Statistica Sinica **7**(1997), 607-630

ESTIMATING RATIOS OF NORMALIZING CONSTANTS FOR DENSITIES WITH DIFFERENT DIMENSIONS

Ming-Hui Chen and Qi-Man Shao

Worcester Polytechnic Institute and University of Oregon

Abstract: In Bayesian inference, a Bayes factor is defined as the ratio of posterior odds versus prior odds where posterior odds is simply a ratio of the normalizing constants of two posterior densities. In many practical problems, the two posteriors have different dimensions. For such cases, the current Monte Carlo methods such as the bridge sampling method (Meng and Wong (1996)), the path sampling method (Gelman and Meng (1994)), and the ratio importance sampling method (Chen and Shao (1997)) cannot directly be applied. In this article, we extend importance sampling, bridge sampling, and ratio importance sampling to problems of different dimensions. Then we find global optimal importance sampling, bridge sampling, and ratio importance sampling in the sense of minimizing asymptotic relative mean-square errors of estimators. Implementation algorithms, which can asymptotically achieve the optimal simulation errors, are developed and two illustrative examples are also provided.

Key words and phrases: Bayesian computation, Bayes factor, bridge sampling, Gibbs sampler, importance sampling, Markov chain Monte Carlo, Metropolis algorithm, ratio importance sampling.

1. Introduction

Kass and Raftery (1995) illustrated a simple problem for testing the two hypotheses H_1 and H_2. When the hypotheses H_1 and H_2 are equally probable *a priori* so that $P(H_1) = P(H_2) = 0.5$, then the Bayes factor is

$$B = \frac{m(x|H_1)}{m(x|H_2)}, \tag{1.1}$$

where x is the data and

$$m(x|H_i) = \int_{R^{d_i}} L(x|\theta_i, H_i)\pi(\theta_i|H_i)d\theta_i,$$

where θ_i is a $d_i \times 1$ parameter vector under H_i, $\pi(\theta_i|H_i)$ is the prior density, $L(x|\theta_i, H_i)$ is the likelihood function of θ_i and $m(x|H_i)$ the marginal likelihood function for $i = 1, 2$. (See Jeffreys (1961), Chap. 5 for various examples of this simple Bayesian hypothesis testing problem.) Clearly, Bayes factor B is a ratio

of two normalizing constants of two unnormalized densities $L(x|\theta_i, H_i)\pi(\theta_i|H_i)$, $i = 1, 2$, respectively. Note that when $d_1 \neq d_2$, we are dealing with a problem of two different dimensions.

Verdinelli and Wasserman (1996) also considered a similar problem for testing precise null hypotheses using Bayes factors when nuisance parameters are present. Consider the parameter $(\theta, \psi) \in \Omega \times \Psi$, where ψ is a nuisance parameter, and test the null hypothesis H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$. Then they obtain the Bayes factor $B = m_0/m$ where $m_0 = \int_\Psi L(x|\theta_0, \psi)\pi(\theta_0)d\psi$ and $m = \int_{\Omega\times\Psi} L(x|\theta, \psi)\pi(\theta, \psi)d\theta d\psi$ (Jeffreys (1961), Chap. 5). Here $L(x|\theta, \psi)$ is the likelihood function given data x and $\pi(\theta_0)$ and $\pi(\theta, \psi)$ are the priors. Therefore, the Bayes factor B is a ratio of two normalizing constants again. In this case, one density is a function of ψ and the other density is a function of θ and ψ.

From the above two examples, we can form a general problem for computing ratios of two normalizing constants with different dimensions. Let $\theta = (\theta_{(1)}, \ldots, \theta_{(k)})$ and $\psi = (\psi_{(1)}, \ldots, \psi_{(d)})$. Also let $\pi_1(\theta)$ be a density which is known up to a normalizing constant:

$$\pi_1(\theta) = \frac{p_1(\theta)}{c_1}, \quad \theta \in \Omega_1, \tag{1.2}$$

where $\Omega_1 \subset R^k$ is the support of π_1 and let $\pi_2(\theta, \psi)$ be another density which is known up to a normalizing constant:

$$\pi_2(\theta, \psi) = \frac{p_2(\theta, \psi)}{c_2}, \quad (\theta, \psi) \in \Theta_2, \tag{1.3}$$

where $\Theta_2 \subset R^{k+d}$ $(d \geq 1)$ is the support of π_2. We also denote

$$\Omega_2 = \Big\{\theta : \exists\, \psi \in R^d \text{ such that } (\theta, \psi) \in \Theta_2\Big\} \text{ and } \Psi(\theta) = \Big\{\psi : (\theta, \psi) \in \Theta_2\Big\} \text{ for } \theta \in \Omega_2. \tag{1.4}$$

Then the ratio of two normalizing constants is defined as

$$r = \frac{c_1}{c_2}. \tag{1.5}$$

As Gelman and Meng (1994) pointed out, analytic approximation, numerical integration, and Monte Carlo simulation are three common approaches for computing the above intractable ratio of normalizing constants. However, Monte Carlo simulation is widely used especially in Bayesian statistics, mainly because of its general applicability (for example, no restrictions on the dimensionality). Recently, Meng and Wong (1996) proposed bridge sampling, Gelman and Meng

(1994) developed path sampling for estimating the ratio of two normalizing constants and Geyer (1994) proposed reverse logistic regression for obtaining normalizing constants. Chen and Shao (1997) gave a brief overview of current Monte Carlo methods and they further developed ratio importance sampling for estimating the ratio of two normalizing constants. However, all the previous methods cannot be directly applied to cases where the two densities have different dimensions. To see this fact, we can check the simplest importance sampling method (see, for example, Meng and Wong (1996) or Chen and Shao (1997)). The key identity for the simplest importance sampling method

$$r = \frac{c_1}{c_2} = E_{\pi_2}\Big\{\frac{p_1(\theta)}{p_2(\theta,\psi)}\Big\} \tag{1.6}$$

does not hold in general, unless under certain conditions, for example, $\int_{\Psi(\theta)} d\psi = 1$ for all $\theta \in \Omega_2$. Here, E_{π_2} denotes the expectation with respect to π_2. This convention will be used throughout this paper. Further, it is difficult to construct a path to link π_1 and π_2 due to different dimensionality. Therefore, it is not feasible to apply path sampling for estimating the ratio r given in (1.5).

In order to compute the Bayes factor given in (1.1), Newton and Raftery (1994) proposed several Monte Carlo methods to estimate $m(x|H_1)$ and $m(x|H_2)$ individually and then to estimate the Bayes factor. Their methods are essentially special cases of ratio importance sampling (Chen and Shao (1997)). If the main interest is to compute the Bayes factor, their methods might not be efficient.

The problems of different dimensions were also considered by Carlin and Chib (1995) in the context of Bayesian model choice. Instead of computing marginal likelihoods, they developed a Markov chain Monte Carlo algorithm that does not suffer from convergence difficulties; and then the outputs from their algorithm can be directly used to estimate posterior model probabilities.

Note that if the conditional density of ψ given θ is completely known, the problem of different dimensions disappears. We present further explanation as follows. First we denote $\pi_2(\psi|\theta)$ to be the conditional density of ψ given θ, that is,

$$\pi_2(\psi|\theta) = \frac{p_2(\theta,\psi)}{\int_{\Psi(\theta)} p_2(\theta,\psi')d\psi'},\ \psi \in \Psi(\theta)\ \text{ for } \theta \in \Omega_2. \tag{1.7}$$

Then

$$\pi_2(\theta,\psi) = \frac{p_2(\theta,\psi)}{c_2} = \frac{p_2(\theta)}{c_2} \cdot \pi_2(\psi|\theta),$$

where $p_2(\theta)$ is a (completely known) unnormalized marginal density of θ. Thus, one can directly apply the same-dimension identities to the problem that only

involves $p_1(\theta)$ and $p_2(\theta)$. Therefore, we assume that $\pi_2(\psi|\theta)$ is known only up to a normalizing constant

$$c(\theta) = \int_{\Psi(\theta)} p_2(\theta, \psi) d\psi.$$

This assumption will be made throughout this paper. Since $c(\theta)$ depends on θ, the different-dimension problem is a challenging and difficult one.

The outline of this article is as follows. In Section 2 we present the generalized versions of importance sampling, bridge sampling and ratio importance sampling for estimating r given in (1.5). In Section 3 we derive global optimal importance sampling, bridge sampling and ratio importance sampling in the sense of minimizing asymptotic relative mean-square errors of the estimators. In Section 4 we develop detailed implementation procedures. Two illustrative examples are provided in Section 5 and in the final section we give a brief conclusion.

2. Monte Carlo Estimators

In this section, we present generalized versions of importance sampling (IS), bridge sampling (BS), and ratio importance sampling (RIS) for estimating r given in (1.5) when two unnormalized densities have different dimensions.

As discussed in Section 1, we cannot directly use IS, BS, and RIS for estimating r since $\pi(\theta)$ and $\pi(\theta, \psi)$ are defined on two different dimensional parameter spaces. However, this unequal dimensions problem can be resolved by augmenting the lower dimensional density into one that has the same dimension as the higher one by introducing a weight function. To illustrate the idea, let $p_1^*(\theta, \psi) = p_1(\theta) w(\psi|\theta)$ and

$$\pi_1^*(\theta, \psi) = \frac{p_1^*(\theta, \psi)}{c_1^*}, \tag{2.1}$$

where $w(\psi|\theta)$ is a completely known weight density function so that $\int_{\Psi(\theta)} w(\psi|\theta)$ $d\psi = 1$ and c_1^* is the normalizing constant of $\pi_1^*(\theta, \psi)$. Then, it is easy to show that $c_1^* = c_1$. Thus, we can view $r = c_1/c_2$ as the ratio of the two normalizing constants of $\pi_1^*(\theta, \psi)$ and $\pi_2(\theta, \psi)$ and henceforth, we can directly apply the IS, BS, and RIS identities (Meng and Wong (1996) and Chen and Shao (1997)) on the (θ, ψ) space for estimating r. We summarize the IS, BS and RIS estimators of r as follows.

