The Annals of Statistics
1997, Vol. 25, No. 4, 1563–1594

ON MONTE CARLO METHODS FOR ESTIMATING RATIOS OF NORMALIZING CONSTANTS

BY MING-HUI CHEN AND QI-MAN SHAO

Worcester Polytechnic Institute and University of Oregon

Recently, estimating ratios of normalizing constants has played an important role in Bayesian computations. Applications of estimating ratios of normalizing constants arise in many aspects of Bayesian statistical inference. In this article, we present an overview and discuss the current Monte Carlo methods for estimating ratios of normalizing constants. Then we propose a new ratio importance sampling method and establish its theoretical framework. We find that the ratio importance sampling method can be better than the current methods, for example, the bridge sampling method (Meng and Wong) and the path sampling method (Gelman and Meng), in the sense of minimizing asymptotic relative mean-square errors of estimators. An example is given for illustrative purposes. Finally, we present two special applications and the general implementation issues for estimating ratios of normalizing constants.

1. Introduction. Let $\pi_i(\boldsymbol{\theta})$, $i = 1, 2$, be two densities, each of which is known up to a normalizing constant:

$$\pi_i(\boldsymbol{\theta}) = \frac{p_i(\boldsymbol{\theta})}{c_i}, \qquad \boldsymbol{\theta} \in \Omega_i, \tag{1.1}$$

where Ω_i is the support of π_i for $i = 1, 2$. Then, the ratio of two normalizing constants is defined as

$$r = \frac{c_1}{c_2}. \tag{1.2}$$

In this article, we also use the parameter $\boldsymbol{\lambda}$ to index different densities:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_i) = \frac{p(\boldsymbol{\theta}|\boldsymbol{\lambda}_i)}{c(\boldsymbol{\lambda}_i)} \quad \text{for } i = 1, 2$$

and then the ratio is

$$r = \frac{c(\boldsymbol{\lambda}_1)}{c(\boldsymbol{\lambda}_2)}. \tag{1.3}$$

Estimating ratios of normalizing constants is extremely challenging and very important, particularly in Bayesian computations. Such problems typically arise in likelihood inference, especially in the presence of missing data

Received August 1994; revised December 1996.

AMS 1991 *subject classifications*. Primary 62E25; secondary 62A15, 62A99.

Key words and phrases. Bayesian computation, bridge sampling, Gibbs sampling, importance sampling, Markov chain Monte Carlo, Metropolis–Hastings algorithm, path sampling, ratio importance sampling.

[cf. Meng and Wong (1996)], in computing the intrinsic Bayes factors [see Berger and Pericchi (1996)], in the Bayesian comparison of econometric models considered by Geweke (1994), in Gibbs sampling [see Chen (1994a)], and in estimating marginal likelihood [cf. Chib (1995)]. For example, in likelihood inference, this ratio is viewed as the likelihood ratio and in the Bayesian model selection, the ratio is called the Bayes factor.

The $\pi_i(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_i)$ are often very complicated and therefore, the ratio defined by either (1.2) or (1.3) is analytically intractable [see Meng and Wong (1996), Gelman and Meng (1994), and Geyer (1994)]. However, without knowing the normalizing constants, c_i or $c(\boldsymbol{\lambda}_i)$, $i = 1, 2$, the distributions, $\pi_i(\boldsymbol{\theta})$ or $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_i)$, $i = 1, 2$, can be easily sampled by means of the Markov chain Monte Carlo (MCMC) methods, for example, the Metropolis–Hastings algorithm [see Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) and Hastings (1970)], the Gibbs sampler [cf. Geman and Geman (1984) and Gelfand and Smith (1990)], and the various hybrid algorithms [cf. Chen and Schmeiser (1993), Müller (1991) and Tierney (1994)]. Therefore, simulation-based methods for estimating the ratio, r, seem to be very attractive because of their general applicability.

Recently, several Monte Carlo methods for estimating normalizing constants have been developed, which include bridge sampling of Meng and Wong (1996), path sampling of Gelman and Meng (1994), a data augmentation-based method of Chib (1995), and reverse logistic regression of Geyer (1994). Section 2 gives a brief summary of these methods. In Section 2, we also discuss how the importance-weighted marginal density estimation of Chen (1994b) and the reverse logistic regression of Geyer (1994) can be adapted for estimating ratios of normalizing constants and we further find that the reverse logistic regression is essentially equivalent to optimal bridge sampling.

The remainder of this article is organized as follows. In Section 3 we propose a new ratio importance sampling method for estimating r and explore the properties of the ratio importance sampling estimators. We show that ratio importance sampling is better than either bridge sampling or path sampling in the sense of minimizing asymptotic relative mean-square errors (or variances) of estimators. In Section 4, a theoretical example is used for comparing ratio importance sampling with simple importance sampling, bridge sampling, and path sampling. Two special applications and implementation issues are presented in Section 5. Finally, brief concluding remarks are provided in Section 6.

2. Current Monte Carlo methods. In this section, we give a brief overview of the current Monte Carlo methods for estimating the ratios of normalizing constants and discuss their properties.

2.1. *Importance sampling*. The standard method for estimating the ratios of normalizing constants is importance sampling [see, e.g., Geweke (1989)]. We present two versions of the importance sampling methods.

VERSION 1. Choose two importance sampling densities $\pi_i^I(\boldsymbol{\theta})$, $i = 1, 2$, which are completely known, for $\pi_i(\boldsymbol{\theta})$, $i = 1, 2$, respectively. Let $\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}, \ldots, \boldsymbol{\theta}_{in_i}$, $i = 1, 2$, be two independent draws from $\pi_i^I(\boldsymbol{\theta})$, $i = 1, 2$. Then a consistent estimator of r is

$$\hat{r} = \frac{(1/n_1)\sum_{j=1}^{n_1} p_1(\boldsymbol{\theta}_{1j})/\pi_1^I(\boldsymbol{\theta}_{1j})}{(1/n_2)\sum_{j=1}^{n_2} p_2(\boldsymbol{\theta}_{2j})/\pi_2^I(\boldsymbol{\theta}_{2j})}. \tag{2.1}$$

The performance of the estimator, $\hat{r}$, depends heavily on the choices of the $\pi_i^I(\boldsymbol{\theta})$. If the $\pi_i^I(\boldsymbol{\theta})$ are good approximations of the $\pi_i(\boldsymbol{\theta})$, this importance sampling method works well. However, it is often difficult to find $\pi_i^I(\boldsymbol{\theta})$ to serve as good importance sampling densities [see Geyer (1994), Green (1992) and Gelman and Meng (1994)]. When the parameter spaces, Ω_i, $i = 1, 2$, are constrained, good completely known importance sampling densities, $\pi_i^I(\boldsymbol{\theta})$, $i = 1, 2$, are not available or are extremely difficult to obtain [see Chen (1994a) or Gelfand, Smith, and Lee (1992) for practical examples].

VERSION 2. Let $\boldsymbol{\Theta}$ be a random variable from π_2. When $\Omega_1 \subset \Omega_2$, we have the identity:

$$r = \frac{c_1}{c_2} = E_2\left\{\frac{p_1(\boldsymbol{\Theta})}{p_2(\boldsymbol{\Theta})}\right\}, \tag{2.2}$$

where E_i denotes the expectation with respect to π_i $(i = 1, 2)$. Let $\boldsymbol{\theta}_{21}$, $\boldsymbol{\theta}_{22}, \ldots, \boldsymbol{\theta}_{2n}$ be a random draw from π_2. Then the ratio r can be estimated by

$$\hat{r} = \frac{1}{n}\sum_{i=1}^{n} \frac{p_1(\boldsymbol{\theta}_{2i})}{p_2(\boldsymbol{\theta}_{2i})}. \tag{2.3}$$

It is easy to see that when the two densities π_i have very little overlap [here, meaning $E_2(\pi_1(\boldsymbol{\Theta}))$ is very small], this importance sampling-based method will work poorly.

2.2. *Bridge sampling*. The generalization of (2.2) given by Meng and Wong (1996) is

$$\frac{c_1}{c_2} = \frac{E_2\{p_1(\boldsymbol{\Theta})\,\alpha(\boldsymbol{\Theta})\}}{E_1\{p_2(\boldsymbol{\Theta})\,\alpha(\boldsymbol{\Theta})\}}, \tag{2.4}$$

where $\alpha(\boldsymbol{\theta})$ is an arbitrary function defined on $\Omega_1 \cap \Omega_2$ such that

$$0 < \left| \int_{\Omega_1 \cap \Omega_2} \alpha(\boldsymbol{\theta})\, p_1(\boldsymbol{\theta})\, p_2(\boldsymbol{\theta})\, d\boldsymbol{\theta} \right| < \infty.$$

Then, letting $\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}, \ldots, \boldsymbol{\theta}_{in_i}$ be a random draw from π_i for $i = 1, 2$, an estimator $\hat{r}_\alpha$ of r is given by

$$\hat{r}_\alpha = \frac{n_2^{-1}\sum_{i=1}^{n_2} p_1(\boldsymbol{\theta}_{2i})\,\alpha(\boldsymbol{\theta}_{2i})}{n_1^{-1}\sum_{i=1}^{n_1} p_2(\boldsymbol{\theta}_{1i})\,\alpha(\boldsymbol{\theta}_{1i})}. \tag{2.5}$$

Let $n = n_1 + n_2$ and $s_i = n_i/n$ and assume $\lim_{n\to\infty} s_i > 0$, $i = 1, 2$. Meng and Wong (1996) showed that the optimal choice for α is given by

$$\alpha_{\text{opt}}(\boldsymbol{\theta}) = \frac{c}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}, \qquad \boldsymbol{\theta} \in \Omega \cap \Omega_2,\ c \neq 0, \tag{2.6}$$

which minimizes the relative mean-square error

$$RE^2(\hat{r}_\alpha) = \frac{E(\hat{r}_\alpha - r)^2}{r^2}, \tag{2.7}$$

where E denotes the expectation taken over all random draws, and the asymptotic minimal relative mean-square error is

$$(ns_1s_2)^{-1}\left[\left\{\int_{\Omega_1\cap\Omega_2} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right\}^{-1} - 1\right]. \tag{2.8}$$

Because c_1 and c_2 are unknown, the optimal α_{opt} is not directly usable. Meng and Wong (1996) constructed the following iterative estimator:

$$\hat{r}_{\text{opt}}^{(t+1)} = \frac{(1/n_2)\sum_{i=1}^{n_2} p_1(\boldsymbol{\theta}_{2i})/\left(s_1 p_1(\boldsymbol{\theta}_{2i}) + s_2\hat{r}_{\text{opt}}^{(t)} p_2(\boldsymbol{\theta}_{2i})\right)}{(1/n_1)\sum_{i=1}^{n_1} p_2(\boldsymbol{\theta}_{1i})/\left(s_1 p_1(\boldsymbol{\theta}_{1i}) + s_2\hat{r}_{\text{opt}}^{(t)} p_2(\boldsymbol{\theta}_{1i})\right)}. \tag{2.9}$$

They showed that for each $t\ (\geq 0)\hat{r}_{\text{opt}}^{(t+1)}$ provides a consistent estimator of r and that the unique limit, $\hat{r}_{\text{opt}}$, achieves the asymptotic minimal relative mean-squared error given in (2.8). Meng and Wong (1996) also considered several noniterative choices of α such as $\alpha = 1$, $\alpha = (p_1, p_2)^{-1/2}$ and $\alpha = (p_1 + p_2)^{-1}$.

As in the second version of the importance sampling method, the bridge sampling estimator $\hat{r}_\alpha$ given in (2.5) will be unstable when π_1 and π_2 have little overlap. For such cases, the following path sampling method of Gelman and Meng (1994) will substantially improve the simulation efficiency.

