2
Markov Chain Monte Carlo Sampling

Recently, Monte Carlo (MC) based sampling methods for evaluating high-dimensional posterior integrals have been rapidly developing. Those sampling methods include MC importance sampling (Hammersley and Handscomb 1964; Ripley 1987; Geweke 1989; Wolpert 1991), Gibbs sampling (Geman and Geman 1984; Gelfand and Smith 1990), Hit-and-Run sampling (Smith 1984; Bélisle, Romeijn, and Smith 1993; Chen 1993; Chen and Schmeiser 1993 and 1996), Metropolis–Hastings sampling (Metropolis et al. 1953; Hastings 1970; Green 1995), and hybrid methods (e.g., Müller 1991; Tierney 1994; Berger and Chen 1993). A general discussion of the Gibbs sampler and other Markov chain Monte Carlo (MCMC) methods is given in the *Journal of the Royal Statistical Society, Series B* (1993), and an excellent roundtable discussion on the practical use of MCMC can be found in Kass et al. (1998). Other discussions or instances of the use of MCMC sampling can be found in Tanner and Wong (1987), Tanner (1996), Geyer (1992), Gelman and Rubin (1992), Gelfand, Smith, and Lee (1992), Gilks and Wild (1992), and many others. Further development of state-of-the-arts MCMC sampling techniques include the accelerated MCMC sampling of Liu and Sabatti (1998, 1999), Liu (1998), and Liu and Wu (1997), and the exact MCMC sampling of Green and Murdoch (1999). Comprehensive accounts of MCMC methods and their applications may also be found in Meyn and Tweedie (1993), Tanner (1996), Gilks, Richardson, and Spiegelhalter (1996), Robert and Casella (1999), and Gelfand and Smith (2000). The purpose of this chapter is to give a brief overview of several commonly used MCMC sampling algorithms as well as to present selectively several newly developed computational tools for MCMC sampling.

2.1 Gibbs Sampler

The Gibbs sampler may be one of the best known MCMC sampling algorithms in the Bayesian computational literature. As discussed in Besag and Green (1993), the Gibbs sampler is founded on the ideas of Grenander (1983), while the formal term is introduced by Geman and Geman (1984). The primary bibliographical landmark for Gibbs sampling in problems of Bayesian inference is Gelfand and Smith (1990). A similar idea termed as *data augmentation* is introduced by Tanner and Wong (1987). Casella and George (1992) provide an excellent tutorial on the Gibbs sampler.

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ be a p-dimensional vector of parameters and let $\pi(\boldsymbol{\theta}|D)$ be its posterior distribution given the data D. Then, the basic scheme of the Gibbs sampler is given as follows:

Gibbs Sampling Algorithm

Step 0. Choose an arbitrary starting point $\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{p,0})'$, and set $i = 0$.

Step 1. Generate $\boldsymbol{\theta}_{i+1} = (\theta_{1,i+1}, \theta_{2,i+1}, \ldots, \theta_{p,i+1})'$ as follows:

- Generate $\theta_{1,i+1} \sim \pi(\theta_1|\theta_{2,i}, \ldots, \theta_{p,i}, D)$;
- Generate $\theta_{2,i+1} \sim \pi(\theta_2|\theta_{1,i+1}, \theta_{3,i}, \ldots, \theta_{p,i}, D)$;
  ... ...
- Generate $\theta_{p,i+1} \sim \pi(\theta_p|\theta_{1,i+1}, \theta_{2,i+1}, \ldots, \theta_{p-1,i+1}, D)$.

Step 2. Set $i = i + 1$, and go to Step 1.

Thus each component of $\boldsymbol{\theta}$ is visited in the natural order and a cycle in this scheme requires generation of p random variates. Gelfand and Smith (1990) show that under certain regularity conditions, the vector sequence $\{\boldsymbol{\theta}_i,\ i = 1, 2, \ldots\}$ has a stationary distribution $\pi(\boldsymbol{\theta}|D)$. Schervish and Carlin (1992) provide a sufficient condition that guarantees geometric convergence. Other properties regarding geometric convergence are discussed in Roberts and Polson (1994). To illustrate the Gibbs sampler, we consider the following two simple examples:

Example 2.1. Bivariate normal model. The purpose of this example is to examine the exact correlation structure of the Markov chain induced by the Gibbs sampler. Assume that the posterior distribution $\pi(\boldsymbol{\theta}|D)$ is a bivariate normal distribution $N_2(\boldsymbol{\mu}, \Sigma)$ with

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

where μ_j, σ_j, $j = 1, 2$, and ρ are known. Then the Gibbs sampler requires sampling from

$$\theta_1 \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(\theta_2 - \mu_2), \sigma_1^2(1-\rho^2)\right)$$

and

$$\theta_2 \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\theta_1 - \mu_1), \sigma_2^2(1-\rho^2)\right).$$

Let $\{\boldsymbol{\theta}_i = (\theta_{1,i}, \theta_{2,i})', \ i \geq 0\}$ denote the Markov chain induced by the Gibbs sampler for the above bivariate normal distribution. If we start from the stationary distribution, i.e., $\boldsymbol{\theta}_0 \sim N(\mu, \Sigma)$, then each of $\{\theta_{1,i}, \ i \geq 0\}$ and $\{\theta_{2,i}, \ i \geq 0\}$ is an AR(1) process.

To see this, let $\{z_{1,i}, z_{2,i}, \ i \geq 0\}$ be an i.i.d. $N(0,1)$ random variable sequence. Then the structure of the Gibbs sampler implies

$$\begin{aligned}
\theta_{1,0} &= \mu_1 + \sigma_1 z_{1,0}, \\
\theta_{2,0} &= \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\theta_{1,0} - \mu_1) + \sigma_2\sqrt{1-\rho^2}z_{2,0},
\end{aligned}$$

and

$$\begin{aligned}
\theta_{1,i+1} &= \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(\theta_{2,i} - \mu_2) + \sigma_1\sqrt{1-\rho^2}z_{1,i+1}, \\
\theta_{2,i+1} &= \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(\theta_{1,i+1} - \mu_1) + \sigma_2\sqrt{1-\rho^2}z_{2,i+1},
\end{aligned} \tag{2.1.1}$$

for $i \geq 0$, Now, we consider the first component $\theta_{1,i+1}$. From (2.1.1), for $i \geq 0$,

$$\begin{aligned}
\theta_{1,i+1} =& \mu_1 + \rho\frac{\sigma_1}{\sigma_2}\left[\rho\frac{\sigma_2}{\sigma_1}(\theta_{1,i} - \mu_1) + \sigma_2\sqrt{1-\rho^2}z_{2,i}\right] \\
&+ \sigma_1\sqrt{1-\rho^2}z_{1,i+1} \\
=& \mu_1 + \rho^2(\theta_{1,i} - \mu_1) + \rho\sigma_1\sqrt{1-\rho^2}z_{2,i} \\
&+ \sigma_1\sqrt{1-\rho^2}z_{1,i+1}.
\end{aligned} \tag{2.1.2}$$

Let $\psi = \rho^2$ and $\sigma_1^{*2} = \sigma_1^2(1-\rho^4)$. Let $\{z_i^*, \ i \geq 0\}$ denote an i.i.d. $N(0,1)$ random variable sequence. Since $z_{1,i}$ and $z_{2,i+1}$ are independently and identically distributed as $N(0,1)$, then we can rewrite (2.1.2) as

$$\theta_{1,0} = \mu_1 + \sigma_1 z_0^*, \tag{2.1.3}$$

$$\theta_{1,i+1} = \mu_1 + \psi(\theta_{1,i} - \mu_1) + \sigma_1^* z_{i+1}^* \quad \text{for } i \geq 0. \tag{2.1.4}$$

Thus, $\{\theta_{1,i}, \ i \geq 0\}$ is an AR(1) process with lag-one autocorrelation $\psi = \rho^2$. Similarly, $\{\theta_{2,i}, \ i \geq 0\}$ is also an AR(1) process with lag-one autocorrelation $\psi = \rho^2$. The only difference is that we use $\sigma_2^* = \sigma_2\sqrt{1-\rho^4}$ instead of σ_1^* in (2.1.4), and use μ_2 and σ_2 instead of μ_1 and σ_1 in (2.1.3).

Roberts and Sahu (1997) obtain a similar result for a general multivariate normal target distribution $\pi(\theta|D)$, that is, the Markov chain induced by the Gibbs sampler is a multivariate AR(1) process.

Example 2.2. Constrained multiple linear regression model. We consider a constrained multiple linear regression model given by (1.3.1) to model the New Zealand apple data described in Example 1.1. Let

$$\Omega = \{\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{10})' : 0 \leq \beta_1 \leq \beta_2 \leq \ldots \leq \beta_{10}, \boldsymbol{\beta} \in R^{10}\} \quad (2.1.5)$$

denote the constraints given in (1.3.2). We take a joint prior for $(\boldsymbol{\beta}, \sigma^2)$ of the form

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}\pi(\beta_{10}|\mu_{10}, \sigma^2_{10}), \quad (2.1.6)$$

for $\sigma^2 > 0$ and $\boldsymbol{\beta} \in \Omega$, where $\pi(\beta_{10}|\mu_{10}, \sigma^2_{10})$ is a normal density with mean μ_{10} and variance σ^2_{10}. The modification of the usual flat noninformative prior to include the informative distribution on β_{10} is necessary to prevent too much weight being given to the unconstrained and therefore unbounded parameter β_{10}. Chen and Deely (1996) specify $\mu_{10} = 0.998$, and $\sigma^2_{10} = 0.089$ by using method-of-moments, a well–known type of empirical Bayes estimation, from the data on growers with mature trees only. Using (2.1.6), the posterior distribution for $(\boldsymbol{\beta}, \sigma^2)$ based on the New Zealand apple data D is given by

$$\pi(\boldsymbol{\beta}, \sigma^2|D) = \frac{\exp\{-(\beta_{10} - \mu_{10})^2/2\sigma^2_{10}\}}{c(D)(\sigma^2)^{(n+1)/2}} \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{10} x_{ij}\beta_j\right)^2\right\}, \quad (2.1.7)$$

for $\sigma^2 > 0$ and $\boldsymbol{\beta} \in \Omega$, where y_i is the total number of cartons of fruit produced and x_{ij} = number of trees at age $j+1$ for $j = 1, 2, \ldots, 10$ for the i^{th} grower, $c(D)$ is the normalizing constant, and $n = 207$ denotes the sample size. Due to the constraints, the analytical evaluation of posterior quantities such as the posterior mean and posterior standard deviation of β_j does not appear possible. However, the implementation of the Gibbs sampler for sampling from the posterior (2.1.7) is straightforward. More specifically, we run the Gibbs sampler by taking

$$\beta_j|\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_{10}, \sigma^2, D \sim N(\theta_j, \delta_j^2) \quad (2.1.8)$$

subject to $\beta_{j-1} \leq \beta_j \leq \beta_{j+1}$ $(\beta_0 = 0)$ for $j = 1, 2, \ldots, 9$,

$$\beta_{10}|\beta_1, \ldots, \beta_9, \sigma^2, D \sim N(\psi\theta_{10} + (1-\psi)\mu_{10}, (1-\psi)\sigma^2_{10}) \quad (2.1.9)$$

subject to $\beta_{10} \geq \beta_9$ and

$$\sigma^2|\boldsymbol{\beta}, D \quad \sim \quad \mathcal{IG}\left(\frac{n}{2}, \frac{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{10} x_{ij}\beta_j)^2}{2}\right), \tag{2.1.10}$$

where in (2.1.8) and (2.1.9), $\psi = \sigma_{10}^2/(\sigma_{10}^2 + \delta_{10}^2)$,

$$\theta_j = \left(\sum_{i=1}^{n} x_{ij}^2\right)^{-1} \left[\sum_{i=1}^{n}\left(y_i - \sum_{l \neq j} x_{il}\beta_l\right) x_{ij}\right], \tag{2.1.11}$$

and

$$\delta_j^2 = \left(\sum_{i=1}^{n} x_{ij}^2\right)^{-1} \sigma^2 \tag{2.1.12}$$

for $j = 1, \ldots, 10$, and $\mathcal{IG}(\xi, \eta)$ denotes the inverse gamma distribution with parameters (ξ, η), whose density is given by

$$\pi(\sigma^2|\xi, \eta) \;\propto\; (\sigma^2)^{-(\xi+1)} e^{-\eta/\sigma^2}.$$

2.2 Metropolis–Hastings Algorithm

The Metropolis–Hastings algorithm is developed by Metropolis et al. (1953) and subsequently generalized by Hastings (1970). Tierney (1994) gives a comprehensive theoretical exposition of this algorithm, and Chib and Greenberg (1995) provide an excellent tutorial on this topic.

Let $q(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ be a proposal density, which is also termed as a *candidate-generating density* by Chib and Greenberg (1995), such that

$$\int q(\boldsymbol{\theta}, \boldsymbol{\vartheta})\, d\boldsymbol{\vartheta} = 1.$$

Also let $U(0, 1)$ denote the uniform distribution over $(0, 1)$. Then, a general version of the Metropolis–Hastings algorithm for sampling from the posterior distribution $\pi(\boldsymbol{\theta}|D)$ can be described as follows:

Metropolis–Hastings Algorithm

Step 0. Choose an arbitrary starting point $\boldsymbol{\theta}_0$ and set $i = 0$.

Step 1. Generate a candidate point $\boldsymbol{\theta}^*$ from $q(\boldsymbol{\theta}_i, \cdot)$ and u from $U(0, 1)$.

Step 2. Set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$ if $u \leq a(\boldsymbol{\theta}_i, \boldsymbol{\theta}^*)$ and $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$ otherwise, where the acceptance probability is given by

$$a(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = \min\left\{\frac{\pi(\boldsymbol{\vartheta}|D)q(\boldsymbol{\vartheta}, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|D)q(\boldsymbol{\theta}, \boldsymbol{\vartheta})}, 1\right\}. \tag{2.2.1}$$

Step 3. Set $i = i + 1$, and go to Step 1.

