5
Estimating Ratios of Normalizing Constants

5.1 Introduction

A computational problem arising frequently in Bayesian inference is the computation of normalizing constants for posterior densities from which we can sample. Typically, we are interested in the ratios of such normalizing constants. For example, 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. Mathematically, this problem can be formulated as follows. Let $\pi_l(\boldsymbol{\theta})$, $l = 1, 2$, be two densities, each of which is known up to a normalizing constant:

$$\pi_l(\boldsymbol{\theta}) = \frac{q_l(\boldsymbol{\theta})}{c_l}, \quad \boldsymbol{\theta} \in \Omega_l,$$

where Ω_l is the support of π_l, and the unnormalized density $q_l(\boldsymbol{\theta})$ can be evaluated at any $\boldsymbol{\theta} \in \Omega_l$ for $l = 1, 2$. Then, the ratio of two normalizing constants is defined as

$$r = \frac{c_1}{c_2}. \tag{5.1.1}$$

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

$$\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_l) = \frac{q(\boldsymbol{\theta}|\boldsymbol{\lambda}_l)}{c(\boldsymbol{\lambda}_l)} \text{ for } l = 1, 2,$$

where $q(\boldsymbol{\theta}|\boldsymbol{\lambda}_l)$ is known, and the ratio is

$$r = \frac{c(\boldsymbol{\lambda}_1)}{c(\boldsymbol{\lambda}_2)}. \tag{5.1.2}$$

Estimating ratios of normalizing constants is extremely challenging and very important, particularly in Bayesian computations. Such problems often arise in likelihood inference, especially in the presence of missing data (Meng and Wong 1996), in computing intrinsic Bayes factors (Berger and Pericchi 1996), in the Bayesian comparison of econometric models considered by Geweke (1994), and in estimating marginal likelihood (Chib 1995). For example, in likelihood inference, this ratio is viewed as the likelihood ratio and in Bayesian model selection, the ratio is called the Bayes factor.

The $\pi_l(\boldsymbol{\theta})$ or $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_l)$ are often very complicated and therefore, the ratio defined by either (5.1.1) or (5.1.2) is analytically intractable (Meng and Wong 1996; Gelman and Meng 1998; Geyer 1994). However, without knowing the normalizing constants, c_l or $c(\boldsymbol{\lambda}_l)$, $l = 1, 2$, the distributions, $\pi_l(\boldsymbol{\theta})$ or $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_l)$, $l = 1, 2$, can be sampled by means of MCMC methods, for example, the Metropolis–Hastings algorithm, the Gibbs sampler, and the various hybrid algorithms (Chen and Schmeiser 1993; Müller 1991; 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 (MC) methods for estimating normalizing constants have been developed, which include bridge sampling of Meng and Wong (1996), path sampling of Gelman and Meng (1998), ratio importance sampling of Chen and Shao (1997a), Chib's method for computing marginal likelihood (Chib 1995), and reverse logistic regression of Geyer (1994). We start with importance sampling (IS) in Section 5.2. Sections 5.3–5.5 present bridge sampling (BS), path sampling (PS), and ratio importance sampling (RIS). A theoretical illustration is given in Section 5.6 and extensions to posterior densities with different dimensions are considered in Section 5.8. Section 5.7 presents a comprehensive treatment of how to compute simulation standard errors. The estimation of normalizing constants after transformation as well as some other related MC methods are discussed in Sections 5.9 and 5.10. An application of the weighted MC estimators discussed in Section 3.4.2 to the computation of the ratio of normalizing constants is given in Section 5.11. We conclude this chapter with a brief discussion in Section 5.12.

5.2 Importance Sampling

A standard and simple 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.

5.2.1 *Importance Sampling–Version 1*

Choose two importance sampling densities $\pi_l^{\mathrm{I}}(\boldsymbol{\theta})$, $l = 1, 2$, which are completely known, for $\pi_i(\boldsymbol{\theta})$, $l = 1, 2$, respectively. Let $\{\boldsymbol{\theta}_{l,1}, \boldsymbol{\theta}_{l,2}, \ldots, \boldsymbol{\theta}_{l,n_l}\}$, $l = 1, 2$, be two independent samples from $\pi_l^{\mathrm{I}}(\boldsymbol{\theta})$, $l = 1, 2$, respectively. Then an IS estimator of r is defined as

$$\hat{r}_{\mathrm{IS}_1} = \frac{(1/n_1)\sum_{i=1}^{n_1} q_1(\boldsymbol{\theta}_{1,i})/\pi_1^{\mathrm{I}}(\boldsymbol{\theta}_{1,i})}{(1/n_2)\sum_{i=1}^{n_2} q_2(\boldsymbol{\theta}_{2,i})/\pi_2^{\mathrm{I}}(\boldsymbol{\theta}_{2,i})}. \tag{5.2.1}$$

From the law of large numbers, it is easy to see that

$$\hat{r}_{\mathrm{IS}_1} \;\to\; r \;\text{ a.s. }\; \text{as } n_1, n_2 \to \infty.$$

To examine the performance of the estimator, $\hat{r}$, we introduce the relative mean-square error (RE^2) as a measure of accuracy:

$$\mathrm{RE}^2(\hat{r}_{\mathrm{IS}_1}) = \frac{E(\hat{r}_{\mathrm{IS}_1} - r)^2}{r^2}, \tag{5.2.2}$$

where the expectation is taken over all random samples. The exact calculation of (5.2.2) does not appear possible since it depends on the choice of the $\pi_i^{\mathrm{I}}(\boldsymbol{\theta})$. However, when both n_1 and n_2 are large, we can approximate (5.2.2) by the first-order term of its asymptotic expansion.

Theorem 5.2.1 *Let* $n = n_1 + n_2$, $s_{l,n} = n_l/n$. *Suppose that* $\lim_{n\to\infty} s_{l,n} > 0$ *for* $l = 1, 2$. *Then we have*

$$\mathrm{RE}^2(\hat{r}_{\mathrm{IS}_1}) = \sum_{l=1}^{2} \frac{1}{n_l} E_l^{\mathrm{I}} \left(\frac{\pi_l(\boldsymbol{\theta}) - \pi_l^{\mathrm{I}}(\boldsymbol{\theta})}{\pi_l^{\mathrm{I}}(\boldsymbol{\theta})} \right)^2 + o\Big(\frac{1}{n}\Big), \tag{5.2.3}$$

where the expectation E_l^{I} *is taken with respect to* $\pi_l^{\mathrm{I}}(\boldsymbol{\theta})$ *for* $l = 1, 2$.

The proof of Theorem 5.2.1 follows directly from the δ-method. From (5.2.3), it is easy to observe that the performance of the estimator, $\hat{r}_{\mathrm{IS}_1}$, depends heavily on the choice of $\pi_l^{\mathrm{I}}(\boldsymbol{\theta})$. If $\pi_l^{\mathrm{I}}(\boldsymbol{\theta})$ is a good approximation to $\pi_l(\boldsymbol{\theta})$, this IS method works well. However, it is often difficult to find $\pi_l^{\mathrm{I}}(\boldsymbol{\theta})$, $l = 1, 2$, which serve as good IS densities (see Geyer 1994; Green 1992; Gelman and Meng 1998). When the parameter spaces, Ω_l, $l = 1, 2$, are constrained, good completely known IS densities, $\pi_l^{\mathrm{I}}(\boldsymbol{\theta})$, $l = 1, 2$, are not available or are extremely difficult to obtain (see Gelfand, Smith, and Lee 1992 for practical examples).

5.2.2 *Importance Sampling–Version 2*

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

$$r = \frac{c_1}{c_2} = E_2 \left\{ \frac{q_1(\boldsymbol{\theta})}{q_2(\boldsymbol{\theta})} \right\}. \tag{5.2.4}$$

Here, and in the sequel, E_2 denotes the expected value with respect to π_2. Let $\{\boldsymbol{\theta}_{2,1}, \boldsymbol{\theta}_{2,2}, \ldots, \boldsymbol{\theta}_{2,n}\}$ be a random sample from π_2. Then the ratio r can be estimated by

$$\hat{r}_{\text{IS}_2} = \frac{1}{n} \sum_{i=1}^{n} \frac{q_1(\boldsymbol{\theta}_{2,i})}{q_2(\boldsymbol{\theta}_{2,i})}. \tag{5.2.5}$$

Unlike the estimator $\hat{r}_{\text{IS}_1}$ of r given in (5.2.1), it is easy to show that $\hat{r}_{\text{IS}_2}$ is an unbiased and consistent estimator of r and direct calculations yield

$$\text{RE}^2(\hat{r}_{\text{IS}_2}) = \frac{\text{Var}(\hat{r}_{\text{IS}_2})}{r^2} = \frac{1}{n} E_2 \left(\frac{\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})}{\pi_2(\boldsymbol{\theta})} \right)^2. \tag{5.2.6}$$

Thus it is easy to see that when the two densities π_1 and π_2 have very little overlap (i.e., $E_2(\pi_1(\boldsymbol{\theta}))$ is very small), this IS-based method will work poorly.

5.3 Bridge Sampling

The generalization of (5.2.4) given by Meng and Wong (1996) is

$$r = \frac{c_1}{c_2} = \frac{E_2\{q_1(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})\}}{E_1\{q_2(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})\}}, \tag{5.3.1}$$

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

$$0 < \left| \int_{\Omega_1 \cap \Omega_2} \alpha(\boldsymbol{\theta}) q_1(\boldsymbol{\theta}) q_2(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \right| < \infty. \tag{5.3.2}$$

The identity given in (5.3.1) unifies many identities used in the literature for simulating normalizing constants or other similar computations. As discussed in Meng and Wong (1996), the most general one is given by Bennett (1976), who proposes (5.3.1) in the context of simulating free-energy differences with $q_l = \exp(-U_l)$, where U_l is the temperature-scaled potential energy and $l = 1, 2$ indexes two canonical ensembles on the same configuration space. Taking $\alpha(\boldsymbol{\theta}) = q_2^{-1}(\boldsymbol{\theta})$ leads to (5.2.4), assuming $\Omega_1 \subset \Omega_2$. When $\Omega_1 = \Omega_2$ and Ω_1 has a finite Lebesgue measure, taking $\alpha(\boldsymbol{\theta}) = [q_1(\boldsymbol{\theta}) q_2(\boldsymbol{\theta})]^{-1}$ leads to a generalization of the "harmonic rule" given in Newton and Raftery (1994) and Gelfand and Dey (1994):

$$r = \frac{E_2[q_2^{-1}(\boldsymbol{\theta})]}{E_1[q_1^{-1}(\boldsymbol{\theta})]}.$$

Before discussing the optimal choice of $\alpha(\boldsymbol{\theta})$, we first define the BS estimator, denoted by $\hat{r}_{\text{BS}}(\alpha)$, of r. Letting $\{\boldsymbol{\theta}_{l,1}, \boldsymbol{\theta}_{l,2}, \ldots, \boldsymbol{\theta}_{l,n_l}\}$ be a random sample from π_l for $l = 1, 2$, a BS estimator of r is given by

$$\hat{r}_{\text{BS}} = \hat{r}_{\text{BS}}(\alpha) = \frac{(1/n_2)\sum_{i=1}^{n_2} q_1(\boldsymbol{\theta}_{2,i})\alpha(\boldsymbol{\theta}_{2,i})}{(1/n_1)\sum_{i=1}^{n_1} q_2(\boldsymbol{\theta}_{1,i})\alpha(\boldsymbol{\theta}_{1,i})}. \tag{5.3.3}$$

Similar to $\hat{r}_{\text{IS}_1}$ in (5.2.1), the law of large numbers yields that $\hat{r}_{\text{BS}}$ is a consistent estimator of r. Let $n = n_1 + n_2$ and $s_{l,n} = n_l/n$, and assume $s_l = \lim_{n\to\infty} s_{l,n} > 0$, $l = 1, 2$. Analogous to Theorem 5.2.1, the δ-method yields

$$\begin{aligned}\text{RE}^2(\hat{r}_{\text{BS}}) =& \frac{1}{ns_1s_2}\left\{\frac{\int_{\Omega_1\cap\Omega_2}\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})(s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta}))\alpha^2(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{(\int_{\Omega_1\cap\Omega_2}\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})\, d\boldsymbol{\theta})^2} - 1\right\}\\ &+ o\Big(\frac{1}{n}\Big).\end{aligned} \tag{5.3.4}$$

Meng and Wong (1996) provide the so-called (asymptotically) optimal choice of α, which is given by the following theorem:

Theorem 5.3.1 *The first term of the right side of* (5.3.4), *as a function of* α, *is minimized at*

$$\alpha_{\text{opt}}(\boldsymbol{\theta}) \;\propto\; \frac{1}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}, \quad \boldsymbol{\theta} \in \Omega_1 \cap \Omega_2, \tag{5.3.5}$$

with the minimum value

$$\frac{1}{ns_1s_2}\left[\left\{\int_{\Omega_1\cap\Omega_2}\frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right\}^{-1} - 1\right]. \tag{5.3.6}$$

The proof of the theorem is given in the Appendix. This asymptotically optimal choice is intuitively appealing. It represents the inverse of the mixture of π_1 and π_2 with mixture proportions determined by the sampling rates of the two distributions. But, it is not of direct use because α_{opt} depends on the unknown ratio $r = c_1/c_2$. Furthermore, it depends on the ratio of the two sample sizes, because $\alpha_{\text{opt}}(\boldsymbol{\theta}) \propto 1/(\pi_1(\boldsymbol{\theta}) + (n_2/n_1)\pi_2(\boldsymbol{\theta}))$. To overcome this problem, Meng and Wong (1996) construct the following iterative estimator:

$$\hat{r}^{(t+1)}_{\text{BS,opt}} = \frac{(1/n_2)\sum_{i=1}^{n_2} q_1(\boldsymbol{\theta}_{2,i})/(s_1q_1(\boldsymbol{\theta}_{2,i}) + s_2\hat{r}^{(t)}_{\text{BS,opt}}q_2(\boldsymbol{\theta}_{2,i}))}{(1/n_1)\sum_{i=1}^{n_1} q_2(\boldsymbol{\theta}_{1,i})/(s_1q_1(\boldsymbol{\theta}_{1,i}) + s_2\hat{r}^{(t)}_{\text{opt}}q_2(\boldsymbol{\theta}_{1,i}))}, \tag{5.3.7}$$

with an initial guess of r, $\hat{r}^{(0)}_{\text{BS,opt}}$. They show that for each $t \geq 0$, $\hat{r}^{(t+1)}_{\text{BS,opt}}$ provides a consistent estimator of r and that the unique limit, $\hat{r}_{\text{BS,opt}}$, achieves the asymptotic minimal relative mean-square error with the first-order term given in (5.3.6). By the construction of $\hat{r}^{(t+1)}_{\text{BS,opt}}$ given in (5.3.7), it can be shown that $\hat{r}_{\text{BS,opt}}$ must be a root of the following "score" function:

$$S(r) = \sum_{i=1}^{n_1}\frac{s_2rq_2(\boldsymbol{\theta}_{1,i})}{s_1q_1(\boldsymbol{\theta}_{1,i}) + s_2rq_2(\boldsymbol{\theta}_{1,i})} - \sum_{i=1}^{n_2}\frac{s_1q_1(\boldsymbol{\theta}_{2,i})}{s_1q_1(\boldsymbol{\theta}_{2,i}) + s_2rq_2(\boldsymbol{\theta}_{2,i})}. \tag{5.3.8}$$

Since $S(0) = -n_2 < 0$, $S(\infty) = n_1 > 0$, and

$$\frac{dS(r)}{dr} = \sum_{i=1}^{n_1} \frac{s_1 s_2 q_1(\boldsymbol{\theta}_{1,i}) q_2(\boldsymbol{\theta}_{1,i})}{[s_1 q_1(\boldsymbol{\theta}_{1,i}) + s_2 r q_2(\boldsymbol{\theta}_{1,i})]^2} + \sum_{i=1}^{n_2} \frac{s_1 s_2 q_1(\boldsymbol{\theta}_{2,i}) q_2(\boldsymbol{\theta}_{2,i})}{[s_1 q_1(\boldsymbol{\theta}_{2,i}) + s_2 r q_2(\boldsymbol{\theta}_{2,i})]^2} > 0$$

for all $r \geq 0$, $S(r)$ has a unique root. This property yields another approach to finding $\hat{r}_{\text{BS,opt}}$ instead of using the iterative procedure of Meng and Wong (1996), which requires an initial guess for an estimator of r. We solve the equation

$$S(r) = 0$$

to get $\hat{r}_{\text{BS,opt}}$ by, for example, a simple bisection method. Now, the only issue for a BS estimator is the choice of the sample sizes n_l. This issue is discussed in detail in Meng and Wong (1996), and it is shown that when $\Omega_1 = \Omega_2$ and $\alpha(\boldsymbol{\theta}) = [q_1(\boldsymbol{\theta}) q_2(\boldsymbol{\theta})]^{-1/2}$ is used, the optimal allocation of sample sizes, given $n_1 + n_2 = n$, is $n_1 = n_2 = n/2$. When sampling from the two densities requires a similar amount of time per sample, equal-sample-size allocation is also recommended by Bennett (1976). To obtain a simulation efficient BS estimator, the optimal choice of α is often more essential than the optimal allocation of sample sizes. However, equal-sample-size allocation may not be a good idea for the cases in which we know that the locations of both densities are roughly the same while one density has heavier tails than the other. Sometimes, it is even better that we just take random samples only from one density if it has extremely heavier tails. See Section 5.6 for an illustrative example.

Similar to the IS estimator $\hat{r}_{\text{IS}_2}$, the BS estimator $\hat{r}_{\text{BS}}$ given in (5.3.3) will become inefficient when π_1 and π_2 have little overlap; see Section 5.4.3 for further explanation. For such cases, the PS method of Gelman and Meng (1998) presented in Section 5.4, as well as the BS method after transformation as given in Section 5.9, will substantially improve the simulation efficiency.

5.4 Path Sampling

In this section, we let $q(\boldsymbol{\theta}|\boldsymbol{\lambda}_l)$ denote the unnormalized density and denote Ω to be the support of $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_l)$ for $l = 1, 2$. As discussed in Gelman and Meng (1998), we can often construct a continuous path to link $q(\boldsymbol{\theta}|\boldsymbol{\lambda}_1)$ and $q(\boldsymbol{\theta}|\boldsymbol{\lambda}_2)$. Instead of directly working on r, Gelman and Meng (1998) propose the PS method to estimate the natural logarithm of r, i.e.,

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

5.4.1 Univariate Path Sampling

We first consider λ to be a scalar quantity, i.e., λ is one dimensional. Without loss of generality, assume that $\lambda_1 < \lambda_2$. Gelman and Meng (1998) develop the following identity:

$$\xi = -\ln\left\{\frac{c(\lambda_1)}{c(\lambda_2)}\right\} = E\left[\frac{U(\boldsymbol{\theta},\lambda)}{\pi_\lambda(\lambda)}\right], \tag{5.4.1}$$

where $U(\boldsymbol{\theta},\lambda) = (d/d\lambda)\ln(q(\boldsymbol{\theta}|\lambda))$, $\pi_\lambda(\lambda)$ is a prior density (completely known) for $\lambda \in [\lambda_1,\lambda_2]$, and the expectation is taken with respect to the joint density $\pi(\boldsymbol{\theta},\lambda) = \pi(\boldsymbol{\theta}|\lambda)\pi_\lambda(\lambda)$, where $\pi(\boldsymbol{\theta}|\lambda) = q(\boldsymbol{\theta}|\lambda)/c(\lambda)$ for $\lambda = \lambda_1$ or λ_2. Let $\{(\boldsymbol{\theta}_i,\lambda_i), i = 1,2,\ldots,n\}$, be a random sample from $\pi(\boldsymbol{\theta},\lambda)$. Then, a PS estimator of ξ is given by

$$\hat{\xi}_{\text{PS}} = \frac{1}{n}\sum_{i=1}^{n}\frac{U(\boldsymbol{\theta}_i,\lambda_i)}{\pi_\lambda(\lambda_i)}. \tag{5.4.2}$$

It can be shown that $\hat{\xi}_{\text{PS}}$ is unbiased and consistent. The MC variance of $\hat{\xi}_{\text{PS}}$ is

$$\text{Var}(\hat{\xi}_{\text{PS}}) = \frac{1}{n}\left[\int_{\lambda_1}^{\lambda_2}\frac{E_\lambda\{U^2(\boldsymbol{\theta},\lambda)\}}{\pi_\lambda(\lambda)}\,d\lambda - \xi^2\right], \tag{5.4.3}$$

where the expectation E_λ is taken with respect to $\pi(\boldsymbol{\theta}|\lambda)$.

In (5.4.2), the choice of $\pi_\lambda(\lambda)$ is somehow arbitrary. However, the following result gives the optimal choice of $\pi_\lambda(\lambda)$ in the sense of minimizing the MC variance $\text{Var}(\hat{\xi}_{\text{PS}})$.

Theorem 5.4.1 *The optimal prior density* $\pi_\lambda^{\text{opt}}(\lambda)$ *given by*

$$\pi_\lambda^{\text{opt}}(\lambda) = \frac{\sqrt{E_\lambda\{U^2(\boldsymbol{\theta},\lambda)\}}}{\int_{\lambda_1}^{\lambda_2}\sqrt{E_\eta\{U^2(\boldsymbol{\theta},\eta)\}}\,d\eta}, \tag{5.4.4}$$

minimizes the MC variance $\text{Var}(\hat{\xi}_{\text{PS}})$ *given in* (5.4.3)*. The minimum value of* $\text{Var}(\hat{\xi})$ *is*

$$\text{Var}_{\text{opt}}(\hat{\xi}_{\text{PS}}) = \frac{1}{n}\left[\left(\int_{\lambda_1}^{\lambda_2}\sqrt{E_\lambda\{U^2(\boldsymbol{\theta},\lambda)\}}\,d\lambda\right)^2 - \xi^2\right]. \tag{5.4.5}$$

The proof of Theorem 5.4.1 is analogous to the one of Theorem 5.3.1 by the Cauchy–Schwarz inequality, and thus is left as an exercise. Interestingly, when $c(\lambda)$ is independent of λ, the optimal density given in (5.4.4) is exactly the Jeffreys' prior density based on $\pi(\boldsymbol{\theta}|\lambda)$ restricted to $\lambda \in [\lambda_1,\lambda_2]$; see Gelman and Meng (1998) for further explanation of the optimal prior density in general cases.

Gelman and Meng (1998) conjecture that the optimal MC 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 result confirms their conjecture:

Theorem 5.4.2 *Under certain regularity conditions, we have*

$$\mathrm{Var}(\hat{\xi}_{\mathrm{PS}}) \geq \frac{4}{n} \int_{\Omega} \left[\sqrt{\pi(\boldsymbol{\theta}|\lambda_1)} - \sqrt{\pi(\boldsymbol{\theta}|\lambda_2)} \right]^2 d\boldsymbol{\theta} \qquad (5.4.6)$$

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

The proof of Theorem 5.4.2 is given in the Appendix. It is interesting to see that the lower bound of $\mathrm{Var}(\hat{\xi}_{\mathrm{PS}})$ given in (5.4.6) indeed equals $(4/n)H^2(\pi_1, \pi_2)$, where

$$H(\pi_1, \pi_2) = \left\{ \int_{\Omega} \left[\sqrt{\pi_1(\boldsymbol{\theta})} - \sqrt{\pi_2(\boldsymbol{\theta})} \right]^2 d\boldsymbol{\theta} \right\}^{1/2} \qquad (5.4.7)$$

is the Hellinger divergence between two densities π_1 and π_2, and $\pi_l(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\lambda_l)$ for $l = 1, 2$.