2.3. *Path sampling*. In this subsection, we use $p(\boldsymbol{\theta}|\boldsymbol{\lambda}_i)$ for the unnormalized density for $i = 1, 2$. As discussed in Gelman and Meng (1994), we can often construct a continuous path to link $p(\boldsymbol{\theta}|\boldsymbol{\lambda}_1)$ and $p(\boldsymbol{\theta}|\boldsymbol{\lambda}_2)$. Instead of directly working on r, Gelman and Meng (1994) proposed the path sampling method to estimate the logarithm of r, that is,

$$\xi = -\log(r) = -\log\left(c(\boldsymbol{\lambda}_1)/c(\boldsymbol{\lambda}_2)\right).$$

First, we consider $\boldsymbol{\lambda}$ to be a scalar quantity and use λ instead of $\boldsymbol{\lambda}$. Without loss of generality, assume that $\lambda \in [0, 1]$, $\lambda_1 = 0$ and $\lambda_2 = 1$. Gelman and Meng (1994) developed the following identity:

$$\xi = -\log\left\{\frac{c(\boldsymbol{\lambda}_1)}{c(\boldsymbol{\lambda}_2)}\right\} = E\left[\frac{U(\boldsymbol{\Theta}, \Lambda)}{\pi_\lambda(\Lambda)}\right], \tag{2.10}$$

where $U(\boldsymbol{\theta}, \lambda) = (d/d\lambda)\log(p(\boldsymbol{\theta}|\lambda))$, $\pi_\lambda(\lambda)$ is a prior density (completely known) for $\lambda \in [0, 1]$, and the expectation is taken with respect to the joint density

$\pi(\boldsymbol{\theta}, \lambda) = \pi(\boldsymbol{\theta}|\lambda)\pi_\lambda(\lambda)$. Let $(\boldsymbol{\theta}_i, \Lambda_i)$, $i = 1, 2, \ldots, n$, be a random draw from $\pi(\boldsymbol{\theta}, \lambda)$. Then, under certain regularity conditions, they derived the following consistent estimator of ξ:

$$\hat{\xi} = \frac{1}{n} \sum_{i=1}^{n} \frac{U(\boldsymbol{\theta}_i, \Lambda_i)}{\pi_\lambda(\Lambda_i)}. \tag{2.11}$$

The Monte Carlo variance of $\hat{\xi}$ is

$$\operatorname{Var}(\hat{\xi}) = \frac{1}{n}\left[\int_0^1 \frac{E_\lambda\{U^2(\boldsymbol{\theta}, \lambda)\}}{\pi_\lambda(\lambda)}\, d\lambda - \xi^2\right], \tag{2.12}$$

where the expectation E_λ is taken with respect to $\pi(\boldsymbol{\theta}|\lambda)$. They found the optimal prior density $\pi_\lambda^{\text{opt}}(\lambda)$ given by

$$\pi_\lambda^{\text{opt}}(\lambda) = \frac{\sqrt{E_\lambda\{U^2(\boldsymbol{\Theta}, \lambda)\}}}{\int_0^1 \sqrt{E_{\lambda'}\{U^2(\boldsymbol{\Theta}, \lambda')\}}\, d\lambda'}, \tag{2.13}$$

which minimizes the Monte Carlo variance $\operatorname{Var}(\hat{\xi})$ given in (2.12). The minimum value of $\operatorname{Var}(\hat{\xi})$ is

$$\operatorname{Var}_{\text{opt}}(\hat{\xi}) = \frac{1}{n}\left[\left(\int_0^1 \sqrt{E_\lambda\{U^2(\boldsymbol{\theta}, \lambda)\}}\, d\lambda\right)^2 - \xi^2\right]. \tag{2.14}$$

As discussed by Gelman and Meng (1994), it is difficult in general to find the optimal path, it is intuitive that the optimal Monte Carlo variance cannot be arbitrary small, and must be bounded below by a distance between $\pi(\boldsymbol{\theta}|\lambda_1)$ and $\pi(\boldsymbol{\theta}|\lambda_2)$. The following lemma confirms this conjecture.

LEMMA 2.1. *Under certain regularity conditions, we have*

$$\operatorname{Var}(\hat{\xi}) \geq \frac{4}{n}\int\left[\sqrt{\pi(\boldsymbol{\theta}|\lambda_1)} - \sqrt{\pi(\boldsymbol{\theta}|\lambda_2)}\right]^2 d\boldsymbol{\theta} \tag{2.15}$$

for any prior density $\pi_\lambda(\lambda)$ *with support* $[\lambda_1, \lambda_2]$.

PROOF. Without loss of generality, assume $\lambda_1 = 0$ and $\lambda_2 = 1$. Letting $c(\lambda) = \int p(\boldsymbol{\theta}|\lambda)\, d\boldsymbol{\theta}$, we have

$$\xi = \int_0^1 \left[\frac{d}{d\lambda}\log c(\lambda)\right] d\lambda$$

and

$$E_\lambda\{U^2(\boldsymbol{\Theta}, \lambda)\} = \int\left[\frac{d}{d\lambda}\log \pi(\boldsymbol{\theta}|\lambda)\right]^2 \pi(\boldsymbol{\theta}|\lambda)\, d\boldsymbol{\theta} + \left[\frac{d}{d\lambda}\log c(\lambda)\right]^2. \tag{2.16}$$

Equations (2.12) and (2.16) lead to

$$\begin{aligned} n\operatorname{Var}(\hat{\xi}) = {} & \int_0^1 \int\left[\frac{d}{d\lambda}\log \pi(\boldsymbol{\theta}|\lambda)\right]^2 \frac{\pi(\boldsymbol{\theta}|\lambda)}{\pi_\lambda(\lambda)}\, d\boldsymbol{\theta}\, d\lambda \\ & + \left[\int_0^1 \left[\frac{d}{d\lambda}\log c(\lambda)\right]^2 \frac{1}{\pi_\lambda(\lambda)}\, d\lambda - \xi^2\right]. \end{aligned} \tag{2.17}$$

Using the Cauchy–Schwarz inequality and $\int_0^1 \pi_\lambda(\lambda)\,d\lambda = 1$, we have

$$\begin{aligned}
&\int_0^1 \left[\frac{d}{d\lambda}\log c(\lambda)\right]^2 \frac{1}{\pi_\lambda(\lambda)}\,d\lambda - \xi^2 \\
&\qquad \geq \left[\int_0^1 \frac{(d/d\lambda)\log c(\lambda)}{\sqrt{\pi_\lambda(\lambda)}}\sqrt{\pi_\lambda(\lambda)}\,d\lambda\right]^2 - \xi^2 = 0.
\end{aligned} \tag{2.18}$$

Similarly,

$$\begin{aligned}
&\int_0^1\int \left[\frac{d}{d\lambda}\log \pi(\boldsymbol{\theta}|\lambda)\right]^2 \frac{\pi(\boldsymbol{\theta}|\lambda)}{\pi_\lambda(\lambda)}\,d\boldsymbol{\theta}\,d\lambda \\
&\quad = \int_0^1\int 4\left[\frac{d}{d\lambda}\sqrt{\pi(\boldsymbol{\theta}|\lambda)}\right]^2 \frac{1}{\pi_\lambda(\lambda)}\,d\boldsymbol{\theta}\,d\lambda \\
&\quad \geq 4\int\left[\int_0^1 \frac{(d/d\lambda)\sqrt{\pi(\boldsymbol{\theta}|\lambda)}}{\sqrt{\pi_\lambda(\lambda)}}\sqrt{\pi_\lambda(\lambda)}\,d\lambda\right]^2 d\boldsymbol{\theta} \\
&\quad = 4\int\left[\int_0^1 \frac{d}{d\lambda}\sqrt{\pi(\boldsymbol{\theta}|\lambda)}\,d\lambda\right]^2 d\boldsymbol{\theta} \\
&\quad = 4\int\left[\sqrt{\pi(\boldsymbol{\theta}|\lambda_2)} - \sqrt{\pi(\boldsymbol{\theta}|\lambda_1)}\right]^2 d\boldsymbol{\theta}.
\end{aligned} \tag{2.19}$$

Thus, the lemma follows from (2.17), (2.18) and (2.19).

REMARK 2.1. The lower bound of $\mathrm{Var}(\hat{\xi})$ given in (2.15) indeed equals $(4/n)H^2(\pi_1, \pi_2)$, where $H(\pi_1, \pi_2)$ is the Hellinger distance between π_1 and π_2. We note that Gelman and Meng (1994) proved inequality (2.15) when $\pi_\lambda(\lambda)$ is a uniform prior density on λ.

Second, we consider $\boldsymbol{\lambda}$ to be d-dimensional. Assume that a continuous path in the d-dimensional parameter space that links $p(\boldsymbol{\theta}|\boldsymbol{\lambda}_1)$ and $p(\boldsymbol{\theta}|\boldsymbol{\lambda}_2)$ is given by

$$\boldsymbol{\lambda}(t) = (\boldsymbol{\lambda}_1(t), \ldots, \boldsymbol{\lambda}_d(t)) \quad \text{for } t \in [0,1]; \qquad \boldsymbol{\lambda}(0) = \boldsymbol{\lambda}_1 \quad \text{and} \quad \boldsymbol{\lambda}(1) = \boldsymbol{\lambda}_2.$$

Under some regularity conditions, Gelman and Meng (1994) obtained the identity

$$\xi = -\log\left\{\frac{c(\boldsymbol{\lambda}_1)}{c(\boldsymbol{\lambda}_2)}\right\} = \int_0^1 E_{\boldsymbol{\lambda}(t)}\left[\sum_{k=1}^d \dot{\lambda}_k(t) U_k(\boldsymbol{\theta}, \boldsymbol{\lambda}(t))\right] dt, \tag{2.20}$$

where $\dot{\lambda}_k(t) = (d\lambda_k(t)/dt)$ and $U_k(\boldsymbol{\theta}, \boldsymbol{\lambda}(t)) = (\partial \log p(\boldsymbol{\theta}|\boldsymbol{\lambda}))/\partial\lambda_k$ for $k = 1, 2, \ldots, d$. Then, a corresponding path sampling estimator for ξ is given by

$$\hat{\xi} = \frac{1}{n}\sum_{i=1}^n \left[\sum_{k=1}^d \dot{\lambda}_k(t_i) U_k(\boldsymbol{\theta}_i, \boldsymbol{\lambda}(t_i))\right], \tag{2.21}$$

where the t_i's are sampled uniformly from $[0, 1]$ and $\boldsymbol{\theta}_i$ is a draw from $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}(t_i))$. The variance of $\hat{\xi}$ is

$$\text{(2.22)} \qquad \operatorname{Var}(\hat{\xi}) = \frac{1}{n}\left[\int_0^1 \left(\sum_{i,j=1}^{d} g_{ij}(\boldsymbol{\lambda}(t))\,\dot{\lambda}_i(t)\,\dot{\lambda}_j(t)\right) dt - \xi^2\right],$$

where $g_{ij}(\boldsymbol{\lambda}(t)) = E_{\boldsymbol{\lambda}(t)}\{U_i(\boldsymbol{\Theta}, \boldsymbol{\lambda}(t))U_j(\boldsymbol{\Theta}, \boldsymbol{\lambda}(t))\}$. The path function $\boldsymbol{\lambda}(t)$ that minimizes the first term on the right-hand side of (2.22) is the solution of the following Euler–Lagrange equations with the boundary conditions $\boldsymbol{\lambda}(t) = \boldsymbol{\lambda}_t$ for $t = 1, 2$:

$$\text{(2.23)} \quad \sum_{i=1}^{d} g_{ij}(\boldsymbol{\lambda}(t))\,\ddot{\lambda}_i(t) + \sum_{i,j=1}^{d} [ij, k]\,\dot{\lambda}_i(t)\,\dot{\lambda}_j(t) = 0 \quad \text{for } k = 1, 2, \ldots, d,$$

where $\ddot{\boldsymbol{\lambda}}(t)$ denotes the second derivative with respect to t and $[ij, k]$ is the Christoffel symbol of the first kind:

$$[ij, k] = \frac{1}{2}\left[\frac{\partial g_{ik}(\boldsymbol{\lambda})}{\partial \lambda_j} + \frac{\partial g_{jk}(\boldsymbol{\lambda})}{\partial \lambda_i} - \frac{\partial g_{ij}(\boldsymbol{\lambda})}{\partial \lambda_k}\right], \qquad i, j, k = 1, 2, \ldots, d.$$

In general, it is not trivial to find the optimal path from (2.23). Also, if $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are far away from each other, the path sampling method might work poorly since simulation efficiency could be lost in averaging over a "long" path.

2.4. *Marginal likelihood.* In the context of Bayesian inference, the posterior is typically of the form

$$\pi(\boldsymbol{\theta}|x) = L(x, \boldsymbol{\theta})\pi(\boldsymbol{\theta})/m(x),$$

where $L(x, \boldsymbol{\theta})$ is the likelihood function, x is the data, $\boldsymbol{\theta}$ is the parameter vector, $\pi(\boldsymbol{\theta})$ is a prior and $m(x)$ is the marginal density (marginal likelihood). Clearly, $m(x)$ is the normalizing constant of the posterior distribution $\pi(\boldsymbol{\theta}|x)$. Calculating the marginal likelihood, $m(x)$, plays an important role in the computation of Bayes factors.

Chib (1995) considered the following identity:

$$\text{(2.24)} \qquad m(x) = \frac{L(x, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|x)}.$$

Let $\boldsymbol{\theta}^*$ be the posterior mean or the posterior mode and also let $\hat{\pi}(\boldsymbol{\theta}^*|x)$ be an estimator of the joint posterior density evaluated at $\boldsymbol{\theta}^*$. Chib (1995) obtained the following estimator for $m(x)$:

$$\text{(2.25)} \qquad \hat{m}(x) = \frac{L(x, \boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)}{\hat{\pi}(\boldsymbol{\theta}^*|x)}.$$

Then, Chib (1995) developed a data augmentation technique [cf. Tanner and Wong (1987)] to estimate $\hat{\pi}(\boldsymbol{\theta}^*|x)$ by introducing latent variables. Chib's method is particularly useful for multivariate problems when the full condi-

tional densities are completely known. Indeed, $\hat{\pi}(\boldsymbol{\theta}^*|x)$ can also be estimated by using the importance-weighted marginal density estimation (IWMDE) method of Chen (1994b). The IWMDE method does not require completely known full conditional densities.

Further, the IWMDE method can be used to estimate $m(x)$ directly. Let $\boldsymbol{\theta}_i$, $i = 1, 2, \ldots, n$, be a sample from $\pi(\boldsymbol{\theta}|x)$. Such a sample can be obtained by employing MCMC methods (for example, the Gibbs sampler or a Metropolis–Hastings algorithm). Then, IWMDE yields a consistent estimator for $m(x)$,

$$\tilde{m}_{\mathrm{IWMDE}}(x) = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{w(\boldsymbol{\theta}_i)}{L(x,\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i)}\right]^{-1}, \tag{2.26}$$

where $w(\boldsymbol{\theta})$ is a weighted density function (completely known) with the support $\Omega_w \subset \Omega_{\pi(\cdot|x)}$ [the support of the posterior distribution $\pi(\cdot|x)$.] Chen (1994b) also provided an empirical procedure to achieve a fairly good w. Compared to (2.25), (2.26) does not require the selection of $\boldsymbol{\theta}^*$ and thus (2.26) might lead to a more efficient estimator for $m(x)$, especially when parameter spaces are constrained.