The above algorithm is very general. When $q(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = q(\boldsymbol{\vartheta})$, the Metropolis–Hastings algorithm reduces to the *independence chain* Metropolis algorithm (see Tierney 1994). More interestingly, the Gibbs sampler is obtained as a special case of the Metropolis–Hastings algorithm by choosing an appropriate $q(\boldsymbol{\theta}, \boldsymbol{\vartheta})$. This relationship is first pointed out by Gelman (1992) and further elaborated on by Chib and Greenberg (1995).

Another family of proposal densities is given by the form $q(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = q_1(\boldsymbol{\vartheta} - \boldsymbol{\theta})$, where $q_1(\cdot)$ is a multivariate density (see Müller 1991). The candidate $\boldsymbol{\theta}^*$ is thus drawn according to the process $\boldsymbol{\theta}^* = \boldsymbol{\theta} + \boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is called the increment random variable and follows the distribution q_1. Because the candidate is equal to the current value plus noise, Chib and Greenberg (1995) call this case a *random walk chain.* Many other algorithms such as the Hit-and-Run algorithm and dynamic weighting algorithm, which will be presented later in this chapter, are also special cases of this general algorithm.

The performance of a Metropolis–Hastings algorithm depends on the choice of a proposal density q. As discussed in Chib and Greenberg (1995), the spread of the proposal density q affects the behavior of the chain in at least two dimensions: one is the "acceptance rate" (the percentage of times a move to a new point is made), and the other is the region of the sample space that is covered by the chain. If the spread is extremely large, some of the generated candidates will have a low probability of being accepted. On the other hand, if the spread is chosen too small, the chain will take longer to traverse the support of the density. Both of these situations are likely to be reflected in high autocorrelations across sample values. In the context of q_1 (the random walk proposal density), Roberts, Gelman, and Gilks (1997) show that if the target and proposal densities are normal, then the scale of the latter should be tuned so that the acceptance rate is approximately 0.45 in one-dimensional problems and approximately 0.23 as the number of dimensions approaches infinity, with the optimal acceptance rate being around 0.25 in six dimensions. For the *independence chain*, in which we take $q(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = q(\boldsymbol{\vartheta})$, it is important to ensure that the tails of the proposal density $q(\boldsymbol{\vartheta})$ dominate those of the target density $\pi(\boldsymbol{\theta}|D)$, which is similar to a requirement on the importance sampling function in Monte Carlo integration with importance sampling (Geweke 1989).

To illustrate the Metropolis–Hastings algorithm, we consider a problem of sampling a correlation coefficient ρ from its posterior distribution.

Example 2.3. An algorithm for sampling a correlation ρ. Assume that $D = \{\boldsymbol{y}_i = (y_{1i}, y_{2i})', \ i = 1, 2, \ldots, n\}$ is a random sample from a

bivariate normal distribution $N_2(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Assuming a uniform prior $U(-1, 1)$ for ρ, the posterior density for ρ is given by

$$\pi(\rho|D) \propto (1-\rho^2)^{-n/2} \exp\left\{ -\frac{1}{2(1-\rho^2)}(S_{11} - 2\rho S_{12} + S_{22}) \right\}, \quad (2.2.2)$$

where $-1 < \rho < 1$, and $S_{rs} = \sum_{i=1}^{n} y_{ri} y_{si}$ for $r, s = 1, 2$. Generating ρ from (2.2.2) is not trivial since $\pi(\rho|D)$ is not log-concave. Therefore, we consider the following Metropolis–Hastings algorithm with a "de-constraint" transformation to sample ρ. Since $-1 < \rho < 1$, we let

$$\rho = \frac{-1 + e^{\xi}}{1 + e^{\xi}}, \quad -\infty < \xi < \infty. \quad (2.2.3)$$

Then

$$\pi(\xi|D) = \pi(\rho|D) \frac{2e^{\xi}}{(1+e^{\xi})^2}.$$

Instead of directly sampling ρ, we generate ξ by choosing a normal proposal $N(\hat{\xi}, \hat{\sigma}^2_{\hat{\xi}})$, where $\hat{\xi}$ is a maximizer of the logarithm of $\pi(\xi|D)$, which can be obtained by, for example, the standard Newton–Raphson algorithm or the Nelder–Mead algorithm implemented by O'Neill (1971), and $\hat{\sigma}^2_{\hat{\xi}}$ is minus the inverse of the second derivative of $\log \pi(\xi|D)$ evaluated at $\xi = \hat{\xi}$, that is,

$$\hat{\sigma}^{-2}_{\hat{\xi}} = - \left. \frac{d^2 \log \pi(\xi|D)}{d\xi^2} \right|_{\xi=\hat{\xi}}.$$

The algorithm to generate ξ operates as follows:

Step 1. Let ξ be the current value.

Step 2. Generate a proposal value ξ^* from $N(\hat{\xi}, \hat{\sigma}^2_{\hat{\xi}})$.

Step 3. A move from ξ to ξ^* is made with probability

$$\min \left\{ \frac{\pi(\xi^*|D)\phi\left(\frac{\xi - \hat{\xi}}{\hat{\sigma}_{\hat{\xi}}}\right)}{\pi(\xi|D)\phi\left(\frac{\xi^* - \hat{\xi}}{\hat{\sigma}_{\hat{\xi}}}\right)}, 1 \right\}, \quad (2.2.4)$$

where ϕ is the standard normal probability density function.

After we obtain ξ, we compute ρ by using (2.2.3).

Since the above algorithm does not use a random walk proposal density, the optimal acceptance rate, 0.23, of Roberts, Gelman, and Gilks (1997)

cannot be applied here. A detailed study of how this algorithm performs is thus left as an exercise. The above algorithm can also be extended to the cases where $\pi(\rho|D)$ is a conditional posterior distribution that depends on other parameters. For example, the conditional posterior distribution for ρ may be written as $\pi(\rho|\boldsymbol{\theta}, D)$. Then, the Metropolis–Hastings algorithm to sample from $\pi(\rho|\boldsymbol{\theta}, D)$ proceeds in a similar way. The idea of a normal proposal that is matched to the conditional posterior appears for the first time in Chib and Greenberg (1994). A nice feature of this extension is that the normal proposal density for this more general case becomes adaptive since it depends on the values of the other parameters from the current and previous iterations. This semiautomatic updating feature makes the proposal density closer to the true conditional posterior, which may lead to a more efficient Metropolis–Hastings algorithm.

2.3 Hit-and-Run Algorithm

The Hit-and-Run (H&R) algorithm is a special case of the Metropolis–Hastings algorithm. Its original form is proposed independently by Boneh and Golan (1979) and Smith (1980) for generating points uniformly distributed over bounded regions in mathematical programming problems. Smith (1984) calls the H&R a "Mixing Algorithm" and he then proves the convergence of the algorithm. This algorithm has not been studied for about 10 years until Bélisle, Romeijn, and Smith (1993) propose a more general form of the H&R algorithm that generates a sample of points from an arbitrary continuous target distribution. However, Bélisle, Romeijn, and Smith (1993) prove the convergence assuming that the target density is bounded and has bounded support. Chen and Schmeiser (1996) further generalize the H&R algorithm to a general target density for evaluating multidimensional integrals and Chen and Schmeiser (1993) also consider the performance of H&R compared to the Gibbs sampler. In the context of Bayesian computation, Berger and Chen (1993) use the H&R for sampling from a multinomial distribution with a constrained parameter space; Yang and Berger (1994) apply the H&R algorithm for estimation of a covariance matrix using the reference prior; and Yang and Chen (1995) employ the H&R algorithm with parameter transformations for Bayesian analysis of random coefficient regression models using noninformative priors. A slightly different but related algorithm termed *adaptive direction sampling* can be found in Gilks, Roberts, and George (1994) and Roberts and Gilks (1994).

Assume that the posterior distribution $\pi(\boldsymbol{\theta}|D)$ has support Ω. Then, the general H&R algorithm, requiring a distribution for the direction, a density g_i for the signed distance, and an acceptance probability a_i, can be stated as follows:

Hit-and-Run Algorithm

Step 0. Choose an arbitrary starting point $\boldsymbol{\theta}_0$ and set $i = 0$.

Step 1. Generate a direction $\boldsymbol{d}_i$ from a distribution on the surface of the unit sphere.

Step 2. Find the set $\Omega_i = \Omega_i(\boldsymbol{d}_i, \boldsymbol{\theta}_i) = \{\lambda \in R | \boldsymbol{\theta}_i + \lambda \boldsymbol{d}_i \in \Omega\}$.

Step 3. Generate a signed distance λ_i from density $g_i(\lambda | \, \boldsymbol{d}_i, \boldsymbol{\theta}_i)$, where $\lambda_i \in \Omega_i$.

Step 4. Set $\boldsymbol{\theta}^* = \boldsymbol{\theta}_i + \lambda_i \boldsymbol{d}_i$. Then set

$$\boldsymbol{\theta}_{i+1} = \begin{cases} \boldsymbol{\theta}^*, & \text{with the probability } \ a_i(\boldsymbol{\theta}^* | \boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i, & \text{otherwise.} \end{cases} \tag{2.3.1}$$

Step 5. Set $i = i + 1$, and go to Step 1.

Chen and Schmeiser (1996) discuss various choices for the distributions of $\boldsymbol{d}_i$, the densities g_i, and the probabilities a_i. Let the distribution of the direction $\boldsymbol{d}_i$, as used in Step 2 of H&R, have density $r(\boldsymbol{d}_i)$, with the surface of the unit sphere as its support. Then, assume that:

(i) for any density $g_i(\lambda | \boldsymbol{d}_i, \boldsymbol{\theta}_i)$ in Step 3, $g_i(\lambda | \boldsymbol{d}_i, \boldsymbol{\theta}_i) > 0$ and

$$g_i(-\lambda | \, -\boldsymbol{d}_i, \boldsymbol{\theta}_i) = g_i(\lambda | \boldsymbol{d}_i, \boldsymbol{\theta}_i);$$

(ii) for the distribution of the direction, $r(\boldsymbol{d}_i) > 0$;

(iii) for any a_i in Step 4, $0 < a_i(\boldsymbol{\theta}^* | \boldsymbol{\theta}_i) \leq 1$; and

(iv) for any $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Omega$

$$g_i\left(\| \, \boldsymbol{\theta} - \boldsymbol{\theta}^* \, \| \, \left| \frac{\boldsymbol{\theta}^* - \boldsymbol{\theta}}{\| \, \boldsymbol{\theta}^* - \boldsymbol{\theta} \, \|}, \theta \right. \right) \cdot a_i(\boldsymbol{\theta}^* | \boldsymbol{\theta}) \pi(\boldsymbol{\theta} | D)$$
$$= g_i\left(\| \, \boldsymbol{\theta}^* - \boldsymbol{\theta} \, \| \, \left| \frac{\boldsymbol{\theta} - \boldsymbol{\theta}^*}{\| \, \boldsymbol{\theta} - \boldsymbol{\theta}^* \, \|}, \boldsymbol{\theta}^* \right. \right) \cdot a_i(\boldsymbol{\theta} | \boldsymbol{\theta}^*) \pi(\boldsymbol{\theta}^* | D).$$

Under the assumptions above, the Markov chain $\{\boldsymbol{\theta}_i, \ i = 0, 1, 2, \ldots\}$ converges to its stationary distribution $\pi(\boldsymbol{\theta} | D)$.

The most common choice of $r(\boldsymbol{d}_i)$ is a uniform distribution on the surface of the unit sphere. Common choices of g_i and a_i are given as follows:

Choice I:

$$g_i^{\mathrm{I}}(\lambda | \boldsymbol{d}_i, \boldsymbol{\theta}_i) = \frac{\pi(\boldsymbol{\theta}_i + \lambda \boldsymbol{d}_i | D)}{\int_{\Omega_i(\boldsymbol{d}_i, \boldsymbol{\theta}_i)} \pi(\boldsymbol{\theta}_i + u \boldsymbol{d}_i | D) \, du} \quad \text{for } \lambda \in \Omega_i(\boldsymbol{d}_i, \boldsymbol{\theta}_i),$$

and

$$a_i^{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i) = a_i^{\mathrm{I}}(\boldsymbol{\theta}_i|\boldsymbol{\theta}^*), \quad 0 < a^{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i) \leq 1 \text{ for all } \boldsymbol{\theta}_i, \boldsymbol{\theta}^* \in \Omega.$$

Typically $a_i^{\mathrm{I}}(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i) = 1$.

Choice II:

Choose $g_i(\lambda|\boldsymbol{d}_i, \boldsymbol{\theta}_i)$ to be one of the following:

(a) If Ω is bounded, then

$$g_i^{\mathrm{II}}(\lambda|\boldsymbol{d}_i, \boldsymbol{\theta}_i) = \frac{1}{m(\Omega_i(\boldsymbol{d}_i, \boldsymbol{\theta}_i))} \quad \text{for } \lambda \in \Omega_i(\boldsymbol{d}_i, \boldsymbol{\theta}_i),$$

where m denotes Lebesgue measure.

(b) If Ω is unbounded, then choose $g_i^{\mathrm{II}}(\lambda|\boldsymbol{d}_i, \boldsymbol{\theta}_i)$ to be a symmetric-about-zero, continuous distribution with unbounded support $\Omega_i(\boldsymbol{d}_i, \boldsymbol{\theta}_i)$ and shape depending only on $\Omega_i(\boldsymbol{d}_i, \boldsymbol{\theta}_i)$. For example, g_i^{II} can be a normal distribution, Cauchy distribution, or double-exponential distribution with location parameter zero and scale parameter depending only on $\Omega_i(\boldsymbol{d}_i, \boldsymbol{\theta}_i)$.