5.4.2 Multivariate Path Sampling

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

$$\boldsymbol{\lambda}(t) = (\lambda_1(t), \ldots, \lambda_k(t)) \text{ for } t \in [0,1]; \quad \boldsymbol{\lambda}(0) = \boldsymbol{\lambda}_1 \text{ and } \boldsymbol{\lambda}(1) = \boldsymbol{\lambda}_2.$$

Under some regularity conditions, Gelman and Meng (1998) obtain the identity

$$\xi = -\ln \left\{ \frac{c(\boldsymbol{\lambda}_1)}{c(\boldsymbol{\lambda}_2)} \right\} = \int_0^1 E_{\boldsymbol{\lambda}(t)} \left[\sum_{j=1}^{k} \dot{\boldsymbol{\lambda}}_j(t) U_j(\boldsymbol{\theta}, \boldsymbol{\lambda}(t)) \right] dt,$$

where $\dot{\boldsymbol{\lambda}}_j(t) = d\boldsymbol{\lambda}_j(t)/dt$ and $U_j(\boldsymbol{\theta}, \boldsymbol{\lambda}(t)) = \partial \ln q(\boldsymbol{\theta}|\boldsymbol{\lambda})/\partial \boldsymbol{\lambda}_j$ for $j = 1, 2, \ldots, k$. Then, a corresponding PS estimator for ξ is given by

$$\hat{\xi}_{\mathrm{PS}} = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{k} \dot{\boldsymbol{\lambda}}_j(t_i) U_j(\boldsymbol{\theta}_i, \boldsymbol{\lambda}(t_i)) \right],$$

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

$$\mathrm{Var}(\hat{\xi}_{\mathrm{PS}}) = \frac{1}{n} \left[\int_0^1 \left(\sum_{i,j=1}^{k} g_{ij}(\boldsymbol{\lambda}(t)) \dot{\boldsymbol{\lambda}}_i(t) \dot{\boldsymbol{\lambda}}_j(t) \right) dt - \xi^2 \right], \qquad (5.4.8)$$

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 optimal path function $\boldsymbol{\lambda}(t)$ that minimizes the first term on the right side of (5.4.8) is the solution of the following Euler–Lagrange equations (e.g., see Atkinson and Mitchell 1981) with the boundary conditions $\boldsymbol{\lambda}(0) = \boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}(1) = \boldsymbol{\lambda}_2$:

$$\sum_{i=1}^{k} g_{ij}(\boldsymbol{\lambda}(t))\ddot{\boldsymbol{\lambda}}_i(t) + \sum_{i,j=1}^{k} [ij, l]\dot{\boldsymbol{\lambda}}_i(t)\dot{\boldsymbol{\lambda}}_j(t) = 0 \quad \text{for } l = 1, 2, \ldots, k, \tag{5.4.9}$$

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

$$[ij, l] = \frac{1}{2}\left[\frac{\partial g_{il}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}_j} + \frac{\partial g_{jl}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}_i} - \frac{\partial g_{ij}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}_l}\right], \quad i, j, l = 1, 2, \ldots, k.$$

5.4.3 Connection Between Path Sampling and Bridge Sampling

The fundamental idea underlying the BS approach is to take advantage of the "overlap" of the two densities. Indeed, a crucial (implicit) condition behind (5.3.2) is that $\Omega_1 \cap \Omega_2$ is nonempty: the more the overlap is, the more efficient the BS estimates are. To see this idea more clearly, Gelman and Meng (1998) consider a reexpression of (5.3.1) by taking $\alpha = q_{3/2}/(q_1 q_2)$ where $q_{3/2}$ is an arbitrary unnormalized density having support $\Omega_1 \cap \Omega_2$ while the subscript "3/2" indicates a density that is "between" π_1 and π_2. Substituting this α into (5.3.1) yields

$$r = \frac{c_1}{c_2} = \frac{E_2[q_{3/2}/q_2]}{E_1[q_{3/2}/q_1]}. \tag{5.4.10}$$

Comparing (5.4.10) to (5.2.4), we see that estimating r with (5.2.4) requires random samples from π_2 to "reach" π_1, whereas with (5.4.10) random samples from both q_1 and q_2 with $q_{3/2}$ as a connecting "bridge" can be used to estimate r. Thus, use of (5.4.10) effectively shortens the distance between the two densities. This idea essentially leads to extensions using multiple bridges, that is, by applying (5.4.10) in a "chain" fashion. Gelman and Meng (1998) show that the limit from using infinitely many bridges leads to the PS identity given in (5.4.1). Thus, BS is a natural extension of IS while PS is a further extension of BS.

5.5 Ratio Importance Sampling

5.5.1 The Method

In the same spirit as reducing the distance between two densities, Torrie and Valleau (1977) and Chen and Shao (1997a) propose another MC method for estimating a ratio of two normalizing constants. Their method is based on the following identity:

$$r = \frac{c_1}{c_2} = \frac{E_\pi\{q_1(\boldsymbol{\theta})/\pi(\boldsymbol{\theta})\}}{E_\pi\{q_2(\boldsymbol{\theta})/\pi(\boldsymbol{\theta})\}}, \tag{5.5.1}$$

where the expectation E_π is taken with respect to π and $\pi(\boldsymbol{\theta})$ is an arbitrary density with the support $\Omega = \Omega_1 \cup \Omega_2$. In (5.5.1), π serves as a "middle" density between π_1 and π_2. It is interesting to see that (5.5.1) is "opposite" to (5.4.10). With (5.4.10), we need random samples from both π_1 and π_2 while with (5.5.1), only one random sample from the "middle" density π is required for estimating r. This is advantageous in the context of computing posterior model probabilities since many normalizing constants need to be estimated simultaneously (see Chapters 8 and 9 for more details). It can also be observed that (5.5.1) is an extension of (5.2.4) since (5.5.1) reduces to (5.2.4) by taking $\pi = \pi_2$.

Torrie and Valleau (1977) call this method "umbrella sampling," conveying the intention of constructing a middle density that "covers" both ends. However, Chen and Shao (1997a) term this method RIS because:

(i) it is a natural extension of IS;

(ii) the identity given in (5.5.1) contains the "middle" density π in both numerator and denominator in a ratio fashion; and

(iii) most importantly, this method is used for estimating a ratio of two normalizing constants.

Although this method is initially proposed by Torrie and Valleau (1977), the theoretical properties of this method are explored by Chen and Shao (1997a) and extensions of this method to Bayesian variable selection are considered by Ibrahim, Chen, and MacEachern (1999) and Chen, Ibrahim, and Yiannoutsos (1999). Given a random sample $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_n\}$ from π, a RIS estimator of r is given by

$$\hat{r}_{\text{RIS}} = \hat{r}_{\text{RIS}}(\pi) = \frac{\sum_{i=1}^n q_1(\boldsymbol{\theta}_i)/\pi(\boldsymbol{\theta}_i)}{\sum_{i=1}^n q_2(\boldsymbol{\theta}_i)/\pi(\boldsymbol{\theta}_i)}. \tag{5.5.2}$$

For any π with the support Ω, $\hat{r}_{\text{RIS}}$ is a consistent estimator of r. To explore further properties of $\hat{r}_{\text{RIS}}$, we let

$$\text{RE}^2(\hat{r}_{\text{RIS}}) = \frac{E_\pi(\hat{r}_{\text{RIS}} - r)^2}{r^2} \tag{5.5.3}$$

denote the relative mean-square error which is similar to (5.2.2). The analytical calculation of (5.5.3) is typically intractable. However, under the assumption that the $\boldsymbol{\theta}_i$ are independent and identically distributed (i.i.d.) from π, we can obtain the asymptotic form of $\text{RE}^2(\hat{r}_{\text{RIS}})$. Let $f_1(\boldsymbol{\theta}) = q_1(\boldsymbol{\theta})/\pi(\boldsymbol{\theta})$ and $f_2(\boldsymbol{\theta}) = q_2(\boldsymbol{\theta})/\pi(\boldsymbol{\theta})$. Then, we have $E_\pi[f_1(\boldsymbol{\theta})] = c_1$ and $E_\pi[f_2(\boldsymbol{\theta})] = c_2$. We are led to the following theorem:

Theorem 5.5.1 *Let* $\{\boldsymbol{\theta}_i,\ i = 1, 2, \ldots\}$ *be* i.i.d. *random samples from* π. *Assume* $\int_\Omega |q_1(\boldsymbol{\theta}) - aq_2(\boldsymbol{\theta})|\ d\boldsymbol{\theta} > 0$ *for every* $a > 0$,

$$E_\pi\left(\frac{f_1(\boldsymbol{\theta})}{c_1} - \frac{f_2(\boldsymbol{\theta})}{c_2}\right)^2 < \infty, \quad \text{and } E_\pi\{f_1(\boldsymbol{\theta})/f_2(\boldsymbol{\theta})\}^2 < \infty.$$

Then

$$\lim_{n\to\infty} n\, \mathrm{RE}^2(\hat{r}_{\mathrm{RIS}}) = E_\pi\left\{\frac{f_1(\boldsymbol{\theta})}{c_1} - \frac{f_2(\boldsymbol{\theta})}{c_2}\right\}^2, \tag{5.5.4}$$

and

$$\sqrt{n}(\hat{r}_{\mathrm{RIS}} - r) \xrightarrow{\mathcal{D}} N\left(0, r^2 E_\pi\left\{\frac{f_1(\boldsymbol{\theta})}{c_1} - \frac{f_2(\boldsymbol{\theta})}{c_2}\right\}^2\right) \quad \text{as } n \to \infty. \tag{5.5.5}$$

If, in addition, $E_\pi(f_1(\boldsymbol{\theta})/c_1 - f_2(\boldsymbol{\theta})/c_2)^4 < \infty$ *and* $E_\pi f_2^4(\boldsymbol{\theta}) < \infty$, *then*

$$\mathrm{RE}^2(\hat{r}_{RIS}) = \frac{1}{n}E_\pi\left\{\frac{f_1(\boldsymbol{\theta})}{c_1} - \frac{f_2(\boldsymbol{\theta})}{c_2}\right\}^2 + O\left(\frac{1}{n^2}\right) \quad \text{as } n \to \infty. \tag{5.5.6}$$

The proof of Theorem 5.5.1 is given in the Appendix. By (5.5.4), we have the asymptotic form of $\mathrm{RE}^2(\hat{r}_{\mathrm{RIS}})$:

$$\mathrm{RE}^2(\hat{r}_{\mathrm{RIS}}) = \frac{1}{n}E_\pi\left[\frac{\{\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})\}^2}{\pi^2(\boldsymbol{\theta})}\right] + o\Big(\frac{1}{n}\Big). \tag{5.5.7}$$

When $\Omega_1 \subset \Omega_2$ and $\pi(\boldsymbol{\theta}) = \pi_2(\boldsymbol{\theta}) = q_2(\boldsymbol{\theta})/c_2$, (5.5.2) becomes the importance sampling estimator (5.2.5) for r, and the corresponding relative mean-square error is

$$\mathrm{RE}^2(\hat{r}_{\mathrm{IS}_2}) = \frac{1}{n}\int_{\Omega_2} \frac{(\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta}))^2}{\pi_2(\boldsymbol{\theta})}\ d\boldsymbol{\theta}, \tag{5.5.8}$$

which is the χ^2-divergence, denoted by $\chi^2(\pi_2, \pi_1)$, between π_2 and π_1.

Since the RIS estimator $\hat{r}_\pi$ depends on π, it is of interest to determine the optimal RIS density π_{opt} of π. The result is given in the following theorem:

Theorem 5.5.2 *Assume* $\int_\Omega |q_1(\boldsymbol{\theta}) - aq_2(\boldsymbol{\theta})|\ d\boldsymbol{\theta} > 0$ *for every* $a > 0$. *The first term of the right side of* (5.5.7) *is minimized at*

$$\pi_{\mathrm{opt}}(\boldsymbol{\theta}) = \frac{|\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})|}{\int_\Omega |\pi_1(\boldsymbol{\delta}) - \pi_2(\boldsymbol{\delta})|\, d\boldsymbol{\delta}} \tag{5.5.9}$$

with a minimal value

$$\frac{1}{n}\left[\int_\Omega |\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})|\ d\boldsymbol{\theta}\right]^2. \tag{5.5.10}$$

The proof of Theorem 5.5.2 is given in the Appendix. It is interesting to note that (5.5.10) is $(1/n)L_1^2(\pi_1, \pi_2)$, where $L_1(\pi_1, \pi_2)$ is the L_1-divergence between π_1 and π_2. From Theorem 5.5.2, and (5.5.8) and (5.5.10), we also have $L_1^2(\pi_1, \pi_2) \le \chi^2(\pi_2, \pi_1)$.

Now, we compare the RIS method with the BS method. The following theorem states that the RIS estimator (5.5.2) with the optimal π_{opt} given in (5.5.9) has a smaller asymptotic relative mean-square error than the BS estimator (5.3.3) with the optimal choice α_{opt} given in (5.3.5).

Theorem 5.5.3 *For* $0 < s_1, s_2 < 1$, *and* $s_1 + s_2 = 1$, *we have*

$$\left[\int_{\Omega} |\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})|\, d\boldsymbol{\theta}\right]^2 \leq (s_1 s_2)^{-1} \left[\left\{\int_{\Omega_1 \cap \Omega_2} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right\}^{-1} - 1\right]. \tag{5.5.11}$$

The proof of the theorem given in the Appendix.

Next, we compare the RIS method with the PS method. Gelman and Meng (1998) point out that the asymptotic variance $\hat{\xi}_{\text{PS}}$ is the same as the asymptotic relative mean-square error of $\hat{r}$, i.e.,

$$\lim_{n\to\infty} n \operatorname{Var}(\hat{\xi}_{\text{PS}}) = \lim_{n\to\infty} nE(\hat{r}_{\text{PS}} - r)^2/r^2,$$

where $\hat{r}_{\text{PS}} = \exp(-\hat{\xi}_{\text{PS}})$. Thus, the next theorem shows that the asymptotic relative mean-square error of the RIS estimator (5.5.2) with the optimal π_{opt} is less than the lower bound, given on the right side of (5.4.6), of the variance of $\hat{\xi}_{\text{PS}}$ given in (5.4.3).

Theorem 5.5.4 *Defining* $\pi_l(\boldsymbol{\theta}) = q_l(\boldsymbol{\theta})/c_l = \pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_l)$ *for* $l = 1, 2$, *we have*

$$\left[\int_{\Omega} |\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})|\, d\boldsymbol{\theta}\right]^2 \leq 4 \int \left[\sqrt{\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_1)} - \sqrt{\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}_2)}\right]^2 d\boldsymbol{\theta}. \tag{5.5.12}$$

The proof of Theorem 5.5.4 is given in the Appendix. From Theorem 5.5.4, we can see that $L_1^2(\pi_1, \pi_2) \leq 4H^2(\pi_1, \pi_2)$ and that the optimal RIS estimator $\hat{r}_{\text{RIS}}(\pi_{\text{opt}})$ is always better than the BS estimator, and $\hat{r}_{\text{RIS}}(\pi_{\text{opt}})$ is also better than any PS 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 the next subsection.

5.5.2 *Implementation*

In this subsection, we present two approaches to implement the optimal RIS estimators. We also discuss other "nonoptimal" implementation schemes.

Exact Optimal Scheme

Let $\pi(\boldsymbol{\theta})$ be an arbitrary density over Ω such that $\pi(\boldsymbol{\theta}) > 0$ for $\boldsymbol{\theta} \in \Omega$.

Given a random sample $\{\boldsymbol{\theta}_i,\ i=1,2,\ldots,n\}$ from π, define

$$\tau_n = \frac{\sum_{i=1}^n q_1(\boldsymbol{\theta}_i)/\pi(\boldsymbol{\theta}_i)}{\sum_{i=1}^n q_2(\boldsymbol{\theta}_i)/\pi(\boldsymbol{\theta}_i)} \tag{5.5.13}$$

and let

$$\psi_n(\boldsymbol{\theta}) = \frac{|q_1(\boldsymbol{\theta}) - \tau_n q_2(\boldsymbol{\theta})|}{\int_\Omega |q_1(\boldsymbol{\delta}) - \tau_n q_2(\boldsymbol{\delta})|\, d\boldsymbol{\delta}}. \tag{5.5.14}$$

Then, take a random sample $\{\boldsymbol{\vartheta}_{n,1}, \boldsymbol{\vartheta}_{n,2}, ..., \boldsymbol{\vartheta}_{n,n}\}$ from ψ_n and define the "optimal" estimator $\hat{r}_{\text{RIS},n}$ as follows:

$$\hat{r}_{\text{RIS},n} = \frac{\sum_{i=1}^n q_1(\boldsymbol{\vartheta}_{n,i})/\psi_n(\boldsymbol{\vartheta}_{n,i})}{\sum_{i=1}^n q_2(\boldsymbol{\vartheta}_{n,i})/\psi_n(\boldsymbol{\vartheta}_{n,i})}. \tag{5.5.15}$$

Then, we have the following result:

Theorem 5.5.5 *Suppose that there exists a neighborhood U_r of r such that the following conditions are satisfied:*

(i) $\inf_{a\in U_r} \int_\Omega |q_1(\boldsymbol{\theta}) - a q_2(\boldsymbol{\theta})|\, d\boldsymbol{\theta} > 0$;

(ii) $\int_\Omega \sup_{a\in U_r} \dfrac{q_1^2(\boldsymbol{\theta}) + q_2^2(\boldsymbol{\theta})}{|q_1(\boldsymbol{\theta}) - a q_2(\boldsymbol{\theta})|}\, d\boldsymbol{\theta} < \infty$; *and*

(iii) $\sup_{a\in U_r} \int_\Omega \dfrac{q_1^2(\boldsymbol{\theta})|q_1(\boldsymbol{\theta}) - a q_2(\boldsymbol{\theta})|}{q_2^2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} < \infty.$

Then

$$\lim_{n\to\infty} nE\left(\frac{(\hat{r}_{\text{RIS},n} - r)^2}{r^2} \,\middle|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_n\right) = \left[\int_\Omega |\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})|\, d\boldsymbol{\theta}\right]^2 \quad \text{a.s.} \tag{5.5.16}$$

The proof of Theorem 5.5.5 is given in the Appendix. Theorem 5.5.5 says that the "optimal" estimator $\hat{r}_{\text{RIS},n}$ obtained by the two-stage sampling scheme has the same optimal relative mean-square error as $\hat{r}_{\text{RIS}}(\pi_{\text{opt}})$. 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 (5.5.13) and (5.5.14) (the first-stage sample size) and n_2 in (5.5.15) (the second-stage sample size). Then, (5.5.16) still holds as long as $n_1 = o(n)$ and $n_1 \to \infty$, where $n = n_1 + n_2$.

Approximate Optimal Scheme

Let $\pi_l^{\text{I}}(\boldsymbol{\theta})$, $l = 1, 2$, be good importance sampling densities for $\pi_l(\boldsymbol{\theta})$, $l = 1, 2$, respectively. Then, the optimal RIS density, π_{opt}, can be approximated by

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

Let $\{\boldsymbol{\theta}_i,\ i = 1, 2, \ldots, n\}$ be a random sample from $\pi^{\mathrm{I}}_{\mathrm{opt}}$. Then an approximate optimal RIS estimator is given by

$$\hat{r}_{\pi^{\mathrm{I}}_{\mathrm{opt}}} = \frac{\sum_{i=1}^{n} q_1(\boldsymbol{\theta}_i) / \left|\pi^{\mathrm{I}}_1(\boldsymbol{\theta}_i) - \pi^{\mathrm{I}}_2(\boldsymbol{\theta}_i)\right|}{\sum_{i=1}^{n} q_2(\boldsymbol{\theta}_i) / \left|\pi^{\mathrm{I}}_1(\boldsymbol{\theta}_i) - \pi^{\mathrm{I}}_2(\boldsymbol{\theta}_i)\right|}.$$

Note that when π_1 and π_2 do not overlap, we can choose $\pi^{\mathrm{I}}_{\mathrm{opt}}(\boldsymbol{\theta}) = \{\pi^{\mathrm{I}}_1(\boldsymbol{\theta}) + \pi^{\mathrm{I}}_2(\boldsymbol{\theta})\}/2$ because $\pi_{\mathrm{opt}}(\boldsymbol{\theta}) = \{\pi_1(\boldsymbol{\theta}) + \pi_2(\boldsymbol{\theta})\}/2$. For such cases, sampling from $\pi^{\mathrm{I}}_{\mathrm{opt}}$ is straightforward.

Other "Nonoptimal" Schemes

First, assume that π_1 and π_2 do not overlap, i.e., $\int_\Omega q_1(\boldsymbol{\theta}) q_2(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 0$. For this case, the IWMDE method of Chen (1994) will give a reasonably good estimator of r. Let $w_l(\boldsymbol{\theta})$ be a weighted density with a shape roughly similar to q_l, for $l = 1, 2$. Also let $\{\boldsymbol{\theta}_{l,i}, i = 1, 2, \ldots, n_l\}$, $l = 1, 2$, be independent random samples from π_l, $l = 1, 2$, respectively. Then, a consistent estimator of r is

$$\hat{r}_{\mathrm{IWMDE}} = \frac{(1/n_2)\sum_{i=1}^{n_2} w_2(\boldsymbol{\theta}_{2,i}) / q_2(\boldsymbol{\theta}_{2,i})}{(1/n_1)\sum_{i=1}^{n_1} w_1(\boldsymbol{\theta}_{1,i}) / q_1(\boldsymbol{\theta}_{1,i})}.$$

In this case, PS is also useful (if it is applicable).

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

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

where $0 < \psi < 1$ is known (e.g., $\psi = \frac{1}{2}$). Note that we can straightforwardly sample from $\pi_{\mathrm{mix}}(\boldsymbol{\theta})$ by a composition method without knowing c_1 and c_2. Let

$$S_n(r) = \sum_{i=1}^{n} \frac{r q_2(\boldsymbol{\theta}_i)}{\psi q_1(\boldsymbol{\theta}_i) + r \cdot (1 - \psi) q_2(\boldsymbol{\theta}_i)} - \sum_{i=1}^{n} \frac{q_1(\boldsymbol{\theta}_i)}{\psi q_1(\boldsymbol{\theta}_i) + r \cdot (1 - \psi) q_2(\boldsymbol{\theta}_i)}.$$

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

$$S_n(r) = 0. \tag{5.5.17}$$

Similar to (5.3.8), it can be shown that there exists a unique solution of (5.5.17). The asymptotic properties of $\hat{r}_{\mathrm{BS},n}$ are given in the next theorem.

Theorem 5.5.6 *Suppose that* $\int_\Omega q_1(\boldsymbol{\theta}) q_2(\boldsymbol{\theta})\, d\boldsymbol{\theta} > 0$. *Then*

$$\hat{r}_{\mathrm{BS},n} \xrightarrow{\text{a.s.}} r \quad \text{as } n \to \infty. \tag{5.5.18}$$

If, in addition, $E_{\pi_{\text{mix}}}(q_1(\boldsymbol{\theta})/q_2(\boldsymbol{\theta}))^2 < \infty$, *then*

$$\lim_{n\to\infty} n\, E_{\pi_{\text{mix}}} \frac{(\hat{r}_{\text{BS},n} - r)^2}{r^2}$$
$$= \int_\Omega \frac{(\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta}))^2}{\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \cdot \left\{ \int_\Omega \frac{\pi_1(\boldsymbol{\theta}) \cdot \pi_2(\boldsymbol{\theta})}{\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \right\}^{-2}. \tag{5.5.19}$$

The proof of this theorem is given in the Appendix.

5.6 A Theoretical Illustration

To get a better understanding of IS, BS, PS, and RIS, we conduct two theoretical case studies based on two normal densities where we know the exact values of the two normalizing constants.

CASE 1. $N(0,1)$ *and* $N(\delta,1)$

Let $q_1(\theta) = \exp(-\theta^2/2)$ and $q_2(\theta) = \exp(-(\theta-\delta)^2/2)$ with δ a known positive constant. In this case, $c_1 = c_2 = \sqrt{2\pi}$ and, therefore, $r = 1$ and $\xi = -\ln(r) = 0$. For PS, we consider q_1 and q_2 as two points in the family of unnormalized normal densities: $q(\theta|\boldsymbol{\lambda}) = \exp\{-(\theta-\mu)^2/2\sigma^2\}$, with $\boldsymbol{\lambda} = (\mu,\sigma)'$, $\boldsymbol{\lambda}_1 = (0,1)'$, and $\boldsymbol{\lambda}_2 = (\delta,1)'$.

As discussed in Gelman and Meng (1998), in order to make fair comparisons, we assume that:

(i) with IS–version 2, we sample n i.i.d. observations from $N(\delta,1)$;

(ii) with BS, we sample $n/2$ (assume n is even) i.i.d. observations from each of $N(0,1)$ and $N(\delta,1)$;

(iii) with PS, we first sample t_i, $i = 1,2,\ldots,n$, uniformly from $(0,1)$ and then sample 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 RIS, we sample n i.i.d. observations from the optimal RIS density:

$$\pi_{\text{opt}}(\theta) = \frac{|\phi(\theta) - \phi(\theta-\delta)|}{c_{\text{opt}}(\delta)}, \tag{5.6.1}$$

where

$$\begin{aligned} c_{\text{opt}}(\delta) &= \int_{-\infty}^{\infty} |\phi(\theta) - \phi(\theta-\delta)|\, d\theta \\ &= 2(\Phi(\delta/2) - \Phi(-\delta/2)) \\ &= 2(2\Phi(\delta/2) - 1), \end{aligned} \tag{5.6.2}$$

and ϕ and Φ are the $N(0,1)$ probability density function and cumulative distribution function, respectively.

Since the cumulative distribution function (cdf) for $\pi_{\text{opt}}(\theta)$ is

$$\Pi_{\text{opt}}(\theta) = \begin{cases} (\Phi(\theta) - \Phi(\theta-\delta))\,/2\,(2\Phi(\delta/2)-1) & \text{for } \theta \le \delta/2, \\ 1 - (\Phi(\theta) - \Phi(\theta-\delta))\,/2\,(2\Phi(\delta/2)-1) & \text{for } \theta > \delta/2, \end{cases} \tag{5.6.3}$$

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

Since the asymptotic variance of $\hat{\xi}_{\text{PS}}$ is the same as the asymptotic relative mean-square error of $\hat{r}_{\text{PS}} = \exp(-\hat{\xi}_{\text{PS}})$, that is,

$$\lim_{n\to\infty} n \operatorname{Var}(\hat{\xi}_{\text{PS}}) = \lim_{n\to\infty} nE(\hat{r}_{\text{PS}} - r)^2/r^2,$$

using (5.5.10), (5.6.2), and the results given by Gelman and Meng (1998), we obtain Table 5.1.

TABLE 5.1. Comparison of Asymptotic Relative Mean-Square Errors (I).

