that if $h(\theta)$ is continuous, $h(\theta)$ is approximately constant over A_l as long as the partition $\{h_l, l = 0, 1, \ldots, k\}$ is fine enough.

In (2.12), the optimal weights, $a_{opt,l}$'s, depend on the unknown parameter μ. The optimal partition-weighted sample mean $\hat{E}_{a_{opt}}(h)$ appears to be not directly useful. However, if $\mu_l = p_l\mu$, where p_l is known or it can be estimated, $\hat{E}_{a_{opt}}(h)$ can be attractive. In fact, this will be the case when our interest is in estimating ratios of normalizing constants. We will discuss this application in the next section in detail. Finally, we note that when $\theta_1, \theta_2, \ldots, \theta_n$ are not independent, a similar result can still be obtained. We give a brief explanation as follows. Let

$$\sigma_{l,v} = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(h(\theta_i)1\{\theta_i \in A_l\}, h(\theta_j)1\{\theta_j \in A_v\})$$

and put $\Sigma = (\sigma_{l,v}, 1 \leq l, v \leq k)$. It is easy to see that

$$\mathrm{Var}(\hat{E}_a(h)) = \frac{1}{n^2} a'\Sigma a. \tag{2.17}$$

With constraint (2.8), it can be shown that the value of $a = (a_1, a_2, \ldots, a_k)'$ that minimizes the variance of $\hat{E}_a(h)$ in (2.17) is given by

$$a_{opt} = \frac{\mu\Sigma^{-1}d}{d'\Sigma^{-1}d}, \tag{2.18}$$

where $d = (\mu_1, \mu_2, \ldots, \mu_k)'$. Moreover, the optimal partition-weighted sample mean $\hat{E}_{a_{opt}}(h)$ has variance

$$\mathrm{Var}(\hat{E}_{a_{opt}}(h)) = \frac{\mu^2}{n^2 d'\Sigma^{-1}d}. \tag{2.19}$$

3. Computing ratios of normalizing constants

Computation of normalizing constants for posterior densities from which we can sample frequently arises in Bayesian inference. Typically, we are interested in ratios of such normalizing constants. For example, suppose we wish to compare two models $\mathcal{M}_1$ and $\mathcal{M}_2$. Let $L(\theta \mid D, \mathcal{M}_j)$ and $\pi(\theta \mid \mathcal{M}_j)$ denote the likelihood function and the prior distribution under model $\mathcal{M}_j$ for $j = 1, 2$. Then, the Bayes factor for comparing model $\mathcal{M}_1$ to model $\mathcal{M}_2$ is given by

$$B = \frac{\int_{R^d} L(\theta \mid D, \mathcal{M}_1)\pi(\theta \mid \mathcal{M}_1)d\theta}{\int_{R^d} L(\theta \mid D, \mathcal{M}_2)\pi(\theta \mid \mathcal{M}_2)d\theta}. \tag{3.1}$$

See Kass and Raftery (1995) for more details. From (3.1), it is easy to see that the Bayes factor is simply a ratio of the normalizing constants of two posterior densities. Estimating

ratios of normalizing constants is extremely challenging and very important, particularly in Bayesian computation. Such problems often arise in likelihood inference, especially in the presence of missing data (Meng and Wong (1996)), in computing intrinsic Bayes factors (Berger and Pericchi (1996)), in Bayesian comparison of econometric models considered by Geweke (1994), and in estimating marginal likelihood (Chib (1995)). Recently, several Monte Carlo methods for estimating normalizing constants have been developed, which include, for example, bridge sampling of Meng and Wong (1996), ratio importance sampling of Chen and Shao (1997), Chib's method for computing marginal likelihood (Chib (1995)), and reverse logistic regression of Geyer (1994).

In this section, we aim to illustrate how the partition-weighted sample mean given by (2.7) or (2.11) can be used for computing the ratio of normalizing constants. Let $\pi_j(\theta \mid D)$, $j = 1, 2$, be two densities, each of which is known up to a normalizing constant:

$$\pi_j(\theta \mid D) = \frac{q_j(\theta)}{c_j}, \qquad \theta \in \Omega_j, \tag{3.2}$$

where D denotes the data, $\Omega_j \subset R^d$ is the support of π_j, and the unnormalized density $q_j(\theta)$ can be evaluated at any $\theta \in \Omega_j$ for $j = 1, 2$. Then, the ratio of two normalizing constants is defined as

$$r = \frac{c_1}{c_2}. \tag{3.3}$$

Let θ be a random variable from π_2. When $\Omega_1 \subset \Omega_2$, we have the identity:

$$r = \frac{c_1}{c_2} = E_2\left\{\frac{q_1(\theta)}{q_2(\theta)}\right\}, \tag{3.4}$$

here and in the sequel, E_2 denotes the expected value with respective to π_2. Let $\{\theta_{2,1}, \theta_{2,2}, \ldots, \theta_{2,n}\}$ be an i.i.d. sample from π_2. Then, the ratio r can be estimated by

$$\hat{r} = \frac{1}{n}\sum_{i=1}^{n} \frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})}. \tag{3.5}$$