Independent of the choice (a) or (b), choose $a_i(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i)$ to be either:

(c) Barker's method (Barker 1965)

$$a_i^{\mathrm{II}}(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i) = \frac{\pi(\boldsymbol{\theta}^*|D)}{\pi(\boldsymbol{\theta}_i|D) + \pi(\boldsymbol{\theta}^*|D)}.$$

or

(d) Metropolis's method

$$a_i^{\mathrm{II}}(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i) = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*|D)}{\pi(\boldsymbol{\theta}_i|D)}\right).$$

Choice III:

Choose $g_i^{\mathrm{III}}(\lambda|\boldsymbol{d}_i, \boldsymbol{\theta}_i) = g_i(\boldsymbol{\theta}_i + \lambda \boldsymbol{d}_i)$, where g_i depends only on $\Omega_i(\boldsymbol{d}_i, \boldsymbol{\theta}_i)$, and

$$a_i^{\mathrm{III}}(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i) = \min\{\omega(\boldsymbol{\theta}_i + \lambda \boldsymbol{d}_i)/\omega(\boldsymbol{\theta}_i), 1\},$$

where $\omega(\boldsymbol{\theta}_i) = \pi(\boldsymbol{\theta}_i|D)/g_i(\boldsymbol{\theta}_i)$.

These choices are motivated by Hastings (1970). For a given g_i in Choice III, the results of Peskun (1973) imply that when Ω is a finite set, the choice of a_i^{III} is optimal in the sense of minimizing the asymptotic variance of the sample average $(1/n)\sum_{i=1}^n h(\boldsymbol{\theta}_i)$, where $h(\cdot)$ is a real function of $\boldsymbol{\theta}$

satisfying

$$\int_{R^p} |h(\boldsymbol{\theta})|\pi(\boldsymbol{\theta}|D)\ d\boldsymbol{\theta} < \infty.$$

With Choice I, Kaufman and Smith (1998) develop an optimal direction choice algorithm for H&R and prove that there exists a unique optimal direction choice distribution for $r(\cdot)$. The other theoretical properties of H&R can be found in Bélisle, Romeijn, and Smith (1993), and Chen and Schmeiser (1993, 1996). Regarding applications of H&R to Bayesian computation, Berger (1993) comments that

> *"This method is particularly useful when $\boldsymbol{\theta}$ has a sharply constrained parameter space."*

To illustrate the H&R algorithm, we revisit the constrained multiple linear regression model discussed in Section 2.1.

Example 2.4. Constrained multiple linear regression model (Example 2.2 continued). Instead of using the Gibbs sampler to sample $(\boldsymbol{\beta}, \sigma^2)$ from $\pi(\boldsymbol{\beta}, \sigma^2|D)$ given in (2.1.7), we use the H&R algorithm. All eleven dimensions are sampled within a Gibbs sampling framework, with the ten $\boldsymbol{\beta}$ dimensions sampled with H&R and σ^2 sampled from its known conditional gamma density in the Gibbs step. For illustrative purposes, we state the H&R logic for sampling $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{10})'$ from its conditional posterior distribution for a given value of σ^2 and D:

Step 0. Choose a starting point $\boldsymbol{\beta}_0 \in \Omega$ and set $i = 0$.

Step 1. Generate a uniformly distributed unit-length direction $\boldsymbol{d}_i = (d_{1,i}, d_{2,i}, \ldots, d_{10,i})'$.

Step 2. Find the set $\Omega_i = (R_1^i, R_2^i)$, where

$$R_1^i = \inf_{\lambda}\{\lambda : \boldsymbol{\beta}_i + \lambda \boldsymbol{d}_i \in \Omega\} \text{ and } R_2^i = \sup_{\lambda}\{\lambda : \boldsymbol{\beta}_i + \lambda \boldsymbol{d}_i \in \Omega\}.$$

Step 3. Generate a signed distance λ_i from the density

$$\pi_i(\lambda) = \frac{\pi(\boldsymbol{\beta}_i + \lambda \boldsymbol{d}_i, \sigma^2|D)}{\int_{R_1^i}^{R_2^i} \pi(\boldsymbol{\beta}_i + u\boldsymbol{d}_i, \sigma^2|D)\ du}, \quad \lambda \in (R_1^i, R_2^i). \tag{2.3.2}$$

Step 4. Set $\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i + \lambda_i \boldsymbol{d}_i$.

Step 5. Set $i = i + 1$ and go to Step 1.

Here we use the probability $a_i = 1$. Sampling in each step is straightforward. A random unit-length direction $\boldsymbol{d}_i$ can be generated in Step 2 by

independently generating $z_l \sim N(0,1)$ and setting

$$d_{l,i} = z_l \left(\sum_{j=1}^{10} z_j^2 \right)^{-1/2}$$

for $l = 1, 2, ..., 10$; see, for example, Devroye (1986). The density given in (2.3.2) is a truncated normal probability density function, where the mean and variance are easy-to-compute functions of σ^2, $\boldsymbol{\beta}_i$, $\boldsymbol{d}_i$, and D. Computationally, the H&R algorithm is slightly more efficient than the usual (one-coordinate-at-a-time) Gibbs sampler. Implementation difficulty of the two sampling algorithms is similar.

2.4 Multiple-Try Metropolis Algorithm

Liu, Liang, and Wong (1998a) propose a novel algorithm, called the Multiple-Try Metropolis (MTM) algorithm. The algorithm proceeds as follows. Let $T(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ be a proposal transition density function, which may or may not be symmetric. A requirement for $T(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is that $T(\boldsymbol{\theta}, \boldsymbol{\vartheta}) > 0$ if and only if $T(\boldsymbol{\vartheta}, \boldsymbol{\theta}) > 0$. Furthermore, define

$$w(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = \pi(\boldsymbol{\theta}|D) T(\boldsymbol{\theta}, \boldsymbol{\vartheta}) \lambda(\boldsymbol{\theta}, \boldsymbol{\vartheta}), \tag{2.4.1}$$

where $\lambda(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is a nonnegative symmetric function in $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ so that $\lambda(\boldsymbol{\theta}, \boldsymbol{\vartheta}) > 0$ whenever $T(\boldsymbol{\theta}, \boldsymbol{\vartheta}) > 0$. Suppose the current state is $\boldsymbol{\theta}_i$. In an MTM transition, the next state is generated as follows:

Multiple-Try Metropolis

Step 1. Generate k trials $\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, \ldots, \boldsymbol{\vartheta}_k$ from the proposal distribution $T(\boldsymbol{\theta}_i, \boldsymbol{\vartheta})$. Compute $w(\boldsymbol{\vartheta}_j, \boldsymbol{\theta}_i)$ for $j = 1, 2, \ldots, k$.

Step 2. Select $\boldsymbol{\vartheta}_l$ among the $\boldsymbol{\vartheta}_j$'s with probability proportional to $w(\boldsymbol{\vartheta}_j, \boldsymbol{\theta}_i)$, $j = 1, 2, \ldots, k$. Then draw $\boldsymbol{\vartheta}_1^*, \boldsymbol{\vartheta}_2^*, \ldots, \boldsymbol{\vartheta}_{k-1}^*$ from the distribution $T(\boldsymbol{\vartheta}_l, \boldsymbol{\vartheta}^*)$, and let $\boldsymbol{\vartheta}_k^* = \boldsymbol{\theta}_i$.

Step 3. Generate u from $U(0,1)$. Set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\vartheta}_l$ if $u \leq a$ and $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$ otherwise, where the acceptance probability is given by

$$a = \min \left\{ 1, \frac{w(\boldsymbol{\vartheta}_1, \boldsymbol{\theta}_i) + w(\boldsymbol{\vartheta}_2, \boldsymbol{\theta}_i) + \cdots + w(\boldsymbol{\vartheta}_k, \boldsymbol{\theta}_i)}{w(\boldsymbol{\vartheta}_1^*, \boldsymbol{\vartheta}_l) + w(\boldsymbol{\vartheta}_2^*, \boldsymbol{\vartheta}_l) + \cdots + w(\boldsymbol{\vartheta}_k^*, \boldsymbol{\vartheta}_l)} \right\}.$$

Liu, Liang, and Wong (1998a) show that the MTM transition rule satisfies the detailed balance, and hence, induces a reversible MC with $\pi(\boldsymbol{\theta}|D)$ as its equilibrium distribution. They also present several choices of $\lambda(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ in (2.4.1). When $T(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is symmetric and $\lambda(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = [T(\boldsymbol{\theta}, \boldsymbol{\vartheta})]^{-1}$, the MTM

algorithm reduces to the method of "orientation biased-Monte Carlo" described in Frenkel and Smit (1996), where they provide a specialized proof in the context of simulating molecular structures of materials. As discussed in Liu, Liang, and Wong (1998a), the MTM algorithm is more advantageous, since it allows one to explore more thoroughly the "neighboring region" defined by $T(\boldsymbol{\theta}, \boldsymbol{\vartheta})$, and it is particularly useful when one identifies certain directions of interest but has difficulty implementing a Gibbs sampling type move due to unfavorable conditional distributions. Liu, Liang, and Wong (1998a) also propose several variations of the MTM algorithm. These include a conjugate-gradient MC algorithm, a random-ray algorithm, and a Griddy–Gibbs MTM, which are closely related to the adaptive direction sampling algorithm of Gilks, Roberts, and George (1994) and Roberts and Gilks (1994), the H&R algorithm of Chen and Schmeiser (1993, 1996), and the Griddy–Gibbs algorithm of Ritter and Tanner (1992). For illustrative purposes, we briefly describe the random-ray algorithm as follows. Suppose the current state is $\boldsymbol{\theta}_i$. The random-ray algorithm executes the following update:

- Randomly generate a unit-length direction $\boldsymbol{d}$.
- Draw $\boldsymbol{\vartheta}_1$, $\boldsymbol{\vartheta}_2$, $\ldots$, $\boldsymbol{\vartheta}_k$ from the proposal transition $T_{\boldsymbol{d}}(\boldsymbol{\theta}_i, \boldsymbol{\vartheta})$ along the direction $\boldsymbol{d}$. One possible way to do this is to generate a random sample $\{r_1, r_2, \ldots, r_k\}$ from $N(0, \sigma^2)$, where σ^2 can be chosen large, and set $\boldsymbol{\vartheta}_k = \boldsymbol{\theta}_i + r_j \boldsymbol{d}$. Another approach is to generate $r_j \sim U[-\sigma, \sigma]$.
- Conduct the other MTM steps as described in the Multiple-Try Metropolis algorithm.

The implementational details for the other variations can be found in Liu, Liang, and Wong (1998a), and are omitted here for brevity.

2.5 Grouping, Collapsing, and Reparameterizations

In this section, we discuss several useful tools to improve convergence of MCMC sampling. In particular, we focus on the grouped and collapsed Gibbs techniques of Liu (1994) and Liu, Wong, and Kong (1994), and the hierarchical centering method of Gelfand, Sahu, and Carlin (1995, 1996).

2.5.1 Grouped and Collapsed Gibbs

Liu (1994) proposes a method of "grouping" and "collapsing" when using the Gibbs sampler in which he shows that both grouping and collapsing are beneficial based on operator theory. To illustrate his idea, we consider a three-dimensional posterior distribution $\pi(\boldsymbol{\theta}|D)$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$.

Liu (1994) considers the following three variations of the Gibbs sampler to sample from $\pi(\boldsymbol{\theta}|D)$:

Algorithm 1: *Standard (Original) Gibbs Sampler*

The standard Gibbs sampler requires drawing:

(i) $\theta_1 \sim \pi(\theta_1|\theta_2, \theta_3, D)$;

(ii) $\theta_2 \sim \pi(\theta_2|\theta_1, \theta_3, D)$;

(iii) $\theta_3 \sim \pi(\theta_3|\theta_1, \theta_2, D)$.

Algorithm 2: *Grouped Gibbs Sampler*

The grouped Gibbs sampler requires drawing:

(i) $(\theta_1, \theta_2) \sim \pi(\theta_1, \theta_2|\theta_3, D)$;

(ii) $\theta_3 \sim \pi(\theta_3|\theta_1, \theta_2, D)$.

In Algorithm 2, we first group (θ_1, θ_2) together and then simultaneously draw (θ_1, θ_2) from their joint conditional posterior distribution $\pi(\theta_1, \theta_2|\theta_3, D)$.

Algorithm 3: *Collapsed Gibbs Sampler*

The collapsed Gibbs sampler requires drawing:

(i) $(\theta_1, \theta_2) \sim \pi(\theta_1, \theta_2|D)$;

(ii) $\theta_3 \sim \pi(\theta_3|\theta_1, \theta_2, D)$.

The main difference between Algorithms 2 and 3 is the implementation of step (i). In particular, the collapsed Gibbs draws (θ_1, θ_2) from their marginal posterior distribution instead of the conditional posterior distribution as in Algorithm 2. Liu (1994) also mentions that if one uses a "mini-Gibbs" to draw (θ_1, θ_2) in step (i), that is, to sample $\theta_1 \sim \pi(\theta_1|\theta_2, D)$ and then $\theta_2 \sim \pi(\theta_2|\theta_1, D)$, the collapsed Gibbs requires that the chain from the mini-Gibbs sampler converges before step (ii). In practice, it may be difficult or expensive to directly draw (θ_1, θ_2) jointly from $\pi(\theta_1, \theta_2|D)$. In this case, we consider the following modified version of the collapsed Gibbs sampler:

Algorithm 3(a): *Modified Collapsed Gibbs Sampler*

The modified collapsed Gibbs sampler is similar to the original version by changing step (i) to:

(ia) $\theta_1 \sim \pi(\theta_1|\theta_2, D)$;

(ib) $\theta_2 \sim \pi(\theta_2|\theta_1, D)$.

We can show that the modified Gibbs sampler still leaves the target posterior distribution invariant. To see this, let $\boldsymbol{\theta}_i = (\theta_{1,i}, \theta_{2,i}, \theta_{3,i})'$ and $\boldsymbol{\theta}_{i+1} = (\theta_{1,i+1}, \theta_{2,i+1}, \theta_{3,i+1})'$ be two consecutive states. Then the construction of Algorithm 3(a) yields the following transition probability kernel:

$$T(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1}) = \pi(\theta_{1,i+1}|\theta_{2,i}, D)\pi(\theta_{2,i+1}|\theta_{1,i+1}, D) \times \pi(\theta_{3,i+1}|\theta_{1,i+1}, \theta_{2,i+1}, D). \tag{2.5.1}$$

It follows that

$$\int_{R^3} T(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1})\pi(\boldsymbol{\theta}_i|D)\, d\boldsymbol{\theta}_i = \pi(\boldsymbol{\theta}_{i+1}|D). \tag{2.5.2}$$

(The proof of (2.5.2) is left as an exercise.) Thus, $\pi(\boldsymbol{\theta}|D)$ is invariant with respect to the transition probability kernel $T(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1})$. The modified version of the collapsed Gibbs sampler is useful and practically advantageous since drawing from the conditional posterior distributions is usually easier than sampling from the joint unconditional one. This is particularly true when dealing with higher-dimensional problems.