Chib's estimator works if a good approximation $\hat{\pi}(\boldsymbol{\theta}^*|x)$ and a good point $\boldsymbol{\theta}^*$ are chosen and the IWMDE method works if a good weighted density function w is selected. Obviously, the above methods can be used to estimate the ratio of two marginal likelihoods, and this is useful in the calculation of Bayes factors.

2.5. *Reverse logistic regression.* In this subsection, we discuss how reverse logistic regression of Geyer (1994) can be adapted for estimating ratios of normalizing constants.

Let $\{\Theta_{ij},\ j = 1, \ldots, n_i\}$, $i = 1, 2$, be independent random draws from π_i, $i = 1, 2$, respectively. Also let $n = n_1 + n_2$, $s_1 = n_1/n$ and $s_2 = n_2/n$. Consider a mixture distribution with density

$$\pi_{\mathrm{mix}}(\boldsymbol{\theta}) = s_1\frac{p_1(\boldsymbol{\theta})}{c_1} + s_2\frac{p_2(\boldsymbol{\theta})}{c_2}. \tag{2.27}$$

Define

$$\begin{aligned} q_1(\boldsymbol{\theta}, r) &= \frac{s_1 p_1(\boldsymbol{\theta})/c_1}{s_1 p_1(\boldsymbol{\theta})/c_1 + s_2 p_2(\boldsymbol{\theta})/c_2} = \frac{s_1 p_1(\boldsymbol{\theta})}{s_1 p_1(\boldsymbol{\theta}) + r s_2 p_2(\boldsymbol{\theta})}, \\ q_2(\boldsymbol{\theta}, r) &= \frac{s_2 p_2(\boldsymbol{\theta})/c_2}{s_1 p_1(\boldsymbol{\theta})/c_1 + s_2 p_2(\boldsymbol{\theta})/c_2} = \frac{r s_2 p_2(\boldsymbol{\theta})}{s_1 p_1(\boldsymbol{\theta}) + r s_2 p_2(\boldsymbol{\theta})}, \end{aligned} \tag{2.28}$$

and also define the log quasi-likelihood as

$$l_n(r) = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\log q_i(\boldsymbol{\Theta}_{ij}, r). \tag{2.29}$$

Then the reverse logistic regression estimator, $\hat{r}_{\mathrm{RLR}}$, of r is obtained by maximizing the log quasi-likelihood $l_n(r)$ in (2.29). Clearly, $\hat{r}_{\mathrm{RLR}}$ satisfies the following equation:

$$\sum_{j=1}^{n_2} \frac{s_1 p_1(\boldsymbol{\Theta}_{2j})}{\hat{r}_{\mathrm{RLR}}\big(s_1 p_1(\boldsymbol{\Theta}_{2j}) + \hat{r}_{\mathrm{RLR}} s_2 p_2(\boldsymbol{\Theta}_{2j})\big)} - \sum_{j=1}^{n_1} \frac{s_2 p_2(\boldsymbol{\Theta}_{1j})}{s_1 p_1(\boldsymbol{\Theta}_{1j}) + \hat{r}_{\mathrm{RLR}} s_2 p_2(\boldsymbol{\Theta}_{1j})} = 0. \tag{2.30}$$

Therefore, when π_1 and π_2 overlap, that is,

$$\int_{\Omega} \pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})\, d\boldsymbol{\theta} > 0,$$

and under some regularity conditions, we have

$$\hat{r}_{\mathrm{RLR}} \to r \quad \text{a.s. as } n \to \infty, \tag{2.31}$$

and the asymptotic value of $E((\hat{r}_{\mathrm{RLR}} - r)^2/r^2)$ is

$$\frac{1}{n s_1 s_2}\left[\left\{\int_{\Omega} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right\}^{-1} - 1\right]. \tag{2.32}$$

From (2.30) and (2.32), we can see that the reverse logistic regression estimator, $\hat{r}_{\mathrm{RLR}}$, is exactly the same as the optimal bridge sampling estimator, $\hat{r}_{\alpha_{\mathrm{opt}}}$, given by (2.5) and (2.6). Note that when π_1 and π_2 do not overlap, the reverse logistic regression method does not work directly.

3. A new ratio importance sampling identity and theory. In this section we present a new ratio importance sampling identity for the ratio $r = c_1/c_2$. Based on this simple identity, we propose a ratio importance sampling estimator for r, and then explore its theoretical properties.

Denote $\Omega = \Omega_1 \cup \Omega_2$. Let $\pi(\boldsymbol{\theta})$ be an arbitrary density over Ω such that $\pi(\boldsymbol{\theta}) > 0$ for $\boldsymbol{\theta} \in \Omega$. Then, we have the identity

$$r = \frac{c_1}{c_2} = \frac{E_\pi\{p_1(\boldsymbol{\Theta})/\pi(\boldsymbol{\Theta})\}}{E_\pi\{p_2(\boldsymbol{\Theta})/\pi(\boldsymbol{\Theta})\}}, \tag{3.1}$$

where E_π denotes the expectation with respect to π. We call (3.1) and $\pi(\boldsymbol{\theta})$ the ratio importance sampling identity and the ratio importance sampling density, respectively. Note that if $\pi = p_2/c_2$, then (3.1) leads to the importance sampling identity (2.2). Therefore, the ratio importance sampling identity is a generalization of the simple importance sampling identity (2.2).

Given a random draw $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ from π, which can often be facilitated by the MCMC methods (e.g., the Gibbs sampler or a Metropolis–Hastings algorithm), the ratio importance sampling estimator for $r = c_1/c_2$ is

$$\hat{r}_\pi = \frac{\sum_{i=1}^n p_1(\boldsymbol{\theta}_i)/\pi(\boldsymbol{\theta}_i)}{\sum_{i=1}^n p_2(\boldsymbol{\theta}_i)/\pi(\boldsymbol{\theta}_i)}. \tag{3.2}$$

For any π with the support Ω, $\hat{r}_\pi$ is a consistent estimator of r. Even for a dependent sample $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n\}$, under mild conditions (for example, ergodicity) the consistency for $\hat{r}_\pi$ still holds. One feature of (3.2) is that the estimator $\hat{r}_\pi$ can be obtained by using one single random draw from π instead of π_1 or π_2. Such a property is useful in Gibbs sampling when the posterior density contains an analytically intractable constant that depends on the hyper-parameters [see Gelfand, Smith and Lee (1992) and Chen (1994a)].

Because the ratio importance sampling estimator $\hat{r}_\pi$ depends on π, it is interesting to determine the optimal ratio importance sampling density π_{opt} of π. We use the relative mean-square error similar to (2.7):

$$RE^2(\hat{r}_\pi) = \frac{E_\pi(\hat{r}_\pi - r)^2}{r^2}. \tag{3.3}$$

The analytical calculation of (3.3) is typically intractable. However, under the assumption that the $\boldsymbol{\theta}_i$ are independently and identically distributed from π, we can obtain the asymptotic form of $\mathrm{RE}^2(\hat{r}_\pi)$. Let $f(\boldsymbol{\theta}) = p_1(\boldsymbol{\theta})/\pi(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta}) = p_2(\boldsymbol{\theta})/\pi(\boldsymbol{\theta})$. Then, we have $E_\pi f(\boldsymbol{\Theta}) = c_1$ and $E_\pi g(\boldsymbol{\Theta}) = c_2$.

THEOREM 3.1. *Let $\{\boldsymbol{\Theta}_i,\ i = 1, 2, \ldots, n\}$ be i.i.d. random variables from π. Assume $\int_\Omega |p_1(\boldsymbol{\theta}) - a\, p_2(\boldsymbol{\theta})| d\boldsymbol{\theta} > 0$ for every $a > 0$, $E_\pi((f(\boldsymbol{\Theta})/c_1) - (g(\boldsymbol{\Theta})/c_2))^2 < \infty$ and $E_\pi\{f(\boldsymbol{\Theta})/g(\boldsymbol{\Theta})\}^2 < \infty$. Then we have*

$$\begin{aligned} \lim_{n\to\infty} \{nRE^2(\hat{r}_\pi)\} &= \lim_{n\to\infty} n\left\{\frac{E_\pi(\hat{r}_\pi - r)^2}{r^2}\right\} \\ &= E_\pi\left\{\frac{f(\boldsymbol{\Theta})}{c_1} - \frac{g(\boldsymbol{\Theta})}{c_2}\right\}^2 \end{aligned} \tag{3.4}$$

and

$$\sqrt{n}\,(\hat{r}_\pi - r) \to_{\mathscr{D}} N\left(0, r^2 E_\pi\left\{\frac{f(\boldsymbol{\Theta})}{c_1} - \frac{g(\boldsymbol{\Theta})}{c_2}\right\}^2\right) \quad \text{as } n \to \infty. \tag{3.5}$$

If, in addition, $E_\pi((f(\boldsymbol{\Theta}))/c_1 - (g(\boldsymbol{\Theta}))/c_2)^4 < \infty$ and $E_\pi g^4(\boldsymbol{\Theta}) < \infty$, then

$$\begin{aligned} RE^2(\hat{r}_\pi) &= \frac{E_\pi(\hat{r}_\pi - r)^2}{r^2} \\ &= \frac{1}{n} E_\pi\left\{\frac{f(\boldsymbol{\Theta})}{c_1} - \frac{g(\boldsymbol{\Theta})}{c_2}\right\}^2 + O\left(\frac{1}{n^2}\right) \quad \text{as } n \to \infty. \end{aligned} \tag{3.6}$$

The proof is given in Appendix A. Note that if $\int_\Omega |p_1(\boldsymbol{\Theta}) - a p_2(\boldsymbol{\theta})| d\boldsymbol{\theta} = 0$ for some $a > 0$, then $c_1 = a c_2$, and therefore $\hat{r}_\pi = r = a$.

Using (3.4) or (3.6), we have the asymptotic form of $\mathrm{RE}^2(\hat{r}_\pi)$:

$$\frac{1}{n}E_\pi\left[\frac{\{p_1(\Theta)/c_1 - p_2(\Theta)/c_2\}^2}{\pi^2(\Theta)}\right]. \tag{3.7}$$

Note that when $\Omega_1 \subset \Omega_2$ and $\pi(\theta) = \pi_2(\theta) = p_2(\theta)/c_2$, (3.7) returns the exact value of $\mathrm{RE}^2(\hat{r}_\pi)$. For this case, (3.2) becomes the importance sampling estimator (2.3) for r, and the corresponding relative mean-square error is

$$RE_I^2(\hat{r}) = \frac{1}{n}\int_{\Omega_2}\frac{(\pi_1(\theta) - \pi_2(\theta))^2}{\pi_2(\theta)}\,d\theta, \tag{3.8}$$

which is the chi-squared distance $\chi^2(\pi_2, \pi_1)$ between π_2 and π_1.

The following theorem gives the optimal ratio importance sampling density π_{opt} that minimizes the asymptotic relative mean-square error (3.7).

THEOREM 3.2. *Assume* $\int_\Omega |p_1(\theta) - ap_2(\theta)|d\theta > 0$ *for every* $a > 0$. *The asymptotic relative mean-square error* (3.7) *is minimized at*

$$\pi_{\mathrm{opt}}(\theta) = \frac{|p_1(\theta)/c_1 - p_2(\theta)/c_2|}{\int_\Omega |p_1(\theta')/c_1 - p_2(\theta')/c_2|d\theta'} \tag{3.9}$$

with the asymptotic minimal value

$$\frac{1}{n}\left[\int_\Omega \left|\frac{p_1(\theta)}{c_1} - \frac{p_2(\theta)}{c_2}\right| d\theta\right]^2. \tag{3.10}$$

PROOF. By the Cauchy–Schwarz inequality, for an arbitrary density $\pi(\cdot)$,

$$\left[\int_\Omega \left|\frac{p_1(\theta)}{c_1} - \frac{p_2(\theta)}{c_2}\right| d\theta\right]^2 \le \int_\Omega \frac{[p_1(\theta)/c_1 - p_2(\theta)/c_2]^2}{\pi(\theta)}\,d\theta \int_\Omega \pi(\theta)\,d\theta.$$

Thus,

$$E_\pi\left[\frac{\{p_1(\Theta)/c_1 - p_2(\Theta)/c_2\}^2}{\pi^2(\Theta)}\right] \ge \left[\int_\Omega \left|\frac{p_1(\theta)}{c_1} - \frac{p_2(\theta)}{c_2}\right| d\theta\right]^2 \tag{3.11}$$

with equality holding if and only if (up to a zero-measure set)

$$\pi(\theta) \propto \left|\frac{p_1(\theta)}{c_1} - \frac{p_2(\theta)}{c_2}\right|,$$

that is, $\pi(\theta) = \pi_{\mathrm{opt}}(\theta)$. This proves (3.9). Replacing π by π_{opt} in (3.7) gives (3.10). □

It is interesting to see that (3.10) is $(1/n)L_1^2(\pi_1, \pi_2)$ where $L_1(\pi_1, \pi_2)$ is the L_1-distance between π_1 and π_2. From Theorem 3.2 and Equations (3.8) and (3.10), we also have $L_1^2(\pi_1, \pi_2) \le \chi^2(\pi_2, \pi_1)$.