Importance Sampling

Assume $\Omega_1 \subset \Omega_2$. Let (θ_{21}, ψ_{21}), ..., (θ_{2n}, ψ_{2n}) be a random draw from π_2. Then, on the (θ, ψ) space, using the IS identity

$$r = \frac{c_1}{c_2} = E_{\pi_2}\Big\{\frac{p_1(\theta) w(\psi|\theta)}{p_2(\theta, \psi)}\Big\}, \tag{2.2}$$

the ratio r can be estimated by

$$\hat{r}_{IS}(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{p_1(\theta_{2i})w(\psi_{2i}|\theta_{2i})}{p_2(\theta_{2i},\psi_{2i})}. \tag{2.3}$$

Bridge Sampling

Using the BS identity on the (θ,ψ) space (Meng and Wong (1996)), we have

$$r = \frac{c_1}{c_2} = \frac{E_{\pi_2}\{\, p_1(\theta)w(\psi|\theta)\alpha(\theta,\psi)\}}{E_{\pi_1^*}\{\, p_2(\theta,\psi)\alpha(\theta,\psi)\}}, \tag{2.4}$$

where $\pi_1^*(\theta,\psi)$ is defined by (2.1) with the support of $\Theta_1 = \{(\theta,\psi) : \psi \in \Psi_1(\theta),\ \theta \in \Omega_1\}$ and $\alpha(\theta,\psi)$ is an arbitrary function defined on $\Theta_1 \cap \Theta_2$ such that

$$0 < \Big| \int_{\Theta_1\cap\Theta_2} \alpha(\theta,\psi)p_1(\theta)w(\psi|\theta)p_2(\theta,\psi)d\theta d\psi \Big| < \infty.$$

Then using two random draws $(\theta_{i1},\psi_{i1}), \ldots, (\theta_{in_i},\psi_{in_i})$, $i = 1, 2$, from π_1^* and π_2 respectively, we obtain consistent estimator of r as follows

$$\hat{r}_{BS}(w,\alpha) = \frac{n_2^{-1}\sum_{i=1}^{n_2} p_1(\theta_{2i})w(\psi_{2i}|\theta_{2i})\alpha(\theta_{2i},\psi_{2i})}{n_1^{-1}\sum_{i=1}^{n_1} p_2(\theta_{1i},\psi_{1i})\alpha(\theta_{1i},\psi_{1i})}. \tag{2.5}$$

Ratio Importance Sampling

Using the RIS identity on the (θ,ψ) space (Chen and Shao (1997)), we have

$$r = \frac{c_1}{c_2} = \frac{E_\pi\Big\{p_1(\theta)w(\psi|\theta)/\pi(\theta,\psi)\Big\}}{E_\pi\Big\{p_2(\theta,\psi)/\pi(\theta,\psi)\Big\}}, \tag{2.6}$$

where π is an arbitrary density over Θ such that $\pi(\theta,\psi) > 0$ for $(\theta,\psi) \in \Theta = \Theta_1 \cup \Theta_2$. We remark that in (2.6), it is not necessary for π to be completely known, i.e., π can be known up to an unknown normalizing constant: $\pi(\theta,\psi) = p(\theta,\psi)/c$. Given a random draw $(\theta_1,\psi_1), \ldots, (\theta_n,\psi_n)$ from π, the ratio importance sampling estimator of r is

$$\hat{r}_{RIS}(w,\pi) = \frac{\sum_{i=1}^{n} p_1(\theta_i)w(\psi_i|\theta_i)/\pi(\theta_i,\psi_i)}{\sum_{i=1}^{n} p_2(\theta_i,\psi_i)/\pi(\theta_i,\psi_i)}. \tag{2.7}$$

Even without knowing the normalizing constants, c_i, $i = 1, 2$, or c, the distributions $\pi_2(\theta,\psi)$, $\pi_1(\theta)$, or $\pi(\theta,\psi)$ for IS, BS or RIS can be sampled, for example, by means of the Markov chain Monte Carlo (MCMC) methods such as

the Metropolis-Hastings algorithm (Metropolis et al. (1953); Hastings (1970)), the Gibbs sampler (Geman and Geman (1984); Gelfand and Smith (1990); Tanner and Wong (1987)), and various hybrid algorithms (Chen and Schmeiser (1993); Müller (1991); Tierney (1994)).

In the next section, we will discuss what the optimal choices of w, α and π are so that $\hat{r}_{IS}(w)$, $\hat{r}_{BS}(w,\alpha)$ and $\hat{r}_{RIS}(w,\pi)$ have the smallest asymptotic relative mean-square errors.

3. Global Optimal Monte Carlo Methods

In this section, we explore the properties of three estimators, namely, $\hat{r}_{IS}(w)$, $\hat{r}_{BS}(w,\alpha)$ and $\hat{r}_{RIS}(w,\pi)$.

We use the following notation. Let $\pi_{21}(\theta)$ be the marginal density of θ defined on Ω_2. Then

$$\pi_{21}(\theta) = \int_{\Psi(\theta)} \frac{p_2(\theta,\psi)}{c_2} d\psi \quad \text{for} \quad \theta \in \Psi(\theta), \tag{3.1}$$

where Ω_2 and $\Psi(\theta)$ are defined in (1.4). Denote $\hat{r}$ as an estimator of r. Then the relative mean-square error (RE) is defined as

$$RE^2(\hat{r}) = \frac{E(\hat{r}-r)^2}{r^2} \tag{3.2}$$

and the asymptotic relative mean-square error (ARE) is defined as

$$ARE^2(\hat{r}) = \lim_{n\to\infty} nRE^2(\hat{r}). \tag{3.3}$$

On the (θ,ψ) space, for a given weight density function $w(\psi|\theta)$, the RE's and ARE's of $\hat{r}_{IS}(w)$, $\hat{r}_{BS}(w,\alpha)$ and $\hat{r}_{RIS}(w,\pi)$ can be directly obtained from Meng and Wong (1996) and Chen and Shao (1997). The results are presented in the following three lemmas.

Lemma 3.1. *Assume* $\Omega_1 \subset \Omega_2$ *and* $\int_{\Theta_2}\{p_1^2(\theta)w^2(\psi|\theta)/p_2(\theta,\psi)\}d\theta d\psi < \infty$. *Then we have*

$$RE^2\Big(\hat{r}_{IS}(w)\Big) = \frac{1}{r^2}\mathrm{Var}(\hat{r}_{IS}(w)) = \frac{1}{n}\Big[\int_{\Theta_2}\frac{\pi_1^2(\theta)w^2(\psi|\theta)}{\pi_2(\theta,\psi)}d\theta d\psi - 1\Big] \tag{3.4}$$

and

$$ARE^2(\hat{r}_{IS}(w)) = \int_{\Theta_2}\frac{\pi_1^2(\theta)w^2(\psi|\theta)}{\pi_2(\theta,\psi)}d\theta d\psi - 1. \tag{3.5}$$

Lemma 3.2. *Let* $n = n_1 + n_2$ *and* $s_{i,n} = n_i/n$ *for* $i = 1,2$. *Assume that* $s_i = \lim_{n\to\infty} s_{i,n} > 0$ $(i = 1,2)$, $E_{\pi_2}\{\,p_1(\theta)w(\psi|\theta)\alpha(\theta,\psi)\}^2 < \infty$ *and*

$$E_{\pi_1^*}\Big\{\,(\,p_2(\theta,\psi)\alpha(\theta,\psi))^2 + 1/(\,p_2(\theta,\psi)\alpha(\theta,\psi))^2\Big\} < \infty.$$

Then we have

$$RE^2\Big(\hat{r}_{BS}(w,\alpha)\Big) = \frac{1}{ns_{1,n}s_{2,n}}$$
$$\cdot\Big\{\frac{\int_{\Theta_1\cap\Theta_2}\pi_1(\theta)w(\psi|\theta)\pi_2(\theta,\psi)(s_{1,n}\pi_1(\theta)w(\psi|\theta)+s_{2,n}\pi_2(\theta,\psi))\alpha^2(\theta,\psi)d\theta d\psi}{\Big(\int_{\Theta_1\cap\Theta_2}\pi_1(\theta)w(\psi|\theta)\pi_2(\theta,\psi)\alpha(\theta,\psi)d\theta d\psi\Big)^2}-1\Big\}$$
$$+\,o(\frac{1}{n})$$

and

$$ARE^2\Big(\hat{r}_{BS}(w,\alpha)\Big)=\frac{1}{s_1s_2}$$
$$\cdot\Big\{\frac{\int_{\Theta_1\cap\Theta_2}\pi_1(\theta)w(\psi|\theta)\pi_2(\theta,\psi)(s_1\pi_1(\theta)w(\psi|\theta)+s_2\pi_2(\theta,\psi))\alpha^2(\theta,\psi)d\theta d\psi}{\Big(\int_{\Theta_1\cap\Theta_2}\pi_1(\theta)w(\psi|\theta)\pi_2(\theta,\psi)\alpha(\theta,\psi)d\theta d\psi\Big)^2}-1\Big\}. \quad (3.6)$$

Lemma 3.3. *Assume that* $E_\pi\{(\pi_1(\theta)w(\psi|\theta)-\pi_2(\theta,\psi))/\pi(\theta,\psi)\}^2<\infty$ *and*

$$E_\pi\Big\{p_1(\theta)w(\psi|\theta)/p_2(\theta,\psi)\Big\}^2<\infty.$$

Then we have

$$RE^2\Big(\hat{r}_{RIS}(w,\pi)\Big)=\frac{1}{n}E_\pi\Big\{\frac{(\pi_1(\theta)w(\psi|\theta)-\pi_2(\theta,\psi))^2}{\pi^2(\theta,\psi)}\Big\}+o(\frac{1}{n})$$

and

$$ARE^2\Big(\hat{r}_{RIS}(w,\pi)\Big)=\int_{\Theta_1\cup\Theta_2}\frac{(\pi_1(\theta)w(\psi|\theta)-\pi_2(\theta,\psi))^2}{\pi(\theta,\psi)}d\theta d\psi. \quad (3.7)$$

Now we present a general result that will be essentially used for deriving optimal choices of $w(\psi|\theta)$, $\alpha(\theta,\psi)$ and $\pi(\theta,\psi)$ for IS, BS and RIS.