Using norms of the forward and backward operators of the induced Markov chain, Liu (1994) shows that the collapsed Gibbs works better than the grouped Gibbs, while the latter is better than the original Gibbs. It is expected that the collapsed Gibbs may work better than the modified collapsed Gibbs, while the modified version of collapsed Gibbs may be more beneficial than the original Gibbs. However, between the modified collapsed Gibbs and the grouped Gibbs, it is not straightforward to see which one works better. The performance of these two algorithms may depend on the correlations between θ_i and θ_j. If θ_1 and θ_2 are highly correlated, the grouped Gibbs is expected to work better. Otherwise, the modified collapsed Gibbs may have better performance.

The above three-component Gibbs sampler is also studied by Liu, Wong, and Kong (1994) and further discussed by Roberts and Sahu (1997), when the target distribution $\pi(\boldsymbol{\theta}|D)$ is normal. Regarding the grouping or blocking strategy for the Gibbs sampler, Roberts and Sahu (1997) provide a comprehensive study by comparing rates of convergence of various blocking combinations, and thus we refer the reader to their paper for further discussion. In general, grouping or blocking is beneficial, but often more computationally demanding. In particular, Roberts and Sahu (1997) show that if all partial correlations of a normal (Gaussian) target distribution are nonnegative, i.e., all of the off-diagonal elements of the inverse covariance matrix are nonpositive, then the grouped (blocked) Gibbs sampler has a faster rate of convergence than the standard (original) Gibbs sampler. That is, grouping positively correlated parameters in Gibbs sampling is always beneficial. However, Roberts and Sahu (1997) also find some ex-

amples showing that blocking can also make an algorithm converge more slowly.

2.5.2 *Reparameterizations: Hierarchical Centering and Rescaling*

As pointed out by Roberts and Sahu (1997), high correlations among the coordinates of θ diminish the speed of convergence of the Gibbs sampler (see also Hills and Smith 1992). The correlations among the coordinates are determined by the particular parameterization of the problem. Gelfand, Sahu, and Carlin (1995, 1996) argue that a hierarchically centered parameterization leads to faster mixing and convergence because it generally leads to smaller intercomponent correlations among the coordinates in Bayesian linear models. Roberts and Sahu (1997) further examine the hierarchically centered parameterization and they demonstrate that hierarchical centering yields faster mixing Gibbs samplers.

Here we illustrate this idea with a one-way analysis of variance model with random effects.

Example 2.5. One-way analysis of variance with random effects. Gelfand, Sahu, and Carlin (1996) and Roberts and Sahu (1997) consider the following one-way analysis of variance model. Assume that the error variance σ_e^2 is known and suppose that we have a single observation y_i for each population, i.e.,

$$y_i = \mu + \alpha_i + \epsilon_i, \quad i = 1, 2, \ldots, m, \tag{2.5.3}$$

where $\epsilon_i \sim N(0, \sigma_e^2)$, $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\mu \sim N(\mu_0, \sigma_\mu^2)$, and σ_α^2 is also known. We denote the data by $D = \boldsymbol{y} = (y_1, y_2, \ldots, y_m)'$. Gelfand, Sahu, and Carlin (1996) rewrite (2.5.3) in a hierarchical form. Defining $\eta_i = \mu + \alpha_i$, we have

$$y_i|\eta_i \sim N(\eta_i, \sigma_e^2), \quad \eta_i|\mu \sim N(\mu, \sigma_\alpha^2), \quad \text{and} \quad \mu \sim N(\mu_0, \sigma_\mu^2).$$

This transformation from $(\alpha_1, \alpha_2, \ldots, \alpha_m)'$ to $(\eta_1, \eta_2, \ldots, \eta_m)'$ is thus referred to as hierarchical centering. Working in μ–η space, Gelfand, Sahu, and Carlin (1996) obtain

$$\text{corr}(\eta_i, \mu|D) = \left(1 + \frac{b\sigma_\alpha^2}{\sigma_e^2 \sigma_\mu^2}\right)^{-1/2} \tag{2.5.4}$$

and

$$\text{corr}(\eta_i, \eta_j|D) = \left(1 + \frac{b\sigma_\alpha^2}{\sigma_e^2 \sigma_\mu^2}\right)^{-1}, \tag{2.5.5}$$

where $b = \sigma_e^2 + \sigma_\alpha^2 + m\sigma_\mu^2$. The correlations given in (2.5.4) and (2.5.5) tend to 0 for fixed σ_e^2 and σ_μ^2 if $\sigma_\alpha^2 \to \infty$. On the other hand, if $\sigma_e^2 \to \infty$, the correlations do not approach 0, and in fact will tend to 1 if $\sigma_\mu^2 \to \infty$.

In μ–α space, we can obtain

$$\text{corr}(\alpha_i, \mu|D) = \left(1 + \frac{b\sigma_e^2}{\sigma_\alpha^2\sigma_\mu^2}\right)^{-1/2} \tag{2.5.6}$$

and

$$\text{corr}(\alpha_i, \alpha_j|D) = \left(1 + \frac{b\sigma_e^2}{\sigma_\alpha^2\sigma_\mu^2}\right)^{-1}. \tag{2.5.7}$$

The correlations given in (2.5.6) and (2.5.7) tend to 0 as $\sigma_e^2 \to \infty$, but do not approach 0 as $\sigma_\alpha^2 \to \infty$, and in fact will tend to 1 if $\sigma_\mu^2 \to \infty$ as well. In practice, when the random effects are needed, the error variance is much reduced. Thus σ_e^2 will rarely dominate the variability, so that the centered parameterization will likely be preferred. Roberts and Sahu (1997) obtain similar results by studying the rate of convergence of the Gibbs sampler.

Hierarchical centering is also useful for Bayesian nonlinear models. We will address this issue in Section 2.5.4 below. In the same spirit as hierarchical centering, hierarchical rescaling is another useful tool to reduce the correlations between the location coordinates and the scalar coordinates. We will illustrate hierarchical rescaling in the next subsection using ordinal response models.

2.5.3 Collapsing and Reparameterization for Ordinal Response Models

Consider a probit model for ordinal response data. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ denote an $n \times 1$ vector of n independent ordinal random variables. Assume that y_i takes a value of l $(1 \le l \le L,\ L > 2)$ with probability

$$p_{il} = \Phi(\gamma_l + \boldsymbol{x}_i'\boldsymbol{\beta}) - \Phi(\gamma_{l-1} + \boldsymbol{x}_i'\boldsymbol{\beta}), \tag{2.5.8}$$

for $i = 1, ..., n$ and $l = 1, \ldots, L$, where $-\infty = \gamma_0 < \gamma_1 \le \gamma_2 < \ldots < \gamma_{L-1} < \gamma_L = \infty$, $\Phi(\cdot)$ denotes the $N(0,1)$ cumulative distribution function (cdf), which defines the link, $\boldsymbol{x}_i$ is a $p \times 1$ column vector of covariates, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ column vector of regression coefficients. To ensure identifiability, we take $\gamma_1 = 0$. Let $\boldsymbol{\gamma} = (\gamma_2, \ldots, \gamma_{L-1})'$ and $D = (\boldsymbol{y}, X, n)$ denote the data, where X is the $n \times p$ design matrix with $\boldsymbol{x}_i'$ as its i^{th} row. Thus, the likelihood function is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}|D) = \prod_{i=1}^{n} \left[\Phi(\gamma_{y_i} + \boldsymbol{x}_i'\boldsymbol{\beta}) - \Phi(\gamma_{y_i-1} + \boldsymbol{x}_i'\boldsymbol{\beta})\right]. \tag{2.5.9}$$

We further assume that $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ has an improper uniform prior, i.e., $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) \propto 1$. The posterior distribution for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ takes the form

$$\begin{aligned}\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}|D) &\propto L(\boldsymbol{\beta}, \boldsymbol{\gamma}|D)\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) \\ &= \prod_{i=1}^{n} [\Phi(\gamma_{y_i} + \boldsymbol{x}_i'\boldsymbol{\beta}) - \Phi(\gamma_{y_i-1} + \boldsymbol{x}_i'\boldsymbol{\beta})]\,. \end{aligned} \tag{2.5.10}$$

Chen and Shao (1999a) obtain necessary and sufficient conditions for the propriety of the posterior defined by (2.5.10). To facilitate the Gibbs sampler, Albert and Chib (1993) introduce latent variables z_i such that

$$y_i = l \text{ iff } \gamma_{l-1} \le z_i < \gamma_l,$$

for $l = 1, 2, \ldots, L$. Let $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)'$. The complete-data likelihood is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{z}|D) \propto \prod_{i=1}^{n} [\exp\{-\tfrac{1}{2}(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\} 1\{\gamma_{y_i-1} \le z_i < \gamma_{y_i}\}], \tag{2.5.11}$$

where $1\{\gamma_{y_i-1} \le z_i < \gamma_{y_i}\}$ is the indicator function, and the joint posterior density for $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{z})$ is given by

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{z}|D) \propto \left\{ \prod_{i=1}^{n} [\exp\{-\tfrac{1}{2}(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\} 1\{\gamma_{y_i-1} \le z_i < \gamma_{y_i}\}] \right\}. \tag{2.5.12}$$

Then, Albert and Chib (1993) incorporate the unknown latent variables $\boldsymbol{z}$ as additional parameters to run the Gibbs sampler. The original Gibbs sampler for the ordinal probit model proposed by Albert and Chib (1993), which is referred to as the Albert–Chib algorithm thereafter, may be implemented as follows:

Albert–Chib Algorithm

Step 1. Draw $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{\gamma} \;\sim\; N((X'X)^{-1}X'\boldsymbol{z}, (X'X)^{-1}).$$

Step 2. Draw z_i from

$$z_i \;\sim\; N(\boldsymbol{x}_i'\boldsymbol{\beta}, 1), \;\; \gamma_{y_i-1} \le z_i \le \gamma_{y_i}.$$

Step 3. Draw $\boldsymbol{\gamma}$ from

$$\gamma_l|\boldsymbol{\gamma}^{(-l)}, \boldsymbol{\beta}, \boldsymbol{z} \;\sim\; U[a_l, b_l], \tag{2.5.13}$$

where $a_l = \max\left\{\gamma_{l-1}, \max_{y_i=l} z_i\right\}$, $b_l = \min\left\{\gamma_{l+1}, \min_{y_i=l+1} z_i\right\}$, and $\boldsymbol{\gamma}^{(-l)}$ is $\boldsymbol{\gamma}$ with γ_l deleted.

Since, in Step 1 all p components of the regression coefficient vector are drawn simultaneously, the Albert–Chib algorithm is indeed a grouped Gibbs sampler. The implementation of the Albert–Chib algorithm is straightforward since the conditional posterior distributions are normal, truncated normal, or uniform. When the sample size n is not too big, the Albert–Chib algorithm works reasonably well. However, when n is large, say $n \geq 50$, slow convergence of the Albert–Chib algorithm may occur. Cowles (1996) points out this slow convergence problem. Because the interval (a_l, b_l) within which each γ_l must be generated from its full conditional can be very narrow, the cutpoint values may change very little between successive iterations, making the iterates highly correlated. Of course, slower convergence of the γ_l is also associated with the fact that the variance of the latent variables is fixed at one. The empirical study of Cowles (1996) further shows that the slow convergence of the cutpoints may also seriously affect the convergence of $\boldsymbol{\beta}$. To improve convergence of the original Gibbs sampler, she proposes a Metropolis–Hastings algorithm to generate the cutpoints from their conditional distributions; henceforth, this algorithm is called the Cowles algorithm. Instead of directly generating γ_l from (2.5.13), the Cowles algorithm generates $(\boldsymbol{\gamma}, \boldsymbol{z})$ jointly, which is essentially the same idea as the (modified) collapsed Gibbs sampler described in Section 2.5.1. The joint conditional distribution $\pi(\boldsymbol{\gamma}, \boldsymbol{z}|\boldsymbol{\beta}, D)$ can be expressed as the product of the marginal conditional distributions $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta}, D)$ and $\pi(\boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\beta}, D)$. The Cowles algorithm can be described as follows:

Cowles Algorithm

Step 1. Draw $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{\gamma} \ \sim \ N((X'X)^{-1}X'\boldsymbol{z}, (X'X)^{-1}).$$

Step 2. Draw z_i from

$$z_i \ \sim \ N(\boldsymbol{x}_i'\boldsymbol{\beta}, 1), \ \ \gamma_{y_i-1} \leq z_i \leq \gamma_{y_i}.$$

Step 3. Draw $\boldsymbol{\gamma}$ from the conditional distribution

$$\pi(\boldsymbol{\gamma}|\boldsymbol{\beta}, D) \propto \prod_{i=1}^{n} \left[\Phi(\gamma_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}) - \Phi(\gamma_{y_i-1} - \boldsymbol{x}_i'\boldsymbol{\beta})\right]. \qquad (2.5.14)$$

In the Cowles algorithm, a Metropolis–Hastings scheme is used to draw $\boldsymbol{\gamma}$. That is, given the value $\boldsymbol{\gamma}_{j-1}$ from the previous iteration, a vector of proposal cutpoint values, γ_l^*, $l = 2, 3, \ldots, L-1$, is generated from a truncated normal distribution

$$\gamma_l^*|\gamma_{l-1}^*, \gamma_{l+1,j-1} \ \sim \ N(\gamma_{l,j-1}, \sigma_\gamma^2), \qquad (2.5.15)$$

where $\gamma_{l-1}^* \leq \gamma_l^* \leq \gamma_{l+1,j-1}$. The acceptance probability for the vector $\boldsymbol{\gamma}^*$ of new cutpoints is $a = \min\{1, R\}$, where

$$R = \prod_{l=2}^{L-1} \frac{\{\Phi\{(\gamma_{l+1,j-1} - \gamma_{l,j-1})/\sigma_\gamma\} - \Phi\{(\gamma_{l-1}^* - \gamma_{l,j-1})/\sigma_\gamma\}}{\Phi\{(\gamma_{l+1}^* - \gamma_l^*)/\sigma_\gamma\} - \Phi\{(\gamma_{l-1,j-1} - \gamma_l^*)/\sigma_\gamma\}}$$
$$\times \prod_{i=1}^{n} \frac{\Phi(\gamma_{y_i}^* - \boldsymbol{x}_i'\boldsymbol{\beta}) - \Phi(\gamma_{y_i-1}^* - \boldsymbol{x}_i'\boldsymbol{\beta})}{\Phi(\gamma_{y_i,j-1} - \boldsymbol{x}_i'\boldsymbol{\beta}) - \Phi(\gamma_{y_i-1,j-1} - \boldsymbol{x}_i'\boldsymbol{\beta})}. \tag{2.5.16}$$

However, our experience suggests that in the Cowles algorithm, the truncated normal-distribution in (2.5.15) might not serve as a good proposal density for the conditional posterior density in (2.5.14), since it is not spread out enough (see Tierney (1994), and Section 2.2). Further, Cowles (1996) points out that a good σ_γ^2 in (2.5.15) is difficult to obtain even when using a conventional updating scheme.