Now, we compare the ratio importance sampling method with the bridge sampling method. The following theorem says that the ratio importance

sampling estimator (3.2) with the optimal π_{opt} given in (3.9) has a smaller asymptotic relative mean-square error than the bridge sampling estimator (2.5) with the optimal choice α_{opt} given in (2.6).

THEOREM 3.3. *For* $0 < s_1, s_2 < 1$ *and* $s_1 + s_2 = 1$, *we have*

$$\begin{aligned}&\left[\int_\Omega \left|\frac{p_1(\boldsymbol{\theta})}{c_1} - \frac{p_2(\boldsymbol{\theta})}{c_2}\right| d\boldsymbol{\theta}\right]^2 \\ &\quad \le (s_1 s_2)^{-1}\left[\left\{\int_{\Omega_1\cap\Omega_2} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right\}^{-1} - 1\right].\end{aligned} \tag{3.12}$$

PROOF. Noting that

$$\begin{aligned}&1 - \int_{\Omega_1\cap\Omega_2} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\ &\quad = \int_\Omega \frac{(s_2\pi_1(\boldsymbol{\theta}) + s_1\pi_2(\boldsymbol{\theta}))(s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})) - \pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\ &\quad = \int_\Omega \frac{\big(s_1 s_2\pi_1^2(\boldsymbol{\theta}) + s_1 s_2\pi_2^2(\boldsymbol{\theta}) + \big(s_1^2 + s_2^2 - 1\big)\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})\big)}{(s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta}))}\, d\boldsymbol{\theta} \\ &\quad = s_1 s_2 \int_\Omega \frac{(\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta}))^2}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta},\end{aligned} \tag{3.13}$$

we have that the right-hand side of (3.12) equals

$$\begin{aligned}&\int_\Omega \frac{(\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta}))^2}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \Big/ \int_\Omega \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\ &\quad = \int_\Omega \frac{(\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta}))^2}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\ &\qquad \times \int_\Omega \big(s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})\big)\, d\boldsymbol{\theta} \Big/ \int_\Omega \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \qquad (3.14) \\ &\quad \ge \left[\int_\Omega \frac{|\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})|}{\sqrt{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}} \right. \\ &\qquad \left. \times \sqrt{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right]^2 \Big/ \int_\Omega \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\ &\quad = \left[\int_\Omega |\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})|\, d\boldsymbol{\theta}\right]^2 \Big/ \int_\Omega \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}, \qquad (3.15)\end{aligned}$$

where (3.14) is obtained by the Cauchy–Schwarz inequality. From (3.13) it is easy to see that

$$\int_{\Omega_1\cap\Omega_2}\frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta})+s_2\pi_2(\boldsymbol{\theta})}\,d\boldsymbol{\theta}\le 1. \tag{3.16}$$

Now (3.12) follows from (3.15) and (3.16). □

Next, we compare the ratio importance sampling method with the path sampling method. Gelman and Meng (1994) pointed out that the asymptotic variance $\hat{\xi}$ is the same as the asymptotic relative mean-squared error of $\hat{r}$, that is, $\lim_{n\to\infty} n\,\mathrm{Var}(\hat{\xi}) = \lim_{n\to\infty} nE(\hat{r}-r)^2/r^2$. Thus, the next theorem shows that the asymptotic relative mean-square error of the ratio importance sampling estimator (3.2) with the optimal π_{opt} is less than the lower bound, given on the right-hand side of (2.15), of the variance of $\hat{\xi}$ given in (2.12).

THEOREM 3.4. *Defining* $\pi_i(\boldsymbol{\theta}) = -p_i(\boldsymbol{\theta})/c_i = \pi(\boldsymbol{\theta}|\lambda_i)$ *for* $i = 1, 2$, *we have*

$$\left[\int_\Omega\left|\frac{p_1(\boldsymbol{\theta})}{c_1}-\frac{p_2(\boldsymbol{\theta})}{c_2}\right|d\boldsymbol{\theta}\right]^2 \le 4\int\left[\sqrt{\pi(\boldsymbol{\theta}|\lambda_1)}-\sqrt{\pi(\boldsymbol{\theta}|\lambda_2)}\right]^2 d\boldsymbol{\theta}. \tag{3.17}$$

PROOF. By the Cauchy–Schwarz inequality, we have that the left-hand side of (3.17) equals

$$\begin{aligned}&\left[\int_\Omega\left|\sqrt{\pi_1(\boldsymbol{\theta})}-\sqrt{\pi_2(\boldsymbol{\theta})}\right|\left(\sqrt{\pi_1(\boldsymbol{\theta})}+\sqrt{\pi_2(\boldsymbol{\theta})}\right)d\boldsymbol{\theta}\right]^2\\ &\quad\le\int_\Omega\left[\sqrt{\pi_1(\boldsymbol{\theta})}-\sqrt{\pi_2(\boldsymbol{\theta})}\right]^2 d\boldsymbol{\theta}\int_\Omega\left[\sqrt{\pi_1(\boldsymbol{\theta})}+\sqrt{\pi_2(\boldsymbol{\theta})}\right]^2 d\boldsymbol{\theta}.\end{aligned} \tag{3.18}$$

It is easy to see that

$$\begin{aligned}&\int_\Omega\left[\sqrt{\pi_1(\boldsymbol{\theta})}+\sqrt{\pi_2(\boldsymbol{\theta})}\right]^2 d\boldsymbol{\theta}\\ &\quad=\int_\Omega\left[\pi_1(\boldsymbol{\theta})+2\sqrt{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}+\pi_2(\boldsymbol{\theta})\right]d\boldsymbol{\theta}\le 4.\end{aligned} \tag{3.19}$$

Thus, (3.17) follows from (3.18) and (3.19). □

From Theorem 3.4, we can see that $L_1^2(\pi_1,\pi_2)\le 4H^2(\pi_1,\pi_2)$. From Theorem 3.4, the optimal ratio importance sampling estimator $\hat{r}_{\pi_{\mathrm{opt}}}$ is always

better than the bridge sampling estimator and $\hat{r}_{\pi_{\text{opt}}}$ is also better than any path sampling estimator. However, π_{opt} depends on the unknown normalizing constants c_1 and c_2. Therefore, $\hat{r}_{\pi_{\text{opt}}}$ is not directly usable. We will address implementation issues in Section 5.

4. A simple example. In this example, we consider two case studies; the first one was also used in Meng and Wong (1996) to illustrate bridge sampling and in Gelman and Meng (1994) to illustrate path sampling.

CASE 1 [$N(0,1)$ and $N(D,1)$]. Let $p_1(\theta) = \exp(-\theta^2/2)$ and $p_2(\theta) = \exp(-(\theta - D)^2/2)$ with D a known positive constant. In this case, $c_1 = c_2 = \sqrt{2\pi}$ and, therefore, $r = 1$ and $\xi = -\log(r) = 0$. For path sampling, we consider p_1 and p_2 as two points in the family of unnormalized normal densities: $p(\theta|\boldsymbol{\lambda}) = \exp\{-(\theta - \mu)^2/2\sigma^2\}$, with $\boldsymbol{\lambda} = (\mu, \sigma)'$, $\boldsymbol{\lambda}_1 = (0,1)'$, and $\boldsymbol{\lambda}_2 = (D,1)'$.

As discussed in Gelman and Meng (1994), in order to make fair comparisons, we assume that (i) with importance sampling-version 2, we draw n i.i.d. observations from $N(D,1)$; (ii) with bridge sampling, we draw $n/2$ (assume n is even) i.i.d. observations from each of $N(0,1)$ and $N(D,1)$; (iii) with path sampling, we first draw t_i, $i = 1, 2, \ldots, n$ uniformly from $(0,1)$ and then draw an observation from $N(\mu(t_i), \sigma^2(t_i))$ where $\boldsymbol{\lambda}(t) = (\mu(t), \sigma(t))'$ is a given path and (iv) with ratio importance sampling, we draw n i.i.d. observations from the optimal ratio importance sampling density:

$$\pi_{\text{opt}}(\theta) = \frac{|\phi(\theta) - \phi(\theta - D)|}{c_{\text{opt}}(D)}, \tag{4.1}$$

where

$$\begin{aligned} c_{\text{opt}}(D) &= \int_{-\infty}^{\infty} |\phi(\theta) - \phi(\theta - D)|\, d\theta = 2(\Phi(D/2) - \Phi(-D/2)) \\ &= 2(2\Phi(D/2) - 1), \end{aligned} \tag{4.2}$$

and ϕ and Φ are the $N(0,1)$ probability density function and cumulative distribution function, respectively. Since the cumulative distribution function for $\pi_{\text{opt}}(\theta)$ is

$$\Pi_{\text{opt}}(\theta) = \begin{cases} (\Phi(\theta) - \Phi(\theta - D))/2(2\Phi(D/2) - 1), & \text{for } \theta \le D/2, \\ 1 - (\Phi(\theta) - \Phi(\theta - D))/2(2\Phi(D/2) - 1), & \\ & \text{for } \theta > D/2, \end{cases} \tag{4.3}$$

then the generation from π_{opt} can be done easily by the inversion method [see, e.g., Devroye (1986), pages 27–35].

Since the asymptotic variance of $\hat{\xi}$ is the same as the asymptotic relative mean-square error of $\hat{r}$ (i.e., $\lim_{n\to\infty} n\,\mathrm{Var}(\hat{\xi}) = \lim_{n\to\infty} nE(\hat{r}-r)^2/r^2$), then, using the results given by Gelman and Meng (1994), and (3.10) and (4.2), we obtain Table 1.

We define the relative simulation efficiency as follows:

$$e(i,j) = \frac{\lim_{n\to\infty}\sqrt{nE(\hat{r}-r)^2/r^2} \text{ for method } j}{\lim_{n\to\infty}\sqrt{nE(\hat{r}-r)^2/r^2} \text{ for method } i} \quad \text{for } i,j = 1,2,\ldots,7. \tag{4.4}$$

Then, $e(7,j)$, $j=1,\ldots,6$, versus D are plotted in Figure 1. Note that when $e(i,j)\geq 1$, method j has a greater asymptotic relative mean-square error than method i, and therefore, method i is more efficient than method j. It is easy to verify that $e(7,j)\geq\sqrt{2\pi}/2 = 1.2533$ for $j=1,2,\ldots,6$. It is interesting to note that $\lim_{D\to 0} e(7,j) = \sqrt{2\pi}/2 = 1.2533$ for all $j=1,2,\ldots,6$. The lower bound of path sampling in (2.15) is quite close to the asymptotic relative mean-square error of the ratio importance sampling method with the optimal π_{opt}. The ratio importance sampling method is significantly better than the bridge sampling method, especially for $D>3$, and it is also better than the path sampling method. In this case, both ratio importance sampling and path sampling are much better than importance sampling-version 2.

CASE 2 [$N(0,1)$ and $N(0,\Delta^2)$]. Without loss of generality, we consider $\Delta>1$ only. Let $p_1(\theta)=\exp(-\theta^2/2)$ and $p_2(\theta)=\exp(-\theta^2/2\Delta^2)$ with Δ a known positive constant. In this case, $c_1=\sqrt{2\pi}$, $c_2=\sqrt{2\pi}\Delta$ and, therefore, the ratio $r=c_1/c_2=1/\Delta$. For path sampling, $\xi=\log\Delta$, let $p(\theta|\boldsymbol{\lambda}_1)=p_1(\theta)$ and $p(\theta|\boldsymbol{\lambda}_2)=p_2(\theta)$ with $\boldsymbol{\lambda}_1=(0,1)'$ and $\boldsymbol{\lambda}_2=(0,\Delta)'$.

TABLE 1
Comparison of asymptotic relative mean-square errors (I)

Index	Method[1]	$\lim_{n\to\infty}\sqrt{nE(\hat{r}-r)^2/r^2}$
1	Importance sampling-version 2	$\{\exp(D^2)-1\}^{1/2}$
2	Bridge sampling with $\alpha=(p_1p_2)^{-1/2}$ (geometric bridge)	$2\left\{\exp\left(\frac{D^2}{4}\right)-1\right\}^{1/2}$
3	Bridge sampling with optimal bridge α_{opt}	$2\left\{\frac{D\exp(D^2/8)}{\beta(D)\sqrt{2\pi}}-1\right\}^{1/2}$
4	Path sampling with optimal path in μ-space	D
5	Path sampling with optimal path in $(\mu,\sigma)'$-space	$\sqrt{12}\left\{\log\left(\frac{D}{\sqrt{12}}+\sqrt{1+\frac{D^2}{12}}\right)\right\}$
6	Lower bound of path sampling in (2.15)	$\sqrt{8}(1-\exp(-D^2/8))^{1/2}$
7	Ratio importance sampling with optimal π_{opt}	$2(2\Phi(D/2)-1)$

[1]In method 3, $\beta(D)=(1/\pi)\int_0^\infty(\exp(-\theta^2/2D^2)/\cosh(\theta/2))\,d\theta$.