A direct calculation yields

$$\mathrm{Var}\,(\hat{r}) = \frac{r^2}{n} E_2\left(\frac{\pi_1(\theta \mid D) - \pi_2(\theta \mid D)}{\pi_2(\theta \mid D)}\right)^2. \tag{3.6}$$

This method is simple and easy to implement. As pointed out by Chen and Shao (1997), $\hat{r}$ is efficient when $\pi_2(\theta \mid D)$ has heavier tails than $\pi_1(\theta \mid D)$. However, when the two densities π_1 and π_2 have very little overlap (i.e., $E_2(\pi_1(\theta \mid D))$ is very small), this method performs poorly.

To improve the simulation efficiency of $\hat{r}$, we use the partition-weighted estimator defined by (2.7) with the optimal weight $a_{opt,l}$ given in (2.12). Let $\{A_l, l = 1, 2, \ldots, k\}$ denote a partition of Ω_2. Using (2.6), we have

$$\mu_l = E_2\left[\frac{q_1(\theta)}{q_2(\theta)} 1\{\theta \in A_l\}\right] = r\int_{A_l} \pi_1(\theta \mid D)d\theta = r\pi_1(A_l \mid D),$$

where $\pi_1(A_l \mid D)$ is the probability of set A_l with respect to π_1. Let $p_l = \pi_1(A_l \mid D)$ for $l = 1, 2, \ldots, k$. The constraint given in (2.8) becomes

$$\sum_{l=1}^{k} a_l p_l = 1. \tag{3.7}$$

By taking $h(\theta) = q_1(\theta)/q_2(\theta)$ in (2.7) and (2.12), the partition-weighted sample mean with the optimal weight a_{opt} reduces to

$$\hat{r}(a_{opt}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{k} a_{opt,l} \left[\frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})} \right] 1\{\theta_{2,i} \in A_l\}, \tag{3.8}$$

where

$$a_{opt,l} = \frac{p_l}{b_l} \frac{1}{\sum_{j=1}^{k} p_j^2/b_j}, \tag{3.9}$$

and

$$b_l = E_2 \left[\left(\frac{q_1(\theta)}{q_2(\theta)} \right)^2 1\{\theta \in A_l\} \right]. \tag{3.10}$$

The variance given by (2.13) can be simplified to

$$\mathrm{Var}(\hat{r}(a_{opt})) = \frac{1}{n} \left(\frac{1}{\sum_{l=1}^{k} p_l^2/b_l} - r^2 \right). \tag{3.11}$$

From (3.11), it can be observed that in the partition-weighted sample mean $\hat{r}(a_{opt})$, the observations with larger probabilities, p_l's, and smaller second moments are assigned more weights. As a contrast, the same weight is assigned to each observation in the estimator $\hat{r}$. In practice, p_l and b_l are unknown. However, they can be estimated by using the standard Monte Carlo method. Suppose $\{\theta_{1,i}, i = 1, 2, \ldots, m\}$ is a random sample from π_1. Then, p_l can be estimated by

$$\hat{p}_l = \frac{1}{m} \sum_{i=1}^{m} 1\{\theta_{1,i} \in A_l\}.$$

For b_l, we can simply use the random sample $\{\theta_{2,i}, i = 1, 2, \ldots, n\}$ to obtain an estimated value. That is,

$$\hat{b}_l = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})} \right]^2 1\{\theta_{2,i} \in A_l\}. \tag{3.12}$$

Replacing p_l and b_l by $\hat{p}_l$ and $\hat{b}_l$ in (3.9), an estimate of $a_{opt,l}$ is given by

$$\hat{a}_{opt,l} = \frac{\hat{p}_l}{\hat{b}_l} \frac{1}{\sum_{j=1}^{k} \hat{p}_j^2/\hat{b}_j}. \tag{3.13}$$

Plugging $\hat{a}_{opt,l}$ into (3.8) yields

$$\hat{r}(\hat{a}_{opt}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{k} \hat{a}_{opt,l} \left[\frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})} \right] 1\{\theta_{2,i} \in A_l\}. \tag{3.14}$$

Here we note that unlike $\hat{r}(a_{opt})$, $\hat{r}(\hat{a}_{opt})$ is an estimator. Thus, we shall refer to $\hat{r}(\hat{a}_{opt})$ as the partition-weighted estimator. It can be shown that $\hat{r}(\hat{a}_{opt})$ is consistent as $n \to \infty$ and $m \to \infty$. Moreover, the next theorem shows that $\hat{r}(\hat{a}_{opt})$ achieves the same variance as that of $\hat{r}(a_{opt})$ given in (3.11) asymptotically.