Index	Method	$\lim_{n\to\infty} \sqrt{nE(\hat{r}-r)^2/r^2}$
1	IS–version 2	$\left\{\exp(\delta^2)-1\right\}^{1/2}$
2	BS with $\alpha = (q_1 q_2)^{-1/2}$	$2\left\{\exp\left(\frac{\delta^2}{4}\right)-1\right\}^{1/2}$
3	Optimal BS with α_{opt}	$2\left\{\frac{\delta\exp\left(\delta^2/8\right)}{\beta(\delta)\sqrt{2\pi}}-1\right\}^{1/2}$
4	Optimal PS in μ-space	δ
5	Optimal PS in $(\mu,\sigma)'$-space	$\sqrt{12}\left\{\ln\left(\frac{\delta}{\sqrt{12}}+\sqrt{1+\frac{\delta^2}{12}}\right)\right\}$
6	Lower bound of PS in (5.4.6)	$\sqrt{8}\left(1-\exp(-\delta^2/8)\right)^{1/2}$
7	Optimal RIS with π_{opt}	$2\left(2\Phi(\delta/2)-1\right)$

In Table 5.1, for optimal BS,

$$\beta(\delta) = \frac{1}{\pi}\int_0^\infty \exp(-\theta^2/2\delta^2)/\cosh(\theta/2)\, d\theta.$$

For the normal family $N(\mu(t), \sigma^2(t))$, the optimal path for PS in μ-space is the solution of the Euler–Lagrange equation given in (5.4.9) with $k=1$ and

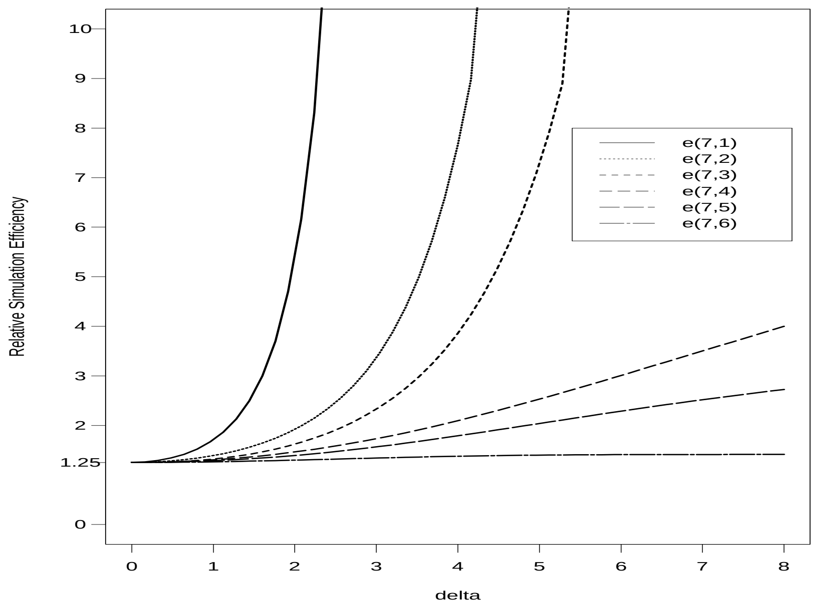

FIGURE 5.1. Relative simulation efficiency plot (I).

boundary conditions $\mu(0) = 0$ and $\mu(1) = \delta$ when we treat a fixed $\sigma^2(t) \equiv 1$, and the optimal path in $(\mu, \sigma)'$-space is the Euler–Lagrange equation with $k = 2$ while both $\mu(t)$ and $\sigma^2(t)$ are functions of t and boundary conditions are $(\mu(0), \sigma^2(0))' = (0, 1)'$ and $(\mu(1), \sigma^2(1))' = (\delta, 1)'$. The derivation of Table 5.1 is left as an exercise.

We define the relative simulation efficiency as follows:

$$e(i,j) = \frac{\lim_{n\to\infty} \sqrt{nE(\hat{r}-r)^2/r^2} \text{ for method } j}{\lim_{n\to\infty} \sqrt{nE(\hat{r}-r)^2/r^2} \text{ for method } i} \quad \text{for } i,j = 1,2,\ldots,7, \tag{5.6.4}$$

where $\hat{r}$ is an estimator of r. Then, $e(7, j)$, $j = 1, \ldots, 6$, versus δ are plotted in Figure 5.1. 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 $j = 1, 2, \ldots, 6$, and

$$\lim_{\delta\to 0} e(7,j) = \sqrt{2\pi}/2 = 1.2533$$

for all $j = 1, 2, \ldots, 6$. Therefore, the lower bound of PS in (5.4.6) is quite close to the asymptotic relative mean-square error of the RIS method with the optimal π_{opt}. The RIS method is significantly better than the BS method, especially for $\delta > 3$, and it is also better than the PS method. In this case, both RIS and PS are much better than IS–version 2.

CASE 2. *$N(0,1)$ and $N(0,\Delta^2)$*
Without loss of generality, we consider $\Delta > 1$ only. Let $q_1(\theta) = \exp(-\theta^2/2)$ and $q_2(\theta) = \exp(-\theta^2/2\Delta^2)$ 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 PS, $\xi = \ln\Delta$. Let $q(\theta|\boldsymbol{\lambda}_1) = q_1(\theta)$ and $q(\theta|\boldsymbol{\lambda}_2) = q_2(\theta)$ with $\boldsymbol{\lambda}_1 = (0,1)'$ and $\boldsymbol{\lambda}_2 = (0,\Delta)'$.

For IS-version 2, BS, and PS, we use the sampling schemes similar to those in Case 1 by using $N(0,\Delta^2)$ to replace $N(\delta,1)$. For RIS, the optimal density is

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

where

$$\begin{aligned} c_{\text{opt}}(\Delta) &= \int_{-\infty}^{\infty} \left| \phi(\theta) - \frac{1}{\Delta}\phi\left(\frac{\theta}{\Delta}\right) \right| d\theta \\ &= 4\left[\Phi\left(\sqrt{\frac{2\ln\Delta}{1-1/\Delta^2}}\right) - \Phi\left(\frac{1}{\Delta}\sqrt{\frac{2\ln\Delta}{1-1/\Delta^2}}\right)\right]. \end{aligned} \tag{5.6.5}$$

The corresponding optimal cumulative distribution is

$$\Pi_{\text{opt}}(\theta) = \begin{cases} \dfrac{\Phi(\theta/\Delta) - \Phi(\theta)}{c_{\text{opt}}(\Delta)} & \text{for } \theta \le -\sqrt{\dfrac{2\ln\Delta}{1-1/\Delta^2}}, \\ \dfrac{1}{2} + \dfrac{\Phi(\theta) - \Phi(\frac{\theta}{\Delta})}{c_{\text{opt}}(\Delta)} & \text{for } -\sqrt{\dfrac{2\ln\Delta}{1-1/\Delta^2}} < \theta \le \sqrt{\dfrac{2\ln\Delta}{1-1/\Delta^2}}, \\ 1 - \dfrac{\Phi(\theta) - \Phi(\frac{\theta}{\Delta})}{c_{\text{opt}}(\Delta)} & \text{for } \theta < \sqrt{\dfrac{2\ln\Delta}{1-1/\Delta^2}}. \end{cases}$$

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

In this case, the optimal path in $(\mu,\sigma)'$-space with boundary conditions $\mu(t) = 0$ and $\sigma(t) = \Delta^t$ for $0 \le t \le 1$ can be obtained using Problem 4.12 in the exercises. Then, using (5.3.6), (5.4.5), (5.4.8), (5.5.10), and (5.6.5), we derive the asymptotic relative mean-square errors (variances) for IS, BS, PS, and RIS, which are reported in Table 5.2. In Table 5.2,

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

and $h(\Delta) = (2\ln\Delta/(1-1/\Delta^2))^{1/2}$.

TABLE 5.2. Comparison of Asymptotic Relative Mean-Square Errors (II).

Index	Method	$\lim_{n\to\infty} \sqrt{nE(\hat{r}-r)^2/r^2}$
1	IS–version 2	$\sqrt{(\Delta^2/\sqrt{2\Delta^2-1})-1}$
2	BS with $\alpha = (q_1 q_2)^{-1/2}$	$\sqrt{2}(\Delta-1)/\sqrt{\Delta}$
3	Optimal BS α_{opt}	$2b(\Delta)$
4	Optimal PS in μ-space	$\sqrt{2}\ln\Delta$
5	Optimal PS in $(\mu,\sigma)'$-space	$\sqrt{2}\ln\Delta$
6	Lower bound of PS in (5.4.6)	$2\sqrt{2}\left(1-\sqrt{2\Delta/(1+\Delta^2)}\right)^{1/2}$
7	Optimal RIS with π_{opt}	$4\left[\Phi\left(h(\Delta)\right)-\Phi\left(\frac{1}{\Delta}h(\Delta)\right)\right]$

The relative simulation efficiencies defined in (5.6.4) are calculated and $e(7,j)$, $j = 1,2,\ldots,6$, versus Δ are also plotted in Figure 5.2. It can be shown 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 RIS method is better than all five counterparts. Once again, the lower bound of PS and the asymptotic relative mean-square error of optimal RIS are very close. Note that it is not necessarily true that optimal BS is better than IS–version 2 because of our sampling scheme. However, it is true that

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

Thus, when one density has a heavier tail than another, taking samples from the heavier-tailed one is always more beneficial. For example, when one is a normal density and another is a Student t density, we recommend that a random sample be taken from the Student t distribution. Furthermore, for this case, we can see that even the simple IS method (version 2) is better than the optimal PS method. Therefore, PS is advantageous only for the cases where the two modes of π_1 and π_2 are far away from each other. Finally, we note that reverse logistic regression (see Section 5.10.2)

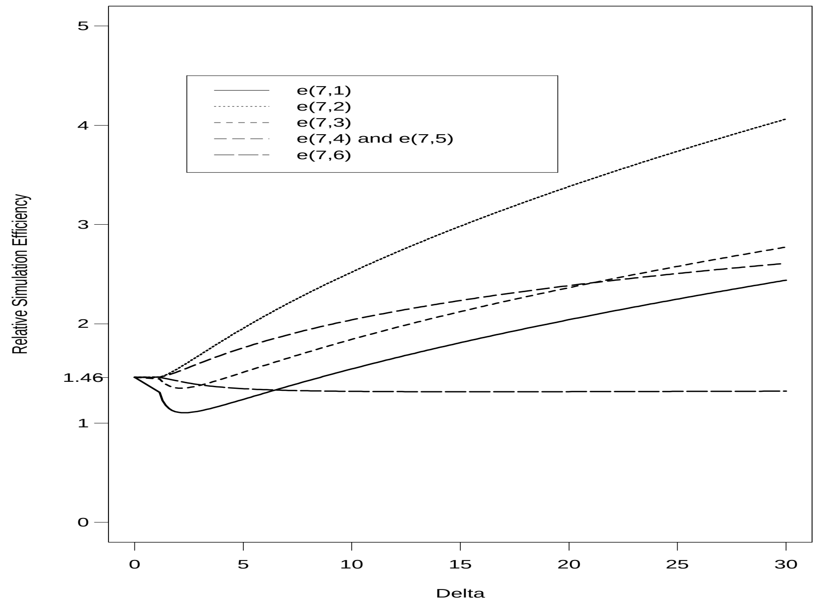

FIGURE 5.2. Relative simulation efficiency plot (II).

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

5.7 Computing Simulation Standard Errors

In Sections 5.2–5.5, IS, BS, PS, and RIS are used to obtain MC estimates of the ratio of the two normalizing constants. In order to assess the simulation accuracy of each estimate, it is important to obtain its associated simulation standard error. In this section, we discuss how to use the asymptotic relative mean-square errors to obtain an approximation of the simulation standard error. Other methods for calculating the simulation standard errors can be found in Section 3.3.

We first start with the importance sampling estimates of r, which are given by (5.2.1) and (5.2.5), respectively. Using (5.2.3) and two independent random samples $\{\boldsymbol{\theta}_{l,1}, \boldsymbol{\theta}_{l,2}, \ldots, \boldsymbol{\theta}_{l,n_l}\}$, $l = 1, 2$, the simulation standard error of $\hat{r}_{\text{IS}_1}$ given in (5.2.1) can be approximated by

$$\text{se}(\hat{r}_{\text{IS}_1}) = \hat{r}_{\text{IS}_1}\left\{\sum_{l=1}^{2}\frac{1}{n_l^2}\sum_{i=1}^{n_l}\left(\frac{q_l(\boldsymbol{\theta}_{l,i})/\hat{c}_l - \pi_l^{\text{I}}(\boldsymbol{\theta}_{l,i})}{\pi_l^{\text{I}}(\boldsymbol{\theta}_{l,i})}\right)^2\right\}^{1/2}, \quad (5.7.1)$$

where $\hat{c}_l = (1/n_l)\sum_{i=1}^{n_l} q_l(\boldsymbol{\theta}_{l,i})/\pi_l^{\mathrm{I}}(\boldsymbol{\theta}_{l,i})$ for $l = 1, 2$. Similarly, using (5.2.6) and $\{\boldsymbol{\theta}_{2,1}, \boldsymbol{\theta}_{2,2}, \dots, \boldsymbol{\theta}_{2,n}\}$, the simulation standard error of $\hat{r}_{\mathrm{IS}_2}$ is given by

$$\mathrm{se}(\hat{r}_{\mathrm{IS}_2}) = \frac{1}{\sqrt{n}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_1(\boldsymbol{\theta}_{2,i}) - \hat{r}_{\mathrm{IS}_2} q_2(\boldsymbol{\theta}_{2,i})}{q_2(\boldsymbol{\theta}_{2,i})}\right)^2\right\}^{1/2}. \tag{5.7.2}$$

Next, we consider the BS estimate $\hat{r}_{\mathrm{BS}}$. We have two approaches to compute $\mathrm{se}(\hat{r}_{\mathrm{BS}})$. Using $\{\boldsymbol{\theta}_{2,1}, \boldsymbol{\theta}_{2,2}, \dots, \boldsymbol{\theta}_{2,n_2}\}$ and $\hat{r}_{\mathrm{BS}}$, an approximation of the simulation standard error is

$$\begin{aligned}\mathrm{se}(\hat{r}_{\mathrm{BS}}) = \frac{\hat{r}_{\mathrm{BS}}}{\sqrt{n s_1 s_2}}\Bigg\{&\frac{1}{n_2}\sum_{i=1}^{n_2} q_1(\boldsymbol{\theta}_{2,i})(s_1 q_1(\boldsymbol{\theta}_{2,i}) + s_2 \hat{r}_{\mathrm{BS}} q_2(\boldsymbol{\theta}_{2,i}))\alpha^2(\boldsymbol{\theta}_{2,i}) \\ &\times \left(\frac{1}{n_2}\sum_{i=1}^{n_2} q_1(\boldsymbol{\theta}_{2,i})\alpha(\boldsymbol{\theta}_{2,i})\right)^{-2} - 1\Bigg\}^{1/2}.\end{aligned} \tag{5.7.3}$$

With $\{\boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{1,2}, \dots, \boldsymbol{\theta}_{1,n_1}\}$ and $\hat{r}_{\mathrm{BS}}$, we obtain

$$\begin{aligned}\mathrm{se}(\hat{r}_{\mathrm{BS}}) = \frac{\hat{r}_{\mathrm{BS}}}{\sqrt{n s_1 s_2}}\Bigg\{&\frac{1}{n_1}\sum_{i=1}^{n_1} q_2(\boldsymbol{\theta}_{2,i})(s_1 q_1(\boldsymbol{\theta}_{1,i}) + s_2 \hat{r}_{\mathrm{BS}} q_2(\boldsymbol{\theta}_{1,i}))\alpha^2(\boldsymbol{\theta}_{1,i}) \\ &\times \left[\hat{r}_{\mathrm{BS}}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} q_2(\boldsymbol{\theta}_{1,i})\alpha(\boldsymbol{\theta}_{1,i})\right)^2\right]^{-1} - 1\Bigg\}^{1/2}.\end{aligned} \tag{5.7.4}$$

In practice, we recommend that one may use (5.7.3) when $n_2 > n_1$ and (5.7.4) when $n_2 < n_1$. When $n_2 = n_1$, one can use either (5.7.3) or (5.7.4). Analogous to $\hat{r}_{\mathrm{BS}}$, using (5.3.6), an approximation of the simulation standard error for the optimal BS estimate $\hat{r}_{\mathrm{BS,opt}}$ can be written as

$$\mathrm{se}(\hat{r}_{\mathrm{BS,opt}}) = \frac{\hat{r}_{\mathrm{BS,opt}}}{\sqrt{n s_1 s_2}}\left[\left\{\frac{1}{n_2}\sum_{i=1}^{n_2}\frac{q_1(\boldsymbol{\theta}_{2,i})}{s_1 q_1(\boldsymbol{\theta}_{2,i}) + s_2 \hat{r}_{\mathrm{BS,opt}} q_2(\boldsymbol{\theta}_{2,i})}\right\}^{-1} - 1\right]^{1/2}. \tag{5.7.5}$$

For PS, since the variance of $\hat{\xi}_{\mathrm{PS}}$ has a closed form, a derivation of the formula for the simulation standard error of $\hat{\xi}_{\mathrm{PS}}$ is straightforward. In particular, the method for IS–version 2 can be exactly applied.

To compute the simulation standard error for a RIS estimate $\hat{r}_{\mathrm{RIS}}$, we write $\pi(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/c_\pi$, where $q(\boldsymbol{\theta})$ is completely known, but c_π is an unknown quantity. Then, we can express the first-order term of $\mathrm{RE}^2(\hat{r}_{\mathrm{RIS}})$ in (5.5.7) as

$$\frac{1}{n}E_\pi\left[\frac{\{\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})\}^2}{\pi^2(\boldsymbol{\theta})}\right] = \frac{1}{n}\left(\frac{c_\pi}{c_1}\right)^2 E_\pi\left[\left\{\frac{q_1(\boldsymbol{\theta}) - r q_2(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\}^2\right]. \tag{5.7.6}$$

Using (5.2.5), a consistent estimate of (c_1/c_π) in (5.7.6) is given by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{q_1(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}, \tag{5.7.7}$$

where $\{\boldsymbol{\theta}_i,\ i=1,2,\ldots,n\}$ is a random sample from $\pi(\boldsymbol{\theta})$. Also, we can use the same random sample from π to obtain a consistent estimate for $E_\pi[\{(q_1(\boldsymbol{\theta})-rq_2(\boldsymbol{\theta}))/q(\boldsymbol{\theta})\}^2]$, which is given by

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{q_1(\boldsymbol{\theta}_i)-\hat{r}_{\mathrm{RIS}}q_2(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}\right\}^2, \tag{5.7.8}$$

where $\hat{r}_{\mathrm{RIS}}$ is defined by (5.5.2). Since q_1, q_2, and q are completely known, (5.7.7) and (5.7.7) are readily computed. Combining (5.7.6), (5.7.7), and (5.7.8) together gives a first-order approximation of the simulation standard error for $\hat{r}_{\mathrm{RIS}}$ as follows:

$$\mathrm{se}(\hat{r}_{\mathrm{RIS}})=\frac{\hat{r}_{\mathrm{RIS}}}{\sqrt{n}}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{q_1(\boldsymbol{\theta}_i)-\hat{r}_{\mathrm{RIS}}q_2(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}\right\}^2\right]^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{q_1(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)}\right]^{-1}. \tag{5.7.9}$$

From the derivation of the approximation of the simulation standard error for an estimate of r, we observe an interesting feature. That is, the same random sample(s) can be used for computing both the estimate of r and its simulation standard error. This feature is important since it indicates that computing the simulation standard error does not require any additional random samples. On the other hand, we also observe that our derivation of the simulation standard error is based on a first-order asymptotic approximation. Hence, one may wonder how accurate this type of approximation is. To examine this, several simulation studies were conducted by Chen and Shao (1997b). Their simulation results indicate that the simulation standard error based on the first-order approximation is indeed quite accurate as long as the MCMC sample size is greater than 1000. However, a suggested MCMC sample size is 5000 or larger to ensure that a reliable approximation of the simulation standard error can be obtained.

5.8 Extensions to Densities with Different Dimensions

5.8.1 Why Different Dimensions?

Kass and Raftery (1995) illustrate a simple problem for testing the two hypotheses H_1 and H_2. Given data D, the Bayes factor is defined by

$$B=\frac{m(D|H_1)}{m(D|H_2)},$$

where the marginal likelihood function

$$m(D|H_l) = \int_{\Omega_l} L(\boldsymbol{\theta}_l|D, H_l)\pi(\boldsymbol{\theta}_l|H_l)\ d\boldsymbol{\theta}_l,$$

$\boldsymbol{\theta}_l$ is a $p_l \times 1$ parameter vector under H_l, $\pi(\boldsymbol{\theta}_l|H_l)$ is the prior density, $L(\boldsymbol{\theta}_l|D, H_l)$ is the likelihood function of $\boldsymbol{\theta}_l$, and Ω_l is the support of the posterior density that is proportional to $L(\boldsymbol{\theta}_l|D, H_l)\pi(\boldsymbol{\theta}_l|H_l)$ for $l = 1, 2$. (See Jeffreys (1961, Chap. 5) for several examples of this simple Bayesian hypothesis testing problem.) Clearly, the Bayes factor B is a ratio of two normalizing constants of two unnormalized densities $L(\boldsymbol{\theta}_l|D, H_l)\pi(\boldsymbol{\theta}_l|H_l)$, $l = 1, 2$, respectively. Note that when $p_1 \neq p_2$, we are dealing with a problem of two different dimensions.

Verdinelli and Wasserman (1996) also consider a similar problem for testing precise null hypotheses using the Bayes factors when nuisance parameters are present. Consider the parameter $(\boldsymbol{\theta}, \boldsymbol{\psi}) \in \Omega \times \boldsymbol{\psi}$, where $\boldsymbol{\psi}$ is a nuisance parameter, and suppose we wish to test the null hypothesis H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus H_1: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Then they obtain the Bayes factor $B = m_0/m$ where $m_0 = \int_\Psi L(\boldsymbol{\theta}_0, \boldsymbol{\psi}|D)\pi_0(\boldsymbol{\psi})\ d\boldsymbol{\psi}$ and $m = \int_{\Omega\times\Psi} L(\boldsymbol{\theta}, \boldsymbol{\psi}|D)\pi(\boldsymbol{\theta}, \boldsymbol{\psi})\ d\boldsymbol{\theta}\ d\boldsymbol{\psi}$ (Jeffreys 1961, Chap. 5). Here $L(\boldsymbol{\theta}, \boldsymbol{\psi}|D)$ is the likelihood function, and $\pi_0(\boldsymbol{\psi})$ and $\pi(\boldsymbol{\theta}, \boldsymbol{\psi})$ 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 $\boldsymbol{\psi}$ and the other density is a function of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$.

5.8.2 General Formulation

From the two illustrative examples given in Section 5.7.1, we can formulate the general problem of computing ratios of two normalizing constants with different dimensions. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)$. Also let $\pi_1(\boldsymbol{\theta})$ be a density which is known up to a normalizing constant:

$$\pi_1(\boldsymbol{\theta}) = \frac{q_1(\boldsymbol{\theta})}{c_1}, \quad \boldsymbol{\theta} \in \Omega_1,$$

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

$$\pi_2(\boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{q_2(\boldsymbol{\theta}, \boldsymbol{\psi})}{c_2}, \quad (\boldsymbol{\theta}, \boldsymbol{\psi}) \in \Theta_2,$$

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

$$\Omega_2 = \{\boldsymbol{\theta} : \exists\ \boldsymbol{\psi} \in R^k \text{ such that } (\boldsymbol{\theta}, \boldsymbol{\psi}) \in \Theta_2\} \tag{5.8.1}$$

and $\Psi(\boldsymbol{\theta}) = \{\boldsymbol{\psi} : (\boldsymbol{\theta}, \boldsymbol{\psi}) \in \Theta_2\}$ for $\boldsymbol{\theta} \in \Omega_2$. Then the ratio of two normalizing constants is defined as $r = c_1/c_2$, which is (5.1.1).

Since the two densities of interest have different dimensions, the MC methods for estimating a ratio of two normalizing constants described in

Sections 5.2–5.5, which include IS, BS, PS, as well as RIS, cannot work directly here. To see this, we consider IS–version 2. The key identity for IS–version 2 is

$$r = \frac{c_1}{c_2} = E_{\pi_2}\left\{\frac{q_1(\boldsymbol{\theta})}{q_2(\boldsymbol{\theta}, \boldsymbol{\psi})}\right\},$$

which does not hold in general, unless under certain conditions such as $\int_{\Psi(\boldsymbol{\theta})} d\boldsymbol{\psi} = 1$ for all $\boldsymbol{\theta} \in \Omega_2$. Since IS–version 1 described in Section 5.2.1 depends highly on the choices of the two IS densities, we consider only IS–version 2 in this section. It is inconvenient here to construct a path to link π_1 and π_2 due to different dimensionality. Therefore, it is not feasible to apply PS for problems with different dimensions. On the other hand, if the conditional density of $\boldsymbol{\psi}$ given $\boldsymbol{\theta}$ is completely known, the problem of different dimensions disappears. This can be explained as follows. Let $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$ denote the conditional density of $\boldsymbol{\psi}$ given $\boldsymbol{\theta}$,

$$\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta}) = \frac{q_2(\boldsymbol{\theta}, \boldsymbol{\psi})}{\int_{\Psi(\boldsymbol{\theta})} q_2(\boldsymbol{\theta}, \boldsymbol{\psi}^*)\ d\boldsymbol{\psi}^*}, \quad \boldsymbol{\psi} \in \Psi(\boldsymbol{\theta}) \text{ for } \boldsymbol{\theta} \in \Omega_2.$$

Then

$$\pi_2(\boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{q_2(\boldsymbol{\theta}, \boldsymbol{\psi})}{c_2} = \frac{q_2(\boldsymbol{\theta})}{c_2} \cdot \pi_2(\boldsymbol{\psi}|\boldsymbol{\theta}),$$

where $q_2(\boldsymbol{\theta})$ is a completely known unnormalized marginal density of $\boldsymbol{\theta}$. Thus, one can directly apply the same-dimension identities to the problem that only involves $q_1(\boldsymbol{\theta})$ and $q_2(\boldsymbol{\theta})$. Therefore, we assume that $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$ is known only up to a normalizing constant

$$c(\boldsymbol{\theta}) = \int_{\Psi(\boldsymbol{\theta})} q_2(\boldsymbol{\theta}, \boldsymbol{\psi})\ d\boldsymbol{\psi}.$$

This assumption will be made throughout this section. Since $c(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$, the different-dimension problem is challenging and difficult.