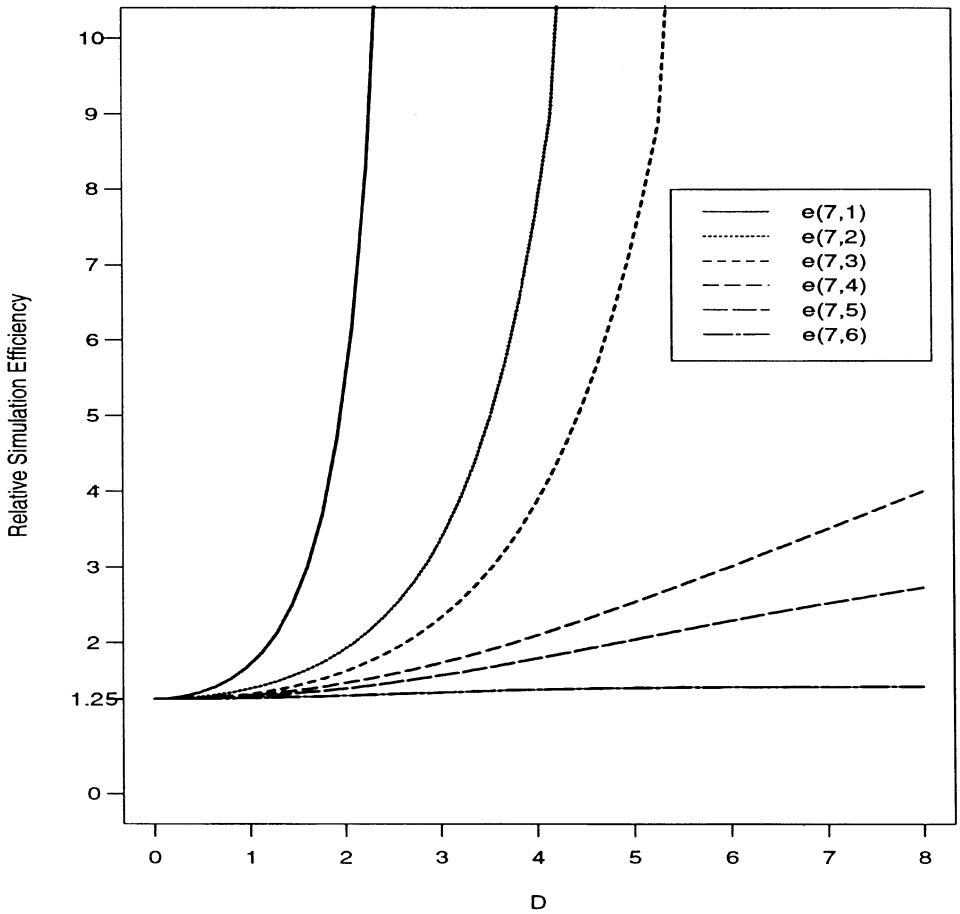

FIG. 1. *Relative simulation efficiency plot* (I).

For importance sampling-version 2, bridge sampling and path sampling, we use the sampling schemes similar to those in Case 1 by using $N(0, \Delta^2)$ to replace $N(D, 1)$. For ratio importance sampling, the optimal density is

$$\pi_{\text{opt}}(\theta) = \frac{|\phi(\theta) - (1/\Delta)\phi(\theta/\Delta)|}{c_{\text{opt}}(\Delta)}, \tag{4.5}$$

where

$$\begin{aligned} c_{\text{opt}}(\Delta) &= \int_{-\infty}^{\infty} \left| \phi(\theta) - \frac{1}{\Delta}\phi\left(\frac{\theta}{\Delta}\right) \right| d\theta \\ &= 4\left[\Phi\left(\sqrt{\frac{2\log\Delta}{1 - 1/\Delta^2}} \right) - \Phi\left(\frac{1}{\Delta}\sqrt{\frac{2\log\Delta}{1 - 1/\Delta^2}} \right) \right]. \end{aligned} \tag{4.6}$$

The corresponding optimal cumulative distribution is

$$\Pi_{\mathrm{opt}}(\theta) = \begin{cases} \left(\Phi\left(\dfrac{\theta}{\Delta}\right) - \Phi(\theta)\right) \Big/ c_{\mathrm{opt}}(\Delta), & \text{for } \theta \le -\sqrt{\dfrac{2\log\Delta}{1-1/\Delta^2}}\,, \\ \dfrac{1}{2} + \left(\Phi(\theta) - \Phi\left(\dfrac{\theta}{\Delta}\right)\right) \Big/ c_{\mathrm{opt}}(\Delta), & \\ \qquad\qquad \text{for } -\sqrt{\dfrac{2\log\Delta}{1-1/\Delta^2}} < \theta \le \sqrt{\dfrac{2\log\Delta}{1-1/\Delta^2}}\,, & \\ 1 - \left(\Phi(\theta) - \Phi\left(\dfrac{\theta}{\Delta}\right)\right) \Big/ c_{\mathrm{opt}}(\Delta), & \text{for } \theta > \sqrt{\dfrac{2\log\Delta}{1-1/\Delta^2}}\,. \end{cases} \tag{4.7}$$

Thus, the inversion method can be employed for generating a random variate Θ from Π_{opt}.

In this case Gelman and Meng (1994) derived the optimal path in $(\mu, \sigma)'$-space; that is, $\mu(t) = 0$ and $\sigma(t) = \Delta^t$ for $0 \le t \le 1$. Then, using (2.8), (2.14), (2.22), (3.10), (4.6) and algebra, we derive the asymptotic relative mean-square errors (variances) for importance sampling, bridge sampling, path sampling, and ratio importance sampling, which are reported in Table 2.

The relative simulation efficiencies defined in (4.4) are calculated and $e(7, j)$, $j = 1, 2, \ldots, 6$, versus Δ are also plotted in Figure 2.

It is easy to verify that $\lim_{\Delta \to 1} e(7, j) = \sqrt{e\pi}/2 = 1.461$ and $e(7, j) > 1$ for all $j = 1, 2, \ldots, 6$. Therefore, the optimal ratio importance sampling method is better than all five counterparts. Once again, the lower bound of path sampling and the asymptotic relative mean-square error of optimal ratio importance sampling are very close. Note that it is not necessary that optimal bridge sampling be better than importance sampling-version 2 due to our sampling scheme. However, it is true that

$$2\left[\frac{\sqrt{2\pi}}{2\displaystyle\int_{-\infty}^{\infty}\left(\exp(\theta^2/2) + \Delta\exp(\theta^2/2\Delta^2)\right)^{-1} d\theta} - 1\right]^{1/2} \le \sqrt{2}\sqrt{\left(\Delta^2/\sqrt{2\Delta^2 - 1}\right) - 1}\,.$$

So, when one density has a heavier tail than another, drawing samples from the heavier tail one is always beneficial. Furthermore, for this case, we can see that even the simple importance sampling method (version 2) is better than the optimal path sampling method. Therefore, path sampling is advantageous only for the cases where the two modes of π_1 and π_2 are far away from each other. Finally, we notice that reverse logistic regression has the

TABLE 2
Comparison of asymptotic relative mean-square errors (II)

Index	Method	$\lim_{n\to\infty}\sqrt{nE(\hat{r}-r)^2/r^2}$
1	Importance sampling-version 2	$\sqrt{\left(\Delta^2/\sqrt{2\Delta^2-1}\right)-1}$
2	Bridge sampling with $\alpha=(p_1p_2)^{-1/2}$	$\dfrac{\sqrt{2}\,(\Delta-1)}{\sqrt{\Delta}}$
3	Bridge sampling with optimal bridge α_{opt}	$2\left[\dfrac{\sqrt{2\pi}}{2\int_{-\infty}^{\infty}\left(\exp(\theta^2/2)+\Delta\exp(\theta^2/2\Delta^2)\right)^{-1}d\theta}-1\right]^{1/2}$
4	Path sampling with optimal path in μ-space	$\sqrt{2}\log\Delta$
5	Path sampling with optimal path in $(\mu,\sigma)'$-space	$\sqrt{2}\log\Delta$
6	Lower bound of path sampling in (2.15)	$2\sqrt{2}\left(1-\sqrt{2\Delta/(1+\Delta^2)}\right)^{1/2}$
7	Ratio importance sampling with optimal π_{opt}	$4\left[\Phi\left(\sqrt{\dfrac{2\log\Delta}{1-1/\Delta^2}}\right)-\Phi\left(\dfrac{1}{\Delta}\sqrt{\dfrac{2\log\Delta}{1-1/\Delta^2}}\right)\right]$

same $\lim_{n\to\infty}\sqrt{nE(\hat{r}-r)^2/r^2}$ as bridge sampling with optimal bridge α_{opt} for both cases.

5. Applications and implementation. In this section we present two interesting applications and discuss general implementation issues for estimating ratios of normalizing constants.

5.1. *Applications*. In Section 1 we discuss many applications of estimating ratios of normalizing constants. Here, we consider two special practical problems.

The first problem arises in Markov chain Monte Carlo (MCMC) sampling from a posterior distribution for a Bayesian hierarchical model with constrained parameter spaces. Let the posterior distribution be of the form

$$\pi(\boldsymbol{\theta},\boldsymbol{\lambda}|x)\propto L(x,\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})/c(\boldsymbol{\lambda}), \tag{5.1}$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ and $\pi(\boldsymbol{\lambda})$ are priors, $c(\boldsymbol{\lambda})=\int_S\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})\,d\boldsymbol{\theta}$, and S is the constrained space for $\boldsymbol{\theta}$. In (5.1), $\boldsymbol{\lambda}$ is a hyper-parameter vector. Notice that $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is a proper probability density function over the whole parameter space. Because analytical evaluation of $c(\boldsymbol{\lambda})$ is often not available, the Gibbs

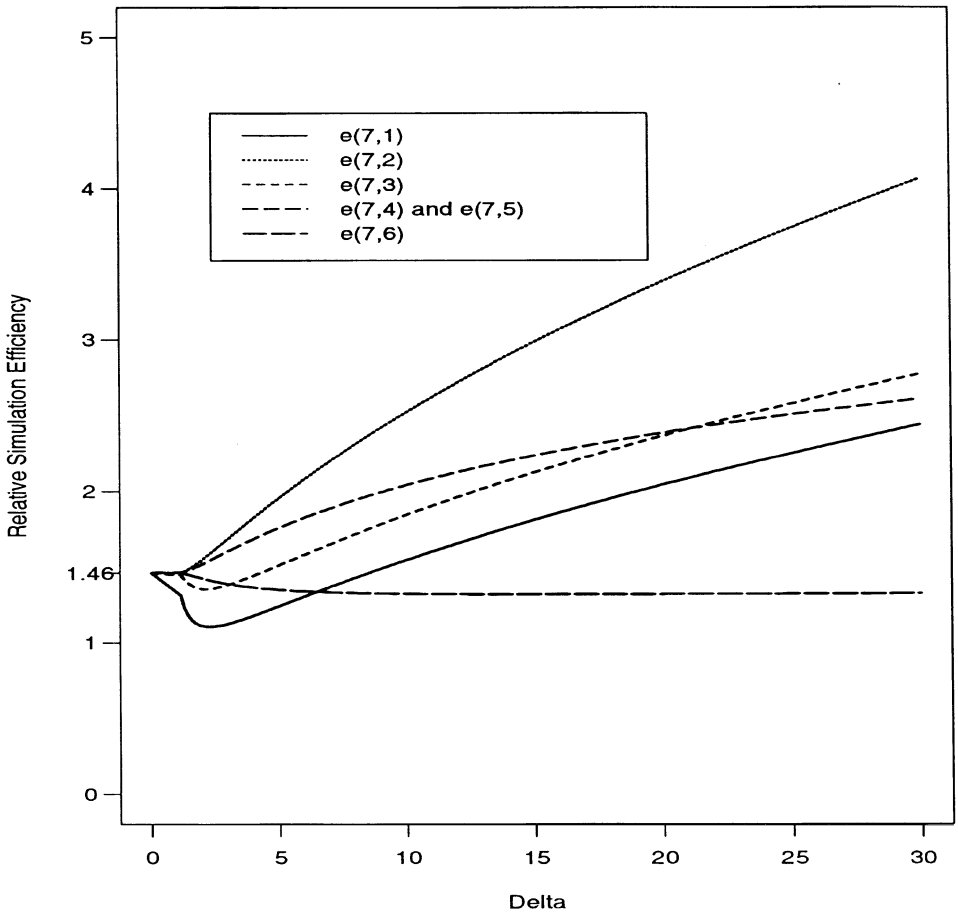

FIG. 2. *Relative simulation efficiency plot (II).*

sampler cannot be used directly for generating $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ from the posterior distribution (5.1). A natural alternative is the Metropolis-within-Gibbs sampler [cf. Müller (1991)]. But this sampling scheme needs to evaluate ratios of normalizing constants for each Metropolis step within each Gibbs iteration.

The second problem arises in estimating marginal Bayesian posterior densities. Chen (1994b) proposed the importance-weighted marginal density estimation (IWMDE). Let $\{(\boldsymbol{\theta}_i, \boldsymbol{\lambda}_i),\ 1 \le i \le m\}$ be a realization of the Markov chain Monte Carlo sample from $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}|\text{data})$ given in (5.1). Then the value of the joint marginal density of $\boldsymbol{\lambda}$ at $\boldsymbol{\lambda}^*$ can be estimated by

$$\hat{\pi}(\boldsymbol{\lambda}^*|\text{data}) = \frac{1}{m}\sum_{i=1}^{m} w(\boldsymbol{\lambda}_i|\boldsymbol{\theta}_i)\frac{\pi(\boldsymbol{\theta}_i|\boldsymbol{\lambda}^*)}{\pi(\boldsymbol{\theta}_i|\boldsymbol{\lambda}_i)}\frac{\pi(\boldsymbol{\lambda}^*)}{\pi(\boldsymbol{\lambda}_i)}\frac{c(\boldsymbol{\lambda}_i)}{c(\boldsymbol{\lambda}^*)}, \tag{5.2}$$

where $w(\boldsymbol{\lambda}|\boldsymbol{\theta})$ is a conditional density given $\boldsymbol{\theta}$. See Chen (1994b) for details about IWMDE. Once again, IWMDE requires evaluating ratios of two normalizing constants $c(\boldsymbol{\lambda}^*)$ and $c(\boldsymbol{\lambda}_i)$ at different grid points $\boldsymbol{\lambda}^*$ and different observed values $\boldsymbol{\lambda}_i$.