THEOREM 3.1. *Assume that* $\{\theta_{1,i}, i = 1,2,\ldots,m\}$ *and* $\{\theta_{2,i}, i = 1,2,\ldots,n\}$ *are two independent random samples. If* $n = o(m)$, *then*

$$\lim_{n\to\infty} nE(\hat{r}(\hat{a}_{opt}) - r)^2 = \frac{1}{\sum_{l=1}^{k} p_l^2/b_l} - r^2. \tag{3.15}$$

PROOF. Write

$$\begin{aligned}\hat{r}(\hat{a}_{opt}) - r &= \frac{1}{c_2 \sum_{j=1}^{k} \hat{p}_j^2/\hat{b}_j} \sum_{l=1}^{k} \frac{\hat{p}_l}{\hat{b}_l} \left(\frac{1}{n} \sum_{i=1}^{n} c_2 \hat{a}_{opt,l} \left[\frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})} \right] 1\{\theta_{2,i} \in A_l\} - c_1 \hat{p}_l \right) \\ &:= \frac{1}{c_2 \sum_{j=1}^{k} \hat{p}_j^2/\hat{b}_j} \times R\end{aligned}$$

and

$$\begin{aligned}R &= \sum_{l=1}^{k} \frac{\hat{p}_l}{\hat{b}_l} \left(\frac{1}{n} \sum_{i=1}^{n} c_2 \hat{a}_{opt,l} \left[\frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})} \right] 1\{\theta_{2,i} \in A_l\} - c_1 p_l \right) + c_1 \sum_{l=1}^{k} \frac{\hat{p}_l}{\hat{b}_l} (p_l - \hat{p}_l) \\ &= \sum_{l=1}^{k} \frac{p_l}{\hat{b}_l} \left(\frac{1}{n} \sum_{i=1}^{n} c_2 \hat{a}_{opt,l} \left[\frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})} \right] 1\{\theta_{2,i} \in A_l\} - c_1 p_l \right) \\ &\quad + \sum_{l=1}^{k} \frac{\hat{p}_l - p_l}{\hat{b}_l} \left(\frac{1}{n} \sum_{i=1}^{n} c_2 \hat{a}_{opt,l} \left[\frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})} \right] 1\{\theta_{2,i} \in A_l\} - c_1 p_l \right) + c_1 \sum_{l=1}^{k} \frac{\hat{p}_l}{\hat{b}_l} (p_l - \hat{p}_l) \\ &= \sum_{l=1}^{k} \frac{p_l}{b_l} \left(\frac{1}{n} \sum_{i=1}^{n} c_2 \hat{a}_{opt,l} \left[\frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})} \right] 1\{\theta_{2,i} \in A_l\} - c_1 p_l \right) \\ &\quad + \sum_{l=1}^{k} \left(\frac{p_l}{\hat{b}_l} - \frac{p_l}{b_l} \right) \left(\frac{1}{n} \sum_{i=1}^{n} c_2 \hat{a}_{opt,l} \left[\frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})} \right] 1\{\theta_{2,i} \in A_l\} - c_1 p_l \right) \\ &\quad + \sum_{l=1}^{k} \frac{\hat{p}_l - p_l}{\hat{b}_l} \left(\frac{1}{n} \sum_{i=1}^{n} c_2 \hat{a}_{opt,l} \left[\frac{q_1(\theta_{2,i})}{q_2(\theta_{2,i})} \right] 1\{\theta_{2,i} \in A_l\} - c_1 p_l \right) + c_1 \sum_{l=1}^{k} \frac{\hat{p}_l}{\hat{b}_l} (p_l - \hat{p}_l) \\ &:= R_1 + R_2 + R_3 + R_4.\end{aligned}$$

It follows from the law of large numbers that

$$\frac{1}{c_2 \sum_{j=1}^{k} \hat{p}_j^2/\hat{b}_j} \to \frac{1}{c_2 \sum_{j=1}^{k} p_j^2/b_j} \quad \text{a.s.}$$

By the assumption that $n = o(m)$, we have $E(R_2^2) + E(R_3^2) + E(R_4^2) = o(1/n)$ and

$$\frac{E(R_1^2)}{(c_2 \sum_{j=1}^{k} p_j^2/b_j)^2} = \frac{1}{n} \left(\frac{1}{\sum_{l=1}^{k} p_l^2/b_l} - r^2 \right)$$

by (3.11). This proves (3.15) by the above inequalities. □

The partition-weighted estimator $\hat{r}(\hat{a}_{opt})$ is always better than $\hat{r}$ asymptotically. However, the trade-off here is that we have to pay a price to obtain an additional sample from π_1. Since it is relatively easy to compute $\hat{p}_l$ and $\hat{r}(\hat{a}_{opt})$, the partition-weighted estimator is potentially useful, if $\hat{r}(\hat{a}_{opt})$ leads to a substantial gain in simulation efficiency. We will empirically study the performance of $\hat{r}(\hat{a}_{opt})$ in the next section. We conclude this section with the following remarks.