5.8.3 Extensions of the Previous Monte Carlo Methods

Although we cannot directly use IS, BS, and RIS for estimating r since $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}, \boldsymbol{\psi})$ are defined on two different dimensional parameter spaces, this different 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

$$q_1^*(\boldsymbol{\theta}, \boldsymbol{\psi}) = q_1(\boldsymbol{\theta}) w(\boldsymbol{\psi}|\boldsymbol{\theta})$$

and

$$\pi_1^*(\boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{q_1^*(\boldsymbol{\theta}, \boldsymbol{\psi})}{c_1^*}, \tag{5.8.2}$$

where $w(\boldsymbol{\psi}|\boldsymbol{\theta})$ is a completely known weight density function so that

$$\int_{\boldsymbol{\psi}(\boldsymbol{\theta})} w(\boldsymbol{\psi}|\boldsymbol{\theta})\ d\boldsymbol{\psi} = 1,$$

and c_1^* is the normalizing constant of $\pi_1^*(\boldsymbol{\theta}, \boldsymbol{\psi})$. 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 $\pi_1^*(\boldsymbol{\theta}, \boldsymbol{\psi})$ and $\pi_2(\boldsymbol{\theta}, \boldsymbol{\psi})$. Therefore, we can directly apply the IS, BS, and RIS identities given in (5.2.4), (5.3.1), and (5.5.1) on the $(\boldsymbol{\theta}, \boldsymbol{\psi})$ space for estimating r. We summarize the IS, BS, and RIS estimators of r as follows.

First, we consider IS–version 2. Assume $\Omega_1 \subset \Omega_2$. Let $\{(\boldsymbol{\theta}_{2,1}, \boldsymbol{\psi}_{2,1}), \ldots, (\boldsymbol{\theta}_{2,n}, \boldsymbol{\psi}_{2,n})\}$ be a random sample from π_2. Then, on the $(\boldsymbol{\theta}, \boldsymbol{\psi})$ space, using the IS identity

$$r = \frac{c_1}{c_2} = E_{\pi_2}\left\{\frac{q_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta})}{q_2(\boldsymbol{\theta}, \boldsymbol{\psi})}\right\},$$

and r can be estimated by

$$\hat{r}_{\text{IS}}(w) = \frac{1}{n}\sum_{i=1}^{n} \frac{q_1(\boldsymbol{\theta}_{2,i})w(\boldsymbol{\psi}_{2,i}|\boldsymbol{\theta}_{2,i})}{q_2(\boldsymbol{\theta}_{2,i}, \boldsymbol{\psi}_{2,i})}. \tag{5.8.3}$$

Second, we extend BS. Using the BS identity given in (5.3.1) on the $(\boldsymbol{\theta}, \boldsymbol{\psi})$ space, we have

$$r = \frac{c_1}{c_2} = \frac{E_{\pi_2}\{q_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta})\alpha(\boldsymbol{\theta}, \boldsymbol{\psi})\}}{E_{\pi_1^*}\{q_2(\boldsymbol{\theta}, \boldsymbol{\psi})\alpha(\boldsymbol{\theta}, \boldsymbol{\psi})\}},$$

where $\pi_1^*(\boldsymbol{\theta}, \boldsymbol{\psi})$ is defined by (5.8.2) with the support of $\boldsymbol{\theta}_1 = \{(\boldsymbol{\theta}, \boldsymbol{\psi}) : \boldsymbol{\psi} \in \Psi_1(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Omega_1\}$ and $\alpha(\boldsymbol{\theta}, \boldsymbol{\psi})$ is an arbitrary function defined on $\Theta_1 \cap \Theta_2$ such that

$$0 < \left|\int_{\Theta_1 \cap \Theta_2} \alpha(\boldsymbol{\theta}, \boldsymbol{\psi})q_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta})q_2(\boldsymbol{\theta}, \boldsymbol{\psi})\ d\boldsymbol{\theta}\ d\boldsymbol{\psi}\right| < \infty.$$

Then using two random samples $\{(\boldsymbol{\theta}_{l,1}), \boldsymbol{\psi}_{l,1}), \ldots, (\boldsymbol{\theta}_{l,n_l}, \boldsymbol{\psi}_{l,n_l})\}$, $l = 1, 2$, from π_1^* and π_2, respectively, we obtain a consistent estimator of r as

$$\hat{r}_{\text{BS}}(w, \alpha) = \frac{n_2^{-1}\sum_{i=1}^{n_2} q_1(\boldsymbol{\theta}_{2,i})w(\boldsymbol{\psi}_{2,i}|\boldsymbol{\theta}_{2,i})\alpha(\boldsymbol{\theta}_{2,i}, \boldsymbol{\psi}_{2,i})}{n_1^{-1}\sum_{i=1}^{n_1} q_2(\boldsymbol{\theta}_{1,i}, \boldsymbol{\psi}_{1,i})\alpha(\boldsymbol{\theta}_{1,i}, \boldsymbol{\psi}_{1,i})}. \tag{5.8.4}$$

Finally, we generalize RIS. Using the RIS identity given in (5.5.1) on the $(\boldsymbol{\theta}, \boldsymbol{\psi})$ space, we have

$$r = \frac{c_1}{c_2} = \frac{E_{\pi}\{q_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta})/\pi(\boldsymbol{\theta}, \boldsymbol{\psi})\}}{E_{\pi}\{q_2(\boldsymbol{\theta}, \boldsymbol{\psi})/\pi(\boldsymbol{\theta}, \boldsymbol{\psi})\}}, \tag{5.8.5}$$

where π is an arbitrary density over $\boldsymbol{\theta}$ such that $\pi(\boldsymbol{\theta}, \boldsymbol{\psi}) > 0$ for $(\boldsymbol{\theta}, \boldsymbol{\psi}) \in \boldsymbol{\theta} = \Theta_1 \cup \Theta_2$. We mention that in (5.8.5), it is not necessary for π to be completely known, i.e., π can be known up to an unknown normalizing constant:

$$\pi(\boldsymbol{\theta}, \boldsymbol{\psi}) = \frac{q(\boldsymbol{\theta}, \boldsymbol{\psi})}{c}.$$

Given a random sample $\{(\boldsymbol{\theta}_1, \boldsymbol{\psi}_1), \ldots, (\boldsymbol{\theta}_n, \boldsymbol{\psi}_n)\}$ from π, the RIS estimator of r is

$$\hat{r}_{\text{RIS}}(w, \pi) = \frac{\sum_{i=1}^n q_1(\boldsymbol{\theta}_i) w(\boldsymbol{\psi}_i | \boldsymbol{\theta}_i) / \pi(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i)}{\sum_{i=1}^n q_2(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i) / \pi(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i)}. \tag{5.8.6}$$

5.8.4 Global Optimal Estimators

From (5.8.3), (5.8.4), and (5.8.5), it can be observed that all three estimators, namely, $\hat{r}_{\text{IS}}(w)$, $\hat{r}_{\text{BS}}(w, \alpha)$, and $\hat{r}_{\text{RIS}}(w, \pi)$, depend on w, while $\hat{r}_{\text{BS}}(w, \alpha)$ and $\hat{r}_{\text{RIS}}(w, \pi)$ further depend on α and π, respectively. Thus, a natural question is what are the optimal choices of these parameters? To address this question, we use a conventional criterion for optimality. An estimator is optimal if it minimizes the asymptotic relative mean-square error.

We first introduce some notation. Let $\pi_{21}(\boldsymbol{\theta})$ be the marginal density of $\boldsymbol{\theta}$ defined on Ω_2. Then

$$\pi_{21}(\boldsymbol{\theta}) = \int_{\Psi(\boldsymbol{\theta})} \frac{q_2(\boldsymbol{\theta}, \boldsymbol{\psi})}{c_2} \, d\boldsymbol{\psi} \quad \text{for } \boldsymbol{\theta} \in \Psi(\boldsymbol{\theta}),$$

where Ω_2 and $\Psi(\boldsymbol{\theta})$ are defined in (5.8.1). Let $\hat{r}$ denote the estimator of r. Then the asymptotic relative mean-square error (ARE) is defined as

$$\text{ARE}^2(\hat{r}) = \lim_{n \to \infty} n \, \text{RE}^2(\hat{r}),$$

where $\text{RE}^2(\hat{r})$ is defined in (5.2.2).

For a given weight density function $w(\boldsymbol{\psi}|\boldsymbol{\theta})$ on the $(\boldsymbol{\theta}, \boldsymbol{\psi})$ space, the generalized version of the REs and AREs for $\hat{r}_{\text{IS}}(w)$, $\hat{r}_{\text{BS}}(w, \alpha)$, and $\hat{r}_{\text{RIS}}(w, \pi)$ can be directly obtained from (5.2.6), (5.3.4), and (5.5.7). The results are summarized in the following three lemmas:

Lemma 5.8.1 *Assume $\Omega_1 \subset \Omega_2$ and*

$$\int_{\Theta_2} \{q_1^2(\boldsymbol{\theta}) w^2(\boldsymbol{\psi}|\boldsymbol{\theta}) / q_2(\boldsymbol{\theta}, \boldsymbol{\psi})\} \, d\boldsymbol{\theta} \, d\boldsymbol{\psi} < \infty.$$

Then

$$\text{RE}^2(\hat{r}_{\text{IS}}(w)) = \frac{1}{r^2} \text{Var}(\hat{r}_{\text{IS}}(w)) = \frac{1}{n} \left[\int_{\Theta_2} \frac{\pi_1^2(\boldsymbol{\theta}) w^2(\boldsymbol{\psi}|\boldsymbol{\theta})}{\pi_2(\boldsymbol{\theta}, \boldsymbol{\psi})} \, d\boldsymbol{\theta} \, d\boldsymbol{\psi} - 1 \right]$$

and

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

Lemma 5.8.2 *Let* $n = n_1 + n_2$ *and* $s_{l,n} = n_l/n$ *for* $l = 1, 2$. *Assume that* $s_l = \lim_{n\to\infty} s_{l,n} > 0$ $(l = 1, 2)$, $E_{\pi_2}\{q_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta})\alpha(\boldsymbol{\theta},\boldsymbol{\psi})\}^2 < \infty$, *and*

$$E_{\pi_1^*}\{\big(q_2(\boldsymbol{\theta},\boldsymbol{\psi})\alpha(\boldsymbol{\theta},\boldsymbol{\psi})\big)^2 + 1/\big(q_2(\boldsymbol{\theta},\boldsymbol{\psi})\alpha(\boldsymbol{\theta},\boldsymbol{\psi})\big)^2\} < \infty.$$

Then

$$\begin{aligned}
&\mathrm{RE}^2(\hat{r}_{\mathrm{BS}}(w,\alpha)) \\
&\quad = \frac{1}{ns_{1,n}s_{2,n}}\Bigg\{\bigg(\int_{\Theta_1\cap\Theta_2} \pi_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta})\pi_2(\boldsymbol{\theta},\boldsymbol{\psi})\alpha(\boldsymbol{\theta},\boldsymbol{\psi})\ d\boldsymbol{\theta}\ d\boldsymbol{\psi}\bigg)^{-2} \\
&\qquad \times\bigg(\int_{\Theta_1\cap\Theta_2} \pi_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta})\pi_2(\boldsymbol{\theta},\boldsymbol{\psi})(s_{1,n}\pi_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta}) \\
&\qquad + s_{2,n}\pi_2(\boldsymbol{\theta},\boldsymbol{\psi}))\alpha^2(\boldsymbol{\theta},\boldsymbol{\psi})\ d\boldsymbol{\theta}\ d\boldsymbol{\psi}\bigg) - 1\Bigg\} + o\Big(\frac{1}{n}\Big)
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{ARE}^2(\hat{r}_{\mathrm{BS}}(w,\alpha)) \\
&\quad = \frac{1}{s_1s_2}\Bigg\{\bigg(\int_{\Theta_1\cap\Theta_2} \pi_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta})\pi_2(\boldsymbol{\theta},\boldsymbol{\psi})\alpha(\boldsymbol{\theta},\boldsymbol{\psi})\ d\boldsymbol{\theta}\ d\boldsymbol{\psi}\bigg)^{-2} \\
&\qquad \times\bigg(\int_{\Theta_1\cap\Theta_2} \pi_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta})\pi_2(\boldsymbol{\theta},\boldsymbol{\psi})(s_1\pi_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta}) \\
&\qquad + s_2\pi_2(\boldsymbol{\theta},\boldsymbol{\psi}))\alpha^2(\boldsymbol{\theta},\boldsymbol{\psi})\ d\boldsymbol{\theta}\ d\boldsymbol{\psi}\bigg) - 1\Bigg\}.
\end{aligned}$$

Lemma 5.8.3 *Assume that* $E_\pi\{\big(\pi_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta},\boldsymbol{\psi})\big)/\pi(\boldsymbol{\theta},\boldsymbol{\psi})\}^2 < \infty$ *and*

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

Then

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

and

$$\mathrm{ARE}^2(\hat{r}_{\mathrm{RIS}}(w,\pi)) = \int_{\Theta_1\cup\Theta_2} \frac{(\pi_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta},\boldsymbol{\psi}))^2}{\pi(\boldsymbol{\theta},\boldsymbol{\psi})}\ d\boldsymbol{\theta}\ d\boldsymbol{\psi}. \quad (5.8.7)$$

The proofs of these three lemmas are left as exercises. Now, we present a general result that will be needed for deriving optimal choices of $w(\boldsymbol{\psi}|\boldsymbol{\theta})$, $\alpha(\boldsymbol{\theta},\boldsymbol{\psi})$, and $\pi(\boldsymbol{\theta},\boldsymbol{\psi})$ for IS, BS, and RIS.

Theorem 5.8.1 *Assume there exist functions* h *and* g *such that*:

(I) $\mathrm{ARE}^2(\hat{r}) \geq h\{E_{\pi_2}[g(\pi_1(\boldsymbol{\theta})w(\boldsymbol{\psi}|\boldsymbol{\theta})/\pi_2(\boldsymbol{\theta},\boldsymbol{\psi}))]\}$;

(II) *either* (i) *or* (ii) *holds*:

(i) *h is an increasing function and g is convex; and*

(ii) *h is a decreasing function and g is concave.*

Then for an arbitrary $w(\boldsymbol{\psi}|\boldsymbol{\theta})$ defined on $\Psi(\boldsymbol{\theta})$ or $\Psi_1(\boldsymbol{\theta})$,

$$\mathrm{ARE}^2(\hat{r}) \geq h\{E_{\pi_{21}}[g(\pi_1(\boldsymbol{\theta})/\pi_{21}(\boldsymbol{\theta}))]\}. \tag{5.8.8}$$

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

The proof of (5.8.8) follows from assumptions (i) and (ii) and Jensen's inequality and is thus left as an exercise.

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

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

We are led to the following theorem:

Theorem 5.8.2 *The optimal choices are*

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

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

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

and

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

The optimal AREs *are*

$$\mathrm{ARE}^2(\hat{r}_{\mathrm{OIS}}) = \int_{\Omega_1} \frac{\pi_1^2(\boldsymbol{\theta})}{\pi_{21}(\boldsymbol{\theta})}\, d\boldsymbol{\theta} - 1, \tag{5.8.9}$$

$$\mathrm{ARE}^2(\hat{r}_{\mathrm{GOBS}}) = \frac{1}{s_1 s_2}\left\{\left(\int_{\Omega_1\cap\Omega_2} \frac{\pi_1(\boldsymbol{\theta})\pi_{21}(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_{21}(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right)^{-1} - 1\right\}, \tag{5.8.10}$$

and

$$\mathrm{ARE}^2(\hat{r}_{\mathrm{GORIS}}) = \left[\int_{\Omega_1 \cup \Omega_2} |\pi_1(\boldsymbol{\theta}) - \pi_{21}(\boldsymbol{\theta})| \, d\boldsymbol{\theta} \right]^2 . \tag{5.8.11}$$

The proof of Theorem 5.8.2 is given in the Appendix. It is interesting to mention that the optimal choices of w are the same for all three MC methods (IS, BS, and RIS). The optimal w is the conditional density $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$. These results are consistent with our intuitive guess. We also note that although IS is a special case of BS with $\alpha(\boldsymbol{\theta}, \boldsymbol{\psi}) = 1/\pi_2(\boldsymbol{\theta}, \boldsymbol{\psi})$, the proof for the optimal choice of w for IS cannot simply follow from that of BS because this α is not α_{opt}. With the global optimal choices of w, α, and π, the (asymptotic) relative mean-square errors (AREs) for all three methods depend only on $\pi_1(\boldsymbol{\theta})$ and $\pi_{21}(\boldsymbol{\theta})$, which implies that the extra parameter $\boldsymbol{\psi}$ 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 conclusions are valid only if the optimal solutions can be implemented in practice, since $w(\boldsymbol{\psi}|\boldsymbol{\theta})$ is not completely known. We will discuss implementation issues in the next subsection.

5.8.5 Implementation Issues

In many practical problems, a closed-form of the conditional density $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$ is not available especially when $\Psi(\boldsymbol{\theta})$ is a constrained parameter space (see Chapter 4 for an explanation). Therefore, evaluating ratios of normalizing constants for densities with different dimensions is a challenging problem. Here we present detailed implementation schemes for obtaining $\hat{r}_{\mathrm{OIS}}$, $\hat{r}_{\mathrm{GOBS}}$, and $\hat{r}_{\mathrm{GORIS}}$. We consider our implementation procedures for $k = 1$ and $k > 1$ separately.

First, we consider $k = 1$. In this case,

$$\pi_2(\psi|\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta}, \psi)}{c(\boldsymbol{\theta})},$$

where $c(\boldsymbol{\theta}) = \int_{\Psi(\boldsymbol{\theta})} q(\boldsymbol{\theta}, \psi') \, d\psi'$. Note that the integral in $c(\boldsymbol{\theta})$ is only one dimensional. Since one-dimensional numerical integration methods are well developed and computationally fast, one can use, for example, the IMSL subroutines QDAG or QDAGI; or as Verdinelli and Wasserman (1995) suggest, one can use a grid $\{\psi_1^*, \dots, \psi_M^*\}$ that includes all sample points ψ_1, $\dots$, ψ_n and then use the trapezoidal rule to approximate the integral. In the following three algorithms, we assume that $c(\boldsymbol{\theta})$ will be calculated or approximated by a numerical integration method. Detailed implementation schemes for obtaining $\hat{r}_{\mathrm{OIS}}$, $\hat{r}_{\mathrm{GOBS}}$ and $\hat{r}_{\mathrm{GORIS}}$ are presented as follows.

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

ALGORITHM OIS

Step 1. Generate a random sample $\{(\boldsymbol{\theta}_i, \psi_i),\ i = 1, 2, \ldots, n\}$ from $\pi_2(\boldsymbol{\theta}, \psi)$.

Step 2. Calculate $c(\boldsymbol{\theta}_i)$ and compute

$$\hat{r}_{\text{OIS}} = \frac{1}{n} \sum_{i=1}^{n} \frac{q_1(\boldsymbol{\theta}_i)}{c(\boldsymbol{\theta}_i)}. \tag{5.8.12}$$

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

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

ALGORITHM GOBS

Step 1. Generate random samples $\{(\boldsymbol{\theta}_{l,i}, \psi_{l,i}),\ i = 1, 2 \ldots, n_l\}$, $l = 1, 2$, $(n_1 + n_2 = n)$ as follows:

(i) Generate $\{\boldsymbol{\theta}_{1,i},\ i = 1, 2, \ldots, n_1\}$ from $\pi_1(\boldsymbol{\theta})$ and then generate $\{\boldsymbol{\theta}_{2,i},\ l = 1, 2, \ldots, n_2\}$ from the marginal distribution of $\boldsymbol{\theta}$ with respect to $\pi_2(\boldsymbol{\theta}, \psi)$.

(ii) Generate $\psi_{l,i}$ independently from $\pi_2(\psi|\boldsymbol{\theta}_{l,i})$ for $i = 1, 2, \ldots, n_l$ and $l = 1, 2$.

Step 2. Calculate $c(\boldsymbol{\theta}_{l,i})$ and set $\hat{r}_{\text{GOBS}}$ to be the unique zero root of the "score" function

$$\begin{aligned} S(r) = &\sum_{i=1}^{n_1} \frac{s_2 r}{s_1 q_1(\boldsymbol{\theta}_{1,i})/c(\boldsymbol{\theta}_{1,i}) + s_2 r} \\ &- \sum_{i=1}^{n_2} \frac{s_1 q_1(\boldsymbol{\theta}_{2,i})/c(\boldsymbol{\theta}_{2,i})}{s_1 q_1(\boldsymbol{\theta}_{2,i})/c(\boldsymbol{\theta}_{2,i}) + s_2 r}. \end{aligned} \tag{5.8.13}$$

In Step 1, generating the $\boldsymbol{\theta}_{ij}$ or the ψ_{ij} does not require knowing the normalizing constants since we can use, for example, a rejection/acceptance, Metropolis, or Gibbs sampler method. In Step 2, $\hat{r}_{\text{GOBS}}$ can also be obtained by using an iterative method described in Section 5.3. This method can be implemented as follows. Starting with an initial guess of r, $\hat{r}^{(0)}$, at the $(t+1)^{\text{th}}$ iteration, we compute

$$\hat{r}^{(t+1)} = \left\{ \frac{1}{n_2} \sum_{i=1}^{n_2} \frac{q_1(\boldsymbol{\theta}_{2,i})/c(\boldsymbol{\theta}_{2,i})}{s_1 q_1(\boldsymbol{\theta}_{2,i})/c(\boldsymbol{\theta}_{2,i}) + s_2 \hat{r}^{(t)}} \right\} \times \left\{ \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1/c(\boldsymbol{\theta}_{1,i})}{s_1 q_1(\boldsymbol{\theta}_{1,i})/c(\boldsymbol{\theta}_{1,i}) + s_2 \hat{r}^{(t)}} \right\}^{-1}.$$

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

For RIS, we obtain an approximate $\hat{r}_{\text{GORIS}}$, denoted by $\hat{r}^*_{\text{GORIS}}$, by a two-stage procedure developed Section 5.5.2.

Algorithm GORIS

Step 1. Let $\pi(\boldsymbol{\theta}, \psi)$ be an arbitrary (known up to a normalizing constant) density over $\boldsymbol{\theta}$ such that $\pi(\boldsymbol{\theta}, \psi) > 0$ for $(\boldsymbol{\theta}, \psi) \in \boldsymbol{\theta}$. (For example, $\pi(\boldsymbol{\theta}, \psi) = \pi_2(\boldsymbol{\theta}, \psi)$.) Generate a random sample $\{(\boldsymbol{\theta}_i, \psi_i),\ i = 1, 2, \ldots, n\}$ from π. Calculate $c(\boldsymbol{\theta}_i)$ and compute

$$\tau_{n_1} = \frac{\sum_{i=1}^{n_1} q_1(\boldsymbol{\theta}_i) q_2(\boldsymbol{\theta}_i, \psi_i)/[c(\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i, \psi_i)]}{\sum_{i=1}^{n_1} q_2(\boldsymbol{\theta}_i, \psi_i)/\pi(\boldsymbol{\theta}_i, \psi_i)}. \tag{5.8.14}$$

Step 2. Let

$$\pi^*_{n_1}(\boldsymbol{\theta}, \psi) = \frac{|q_1(\boldsymbol{\theta})\pi_2(\psi|\boldsymbol{\theta}) - \tau_{n_1} q_2(\boldsymbol{\theta}, \psi)|}{\int_{\boldsymbol{\theta}} |q_1(\boldsymbol{\theta}')\pi_2(\psi'|\boldsymbol{\theta}') - \tau_{n_1} q_2(\boldsymbol{\theta}', \psi')|\, d\boldsymbol{\theta}'\, d\psi'}.$$

Then, take a random sample $\{(\boldsymbol{\vartheta}_i, \varphi_i),\ i = 1, 2, \ldots, n_2\}$ from $\pi^*_{n_1}$ $(n_1 + n_2 = n)$.

Step 3. Calculate $c(\boldsymbol{\vartheta}_i)$ and compute

$$\hat{r}^*_{\text{GORIS}} = \frac{\sum_{i=1}^{n_2} q_1(\boldsymbol{\vartheta}_i)/|q_1(\boldsymbol{\vartheta}_i) - \tau_{n_1} c(\boldsymbol{\vartheta}_i)|}{\sum_{i=1}^{n_2} c(\boldsymbol{\vartheta}_i)/|q_1(\boldsymbol{\vartheta}_i) - \tau_{n_1} c(\boldsymbol{\vartheta}_i)|}. \tag{5.8.15}$$

Similar to Theorem 5.5.5, we can prove that $\hat{r}^*_{\text{GORIS}}$ has the same asymptotic relative mean-square error as $\hat{r}_{\text{GORIS}}$ as long as $n_1 \to \infty$ and $n_2 \to \infty$. The most expensive/difficult part of Algorithm GORIS is Step 2. There are two possible approaches to sample $(\boldsymbol{\vartheta}_i, \varphi_i)$ from $\pi^*_{n_1}$. The first approach is the random-direction interior-point (RDIP) sampler given in Section 2.8. The RDIP sampler requires only that $|q_1(\boldsymbol{\theta})\pi_2(\psi|\boldsymbol{\theta}) - \tau_{n_1} q_2(\boldsymbol{\theta}, \psi)|$ can be computed at any point $(\boldsymbol{\theta}, \psi)$. 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 $\pi_2(\boldsymbol{\theta}, \psi)$ has a tail as heavy as the one of $q_1(\boldsymbol{\theta})\pi_2(\psi|\boldsymbol{\theta})$, then one can simply choose $\pi_2(\boldsymbol{\theta}, \psi)$ as a proposal density. Compared to Algorithms OIS and GOBS, Algorithm GORIS requires an evaluation of $c(\boldsymbol{\theta})$ in the sampling step; therefore, Algorithm GORIS is more expensive.