To overcome the difficulties arising in the Cowles algorithm, Nandram and Chen (1996) develop an improved algorithm using a Dirichlet proposal distribution based on a rescaling transformation. Similar to hierarchical centering, the hierarchically rescaled transformations proposed by Nandram and Chen (1996) are

$$\delta = 1/\gamma_{L-1}, \;\; \boldsymbol{\gamma}^* = \delta\boldsymbol{\gamma}, \;\; \boldsymbol{\beta}^* = \delta\boldsymbol{\beta}, \;\; \text{and} \;\; \boldsymbol{z}^* = \delta\boldsymbol{z}. \tag{2.5.17}$$

Note that in (2.5.17), $\gamma_0^* = -\infty < \gamma_1^* = 0 \leq \gamma_2^* \leq \cdots \leq \gamma_{L-2}^* \leq \gamma_{L-1}^* = 1 < \gamma_L^* = \infty$, and that effectively there are only $L-3$ unknown cutpoints in the reparameterized model. Let $\boldsymbol{\gamma}^* = (\gamma_2^*, \gamma_3^*, \ldots, \gamma_{L-2}^*)'$. Thus, when $L = 3$, there are no unknown cutpoints in $\boldsymbol{\gamma}^*$, which is advantageous when one deals with a 3-level ordinal response model. For $L = 3$, the Nandram–Chen algorithm can be implemented as follows:

Nandram–Chen Algorithm

Step 1. Draw $\boldsymbol{\beta}^*$ from

$$\boldsymbol{\beta}^* | \boldsymbol{z}^*, \boldsymbol{\gamma}^* \;\sim\; N((X'X)^{-1}X'\boldsymbol{z}, \delta^2(X'X)^{-1}).$$

Step 2. Draw z_i^* from

$$z_i^* | \boldsymbol{\beta}^*, \delta \;\sim\; N(\boldsymbol{x}_i'\boldsymbol{\beta}^*, \delta^2), \;\; \gamma_{y_i-1}^* \leq z_i^* < \gamma_{y_i}^*.$$

Step 3. Draw δ^2 from

$$\delta^2 | \boldsymbol{\beta}^*, \boldsymbol{z}^* \;\sim\; \mathcal{IG}\left\{\frac{n+p+L-2}{2}, \tfrac{1}{2}[(\boldsymbol{z}^* - X\boldsymbol{\beta}^*)'(\boldsymbol{z}^* - X\boldsymbol{\beta}^*)]\right\}.$$

For $L > 3$, the Nandram–Chen algorithm requires an additional step to draw $\boldsymbol{\gamma}^*$. That is,

Step 4. Draw $\boldsymbol{\gamma}^*$ from the conditional posterior distribution

$$\pi(\boldsymbol{\gamma}^*|\boldsymbol{\beta}^*, \delta^2, D)$$
$$\propto \prod_{i=1}^{n} \left\{ \Phi\left(\frac{\gamma^*_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}^*}{\delta}\right) - \Phi\left(\frac{\gamma^*_{y_i-1} - \boldsymbol{x}_i'\boldsymbol{\beta}^*}{\delta}\right) \right\}. \tag{2.5.18}$$

Instead of using a truncated normal proposal density as in the Cowles algorithm, Nandram and Chen (1996) construct a Dirichlet proposal density for $\pi(\boldsymbol{\gamma}^*|\boldsymbol{\beta}^*, \delta^2, D)$. The motivation for the Dirichlet proposal density is given as follows. Let $q_{l-1} = \gamma^*_l - \gamma^*_{l-1}$, $l = 2, \ldots, L-1$, and let $\boldsymbol{q} = (q_1, q_2, \ldots, q_{L-2})'$, $q_l \geq 0$, $l = 1, 2, \ldots, L-2$ and $\sum_{l=1}^{L-2} q_l = 1$. By the fundamental mean value theorem,

$$\Phi\left(\frac{\gamma^*_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}^*}{\delta}\right) - \Phi\left(\frac{\gamma^*_{y_i-1} - \boldsymbol{x}_i'\boldsymbol{\beta}^*}{\delta}\right) = \frac{1}{\delta}\phi\left(\frac{\xi_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}^*}{\delta}\right) q_{y_i-1}, \tag{2.5.19}$$

where ξ_{y_i} is a real number between $\gamma^*_{y_i}$ and $\gamma^*_{y_i-1}$, $i = 1, 2, 3, \ldots, n$, and $\phi(\cdot)$ is the standard normal density function. Then by (2.5.19),

$$\pi(\boldsymbol{\gamma}^*|\boldsymbol{\beta}^*, \delta^2, D) \propto g_1(\xi) g_2(\boldsymbol{q}), \tag{2.5.20}$$

where

$$g_1(\xi) = \prod_{i=1}^{n} \phi\left(\frac{\xi_{y_i} - \boldsymbol{x}_i'\boldsymbol{\beta}^*}{\delta}\right), \quad g_2(\boldsymbol{q}) = \prod_{l=1}^{L-2} q_l^{n_{l+1}}, \quad \text{and} \quad n_l = \sum_{i=1}^{n} 1\{y_i = l\}.$$

for $l = 1, 2, \ldots, L$. While in the Cowles algorithm the proposal density is based on $g_1(\xi)$, Nandram and Chen (1996) use $g_2(\boldsymbol{q})$ to construct a proposal density. This is quite natural because if there are no covariates, we can associate $\boldsymbol{q}$ with the bin "probabilities." Assigning an improper prior to the bins and treating the bin counts as data, the joint posterior distribution of these bin "probabilities" is a Dirichlet distribution with the bin counts as the posterior parameters.

An approximation of $\pi(\boldsymbol{\gamma}^*|\boldsymbol{\beta}^*, \delta^2, D)$ motivated by (2.5.20) is

$$\pi(\boldsymbol{q}|\boldsymbol{\beta}^*, \delta^2, D) \propto \prod_{l=1}^{L-2} q_l^{\alpha_l n_{l+1} - 1}, \tag{2.5.21}$$

where $0 \leq \alpha_l \leq 1$, $l = 1, \ldots, L-2$, are the tuning parameters. (That is, $\boldsymbol{q}$ has a Dirichlet distribution.) The proposal density (2.5.21) is attractive because we can draw the entire vector $\boldsymbol{q}$ at once, and it does not depend on $\boldsymbol{\beta}^*$ and δ^2. In addition, the Dirichlet proposal density will be more useful when more complex link functions (e.g., logistic and complementary log-log) are used. Moreover, one can choose the α_l in (2.5.21) by taking the α_l so as to make the dispersion of the posterior distribution of $\boldsymbol{q}$ comparable to or at least as large as that of the distribution of $\boldsymbol{\gamma}^*$. The rest of the implementation for drawing $\boldsymbol{\gamma}^*$ simultaneously from its conditional posterior

distribution is the same as the one given in the Cowles algorithm, and thus the details are omitted.

Nandram and Chen (1996) conduct several simulation studies, and their empirical results show that the Nandram–Chen algorithm substantially improves convergence of the Gibbs sampler compared to the Albert–Chib and Cowles algorithms. A partial explanation for this is that:

(a) hierarchical rescaling reduces the correlations between the cutpoints and the latent variables; and

(b) the meaningful choice (from a theoretical or statistical viewpoint) of the proposal density has better properties than the truncated normal proposal density used in the Cowles algorithm.

The Nandram–Chen algorithm works well if the cell counts n_l are relatively balanced. When the cell counts are unbalanced, in particular, if some of those counts are close to zero, $\pi(\boldsymbol{q}|\boldsymbol{\beta}^*, \delta^2, D)$ in (2.5.21) may not serve as a good proposal density. For these cases, Chen and Dey (1996) propose a Metropolis–Hastings algorithm using a "de-constraint" transformation to draw $\boldsymbol{\gamma}^*$. A similar transformation of the cutpoints is also considered in Albert and Chib (1998). Let

$$\gamma_l^* = \frac{\gamma_{l-1}^* + e^{\zeta_l}}{1 + e^{\zeta_l}}, \quad l = 2, \ldots, L-2, \tag{2.5.22}$$

and $\boldsymbol{\zeta} = (\zeta_2, \ldots, \zeta_{L-2})'$. Then the conditional posterior distribution for ζ is

$$\pi(\boldsymbol{\zeta}|\boldsymbol{\beta}^*, \delta^2, D) \propto \pi(\boldsymbol{\gamma}^*|\boldsymbol{\beta}^*, \delta^2, D) \prod_{l=2}^{L-2} \frac{(1-\gamma_{l-1})e^{\zeta_l}}{(1+e^{\zeta_l})^2}, \tag{2.5.23}$$

where $\boldsymbol{\gamma}^*$ is evaluated at $\gamma_l^* = (\gamma_{l-1}^* + e^{\zeta_l})/(1+e^{\zeta_l})$ for $l = 2, 3, \ldots, L-2$. The remaining steps of the Metropolis–Hastings algorithm are the same as the algorithm for sampling ρ as described in Example 2.3. This modified Nandram–Chen algorithm is thus called the *Chen–Dey algorithm.*

2.5.4 Hierarchical Centering for Poisson Random Effects Models

A Poisson regression model with AR(1) random effects is used for modeling the pollen count data in Example 1.3. Using (1.3.3), the complete-data likelihood is given by

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\epsilon}|D) = &\exp\{\boldsymbol{y}'(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) - J_n' Q(\boldsymbol{\beta}, \boldsymbol{\epsilon}) - J_n' C(\boldsymbol{y})\} \\ &\times (2\pi\sigma^2)^{-n/2}(1-\rho^2)^{-(n-1)/2} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\epsilon}'\Sigma^{-1}\boldsymbol{\epsilon}\right\}, \end{aligned} \tag{2.5.24}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$, X is the $n \times 8$ matrix of covariates with the t^{th} row equal to $\boldsymbol{x}_t'$, J_n is an $n \times 1$ vector of ones, and $Q(\boldsymbol{\beta}, \boldsymbol{\epsilon})$ is an $n \times 1$ vector with the t^{th} element equal to $q_t = \exp\{\epsilon_t + \boldsymbol{x}_t'\boldsymbol{\beta}\}$, $C(\boldsymbol{y})$ is an $n \times 1$ vector with the j^{th} element $\log(y_j!)$, and $D = (n, \boldsymbol{y}, X)$. In (2.5.24), $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$, where $\rho^{|i-j|}$ is the correlation between (ϵ_i, ϵ_j), and $-1 \leq \rho \leq 1$. Assume that a noninformative prior for $(\boldsymbol{\beta}, \sigma^2, \rho)$ has the form

$$\pi(\boldsymbol{\beta}, \sigma^2, \rho) \propto (\sigma^2)^{-(\delta_0+1)} \exp(-\sigma^{-2}\gamma_0), \tag{2.5.25}$$

where the hyperparameters $\delta_0 > 0$ and $\gamma_0 > 0$ are prespecified. Then, the joint posterior distribution for $(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\epsilon})$ is given by

$$\pi(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\epsilon}|D) \propto L(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\epsilon}|D)(\sigma^2)^{-(\delta_0+1)} \exp(-\sigma^{-2}\gamma_0), \tag{2.5.26}$$

where the likelihood $L(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\epsilon}|D)$ is defined by (2.5.24). It can be shown that if X^* is of full rank, where X^* is a matrix induced by X and $\boldsymbol{y}$ with its t^{th} row equal to $1\{y_t > 0\}\boldsymbol{x}_t'$, then the posterior distribution $\pi(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\epsilon}|D)$ is proper.

Unlike the one-way analysis of variance model with random effects in Example 2.5, the Poisson regression model with AR(1) random effects is not a linear model. The exact correlations among the parameters $\boldsymbol{\epsilon}$, $\boldsymbol{\beta}$, σ^2, and ρ are not clear. However, it is expected that the correlation patterns in the Poisson regression model are similar to that of the one-way analysis of variance model. Ibrahim, Chen, and Ryan (1999) observe that the original Gibbs sampler without hierarchical centering results in very slow convergence and poor mixing. In particular, the correlation ρ appears to converge the slowest. They further find that the hierarchical centering technique is perfectly suited for this problem, and appears quite crucial for convergence of the Gibbs sampler.