Remark 3.1. The partition-weighted version of the other Monte Carlo methods, such as bridge sampling and ratio importance sampling, can also be developed. As an illustration, we consider the ratio importance sampling (RIS) estimator of Torrie and Valleau (1977) and Chen and Shao (1997). The RIS estimator is based on the following identity:

$$r = \frac{c_1}{c_2} = \frac{E_\pi\{q_1(\theta)/\pi(\theta)\}}{E_\pi\{q_2(\theta)/\pi(\theta)\}}, \tag{3.16}$$

where the expectation E_π is taken with respect to π and $\pi(\theta)$ is an arbitrary density with the support $\Omega = \Omega_1 \cup \Omega_2$. Given a random sample $\{\theta_1, \theta_2, \ldots, \theta_n\}$ from π, an estimator of r denoted by $\hat{r}_{\mathrm{RIS}}$ is given by

$$\hat{r}_{\mathrm{RIS}} = \hat{r}_{\mathrm{RIS}}(\pi) = \frac{\sum_{i=1}^n q_1(\theta_i)/\pi(\theta_i)}{\sum_{i=1}^n q_2(\theta_i)/\pi(\theta_i)}. \tag{3.17}$$

For any π with the support Ω, $\hat{r}_{\mathrm{RIS}}$ is a consistent estimator of r. Let $\{A_l, l = 1, 2, \ldots, k\}$ denote a partition of $\Omega_1 \cup \Omega_2$. Similar to (3.8), the partition-weighted version of the RIS estimator can be written as

$$\hat{r}_{\mathrm{RIS}}(\pi, a) = \frac{\sum_{i=1}^n \sum_{l=1}^k a_{1,l}[q_1(\theta_i)/\pi(\theta_i)]1\{\theta_i \in A_l\}}{\sum_{i=1}^n \sum_{l=1}^k a_{2,l}[q_2(\theta_i)/\pi(\theta_i)]1\{\theta_i \in A_l\}}, \tag{3.18}$$

where $\{\theta_1, \theta_2, \ldots, \theta_n\}$ is a random sample from π, $a = (a_1, a_2)$, and $a_j = (a_{j,1}, a_{j,2}, \ldots, a_{j,k})'$ is subject to

$$\sum_{l=1}^k a_{j,l} p_{j,l} = 1, \tag{3.19}$$

where $p_{j,l} = \int_{A_l} \pi_j(\theta \mid D)d\theta$, for $l = 1, 2, \ldots, k$, and $j = 1, 2$. It is easy to see that (3.19) is an extension to the constraint given by (3.7). The optimal weight is also available via the minimization of the relative mean-square error defined by

$$\mathrm{RE}^2(\hat{r}_{\mathrm{RIS}}) = \frac{E_\pi(\hat{r}_{\mathrm{RIS}} - r)^2}{r^2}.$$

The detail is omitted here for brevity.

Remark 3.2. Peng (1998) developed an efficient weighted Monte Carlo method for computing the normalizing constants, which are essentially the posterior model probabilities resulted from the stochastic search variable selection method of George and McCulloch (1993). She obtained the fixed weight and data dependent weight estimators of the normalizing constants. She also showed that the weighted estimators are better than the ones proposed by George and McCulloch (1997). However, the support of the posterior distribution considered in Peng (1998) is discrete and finite. The main idea of

her method is to partition a Monte Carlo sample (not the support of posterior distribution) into several subsets, and then she assigned a fixed weight or a random weight to each subset. The noticeable difference between her method and the one proposed in this section is that she partitions the sample, and her partition requires that the subsets be not mutually exclusive. Therefore, her results cannot be directly applied to the problem considered here.

4. Examples

Example 1: *A theoretical illustration*

To get a better understanding of the partition-weighted sample mean developed in Section 3, we conduct a theoretical case study based on two normal densities, in which we know the exact values of the two normalizing constants. Let $q_1(\theta) = \exp(-\theta^2/2)$ and $q_2(\theta) = \exp(-(\theta-\delta)^2/2)$ with δ a known positive constant. In this case, $c_1 = c_2 = \sqrt{2\pi}$ and, therefore, $r = 1$. Since θ is one-dimensional, we are able to compute the weights exactly in this example.

For the optimal partition-weighted sample mean $\hat{r}(a_{opt})$ given by (3.8), we consider the following partitions:

(i) $k = 2$, $A_1 = (-\infty, 0]$ and $A_2 = (0, \infty)$;

(ii) $k > 2$, $A_1 = (-\infty, 0]$, $A_l = ((l-2)/(k-2) \times 1.5\delta, (l-1)/(k-2) \times 1.5\delta]$, $l = 2, 3, \ldots, k-1$, and $A_k = (1.5\delta, \infty)$.