Theorem 3.1. *Assume there exist functions* h *and* g *such that*
(I) $ARE^2(\hat{r})\ge h\{E_{\pi_2}[g(\pi_1(\theta)w(\psi|\theta)/\pi_2(\theta,\psi))]\}$,
(II) *either* (i) *or* (ii) *holds:*
(i) h *is an increasing function and* g *is convex;*
(ii) h *is a decreasing function and* g *is concave.*
Then for an arbitrary $w(\psi|\theta)$ *defined on* $\Psi(\theta)$ *or* $\Psi_1(\theta)$,

$$ARE^2(\hat{r})\ge h\Big\{E_{\pi_{21}}\Big[g(\pi_1(\theta)/\pi_{21}(\theta))\Big]\Big\}. \quad (3.8)$$

That is, the lower bound of $ARE^2(\hat{r})$ *is* $h\{E_{\pi_{21}}[g(\pi_1(\theta)/\pi_{21}(\theta))]\}$. *Furthermore, if the equality holds in* (I), *the lower bound of* $ARE^2(\hat{r})$ *is achieved when* $w(\psi|\theta)=\pi_2(\psi|\theta)$.

The proof of (3.8) simply follows assumptions (I) and (II) and Jensen's inequality.

Using the above theorem, we can easily obtain the optimal choices of $w(\psi|\theta)$, $\alpha(\theta,\psi)$ and $\pi(\theta,\psi)$ for IS, BS and RIS in the sense of minimizing their ARE's. These optimal choices are denoted by w_{opt}^{IS} for IS, w_{opt}^{BS} and α_{opt} for BS and w_{opt}^{RIS} and π_{opt} for RIS. IS with $w(\psi|\theta) = w_{opt}^{IS}(\psi|\theta)$, BS with $w = w_{opt}^{BS}$ and $\alpha = \alpha_{opt}$, and RIS with $w = w_{opt}^{RIS}$ and $\pi = \pi_{opt}$ are called optimal importance sampling (OIS), global optimal bridge sampling (GOBS), and global optimal ratio importance sampling (GORIS), respectively. We further denote

$$\hat{r}_{OIS} = \hat{r}_{IS}(w_{opt}^{IS}), \quad \hat{r}_{GOBS} = \hat{r}_{BS}(w_{opt}^{BS}, \alpha_{opt}) \quad \text{and} \quad \hat{r}_{GORIS} = \hat{r}_{RIS}(w_{opt}^{RIS}, \pi_{opt}).$$

Then we have the following results.

Theorem 3.2. *The optimal choices are*

$$w_{opt}^{IS} = w_{opt}^{BS} = w_{opt}^{RIS} = \pi_2(\psi|\theta), \quad \psi \in \Psi(\theta) \quad for \quad \theta \in \Omega_1 \cap \Omega_2$$

and w_{opt}^{BS} and w_{opt}^{RIS} are arbitrary densities for $\theta \in \Omega_1 - \Omega_2$,

$$\alpha_{opt}(\theta,\psi) = \frac{c}{s_1\pi_1(\theta)w_{opt}^{BS}(\psi|\theta) + s_2\pi_2(\theta,\psi)}, \quad (\theta,\psi) \in \Theta_1 \cap \Theta_2, \quad \forall \quad c \neq 0,$$

and

$$\pi_{opt}(\theta,\psi) = \frac{|\pi_1(\theta)w_{opt}^{RIS}(\psi|\theta) - \pi_2(\theta,\psi)|}{\int_{\Theta_1\cup\Theta_2} |\pi_1(\theta')w_{opt}^{RIS}(\psi'|\theta') - \pi_2(\theta',\psi')|d\theta' d\psi'}.$$

The optimal ARE's are

$$ARE^2\big(\hat{r}_{OIS}\big) = \int_{\Omega_1} \frac{\pi_1^2(\theta)}{\pi_{21}(\theta)}d\theta - 1, \tag{3.9}$$

$$ARE^2\big(\hat{r}_{GOBS}\big) = \frac{1}{s_1 s_2}\Big\{\Big(\int_{\Omega_1\cap\Omega_2} \frac{\pi_1(\theta)\pi_{21}(\theta)}{s_1\pi_1(\theta) + s_2\pi_{21}(\theta)}d\theta\Big)^{-1} - 1\Big\}, \tag{3.10}$$

and

$$ARE^2\big(\hat{r}_{GORIS}\big) = \Big[\int_{\Omega_1\cup\Omega_2} |\pi_1(\theta) - \pi_{21}(\theta)|d\theta\Big]^2. \tag{3.11}$$

Proof. We prove the theorem in turn for IS, BS and RIS.

For IS, from Lemma 3.1, we take $h(y) = y-1$, which is an increasing function of y, and $g(x) = x^2$, which is convex. Therefore, Theorem 3.1 implies that the lower bound of $ARE^2(\hat{r}_{IS}(w))$ is $\int_{\Omega_1} \frac{\pi_1^2(\theta)}{\pi_{21}(\theta)}d\theta - 1$. Since the equality holds in (I) of Theorem 3.1, this lower bound is attained at $w = \pi_2(\psi|\theta)$. Thus we have proved the optimal results for IS.

For BS, analogous to the proof given by Meng and Wong (1996), by Lemma 3.2 and the Cauchy-Schwarz inequality, for all $\alpha(\theta,\psi)$

$$ARE^2\Big(\hat{r}_{BS}(w,\alpha)\Big) \geq \frac{1}{s_1 s_2}\Big\{\Big(\int_{\Theta_1\cap\Theta_2} \frac{\pi_1(\theta)w(\psi|\theta)\pi_2(\theta,\psi)}{s_1\pi_1(\theta)w(\psi|\theta)+s_2\pi_2(\theta,\psi)}d\theta d\psi\Big)^{-1}-1\Big\}.$$

We take $h(y)=\frac{1}{s_1 s_2}(\frac{1}{y}-1)$ and $g(x)=\frac{x}{s_1x+s_2}$. Then $h(y)$ is a decreasing function of y and $g''(x)=\frac{-2s_1s_2}{(s_1x+s_2)^3}<0$ which implies that g is concave. Therefore, Theorem 3.1 yields the lower bound of $ARE^2(\hat{r}_{BS}(w,\alpha))$ as

$$\frac{1}{s_1 s_2}\Big\{\Big(\int_{\Omega_1\cap\Omega_2} \frac{\pi_1(\theta)\pi_{21}(\theta)}{s_1\pi_1(\theta)+s_2\pi_{21}(\theta)}d\theta\Big)^{-1}-1\Big\}. \tag{3.12}$$

Although the equality does not hold in (I) of Theorem 3.1, it can be easily verified that the lower bound (3.12) is reached at $w=w_{opt}^{BS}$ and $\alpha=\alpha_{opt}$. Thus, we have proved Theorem 3.2 for BS.

Finally, for RIS, by Lemma 3.3 and the Cauchy-Schwarz inequality, for an arbitrary density π,

$$ARE^2\Big(\hat{r}_{RIS}(w,\pi)\Big) \geq \Big[\int_{\Theta_1\cup\Theta_2} |\pi_1(\theta)w(\psi|\theta)-\pi_2(\theta,\psi)|d\theta d\psi\Big]^2. \tag{3.13}$$

Now we take $h(y)=y^2$ and $g(x)=|x-1|$. Obviously, $h(y)$ is an increasing function of y for $y>0$ and $g(x)$ is convex. Therefore, from Theorem 3.1 the lower bound of $ARE^2(\hat{r}_{RIS}(w,\pi))$ is

$$\Big[\int_{\Omega_1\cup\Omega_2} |\pi_1(\theta)-\pi_{21}(\theta)|d\theta\Big]^2.$$

Note that since the integral region of the right side of Inequality (3.13) is bigger than the support of π_2, Theorem 3.1 needs an obvious adjustment. By algebra, plugging $w=w_{opt}^{RIS}$ and $\pi=\pi_{opt}$ into (3.7) leads to (3.11). Thus we have completed the proof of Theorem 3.2.

It is interesting to mention that the optimal choices of w are the same for all three Monte Carlo methods (IS, BS and RIS). The optimal w is the conditional density $\pi_2(\psi|\theta)$. These results are consistent with our intuitive guess. We conclude this section with the following brief remarks.

Remark 3.1. It is known that IS is a special case of BS with $\alpha(\theta,\psi)=1/\pi_2(\theta,\psi)$. Because this α is not α_{opt}, the proof for the optimal choice of w for IS cannot simply follow that for BS.