Similar to the one-way analysis of variance model, a hierarchically centered reparameterization is given by

$$\boldsymbol{\eta} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.5.27}$$

Using (2.5.26), the reparameterized posterior for $(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\eta})$ is written as

$$\begin{aligned}\pi(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\eta}|D) \propto & \exp\{\boldsymbol{y}'\boldsymbol{\eta} - J_n'Q(\boldsymbol{\eta}) - J_n'C(\boldsymbol{y})\} \\ & \times (2\pi\sigma^2)^{-n/2}(1-\rho^2)^{-(n-1)/2} \\ & \times \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\eta} - X\boldsymbol{\beta})'\Sigma^{-1}(\boldsymbol{\eta} - X\boldsymbol{\beta})\right\},\end{aligned} \tag{2.5.28}$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)'$, and $Q(\boldsymbol{\eta})$ is an $n \times 1$ vector with the t^{th} element equal to $q_t = \exp(\eta_t)$.

The Gibbs sampler for sampling from the reparameterized posterior $\pi(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\eta}|D)$ requires the following steps:

Step 1. Draw $\boldsymbol{\eta}$ from its conditional posterior distribution

$$\pi(\boldsymbol{\eta}|\boldsymbol{\beta},\sigma^2,\rho,D) \propto \exp\left\{\boldsymbol{y}'\boldsymbol{\eta} - J_n' Q(\boldsymbol{\eta}) - \frac{(\boldsymbol{\eta}-X\boldsymbol{\beta})'\Sigma^{-1}(\boldsymbol{\eta}-X\boldsymbol{\beta})}{2\sigma^2}\right\}. \tag{2.5.29}$$

Step 2. Draw $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta}|\boldsymbol{\eta},\sigma^2,\rho,D \;\sim\; N_8((X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\boldsymbol{\eta},\sigma^2(X'\Sigma^{-1}X)^{-1}).$$

Step 3. Draw σ^2 from its conditional posterior

$$\sigma^2|\boldsymbol{\beta},\rho,\boldsymbol{\eta},D \;\sim\; \mathcal{IG}(\delta^*,\gamma^*),$$

where $\delta^* = \delta_0 + n/2$, $\gamma^* = \gamma_0 + \frac{1}{2}(\boldsymbol{\eta}-X\boldsymbol{\beta})'\Sigma^{-1}(\boldsymbol{\eta}-X\boldsymbol{\beta})$, and $\mathcal{IG}(\delta^*,\gamma^*)$ is an inverse gamma distribution.

Step 4. Draw ρ from its conditional posterior

$$\pi(\rho|\sigma^2,\boldsymbol{\beta},\boldsymbol{\eta},D)$$
$$\propto (1-\rho^2)^{-(n-1)/2}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\eta}-X\boldsymbol{\beta})'\Sigma^{-1}(\boldsymbol{\eta}-X\boldsymbol{\beta})\right\}.$$

In Step 1, it can be shown that $\pi(\boldsymbol{\eta}|\boldsymbol{\beta},\sigma^2,\rho,D)$ is log-concave in each component of $\boldsymbol{\eta}$ (see Exercise 2.7). Thus $\boldsymbol{\eta}$ can be drawn using the adaptive rejection sampling algorithm of Gilks and Wild (1992). The implementation of Steps 2 and 3 is straightforward, which may be a bonus of hierarchical centering, since sampling $\boldsymbol{\beta}$ is much more expensive before the reparameterization. In Chapter 9, we will also show that the hierarchical centering reparameterization can greatly ease the computational burden in Bayesian variable selection. In Step 4, we can use the algorithm in Example 2.3 for sampling ρ.

2.6 Acceleration Algorithms for MCMC Sampling

The major problems for many MCMC algorithms are slow convergence and poor mixing. For example, the Gibbs sampler converges slowly even for a simple ordinal response model as discussed in Section 2.5.3. In the earlier sections of this chapter, we discuss several tools for speeding up an MCMC algorithm, which include grouping (blocking) and collapsing (Liu 1994; Liu, Wong, and Kong 1994; Roberts and Sahu 1997), reparameterizations (Gelfand, Sahu, and Carlin 1995 and 1996; Roberts and Sahu 1997). The other useful techniques are adaptive direction sampling (Gilks, Roberts, and George 1994; Roberts and Gilks 1994), Multiple-Try Metropolis (Liu, Liang, and Wong 1998a), auxiliary variable methods (Besag and Green 1993; Damien, Wakefield, and Walker 1999), simulated tempering (Marinari and Parisi 1992; Geyer and Thompson 1995), and working parameter methods (Meng and van Dyk 1999). In this section, we present

two special acceleration MCMC algorithms, i.e., grouped move and Multigrid MC sampling (Liu and Wu 1997; Liu and Sabatti 1998 and 1999) and covariance-adjusted MCMC sampling (Liu 1998), which provide us with a general framework of how to further improve mixing and convergence of an MCMC algorithm.

2.6.1 Grouped Move and Multigrid Monte Carlo Sampling

Goodman and Sokal (1989) present a comparative review of the multigrid Monte Carlo (MGMC) method, which is a stochastic generalization of the multigrid (MG) method for solving finite-difference equations. Liu and Wu (1997) and Liu and Sabatti (1998 and 1999) generalize Goodman and Sokal's MGMC via groups of transformations with applications to MCMC sampling. They propose a Grouped Move Multigrid Monte Carlo (GM-MGMC) algorithm and a generalized version of the MGMC algorithm for sampling from a target posterior distribution.

Assume that the target posterior distribution $\pi(\boldsymbol{\theta}|D)$ is defined on Ω and that an MCMC algorithm such as the Gibbs sampler or Metropolis–Hastings algorithm is used to generate a Markov chain $\{\boldsymbol{\theta}_i,\ i = 0, 1, 2, \dots\}$ from the target distribution $\pi(\boldsymbol{\theta}|D)$. We call the MCMC algorithm used to generate the $\boldsymbol{\theta}_i$ the *parent* MCMC algorithm. Let Γ be a locally compact transformation group (Rao 1987) on Ω. Then the GM-MGMC algorithm of Liu and Wu (1997) and Liu and Sabatti (1998) proceeds as follows:

GM-MGMC Algorithm

MCMC Step. Generate an iteration $\boldsymbol{\theta}_i$ from the parent MCMC.

GM Step. Draw the group element g from

$$g \ \sim \ \pi(g|\boldsymbol{\theta}_i)H(g) \propto \pi(g(\boldsymbol{\theta}_i)|D)J_g(\boldsymbol{\theta}_i)H(dg), \tag{2.6.1}$$

and *adjust*

$$\boldsymbol{\theta}_i \leftarrow g(\boldsymbol{\theta}_i).$$

In (2.6.1) $H(dg)$ is the right-invariant Haar measure on Ω and $J_g(\boldsymbol{\theta}_i)$ is the Jacobian of g evaluated at $\boldsymbol{\theta}_i$. Liu and Wu (1997) show that if Γ is a locally compact group of transformations for $\boldsymbol{\theta} \in \Omega$ with a unimodular right-invariant Haar measure $H(dg)$, then $g(\boldsymbol{\theta}_i) \sim \pi(\boldsymbol{\theta}|D)$, provided $\boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta}|D)$ and $g \sim \pi(g|\boldsymbol{\theta}_i)$. This result ensures that the target posterior distribution $\pi(\boldsymbol{\theta}|D)$ is the stationary distribution of the Markov chain induced by the GM-MGMC algorithm. As discussed in Liu and Sabatti (1998), the GM step is flexible: by selecting appropriate groups of transformations one can achieve either the effect of reparameterization or that of "blocking" or "grouping." Liu and Sabatti use several examples to illustrate these points.

In many cases, directly sampling g in the GM step may be difficult or expensive to achieve. For these cases, Liu and Sabatti (1998, 1999) propose a Markov transition. Assume that $T_{\boldsymbol{\theta}}(g', g)H(dg)$ is a Markov transition function, which leaves (2.6.1) invariant and satisfies the *transformation-invariance*, i.e.,

$$T_{\boldsymbol{\theta}}(g', g) = T_{g_0^{-1}\boldsymbol{\theta}}(g'g_0, gg_0) \tag{2.6.2}$$

for all g, g', and g_0 in Γ. Then, the GM-MGMC algorithm can be extended to the following Generalized MGMC algorithm:

Generalized MGMC Algorithm

MCMC Step. Generate an iteration $\boldsymbol{\theta}_i$ from the parent MCMC.

GM Step. Draw the group element g from

$$g \sim T_{\boldsymbol{\theta}_i}(I, g), \tag{2.6.3}$$

and *adjust*

$$\boldsymbol{\theta}_i \leftarrow g(\boldsymbol{\theta}_i).$$

In (2.6.3), I denotes the identity of the group. Similar to the GM-MGMC algorithm, it can be shown that $g(\boldsymbol{\theta}) \sim \pi(\boldsymbol{\theta}|D)$ provided $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|D)$.

In the GM-MGMC algorithm or the generalized MGMC algorithm, one GM step is used. In some situations, multiple GM steps can also be applied. For example, a three-step GM algorithm can be described by the following cycle. Starting with $\boldsymbol{\theta}_i \in \Omega$, which is drawn from a parent MCMC algorithm:

(i) draw g from a proper $T_{\boldsymbol{\theta}_i}(I, g)$, which leaves (2.6.1) invariant and satisfies (2.6.2), and update $\boldsymbol{\theta}^* = g(\boldsymbol{\theta})$;

(ii) draw g^* from a proper $T_{\boldsymbol{\theta}^*}(I, g^*)$ and update $\boldsymbol{\theta}^{**} = g^*(\boldsymbol{\theta}^*)$; and

(iii) draw g^{**} from $T_{\boldsymbol{\theta}^{**}}(I, g^{**})$ and update $\boldsymbol{\theta}_{i+1} = g^{**}(\boldsymbol{\theta}^{**})$.

Then, if $\boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta}|D)$, then $\boldsymbol{\theta}_{i+1} \sim \pi(\boldsymbol{\theta}|D)$.

The GM-MCMC algorithm is a flexible generalization of the Gibbs sampler or the Metropolis–Hastings algorithm, which enables us to design more efficient MCMC algorithms. The purpose of the GM step is to improve the convergence or mixing rates of the parent MCMC algorithm. The nature of the multiple GM steps allows us to achieve such an improvement in an iterative fashion. That is, if the performance of a parent MCMC algorithm is unsatisfactory, one can design a GM step, and then make an additional draw in each iteration of the parent algorithm. From an implementational standpoint, the GM step requires only adding a subroutine to the existing code and does not require a change in the basic structure of the code. After a one-step adjustment, the new MCMC algorithm induced by the GM step

can be viewed as a new parent algorithm. Then, a similar adjustment can be applied to this new parent algorithm. One can repeat this procedure many times until a satisfactory convergence rate is achieved. Therefore, accelerating an MCMC algorithm can be viewed as a continuous improvement process.

Although the GM-MCMC algorithm provides a general framework for speeding up an MCMC algorithm, finding a computationally feasible group of transformations along with a unimodular right-invariant Haar measure $H(dg)$ is a difficult task. Two simple groups are the multiplicative group, i.e., $g(\boldsymbol{\theta}) = g\boldsymbol{\theta}$, and the additive group, i.e., $g(\boldsymbol{\theta}) = \boldsymbol{\theta} + \boldsymbol{g}$. For these two special cases, the associated unimodular right-invariant Haar measures are $H(dg) = 1/g$ for the multiplicative group and $H(d\boldsymbol{g}) = 1$ for the additive group. Although both the multiplicative group and additive group result in unimodular Haar measures, the linear combination of these two group transformations, i.e., $g(\boldsymbol{\theta}) = g_1\boldsymbol{\theta} + \boldsymbol{g}_2$, does not yield a unimodular Haar measure (see Nachbin 1965). In addition, GM-MGMC may not always improve convergence over the parent MCMC algorithm. In fact, Liu and Sabatti (1998) provide an example showing that GM-MGMC can result in a slower convergence rate than the parent MCMC algorithm. However, in many cases, GM-MGMC can achieve a substantial improvement in the convergence and mixing rate over a parent MCMC algorithm; see Liu and Sabatti (1998) and Chen and Liu (1999) for several illustrative examples. In practice, GM-MGMC can be viewed as an advanced experimental technique for improving convergence of an MCMC algorithm, and it can be helpful in getting a better understanding of the problem. As a practical guideline, we suggest using a GM step as long as it is simple and easy to implement.

2.6.2 Covariance-Adjusted MCMC Algorithm

Liu (1998) provides an alternative method for speeding up an MCMC algorithm using the idea of covariance adjustment. Let $\{\boldsymbol{\theta}_i,\ i = 0, 1, 2, \ldots\}$ be generated by the parent MCMC algorithm, having the stationary distribution $\pi(\boldsymbol{\theta}|D)$. Also let $(\boldsymbol{\xi}, \boldsymbol{\delta}) = \mathcal{M}(\boldsymbol{\theta})$ be a one-to-one mapping from Ω on which the target distribution is defined onto the space $\Xi \times \Delta$. Then, the covariance-adjusted MCMC (CA-MCMC) algorithm at the i^{th} iteration consists of the following two steps:

CA-MCMC Algorithm

MCMC Step. Generate an iteration $\boldsymbol{\theta}_i$ from the parent MCMC and compute $(\boldsymbol{\xi}_i, \boldsymbol{\delta}_i) = \mathcal{M}(\boldsymbol{\theta}_i)$.

CA Step. Draw $\boldsymbol{\delta}_i^*$ from the conditional posterior distribution $\pi(\boldsymbol{\delta}|\boldsymbol{\xi}_i, D)$ and *adjust* $\boldsymbol{\theta}_i$ by

$$\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i^* = \mathcal{M}^{-1}(\boldsymbol{\xi}_i, \boldsymbol{\delta}_i^*), \tag{2.6.4}$$

where $\mathcal{M}^{-1}(\boldsymbol{\xi}, \boldsymbol{\delta})$ is the inverse mapping of $(\boldsymbol{\xi}, \boldsymbol{\delta}) = \mathcal{M}(\boldsymbol{\theta})$.