For (i), it can be shown that

$$n \operatorname{Var}(\hat{r}(a_{opt})) = \exp(\delta^2) 4\Phi(\delta)(1 - \Phi(\delta)) - 1,$$

where Φ is the standard normal ($N(0,1)$) cumulative distribution function (cdf). We note that for $\hat{r}$ given by (3.5),

$$n \operatorname{Var}(\hat{r}) = \exp(\delta^2) - 1.$$

Table 1 shows the values of $n \operatorname{Var}(\hat{r}(a_{opt}))$ and $n \operatorname{Var}(\hat{r})$ for several different choices of δ and k. In addition to $\operatorname{Var}(\hat{r}(a_{opt}))$, we also compute $n \operatorname{Var}(\hat{r}(\hat{a}_{opt}))$, where $\hat{r}(\hat{a}_{opt})$ is given in (3.14), using the usual multiple simulation technique, in order to get a sense of how close to optimal the practical implementation can get. Specifically, we simulate M samples of size n from $N(\delta, 1)$ and compute $\hat{r}(\hat{a}_{opt})$ for each simulated sample. Let $\hat{r}_j(\hat{a}_{opt})$ denote the value of $\hat{r}(\hat{a}_{opt})$ from the j-th simulation for $j = 1, 2, \ldots, M$. Then, an estimate of $n \operatorname{Var}(\hat{r}(\hat{a}_{opt}))$ is given by

$$n\widehat{\operatorname{Var}}(\hat{r}(\hat{a}_{opt})) = \frac{n}{M-1} \sum_{j=1}^{M} (\hat{r}_j(\hat{a}_{opt}) - \bar{\hat{r}}(\hat{a}_{opt}))^2,$$

where $\bar{\hat{r}}(\hat{a}_{opt}) = \frac{1}{M} \sum_{j=1}^{M} \hat{r}_j(\hat{a}_{opt})$. The results based on $M = 5,000$ and $n = 10,000$ are reported in Table 1. Although we use the same sample to estimate b_l, the estimated values $n\widehat{\operatorname{Var}}(\hat{r}(\hat{a}_{opt}))$ are fairly close to the theoretical optimal values $n \operatorname{Var}(\hat{r}(a_{opt}))$ for most cases except for $\delta = 3$ and $k = 2$. We note that $\bar{\hat{r}}(\hat{a}_{opt})$ matches the true value $r = 1$ for all cases. We also tried other values of n. For example, when $n = 1,000$, for $k = 2$ and 5, $n\widehat{\operatorname{Var}}(\hat{r}(\hat{a}_{opt})) = 0.458$ and 0.127 for $\delta = 1$, and $n\widehat{\operatorname{Var}}(\hat{r}(\hat{a}_{opt})) = 4.435$ and 0.441 for $\delta = 2$, respectively. However, when $n = 1,000$, $n\widehat{\operatorname{Var}}(\hat{r}(\hat{a}_{opt}))$ is much larger

Table 1. Comparison of variances.

δ	$n\,\mathrm{Var}(\hat{r})$	k	$n\,\mathrm{Var}(\hat{r}(a_{opt}))$	$n\widehat{\mathrm{Var}}(\hat{r}(\hat{a}_{opt}))$
1	1.718	2	0.451	0.447
		5	0.116	0.118
		10	0.105	0.108
		20	0.103	0.107
2	53.598	2	3.855	3.872
		5	0.343	0.342
		10	0.107	0.112
		20	0.073	0.077
3	8102.084	2	42.694	66.603
		5	1.250	1.418
		10	0.242	0.297
		20	0.069	0.113

than the theoretical optimal value $n\,\mathrm{Var}(\hat{r}(a_{opt}))$, but much smaller than $n\,\mathrm{Var}(\hat{r})$ for $\delta = 3$.

From Table 1, we can see that the partition-weighted sample mean $\hat{r}(a_{opt})$ dramatically improves the simulation efficiency over the sample mean $\hat{r}$. For example, when $\delta = 3$, with $k = 20$, $\mathrm{Var}(\hat{r})/\mathrm{Var}(\hat{r}(a_{opt})) = 117,421.51$, i.e., $\hat{r}(a_{opt})$ is about 117,421 times better than $\hat{r}$. Also, it is interesting to see that a finer partition yields a smaller variance. When the two densities are not far apart from each other, the variances of the partition-weighted sample means are quite robust for $k \geq 5$. However, when the two densities do not have much overlap, which is the case when $\delta = 3$, a substantial gain in simulation efficiency can be achieved by a finer partition.