Second, we consider $k > 1$. In this case, the integral in $c(\boldsymbol{\theta})$ is multidimensional. Therefore, simple numerical integration methods might not be feasible. Instead of directly computing $c(\boldsymbol{\theta})$ in the case of $k = 1$, we develop MC schemes to estimate $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$. However, the basic structures of the implementation algorithms are similar to those for $k = 1$. Thus, in the following presentation, we mainly focus on how to estimate or approximate $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$. We propose "exact" and "approximate" approaches.

We start with an "exact" approach. Using the notation of Schervishand Carlin (1992), we let $\boldsymbol{\psi}^* = (\psi_1^*, \ldots, \psi_k^*)$, $\boldsymbol{\psi}^{*(j)} = (\psi_1, \ldots, \psi_j, \psi_{j+1}^*, \ldots, \psi_k^*)$, and $\boldsymbol{\psi}^{*(k)} = \boldsymbol{\psi}$. We denote a "one-step Gibbs transition" density as

$$\pi_2^{(j)}(\boldsymbol{\psi}|\boldsymbol{\theta}) = \pi_2(\psi_j|\psi_1, \ldots, \psi_{j-1}, \psi_{j+1}, \ldots, \psi_k, \boldsymbol{\theta})$$

and a "transition kernel" as

$$T(\boldsymbol{\psi}^*, \boldsymbol{\psi}|\boldsymbol{\theta}) = \prod_{j=1}^{k} \pi_2^{(j)}(\boldsymbol{\psi}^{*(j)}|\boldsymbol{\theta}).$$

Then we have the following key identity:

$$\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta}) = \int_{\Psi(\boldsymbol{\theta})} T(\boldsymbol{\psi}', \boldsymbol{\psi}|\boldsymbol{\theta})\pi_2(\boldsymbol{\psi}'|\boldsymbol{\theta})\ d\boldsymbol{\psi}'.$$

Now we can obtain an MC estimator of $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$ by

$$\hat{\pi}_2(\boldsymbol{\psi}|\boldsymbol{\theta}) = \frac{1}{m} \sum_{l=1}^{m} T(\boldsymbol{\psi}_l, \boldsymbol{\psi}|\boldsymbol{\theta}), \tag{5.8.16}$$

where $\{\boldsymbol{\psi}_l,\ l = 1, 2, \ldots, m\}$ is a random sample from $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$. The above method is 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 numerical integration method, and sometimes some of the one-dimensional conditional densities are even analytically available or easy to compute. Therefore, (5.8.16) is advantageous. In (5.8.16), sampling from $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$ does not require knowing the normalizing constant $c(\boldsymbol{\theta})$ and convergence of $\hat{\pi}_2(\boldsymbol{\psi}|\boldsymbol{\theta})$ to $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$ is expected to be rapid. Algorithms OIS, GOBS, and GORIS for $k > 1$ are similar to the ones for $k = 1$. We only need the following minor adjustment. Generate $\boldsymbol{\psi}_l$, $l = 1, 2, \ldots, m$, from $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta}_i)$, $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta}_{ij})$, or $\pi_2(\boldsymbol{\psi}|\boldsymbol{\vartheta}_i)$ and compute $\hat{\pi}_2(\boldsymbol{\psi}_i|\boldsymbol{\theta}_i)$, $\hat{\pi}_2(\boldsymbol{\psi}_{ij}|\boldsymbol{\theta}_{ij})$, or $\hat{\pi}_2(\boldsymbol{\varphi}_i \mid \boldsymbol{\vartheta}_i)$ by using (5.8.16). Then, for OIS and GOBS, instead of (5.8.12) and (5.8.13), we use

$$\hat{r}_{\text{OIS}} = \frac{1}{n} \sum_{i=1}^{n} \frac{q_1(\boldsymbol{\theta}_i)\hat{\pi}_2(\boldsymbol{\psi}_i|\boldsymbol{\theta}_i)}{q_2(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i)} \tag{5.8.17}$$

and

$$S(r) = \sum_{i=1}^{n_1} \frac{s_2 r q_2(\boldsymbol{\theta}_{1,i}, \boldsymbol{\psi}_{1,i})}{s_1 q_1(\boldsymbol{\theta}_{1,i}) \hat{\pi}_2(\boldsymbol{\psi}_{1,i} | \boldsymbol{\theta}_{1,i}) + s_2 r q_2(\boldsymbol{\theta}_{1,i}, \boldsymbol{\psi}_{1,i})} - \sum_{i=1}^{n_2} \frac{s_1 q_1(\boldsymbol{\theta}_{2,i}) \hat{\pi}_2(\boldsymbol{\psi}_{2,i} | \boldsymbol{\theta}_{2,i})}{s_1 q_1(\boldsymbol{\theta}_{2,i}) \hat{\pi}_2(\boldsymbol{\psi}_{2,i} | \boldsymbol{\theta}_{2,i}) + s_2 r q_2(\boldsymbol{\theta}_{2,i}, \boldsymbol{\psi}_{2,i})}. \tag{5.8.18}$$

For GORIS, instead of (5.8.14) and (5.8.15), we use

$$\tau_{n_1} = \frac{\sum_{i=1}^{n_1} q_1(\boldsymbol{\theta}_i) \hat{\pi}_2(\boldsymbol{\psi}_i | \boldsymbol{\theta}_i) / \pi(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i)}{\sum_{i=1}^{n_1} q_2(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i) / \pi(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i)} \tag{5.8.19}$$

and

$$\hat{r}^*_{\text{GORIS}} = \frac{\sum_{i=1}^{n_2} q_1(\boldsymbol{\vartheta}_i) \hat{\pi}_2(\boldsymbol{\varphi}_i | \boldsymbol{\vartheta}_i) / |q_1(\boldsymbol{\vartheta}_i) \hat{\pi}_2(\boldsymbol{\varphi}_i | \boldsymbol{\vartheta}_i) - \tau_{n_1} q_2(\boldsymbol{\vartheta}_i, \boldsymbol{\varphi}_i)|}{\sum_{i=1}^{n_2} q_2(\boldsymbol{\vartheta}_i, \boldsymbol{\varphi}_i) / |q_1(\boldsymbol{\vartheta}_i) \hat{\pi}_2(\varphi_i | \boldsymbol{\vartheta}_i) - \tau_{n_1} q_2(\boldsymbol{\vartheta}_i, \boldsymbol{\varphi}_i)|}. \tag{5.8.20}$$

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 as $m \to \infty$.

Finally, we briefly introduce an "approximate" approach that requires less computational effort. Mainly, one needs to find a completely known density $w^*(\boldsymbol{\psi}|\boldsymbol{\theta})$ that has a shape similar to $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})$. The details of how to find a good $w^*(\boldsymbol{\psi}|\boldsymbol{\theta})$ are given in Section 4.3. When a good $w^*(\boldsymbol{\psi}|\boldsymbol{\theta})$ is chosen, we simply replace $\hat{\pi}_2$ by $w^*(\boldsymbol{\psi}|\boldsymbol{\theta})$ in (5.8.17), (5.8.18), (5.8.19), and (5.8.20) and then Algorithms OIS, GOBS, and GORIS give approximate $\hat{r}_{\text{OIS}}$, $\hat{r}_{\text{GOBS}}$, and $\hat{r}_{\text{GORIS}}$.

Chen and Shao (1997b) use two examples to illustrate the methodology as well as the implementation algorithms developed in this section. In their examples, they implement 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 this section may be more attractive since they require much less computational effort. 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 $n_1/n_2 \approx \frac{1}{4}$.

Next, we present an example for testing departures from normality to empirically examine the performance of the OIS, GOBS, and GORIS algorithms.

Example 5.1. Testing departures from normality. As an illustration of our implementation algorithms developed in Section 5.8.5 for $k = 1$, we consider an example given in Section 3.2 of Verdinelli and Wasserman (1995). Suppose that we have observations $y_1, \ldots, y_N$ and we want to test whether the sampling distribution is normal or heavier tailed. We use the Student t distribution with ν degrees of freedom for the data. Using the 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 $\boldsymbol{\theta} = (\mu, \sigma)$, where μ and σ are location and scale parameters and denote $\bar{y}$ and s^2 to be the sample mean and the sample variance of $y_1, \ldots, y_N$. Then using exactly the same choices of priors as in Verdinelli and Wasserman (1995), i.e., $\pi_0(\boldsymbol{\theta}) \propto 1/\sigma$, and independently $\pi_0(\psi) \propto 1$, we have the posteriors denoted by $\pi_1(\boldsymbol{\theta})$ under the null hypothesis and $\pi_2(\boldsymbol{\theta}, \psi)$ under the alternative hypothesis:

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

where

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

and

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

Thus, the Bayes factor is $r = c_1/c_2$. It is easy to see that $\boldsymbol{\theta}$ is two dimensional ($p = 2$) and ψ is one dimensional ($k = 1$).

Now we apply Algorithms OIS, GOBS, and GORIS given in Section 5.8.5 to obtain estimates $\hat{r}_{\text{OIS}}$, $\hat{r}_{\text{GOBS}}$, and $\hat{r}_{\text{GORIS}}$ for the Bayes factor r when $k = 1$. To implement these 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 in Verdinelli and Wasserman (1995), we use the Gibbs sampler by introducing auxiliary variables (latent variables). Note that a Student t distribution is a scale mixture of normal distributions (e.g., see Albert and Chib 1993). Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ and let the joint distribution of $(\boldsymbol{\theta}, \psi, \boldsymbol{\lambda})$ be

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

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

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

$$\mu \sim N\left(\frac{\sum_{j=1}^N \lambda_j y_j}{\sum_{j=1}^N \lambda_j}, \frac{\sigma^2}{\sum_{j=1}^N \lambda_j} \right),$$

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

and

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

where $\mathcal{G}(a, b)$ denotes a gamma distribution. Sampling λ_i, μ, and $1/\sigma^2$ from their corresponding conditional distributions is trivial and we use the adaptive rejection sampling algorithm of Gilks and Wild (1992) to generate $1/2\psi$ from $\pi(1/2\psi)$, since $\pi(1/2\psi)$ is log-concave when $N \geq 4$. Therefore, the Gibbs sampler can be exactly implemented. We believe that this Gibbs sampling scheme is superior to an independence chain Metropolis sampling scheme.

We implement the OIS, GOBS, and GORIS algorithms in double precision Fortran-77 using IMSL subroutines. We follow exactly the steps as the Algorithms OIS, GOBS, and GORIS presented in Section 5.8.5. We obtain a "random" sample $(\boldsymbol{\theta}_1, \psi_1), \ldots, (\boldsymbol{\theta}_n, \psi_n)$ 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 $\boldsymbol{\theta}_i$ disappear. The autocorrelations are calculated by the IMSL subroutine DACF. We use another

IMSL subroutine DQDAG to calculate $c(\boldsymbol{\theta}_i)$. A random sample $\boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{1n_1}$ from π_1 can be obtained by using an exact sampling scheme. For Algorithm GORIS, we choose $\pi_2(\boldsymbol{\theta}, \psi)$ as π in Step 1 and take a "random" sample $\{(\boldsymbol{\theta}_i, \psi_i), i = 1, \ldots, n_1\}$ from π_2 to calculate τ_{n_1} given by (5.8.14). In Step 2, we adopt Metropolis sampling with $\pi_2(\boldsymbol{\theta}, \psi)$ as a proposal density. Let $(\boldsymbol{\theta}_j, \psi_j)$ denote the current values of the parameters. We take candidate values $(\boldsymbol{\theta}_c, \psi_c)$ from every Bth "stationary" Gibbs iterate with the target distribution $\pi_2(\boldsymbol{\theta}, \psi)$. We compute

$$a = \min\left\{\frac{\omega(\boldsymbol{\theta}_c)}{\omega(\boldsymbol{\theta}_j)}, 1\right\},$$

where $\omega(\boldsymbol{\theta}) = |p_1(\boldsymbol{\theta})/c(\boldsymbol{\theta}) - \tau_{n_1}|$. We set $(\boldsymbol{\theta}_{j+1}, \psi_{j+1})$ equal to (θ_c, ψ_c) with acceptance probability a and to $(\boldsymbol{\theta}_j, \psi_j)$ with probability $1-a$. We then take every (B')th Metropolis iteration to obtain a "random" sample $(\boldsymbol{\vartheta}_1, \varphi_1)$, $\ldots$, $(\boldsymbol{\vartheta}_{n_2}, \varphi_{n_2})$. The above sampling schemes may not be the most efficient ones, but they do provide roughly independent samples and they are also straightforward to implement.

In order to obtain informative empirical evidence of the performance of OIS, GOBS, and GORIS, we conduct a small-scale simulation study. We take a dataset of $N = 100$ random numbers from $N(0, 1)$. Using this dataset, first we implement GOBS with $n_1 = n_2 = 50000$ 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" MC sample obtained. (Note that the Gibbs sampler converges earlier than 500 iterations.) Second, we use $n = 1000$ for Algorithm OIS, $n_1 = n_2 = 500$ for Algorithm GOBS, and $n_1 = 200$ and $n_2 = 800$ for Algorithm GORIS. As discussed in Section 5.7, we compute the simulation standard errors based on the estimated first-order approximation of $\mathrm{RE}(\hat{r})$ using the available random samples. (No extra random samples are required for this stage of the computation.) For example, the standard error for $\hat{r}_{\text{GOBS}}$ is given by

$$\begin{aligned} &\mathrm{se}(\hat{r}_{\text{GOBS}}) \\ &= \hat{r}_{\text{GOBS}} \left(\frac{1}{ns_1s_2} \left[\left(\frac{1}{n_2} \sum_{i=1}^{n_2} \frac{p_1(\theta_{2i})}{s_1 p_1(\theta_{2i}) + s_2 \hat{r}_{\text{GOBS}} c(\theta_{2i})} \right)^{-1} - 1 \right] \right)^{-1/2}, \end{aligned}$$

where $n = n_1 + n_2 = 1000$. Third, using the above implementation scheme with the same simulated dataset, we independently replicate the three estimation procedures 500 times. Then, we calculate the averages of $\hat{r}_{\text{OIS}}$, $\hat{r}_{\text{GOBS}}$, and $\hat{r}_{\text{GORIS}}$, simulation standard errors (simulation se), estimated biases $(E(\hat{r}) - r)$, mean-square errors (mse), averages of the approximate standard errors (approx. se), and the average CPU time. (Note that our computation was performed on the DEC-station 5000-260.) The results are summarized in Table 5.3.

TABLE 5.3. Results of Simulation Study.

	Method		
	OIS	GOBS	GORIS
Average of $\hat{r}$'s	6.995	6.971	6.933
Bias	0.037	0.013	−0.025
Mse	0.066	0.063	0.054
Simulation se	0.254	0.250	0.231
Approx. se	0.187	0.193	0.184
Average CPU (in minutes)	1.52	1.22	2.10

From Table 5.3, we see that:

(i) all three averages are close to the "true" value and the biases are relatively small;

(ii) GORIS produces a slightly smaller simulation standard error than the other two;

(iii) all three approximate standard errors are slightly understated, which is intuitively appealing since we use the estimated first-order approximation of RE($\hat{r}$); and

(iv) GOBS uses the least CPU time since sampling from $\pi_2(\boldsymbol{\theta}, \psi)$ is much more expensive than sampling from $\pi_1(\boldsymbol{\theta})$, and GORIS uses the most CPU time since sampling from $\pi^*_{n_1}(\boldsymbol{\theta}, \psi)$ in Step 2 of Algorithm GORIS is relatively more expensive.

Finally, we mention that based on the above-estimated value of r, the normal data results in a posterior marginal that is concentrated near $\psi = 0$, leading to a Bayes factor strongly favoring the null hypothesis of normality.

5.9 Estimation of Normalizing Constants After Transformation

When the "distance" between the two densities π_1 and π_2 gets large, the MC methods such as IS, BS, PS, and RIS will become less efficient. See Section 5.6 for illustrative examples. To remedy this problem, we can use a random variable transformation technique, which can help shorten the distance between the two densities π_1 and π_2, before applying the aforementioned MC methods.

Voter (1985) suggests applying a location shift before using the method of Bennett (1976) (see Section 5.3) to calculate free-energy differences between systems that are highly separated in configuration space. Meng and Schilling (1996a) extend Voter's idea by considering a general transformation before applying bridge sampling. To illustrate this idea, consider the following one-to-one transformation:

$$u = T_l(\boldsymbol{\theta}).$$

After the transformation, $\pi_l(\boldsymbol{\theta})$ can be rewritten as

$$\pi_l^*(u) \equiv \pi_l(T_l^{-1}(u))J_l(u) = \frac{q_l^*(u)}{c_l},$$

where $q_l^*(u) = q_l(T_l^{-1}(u))J_l(u)$ and $J_l(u)$ denotes the Jacobian, that is,

$$J_l(u) = \left|\frac{\partial T_l^{-1}(u)}{\partial u}\right|$$

for $l = 1, 2$. Now it is easy to see that c_i serves as the common normalizing constant for both π_l and π_l^*. Instead of directly working with the π_l, we can apply IS, BS, PS, and RIS to the π_l^*. Thus, the theory developed in Sections 5.2–5.4 remains the same. However, the transformation can greatly improve the simulation precision of an MC estimator of r. To see this, we revisit the two illustrative examples given in Section 5.6. For the case involving two densities from $N(0,1)$ and $N(\delta, 1)$, we let $u = T_1(\boldsymbol{\theta}) = \boldsymbol{\theta}$ for $N(0,1)$ and $u = T_2(\boldsymbol{\theta}) = \boldsymbol{\theta} - \delta$ for $N(\delta, 1)$. After the transformation, the two densities π_i^* are the same and both are $N(0,1)$. Thus all MC methods discussed in Section 5.6 give a precise estimate of r, yielding a zero simulation error. This is also true for the second case where we consider $N(0,1)$ and $N(0, \Delta^2)$ and we take $T_1(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $T_2(\boldsymbol{\theta}) = (\Delta^{-1})\boldsymbol{\theta}$. In these two illustrative examples, we indeed use two useful transformations, that is, recentering and rescaling. In general, the standardization, which is the combination of recentering and rescaling, may be a natural choice for T_l. More specifically, for $l = 1, 2$, we let

$$T_l(\boldsymbol{\theta}) = \Sigma_l^{-1/2}(\boldsymbol{\theta} - \mu_l),$$

where μ_l and Σ_l are the mean and covariance matrix for $\boldsymbol{\theta} \sim \pi_l$. If the analytical evaluation of μ_l and Σ_l does not appear possible, the MC approximation of μ_l and Σ_l can be easily obtained using the techniques described in Section 3.2.

Meng and Schilling (1996b) use a full information item factor model to empirically demonstrate the gain in simulation precision of BS after transformation. We conclude this section with a recommendation from Meng and Schilling (1996b), that one should apply transformations whenever feasible and appropriate.

5.10 Other Methods

In addition to IS, BS, PS, and RIS, several other MC methods have been developed recently. In this section, we briefly summarize some of these.

5.10.1 Marginal Likelihood Approach

In the context of Bayesian inference, the posterior is typically of the form

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

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

Consider the following identity:

$$m(D) = \frac{L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|D)}. \tag{5.10.1}$$

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

$$\hat{m}(D) = \frac{L(\boldsymbol{\theta}^*|D)\pi(\boldsymbol{\theta}^*)}{\hat{\pi}(\boldsymbol{\theta}^*|D)}.$$

He also develops a data augmentation technique of Tanner and Wong (1987) to estimate $\hat{\pi}(\boldsymbol{\theta}^*|D)$ by introducing latent variables. Chib's method is particularly useful for multivariate problems when the full conditional densities are completely known. The technical details and applications of this method are presented in Chapter 8. Another approach to estimating $\hat{\pi}(\boldsymbol{\theta}^*|D)$ is the importance-weighted marginal density estimation (IWMDE) method of Chen (1994), which has been extensively discussed in Chapter 4. Furthermore, the IWMDE method can be used to estimate $m(D)$ directly. Let $\boldsymbol{\theta}_i$, $i = 1, 2, \ldots, n$, be a random sample from $\pi(\boldsymbol{\theta}|D)$. Then, IWMDE yields a consistent estimator for $m(D)$:

$$\hat{m}_{\text{IWMDE}}(D) = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{w(\boldsymbol{\theta}_i)}{L(\boldsymbol{\theta}_i|D)\pi(\boldsymbol{\theta}_i)}\right]^{-1},$$

where $w(\boldsymbol{\theta})$ is a weighted density function (completely known) with support $\Omega_w \subset \Omega_{\pi(\cdot|D)}$ (the support of the posterior distribution $\pi(\cdot|D)$).

DiCiccio, Kass, Raftery, and Wasserman (1997) obtain the Laplace approximation to the normalizing constant $m(D)$ by approximating the posterior with a normal distribution, which is easy to sample from. Let $\boldsymbol{\theta}^*$ be the posterior mode and let Σ^* be minus the inverse of the Hessian of the log-posterior evaluated at $\boldsymbol{\theta}^*$. Then the Laplace approximation to $m(D)$ is given by

$$\hat{m}_L(D) = \frac{L(\boldsymbol{\theta}^*|D)\pi(\boldsymbol{\theta}^*)}{\phi(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*, \Sigma^*)} = (2\pi)^{p/2}|\Sigma^*|^{1/2}L(\boldsymbol{\theta}^*|D)\pi(\boldsymbol{\theta}^*),$$

where p is the dimension of $\boldsymbol{\theta}$ and $\phi(\cdot|\boldsymbol{\theta}^*, \Sigma^*)$ denotes a normal density with mean vector $\boldsymbol{\theta}^*$ and covariance matrix Σ^*. This approximation has error of order $O(1/n)$; that is, $m(D) = \hat{m}_L(D)(1 + O(1/n))$. By (5.10.1), we have

$$m(D) = \frac{L(\boldsymbol{\theta}^*|D)\pi(\boldsymbol{\theta}^*)}{\phi(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*, \Sigma^*)} \frac{\phi(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*, \Sigma^*)}{\pi(\boldsymbol{\theta}^*|D)} \approx \frac{L(\boldsymbol{\theta}^*|D)\pi(\boldsymbol{\theta}^*)}{\phi(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*, \Sigma^*)} \frac{\alpha}{P(B)},$$

where $\alpha = \Phi(B) = \int_B \phi(\boldsymbol{\theta}|\boldsymbol{\theta}^*, \Sigma^*)\ d\boldsymbol{\theta}$, $P(B) = \int_B (\pi(\boldsymbol{\theta}|D)\ d\boldsymbol{\theta}$, and $B = \{\boldsymbol{\theta} : \ ||(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'(\Sigma^*)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)|| \leq \delta\}$. DiCiccio, Kass, Raftery, and Wasserman (1997) suggest the following volume-corrected Laplace approximation estimator for $m(D)$:

$$\hat{m}_L^*(D) = \frac{L(\boldsymbol{\theta}^*|D)\pi(\boldsymbol{\theta}^*)}{\phi(\boldsymbol{\theta}^*|\boldsymbol{\theta}^*, \Sigma^*)} \frac{\alpha}{P(B)}.$$

To improve first-order approximations, they also suggest the Bartlett-adjusted Laplace estimator for $m(D)$, which is given by

$$\hat{m}_B^*(D) = \hat{m}_L(D) \cdot \left\{ \frac{E(W(\boldsymbol{\theta})|D)}{d} \right\}^{d/2},$$

where $W(\boldsymbol{\theta}) = 2\ln[L(\boldsymbol{\theta}^*|D)\pi(\boldsymbol{\theta}^*)/(L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta}))]$ and the expectation is taken with respect to $\pi(\boldsymbol{\theta}|D)$. They further show that this adjusted estimator has error of order $O(n^{-2})$. To completely determine $\hat{m}_L^*(D)$ and $\hat{m}_B^*(D)$, we must compute α, $P(B)$, and $E(W(\boldsymbol{\theta})|D)$. As long as a sample from the posterior distribution $\pi(\boldsymbol{\theta}|D)$ is available, $P(B)$ and $E(W(\boldsymbol{\theta})|D)$ are easy to calculate; see Section 3.2 for details. To compute α, one can use a numerical integration approach or an MC method since the normal distribution is easy to generate.