The main feature of these two applications is that we need to evaluate many ratios of normalizing constants because in the Metropolis-within-Gibbs sampler such ratios change from one Gibbs iteration to another Gibbs iteration and also because IWMDE needs to be evaluated at many points $\boldsymbol{\lambda}^*$ for each $\boldsymbol{\lambda}_i$ in (5.2). Therefore, it is very expensive to employ bridge sampling or path sampling for such practical problems. However, ratio importance sampling is very useful. We can choose a fairly good ratio importance sampling density, π_g, and then generate samples from π_g for estimating all these ratios. Thus, the ratio importance sampling method greatly eases the computational burden. However, a globally good π_g may not exist or may be difficult to obtain. Therefore, an adaptive scheme is often required. Chen (1994a) developed an adaptive scheme for Gibbs sampling, and such an adaptive scheme is also applicable for obtaining IWMDE. See Chen (1994a) for details.

5.2. *Implementation.* In this subsection, we present the exact and approximate optimal schemes for obtaining the optimal ratio importance sampling estimators. We also present other "nonoptimal" implementation schemes.

EXACT OPTIMAL SCHEME. In Sections 3 and 4, we showed that ratio importance sampling is better than importance sampling-version 2, bridge sampling, and path sampling. As we pointed out in Section 3, the optimal ratio importance sampling estimator $\hat{r}_{\pi_{\text{opt}}}$ is not directly usable. However, the following two-stage sampling scheme is a practical method leading essentially to $\hat{r}_{\pi_{\text{opt}}}$.

Let $\pi(\boldsymbol{\theta})$ be an arbitrary density over Ω such that $\pi(\boldsymbol{\theta}) > 0$ for $\boldsymbol{\theta} \in \Omega$. Given a random draw $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ from π, define

$$\tau_n = \frac{\sum_{i=1}^n p_1(\boldsymbol{\theta}_i)/\pi(\boldsymbol{\theta}_i)}{\sum_{i=1}^n p_2(\boldsymbol{\theta}_i)/\pi(\boldsymbol{\theta}_i)}. \tag{5.3}$$

Also let

$$\psi_n(\boldsymbol{\theta}) = \frac{|p_1(\boldsymbol{\theta}) - \tau_n p_2(\boldsymbol{\theta})|}{\int_\Omega |p_1(\boldsymbol{\theta}') - \tau_n p_2(\boldsymbol{\theta}')|\, d\boldsymbol{\theta}'}. \tag{5.4}$$

Then, make a random draw $\vartheta_{n,1}, \vartheta_{n,2}, \ldots, \vartheta_{n,n}$ from ψ_n and define the "optimal" estimator $\hat{r}_n$ as follows:

$$\hat{r}_n = \frac{\sum_{i=1}^n p_1(\boldsymbol{\vartheta}_{n,i})/\psi_n(\boldsymbol{\vartheta}_{n,i})}{\sum_{i=1}^n p_2(\boldsymbol{\vartheta}_{n,i})/\psi_n(\boldsymbol{\vartheta}_{n,i})}. \tag{5.5}$$

THEOREM 5.1. *Suppose that there is a neighborhood U_r of r such that the following conditions are satisfied*:

(i) $$\inf_{a \in U_r} \int_\Omega |p_1(\boldsymbol{\theta}) - a p_2(\boldsymbol{\theta})|\, d\boldsymbol{\theta} > 0;$$

(ii) $$\int_\Omega \sup_{a \in \Omega_r} \frac{p_1^2(\boldsymbol{\theta}) + p_2^2(\boldsymbol{\theta})}{|p_1(\boldsymbol{\theta}) - a p_2(\boldsymbol{\theta})|}\, d\boldsymbol{\theta} < \infty;$$

(iii) $$\sup_{a \in U_r} \int_\Omega \frac{p_1^2(\boldsymbol{\theta})|p_1(\boldsymbol{\theta}) - a p_2(\boldsymbol{\theta})|}{p_2^2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} < \infty.$$

Then we have

$$\lim_{n \to \infty} nE\left(\frac{(\hat{r}_n - r)^2}{r^2} \middle| \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \theta_n \right) = \left[\int_\Omega \left| \frac{p_1(\boldsymbol{\theta})}{c_1} - \frac{p_2(\boldsymbol{\theta})}{c_2} \right| d\boldsymbol{\theta} \right]^2 \quad a.s. \tag{5.6}$$

The proof is given in Appendix B.

REMARK 5.1. Theorem 5.1 says that the "optimal" estimator $\hat{r}_n$ obtained by the two-stage sampling scheme has the same optimal relative mean-square error as $\hat{r}_{\pi_{\text{opt}}}$. Note that sampling from ψ_n can be facilitated by the Metropolis–Hastings algorithm or the Gibbs sampler.

REMARK 5.2. In the two-stage sampling scheme, sample sizes in stage 1 and stage 2 need not be the same. More specifically, we can use n_1 in (5.3) and (5.4) (the first-stage sample size) and n_2 in (5.5) (the second-stage sample size). Then (5.6) still holds as long as $n_1 = O(n)$ and $n_1 \to \infty$, where $n = n_1 + n_2$.

APPROXIMATE OPTIMAL SCHEME. Let $\pi_i^I(\boldsymbol{\theta})$, $i = 1, 2$, be good importance sampling densities for $\pi_i(\boldsymbol{\theta})$, $i = 1, 2$, respectively. Then, the optimal ratio importance sampling density, π_{opt}, can be approximated by

$$\pi_{\text{opt}}^I(\boldsymbol{\theta}) \propto |\pi_1^I(\boldsymbol{\theta}) - \pi_2^I(\boldsymbol{\theta})|. \tag{5.7}$$

When the $\pi_i^I(\boldsymbol{\theta})$ are normal importance sampling densities, sampling from (5.7) can proceed in a way similar to the one we used for sampling from (4.3) and (4.7). Let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ be a random draw from π_{opt}^I. Then an approximate optimal ratio importance sampling estimator is given by

$$\hat{r}_{\pi_{\text{opt}}^I} = \frac{\sum_{i=1}^n p_1(\boldsymbol{\theta}_i)/|\pi_1^I(\boldsymbol{\theta}_i) - \pi_2^I(\boldsymbol{\theta}_i)|}{\sum_{i=1}^n p_2(\boldsymbol{\theta}_i)/|\pi_1^I(\boldsymbol{\theta}_i) - \pi_2^I(\boldsymbol{\theta}_i)|}. \tag{5.8}$$

Note that when π_1 and π_2 do not overlap, we can choose $\pi^I_{\text{opt}}(\boldsymbol{\theta}) = \{\pi^I_1(\boldsymbol{\theta}) + \pi^I_2(\boldsymbol{\theta})\}/2$ because $\pi_{\text{opt}}(\boldsymbol{\theta}) = \{\pi_1(\boldsymbol{\theta}) + \pi_2(\boldsymbol{\theta})\}/2$. For such cases, sampling from π^I_{opt} is straightforward.

OTHER "NONOPTIMAL" SCHEMES. First, assume that π_1 and π_2 do not overlap, that is, $\int_\Omega p_1(\boldsymbol{\theta})p_2(\boldsymbol{\theta})\,d\boldsymbol{\theta} = 0$. For this case, the IWMDE method of Chen (1994b) would give a reasonably good estimator of r. Let $w_i(\boldsymbol{\theta})$ be a weighted density, which roughly mimics p_i, for $i = 1, 2$. An empirical procedure for obtaining such a w is given in Chen (1994b). Also let $\{\boldsymbol{\theta}_{ij},\ j = 1, 2, \ldots, n_i\}$, $i = 1, 2$, be independent random draws from π_i, $i = 1, 2$, respectively. Then, a consistent estimator of r is

$$\hat{r}_{\text{IWMDE}} = \frac{n_2^{-1}\sum_{j=1}^{n_2} w_2(\boldsymbol{\theta}_{2j})/p_2(\boldsymbol{\theta}_{2j})}{n_1^{-1}\sum_{j=1}^{n_1} w_1(\boldsymbol{\theta}_{1j})/p_1(\boldsymbol{\theta}_{1j})}. \tag{5.9}$$

In this case, path sampling is also useful (if it is applicable). See Gelman and Meng (1996) for the implementation details.

Second, assume that $\int_\Omega p_1(\boldsymbol{\theta})p_2(\boldsymbol{\theta})\,d\boldsymbol{\theta} > 0$; that is, π_1 and π_2 do overlap. We propose a bridge sampling type estimator as follows. Let $\{\boldsymbol{\theta}_i,\ i = 1, 2, \ldots, n\}$ be a random draw from a mixture density:

$$\pi_{\text{mix}}(\boldsymbol{\theta}) = \psi\pi_1(\boldsymbol{\theta}) + (1 - \psi)\pi_2(\boldsymbol{\theta}), \tag{5.10}$$

where $0 < \psi < 1$ is known (for example, $\psi = 1/2$). Note that we can easily sample from $\pi_{\text{mix}}(\boldsymbol{\theta})$ by a composition method without knowing c_1 and c_2. Let

$$\begin{aligned} S_n(r) = &\sum_{i=1}^{n} \frac{rp_2(\boldsymbol{\theta}_i)}{\psi p_1(\boldsymbol{\theta}_i) + r(1 - \psi)p_2(\boldsymbol{\theta}_i)} \\ &- \sum_{i=1}^{n} \frac{p_1(\boldsymbol{\theta}_i)}{\psi p_1(\boldsymbol{\theta}_i) + r(1 - \psi)p_2(\boldsymbol{\theta}_i)}. \end{aligned} \tag{5.11}$$

Then, a bridge sampling type estimator $\hat{r}_{b,n}$ of r is the solution of the following equation:

$$\begin{aligned} &\sum_{i=1}^{n} \frac{rp_2(\boldsymbol{\theta}_i)}{\psi p_1(\boldsymbol{\theta}_i) + r(1 - \psi)p_2(\boldsymbol{\theta}_i)} \\ &\quad - \sum_{i=1}^{n} \frac{p_1(\boldsymbol{\theta}_i)}{\psi p_1(\boldsymbol{\theta}_i) + r(1 - \psi)p_2(\boldsymbol{\theta}_i)} = 0. \end{aligned} \tag{5.12}$$

Since $S_n(0) = -n/\psi < 0$, $S_n(\infty) = n/(1 - \psi) > 0$, and

$$\frac{dS_n(r)}{dr} = \sum_{i=1}^{n} \frac{p_1(\boldsymbol{\theta}_i)p_2(\boldsymbol{\theta}_i)}{\{\psi p_1(\boldsymbol{\theta}_i) + r(1 - \psi)p_2(\boldsymbol{\theta}_i)\}^2} > 0, \tag{5.13}$$

there exists a unique solution to the equation (5.12). The solution $\hat{r}_{b,n}$, can be easily obtained by, for example, the bisection method. The asymptotic properties of $\hat{r}_{b,n}$ are given in the next theorem.

THEOREM 5.2. *Suppose that* $\int_\Omega p_1(\boldsymbol{\theta})p_2(\boldsymbol{\theta})\,d\boldsymbol{\theta} > 0$. *Then, we have*

$$\hat{r}_n \to r \quad a.s.\ as\ n \to \infty. \tag{5.14}$$

If, in addition, $E_{\pi_{\text{mix}}}(p_1(\boldsymbol{\Theta})/p_2(\boldsymbol{\Theta}))^2 < \infty$, *then*

$$\lim_{n\to\infty} nE_{\pi_{\text{mix}}}\frac{(\hat{r}_n - r)^2}{r^2} = \int_\Omega \frac{(\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta}))^2}{\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta})}\,d\boldsymbol{\theta} \times \left\{\int_\Omega \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta})}\,d\boldsymbol{\theta}\right\}^{-2}. \tag{5.15}$$

The proof is given in Appendix C.

6. Concluding remarks. In this article, we present an overview of the current Monte Carlo methods for estimating ratios of normalizing constants and we also discuss the relationships among those methods. Then, we propose a ratio importance sampling method and show that ratio importance sampling is better than simple importance sampling, bridge sampling, reverse logistic regression, and path sampling. Finally, we provide practical, easy-to-use implementation schemes. Although all our inferences are based on independent draws, under certain mild conditions (e.g., ergodicity), the consistency for all estimators still holds. Indeed, the "independent" samples are still available even if we use the Gibbs sampler or the Metropolis–Hastings algorithm. Such "independent" observations can be obtained by taking every Bth Gibbs (or Metropolis–Hastings) iterate and B is chosen so that the autocorrelations among those observations disappear, which can be assessed by using, for example, the IMSL subroutine, ACF.