Chen and Shao (1997) also used the same example to study the performance of several Monte Carlo methods for estimating the ratio of normalizing constants. In particular, they compared the importance sampling method, the bridge sampling method of Meng and Wong (1996), the path sampling method of Gelman and Meng (1998), and the ratio importance sampling method. We note that $\hat{r}$ given by (3.5) is indeed the importance sampling estimator. Chen and Shao (1997) showed that the ratio importance sampling estimator $\hat{r}_{\mathrm{RIS}}$ given by (3.17) with the optimal π achieves the smallest asymptotic relative mean-square error, while the importance sampling estimator $\hat{r}$ leads to the worst simulation efficiency. By minimizing $\lim_{n\to\infty} n\mathrm{RE}^2(\hat{r}_{\mathrm{RIS}})$, the cdf corresponding to the optimal density π_{opt} for this example is given by

$$\Pi_{opt}(\theta) = \begin{cases} (\Phi(\theta) - \Phi(\theta-\delta))/2(2\Phi(\delta/2)-1) & \text{for } \theta \leq \delta/2 \\ 1 - (\Phi(\theta) - \Phi(\theta-\delta))/2(2\Phi(\delta/2)-1) & \text{for } \theta > \delta/2. \end{cases}$$

With the optimal density π_{opt}, Chen and Shao (1997) obtained

$$\lim_{n\to\infty} n\mathrm{RE}^2(\hat{r}_{\mathrm{RIS}}(\pi_{opt})) = [2(2\Phi(\delta/2)-1)]^2.$$

It is easy to compute that $\lim_{n\to\infty} n\mathrm{RE}^2(\hat{r}_{\mathrm{RIS}}(\pi_{opt})) = 0.587$, 1.864, and 3.002 for $\delta = 1, 2, 3$, respectively. Thus, from Table 1, it can be observed that $\hat{r}(a_{opt})$ is better than the optimal ratio importance sampling estimator when $k \geq 5$. This theoretical illustration

is quite interesting, which tells us that the weighted version of the worst estimator can be much better than the best estimator in terms of their variances.

Example 2: *AIDS study*

In this example, we consider a data set from the AIDS study ACTG036. The ACTG036 study was a placebo-controlled clinical trial comparing AZT to placebo in patients with hereditary coagulation disorders. The results of this study have been published by Merigan *et al.* (1991). The sample size in this study, excluding cases with missing data, was 183. The response variable (y) for these data is binary with a 1 indicating death, development of AIDS, or AIDS related complex (ARC), and a 0 indicates otherwise. Several covariates were measured for these data. The ones we use here are CD4 count (x_1), age (x_2), treatment (x_3), and race (x_4). A summary of the ACTG036 data can be found in Chen *et al.* (1999). Chen *et al.* (1999) analyzed the ACTG036 data using a logistic regression model.

Here, we use the Bayes factor approach (see, for example, Kass and Raftery (1995)) to compare the logit model to the complementary log-log link model. This comparison is of practical interest, since it is not clear whether a symmetric link model is adequate to fit this data set. Let $F_1(t) = \exp(t)/(1+\exp(t))$ and $F_2(t) = 1 - \exp(-\exp(t))$. Also, let $D = (y, X)$ denote the observed data, where $y = (y_1, y_2, \ldots, y_{183})'$ and X is the design matrix with its i-th row $x_i' = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4})$. The likelihood functions corresponding to these two links can be written as

$$L_j(\theta \mid D) = \prod_{i=1}^{183} F_j^{y_i}(x_i'\theta)[1 - F_j(x_i'\theta)]^{1-y_i},$$

for $j = 1, 2$, where $\theta = (\theta_0, \theta_1, \ldots, \theta_4)'$ denotes a 5×1 vector of regression coefficients. We take the same improper uniform prior for θ under both models. Then, the Bayes factor for comparing F_1 to F_2 can be calculated as follows:

$$B = \frac{\int_{R^5} L_1(\theta \mid D)d\theta}{\int_{R^5} L_2(\theta \mid D)d\theta} \equiv \frac{c_1}{c_2}, \tag{4.1}$$

where c_j is the normalizing constant of the posterior distribution under F_j for $j = 1, 2$. Clearly, the Bayes factor B is a ratio of two normalizing constants.

We use the Gibbs sampler to sample from the posterior distribution $\pi_2(\theta \mid D) \propto L_2(\theta \mid D)$. The autocorrelations for all the parameters disappear after lag 5. We obtain a sample of size $n = 1000$ by taking every 10-th Gibbs iteration. Then, using (3.5) and (3.6), we obtain $\hat{B} = 1.161$ and $n\widehat{\mathrm{Var}}(\hat{B}) = 1.331$. In addition, we compute the ratio $h(\theta_i) = L_1(\theta_i \mid D)/L_2(\theta_i \mid D)$ for each observation. The histogram of these 1000 ratios are displayed in Fig. 1. Figure 1 clearly indicates that the posterior distribution of $h(\theta)$ is very skewed to the right. This suggests that the sample mean $\hat{B}$ cannot be reliable/accurate.

To obtain the partition-weighted estimate of B, we consider the following two partitions:

(i) $k = 5$, $A_1 = \{\theta : 0 < h(\theta) \leq 0.75\}$, $A_2 = \{\theta : 0.75 < h(\theta) \leq 1.5\}$, $A_3 = \{\theta : 1.5 < h(\theta) \leq 2.5\}$, $A_4 = \{\theta : 2.5 < h(\theta) \leq 3.5\}$, and $A_5 = \{\theta : 3.5 < h(\theta)\}$.