5.10.2 Reverse Logistic Regression

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

Let $\{\boldsymbol{\theta}_{l,i}, i = 1, \ldots, n_l\}$, $l = 1, 2$, be independent random samples from π_i, $l = 1, 2$, respectively. Also let $n = n_1 + n_2$, $s_{l,n} = n_l/n$, and $s_l = \lim_{n\to\infty} s_{l,n}$ for $l = 1, 2$. Consider a mixture distribution with density

$$\pi_{\text{mix}}(\boldsymbol{\theta}) = s_1 \frac{q_1(\boldsymbol{\theta})}{c_1} + s_2 \frac{q_2(\boldsymbol{\theta})}{c_2}.$$

Define

$$\begin{aligned} q_1^*(\boldsymbol{\theta}, r) =& \frac{s_1 q_1(\boldsymbol{\theta})/c_1}{s_1 q_1(\boldsymbol{\theta})/c_1 + s_2 q_2(\boldsymbol{\theta})/c_2} = \frac{s_1 q_1(\boldsymbol{\theta})}{s_1 q_1(\boldsymbol{\theta}) + r \cdot s_2 q_2(\boldsymbol{\theta})}, \\ q_2^*(\boldsymbol{\theta}, r) =& \frac{s_2 q_2(\boldsymbol{\theta})/c_2}{s_1 q_1(\boldsymbol{\theta})/c_1 + s_2 q_2(\boldsymbol{\theta})/c_2} = \frac{r s_2 q_2(\boldsymbol{\theta})}{s_1 q_1(\boldsymbol{\theta}) + r \cdot s_2 q_2(\boldsymbol{\theta})}, \end{aligned}$$

and also define the log quasi-likelihood as

$$l_n(r) = \sum_{l=1}^{2} \sum_{i=1}^{n_l} \ln q_l^*(\boldsymbol{\theta}_{l,i}, r). \tag{5.10.2}$$

Then the reverse logistic regression (RLR) estimator, $\hat{r}_{\text{RLR}}$, of r is obtained by maximizing the log quasi-likelihood $l_n(r)$ in (5.10.2). Clearly, $\hat{r}_{\text{RLR}}$ satisfies the following equation:

$$\sum_{i=1}^{n_2} \frac{s_1 q_1(\boldsymbol{\theta}_{2,i})}{\hat{r}_{\text{RLR}}(s_1 q_1(\boldsymbol{\theta}_{2,i}) + \hat{r}_{\text{RLR}} \cdot s_2 q_2(\boldsymbol{\theta}_{2,i}))} - \sum_{i=1}^{n_1} \frac{s_2 q_2(\boldsymbol{\theta}_{1,i})}{s_1 q_1(\boldsymbol{\theta}_{1,i}) + \hat{r}_{\text{RLR}} \cdot s_2 q_2(\boldsymbol{\theta}_{1,i})} = 0. \tag{5.10.3}$$

Therefore, when π_1 and π_2 overlap, i.e.,

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

and under some regularity conditions, we have

$$\hat{r}_{\text{RLR}} \xrightarrow{\text{a.s.}} r \ \text{ as } n \to \infty.$$

The asymptotic value of $E\left((\hat{r}_{\text{RLR}} - r)^2/r^2\right)$ is

$$\frac{1}{n s_1 s_2} \left[\left\{ \int_\Omega \frac{\pi_1(\boldsymbol{\theta}) \pi_2(\boldsymbol{\theta})}{s_1 \pi_1(\boldsymbol{\theta}) + s_2 \pi_2(\boldsymbol{\theta})} \, d\boldsymbol{\theta} \right\}^{-1} - 1 \right]. \tag{5.10.4}$$

From (5.10.3) and (5.10.4), we can see that the reverse logistic regression estimator, $\hat{r}_{\text{RLR}}$, is exactly the same as the optimal BS estimator, $\hat{r}_{\text{BS,opt}}$, given by (5.3.3) and (5.3.5) because (5.10.3) is identical to $S(r) = 0$, where $S(r)$ is given in (5.3.8). When π_1 and π_2 do not overlap, the reverse logistic regression method does not work directly.

5.10.3 The Savage–Dickey Density Ratio

In Section 5.8.1, we introduce a hypothesis testing problem considered by Verdinelli and Wasserman (1996). Suppose that the posterior $\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|D)$ is proportional to $L(\boldsymbol{\theta}, \boldsymbol{\psi}|D) \times \pi(\boldsymbol{\theta}, \boldsymbol{\psi})$, where $(\boldsymbol{\theta}, \boldsymbol{\psi}) \in \Omega \times \boldsymbol{\psi}$, $L(\boldsymbol{\theta}, \boldsymbol{\psi}|D)$ is the likelihood function given data D, and $\pi(\boldsymbol{\theta}, \boldsymbol{\psi})$ is the prior. We wish to test H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus H_1: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. The Bayes factor is

$$B = m_0/m,$$

where $m_0 = \int_\Psi L(\boldsymbol{\theta}_0, \boldsymbol{\psi}|D) \pi_0(\boldsymbol{\psi}) \, d\boldsymbol{\psi}$, $m = \int_{\Omega \times \Psi} L(\boldsymbol{\theta}, \boldsymbol{\psi}|D) \pi(\boldsymbol{\theta}, \boldsymbol{\psi}) \, d\boldsymbol{\theta} \, d\boldsymbol{\psi}$, and $\pi_0(\boldsymbol{\psi})$ is the prior under H_0. As discussed in Section 5.8, B is a ratio of two normalizing constants with different dimensions. In contrast to the MC methods presented in Section 5.8, Verdinelli and Wasserman (1995) suggest a generalization of the Savage–Dickey density ratio for estimating B. Dickey (1971) shows that if

$$\pi(\boldsymbol{\psi}|\boldsymbol{\theta}_0) = \pi_0(\boldsymbol{\psi}),$$

then

$$B = \frac{\pi(\boldsymbol{\theta}_0|D)}{\pi(\boldsymbol{\theta}_0)}, \tag{5.10.5}$$

where $\pi(\boldsymbol{\theta}_0|D) = \int_\Psi \pi(\boldsymbol{\theta}_0, \boldsymbol{\psi}|D)\ d\boldsymbol{\psi}$ and $\pi(\boldsymbol{\theta}) = \int_\Psi \pi(\boldsymbol{\theta}, \boldsymbol{\psi})\ d\boldsymbol{\psi}$. The reduced form of the Bayes factor B given in (5.10.5) is called the "Savage–Dickey density ratio."

In the cases where $\pi(\boldsymbol{\psi}|\boldsymbol{\theta}_0)$ depends on $\boldsymbol{\theta}_0$, Verdinelli and Wasserman (1995) obtain a generalized version of the Savage–Dickey density ratio. Assume that $0 < \pi(\boldsymbol{\theta}_0|D) < \infty$ and $0 < \pi(\boldsymbol{\theta}_0, \boldsymbol{\psi}) < \infty$ for almost all $\boldsymbol{\psi}$. Then the generalized Savage–Dickey density ratio is given by

$$B = \pi(\boldsymbol{\theta}_0|D) E\left[\frac{\pi_0(\boldsymbol{\psi})}{\pi(\boldsymbol{\theta}_0, \boldsymbol{\psi})}\right] = \frac{\pi(\boldsymbol{\theta}_0|D)}{\pi(\boldsymbol{\theta}_0)} E\left[\frac{\pi_0(\boldsymbol{\psi})}{\pi(\boldsymbol{\psi}|\boldsymbol{\theta}_0)}\right], \tag{5.10.6}$$

where the expectation is taken with respect to $\pi(\boldsymbol{\psi}|\boldsymbol{\theta}_0, D)$ (the conditional posterior density of $\boldsymbol{\psi}$ given $\boldsymbol{\theta} = \boldsymbol{\theta}_0$). To evaluate the generalized density ratio, we must compute $\pi(\boldsymbol{\theta}_0|D)$ and $E[\pi_0(\boldsymbol{\psi})/\pi(\boldsymbol{\theta}_0, \boldsymbol{\psi})]$. If a sample from the posterior $\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|D)$ is available, and closed forms of $\pi_0(\boldsymbol{\psi})$ and $\pi(\boldsymbol{\theta}_0, \boldsymbol{\psi})$ are also available (see Section 3.2 for details), computing $E[\pi_0(\boldsymbol{\psi})/\pi(\boldsymbol{\theta}_0, \boldsymbol{\psi})]$ is trivial. If closed forms for $\pi_0(\boldsymbol{\psi})$ and $\pi(\boldsymbol{\theta}_0, \boldsymbol{\psi})$ are not available, $\pi(\boldsymbol{\theta}_0|D)$ can be estimated by, for example, the IWMDE method discussed in Section 4.3. The application of the Savage–Dickey density ratio to the computation involving Bayesian model comparisons and Bayesian variable selection will be discussed in detail in Chapters 8 and 9.

5.11 An Application of Weighted Monte Carlo Estimators

In this section, we illustrate how the new weighted MC estimator given by (3.4.15) can be used for computing the ratio of normalizing constants. For illustrative purposes, we only consider the development of the weighted version of the importance sampling estimator $\hat{r}_{\text{IS}_2}$ given by (5.2.5).

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

$$\pi_j(\boldsymbol{\theta}) = \frac{q_j(\boldsymbol{\theta})}{c_j}, \quad \boldsymbol{\theta} \in \Omega_j, \tag{5.11.1}$$

where $\Omega_j \subset R^p$ is the support of π_j, and the unnormalized density $q_j(\boldsymbol{\theta})$ can be evaluated at any $\boldsymbol{\theta} \in \Omega_j$ for $j = 1, 2$. Our objective is to estimate the ratio of two normalizing constants defined as

$$r = \frac{c_1}{c_2}. \tag{5.11.2}$$

Let $\{\boldsymbol{\theta}_{2,1}, \boldsymbol{\theta}_{2,2}, \ldots, \boldsymbol{\theta}_{2,n}\}$ be a random sample from π_2. Then the IS estimator of r denoted by $\hat{r}_{\text{IS}_2}$ and its variance, $\text{Var}(\hat{r}_{\text{IS}_2})$, are given by (5.2.5) and (5.2.6), respectively. As discussed in Sections 5.2.2 and 5.6, $\hat{r}_{\text{IS}_2}$ is efficient when $\pi_2(\boldsymbol{\theta})$ has tails that are heavier than those of $\pi_1(\boldsymbol{\theta})$. However, when the two densities π_1 and π_2 have very little overlap (i.e., $E_2(\pi_1(\boldsymbol{\theta}))$ is very small), this method will work poorly.

To improve the simulation efficiency of $\hat{r}_{\text{IS}_2}$, we use the weighted estimator defined by (3.4.15) with the optimal weight $a_{\text{opt},l}$ given in (3.4.18). Let $\{A_l,\ l = 1, 2, \ldots, \kappa\}$ denote a partition of Ω_2. Using (3.4.14), we have

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

where $\pi_1(A_l|D)$ is the probability of set A_l with respect to π_1. Let $p_l = \pi_1(A_l)$ for $l = 1, 2, \ldots, \kappa$. The constraint given in (3.4.16) becomes

$$\sum_{l=1}^{\kappa} a_l p_l = 1. \tag{5.11.3}$$

The weighted estimator defined by (3.4.15) with the optimal weight a_{opt} reduces to

$$\hat{r}(a_{\text{opt}}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{k} a_{\text{opt},l} \left[\frac{q_1(\boldsymbol{\theta}_{2,i})}{q_2(\boldsymbol{\theta}_{2,i})}\right] 1\{\boldsymbol{\theta}_{2,i} \in A_l\}, \tag{5.11.4}$$

where

$$a_{\text{opt},l} = \frac{p_l}{b_l} \frac{1}{\sum_{j=1}^{\kappa} p_j^2 / b_j}, \tag{5.11.5}$$

and

$$b_l = E_2\left[\left(\frac{q_1(\boldsymbol{\theta})}{q_2(\boldsymbol{\theta})}\right)^2 1\{\boldsymbol{\theta} \in A_l\}\right]. \tag{5.11.6}$$

The variance given by (3.4.19) can be simplified to

$$\text{Var}(\hat{r}(a_{\text{opt}})) = \frac{1}{n}\left(\frac{1}{\sum_{l=1}^{\kappa} p_l^2 / b_l} - r^2\right). \tag{5.11.7}$$

It is easy to see that $\hat{r}(a_{\text{opt}})$ is an unbiased estimator of r. Also, it directly follows from Theorem 3.4.2 that $\hat{r}(a_{\text{opt}})$ is always better than $\hat{r}_{\text{IS}_2}$. We also note that in the weighted estimator $\hat{r}(a_{\text{opt}})$, the observations with larger probabilities, p_l's, and smaller second moments are assigned more weight. In contrast, the same weight is assigned to each observation in the estimator $\hat{r}_{\text{IS}_2}$. In addition, the weighted estimator $\hat{r}(a_{\text{opt}})$ combines information from both densities.

In practice, p_l and b_l are unknown. However, the computation of p_l is relatively easy if a random sample from $\pi_1(\boldsymbol{\theta})$ is available. More specifically, if $\{\boldsymbol{\theta}_{1,i},\ i = 1, 2, \ldots, m\}$ is a random sample from π_1, an estimator of p_l is given by

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

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

$$\hat{b}_l = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{q_1(\boldsymbol{\theta}_{2,i})}{q_2(\boldsymbol{\theta}_{2,i})} \right]^2 1\{\boldsymbol{\theta}_{2,i} \in A_l\}. \tag{5.11.8}$$

Replacing p_l and b_l by $\hat{p}_l$ and $\hat{b}_l$ in (5.11.5), an estimate of $a_{\mathrm{opt},l}$ is given by

$$\hat{a}_{\mathrm{opt},l} = \frac{\hat{p}_l}{\hat{b}_l} \frac{1}{\sum_{j=1}^{\kappa} \hat{p}_j^2 / \hat{b}_j}. \tag{5.11.9}$$

Plugging $\hat{a}_{\mathrm{opt},l}$ into (5.11.4) yields

$$\hat{r}(\hat{a}_{\mathrm{opt}}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{\kappa} \hat{a}_{\mathrm{opt},l} \left[\frac{q_1(\boldsymbol{\theta}_{2,i})}{q_2(\boldsymbol{\theta}_{2,i})} \right] 1\{\boldsymbol{\theta}_{2,i} \in A_l\}. \tag{5.11.10}$$

It is easy to show that $\hat{r}(\hat{a}_{\mathrm{opt}})$ is a consistent estimator as $n \to \infty$ and $m \to \infty$. Moreover, the next theorem shows that $\hat{r}(\hat{a}_{\mathrm{opt}})$ achieves the same variance as that of $\hat{r}(a_{\mathrm{opt}})$ given in (5.11.7) asymptotically.

Theorem 5.11.1 *Assume that* $\{\boldsymbol{\theta}_{1,i},\ i = 1, 2, \ldots, m\}$ *and* $\{\boldsymbol{\theta}_{2,i},\ i = 1, 2, \ldots, n\}$ *are two independent random samples. If* $n = o(m)$, *then*

$$\lim_{n \to \infty} nE\left(\hat{r}(\hat{a}_{\mathrm{opt}}) - r\right)^2 = \frac{1}{\sum_{l=1}^{\kappa} p_l^2 / b_l} - r^2. \tag{5.11.11}$$

The proof of this theorem is given in the Appendix. The weighted estimator $\hat{r}(\hat{a}_{\mathrm{opt}})$ is always better than $\hat{r}$. However, the trade-off here is that we have to pay a price to obtain an additional sample from π_1. Since it is relatively easy to compute $\hat{p}_l$ and $\hat{r}(\hat{a}_{\mathrm{opt}})$, the weighted estimator is potentially useful, if $\hat{r}(\hat{a}_{\mathrm{opt}})$ leads to a substantial gain in simulation efficiency. The following two examples demonstrate how the weighted estimator $\hat{r}(a_{\mathrm{opt}})$ performs.

Example 5.2. A theoretical case study. To get a better understanding of the weighted estimators developed in this section, we conduct a theoretical case study based on two normal densities considered in Section 5.6. Let $q_1(\theta) = \exp(-\theta^2/2)$ and $q_2(\theta) = \exp(-(\theta - \delta)^2/2)$ with δ a known positive constant. In this case, $c_1 = c_2 = \sqrt{2\pi}$ and, therefore, $r = 1$.

TABLE 5.4. Comparison of Variances.

δ	n Var $(\hat{r}_{\text{IS}_2})$	κ	n Var$(\hat{r}(a_{\text{opt}}))$
1	1.718	2	0.451
		5	0.116
		10	0.105
		20	0.103
2	53.598	2	3.855
		5	0.343
		10	0.107
		20	0.073
3	8102.084	2	42.694
		5	1.250
		10	0.242
		20	0.069

For the optimal weighted estimator $\hat{r}(a_{\text{opt}})$ given by (5.11.4), we consider the following partitions:

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

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

For (i), it can be shown that

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

where Φ is the standard normal ($N(0,1)$) cumulative distribution function (cdf). From Table 5.1, the variance of $\hat{r}_{\text{IS}_2}$ is given by

$$\text{Var}(\hat{r}_{\text{IS}_2}) = \frac{1}{n}\left[\exp(\delta^2) - 1\right].$$

Table 5.4 shows the values of n Var$(\hat{r}(a_{\text{opt}}))$ and n Var $(\hat{r}_{\text{IS}_2})$ for several different choices of δ and κ. From Table 5.4, it is easy to see that the weighted estimator $\hat{r}(a_{\text{opt}})$ dramatically improves the simulation efficiency compared to the importance sampling estimator $\hat{r}_{\text{IS}_2}$. For example, when $\delta = 3$ and $\kappa = 20$,

$$\text{Var}(\hat{r}_{\text{IS}_2})/\text{Var}(\hat{r}(a_{\text{opt}})) = 117,421.51,$$

i.e., $\hat{r}(a_{\text{opt}})$ is about 117,421 times better than $\hat{r}_{\text{IS}_2}$. Also, it is interesting to see that a finer partition yields a smaller variance. When the two densities are not far apart from each other, the variances of the weighted estimators are quite robust for $\kappa \geq 5$. However, when the two densities do not have much overlap, which is the case when $\delta = 3$, a substantial gain in simulation efficiency can be achieved by a finer partition.

In Section 5.6, we have shown that the ratio importance sampling estimator $\hat{r}_{\text{RIS}}$ given by (5.5.2) with the optimal π_{opt} given by (5.5.9) achieves the smallest asymptotic relative mean-square error, while the importance sampling estimator $\hat{r}_{\text{IS}_2}$ leads to the worst simulation efficiency. With the optimal density π_{opt}, Table 5.1 gives

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

where $\text{RE}^2(\hat{r}_{\text{RIS}}(\pi_{\text{opt}}))$ is defined by (5.5.3). It is easy to verify that

$$\lim_{n\to\infty} n\ \text{RE}^2(\hat{r}_{\text{RIS}}(\pi_{\text{opt}})) = 0.587,\ 1.864,\ \text{and } 3.002$$

for $\delta = 1$, 2, 3, respectively. Thus, from Table 5.4, it can be observed that $\hat{r}(a_{\text{opt}})$ is better than the optimal RIS estimator when $\kappa \geq 5$. This theoretical illustration is quite interesting, and demonstrates that the weighted version of the worst estimator can be better than the best estimator.

Example 5.3. ACTG036 data. In this example, we consider a data set from the AIDS study ACTG036. A detailed description of the ACTG036 study is given in Example 1.4. The sample size in this study, excluding cases with missing data, was 183. The response variable (y) for these data is binary with a 1 indicating death, development of AIDS, or AIDS related complex (ARC), and a 0 indicates otherwise. Several covariates were measured for these data. The ones we use here are CD4 count (x_1), age (x_2), treatment (x_3), and race (x_4). Chen, Ibrahim, and Yiannoutsos (1999) analyze the ACTG036 data using a logistic regression model.

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

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

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

$$B = \frac{\int_{R^5} L_1(\boldsymbol{\theta}|D)\ d\boldsymbol{\theta}}{\int_{R^5} L_2(\boldsymbol{\theta}|D)\ d\boldsymbol{\theta}} \equiv \frac{c_1}{c_2}, \tag{5.11.12}$$

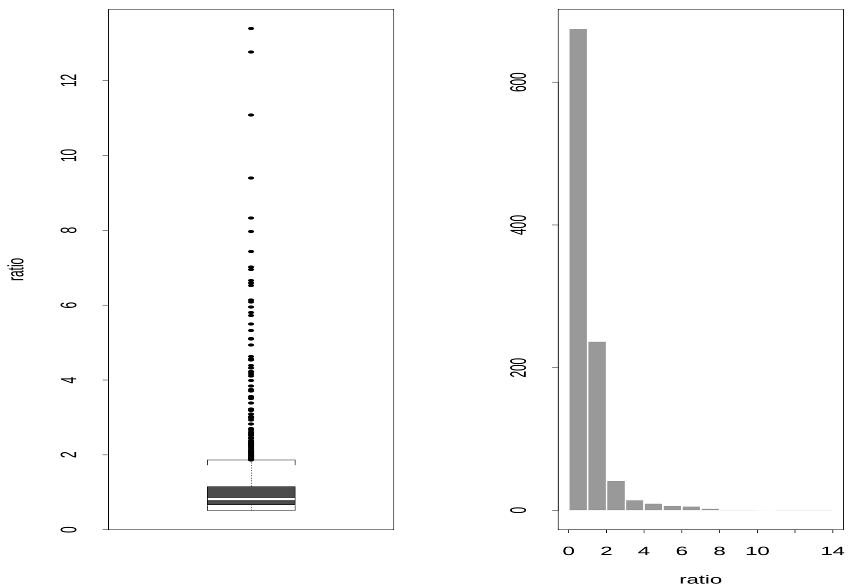

FIGURE 5.3. The box plot and histogram of the ratio $h(\boldsymbol{\theta}_i)$.

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

We use the Gibbs sampler to sample from the posterior distribution $\pi_2(\boldsymbol{\theta}|D) \propto L_2(\boldsymbol{\theta}|D)$. The autocorrelations for all the parameters disappear after lag 5. We obtain a sample of size $n = 1000$ by taking every 10^{th} Gibbs iterate. Then, using (5.2.5) and (5.2.6), we obtain $\hat{B} = 1.161$ and $n\ \widehat{\text{Var}}(\hat{B}) = 1.331$. In addition, we compute the ratio

$$h(\boldsymbol{\theta}_i) = L_1(\boldsymbol{\theta}_i|D)/L_2(\boldsymbol{\theta}_i|D)$$

for each observation. The box plot and histogram of these 1000 ratios are displayed in Figure 5.3.

Figure 5.3 clearly indicates that the posterior distribution of $h(\boldsymbol{\theta})$ is very skewed to the right. This suggests that the importance sampling estimator $\hat{B}$ cannot be reliable or accurate. To obtain a better estimate of B, we use the weighted estimators. We consider the following two partitions:

(i) $\kappa = 5$, $A_1 = \{\boldsymbol{\theta} :\ 0 < h(\boldsymbol{\theta}) \le 0.75\}$, $A_2 = \{\boldsymbol{\theta} :\ 0.75 < h(\boldsymbol{\theta}) \le 1.5\}$, $A_3 = \{\boldsymbol{\theta} :\ 1.5 < h(\boldsymbol{\theta}) \le 2.5\}$, $A_4 = \{\boldsymbol{\theta} :\ 2.5 < h(\boldsymbol{\theta}) \le 3.5\}$, and $A_5 = \{\boldsymbol{\theta} :\ 3.5 < h(\boldsymbol{\theta})\}$; and

(ii) $\kappa = 10$, $A_1 = \{\boldsymbol{\theta} :\ 0 < h(\boldsymbol{\theta}) \le 0.75\}$, $A_2 = \{\boldsymbol{\theta} :\ 0.75 < h(\boldsymbol{\theta}) \le 1.0\}$, $A_3 = \{\boldsymbol{\theta} :\ 1.0 < h(\boldsymbol{\theta}) \le 1.25\}$, $A_4 = \{\boldsymbol{\theta} :\ 1.25 < h(\boldsymbol{\theta}) \le 1.5\}$,

$A_5 = \{\boldsymbol{\theta} : 1.5 < h(\boldsymbol{\theta}) \le 2.0\}$, $A_6 = \{\boldsymbol{\theta} : 2.0 < h(\boldsymbol{\theta}) \le 2.5\}$, $A_7 = \{\boldsymbol{\theta} : 2.5 < h(\boldsymbol{\theta}) \le 3.0\}$, $A_8 = \{\boldsymbol{\theta} : 3.0 < h(\boldsymbol{\theta}) \le 3.5\}$, $A_9 = \{\boldsymbol{\theta} : 3.5 < h(\boldsymbol{\theta}) \le 4.0\}$, and $A_{10} = \{\boldsymbol{\theta} : 4.0 < h(\boldsymbol{\theta})\}$.

We generate a sample of size $m = 50000$ from the posterior distribution $\pi_1(\boldsymbol{\theta}|D) \propto L_1(\boldsymbol{\theta}|D)$ to estimate the probability p_j under each partition. Using (5.11.8), (5.11.9), (5.11.10), and (5.11.7), we obtain that $\hat{B}(\hat{a}_{\text{opt}})$ and $n\,\widehat{\text{Var}}(\hat{B}(\hat{a}_{\text{opt}}))$ are 1.099 and 0.050 for $\kappa = 5$, and 1.100 and 0.030 for $\kappa = 10$. For each observation, we also compute $w_i h(\boldsymbol{\theta}_i)$ (weight-times-ratio) for $\kappa = 10$, where $w_i = \sum_{l=1}^{\kappa} \hat{a}_l 1\{\boldsymbol{\theta}_i \in A_l\}$, and the box plot and the histogram of these 1000 values are displayed in Figure 5.4. From Figure 5.4, the reweighted observations are quite symmetric around the mean value. This result partially explains the reason why the weighted estimate works better. We also record the computing times for $\hat{B}$ and $\hat{B}(\hat{a}_{\text{opt}})$. The computing time for $\hat{B}$ is 137 seconds, and the computing time for $\hat{B}(\hat{a}_{\text{opt}})$ takes an additional 150 seconds on a digital alpha machine. In addition, we run the simulation with a very large number of iterations ($n = 500,000$), and we find that the "golden value" of B is around 1.102, which confirms that the weighted estimate is quite accurate, even when $n = 1000$. Based on the estimated Bayes factor, we can conclude that the logit model is slightly better than the complementary log–log link model.