APPENDIX A

Proof of Theorem 3.1. Write

$$\sqrt{n}\,(\hat{r}_\pi - r) = \frac{c_1 n^{-1/2}\sum_{i=1}^n \{f(\boldsymbol{\Theta}_i)/c_1 - g(\boldsymbol{\Theta}_i)/c_2\}}{(1/n)\sum_{i=1}^n g(\boldsymbol{\Theta}_i)}. \tag{A.1}$$

From the central limit theorem it follows that

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^{n}\left\{\frac{f(\boldsymbol{\Theta}_i)}{c_1}-\frac{g(\boldsymbol{\Theta}_i)}{c_2}\right\} \\ & \quad \to_{\mathscr{D}} N\left(0, E_\pi\left\{\frac{f(\boldsymbol{\Theta})}{c_1}-\frac{g(\boldsymbol{\Theta})}{c_2}\right\}^2\right) \quad \text{as } n\to\infty \end{aligned} \tag{A.2}$$

and by the law of large numbers,

$$(1/n)\sum_{i=1}^{n} g(\boldsymbol{\Theta}_i) \to c_2 \quad \text{a.s. as } n\to\infty. \tag{A.3}$$

Then (3.5) follows from (A.2) and (A.3). To prove (3.4), it suffices to show that $\{n(\hat{r}_\pi - r)^2,\ n\ge 1\}$ is uniformly integrable. In this case, by (3.5), we shall have $E_\pi\{\sqrt{n}\,(\hat{r}_\pi - r)\} = o(1)$ as $n\to\infty$. Thus

$$\frac{1}{r^2}E_\pi\left\{n\left(\hat{r}_\pi - r\right)^2\right\} \to E_\pi\left\{\frac{f(\boldsymbol{\Theta})}{c_1}-\frac{g(\boldsymbol{\Theta})}{c_2}\right\}^2 \quad \text{as } n\to\infty,$$

which gives (3.4). We show below the uniform integrability of $\{n(\hat{r}_\pi - r)^2,\ n\ge 1\}$. Rewrite

$$\sqrt{n}\,(\hat{r}_\pi - r) = \frac{n^{-1/2}\sum_{i=1}^{n}\{c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i)\}}{c_2(1/n)\sum_{i=1}^{n} g(\boldsymbol{\Theta}_i)} \tag{A.4}$$

and let $U_n = n^{-1/2}\sum_{i=1}^{n}\{c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i)\}$ and $V_n = n^{-1}\sum_{i=1}^{n} g(\boldsymbol{\Theta}_i)$. By (A.4), for every $A\ge 2$,

$$\begin{aligned} & E_\pi\left[n\left(\hat{r}_\pi - r\right)^2 I_{\{n(\hat{r}_\pi - r)^2\ge A^2\}}\right] \\ & \quad = E_\pi\left[\frac{U_n^2}{c_2^2 V_n^2} I_{\{|U_n|\ge c_2 A V_n\}}\right] \\ & \quad = E_\pi\left[\frac{U_n^2}{c_2^2 V_n^2} I_{\{|U_n|\ge Ac_2V_n,\ V_n\ge c_2/2\}}\right] + E_\pi\left[\frac{U_n^2}{c_2 V_n^2} I_{\{|U_n|\ge Ac_2V_n,\ V_n< c_2/2\}}\right] \\ & \quad \le 4c_2^{-4}E_\pi\left[U_n^2 I_{\{|U_n|\ge Ac_2^2/2\}}\right] + E_\pi\left[n\left(\hat{r}_\pi - r\right)^2 I_{\{V_n< c_2/2\}}\right], \end{aligned} \tag{A.5}$$

where $I_{\{n(\hat{r}_\pi - r)^2\ge A^2\}}$ is an indicator function. It is well known that $\{U_n^2,\ n\ge 1\}$ is uniformly integrable. Hence

$$\lim_{A\to\infty} E_\pi\left[U_n^2 I_{\{|U_n|\ge Ac_2^2/2\}}\right] = 0. \tag{A.6}$$

Noting that $\hat{r}_\pi \le \sum_{i=1}^n f(\Theta_i)/g(\Theta_i)$, we have

$$
\begin{aligned}
&E_\pi\left[n(\hat{r}_\pi - r)^2 I_{\{V_n < c_2/2\}}\right] \\
&\quad \le nE_\pi\left[(\hat{r}_\pi^2 + r^2) I_{\{V_n < c_2/2\}}\right] \\
&\quad \le nE_\pi\left[\left\{r^2 + n\sum_{i=1}^n (f(\Theta_i)/g(\Theta_i))^2\right\} I_{\{V_n < c_2/2\}}\right] \\
&\quad \le n\left[r^2 P_\pi(V_n < c_2/2) + n\sum_{i=1}^n E_\pi\left\{(f(\Theta_i)/g(\Theta_i))^2 I_{\{\sum_{j\ne i} g(\Theta_j) < nc_2/2\}}\right\}\right] \\
&\quad = n\left[r^2 P_\pi(V_n < c_2/2) + n^2 E_\pi(f(\Theta)/g(\Theta))^2 P_\pi\left(\sum_{j=1}^{n-1} g(\Theta_j) < nc_2/2\right)\right],
\end{aligned}
\tag{A.7}
$$

where P_π is the probability measure with respect to π. Using the Chebyshev inequality, we get

$$
\begin{aligned}
P_\pi(V_n < c_2/2) &= P_\pi\left(\sum_{i=1}^n \{E_\pi g(\Theta_i) - g(\Theta_i)\} > nc_2/2\right) \\
&\le \inf_{t\ge 0} \exp(-tc_2 n/2) E_\pi\left[\exp\left(\sum_{i=1}^n \{E_\pi g(\Theta_i) - g(\Theta_i)\}\right)\right] \\
&= \left(\inf_{t\ge 0} \exp(-tc_2/2) E_\pi \exp\big(t(c_2 - g(\Theta))\big)\right)^n.
\end{aligned}
\tag{A.8}
$$

From $E_\pi(c_2 - g(\Theta)) = 0$, it follows that

$$
\varepsilon = \inf_{t\ge 0} \exp(-tc_2/4) E_\pi \exp\big(t(c_2 - g(\Theta))\big) < 1.
$$

Thus, $P_\pi(V_n < c_2/2) \le \varepsilon^n$. Similarly, for $n \ge 3$, we have

$$
\begin{aligned}
&P_\pi\left(\sum_{j=1}^{n-1} g(\Theta_j) < nc_2/2\right) \\
&\quad = P_\pi\left(\sum_{j=1}^{n-1} \{E_\pi g(\Theta_i) - g(\Theta_i)\} > (n-2)c_2/2\right) \\
&\quad \le \left(\inf_{t\ge 0} \exp(-(n-2)tc_2/2(n-1)) E_\pi \exp\big(t(c_2 - g(\Theta))\big)\right)^{n-1} \\
&\quad \le \varepsilon^{n-1}.
\end{aligned}
\tag{A.9}
$$

Putting together the above inequalities yields

$$
E_\pi\left[n(\hat{r}_\pi - r)^2 I_{\{V_n < c_2/2\}}\right] = O(n^3\varepsilon^n) = o(1). \tag{A.10}
$$

Therefore, (3.4) follows from (A.5), (A.6) and (A.10).

Next we prove (3.6). We have

$$
\begin{aligned}
&nE_\pi(\hat{r}_\pi - r)^2 - c_2^{-4}E_\pi\{c_2 f(\boldsymbol{\Theta}) - c_1 g(\boldsymbol{\Theta})\}^2 \\
&\quad = c_2^{-2} n \Bigg[E_\pi \left\{ \frac{\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))}{\sum_{i=1}^n g(\boldsymbol{\Theta}_i)} \right\}^2 \\
&\qquad\qquad - E_\pi \left\{ \frac{\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))}{nc_2} \right\}^2 \Bigg] \\
&\quad = \frac{c_2^{-4}}{n} \Bigg[E_\pi \Bigg\{ \Bigg(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i)) \Bigg)^2 \sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i)) \\
&\qquad\qquad \times \sum_{i=1}^n (c_2 + g(\boldsymbol{\Theta}_i)) \Bigg(\Bigg(\sum_{i=1}^n g(\boldsymbol{\Theta}_i) \Bigg)^2 \Bigg)^{-1} \Bigg\} \Bigg] \qquad \text{(A.11)} \\
&\quad =_{\text{def}} \frac{c_2^{-4}}{n} \varepsilon_n
\end{aligned}
$$

and

$$
\begin{aligned}
\varepsilon_n &= E_\pi \left\{ \frac{\left(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))\right)^2 \sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i)) 2nc_2}{\left(\sum_{i=1}^n g(\boldsymbol{\Theta}_i)\right)^2} \right\} \\
&\quad - E_\pi \left\{ \frac{\left(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))\right)^2 \left(\sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i))\right)^2}{\left(\sum_{i=1}^n g(\boldsymbol{\Theta}_i)\right)^2} \right\} \\
&= 2E_\pi \left\{ \frac{\left(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))\right)^2 \sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i))}{(nc_2)} \right\} \\
&\quad + 2E_\pi \Bigg\{ \Bigg(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i)) \Bigg)^2 \sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i)) \\
&\qquad \times \Bigg((nc_2)^2 - \Bigg(\sum_{i=1}^n g(\boldsymbol{\Theta}_i) \Bigg)^2 \Bigg) \Bigg(nc_2 \Bigg(\sum_{i=1}^n g(\boldsymbol{\Theta}_i) \Bigg)^2 \Bigg)^{-1} \Bigg\} \qquad \text{(A.12)} \\
&\quad - E_\pi \left\{ \frac{\left(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))\right)^2 \left(\sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i))\right)^2}{\left(\sum_{i=1}^n g(\boldsymbol{\Theta}_i)\right)^2} \right\} \\
&=_{\text{def}} \varepsilon_{n,1} + \varepsilon_{n,2} + \varepsilon_{n,3}.
\end{aligned}
$$

It is easy to see that

$$\begin{aligned}\varepsilon_{n,1} &= 2(nc_2)^{-1}E_\pi\Bigg\{\Bigg(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))^2 \\ &\qquad + 2\sum_{1\le i<j\le n} (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))(c_2 f(\boldsymbol{\Theta}_j) - c_1 g(\boldsymbol{\Theta}_j))\Bigg) \\ &\qquad \times \sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i))\Bigg\} \\ &= 2(nc_2)^{-1}E_\pi\Bigg\{\Bigg(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))^2\Bigg)\sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i))\Bigg\} \\ &= 2(nc_2)^{-1}E_\pi\Bigg\{\Bigg(\sum_{i=1}^n \big\{(c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))^2 \\ &\qquad - E_\pi(c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))^2\big\}\Bigg)\sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i))\Bigg\} \qquad \text{(A.13)} \\ &= (nc_2)^{-1}O\Bigg(\Bigg[E_\pi\Bigg\{\Bigg(\sum_{i=1}^n \big\{(c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))^2 \\ &\qquad - E_\pi(c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))^2\big\}\Bigg)^2 \\ &\qquad \times E_\pi\Bigg\{\sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i))\Bigg\}^2\Bigg]^{1/2}\Bigg) \\ &= O(1).\end{aligned}$$

For $\varepsilon_{n,2}$, we have

$$\begin{aligned}|\varepsilon_{n,2}| &= 2\Bigg|E_\pi\Bigg\{\Bigg(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))\Bigg)^2\Bigg(\sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i))\Bigg)^2 \\ &\qquad \times\Bigg(nc_2 + \sum_{i=1}^n g(\boldsymbol{\Theta}_i)\Bigg)\Bigg(nc_2\Bigg(\sum_{i=1}^n g(\boldsymbol{\Theta}_i)\Bigg)^2\Bigg)^{-1}\Bigg\}\Bigg| \\ &\le 12(nc_2)^{-2}E_\pi\Bigg\{\Bigg(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))\Bigg)^2\Bigg(\sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i))\Bigg)^2\Bigg\} \\ &\quad + 2\Bigg|E_\pi\Bigg\{\Bigg(\sum_{i=1}^n (c_2 f(\boldsymbol{\Theta}_i) - c_1 g(\boldsymbol{\Theta}_i))\Bigg)^2\Bigg(\sum_{i=1}^n (c_2 - g(\boldsymbol{\Theta}_i))\Bigg)^2 \qquad \text{(A.14)}\end{aligned}$$

$$\times\left(nc_2+\sum_{i=1}^{n}g(\boldsymbol{\Theta}_i)\right)\left(nc_2\left(\sum_{i=1}^{n}g(\boldsymbol{\Theta}_i)\right)^2\right)^{-1}I_{\{V_n<c_2/2\}}\Bigg\}\Bigg|$$

$$\le 12(nc_2)^{-2}\Bigg[E_\pi\left\{\sum_{i=1}^{n}(c_2f(\boldsymbol{\Theta}_i)-c_1g(\boldsymbol{\Theta}_i))\right\}^4$$

$$\times E_\pi\left\{\sum_{i=1}^{n}(c_2-g(\boldsymbol{\Theta}_i))\right\}^4\Bigg]^{1/2}$$

$$+4(nc_2)^3E_\pi\left\{\left(c_1+c_2\sum_{i=1}^{n}f(\boldsymbol{\Theta}_i)/g(\boldsymbol{\Theta}_i)\right)^2I_{\{V_n<c_2/2\}}\right\}$$

$$=O(1)+O(n^5\varepsilon^n)=O(1),$$

where the last inequality is from (A.8) and the proof of (A.7). Similarly, we have