(ii) $k = 10$, $A_1 = \{\theta : 0 < h(\theta) \leq 0.75\}$, $A_2 = \{\theta : 0.75 < h(\theta) \leq 1.0\}$, $A_3 = \{\theta : 1.0 < h(\theta) \leq 1.25\}$, $A_4 = \{\theta : 1.25 < h(\theta) \leq 1.5\}$, $A_5 = \{\theta : 1.5 < h(\theta) \leq 2.0\}$,

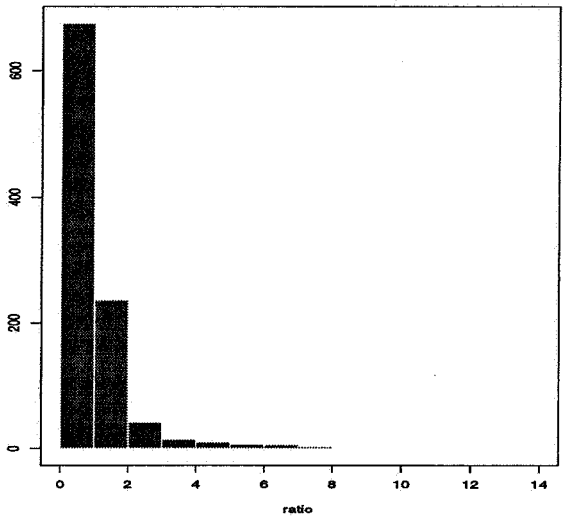

Fig. 1. The histogram of the ratio $h(\theta_i)$.

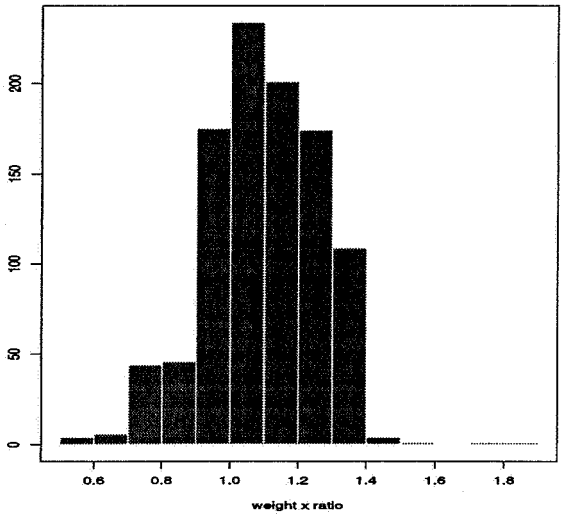

Fig. 2. The histogram of the weight-times-ratio.

$A_6 = \{\theta : 2.0 < h(\theta) \leq 2.5\}$, $A_7 = \{\theta : 2.5 < h(\theta) \leq 3.0\}$, $A_8 = \{\theta : 3.0 < h(\theta) \leq 3.5\}$, $A_9 = \{\theta : 3.5 < h(\theta) \leq 4.0\}$, and $A_{10} = \{\theta : 4.0 < h(\theta)\}$.

Since one assumes that inference for both models 1 and 2 (logit and complementary log-log links) has been conducted, the sample from $\pi_1(\theta \mid D) \propto L_1(\theta \mid D)$ is typically already available. In this regard, we obtain another sample of size $n = 1{,}000$ from $\pi_1(\theta \mid D)$ by taking every 10-th Gibbs iteration, and this sample is then used to estimate the probability p_l under each partition. Now, using the sample from $\pi_2(\theta \mid D)$, which has been already generated earlier, along with (3.12), (3.13), (3.14), and (3.11), we obtain that $\hat{B}(\hat{a}_{opt})$ and $n\widehat{\mathrm{Var}}(\hat{B}(\hat{a}_{opt}))$ are 1.101 and 0.047 for $k = 5$, and 1.098 and 0.027 for

$k = 10$. For each observation, we also compute $w_i h(\theta_i)$ (weight-times-ratio) for $k = 10$, and the histogram of these 1000 values are displayed in Fig. 2. From Fig. 2, the weighted observations are quite symmetric around the mean value.

As discussed in the early sections, both $\frac{1}{n}\sum_{i=1}^{n} h(\theta_i)$ and $\frac{1}{n}\sum_{i=1}^{n}[w_i h(\theta_i)]$ are unbiased or asymptotically unbiased. However, the distribution of $h(\theta_i)$ is skewed, while the distribution of $w_i h(\theta_i)$ is roughly symmetric. The outliers from a skewed distribution can greatly inflate the quality of the sample mean $\frac{1}{n}\sum_{i=1}^{n} h(\theta_i)$. In fact, several big outliers have been observed in the $h(\theta_i)$'s. Moreover, the range of the histogram shown in Fig. 1 is much bigger than the one in Fig. 2. These results partially explain the reason why the weighted method works better.