Remark 3.2. Following the proof of Theorem 3.3 of Chen and Shao (1997), we have

$$ARE^2\Big(\hat{r}_{RIS}(w_{opt}^{RIS},\pi_{opt})\Big) \leq ARE^2\Big(\hat{r}_{BS}(w_{opt}^{BS},\alpha_{opt})\Big).$$

From Section 3 of Chen and Shao, we also have

$$ARE^2\Big(\hat{r}_{RIS}(w_{opt}^{RIS}, \pi_{opt})\Big) \leq ARE^2\Big(\hat{r}_{IS}(w_{opt}^{IS})\Big).$$

Remark 3.3. Under certain conditions (c.f., Theorem 3.1 of Chen and Shao (1997)), the central limit theorem holds for all $\hat{r}_{IS}(w)$, $\hat{r}_{BS}(w,\alpha)$ and $\hat{r}_{RIS}(w,\pi)$. We state the following results without proof:

$$\sqrt{n}\ (\hat{r}_{IS}(w) - r)/r \ \xrightarrow{\mathcal{D}}\ N\Big(0, ARE^2(\hat{r}_{IS}(w))\Big), \quad \text{as} \quad n \to \infty,$$

$$\sqrt{n}\ (\hat{r}_{BS}(w,\alpha) - r)/r \ \xrightarrow{\mathcal{D}}\ N\Big(0, ARE^2(\hat{r}_{BS}(w,\alpha))\Big), \quad \text{as} \quad n \to \infty,$$

and

$$\sqrt{n}\ (\hat{r}_{RIS}(w,\pi) - r)/r \ \xrightarrow{\mathcal{D}}\ N\Big(0, ARE^2(\hat{r}_{RIS}(w,\pi))\Big), \quad \text{as} \quad n \to \infty,$$

where $ARE^2(\hat{r}_{IS}(w))$, $ARE^2(\hat{r}_{BS}(w,\alpha))$ and $ARE^2(\hat{r}_{RIS}(w,\pi))$ are given in (3.5), (3.6) and (3.7) respectively.

Remark 3.4. With the global optimal choices of w, α and π, the (asymptotic) relative mean-square errors (ARE's) for all three methods depend only on $\pi_1(\theta)$ and $\pi_{21}(\theta)$, which implies that the extra parameter ψ does not add any extra simulation variation, i.e., we do not lose any simulation efficiency although the second unnormalized density π_2 has d extra dimensions. However, such a conclusion is valid only if the optimal solutions can be implemented in practice since $w(\psi|\theta)$ is not completely known. We will discuss implementation issues in Section 4.

Remark 3.5. Assuming that $\Psi(\theta) = \Psi \subset R^m$ for all $\theta \in \Omega_2$ and $\Omega_1 \subset \Omega_2$, we have the identity

$$r = E_{\pi_2}\Big\{p_2(\theta^*, \psi)p_1(\theta)/p_2(\theta, \psi)\Big\}/c(\theta^*),$$

where $c(\theta^*) = \int_\Psi p_2(\theta^*, \psi)d\psi$ and $\theta^* \in \Omega_2$ is a fixed point. Thus, a marginal-likelihood estimator of r can be defined by

$$\hat{r}_{ML} = \Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{p_2(\theta^*, \psi_i)p_1(\theta_i)}{p_2(\theta_i, \psi_i)}\Big\} \cdot \Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{w^*(\varphi_i|\theta^*)}{p_2(\theta^*, \varphi_i)}\Big\},$$

where (θ_i, ψ_i), $i = 1, \ldots, n$ and φ_i , $i = 1, \ldots, n$ are two independent random draws from $p_2(\theta, \psi)$ and $\pi_2(\varphi|\theta^*)$, respectively, and $w^*(\varphi|\theta^*)$ is an arbitrary (completely known) density defined on Ψ (see Chib (1995) or Chen and Shao (1997)

for a full description of the marginal likelihood method). Then, we have

$$\mathrm{Var}\Big(\hat{r}_{ML}\Big) = r^2\Big[\frac{1}{n}\Big\{\int_{\Omega_1}\frac{\pi_1^2(\theta)}{\pi_{21}(\theta)}\Big(\int_{\Psi}\frac{\pi_2^2(\psi|\theta^*)}{\pi_2(\psi|\theta)}d\psi\Big)d\theta - 1\Big\} + 1\Big]$$
$$\times\Big[\frac{1}{n}\Big\{\int_{\Psi}\frac{w^{*2}(\varphi|\theta^*)}{\pi_2(\varphi|\theta^*)}d\varphi - 1\Big\} + 1\Big] - r^2.$$

By the Cauchy-Schwarz inequality,

$$\int_{\Psi}\frac{\pi_2^2(\psi|\theta^*)}{\pi_2(\psi|\theta)}d\psi \geq 1 \quad \text{and} \quad \int_{\Psi}\frac{w^{*2}(\varphi|\theta^*)}{\pi_2(\varphi|\theta^*)}d\varphi \geq 1.$$

Thus, for all $w^*(\varphi|\theta^*)$

$$\mathrm{Var}\Big(\hat{r}_{ML}\Big) \geq \mathrm{Var}(\hat{r}_{OIS}), \tag{3.14}$$

where $\mathrm{Var}(\hat{r}_{OIS}) = \frac{r^2}{n}ARE^2(\hat{r}_{OIS})$. Hence, $\hat{r}_{ML}$ is not as good as $\hat{r}_{OIS}$. Note that the optimal choice of w^* is $w^*_{opt}(\varphi|\theta^*) = \pi_2(\varphi|\theta^*)$. Even with this optimal weight density w^*_{opt}, equality in (3.14) still does not hold in general unless θ and ψ are independent.

4. Implementation Issues

In many practical problems, the closed form of the conditional density $\pi_2(\psi|\theta)$ is not available especially when $\Psi(\theta)$ is a constrained parameter space (Chen 1994). (Also see Gelfand, Smith and Lee (1992) for the Bayesian analysis of constrained parameter problems.) Therefore, evaluating ratios of normalizing constants for densities with different dimensions is an important problem. In this section we present detailed implementation schemes for obtaining $\hat{r}_{OIS}$, $\hat{r}_{GOBS}$ and $\hat{r}_{GORIS}$. We consider our implementation procedures for $d = 1$ and $d > 1$ separately.

First, we consider $d = 1$. In this case, $\pi_2(\psi|\theta) = \frac{p(\theta,\psi)}{c(\theta)}$ where $c(\theta) = \int_{\Psi(\theta)} p(\theta,\psi')d\psi'$. Note that the integral in $c(\theta)$ is only one-dimensional. Since one-dimensional numerical integration methods are well-developed and computationally fast, one can use, for example, IMSL subroutine QDAG or QDAGI; or as Verdinelli and Wasserman (1995) suggested, one can use a grid $\{\psi^*_1, \ldots, \psi^*_M\}$ that includes all sample points $\psi_1, \ldots, \psi_n$ and then use the trapezoidal rule to approximate the integral. In the following three algorithms, we assume that $c(\theta)$ will be calculated or approximated by a numerical integration method. Detailed implementation schemes for obtaining $\hat{r}_{OIS}$, $\hat{r}_{GOBS}$ and $\hat{r}_{GORIS}$ are presented as follows.

For IS, $\hat{r}_{OIS}$ is available through the following two step algorithm.

Algorithm: OIS
Step 1 Draw a random sample (θ_1, ψ_1), ..., (θ_n, ψ_n) from $\pi_2(\theta, \psi)$.
Step 2 Calculate $c(\theta_i)$ and compute

$$\hat{r}_{OIS} = \frac{1}{n} \sum_{i=1}^{n} \frac{p_1(\theta_i)}{c(\theta_i)}. \tag{4.1}$$

Note that if one uses a one-dimensional numerical integration subroutine, in Step 1 one needs to draw the θ_i from the marginal distribution of θ. However, drawing θ_i and ψ_i together is often easier than drawing θ_i alone from its marginal distribution. (In such case ψ can be considered as an auxiliary variable or a latent variable. As Besag and Green (1993) and Polson (1996) pointed out, use of latent variables in Monte Carlo sampling will greatly ease implementation difficulty and dramatically accelerate convergence.) Furthermore, if one uses the aforementioned grid numerical integration method to approximate $c(\theta)$, the ψ_i can be used as part of grid points.

For GOBS, similar to Algorithm OIS, we have the following algorithm.

Algorithm: GOBS
Step 1 Draw random samples (θ_{i1}, ψ_{i1}), ..., $(\theta_{in_i}, \psi_{in_i})$, $i = 1, 2$, $(n_1 + n_2 = n)$ as follows:

(i) Draw $\{\theta_{11}, \ldots, \theta_{1n_1}\}$ from $\pi_1(\theta)$ and then draw $\{\theta_{21}, \ldots, \theta_{2n_2}\}$ from the marginal distribution of θ with respect to $\pi_2(\theta, \psi)$.

(ii) Draw ψ_{ij} independently from $\pi_2(\psi|\theta_{ij})$ for $j = 1, \ldots, n_i$ and $i = 1, 2$.

Step 2 Calculate $c(\theta_{ij})$ and set $\hat{r}_{GOBS}$ be the unique zero root of the "score" equation

$$S(r) = \sum_{i=1}^{n_1} \frac{s_2 r}{s_1 p_1(\theta_{1i})/c(\theta_{1i}) + s_2 r} - \sum_{i=1}^{n_2} \frac{s_1 p_1(\theta_{2i})/c(\theta_{2i})}{s_1 p_1(\theta_{2i})/c(\theta_{2i}) + s_2 r}. \tag{4.2}$$

Analogous to Theorem 2 of Meng and Wong (1996), in Step 2 $S(0) = -n_2$, $S(\infty) = n_1$, and

$$\frac{dS(r)}{dr} = \sum_{i=1}^{n_1} \frac{s_1 s_2 p_1(\theta_{1i})/c(\theta_{1i})}{\left\{s_1 p_1(\theta_{1i})/c(\theta_{1i}) + s_2 r\right\}^2} + \sum_{i=1}^{n_2} \frac{s_1 s_2 p_1(\theta_{2i})/c(\theta_{2i})}{\left\{s_1 p_1(\theta_{2i})/c(\theta_{2i}) + s_2 r\right\}^2} > 0.$$

Thus, $S(r) = 0$ has a unique root. Since $S(r)$ is a strictly increasing function, this root can be easily obtained by, for example, the bisection method. Note that in Step 1, drawing the θ_{ij} or the ψ_{ij} does not require knowing normalizing constants since we can use, for example, a rejection/acceptance, Metropolis, or Gibbs sampler method. Also note that in Step 2, $\hat{r}_{GOBS}$ can be obtained by

using an alternative iterative method proposed by Meng and Wong (1996). This method can be implemented as follows. Starting with an initial guess of r, $\hat{r}_{(0)}$, at the $(t+1)$th iteration, we compute

$$\hat{r}_{(t+1)} = \frac{1}{n_2}\sum_{i=1}^{n_2} \frac{p_1(\theta_{2i})/c(\theta_{2i})}{s_1 p_1(\theta_{2i})/c(\theta_{2i}) + s_2\hat{r}_{(t)}} \Big/ \frac{1}{n_1}\sum_{i=1}^{n_1} \frac{1/c(\theta_{1i})}{s_1 p_1(\theta_{1i})/c(\theta_{1i}) + s_2\hat{r}_{(t)}}.$$

Then, the limit of $\hat{r}_{(t)}$ is $\hat{r}_{GOBS}$.