Liu (1998) shows that the CA-MCMC algorithm converges to the target distribution $\pi(\boldsymbol{\theta}|D)$. That is, if the Markov chain induced by an MCMC algorithm is irreducible, aperiodic, and stationary with the equilibrium distribution $\pi(\boldsymbol{\theta}|D)$, so is the covariance-adjusted Markov chain. We refer the reader to Liu's paper for a formal proof. This result ensures that the CA step guarantees the correctness of the CA-MCMC algorithm. In addition, Liu (1998) also proves that the CA-MCMC algorithm converges at least as fast as its parent MCMC algorithm in the sense that the CA-MCMC algorithm results in a smaller reversed Kullback–Leibler information distance (e.g., Liu, Wong, and Kong 1995). This implies that the Markov sequence induced by the CA-MCMC algorithm has less dependence than that induced by the parent MCMC algorithm. This result essentially distinguishes the CA-MCMC algorithm from the GM-MGMC algorithm since the latter does not always guarantee faster convergence than its parent MCMC algorithm.

The key issue in using the CA-MCMC algorithm is how to construct the $\boldsymbol{\delta}$-variable so that the resulting algorithm is efficient and simple to implement. A general strategy proposed by Liu (1998) is to construct the $\boldsymbol{\delta}$-variable based on parameters and their sufficient statistics. We use an example given in Liu (1998) to illustrate this idea.

Example 2.6. One-way analysis of variance with random effects (Example 2.5 continued). Consider the one-way analysis of the variance model given in Example 2.5. Assume that the error variance σ_e^2 is known and that a single observation y_i for each population, i.e.,

$$y_i = \mu + \alpha_i + \epsilon_i, \quad i = 1, 2, \ldots, m, \tag{2.6.5}$$

where $\epsilon_i \sim N(0, \sigma_e^2)$, $\alpha_i \sim N(0, \sigma_\alpha^2)$, and σ_α^2 is also known. We assume that $\pi(\mu) \propto 1$ and let $\bar{y} = (1/m) \sum_{i=1}^m y_i$ and $D = (y_1, y_2, \ldots, y_m)$.

For this one-way analysis of the variance model, the vector of model parameters is $\boldsymbol{\theta} = (\mu, \alpha_1, \alpha_2, \ldots, \alpha_m)'$. We use the Gibbs sampler as the parent MCMC algorithm. To apply the CA-MCMC algorithm, we need to construct $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$. In this example, μ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)'$ may be highly correlated (see (2.5.6) and (2.5.7)), which may cause slow convergence of the original Gibbs sampler. To break down this correlation pattern, we consider the parameter μ. From (2.6.5), it is easy to see that,

a posteriori, the sufficient statistic for μ is

$$\bar{\alpha} = \frac{1}{m}\sum_{i=1}^{m} \alpha_i.$$

Let $\xi_i = \alpha_i - \bar{\alpha}$. Define

$$\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_m)' \text{ and } \boldsymbol{\delta} = (\mu, \bar{\alpha})', \tag{2.6.6}$$

and let $\Xi = \{\xi : \sum_{i=1}^{m} \xi_i = 0,\ \xi_i \in R \text{ for } i = 1, 2, \ldots, m\}$. Then, (2.6.6) clearly defines a one-to-one mapping from R^{m+1} to $\Xi \times R^2$. The Jacobian of this transformation is $J_{(\mu,\alpha)\to(\mu,\bar{\alpha},\xi_1,\ldots,\xi_{m-1})} = 1$. The CA step requires drawing a $(\mu, \bar{\alpha})$ conditional on ξ_i for $i = 1, 2, \ldots, m$. The complete CA-MCMC algorithm can be stated as follows:

CA-MCMC for One-Way Analysis of Variance with Random Effects

Gibbs Step. Draw $(\mu|\boldsymbol{\alpha}, D) \sim N(\bar{y} - \bar{\alpha}, \sigma_e^2/m)$ and

$$(\alpha_i|\mu, D) \sim N\left(\frac{\sigma_\alpha^2}{\sigma_e^2 + \sigma_\alpha^2}(y_i - \mu), \frac{\sigma_e^2\sigma_\alpha^2}{\sigma_e^2 + \sigma_\alpha^2}\right).$$

CA Step. Draw $(\bar{\alpha}^*|\boldsymbol{\xi}, D) \sim N(0, \sigma_\alpha^2/m)$ and

$$(\mu^*|\bar{\alpha}^*, \boldsymbol{\xi}, D) \sim N\left(\bar{y} - \bar{\alpha}^*, \frac{\sigma_e^2}{m}\right),$$

then *adjust*

$$\mu \leftarrow \mu^* \text{ and } \alpha_i \leftarrow \xi_i + \bar{\alpha}^* \text{ for } i = 1, 2, \ldots, m.$$

From the structure of the above CA-MCMC algorithm, it can be seen that the draws of $(\mu^*, \bar{\alpha}^*, \xi_1 + \bar{\alpha}^*, \xi_2 + \bar{\alpha}^*, \ldots, \xi_m + \bar{\alpha}^*)$ are independent. Thus, the rate of convergence of this CA-MCMC algorithm is 0. Roberts and Sahu (1997) show that the rate of convergence of the Markov chain using the Gibbs step only is $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_e^2)$ and that the rate of convergence of the Gibbs sampler based on the hierarchically centered transformation given in Example 2.5 (namely, $\mu = \mu$ and $\eta_i = \mu + \alpha_i$ for $i = 1, 2, \ldots, m$) is $\sigma_e^2/(\sigma_\alpha^2 + \sigma_e^2)$. Thus, CA-MCMC sampling outperforms original Gibbs sampling as well as hierarchical centering. This simple example also illustrates another important feature of the CA-MCMC algorithm, specifically, the concept of sufficient statistics, which can be nicely integrated into MCMC sampling and dramatically improves convergence of the MCMC algorithm.

2.6.3 An Illustration

To illustrate how an MCMC algorithm can be adjusted to achieve faster convergence and better mixing, we consider the following ordinal response data problem. The data are given in Table 2.1.

TABLE 2.1. The Rating Data.

Gender	F	M	F	M	F	M	F	M
Rating	good	fair	good	poor	good	poor	good	good

We code female as $X = 1$ and male as $X = 0$ and we also denote the response (Y) to be 1 for "poor," 2 for "fair," and 3 for "good." We let $\boldsymbol{y} = (y_1, y_2, \ldots, y_8)'$ denote a 8×1 vector of n independent ordinal responses. Assume that y_i takes a value of l $(1 \leq l \leq 3)$ with probability

$$p_{il} = \Phi(\gamma_l + \boldsymbol{x}_i'\boldsymbol{\beta}) - \Phi(\gamma_{l-1} + \boldsymbol{x}_i'\boldsymbol{\beta}),$$

for $i = 1, ..., 8$ and $l = 1, 2, 3$, where $-\infty = \gamma_0 < \gamma_1 \leq \gamma_2 < \gamma_3 = \infty$, x_i is a 2×1 column vector of covariates denoting the intercept and gender, and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is a 2×1 column vector of regression coefficients. To ensure identifiability, we fix $\gamma_1 = 0$. Let $D = (\boldsymbol{y}, X)$, where X is the 8×2 design matrix with $\boldsymbol{x}_i'$ as its i^{th} row.

Using (2.5.12), the complete-data likelihood is

$$L(\boldsymbol{\beta}, \gamma_2, \boldsymbol{z}|D) \propto \prod_{i=1}^{8} [\exp\{-\tfrac{1}{2}(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\} 1\{\gamma_{y_i - 1} \leq z_i < \gamma_{y_i}\}], \qquad (2.6.7)$$

where $\boldsymbol{z} = (z_1, z_2, \ldots, z_8)'$ is the vector of latent variables such that $y_i = l$ if $\gamma_{l-1} \leq z_i < \gamma_l$ for $l = 1, 2, 3$ and $i = 1, 2, \ldots, 8$. Consider a prior distribution for $(\boldsymbol{\beta}, \gamma_2)$ taking the form

$$\pi(\boldsymbol{\beta}, \gamma_2) \propto \exp\Big\{ -\frac{\tau}{2}\boldsymbol{\beta}'\boldsymbol{\beta} \Big\}, \qquad (2.6.8)$$

where $\tau > 0$ is a known precision parameter. Here we take $\tau = 0.001$. Using (2.6.7) and (2.6.8), the posterior for $(\boldsymbol{\beta}, \gamma_2, \boldsymbol{z})$ is given by

$$\pi(\boldsymbol{\beta}, \gamma_2, \boldsymbol{z}|D) \propto L(\boldsymbol{\beta}, \gamma_2, \boldsymbol{z}|D)\pi(\boldsymbol{\beta}, \gamma_2). \qquad (2.6.9)$$

Using the necessary and sufficient conditions of Chen and Shao (1999a), it can be shown that when $\pi(\boldsymbol{\beta}, \gamma_2) \propto 1$, the posterior given in (2.6.9) is improper. With the choice of $\tau = 0.001$, it is expected that the resulting posterior is essentially flat.

We first implement the original Gibbs sampler, which requires the following steps:

Step 1. Sample $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta}|\boldsymbol{z}, \gamma_2 \ \sim \ N_2(\hat{\boldsymbol{\beta}}, B^{-1}),$$

where $B = \tau I_2 + X'X$ and $\hat{\boldsymbol{\beta}} = B^{-1}X'\boldsymbol{z}$.

Step 2. Sample z_i from

$$z_i \ \sim \ N(\boldsymbol{x}_i'\boldsymbol{\beta}, 1), \ \ \gamma_{y_i - 1} \leq z_i \leq \gamma_{y_i}.$$

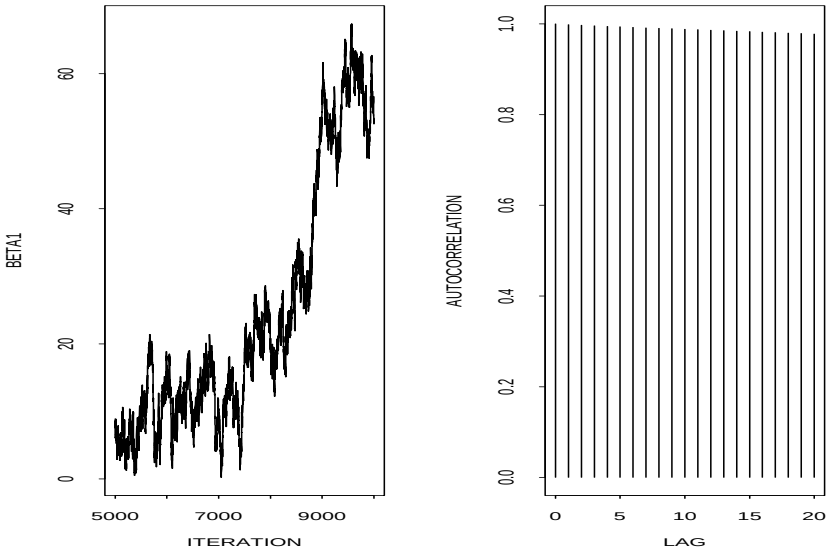

FIGURE 2.1. The original Gibbs sampler sequence of β_1 and its autocorrelation plot.

Step 3. Sample γ_2 from

$$\gamma_2|\boldsymbol{\beta}, \boldsymbol{z} \sim U[a_2, b_2],$$

where $a_2 = \max\left\{0, \max_{y_i=2} z_i\right\}$ and $b_2 = \min_{y_i=3} z_i$.

The trajectory and autocorrelation plots are displayed in Figure 2.1. These plots suggest that the original Gibbs sampler performs very poorly.

To improve the original Gibbs sampling algorithm, we consider the GM-MGMC algorithm. The group transformations proposed by Liu and Sabatti (1998) are

$$g(\boldsymbol{\beta}, \gamma_2, \boldsymbol{z}) = (g\boldsymbol{\beta}, g\gamma_2, g\boldsymbol{z})$$

with $g > 0$. Since the Jacobian $J_g = g^{8+2+1}$ and the Haar measure $H(dg) = dg/g$, the distribution of g is

$$\pi(g|\boldsymbol{\beta}, \gamma_2, \boldsymbol{z})H(dg) \propto g^{10} \exp\{-\tfrac{1}{2}g^2[\tau\boldsymbol{\beta}'\boldsymbol{\beta} + (\boldsymbol{z} - X\boldsymbol{\beta})'(\boldsymbol{z} - X\boldsymbol{\beta})]\}.$$

In addition to the original Gibbs steps, the GM-MGMC algorithm requires the following GM step:

GM Step. Draw the group element g from $\pi(g|\boldsymbol{\beta}, \gamma_2, \boldsymbol{z})H(dg)$ by taking $g = \sqrt{g^2}$, where

$$g^2|\boldsymbol{z}, \boldsymbol{\beta} \sim \mathcal{G}(\tfrac{11}{2}, \tfrac{1}{2}[(\boldsymbol{z} - X\boldsymbol{\beta})'(\boldsymbol{z} - X\boldsymbol{\beta}) + \tau\boldsymbol{\beta}'\boldsymbol{\beta}]),$$

where $\mathcal{G}(\xi, \eta)$ denotes the gamma distribution, whose density is given by

$$\pi(g^2|\xi, \eta) \propto (g^2)^{\xi-1} \exp(-\eta g^2),$$

and *adjust* $\boldsymbol{\beta}$, γ_2, and $\boldsymbol{z}$ by

$$\boldsymbol{\beta} \leftarrow g\boldsymbol{\beta}, \quad \gamma_2 \leftarrow g\gamma_2, \quad \text{and} \quad \boldsymbol{z} \leftarrow g\boldsymbol{z}.$$

The GM-MGMC algorithm has a statistical interpretation in terms of the CA-MCMC of Liu (1998). Given fixed cutpoints, the model reduces to the probit model with the corresponding variance parameter of latent variables fixed at one. The basic idea is to expand this hidden variance by redrawing the following sufficient statistic:

$$S^2 = \sum_{i=1}^{8} (z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2.$$

To make use of the CA-MCMC algorithm, we consider the following one-to-one mapping:

$$s = S = \sqrt{S^2}, \quad e_i = (z_i - \boldsymbol{x}_i'\boldsymbol{\beta})/S, \quad \eta = \gamma_2/S, \quad \text{and} \quad \boldsymbol{\xi} = \boldsymbol{\beta}/S,$$

with the constraint $\sum_{i=1}^{8} e_i^2 = 1$. Since the Jacobian of this transformation (with fixed $\gamma_1 = 0$) is

$$J_{(z_1,\ldots,z_7,\boldsymbol{\beta},\gamma_2,z_8)\to(e_1,\ldots,e_7,\boldsymbol{\xi},\eta,s)} = s^{10}/\sqrt{e_8^2},$$

given $(e_1, ..., e_8, \eta, \boldsymbol{\xi})$ the conditional distribution of s^2 is a gamma distribution:

$$\mathcal{G}(\tfrac{11}{2}, [1 + \tau\boldsymbol{\xi}'\boldsymbol{\xi}]/2).$$

Thus, in addition to the original Gibbs steps, the CA-MCMC algorithm requires the following CA step:

CA Step. Draw s^2 from $\mathcal{G}\left(\frac{11}{2}, [1 + \tau\boldsymbol{\xi}'\boldsymbol{\xi}]/2\right)$ and *adjust* $(\boldsymbol{z}, \boldsymbol{\beta}, \gamma_2)$ by

$$(\boldsymbol{z}, \boldsymbol{\beta}, \gamma_2) \leftarrow (s/S)(\boldsymbol{z}, \boldsymbol{\beta}, \gamma_2),$$

where $S^2 = \sum_{i=1}^{8} (z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$.