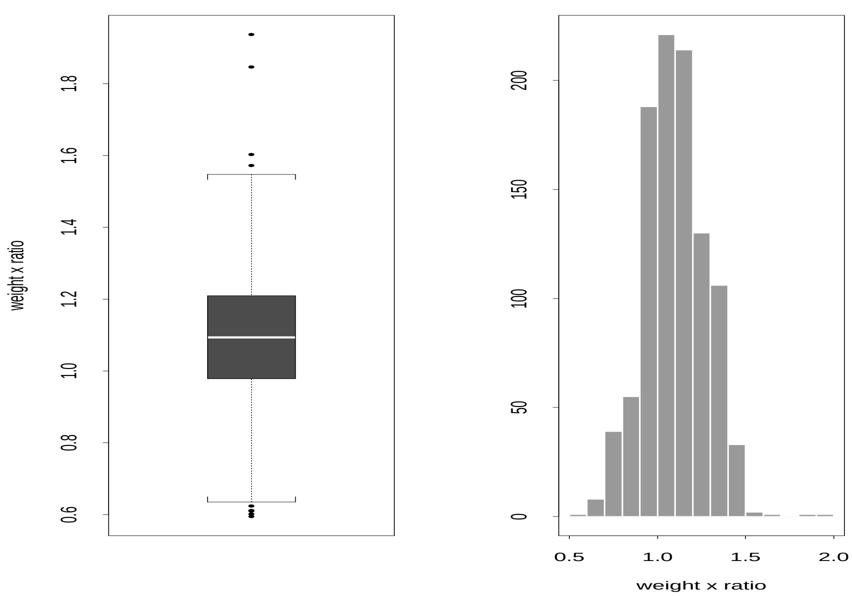

FIGURE 5.4. The box plot and histogram of the weight-times-ratio.

Finally, we note that the weighted estimators for the other MC methods such as BS and RIS can also be developed. The weighted versions of the BS estimator $\hat{r}_{\text{BS}}$ defined in (5.3.3) and the RIS estimator $\hat{r}_{\text{RIS}}$ given by (5.5.2) are analogous to the one for the IS estimator $\hat{r}_{\text{IS}_2}$. The detailed formulations are left as an exercise. We also note that Peng (1998) develops an efficient weighted MC method for computing the normalizing constants, which are essentially the posterior model probabilities obtained from the Stochastic Search Variable Selection method of George and McCulloch (1993). She obtains the fixed weight and data-dependent weight estimators of the normalizing constants. However, the support of the posterior distribution considered in Peng (1998) is discrete and finite. The main idea of her method is to "partition" an MC sample (not the support of the posterior distribution) into several subsets, and then she assigns a fixed or random weight to each subset. The noticeable difference between her method and the one presented in this section is that she partitions the sample, and the subsets in her partition must not be mutually exclusive. Therefore, her method is useful for computing the normalizing constant of a discrete posterior distribution.

5.12 Discussion

In this chapter, we have assumed independence among samples when deriving all theoretical results. However, the samples from a desired distribution using MCMC sampling as described in Chapter 2 are typically dependent. Under certain regularity assumptions, such as ergodicity and weak dependence, the consistency and the central limit theorem of an estimator of r still hold. The only problem is the derivation of the relative mean-square error. One simple remedy is to obtain an approximately random sample by taking every Bth iterate in MCMC sampling, where B is selected so that the autocorrelations are negligible with respect to their standard errors; see, for example, Gelfand and Smith (1990). Other possible approaches are to use the expensive regeneration technique in Markov chain sampling (Mykland, Tierney, and Yu 1995) to obtain a random sample from different regeneration tours, *effective sample sizes* (Meng and Wong 1996) to derive the relative mean-square error, and a coupling-regeneration scheme of Johnson (1998). In addition, Meng and Wong (1996) comment that empirical studies, as reported in DiCiccio, Kass, Raftery, and Wasserman (1997) and in Meng and Schilling (1996a), suggest that the optimal or near-optimal procedures constructed under the independence assumption can work remarkably well in general, providing orders of magnitude improvement over other methods with similar computational effort.

We have shown that RIS with an optimal "middle" density π_{opt} works better than IS, BS, and PS. However, the implementation of the optimal

RIS estimator is expensive, which can be seen from Sections 5.5.2 and 5.8.5. As we discuss in Section 5.5.1, the idea of RIS is useful particularly when one deals with a Bayesian computational problem involving many ratios of normalizing constants. The idea of RIS will be extended to solve computationally intensive problems arising from Bayesian constrained parameter problems in Chapter 6 and Bayesian model comparisons in Chapters 8 and 9.

The different dimensions problems presented in Section 5.8 are important as they often arise in Bayesian model comparison and Bayesian variable selection. The algorithms presented in Section 5.8.5 can asymptotically or approximately achieve the optimal simulation errors, and they can be programmed in a routine manner. The methodology presented in this chapter will also be useful in the computation of Bayes factors (Kass and Raftery 1995), intrinsic Bayes factors (Berger and Pericchi 1996), Bayesian model comparisons (Geweke 1994), and model selection. In particular, the methods developed in this chapter can be directly applied to Bayesian model comparisons, which will be discussed in detail in Chapters 8 and 9.

Appendix

Proof of Theorem 5.3.1. By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\left\{ \int_{\Omega_1 \cap \Omega_2} \pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})\ d\boldsymbol{\theta} \right\}^2 &\leq \left\{ \int_{\Omega_1 \cap \Omega_2} \sqrt{\frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}} \right. \\
&\quad \left. \times \left[\sqrt{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})(s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta}))} |\alpha(\boldsymbol{\theta})| \right] d\boldsymbol{\theta} \right\}^2 \\
&\leq \int_{\Omega_1 \cap \Omega_2} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\ d\boldsymbol{\theta} \\
&\quad \times \int_{\Omega_1 \cap \Omega_2} \pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})(s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta}))\alpha^2(\boldsymbol{\theta})\ d\boldsymbol{\theta}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{\int_{\Omega_1 \cap \Omega_2} \pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})(s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta}))\alpha^2(\boldsymbol{\theta})\ d\boldsymbol{\theta}}{\{\int_{\Omega_1 \cap \Omega_2} \pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})\alpha(\boldsymbol{\theta})\ d\boldsymbol{\theta}\}^2} \\
&\qquad \geq \left[\int_{\Omega_1 \cap \Omega_2} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\ d\boldsymbol{\theta} \right]^{-1},
\end{aligned}$$

where equality holds if and only if (up to a zero-measure set)

$$[\sqrt{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})(s_1\pi_1(\boldsymbol{\theta})+s_2\pi_2(\boldsymbol{\theta}))}]\alpha(\boldsymbol{\theta}) \ \propto \ \sqrt{\frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta})+s_2\pi_2(\boldsymbol{\theta})}},$$

which yields (5.3.5). □

Proof of Theorem 5.4.2. Letting $c(\lambda) = \int_\Omega q(\boldsymbol{\theta}|\lambda)\ d\boldsymbol{\theta}$, we have

$$\xi = \int_{\lambda_1}^{\lambda_2} \left[\frac{d}{d\lambda}\ln c(\lambda)\right]\ d\lambda$$

and

$$E_\lambda\{U^2(\boldsymbol{\theta},\lambda)\} = \int_\Omega \left[\frac{d}{d\lambda}\ln \pi(\boldsymbol{\theta}|\lambda)\right]^2 \pi(\boldsymbol{\theta}|\lambda)\ d\boldsymbol{\theta} + \left[\frac{d}{d\lambda}\ln c(\lambda)\right]^2. \quad (5.A.1)$$

Equations (5.4.3) and (5.A.1) lead to

$$\begin{aligned} n\ \mathrm{Var}(\hat{\xi}_{\mathrm{PS}}) = &\int_{\lambda_1}^{\lambda_2}\int_\Omega \left[\frac{d}{d\lambda}\ln \pi(\boldsymbol{\theta}|\lambda)\right]^2 \frac{\pi(\boldsymbol{\theta}|\lambda)}{\pi_\lambda(\lambda)}\ d\boldsymbol{\theta}\ d\lambda \\ &+ \left[\int_{\lambda_1}^{\lambda_2}\left[\frac{d}{d\lambda}\ln c(\lambda)\right]^2 \frac{1}{\pi_\lambda(\lambda)}\ d\lambda - \xi^2\right]. \end{aligned} \quad (5.A.2)$$

Using the Cauchy–Schwarz inequality and $\int_{\lambda_1}^{\lambda_2} \pi_\lambda(\lambda)\ d\lambda = 1$, we have

$$\begin{aligned} &\int_{\lambda_1}^{\lambda_2}\left[\frac{d}{d\lambda}\ln c(\lambda)\right]^2 \frac{1}{\pi_\lambda(\lambda)}\ d\lambda - \xi^2 \\ &\qquad \geq \left[\int_{\lambda_1}^{\lambda_2} \frac{(d/d\lambda)\ln c(\lambda)}{\sqrt{\pi_\lambda(\lambda)}}\sqrt{\pi_\lambda(\lambda)}\ d\lambda\right]^2 - \xi^2 = 0. \end{aligned} \quad (5.A.3)$$

Similarly,

$$\begin{aligned} &\int_{\lambda_1}^{\lambda_2}\int_\Omega \left[\frac{d}{d\lambda}\ln \pi(\boldsymbol{\theta}|\lambda)\right]^2 \frac{\pi(\boldsymbol{\theta}|\lambda)}{\pi_\lambda(\lambda)}\ d\boldsymbol{\theta}\ d\lambda \\ &\qquad = \int_{\lambda_1}^{\lambda_2}\int_\Omega 4\left[\frac{d}{d\lambda}\sqrt{\pi(\boldsymbol{\theta}|\lambda)}\right]^2 \frac{1}{\pi_\lambda(\lambda)}\ d\boldsymbol{\theta}\ d\lambda \\ &\qquad \geq 4\int_\Omega \left[\int_{\lambda_1}^{\lambda_2} \frac{(d/d\lambda)\sqrt{\pi(\boldsymbol{\theta}|\lambda)}}{\sqrt{\pi_\lambda(\lambda)}}\sqrt{\pi_\lambda(\lambda)}\ d\lambda\right]^2\ d\boldsymbol{\theta} \\ &\qquad = 4\int_\Omega \left[\int_{\lambda_1}^{\lambda_2} \frac{d}{d\lambda}\sqrt{\pi(\boldsymbol{\theta}|\lambda)}\ d\lambda\right]^2\ d\boldsymbol{\theta} \\ &\qquad = 4\int_\Omega \left[\sqrt{\pi(\boldsymbol{\theta}|\lambda_2)} - \sqrt{\pi(\boldsymbol{\theta}|\lambda_1)}\right]^2\ d\boldsymbol{\theta}. \end{aligned} \quad (5.A.4)$$

Thus, the theorem follows from (5.A.2), (5.A.3), and (5.A.4). □

Proof of Theorem 5.5.1. Write

$$\sqrt{n}(\hat{r}_{\text{RIS}} - r) = \frac{c_1 n^{-1/2} \sum_{i=1}^{n} \{f_1(\boldsymbol{\theta}_i)/c_1 - f_2(\boldsymbol{\theta}_i)/c_2\}}{(1/n)\sum_{i=1}^{n} f_2(\boldsymbol{\theta}_i)}. \tag{5.A.5}$$

It follows from the central limit theorem that

$$n^{-1/2}\sum_{i=1}^{n}\{f_1(\boldsymbol{\theta}_i)/c_1 - f_2(\boldsymbol{\theta}_i)/c_2\} \xrightarrow{\mathcal{D}} N\left(0, E_\pi\left\{\frac{f_1(\boldsymbol{\theta})}{c_1} - \frac{f_2(\boldsymbol{\theta})}{c_2}\right\}^2\right) \quad \text{as } n \to \infty \tag{5.A.6}$$

and from the law of large numbers that

$$(1/n)\sum_{i=1}^{n} f_2(\boldsymbol{\theta}_i) \xrightarrow{\text{a.s.}} c_2 \quad \text{as } n \to \infty. \tag{5.A.7}$$

Now (5.5.5) is an immediate consequence of (5.A.6) and (5.A.7). To prove (5.5.4), it suffices to show that $\{n(\hat{r}_{\text{RIS}} - r)^2,\ n \geq 1\}$ is uniformly integrable. In this case, by (5.5.5), we shall have $E\{\sqrt{n}(\hat{r}_{\text{RIS}} - r)\} = o(1)$ as $n \to \infty$. Thus

$$\frac{1}{r^2}E\left\{n(\hat{r}_{\text{RIS}} - r)^2\right\} \to E\left\{\frac{f_1(\boldsymbol{\theta})}{c_1} - \frac{f_2(\boldsymbol{\theta})}{c_2}\right\}^2 \quad \text{as } n \to \infty,$$

which gives (5.5.4). We show below the uniform integrability of $\{n(\hat{r}_{\text{RIS}} - r)^2,\ n \geq 1\}$. Rewrite

$$\sqrt{n}(\hat{r}_{\text{RIS}} - r) = \frac{n^{-1/2}\sum_{i=1}^{n}\{c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i)\}}{c_2(1/n)\sum_{i=1}^{n} f_2(\boldsymbol{\theta}_i)} \tag{5.A.8}$$

and let $U_n = n^{-1/2}\sum_{i=1}^{n}\{c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i)\}$ and $V_n = n^{-1}\sum_{i=1}^{n} f_2(\boldsymbol{\theta}_i)$. By (5.A.8), for every $A \geq 2$,

$$\begin{aligned}
&E[n(\hat{r}_{\text{RIS}} - r)^2 1\{n(\hat{r}_{\text{RIS}} - r)^2 \geq A^2\}] \\
&\quad = E\left[\frac{U_n^2}{c_2^2 V_n^2} 1\{|U_n| \geq c_2 A V_n\}\right] \\
&\quad = E\left[\frac{U_n^2}{c_2^2 V_n^2} 1\{|U_n| \geq A c_2 V_n, V_n \geq c_2/2\}\right] \\
&\quad\quad + E\left[\frac{U_n^2}{c_2 V_n^2} 1\{|U_n| \geq A c_2 V_n, V_n < c_2/2\}\right] \\
&\quad \leq 4c_2^{-4} E[U_n^2 1\{|U_n| \geq A c_2^2/2\}] \\
&\quad\quad + E[n(\hat{r}_{\text{RIS}} - r)^2 1\{V_n < c_2/2\}],
\end{aligned} \tag{5.A.9}$$

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

$$\lim_{A\to\infty} \sup_n E[U_n^2 1\{|U_n| \geq Ac_2^2/2\}] = 0. \tag{5.A.10}$$

Noting that $\hat{r}_{\text{RIS}} \leq \sum_{i=1}^n f_1(\boldsymbol{\theta}_i)/f_2(\boldsymbol{\theta}_i)$, we have

$$\begin{aligned}
&E[n(\hat{r}_{\text{RIS}} - r)^2 1\{V_n < c_2/2\}] \\
&\qquad \leq nE_\pi \left[(\hat{r}_{\text{RIS}}^2 + r^2) 1\{V_n < c_2/2\}\right] \\
&\qquad \leq nE\left[\left\{r^2 + n\sum_{i=1}^n (f_1(\boldsymbol{\theta}_i)/f_2(\boldsymbol{\theta}_i))^2\right\} 1\{V_n < c_2/2\}\right] \\
&\qquad \leq n\Bigg[r^2 P(V_n < c_2/2) \\
&\qquad\quad + n\sum_{i=1}^n E\left\{(f_1(\boldsymbol{\theta}_i)/f_2(\boldsymbol{\theta}_i))^2 1\Big\{\sum_{j\neq i} f_2(\boldsymbol{\theta}_j) < nc_2/2\Big\}\right\}\Bigg] \\
&\qquad = n\Bigg[r^2 P(V_n < c_2/2) \\
&\qquad\quad + n^2 E\left(f_1(\boldsymbol{\theta})/f_2(\boldsymbol{\theta})\right)^2 P\left(\sum_{j=1}^{n-1} f_2(\boldsymbol{\theta}_j) < nc_2/2\right)\Bigg].
\end{aligned} \tag{5.A.11}$$

Using the Chebyshev inequality, we get

$$\begin{aligned}
P(V_n < c_2/2) &= P\left(\sum_{i=1}^n \{Ef_2(\boldsymbol{\theta}_i) - f_2(\boldsymbol{\theta}_i)\} > nc_2/2\right) \\
&\leq \inf_{t\geq 0} \exp(-tc_2 n/2) E\left[\exp\left(\sum_{i=1}^n \{Ef_2(\boldsymbol{\theta}_i) - f_2(\boldsymbol{\theta}_i)\}\right)\right] \\
&= \left(\inf_{t\geq 0} \exp(-tc_2/2) E\exp[t(c_2 - f_2(\boldsymbol{\theta}))]\right)^n.
\end{aligned} \tag{5.A.12}$$

From $E(c_2 - f_2(\boldsymbol{\theta})) = 0$, it follows that

$$\varepsilon = \inf_{t\geq 0} \exp(-tc_2/4) E\exp\{t(c_2 - f_2(\boldsymbol{\theta}))\} < 1.$$

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

$$\begin{aligned}
&P\left(\sum_{j=1}^{n-1} f_2(\boldsymbol{\theta}_j) < nc_2/2\right) \\
&\quad = P\left(\sum_{j=1}^{n-1}\{Ef_2(\boldsymbol{\theta}_j) - f_2(\boldsymbol{\theta}_j)\} > (n-2)c_2/2\right) \\
&\quad \le \left(\inf_{t\ge 0} \exp[-(n-2)tc_2/2(n-1)]E\exp\{t(c_2 - f_2(\boldsymbol{\theta}))\}\right)^{n-1} \\
&\quad \le \varepsilon^{n-1}. \qquad (5.A.13)
\end{aligned}$$

Putting together the above inequalities yields

$$E[n(\hat{r}_{\text{RIS}} - r)^2 1\{V_n < c_2/2\}] = O(n^3\varepsilon^n) = o(1). \qquad (5.A.14)$$

Therefore, (5.5.4) follows from (5.A.9), (5.A.10), and (5.A.14).

Next, we prove (5.5.6). Observe that

$$\begin{aligned}
&nE(\hat{r}_{\text{RIS}} - r)^2 - c_2^{-4}E\{c_2 f_1(\boldsymbol{\theta}) - c_1 f_2(\boldsymbol{\theta})\}^2 \\
&\quad = c_2^{-2}n\left[E\left\{\frac{\sum_{i=1}^n(c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))}{\sum_{i=1}^n f_2(\boldsymbol{\theta}_i)}\right\}^2\right. \\
&\qquad \left. - E\left\{\frac{\sum_{i=1}^n(c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))}{nc_2}\right\}^2\right] \\
&\quad = \frac{c_2^{-4}}{n}\left[E\left\{\frac{\left(\sum_{i=1}^n(c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))\right)^2}{(\sum_{i=1}^n f_2(\boldsymbol{\theta}_i))^2}\right.\right. \\
&\qquad \left.\left. \times \sum_{i=1}^n (c_2 - f_2(\boldsymbol{\theta}_i)) \cdot \sum_{i=1}^n (c_2 + f_2(\boldsymbol{\theta}_i))\right\}\right] \\
&\quad \stackrel{\text{def}}{=} \frac{c_2^{-4}}{n}\varepsilon_n, \qquad (5.A.15)
\end{aligned}$$

where

$$\begin{aligned}
\varepsilon_n &= E\left\{\frac{\left(\sum_{i=1}^n(c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))\right)^2 \cdot \sum_{i=1}^n (c_2 - f_2(\boldsymbol{\theta}_i)) \cdot 2nc_2}{(\sum_{i=1}^n f_2(\boldsymbol{\theta}_i))^2}\right\} \\
&\quad - E\left\{\frac{\left(\sum_{i=1}^n(c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))\right)^2 \cdot \left(\sum_{i=1}^n (c_2 - f_2(\boldsymbol{\theta}_i))\right)^2}{(\sum_{i=1}^n f_2(\boldsymbol{\theta}_i))^2}\right\}.
\end{aligned}$$

After some algebra, we have

$$
\begin{aligned}
\varepsilon_n &= 2E\left\{\frac{\left(\sum_{i=1}^n (c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))\right)^2 \cdot \sum_{i=1}^n (c_2 - f_2(\boldsymbol{\theta}_i))}{(n\,c_2)}\right\} \\
&\quad + 2E\Bigg\{\frac{\left(\sum_{i=1}^n (c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))\right)^2 \cdot \sum_{i=1}^n (c_2 - f_2(\boldsymbol{\theta}_i))}{nc_2(\sum_{i=1}^n f_2(\boldsymbol{\theta}_i))^2} \\
&\quad \times \left((nc_2)^2 - \left(\sum_{i=1}^n f_2(\boldsymbol{\theta}_i)\right)^2\right)\Bigg\} \\
&\quad - E\left\{\frac{\left(\sum_{i=1}^n (c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))\right)^2 \cdot \left(\sum_{i=1}^n (c_2 - f_2(\boldsymbol{\theta}_i))\right)^2}{(\sum_{i=1}^n f_2(\boldsymbol{\theta}_i))^2}\right\} \\
&\stackrel{\text{def}}{=} \varepsilon_{n,1} + \varepsilon_{n,2} + \varepsilon_{n,3}.
\end{aligned}
$$

It is easy to see that

$$
\begin{aligned}
\varepsilon_{n,1} &= 2(nc_2)^{-1}E\Bigg\{\Bigg(\sum_{i=1}^n (c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))^2 \\
&\quad + 2\sum_{1\le i<j\le n} (c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))(c_2 f_1(\boldsymbol{\theta}_j) - c_1 f_2(\boldsymbol{\theta}_j))\Bigg) \\
&\quad \times \sum_{i=1}^n (c_2 - f_2(\boldsymbol{\theta}_i))\Bigg\} \\
&= 2(nc_2)^{-1}E\left\{\left(\sum_{i=1}^n (c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))^2\right) \cdot \sum_{i=1}^n (c_2 - f_2(\boldsymbol{\theta}_i))\right\} \\
&= 2(nc_2)^{-1}E\Bigg\{\left(\sum_{i=1}^n \{(c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))^2 - E(c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))^2\}\right) \\
&\qquad\qquad \times \sum_{i=1}^n (c_2 - f_2(\boldsymbol{\theta}_i))\Bigg\} \\
&\le (nc_2)^{-1}\left[\mathrm{Var}\left(\sum_{i=1}^n (c_2 f_1(\boldsymbol{\theta}_i) - c_1 f_2(\boldsymbol{\theta}_i))^2\right)\right]^{1/2} \\
&\quad \times \left[\mathrm{Var}\left(\sum_{i=1}^n (c_2 - f_2(\boldsymbol{\theta}_i))\right)\right]^{1/2} \\
&= O(1).
\end{aligned}
$$

As for $\varepsilon_{n,2}$, we have

$$\begin{aligned}
|\varepsilon_{n,2}| &= 2\left|E\left\{\frac{\left(\sum_{i=1}^n(c_2f_1(\boldsymbol{\theta}_i)-c_1f_2(\boldsymbol{\theta}_i))\right)^2\cdot\left(\sum_{i=1}^n(c_2-f_2(\boldsymbol{\theta}_i))\right)^2}{nc_2(\sum_{i=1}^n f_2(\boldsymbol{\theta}_i))^2}\right.\right.\\
&\qquad\left.\left.\times\left(nc_2+\sum_{i=1}^n f_2(\boldsymbol{\theta}_i)\right)\right\}\right|\\
&\le 12(nc_2)^{-2}E\left\{\left(\sum_{i=1}^n(c_2f_1(\boldsymbol{\theta}_i)-c_1f_2(\boldsymbol{\theta}_i))\right)^2\cdot\left(\sum_{i=1}^n(c_2-f_2(\boldsymbol{\theta}_i))\right)^2\right\}\\
&\quad+2\left|E\left\{\frac{\left(\sum_{i=1}^n(c_2f_1(\boldsymbol{\theta}_i)-c_1f_2(\boldsymbol{\theta}_i))\right)^2\cdot\left(\sum_{i=1}^n(c_2-f_2(\boldsymbol{\theta}_i))\right)^2}{nc_2(\sum_{i=1}^n f_2(\boldsymbol{\theta}_i))^2}\right.\right.\\
&\qquad\left.\left.\times\left(nc_2+\sum_{i=1}^n f_2(\boldsymbol{\theta}_i)\right)1\{V_n<c_2/2\}\right\}\right|\\
&\le 12(nc_2)^{-2}\left[E\left\{\sum_{i=1}^n(c_2f_1(\boldsymbol{\theta}_i)-c_1f_2(\boldsymbol{\theta}_i))\right\}^4\right.\\
&\qquad\left.\times E_\pi\left\{\sum_{i=1}^n(c_2-f_2(\boldsymbol{\theta}_i))\right\}^4\right]^{1/2}\\
&\quad+4(nc_2)^3E\left\{\left(c_1+c_2\sum_{i=1}^n f_1(\boldsymbol{\theta}_i)/f_2(\boldsymbol{\theta}_i)\right)^2 1\{V_n<c_2/2\}\right\}\\
&= O(1)+O(n^5\varepsilon^n)=O(1),
\end{aligned}$$

where the last inequality is from (5.A.12) and the proof of (5.A.11). Similarly, we have

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

Now (5.5.6) follows from the above inequalities. This proves the theorem. □

Proof of Theorem 5.5.2. By the Cauchy–Schwarz inequality, for an arbitrary density $\pi(\cdot)$,

$$\left[\int_\Omega|\pi_1(\boldsymbol{\theta})-\pi_2(\boldsymbol{\theta})|\ d\boldsymbol{\theta}\right]^2\le\int_\Omega\frac{[\pi_1(\boldsymbol{\theta})-\pi_2(\boldsymbol{\theta})]^2}{\pi(\boldsymbol{\theta})}\ d\boldsymbol{\theta}\cdot\int_\Omega\pi(\boldsymbol{\theta})\ d\boldsymbol{\theta}.$$

Thus,

$$E\left[\frac{\{\pi_1(\boldsymbol{\theta})-\pi_2(\boldsymbol{\theta})\}^2}{\pi^2(\boldsymbol{\theta})}\right]\ge\left[\int_\Omega|\pi_1(\boldsymbol{\theta})-\pi_2(\boldsymbol{\theta})|\ d\boldsymbol{\theta}\right]^2$$