For RIS, we obtain an approximate $\hat{r}_{GORIS}$, denoted by $\hat{r}^*_{GORIS}$, by a two-stage procedure developed by Chen and Shao (1997).

Algorithm: GORIS

Step 1 Let $\pi(\theta,\psi)$ be an arbitrary (known up to a normalizing constant) density over Θ such that $\pi(\theta,\psi) > 0$ for $(\theta,\psi) \in \Theta$. (For example, $\pi(\theta,\psi) = \pi_2(\theta,\psi)$.) Draw a random sample $(\theta_1,\psi_1),\ldots,(\theta_{n_1},\psi_{n_1})$ from π. Calculate the $c(\theta_i)$ and compute

$$\tau_{n_1} = \frac{\sum_{i=1}^{n_1} p_1(\theta_i)p_2(\theta_i,\psi_i)/[c(\theta_i)\pi(\theta_i,\psi_i)]}{\sum_{i=1}^{n_1} p_2(\theta_i,\psi_i)/\pi(\theta_i,\psi_i)}. \tag{4.3}$$

Step 2 Let

$$\pi^*_{n_1}(\theta,\psi) = \frac{|p_1(\theta)\pi_2(\psi|\theta) - \tau_{n_1}p_2(\theta,\psi)|}{\int_\Theta |p_1(\theta')\pi_2(\psi'|\theta') - \tau_{n_1}p_2(\theta',\psi')|d\theta' d\psi'}. \tag{4.4}$$

Then, make a random draw $(\vartheta_1,\varphi_1),\ldots,(\vartheta_{n_2},\varphi_{n_2})$ from $\pi^*_{n_1}$. $(n_1+n_2=n.)$

Step 3 Calculate $c(\vartheta_i)$ and compute

$$\hat{r}^*_{GORIS} = \frac{\sum_{i=1}^{n_2} p_1(\vartheta_i)/|p_1(\vartheta_i) - \tau_{n_1}c(\vartheta_i)|}{\sum_{i=1}^{n_2} c(\vartheta_i)/|p_1(\vartheta_i) - \tau_{n_1}c(\vartheta_i)|}. \tag{4.5}$$

Similar to Theorem 5.1 of Chen and Shao (1997) we can prove that $\hat{r}^*_{GORIS}$ has the same asymptotic relative mean-square error as $\hat{r}_{GORIS}$ as long as $n_1 = o(n)$ and $n_1 \to \infty$. The most expensive/difficult part of Algorithm GORIS is Step 2. There are two possible approaches to draw (ϑ_i,φ_i) from $\pi^*_{n_1}$. The first approach is the random-direction interior-point (RDIP) sampler (Chen and Schmeiser (1994)). RDIP requires only that $|p_1(\theta)\pi_2(\psi|\theta) - \tau_{n_1}p_2(\theta,\psi)|$ can be computed at any point (θ,ψ). Another approach is Metropolis sampling. In Metropolis sampling, one needs to choose a good proposal density that should be spread out enough (Tierney (1994)). For example, if $\pi_2(\theta,\psi)$ has a tail as heavy as the one of $p_1(\theta)\pi_2(\psi|\theta)$, then one can simply choose $\pi_2(\theta,\psi)$ as a proposal density. Compared to Algorithms OIS and GOBS, Algorithm GORIS requires evaluating $c(\theta)$ in the sampling step; hence, Algorithm GORIS is more expensive.

Second, we consider $d > 1$. In this case, the integral in $c(\theta)$ is multidimensional. Therefore, simple numerical integration methods might not be feasible.

Instead of directly computing $c(\theta)$ in the case of $d = 1$, we develop Monte Carlo schemes to estimate $\pi_2(\psi|\theta)$. However, the basic structures of the implementation algorithms are similar to those for $d = 1$. Thus, in the following presentation, we mainly focus on how to estimate or approximate $\pi_2(\psi|\theta)$. We propose "exact" and "approximate" approaches to do so.

We start with an "exact" approach. Using the notation of Schervish and Carlin (1992), we let $\psi' = (\psi'_{(1)}, \ldots, \psi'_{(d)})$, $\psi^{(j')} = (\psi_{(1)}, \ldots, \psi_{(j)}, \psi'_{(j+1)}, \ldots, \psi'_{(d)})$ and $\psi^{(d')} = \psi$. We denote a "one-step Gibbs transition" density as

$$\pi_2^{(j)}(\psi|\theta) = \pi_2(\psi_{(j)}|\psi_{(1)}, \ldots, \psi_{(j-1)}, \psi_{(j+1)}, \ldots, \psi_{(d)}, \theta)$$

and a "transition kernel" as

$$k(\psi', \psi|\theta) = \prod_{j=1}^{d} \pi_2^{(j)}(\psi^{(j')}|\theta).$$

Then we have the following key identity

$$\pi_2(\psi|\theta) = \int_{\Psi(\theta)} k(\psi', \psi|\theta)\pi_2(\psi'|\theta)d\psi'.$$

Now we can obtain a Monte Carlo estimator of $\pi_2(\psi|\theta)$ by

$$\hat{\pi}_2(\psi|\theta) = \frac{1}{m}\sum_{l=1}^{m} k(\psi^l, \psi|\theta), \tag{4.6}$$

where ψ^l, $l = 1, \ldots, m$, is a random draw from $\pi_2(\psi|\theta)$. The above method was originally introduced by Ritter and Tanner (1992) for the Gibbs stopper. Here, we use this method for estimating conditional densities. Although the joint conditional density is not analytically available, one-dimensional conditional densities can be computed by the aforementioned simple numerical integration method and sometimes some of one-dimensional conditional densities are even analytically available or easy to compute (see an illustrated example in Section 5). Therefore, (4.6) is advantageous. In (4.6), sampling from $\pi_2(\psi|\theta)$ does not require knowing the normalizing constant $c(\theta)$ and convergence of $\hat{\pi}_2(\psi|\theta)$ to $\pi_2(\psi|\theta)$ is expected to be rapid. Algorithms OIS, GOBS and GORIS for $d > 1$ are similar to the ones for $d = 1$. We only need the following minor adjustment. Suppose we generate ψ^l, $l = 1, \ldots, m$, from $\pi_2(\psi|\theta_i)$, $\pi_2(\psi|\theta_{ij})$ or $\pi_2(\psi|\vartheta_i)$ and we compute $\hat{\pi}_2(\psi_i|\theta_i)$, $\hat{\pi}_2(\psi_{ij}|\theta_{ij})$, or $\hat{\pi}_2(\varphi_i|\vartheta_i)$ by using (4.6). Then, for OIS and GOBS, instead of (4.1) and (4.2), we use

$$\hat{r}_{OIS} = \frac{1}{n}\sum_{i=1}^{n} \frac{p_1(\theta_i)\hat{\pi}_2(\psi_i|\theta_i)}{p_2(\theta_i, \psi_i)} \tag{4.7}$$

and

$$S(r) = \sum_{i=1}^{n_1} \frac{s_2 r p_2(\theta_{1i}, \psi_{1i})}{s_1 p_1(\theta_{1i})\hat{\pi}_2(\psi_{1i}|\theta_{1i}) + s_2 r p_2(\theta_{1i}, \psi_{1i})} - \sum_{i=1}^{n_2} \frac{s_1 p_1(\theta_{2i})\hat{\pi}_2(\psi_{2i}|\theta_{2i})}{s_1 p_1(\theta_{2i})\hat{\pi}_2(\psi_{2i}|\theta_{2i}) + s_2 r p_2(\theta_{2i}, \psi_{2i})}. \tag{4.8}$$

For GORIS, instead of (4.3) and (4.5), we use

$$\tau_{n_1} = \frac{\sum_{i=1}^{n_1} p_1(\theta_i)\hat{\pi}_2(\psi_i|\theta_i)/\pi(\theta_i, \psi_i)}{\sum_{i=1}^{n_1} p_2(\theta_i, \psi_i)/\pi(\theta_i, \psi_i)} \tag{4.9}$$

and

$$\hat{r}^*_{GORIS} = \frac{\sum_{i=1}^{n_2} p_1(\vartheta_i)\hat{\pi}_2(\varphi_i|\vartheta_i)/|p_1(\vartheta_i)\hat{\pi}_2(\varphi_i|\vartheta_i) - \tau_{n_1} p_2(\vartheta_i, \varphi_i)|}{\sum_{i=1}^{n_2} p_2(\vartheta_i, \varphi_i)/|p_1(\vartheta_i)\hat{\pi}_2(\varphi_i|\vartheta_i) - \tau_{n_1} p_2(\vartheta_i, \varphi_i)|}. \tag{4.10}$$

Although the above method involves extensive computation, it is quite simple especially for OIS and GOBS. More importantly, it achieves the optimal (relative) mean-square errors asymptotically, i.e., as $m \to \infty$.