Since both samples from π_1 and π_2 are available, we can use bridge sampling (BS) of Meng and Wong (1996) for estimating B. Let $\hat{B}_{\mathrm{BS}}$ denote the optimal BS estimate of B. We obtain that $\hat{B}_{\mathrm{BS}} = 1.152$ and $n\widehat{\mathrm{Var}}(\hat{B}_{\mathrm{BS}}) = 0.282$. Thus, the BS estimator is more efficient than the importance sampling estimator given by (3.5), and slightly less efficient than the partition-weighted estimator. Finally, we note that based on the estimated Bayes factor, the logit model is slightly better than the complementary log-log link model.

5. Discussion

In this article, we proposed a partition-weighted sample mean along with an application in computing ratios of normalizing constants. In Section 4, we empirically demonstrated that the partition-weighted sample mean can dramatically improve simulation efficiency. Our proposed weighted method is based on the partition of the support of the posterior distribution. The choice of the partition is somehow arbitrary. Our general recommendation is to form a partition so that the function values of $h(\theta)$ are as close as possible within each subset in the partition. As illustrated in Section 4, we used two different approaches to construct the partition. The first approach works well when θ is a univariate scalar, and the second approach provides an illustration of how to formulate a partition when θ is multidimensional. Our experience also suggests that it be adequate to choose the size of partition (k) to be between 5 to 20.

The weighted method provides us a new approach to improve simulation efficiency in computing posterior quantities of interest. Although we only gave an illustration of how to compute ratios of normalizing constants, the partition-weighted sample mean can be applied to many other Bayesian computations, such as marginal posterior density estimation, and posterior model probability calculation for Bayesian variable selection.

Acknowledgements

The authors wish to thank the Editor and the three referees for their helpful comments and suggestions, which have led to an improvement in this article. Dr Chen's research was partially supported by the National Science Foundation under Grant No. DMS-9702172 and Dr Shao's research was partially supported by the National Science Foundation under Grant No. DMS-9802451.

References

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction, *J. Amer. Statist. Assoc.*, **91**, 109–122.

Casella, G. and Robert, C. P. (1996). Rao-Blackwellization of sampling schemes, *Biometrika*, **83**, 81–94.

Casella, G. and Robert, C. P. (1998). Post-processing accept-reject samples: Recycling and rescaling, *J. Comput. Graph. Statist.*, **7**, 139–157.

Chen, M.-H. and Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants, *Ann. Statist.*, **25**, 1563–1594.

Chen, M.-H., Ibrahim, J. G. and Yiannoutsos, C. (1999). Prior elicitation, variable selection, and Bayesian computation for logistic regression models, *J. Roy. Statist. Soc. Ser. B*, **61**, 223–242.

Chib, S. (1995). Marginal likelihood from the Gibbs output, *J. Amer. Statist. Assoc.*, **90**, 1313–1321.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling, *Statist. Sci.*, **13**, 163–185.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling, *J. Amer. Statist. Assoc.*, **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection, *Statist. Sinica*, **7**, 339–373.

Geweke, J. (1994). Bayesian comparison of econometric models, Tech. Report, No. 532, Federal Reserve Bank of Minneapolis and University of Minnesota.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo, Revision of Tech. Report No. 568, School of Statistics, University of Minnesota.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*, Methuen, London.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *J. Amer. Statist. Assoc.*, **90**, 773–795.

Liu, J. S., Liang, F. and Wong, W. H. (1998). Dynamic weighting in Markov chain Monte Carlo, Tech. Report, Department of Statistics, Stanford University, California.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration, *Statist. Sinica*, **6**, 831–60.

Merigan, T. C., Amato, D. A., Balsley, J., Power, M., Price, W. A., Benoit, S., Perez-Michael, A., Brownstein, A., Kramer, A. S., Brettler, D., Aledort, L., Ragni, M. V., Andes, A. W., Gill, J. C., Goldsmith, J., Stabler, S., Sanders, N., Gjerset, G., Lusher, J. and the NHF-ACTG036 Study Group (1991). Placebo-controlled trial to evaluate zidovudine in treatment of human immunodeficiency virus infection in asymptomatic patients with hemophilia, *Blood*, **78**, 900–906.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087–1092.

Peng, L. (1998). Normalizing constant estimation for discrete distribution simulation, Ph. D. Dissertation, Department of Management Science and Information System, University of Texas at Austin (unpublished).

Thompson, S. K. (1992). *Sampling*, Wiley, New York.

Torrie, G. M. and Valleau, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling, *Journal of Chemical Physics*, **23**, 187–199.

Trotter, H. F. and Tukey, J. W. (1956). Conditional Monte Carlo for normal samples, *Symposium on Monte Carlo Methods* (ed. H. A. Meyer), Wiley, New York.

Wong, W. H. and Liang, F. (1997). Dynamic weighting in Monte Carlo and optimization, *Proceedings of the National Academy of Science*, **94**, 14220–14224.