$$\varepsilon_{n,3}=O(1). \tag{A.15}$$

Thus, (3.6) follows from the above inequalities. □

APPENDIX B

Proof of Theorem 5.1. Write $f_n(\boldsymbol{\theta})=p_1(\boldsymbol{\theta})/\psi_n(\boldsymbol{\theta})$ and $g_n(\boldsymbol{\theta})=p_2(\boldsymbol{\theta})/\psi_n(\boldsymbol{\theta})$. By (A.11), we have

$$nE\left(\frac{(\hat r_n-r)^2}{r^2}\middle|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_n\right)-r^{-2}c_2^{-4}\int_\Omega\{c_2f_n(\boldsymbol{\theta})-c_1g_n(\boldsymbol{\theta})\}^2\psi_n(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

$$=c_2^{-4}r^{-2}n^{-1}E\Bigg\{\left(\sum_{i=1}^{n}(c_2f_n(\boldsymbol{\vartheta}_{n,i})-c_1g_n(\boldsymbol{\vartheta}_{n,i}))\right)^2$$

$$\times\sum_{i=1}^{n}(c_2-g_n(\boldsymbol{\vartheta}_{n,i}))\sum_{i=1}^{n}(c_2+g_n(\boldsymbol{\vartheta}_{n,i}))$$

$$\times\left(\left(\sum_{i=1}^{n}g_n(\boldsymbol{\vartheta}_{n,i})\right)^2\right)^{-1}\Bigg|\tau_n\Bigg\} \tag{B.1}$$

$$=_{\text{def}}c_2^{-4}r^{-2}\eta_n.$$

From the law of large numbers it follows that

$$\tau_n\to r\quad\text{a.s. as } n\to\infty. \tag{B.2}$$

Therefore

$$\lim_{n\to\infty}r^{-2}c_2^{-4}\int_\Omega\{c_2f_n(\boldsymbol{\theta})-c_1g_n(\boldsymbol{\theta})\}^2\psi_n(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

$$=\left[\int_\Omega\left|\frac{p_1(\boldsymbol{\theta})}{c_1}-\frac{p_2(\boldsymbol{\theta})}{c_2}\right|d\boldsymbol{\theta}\right]^2\quad\text{a.s.} \tag{B.3}$$

To complete the proof of the theorem, it suffices to show that

$$\eta_n \to 0 \quad \text{a.s. as } n \to \infty. \tag{B.4}$$

Let

$$G_n = \sum_{i=1}^{n} g_n(\vartheta_{n,i}) \quad \text{and} \quad T_n = \sum_{i=1}^{n} (c_2 f_n(\vartheta_{n,i}) - c_1 g_n(\vartheta_{n,i})).$$

Note that

$$\begin{aligned}
|\eta_n| &= \left| E\left\{ \frac{T_n^2(nc_2 - G_n)(nc_2 + G_n)}{nG_n^2} \tau_n \right\} \right| \\
&\le 6n^{-1} E\{T_n^2 I_{\{|T_n| \ge n^{2/3}\}} | \tau_n\} \\
&\quad + 6(nc_2)^{-1} n^{-1} E\{T_n^2 |nc_2 - G_n| I_{\{G_n \ge nc_2/2\}} I_{\{|T_n| \ge n^{2/3}\}} | \tau_n\} \\
&\quad + 2(nc_2)^2 n^{-1} E\{(T_n/G_n)^2 I_{\{G_n \le nc_2/2\}} | \tau_n\} \\
&\le n^{-1} E\{T_n^2 I_{\{|T_n| \ge n^{2/3}\}} | \tau_n\} + 6c_2^{-1} n^{-2/3} E\{|nc_2 - G_n| \, | \tau_n\} \\
&\quad + 2(nc_2)^2 n^{-1} E\{(T_n/G_n)^2 I_{\{G_n \le nc_2/2\}} | \tau_n\} =_{\text{def}} \eta_{n,1} + \eta_{n,2} + \eta_{n,3}.
\end{aligned} \tag{B.5}$$

Since T_n is a partial sum of i.i.d. random variables under the given τ_n, by (B.2) and (ii), we have

$$\begin{aligned}
\eta_{n,1} &\le K(n^{-1/15}) \\
&\quad + E\left\{ (c_2 f_n(\vartheta_{n,1}) - c_1 g_n(\vartheta_{n,1}))^2 I_{\{|c_2 f_n(\vartheta_{n,1}) - c_1 g_n(\vartheta_{n,1})| \ge n^{1/15}\}} \Big| \tau_n \right\} \Big) \\
&\le K\left(n^{-1/15} + \int_{\{\boldsymbol{\theta}:\, |c_2 p_1(\boldsymbol{\theta}) - c_1 p_2(\boldsymbol{\theta})| \ge n^{1/15} \psi_n(\theta)\}} \frac{|c_2 p_1(\boldsymbol{\theta}) - c_1 p_2(\boldsymbol{\theta})|^2}{\psi_n(\boldsymbol{\theta})} \, d\boldsymbol{\theta} \right) \\
&\to 0 \quad \text{a.s. as } n \to \infty,
\end{aligned} \tag{B.6}$$

where K denotes a positive constant not depending on n. Similarly, one has

$$\lim_{n \to \infty} \eta_{n,2} = 0 \quad \text{a.s.} \tag{B.7}$$

Note that for any positive random variable X with $EX = \mu$ and $EX^2 = \sigma^2$ and for any $0 < t < 1$,

$$\begin{aligned}
&E \exp(t(\mu - X)) \\
&\quad \le E\left\{ 1 + t(\mu - X) + (t(\mu - X))^2/2 + \sum_{k=3}^{\infty} \frac{(t(\mu - X))^k}{k!} I_{\{\mu - X \ge 0\}} \right\} \\
&\quad \le 1 + t^2 EX^2 + (\mu t)^3 \exp(t\mu) \le \exp\{t^2(EX^2 + \exp(4\mu))\}.
\end{aligned}$$

Hence, for $0 < a < EX^2 + e^{4\mu}$,

$$\inf_{t > 0} \exp(-ta) E \exp(t(\mu - X)) \le \exp\left(- \frac{a^2}{4(EX^2 + \exp(4\mu))} \right). \tag{B.8}$$

By (B.8) and similarly to (A.9), we have

$$
\begin{aligned}
&P\left(\sum_{j=1}^{n-1} g_n(\vartheta_{n,j}) \le \frac{nc_2}{2}\middle|\tau_n\right)\\
&\le \left(\inf_{t>0}\exp(-tc_2/4)E\{\exp(c_2 - g_n(\vartheta_{n,1}))|\tau_n\}\right)^{n-1}\\
&\le \exp\left(-\frac{(n-1)c_2^2}{64(\exp(4c_2) + E\{g_n^2(\vartheta_{n,1})|\tau_n\})}\right).
\end{aligned}
\tag{B.9}
$$

Thus, in terms of (B.2) and the conditions (ii) and (iii)

$$
\begin{aligned}
\limsup_{n\to\infty}\eta_{n,3} &\le K\limsup_{n\to\infty} n^3 E\{(f_n(\vartheta_{n,1})/g_n(\vartheta_{n,1}))^2|\tau_n\}\\
&\quad\times\exp\left(-\frac{(n-1)c_2^2}{64(\exp(4c_2) + E\{g_n^2(\vartheta_{n,1})|\tau_n\})}\right) = 0 \quad \text{a.s.}
\end{aligned}
\tag{B.10}
$$

Putting the above inequalities together yields (B.4). □

APPENDIX C

Proof of Theorem 5.2. Let

$$\zeta(x,t) = \frac{p_1(x)}{\psi p_1(x) + (1-\psi)tp_2(x)}.$$

Since $S_n(\hat{r}_n) = 0$, we have

$$\sum_{i=1}^{n}\zeta(\boldsymbol{\theta}_i, \hat{r}_n) = n. \tag{C.1}$$

Note that for each fixed x, $\zeta(x,\cdot)$ is decreasing. Hence, $\forall\ x > 0$,

$$\{\hat{r}_n \ge x\} = \left\{\sum_{i=1}^{n}\zeta(\boldsymbol{\theta}_i, x) \ge n\right\}. \tag{C.2}$$

In particular, $\forall\ 0 < \varepsilon < r$,

$$P(\hat{r}_n \ge r+\varepsilon, i.o.) = P\left(\sum_{i=1}^{n}\zeta(\boldsymbol{\theta}_i, r+\varepsilon) \ge n, i.o.\right) \tag{C.3}$$

and

$$P(\hat{r}_n \le r-\varepsilon, i.o.) = P\left(\sum_{i=1}^{n}\zeta(\boldsymbol{\theta}_i, r-\varepsilon) \le n, i.o.\right). \tag{C.4}$$

Noting that for $x > 0$,

$$
\begin{aligned}
&E_{\pi_{\text{mix}}}\zeta(\boldsymbol{\Theta}, x)\\
&= \int_{\Omega}\frac{p_1(\boldsymbol{\theta})(\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta}))}{\psi p_1(\boldsymbol{\theta}) + (1-\psi)xp_2(\boldsymbol{\theta})}\,d\boldsymbol{\theta}\begin{cases}<1, & \text{if } x > r,\\ =1, & \text{if } x = r,\\ >1, & \text{if } x < r,\end{cases}
\end{aligned}
\tag{C.5}
$$

and by the strong law of large numbers, we have

$$P\left(\sum_{i=1}^{n} \zeta(\boldsymbol{\theta}_i, r+\varepsilon) \geq n, \, i.o.\right) = 0$$

and

$$P\left(\sum_{i=1}^{n} \zeta(\boldsymbol{\theta}_i, r-\varepsilon) \leq n, \, i.o.\right) = 0.$$

This proves (5.14).

Write $\lambda(x) = E_{\pi_{\text{mix}}}(\zeta(\boldsymbol{\Theta}, x) - 1)$. Then, by (C.5), $\lambda(r) = 0$ and

$$\begin{aligned} \dot{\lambda}(x) &=_{\text{def}} \frac{d\lambda(x)}{dx} \\ &= -(1-\psi)\int_{\Omega} \frac{p_1(\boldsymbol{\theta})p_2(\boldsymbol{\theta})(\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta}))}{(\psi p_1(\boldsymbol{\theta}) + (1-\psi)xp_2(\boldsymbol{\theta}))^2}\, d\boldsymbol{\theta}. \end{aligned} \tag{C.6}$$

In particular,

$$\dot{\lambda}(r) = -(1-\psi)\left(\frac{c_2}{c_1}\right)\int_{\Omega} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}. \tag{C.7}$$

By a strong Bahadur representation of He and Shao (1996) [cf. also, Janssen, Jureckova and Veraverbeke (1985)],

$$\hat{r}_n - r = -\frac{1}{n}\sum_{i=1}^{n} (\zeta(\boldsymbol{\theta}_i, r) - 1)/\dot{\lambda}(r) + o(n^{-1}\log^3 n) \quad \text{a.s.},$$

which implies immediately, by the central limit theorem

$$\sqrt{n}\,(\hat{r}_n - r) \to_{\mathscr{D}} N(0, \sigma^2), \tag{C.8}$$

where

$$\begin{aligned} \sigma^2 &= \text{Var}\frac{(\zeta(\boldsymbol{\theta}_1, r))}{(\dot{\lambda}(r))^2} \\ &= r^2\left[\int_{\Omega} \frac{(\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta}))^2}{\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \Big/ \left\{\int_{\Omega} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right\}^2\right]. \end{aligned}$$

In terms of (C.2), as in the proof of Theorem 3.1, one can show that $\{n(\hat{r}_n - r)^2, \, n \geq 1\}$ is uniformly integrable. Thus, (5.15) follows from (C.8). □

Acknowledgments. The authors thank the Editor and an Associate Editor for their helpful comments that improved the presentation.

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.

CHEN, M.-H. (1994a). Simulating ratios of normalization constants for the Gibbs Sampler. Technical report, Dept. Mathematical Sciences, Worcester Polytechnic Inst.

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

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

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

DEVROYE, L. (1986). *Non-Uniform Random Variate Generation*. Springer, New York.

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, Dept. Statistics, Univ. Chicago.

GELMAN, A. and MENG, X.-L. (1996). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Technical Report 440, Dept. Statistics, Univ. Chicago.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

GEWEKE, J. (1989). Bayesian inference in econometrics models using Monte Carlo integration. *Econometrica* **57** 1317–1340.

GEWEKE, J. (1994). Bayesian comparison of econometric models. Technical Report 532, Federal Reserve Bank of Minneapolis and Univ. Minnesota.

GEYER, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Rev. Technical Report 568, School of Statistics, Univ. Minnesota.

GREEN, P. J. (1992). Discussion of "Constrained Monte Carlo maximum likelihood for dependent data," by C. J. Geyer and E. A. Thompson. *J. Roy. Statist. Soc. Ser. B* **54** 657–699.

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

HE, X. and SHAO, Q. M. (1996). A general Bahadur representation of *M*-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630.

JANSSEN, P., JURECKOVA, J. and VERAVERBEKE, N. (1985). Rate of convergence of one- and two-step *M*-estimators with applications to maximum likelihood and Pitman estimators. *Ann. Statist.* **13** 1222–1229.

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, Dept. Statistics, Purdue Univ.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–549.

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

DEPARTMENT OF MATHEMATICAL SCIENCES
WORCESTER POLYTECHNIC INSTITUTE
WORCESTER, MASSACHUSETTS 01609-2280
E-MAIL: mhchen@wpi.edu

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OREGON
EUGENE, OREGON 97403-1222
E-MAIL: qmshao@darkwing.uoregon.edu