Lastly, we briefly introduce an "approximate" approach that requires less computation effort. Mainly, one needs to find a completely known density $w^*(\psi|\theta)$ that has a shape similar to $\pi_2(\psi|\theta)$. Chen (1994) presented detailed guidelines for choosing a good $w^*(\psi|\theta)$. His guidelines are essentially similar to the ones for choosing a good importance sampling density (e.g., see Geweke (1989)) and they can be directly applied to this problem. We use few lines to summarize these guidelines. When the parameter space Θ_2 is unconstrained, we choose a joint importance sampling density, for example, a normal or t density, that has a shape similar to $\pi_2(\theta, \psi)$ by using the method of Laplace approximation or moments estimates. (Note that the posterior moments of $\pi_2(\theta, \psi)$ are quite easy to obtain through, e.g., a Markov chain Monte Carlo method.) Then, $w^*(\psi|\theta)$ is chosen to be the conditional density of the joint importance sampling density. When Θ_2 is a constrained parameter space, we use

$$w^*(\psi|\theta) = w^*(\psi_{(1)}|\theta)w^*(\psi_{(2)}|\psi_{(1)}, \theta) \cdots w^*(\psi_{(d)}|\psi_{(1)}, \ldots, \psi_{(d-1)}, \theta).$$

Then each of the above one-dimensional conditional densities is chosen by method of moments estimates. For example, if the support of the conditional density of $\psi_{(1)}$ given θ is a finite interval, then one can use a beta density as $w^*(\psi_{(1)}|\theta)$ whose mean, variance as well as two endpoints of the interval are determined by posterior moments of $\psi_{(1)}$ and θ (see Chen (1994) for the detailed illustration). When a good $w^*(\psi|\theta)$ is chosen, we simply replace $\hat{\pi}_2$ by $w^*(\psi|\theta)$ in (4.7), (4.8),

(4.9), and (4.10) and then Algorithms OIS, GOBS and GORIS give approximate $\hat{r}_{OIS}$, $\hat{r}_{GOBS}$ and $\hat{r}_{GORIS}$.

In the next section, we present illustrative examples to show how our algorithms can be implemented in practice.

5. Examples

5.1. Testing for departures from normality

As an illustration of our implementation algorithms developed in Section 4 for $d = 1$, we consider an example given in Section 3.2 of Verdinelli and Wasserman (1995). Suppose that we have observations $x_1, \ldots, x_N$ and we would like to test whether the sampling distribution is normal or heavier tailed. We use the t distribution with ν degrees of freedom for the data. Using notation similar to that of Verdinelli and Wasserman (1995), we define $\psi = 1/\nu$ so that $\psi = 0$ corresponds to the null hypothesis of normality and larger values of ψ correspond to heavier-tailed distributions, with $\psi = 1$ corresponding to a Cauchy distribution ($0 \le \psi \le 1$). Let $\theta = (\mu, \sigma)$ where μ and σ are location and scale parameters and denote $\bar{x}$ and s^2 to be the sample mean and the sample variance of $x_1, \ldots, x_N$. Then, using exactly the same choices of priors as in Verdinelli and Wasserman (1995), we have the posteriors denoted by $\pi_1(\theta)$ under the null hypothesis and $\pi_2(\theta, \psi)$ under the alternative hypothesis:

$$\pi_1(\theta) = \frac{p_1(\theta)}{c_1} \quad \text{and} \quad \pi_2(\theta, \psi) = \frac{p_2(\theta, \psi)}{c_2},$$

where

$$\begin{aligned} p_1(\theta) &= \Big[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(x_i - \mu)^2}{2\sigma^2} \Big) \Big] \cdot \frac{1}{\sigma} \\ &= \frac{1}{(\sqrt{2\pi})^N \sigma^{N+1}} \exp\Big(- \frac{(N-1)s^2 + N(\mu - \bar{x})^2}{2\sigma^2} \Big) \end{aligned}$$

and

$$\begin{aligned} p_2(\theta, \psi) &= \Big[\prod_{i=1}^{N} \frac{\Gamma(\frac{1+\psi}{2\psi})\sqrt{\psi}}{\sqrt{\pi}\sigma\Gamma(\frac{1}{2\psi})} \frac{1}{\Big(1 + \frac{\psi(x_i - \mu)^2}{\sigma^2}\Big)^{\frac{1+\psi}{2\psi}}} \Big] \cdot \frac{1}{\sigma} \\ &= \frac{\psi^{\frac{N}{2}}}{(\sqrt{\pi})^N \sigma^{N+1}} \cdot \Big[\frac{\Gamma(\frac{1+\psi}{2\psi})}{\Gamma(\frac{1}{2\psi})} \Big]^N \cdot \prod_{i=1}^{N} \frac{1}{\Big(1 + \frac{\psi(x_i - \mu)^2}{\sigma^2}\Big)^{\frac{1+\psi}{2\psi}}}. \end{aligned}$$

Thus, the Bayes factor is $r = c_1/c_2$. It is easy to see that θ is two-dimensional ($k = 2$) and ψ is one-dimensional ($d = 1$).

Now we apply Algorithms OIS, GOBS and GORIS given in Section 4 to obtain estimates $\hat{r}_{OIS}$, $\hat{r}_{GOBS}$ and $\hat{r}_{GORIS}$ for the Bayes factor r when $d = 1$. It should be mentioned that in this case the generalized Savage-Dickey density ratio estimate of Verdinelli and Wasserman (1995) is exactly the same as the optimal importance sampling estimate $\hat{r}_{OIS}$. In fact, as discussed in Verdinelli and Wasserman (1995), in this case the Savage-Dickey formula holds and the Bayes factor reduces to the posterior marginal for ψ (with respect to $\pi_2(\theta, \psi)$) evaluated at $\psi = 0$. Therefore, the generalized Savage-Dickey density ratio estimate is simply the estimate of this posterior marginal at $\psi = 0$ given by equation (2) of Verdinelli and Wasserman (1995), which is exactly $\hat{r}_{OIS}$.

To implement our three algorithms, we need to sample from π_1 and π_2. Sampling from π_1 is straightforward. To sample from π_2, instead of using an independence chain sampling scheme as in Verdinelli and Wasserman (1995), we use Gibbs sampling by introducing auxiliary variables (latent variables). Note that a t distribution is a scale mixture of normal distributions (e.g., see Albert and Chib (1993)). Let $\lambda = (\lambda_1, \ldots, \lambda_N)$ and let the joint distribution of (θ, ψ, λ) be

$$\pi_2^*(\theta, \psi, \lambda) \propto \Big[\prod_{i=1}^{N} \Big(\frac{\sqrt{\lambda_i}}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{\lambda_i(x_i-\mu)^2}{2\sigma^2}\Big)\Big)\Big(\frac{1}{\Gamma(\frac{1}{2\psi})}(\frac{1}{2\psi})^{\frac{1}{2\psi}} \lambda_i^{\frac{1}{2\psi}-1} \exp\Big(-\frac{1}{2\psi}\lambda_i\Big)\Big)\Big]\frac{1}{\sigma}.$$

Then, the marginal distribution of (θ, ψ) is $\pi_2(\theta, \psi)$. We run the Gibbs sampler by taking

$$\lambda_i \sim \mathcal{G}\Big(\frac{1+\psi}{\psi}, \frac{1}{2\psi} + \frac{(x_i-\mu)^2}{2\sigma^2}\Big) \text{ for } i = 1, \ldots, N,$$

$$\mu \sim N\Big(\frac{\sum_{j=1}^N \lambda_j x_j}{\sum_{j=1}^N \lambda_j}, \frac{\sigma^2}{\sum_{j=1}^N \lambda_j}\Big),$$

$$\frac{1}{\sigma^2} \sim \mathcal{G}\Big(\frac{N}{2}, \frac{\sum_{j=1}^N \lambda_j (x_i-\mu)^2}{2}\Big),$$

and

$$\frac{1}{2\psi} \sim \pi(\frac{1}{2\psi}) \propto \frac{1}{(\frac{1}{2\psi})^2}\Big[\frac{(\frac{1}{2\psi})^{\frac{1}{2\psi}}}{\Gamma(\frac{1}{2\psi})}\Big]^N \Big(\prod_{j=1}^N \lambda_j\Big)^{\frac{1}{2\psi}} \exp\Big(-(\frac{1}{2\psi})\sum_{j=1}^N \lambda_j\Big),$$

where $\mathcal{G}(a, b)$ denotes a Gamma distribution with density $g(\lambda|a, b) \propto \lambda^{a-1} \exp(-b\lambda)$. Sampling λ_i, μ, and $\frac{1}{\sigma^2}$ from their corresponding conditional distributions is trivial and we use the adaptive rejection sampling algorithm of Gilks and

Wild (1992) to generate $\frac{1}{2\psi}$ from $\pi(\frac{1}{2\psi})$ since $\pi(\frac{1}{2\psi})$ is log-concave when $N \geq 4$. Therefore, the Gibbs sampler can be exactly implemented. We believe that our Gibbs sampling scheme is superior to an independence chain Metropolis sampling scheme.