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

$$\pi(\boldsymbol{\theta}) \propto |\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})|,$$

that is, $\pi(\boldsymbol{\theta}) = \pi_{\text{opt}}(\boldsymbol{\theta})$. This proves (5.5.9). Replacing π by π_{opt} in (5.5.7) gives (5.5.10). □

Proof of Theorem 5.5.3. Since

$$\begin{aligned}
1 - &\int_{\Omega_1 \cap \Omega_2} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\
&= \int_\Omega \frac{(s_2\pi_1(\boldsymbol{\theta}) + s_1\pi_2(\boldsymbol{\theta}))(s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})) - \pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\
&= \int_\Omega \frac{(s_1\, s_2\pi_1^2(\boldsymbol{\theta}) + s_1 s_2\pi_2^2(\boldsymbol{\theta}) + (s_1^2 + s_2^2 - 1)\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \\
&= s_1 s_2 \int_\Omega \frac{(\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta}))^2}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}, \qquad (5.A.16)
\end{aligned}$$

the right-hand side of (5.5.11)

$$\begin{aligned}
&= \int_\Omega \frac{(\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta}))^2}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \cdot \left[\int_\Omega \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right]^{-1} \\
&= \int_\Omega \frac{(\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta}))^2}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \cdot \int_\Omega (s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta}))\, d\boldsymbol{\theta} \\
&\quad \times \left[\int_\Omega \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right]^{-1} \\
&\geq \left[\int_\Omega \frac{|\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})|}{\sqrt{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}} \cdot \sqrt{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right]^2 \\
&\quad \times \left[\int_\Omega \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right]^{-1} \qquad (5.A.17) \\
&= \left[\int_\Omega |\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})|\, d\boldsymbol{\theta}\right]^2 \cdot \left[\int_\Omega \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right]^{-1}, \\
&\qquad (5.A.18)
\end{aligned}$$

where (5.A.17) is obtained by the Cauchy–Schwarz inequality. From (5.A.16) it can be shown that

$$\int_{\Omega_1 \cap \Omega_2} \frac{\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta} \leq 1. \qquad (5.A.19)$$

Now (5.5.11) follows from (5.A.18) and (5.A.19). This proves the theorem. □

Proof of Theorem 5.5.4. By the Cauchy–Schwarz inequality, the left side of (5.5.12) equals

$$\left[\int_\Omega \left|\sqrt{\pi_1(\boldsymbol{\theta})} - \sqrt{\pi_2(\boldsymbol{\theta})}\right| \left(\sqrt{\pi_1(\boldsymbol{\theta})} + \sqrt{\pi_2(\boldsymbol{\theta})}\right)\, d\boldsymbol{\theta}\right]^2$$
$$\leq \int_\Omega \left[\sqrt{\pi_1(\boldsymbol{\theta})} - \sqrt{\pi_2(\boldsymbol{\theta})}\right]^2\, d\boldsymbol{\theta} \cdot \int_\Omega \left[\sqrt{\pi_1(\boldsymbol{\theta})} + \sqrt{\pi_2(\boldsymbol{\theta})}\right]^2\, d\boldsymbol{\theta}. \tag{5.A.20}$$

It is easy to see that

$$\int_\Omega \left[\sqrt{\pi_1(\boldsymbol{\theta})} + \sqrt{\pi_2(\boldsymbol{\theta})}\right]^2\, d\boldsymbol{\theta} \leq 2\int_\Omega [\pi_1(\boldsymbol{\theta}) + \pi_2(\boldsymbol{\theta})]\, d\boldsymbol{\theta} = 4. \tag{5.A.21}$$

Thus, (5.5.12) follows from (5.A.20) and (5.A.21). □

Proof of Theorem 5.5.5. 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 (5.A.15), we have

$$\begin{aligned}
&nE\left(\frac{(\hat{r}_{\mathrm{RIS},n} - r)^2}{r^2}\,\middle|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_n\right) \\
&\qquad - r^{-2}\, c_2^{-4} \int_\Omega \{c_2 f_n(\boldsymbol{\theta}) - c_1 g_n(\boldsymbol{\theta})\}^2 \psi_n(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\quad = c_2^{-4} r^{-2}\, n^{-1}\, E\left\{\frac{\left(\sum_{i=1}^n (c_2 f_n(\boldsymbol{\vartheta}_{n,i}) - c_1 g_n(\boldsymbol{\vartheta}_{n,i}))\right)^2}{(\sum_{i=1}^n g_n(\boldsymbol{\vartheta}_{n,i}))^2}\right. \\
&\qquad \left. \times \sum_{i=1}^n (c_2 - g_n(\boldsymbol{\vartheta}_{n,i})) \cdot \sum_{i=1}^n (c_2 + g_n(\boldsymbol{\vartheta}_{n,i}))\,\middle|\, \tau_n\right\} \\
&\quad \stackrel{\text{def}}{=} c_2^{-4} r^{-2} \eta_n.
\end{aligned}$$

By the law of large numbers, we have

$$\tau_n \to r \text{ a.s. as } n \to \infty, \tag{5.A.22}$$

and hence

$$\begin{aligned}
&\lim_{n\to\infty} r^{-2} c_2^{-4} \int_\Omega \{c_2 f_n(\boldsymbol{\theta}) - c_1 g_n(\boldsymbol{\theta})\}^2 \psi_n(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\qquad = \left[\int_\Omega \left|\frac{p_1(\boldsymbol{\theta})}{c_1} - \frac{p_2(\boldsymbol{\theta})}{c_2}\right|\, d\boldsymbol{\theta}\right]^2 \text{ a.s.}
\end{aligned}$$

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

$$\eta_n \to 0 \text{ a.s. as } n \to \infty. \tag{5.A.23}$$

Let

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

Note that

$$\begin{aligned}
|\eta_n| &= \left| E\left\{ \frac{T_n^2 \cdot (nc_2 - G_n) \cdot (nc_2 + G_n)}{nG_n^2} \middle| \tau_n \right\} \right| \\
&\le 6n^{-1} E\{T_n^2 1\{|T_n| \ge n^{2/3}\} | \tau_n\} \\
&\quad + 6(nc_2)^{-1} n^{-1} E\{T_n^2 |nc_2 - G_n| 1\{G_n \ge nc_2/2\} 1\{|T_n| \ge n^{2/3}\} | \tau_n\} \\
&\quad + 2(nc_2)^2 n^{-1} E\{(T_n/G_n)^2 1\{G_n \le nc_2/2\} | \tau_n\} \\
&\le n^{-1} E\{T_n^2 1\{|T_n| \ge n^{2/3}\} | \tau_n\} + 6c_2^{-1} n^{-2/3} E\{|nc_2 - G_n| | \tau_n\} \\
&\quad + 2(nc_2)^2 n^{-1} E\{(T_n/G_n)^2 1\{G_n \le nc_2/2\} | \tau_n\} \\
&\stackrel{\text{def}}{=} \eta_{n,1} + \eta_{n,2} + \eta_{n,3}.
\end{aligned} \tag{5.A.24}$$

Since T_n is a partial sum of i.i.d random variables under the given τ_n, by (5.A.22) and (ii), we have

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

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

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

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

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

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

$$\inf_{t > 0} e^{-ta} E[\exp(t(\mu - X))] \le \exp\left(-\frac{a^2}{4(EX^2 + e^{4\mu})}\right). \tag{5.A.25}$$

By (5.A.25) and similar to (5.A.13), we have

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

Thus, in terms of (5.A.22) and the conditions (ii) and (iii),

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

Putting the above inequalities together yields (5.A.23). This proves the theorem. □

Proof of Theorem 5.5.6. Let

$$
\zeta(x,t) = \frac{q_1(x)}{\psi q_1(x) + (1-\psi) t q_2(x)}.
$$

Since $S_n(\hat{r}_{\mathrm{BS}}, n) = 0$, we have

$$
\sum_{i=1}^{n} \zeta(\boldsymbol{\theta}_i, \hat{r}_{\mathrm{BS},n}) = n.
$$

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

$$
\{\hat{r}_{\mathrm{BS},n} \ge x\} = \left\{\sum_{i=1}^{n} \zeta(\boldsymbol{\theta}_i, x) \ge n\right\}. \tag{5.A.26}
$$

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

$$
P(\hat{r}_{\mathrm{BS},n} \ge r + \varepsilon, \text{i.o.}) = P\left(\sum_{i=1}^{n} \zeta(\boldsymbol{\theta}_i, r+\varepsilon) \ge n, \text{i.o.}\right)
$$

and

$$
P(\hat{r}_{\mathrm{BS},n} \le r - \varepsilon, \text{i.o.}) = P\left(\sum_{i=1}^{n} \zeta(\boldsymbol{\theta}_i, r-\varepsilon) \le n, \text{i.o.}\right).
$$

Noting that for $x > 0$

$$E_{\pi_{\text{mix}}}\zeta(\boldsymbol{\theta}, x) = \int_\Omega \frac{q_1(\boldsymbol{\theta})(\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta}))}{\psi q_1(\boldsymbol{\theta}) + (1-\psi)x\, q_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}$$
$$\begin{cases} < 1 & \text{if } x > r, \\ = 1 & \text{if } x = r, \\ > 1 & \text{if } x < r, \end{cases} \tag{5.A.27}$$

and by the strong law of large numbers, we have

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

and

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

This proves (5.5.18). Write $\lambda(x) = E_{\pi_{\text{mix}}}(\zeta(\boldsymbol{\theta}, x) - 1)$. Then, by (5.A.27), $\lambda(r) = 0$ and

$$\dot{\lambda}(x) = \frac{d\lambda(x)}{dx} = -(1-\psi)\int_\Omega \frac{q_1(\boldsymbol{\theta})q_2(\boldsymbol{\theta})(\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta}))}{(\psi q_1(\boldsymbol{\theta}) + (1-\psi)xq_2(\boldsymbol{\theta}))^2}\, d\boldsymbol{\theta}.$$

In particular,

$$\dot{\lambda}(r) = -(1-\psi)(c_2/c_1)\int_\Omega \frac{\pi_1(\boldsymbol{\theta}) \cdot \pi_2(\boldsymbol{\theta})}{\psi\pi_1(\boldsymbol{\theta}) + (1-\psi)\pi_2(\boldsymbol{\theta})}\, d\boldsymbol{\theta}.$$

By a strong Bahadur representation of He and Shao (1996) or Janssen, Jureckova, and Veraverbeke (1985),

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

which implies immediately, by the central limit theorem,

$$\sqrt{n}(\hat{r}_{\text{BS},n} - r) \xrightarrow{\mathcal{D}} N\big(0, \sigma^2\big), \tag{5.A.28}$$

where

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

In terms of (5.A.26), as in the proof of Theorem 5.5.1, one can show that $\{n(\hat{r}_{\text{BS},n} - r)^2, n \geq 1\}$ is uniformly integrable. Thus, (5.5.19) follows from

(5.A.28). $\square$

Proof of Theorem 5.8.2. We prove the theorem in turn for IS, BS, and RIS.

For IS, from Lemma 5.8.1, we take $h(y) = y - 1$, which is an increasing function of y, and $g(x) = x^2$, which is convex. Therefore, Theorem 5.8.1 implies that the lower bound of $\mathrm{ARE}^2(\hat{r}_{\mathrm{IS}}(w))$ is $\int_{\Omega_1} \pi_1^2(\boldsymbol{\theta})/\pi_{21}(\boldsymbol{\theta})\, d\boldsymbol{\theta} - 1$. Since the equality holds in (I) of Theorem 5.8.1, this lower bound is attained at $w = \pi_2(\psi|\boldsymbol{\theta})$. This proves the optimality result for IS.

For BS, analogous to the proof of Theorem 5.3.1, by Lemma 5.8.2 and the Cauchy–Schwarz inequality, for all $\alpha(\boldsymbol{\theta}, \psi)$,

$$\begin{aligned}&\mathrm{ARE}^2(\hat{r}_{\mathrm{BS}}(w,\alpha))\\ &\quad\geq \frac{1}{s_1 s_2}\left\{\left(\int_{\Theta_1\cap\Theta_2} \frac{\pi_1(\boldsymbol{\theta})w(\psi|\boldsymbol{\theta})\pi_2(\boldsymbol{\theta},\psi)}{s_1\pi_1(\boldsymbol{\theta})w(\psi|\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta},\psi)}\, d\boldsymbol{\theta}\, d\psi\right)^{-1} - 1\right\}.\end{aligned}$$

We take $h(y) = (1/s_1 s_2)(1/y - 1)$ and $g(x) = x/(s_1 x + s_2)$. Then $h(y)$ is a decreasing function of y and $g''(x) = -2s_1 s_2/(s_1 x + s_2)^3 < 0$ which implies that g is concave. Therefore, Theorem 5.8.1 yields that the lower bound of $\mathrm{ARE}^2(\hat{r}_{\mathrm{BS}}(w,\alpha))$ is

$$\frac{1}{s_1 s_2}\left\{\left(\int_{\Omega_1\cap\Omega_2} \frac{\pi_1(\boldsymbol{\theta})\pi_{21}(\boldsymbol{\theta})}{s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_{21}(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right)^{-1} - 1\right\}. \tag{5.A.29}$$

Although the equality does not hold in (I) of Theorem 5.8.1, it can be easily verified that the lower bound (5.A.29) is attained at $w = w_{\mathrm{opt}}^{\mathrm{BS}}$ and $\alpha = \alpha_{\mathrm{opt}}$. This proves Theorem 5.8.2 for BS.

Finally, for RIS, by Lemma 5.8.3 and the Cauchy–Schwarz inequality, for an arbitrary density π,

$$\mathrm{ARE}^2(\hat{r}_{\mathrm{RIS}}(w,\pi)) \geq \left[\int_{\Theta_1\cup\Theta_2} |\pi_1(\boldsymbol{\theta})w(\psi|\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta},\psi)|\, d\boldsymbol{\theta}\, d\psi\right]^2. \tag{5.A.30}$$

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 5.8.1 the lower bound of $\mathrm{ARE}^2(\hat{r}_{\mathrm{RIS}}(w,\pi))$ is

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

Note that since the region of integration on the right side of inequality (5.A.30) is bigger than the support of π_2, Theorem 5.8.1 needs an obvious adjustment. Plugging $w = w_{\mathrm{opt}}^{\mathrm{RIS}}$ and $\pi = \pi_{\mathrm{opt}}$ into (5.8.7) leads to (5.8.11). This completes the proof of Theorem 5.8.2. $\square$

Proof of Theorem 5.11.1. Write

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

and

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

It follows from the law of large numbers that

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

By the assumption that $n = o(m)$, we have

$$E(R_2^2) + E(R_3^2) + E(R_4^2) = o(1/n)$$

and

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

by (5.11.7). This proves (5.11.11) by the above inequalities. □

Exercises

5.1 For $\hat{r}_{\mathrm{IS}_2}$ given in (5.2.5), show that

$$nr^{-2}\,\mathrm{Var}(\hat{r}_{\mathrm{IS}_2}) = E_2\left(\frac{\pi_1(\boldsymbol{\theta}) - \pi_2(\boldsymbol{\theta})}{\pi_2(\boldsymbol{\theta})}\right)^2 \geq \frac{\left[E_1(\sqrt{\pi_1(\boldsymbol{\theta})})\right]^2}{E_2(\pi_1(\boldsymbol{\theta}))} - 1.$$

[*Hint*: Use the Cauchy–Schwartz inequality.]
This result implies that if the two densities π_1 and π_2 have very little overlap, i.e., $E_2(\pi_1(\boldsymbol{\theta}))$ is small, then the variance, $\mathrm{Var}(\hat{r}_{\mathrm{IS}_2})$, of $\hat{r}_{\mathrm{IS}_2}$ is large, and therefore, this importance sampling-based method works poorly.

5.2 Prove the identity given in (5.3.1).

5.3 GEOMETRIC BRIDGE
Let $\alpha_G(\boldsymbol{\theta}) = [q_1(\boldsymbol{\theta})q_2(\boldsymbol{\theta})]^{-1/2}$. With $\alpha(\boldsymbol{\theta}) = \alpha_G(\boldsymbol{\theta})$, the resulting BS estimator $\hat{r}_{\mathrm{BS}}$ given in (5.3.3) is called a geometric bridge sampling (GBS) estimator of r. Show that $\mathrm{RE}^2(\hat{r}_{\mathrm{BS}})$ given in (5.3.4) reduces to

$$\mathrm{RE}_G^2 = \frac{1}{ns_1s_2}\left\{\frac{\int_{\Omega_1\cap\Omega_2}[s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})]\,d\boldsymbol{\theta}}{(\int_{\Omega_1\cap\Omega_2}[\pi_1(\boldsymbol{\theta})\pi_2(\boldsymbol{\theta})]^{1/2}\,d\boldsymbol{\theta})^2} - 1\right\} + o\Big(\frac{1}{n}\Big). \tag{5.E.1}$$

Further show that the first term on the right side of (5.E.1) is equal to

$$\frac{1}{ns_1s_2}\left\{\frac{\int_{\Omega_1\cap\Omega_2}[s_1\pi_1(\boldsymbol{\theta}) + s_2\pi_2(\boldsymbol{\theta})]\,d\boldsymbol{\theta}}{(1 - \frac{1}{2}H^2(\pi_1,\pi_2))^2} - 1\right\},$$

where $H(\pi_1,\pi_2)$ is the Hellinger divergence defined in (5.4.7).

5.4 POWER FAMILY BRIDGE
Let

$$\alpha_{k,A}(\boldsymbol{\theta}) = [q_1^{1/k}(\boldsymbol{\theta}) + (Aq_2(\boldsymbol{\theta}))^{1/k}]^{-k}.$$

With $\alpha(\boldsymbol{\theta}) = \alpha_{k,A}(\boldsymbol{\theta})$, the resulting BS estimator $\hat{r}_{\mathrm{BS}}$ given in (5.3.3) is called a power family bridge sampling (PFBS) estimator of r. Show that:

(i) $\lim_{k\to\infty} 2^k \alpha_{k,A}(\boldsymbol{\theta}) = [Aq_1(\boldsymbol{\theta})q_2(\boldsymbol{\theta})]^{-1/2}$, which implies that when k approaches infinity, the PFBS estimator approaches the GBS estimator.

(ii) $\lim_{k\to 0} \alpha_{k,A}(\boldsymbol{\theta}) = 1/\max\{q_1(\boldsymbol{\theta}), Aq_2(\boldsymbol{\theta})\}$.

5.5 Prove the identity given by (5.4.1).

5.6 Prove Theorem 5.4.1.

5.7 A family of random variables $\{X_t,\ t \in T\}$ is said to be uniformly integrable if

$$\lim_{A\to\infty} \sup_{t\in T} E|X_t| 1\{|X_t| > A\} = 0.$$

Prove that $\{X_t,\ t \in T\}$ is uniformly integrable if $\sup_{t\in T} E|X_t|^q < \infty$ for some $q > 1$.

5.8 Prove that if $\{T_n, n \geq 1\}$ and $\{S_n, n \geq 1\}$ are uniformly integrable, so is $\{T_n + S_n,\ n \geq 1\}$.

5.9 Let $X_1, X_2, \ldots$ be i.i.d. random variables with $EX_i = 0$ and $EX_i^2 = 1$. Prove that $\{n^{-1}S_n^2,\ n \geq 1\}$ is uniformly integrable where $S_n = \sum_{i=1}^n X_i$.

5.10 Verify (5.A.26).

5.11 Consider the normal family $N(\mu(t), \sigma^2(t))$.

(i) Show that the Euler–Lagrange equation given in (5.4.9) with $k = 1$ reduces to

$$\ddot{\mu}(t) = 0. \tag{5.E.2}$$

(ii) Given $\sigma^2 = 1$, and the boundary conditions $\mu(0) = 0$ and $\mu(t) = \delta$, find the solution of the Euler–Lagrange equation given in (5.E.2).

(iii) Show that the Euler–Lagrange equation given in (5.4.9) with $k = 2$ becomes

$$\begin{cases} \dot{\mu}(t) - c_0\sigma^2(t) = 0, \\ 3\ddot{\sigma}(t)\sigma(t) - 3\dot{\sigma}^2(t) + \dot{\mu}^2(t) = 0, \end{cases} \tag{5.E.3}$$

where c_0 is a constant to be determined from the boundary conditions.

(iv) Given the boundary conditions:

$$(\mu(0), \sigma^2(0))' = (0, 1)' \ \text{ and } \ (\mu(1), \sigma^2(1))' = (\delta, 1)',$$

find the solution of the differential equation (5.E.3).

5.12 Derive Table 5.1.

5.13 A SIMULATION STUDY
Consider Case 1 of Section 5.6.

(i) Use the inverse cdf method of Devroye (1986, pp. 27–35) to generate a random sample of size n from the optimal RIS cumulative distribution function $\Pi_{\text{opt}}(\boldsymbol{\theta})$ given in (5.6.3) and compute $\hat{r}_{\text{RIS}}$ given in (5.5.2) with the optimal $\pi_{\text{opt}}(\boldsymbol{\theta})$ given in (5.6.1).

(ii) Repeat (i) m times and then use the standard macro-repetition simulation technique to obtain an estimate of $nE(\hat{r}-r)^2/r^2$. (*Hint*: Here $r=1$.)

(iii) Compare your estimates to the theoretical result given in Table 5.1 for different values of n and m. Discuss your findings from this simulation study.

5.14 Derive Table 5.2.

5.15 Prove Lemmas 5.8.1, 5.8.2, and 5.8.3.

5.16 Prove Theorem 5.8.1.

5.17 Assuming that $\Psi(\boldsymbol{\theta}) = \Psi \subset R^m$ for all $\boldsymbol{\theta} \in \Omega_2$ and $\Omega_1 \subset \Omega_2$, we have the identity

$$r = E_{\pi_2}\{q_2(\boldsymbol{\theta}^*, \boldsymbol{\psi})q_1(\boldsymbol{\theta})/q_2(\boldsymbol{\theta}, \boldsymbol{\psi})\}/c(\boldsymbol{\theta}^*),$$

where $\pi_2(\boldsymbol{\theta}, \boldsymbol{\psi}) \propto q_2(\boldsymbol{\theta}, \boldsymbol{\psi})$, $c(\boldsymbol{\theta}^*) = \int_\Psi q_2(\boldsymbol{\theta}^*, \boldsymbol{\psi})\, d\boldsymbol{\psi}$, and $\boldsymbol{\theta}^* \in \Omega_2$ is a fixed point. Thus, a marginal-likelihood estimator of r can be defined by

$$\hat{r}_{\text{ML}} = \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{q_2(\boldsymbol{\theta}^*, \boldsymbol{\psi}_i)q_1(\boldsymbol{\theta}_i)}{q_2(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i)}\right\} \cdot \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{w^*(\boldsymbol{\psi}_i^*|\boldsymbol{\theta}^*)}{p_2(\boldsymbol{\theta}^*, \boldsymbol{\psi}_i^*)}\right\},$$

where $\{(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i),\ i=1,2,\ldots,n\}$ and $\{\boldsymbol{\psi}_i^*,\ i=1,2,\ldots,n\}$ are two independent random samples from $\pi_2(\boldsymbol{\theta}, \boldsymbol{\psi})$ and $\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta}^*)$ (the conditional density of $\boldsymbol{\psi}$ given $\boldsymbol{\theta}^*$), respectively, and $w^*(\boldsymbol{\psi}|\boldsymbol{\theta}^*)$ is an arbitrary (completely known) density defined on Ψ.

(a) Verify that

$$\begin{aligned}\text{Var}(\hat{r}_{\text{ML}}) = r^2 &\left[\frac{1}{n}\left\{\int_{\Omega_1} \frac{\pi_1^2(\boldsymbol{\theta})}{\pi_{21}(\boldsymbol{\theta})}\left(\int_\Psi \frac{\pi_2^2(\boldsymbol{\psi}|\boldsymbol{\theta}^*)}{\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta})}\, d\boldsymbol{\psi}\right) d\boldsymbol{\theta} - 1\right\} + 1\right] \\ &\times \left[\frac{1}{n}\left\{\int_\Psi \frac{w^{*2}(\boldsymbol{\psi}|\boldsymbol{\theta}^*)}{\pi_2(\boldsymbol{\psi}|\boldsymbol{\theta}^*)}\, d\boldsymbol{\psi} - 1\right\} + 1\right] - r^2.\end{aligned}$$

(b) Further show that for all $w^*(\boldsymbol{\psi}|\boldsymbol{\theta}^*)$

$$\text{Var}(\hat{r}_{\text{ML}}) \geq \text{Var}(\hat{r}_{\text{OIS}}),$$

where $\text{Var}(\hat{r}_{\text{OIS}}) = (r^2/n)\text{ARE}^2(\hat{r}_{\text{OIS}})$ given in (5.8.9). Hence, $\hat{r}_{\text{ML}}$ is not as good as $\hat{r}_{\text{OIS}}$.

5.18 Prove the Savage–Dickey density ratio given in (5.10.5) and the generalized Savage–Dickey density ratio given in (5.10.6). Also show that (5.10.5) is a special case of (5.10.6).

5.19 Similar to the IS estimator $\hat{r}_{\text{IS}_2}$, derive the weighted versions of the BS estimator $\hat{r}_{\text{BS}}$ and the RIS estimator $\hat{r}_{\text{RIS}}$ given by (5.3.3) and (5.5.2), respectively.

http://www.springer.com/978-0-387-98935-8

Monte Carlo Methods in Bayesian Computation
Chen, M.-H.; Shao, Q.-M.; Ibrahim, J.G.
2000, XIII, 387 p., Hardcover
ISBN: 978-0-387-98935-8