We implement our three algorithms in double precision Fortran-77 using the IMSL subroutines. We follow exactly the steps for Algorithms OIS, GOBS and GORIS presented in Section 4. We obtain a "random" draw (θ_1, ψ_1), ..., (θ_n, ψ_n) from π_2 by using the aforementioned Gibbs sampling scheme. First, we use several diagnostic methods to check convergence of the Gibbs sampler recommended by Cowles and Carlin (1996). Second, we take every Bth "stationary" Gibbs iterate so that the autocorrelations for the two components of θ_i disappear. The autocorrelations are calculated by an IMSL subroutine DACF. We use another IMSL subroutine DQDAG to calculate $c(\theta_i)$. A random draw θ_{11}, ..., θ_{1n_1} from π_1 can be obtained by using an exact sampling scheme. For Algorithm GORIS, we choose $\pi_2(\theta, \psi)$ as π in Step 1 and take a "random" sample $\{(\theta_i, \psi_i), i = 1, \ldots, n_1\}$ from π_2 to calculate τ_{n_1}. In Step 2, we adopt Metropolis sampling with $\pi_2(\theta, \psi)$ as a proposal density. Let (θ_j, ψ_j) denote the current values of the parameters. We take candidate values (θ_c, ψ_c) from every Bth "stationary" Gibbs iterate with the target distribution $\pi_2(\theta, \psi)$. We compute $a = \min\{\frac{\omega(\theta_c)}{\omega(\theta_j)}, 1\}$ where $\omega(\theta) = |p_1(\theta)/c(\theta) - \tau_{n_1}|$. We set $(\theta_{j+1}, \psi_{j+1})$ equal to (θ_c, ψ_c) with acceptance probability a and to (θ_j, ψ_j) with probability $1 - a$. We then take every (B')th Metropolis iterate to obtain a "random" draw (ϑ_1, φ_1), ..., $(\vartheta_{n_2}, \varphi_{n_2})$. We make no claim that the above sampling schemes are the most efficient ones, but they provide roughly independent samples and they are also straightforward.

In order to obtain informative empirical evidence of the performance of OIS, GOBS, and GORIS, we conducted a small scale simulation study. We took a dataset of $N = 100$ random numbers from $N(0, 1)$. Using this dataset, first we implemented GOBS with $n_1 = n_2 = 50,000$ to obtain an approximate "true" value of the Bayes factor r, which gives $r = 6.958$. In our implementation, we took $B = 30$ for Gibbs sampling and $B' = 10$ for Metropolis sampling to ensure an approximately "independent" Monte Carlo sample obtained. (Note that the Gibbs sampler converged earlier than 500 iterations.) Second, we used $n = 1,000$ for Algorithm OIS, $n_1 = n_2 = 500$ for Algorithm GOBS and $n_1 = 200$ and $n_2 = 800$ for Algorithm GORIS and we estimated the Monte Carlo standard errors based on the estimated first-order approximation of $\mathrm{RE}(\hat{r})$ using the available random draws. (No extra random draws are required for this stage

of the computation.) For example, the standard error for $\hat{r}_{GOBS}$ is given by

$$\mathrm{se}(\hat{r}_{GOBS}) = \hat{r}_{GOBS}\Big(\frac{1}{ns_1s_2}\Big[\Big(\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{p_1(\theta_{2i})}{s_1p_1(\theta_{2i}) + s_2\hat{r}_{GOBS}c(\theta_{2i})}\Big)^{-1} - 1\Big]\Big)^{-1/2},$$

where $n = n_1 + n_2 = 1,000$. Third, using the above implementation scheme with the same simulated dataset, we independently replicated the three estimation procedures 500 times. Then, we calculated the averages of $\hat{r}_{OIS}$, $\hat{r}_{GOBS}$, and $\hat{r}_{GORIS}$, Monte Carlo standard errors (MC S.E.), estimated biases ($E(\hat{r}) - r$), mean-squared errors (MSE), averages of the approximate standard errors (Approx. S.E.), and the average CPU time. (Note that our computation was performed on the DEC-station 5000-260.) The results are summarized in Table 1.

Table 1. The results of simulation study

Method	Average of $\hat{r}$'s	Bias	MSE	MC S.E.	Approx. S.E.	Average CPU in Minutes
OIS	6.995	0.037	0.066	0.254	0.187	1.52
GOBS	6.971	0.013	0.063	0.250	0.193	1.22
GORIS	6.933	-0.025	0.054	0.231	0.184	2.10

From Table 1, we see that (i) all three averages are close to the "true" value and the biases are relatively small; (ii) GORIS produced a slightly smaller Monte Carlo standard error than the other two; (iii) all three approximate standard errors are slightly understated, which has appeal since we used the estimated first-order approximation of RE($\hat{r}$); (iv) GOBS used the least CPU time since sampling from $\pi_2(\theta, \psi)$ is much more expensive than sampling from $\pi_1(\theta)$ and GORIS used the most CPU time since sampling from $\pi^*_{n_1}(\theta, \psi)$ in Step 2 of Algorithm GORIS is relatively more expensive. Finally, we notice that based on the above estimated value of r, the normal data results in a posterior marginal that is concentrated near $\psi = 0$ and leads to a Bayes factor strongly favoring the null hypothesis of normality and we also note that our estimated value of the Bayes factor disagrees with Verdinelli and Wasserman (1995), probably due to (i) use of a different simulated dataset, (ii) use of an exact Gibbs sampling scheme, and (iii) use of an IMSL numerical integration subroutine.

5.2. Testing for ordered alternatives of normal means

The aim of our second example is to illustrate our implementation algorithms when $d > 1$. Suppose that we have observations $\{x_i = (x_{i1}, \ldots, x_{it}),\ \ i = 1, \ldots, N\}$ where the x_{ij} are independently from $N(\mu_j, \sigma^2)$. We would like to

test whether $\mu_1, \ldots, \mu_t$ are ordered. For simplicity, we consider a special monotone ordering, namely, $\mu_1 < \cdots < \mu_t$. Therefore, our null hypothesis is $H_1 : \mu_1 = \cdots = \mu_t$. Let $\theta = (\mu_1, \sigma)$ and $\psi = (\mu_2, \ldots, \mu_t)$. Under the null and alternative hypotheses (H_1 and H_2), we take independent priors for the location parameters μ_i and scale parameter σ so that $\pi^{H_1}(\theta) = \pi^{H_1}(\mu_1)\pi^{H_1}(\sigma)$ where $\pi^{H_1}(\mu_1) = \frac{D_1}{\sqrt{2\pi}} \exp(-\frac{D_1^2\mu_1^2}{2})$ and $\pi^{H_1}(\sigma) \propto \sigma^{-1}$ and $\pi^{H_2}(\theta, \psi) = \pi^{H_2}(\mu_1, \ldots, \mu_t)\pi^{H_2}(\sigma)$ where $\pi^{H_2}(\mu_1, \ldots, \mu_t) = \frac{1}{c_N} \prod_{j=1}^t \frac{D_2}{\sqrt{2\pi}} \exp(-\frac{D_2^2\mu_j^2}{2})$ with the restriction $\mu_1 < \cdots < \mu_t$, $c_N = \int_{\mu_1<\cdots<\mu_t} \prod_{j=1}^t \frac{D_2}{\sqrt{2\pi}} \exp(-\frac{D_2^2\mu_j^2}{2}) d\mu_1, \ldots, d\mu_t$, and $\pi^{H_2}(\sigma) \propto \sigma^{-1}$. Following Laud and Ibrahim (1996), we take D_1 and D_2 to be of the form: $D_1 = d_0^{(1)} D$ and $D_2 = d_0^{(4)} D$ where

$$d_0^{(1)} = \frac{ba}{1 - ba} \quad \text{and} \quad d_0^{(4)} = \frac{ba^{1/4}}{1 - ba^{1/4}}. \tag{5.1}$$

In (5.1) b and a are determined by

$$ba = u_1 \quad \text{and} \quad ba^{1/4} = u_2.$$

In order to complete prior elicitation, we need to specify values of D and $0 < u_1 < u_2 < 1$ where D, u_1 and u_2 reflect sharp or vague prior beliefs (see, e.g., Laud and Ibrahim (1996)). With the above prior specification, we have the posteriors denoted by $\pi_1(\theta)$ under H_1 and $\pi_2(\theta, \psi)$ under H_2:

$$\pi_1(\theta) = \frac{p_1(\theta)}{c_1} \quad \text{and} \quad \pi_2(\theta, \psi) = \frac{p_2(\theta, \psi)}{c_2}. \tag{5.2}$$

In (5.2),

$$\begin{aligned} p_1(\theta) &= \Big[\prod_{i=1}^N \prod_{j=1}^t \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x_{ij} - \mu_1)^2}{2\sigma^2}\Big)\Big] \frac{D_1}{\sqrt{2\pi}} \exp\Big\{-\frac{D_1^2\mu_1^2}{2}\Big\} \frac{1}{\sigma} \\ &= \frac{D_1}{(\sqrt{2\pi})^{tN+1}\sigma^{tN+1}} \exp\Big\{-\Big[\frac{(tN-1)s^2 + tN(\mu_1 - \bar{x})^2}{2\sigma^2} + \frac{D_1^2\mu_1^2}{2}\Big]\Big\} \end{aligned}$$

and

$$\begin{aligned} p_2(\theta, \psi) &= \Big[\prod_{i=1}^N \prod_{j=1}^t \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x_{ij} - \mu_j)^2}{2\sigma^2}\Big)\Big] \frac{1}{c_N} \prod_{j=1}^t \frac{D_2}{\sqrt{2\pi}} \exp\Big(-\frac{D_2^2\mu_j^2}{2}\Big) \frac{1}{\sigma} \\ &= \frac{D_2^t}{c_N(\sqrt{2\pi})^{tN+t}\sigma^{tN+1}} \\ &\quad \exp\Big\{-\Big[\frac{(tN-1)s^2 - N\sum_{j=1}^t (\bar{x}_j - \bar{x})^2 + N\sum_{j=1}^t (\bar{x}_j - \mu_j)^2}{2\sigma^2} + \frac{D_2^2 \sum_{j=1}^t \mu_j^2}{2}\Big]\Big\}, \end{aligned}$$

where $\bar{x}$ and s^2 are the sample mean and the sample variance of all the x_{ij}'s and $\bar{x}_j$ is the sample mean of $x_{1j}, \ldots, x_{Nj}$ for $j = 1, \ldots, t$. We choose $P(H_1) = P(H_2) = 0.5$ *a priori*. Thus, the Bayes factor is $r = c_1/c_2$. In this case, θ is two-dimensional ($k = 2$) and ψ is $(t-1)$-dimensional ($d = t-1$). When $t > 2$, we have $d > 1$.

The implementation of our three algorithms is almost exactly the same as that in the first example. We sample from $\pi_1(\theta)$, $\pi_2(\theta, \psi)$ and $\pi_2(\psi|\theta)$ using Gibbs sampling. For example, to sample from $\pi_2(\theta, \psi)$, we run the Gibbs sampler by drawing

$$\mu_j \sim N\Big(\frac{N\bar{x}_j}{N + D_2^2\sigma^2}, \frac{\sigma^2}{N + \sigma^2 D_2^2}\Big),$$

where $\mu_{j-1} < \mu_j < \mu_{j+1}$ ($\mu_0 = -\infty$ and $\mu_{t+1} = \infty$) and

$$\frac{1}{\sigma^2} \sim \mathcal{G}\Big(\frac{tN+t}{2}, \frac{(tN-1)s^2 - N\sum_{j=1}^t(\bar{x}_j - \bar{x})^2 + N\sum_{j=1}^t(\bar{x}_j - \mu_j)^2}{2} + \frac{D_2^2\sum_{j=1}^t \mu_j^2}{2}\Big).$$

Then, we take every Bth "stationary" Gibbs iterate to obtain an approximately "independent" Monte Carlo sample. To calculate estimates $\hat{r}_{OIS}$, $\hat{r}_{GOBS}$ and $\hat{r}_{GORIS}$, we use (4.7), (4.8), (4.9) and (4.10) instead of using (4.1), (4.2), (4.3) and (4.5) in the first example. Note that when $d > 1$, $\pi_2(\psi|\theta)$ is not analytically available and when $d > 2$, it is difficult or expensive to use a numerical integration method to evaluate $c(\theta)$. Therefore, we use (4.6) to obtain $\hat{\pi}_2(\psi|\theta)$, an estimate of $\pi_2(\psi|\theta)$. Although $\pi_2(\psi|\theta)$ is not available in closed form, we have an explicit expression of $\pi_2^{(j)}(\psi|\theta)$, that is,

$$\begin{aligned}\pi_2^{(j)}(\psi|\theta) &= \pi_2(\mu_j|\mu_1, \ldots, \mu_{j-1}, \ldots, \mu_{j+1}, \ldots, \mu_t, \sigma) \\ &= \frac{\frac{1}{\sqrt{2\pi}\sigma_j^*}\exp(-\frac{(\mu_j - \xi_j)^2}{2\sigma^{*2}_j})}{\Phi\Big(\frac{\mu_{j+1}-\xi_j}{\sigma_j^*}\Big) - \Phi\Big(\frac{\mu_{j-1}-\xi_j}{\sigma_j^*}\Big)},\end{aligned}$$

for $\mu_{j-1} < \mu_j < \mu_{j+1}$ and $j = 2, 3, \ldots, t$ where $\xi_j = \frac{N\bar{x}_j}{N + D_2^2\sigma^2}$, $\sigma^{*2}_j = \frac{\sigma^2}{N + \sigma^2 D_2^2}$, and Φ is the standard normal cumulative distribution function.

We generated a dataset of $N \times t = 100 \times 4$ random numbers from $N(0, 1)$. We took $B = 5$ for Gibbs sampling, $B' = 5$ for Metropolis sampling, and $m = 1,000$ for estimating $\pi_2(\psi|\theta)$. We used $n = 1,000$ for Algorithm OIS, $n_1 = n_2 = 500$ for Algorithm GOBS, and $n_1 = 200$ and $n_2 = 800$ for Algorithm GORIS. We obtained c_N with a value of 0.0417 by using a Monte Carlo method with 5,000,000 replicates. Since the number of replicates is so large that the Monte Carlo error

is negligible. To specify the prior distributions, we took $D = 1$, $u_1 = 0.2$ and $u_2 = 0.5$. Then, we obtained $\hat{r}_{OIS}$, $\hat{r}_{GOBS}$ and $\hat{r}_{GORIS}$ with the standard errors in parentheses to be 46.05 (7.96), 28.52 (3.01) and 34.99 (1.94) respectively. (Note that the reported standard errors are based on the first-order approximation of $\mathrm{RE}(\hat{r})$.) Based on the above estimates, the normal data with unordered means yields a Bayes factor strongly favoring the null hypothesis. Note that different choices of D, u_1 and u_2 might lead to different values of Bayes factor. (See Ibrahim, Chen, and MacEachern (1996) for a comprehensive sensitivity study for the prior parameters). Also note that in order to obtain an approximate "true" value of the Bayes factor r, using the same dataset we implemented Algorithm GOBS with $n_1 = n_2 = 10,000$ and we obtained $\hat{r}_{GOBS} = 32.96$ with a standard error of 0.73. Therefore, OIS seems very unreliable in this case. Finally, we note that similar to Example 5.1, GOBS used the least CPU time and GORIS used as twice amount of CPU time as GOBS.

6. Conclusions

In this article, we extended importance sampling, bridge sampling, and ratio importance sampling to the cases where two densities have different dimensions and we found the global optimal solutions of such extensions. We also provided practically useful implementation algorithms for obtaining these global optimal estimators.

We used two examples to illustrate the methodology as well as the implementation algorithms developed in this paper. In both examples, we implemented the asymptotically optimal versions of Algorithms OIS, GOBS and GORIS, which are relatively computationally intensive. However, for higher dimensional or more complex problems, "approximate" optimal approaches proposed in Section 4 may be more attractive since they require much less computation effort. Finally, we note that the two-stage GORIS algorithm typically performs better when a small sample size n_1 in Step 1 is chosen. A rule of thumb of choosing n_1 and n_2 is that $n_1/n_2 \approx 1/4$.

The different dimensions problems are the important ones as they often arise in Bayesian model comparison and Bayesian variable selection. As our algorithms can asymptotically or approximately achieve the optimal simulation errors and they can be programmed in a routine manner, our methodology developed in this paper will be useful in computing Bayes factors (Kass and Raftery (1995)) or intrinsic Bayes factors (Berger and Pericchi (1996)) and in Bayesian comparisons (Geweke (1994)) or model selection. In fact, our methods have been successfully applied to Bayesian variable selection for proportional hazards models (Ibrahim, Chen, and MacEachern (1996)) and Bayesian analysis of correlated Binary data models (Dey and Chen (1996)).

Acknowledgements

The work of the second-named author was partially supported by a National University of Singapore Research Project. The authors thank Professors James O. Berger and Bradley P. Carlin for many helpful suggestions. We are also grateful to the associate editor and the two referees for their generous comments which greatly improved the quality of the paper.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669-679.

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

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B* **55**, 25-37.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **57**, 473-484.

Chen. M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *J. Amer. Statist. Assoc.* **89**, 818-824.

Chen, M.-H. and Schmeiser, B. W. (1994). Random-direction interior-point Markov chains: a family of black-box samplers. *Proceedings of the Section on Bayesian Statistical Science.* American Statistical Association, Toronto, Canada, 1-6.

Chen, M.-H. and Schmeiser, B. W. (1993). Performance of the Gibbs, hit-and-run, and Metropolis samplers. *The J. Comput. Graph. Statist.* **2**, 251-272.

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

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

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Assoc.* **91**, 883-904.

Dey, D. K. and Chen, M.-H. (1996). Bayesian analysis of correlated binary data models. Technical report 96-02, Department of Statistics, University of Connecticut.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Gelfand, A. E., Smith, A. F. M. and Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Amer. Statist. Assoc.* **87**, 523-532.

Gelman, A. and Meng, X.-L. (1994). Path sampling for computing normalizing constants: identities and theory. Technical Report 377, Department of Statistics, The University of Chicago.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence* **6**, 721-741.

Geweke, J. (1989). Bayesian inference in econometrics models using Monte Carlo integration. *Econometrica* **57**, 1317-1339.

Geweke, J. (1994). Bayesian comparison of econometric models. Technical Report 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 Technical Report No. 568, School of Statistics, University of Minnesota.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337-348.

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

Ibrahim, J. G., Chen, M.-H. and MacEachern, S. N. (1996). Bayesian variable selection for proportional hazards models. Technical Report, Department of Biostatistics, Harvard School of Public Health.

Jeffreys, H. (1961). *Theory of Probability*. Third Edition. Clarendon Press: Oxford.

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

Laud, P. W. and Ibrahim, J. G. (1996). Predictive specification of prior model probabilities in variable selection. *Biometrika* **83**, 267-274.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1092.

Müller, P. (1991). A generic approach to posterior integration and Gibbs sampling. Technical Report #91-09, Department of Statistics, Purdue University.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *J. Roy. Statist. Soc. Ser. B* **56**, 3-48.

Polson, N. G. (1996). Convergence of Markov chain Monte Carlo algorithms. In *Bayesian Statistics* **5** (Edited by J. M. Bernado, J. O. Berger, A. P. Dawid and A. F. M. Smith), 297-322. Oxford University Press.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. *J. Amer. Statist. Assoc.* **87**, 861-868.

Schervish, M. J. and Carlin, B. P. (1992). On the convergence of successive substitution sampling. *J. Comput. Graphical Statist.* **1**, 111-127.

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

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22**, 1701-1762.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Amer. Statist. Assoc.* **90**, 614-618.

Verdinelli, I. and Wasserman, L. (1996). Bayes factors, nuisance parameters and imprecise tests. In *Bayesian Statistics* **5** (Edited by J. M. Bernado, J. O. Berger, A. P. Dawid and A. F. M. Smith), 765-772. Oxford University Press.

Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280, U.S.A.

E-mail: mhchen@wpi.edu

Department of Mathematics, University of Oregon, Eugene, OR 97403-1222, U.S.A.

E-mail: qmshao@darkwing.uoregon.edu

(Received January 1995; accepted June 1996)