This version of CA-MCMC leads to the same result as that of the GM-MGMC algorithm.

The trajectory and autocorrelation plots of the GM-MGMC algorithm are displayed in Figure 2.2. From these plots, it is clear that the GM-MGMC algorithm substantially improves the original Gibbs sampler. However, the autocorrelations are still large. For example, the autocorrelation of β_1 at lag 10 is 0.404. This may be mainly due to the lack of information to estimate β_1, the regression coefficient for gender, which results in slow convergence of $\mu = \beta_0 + \beta_1$. To speed up the GM-MGMC algorithm further, we add another CA step that draws the parameter $\mu = \beta_0 + \beta_1$ jointly with its

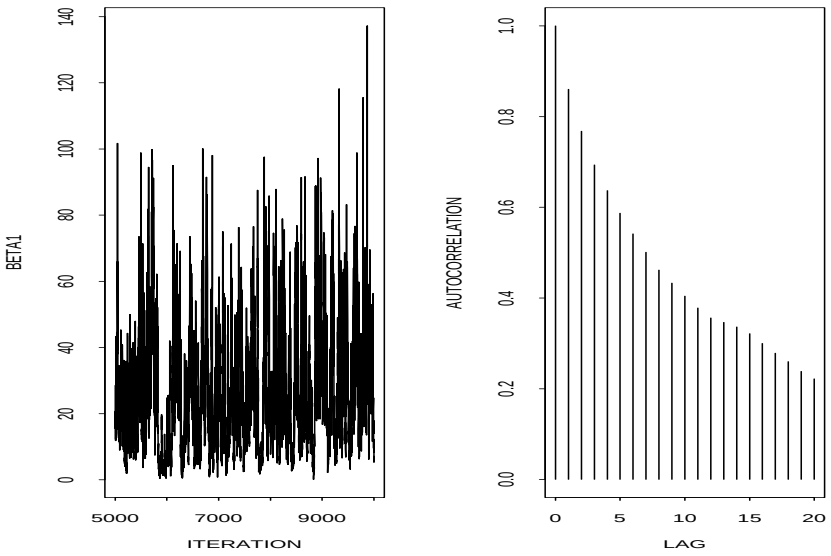

FIGURE 2.2. The GM-MGMC sequence of β_1 and its autocorrelation plot.

sufficient statistic $T = \frac{1}{4}\sum_{x_i=1} z_i$, conditioning on the current draws of $\{z_i : x_i = 0\}$, γ_2, β_0, and $\{z_i^* = z_i - T : x_i = 1\}$. Since the conditional distribution of μ given β_0 from the prior distribution of β is $N(\beta_0, 1/\tau)$, the conditional posterior distribution of (μ, T) given $\{z_i : x_i = 0\}$, γ_2, β_0, and $\{z_i^* = z_i - T : x_i = 1\}$ is

$$N_2\left(\begin{bmatrix}\beta_0\\ \beta_0\end{bmatrix}, \begin{bmatrix}1/\tau & 1/\tau\\ 1/\tau & 1/\tau + \frac{1}{4}\end{bmatrix}\right),$$

where $\max_{x_i=1}(\gamma_2 - z_i^*) \le T < \infty$. Thus, the corresponding CA step in this CA GM-MGMC algorithm can be accomplished by: (i) drawing T from

$$N(\beta_0, 1/\tau + \tfrac{1}{4}),$$

where $\max_{x_i=1}(\gamma_2 - z_i^*) \le T < \infty$, then drawing μ from

$$N\left(\beta_0 + \frac{(1/\tau)}{1/\tau + \frac{1}{4}}(T - \beta_0), \frac{1}{\tau} - \frac{(1/\tau^2)}{1/\tau + \frac{1}{4}}\right),$$

and (ii) adjusting β_1 and $\{z_i : x_i = 1\}$ by

$$\beta_1 \leftarrow \mu - \beta_0 \text{ and } \{z_i : x_i = 1\} \leftarrow \{z_i^* + T : x_i = 1\}.$$

Figure 2.3 indicates that the autocorrelations of β_1 from the CA GM-MGMC algorithm disappear even at lag 1. This simple example illustrates three important points:

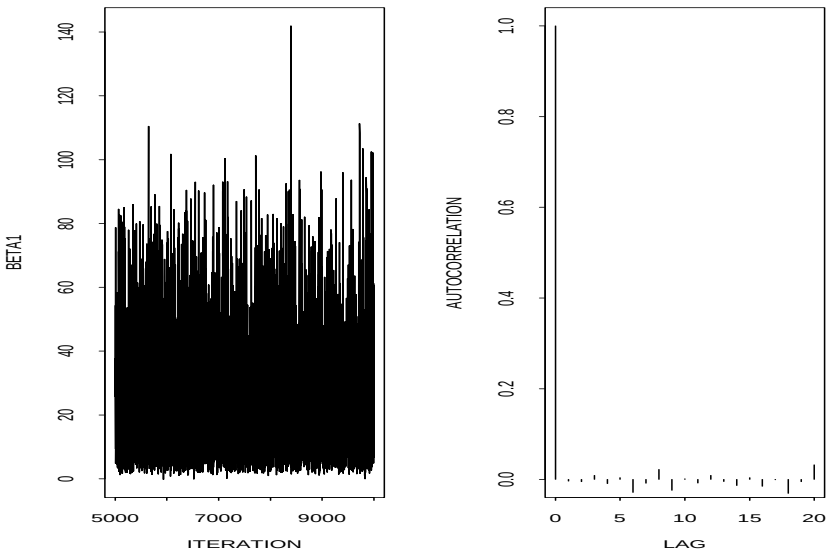

FIGURE 2.3. The CA GM-MGMC sequence of β_1 and its autocorrelation plot.

(i) the adjustment steps can dramatically improve convergence of an MCMC algorithm;

(ii) accelerating an MCMC algorithm is a continuous process; and

(iii) conditioning on sufficient statistics may play a key role in accelerating an MCMC algorithm.

2.7 Dynamic Weighting Algorithm

The *dynamic weighting method* is first introduced by Wong and Liang (1997) and further examined by Liu, Liang, and Wong (1998b). As pointed out by Liu, Liang, and Wong (1998b), the method extends the basic Markov chain equilibrium concept of Metropolis et al. (1953) to a more general weighted equilibrium of a Markov chain. The basic idea of dynamic weighting is to augment the original sample space by a positive scalar w, called a weight function, which can automatically adjust its own value to help the sampler move more freely.

Introducing the importance weights into the dynamic MC process helps make large transitions which are not allowed by the standard Metropolis transition rules. When the distribution has regions of "high" density separated by barriers of very "low" density, for example, when the tar-

get distribution is multimodal, the waiting time for a Metropolis process to cross over the barriers will be essentially infinite. In the dynamically weighted Monte Carlo, the process can often move against very steep probability barriers, which apparently violates the Metropolis rule. The weight variable is updated in a way that allows for an adjustment of the bias induced by such non-Metropolis moves.

Similar to the Metropolis algorithm, dynamic weighting starts with an arbitrary Markov transition kernel $T(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ from which the next candidate move is suggested. Suppose the current state is $(\boldsymbol{\theta}, w)$. Liu, Liang, and Wong (1998b) propose two dynamic weighting moves, called the Q-type move and the R-type move. Assume that the target distribution is $\pi(\boldsymbol{\theta}|D)$. Then, these two dynamic weighting schemes are given as follows:

Q-Type Move

Step 1. (Candidate state.) Draw $\boldsymbol{\vartheta} \sim T(\boldsymbol{\theta}, \boldsymbol{\vartheta})$, and compute the Metropolis ratio

$$r(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = \frac{\pi(\boldsymbol{\vartheta}|D)T(\boldsymbol{\vartheta}, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|D)T(\boldsymbol{\theta}, \boldsymbol{\vartheta})}.$$

Step 2. (Move?) Choose $\alpha = \alpha(w, \boldsymbol{\theta}) > 0$ and draw $u \sim U(0, 1)$. Update $(\boldsymbol{\theta}, w)$ to $(\boldsymbol{\theta}^*, w^*)$ as

$$(\boldsymbol{\theta}^*, w^*) = \begin{cases} (\boldsymbol{\vartheta}, \max\{\boldsymbol{\alpha}, wr(\boldsymbol{\theta}, \boldsymbol{\vartheta})\}) & \text{if } u \le \min\{1, wr(\boldsymbol{\theta}, \boldsymbol{\vartheta})/\alpha\}, \\ (\boldsymbol{\theta}, aw) & \text{otherwise,} \end{cases} \tag{2.7.1}$$

where $a > 1$ can be either a constant or an independent random variable.

R-Type Move

Step 1. (Candidate state.) The same as the Q-type move.

Step 2. (Move?) Choose $\alpha = \alpha(w, \boldsymbol{\theta}) > 0$ and draw $u \sim U(0, 1)$. Update $(\boldsymbol{\theta}, w)$ to $(\boldsymbol{\theta}^*, w^*)$ as

$$(\boldsymbol{\theta}^*, w^*) = \begin{cases} (\boldsymbol{\vartheta}, wr(\boldsymbol{\theta}, \boldsymbol{\vartheta}) + \alpha) & \text{if } u \le \dfrac{wr(\boldsymbol{\theta}, \boldsymbol{\vartheta})}{wr(\boldsymbol{\theta}, \boldsymbol{\vartheta}) + \alpha}, \\ (\boldsymbol{\theta}, w(wr(\boldsymbol{\theta}, \boldsymbol{\vartheta}) + \alpha)/\alpha)) & \text{otherwise.} \end{cases} \tag{2.7.2}$$

For practical use of the two dynamic weighting moves, Liu, Liang, and Wong (1998b) suggest that they be applied in a compact space. This can be achieved by preventing the sampler from visiting exceedingly low-probability space. Furthermore, to guard against a possible boundary effect

caused by exceedingly small $r(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ (i.e., practically 0), one can modify the weight updating as follows: if $r(\boldsymbol{\theta}, \boldsymbol{\vartheta}) < \epsilon$ for a proposal $\boldsymbol{\vartheta}$, rejection does not induce any change of the weights.

The behavior of both dynamic weighting moves is controlled by the parameter α and the transition kernel $T(\boldsymbol{\theta}, \boldsymbol{\vartheta})$. For example, when $\alpha \to 0$, the Q-type move is identical to the R-type move, and every candidate move will be accepted. Two special cases are of great interest. First, when $T(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is reversible and stationary with equilibrium distribution $\pi(\boldsymbol{\theta}|D)$, both moves reduce to the standard Metropolis algorithm. Thus, the dynamic weighting method can be viewed as an extension of the standard Metropolis algorithm. Second, when $T(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ is reversible and stationary with equilibrium distribution $g(\boldsymbol{\theta})$,

$$r(\boldsymbol{\theta}, \boldsymbol{\vartheta}) = \frac{\pi(\boldsymbol{\vartheta}|D)g(\boldsymbol{\theta})}{g(\boldsymbol{\vartheta})\pi(\boldsymbol{\theta}|D)} \quad \text{and} \quad w^* = w\frac{\omega(\boldsymbol{\vartheta})}{\omega(\boldsymbol{\theta})},$$

where $w(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|D)/g(\boldsymbol{\theta})$. Hence, if we start with $\boldsymbol{\theta}_0$ and $w_0 = c_0\omega(\boldsymbol{\theta}_0)$, then for any $i > 0$, $w_i = c_0\omega(\boldsymbol{\theta}_i)$. These weights are identical to those from the standard importance sampling method with an importance sampling distribution $g(\boldsymbol{\theta})$. Liu, Liang, and Wong (1998b) also study the behavior of the Q-type and R-type moves for several other choices of α and T.

In general, for either the Q-type move or the R-type move, the equilibrium distribution of $\boldsymbol{\theta}$ (if it exists) is not $\pi(\boldsymbol{\theta}|D)$. In this regard, Wong and Liang (1997) propose to use *invariance with respect to importance-weighting* (IWIW) as a principle for validating the above scheme and for designing new transition rules. The formal definition of IWIW is given as follows:

> *The joint distribution $\pi(\boldsymbol{\theta}, w)$ of $(\boldsymbol{\theta}, w)$ is said to be correctly weighted with respect to $\pi(\boldsymbol{\theta}|D)$ if*
>
> $$\int w\pi(\boldsymbol{\theta}, w)\, dw \propto \pi(\boldsymbol{\theta}|D). \tag{2.7.3}$$
>
> *A transition rule is said to satisfy IWIW if it maintains the correctly weighted property for the joint distribution of $(\boldsymbol{\theta}, w)$ whenever the initial joint distribution is correctly weighted.*

Suppose the starting joint distribution $\pi_1(\boldsymbol{\theta}, w)$ for $(\boldsymbol{\theta}, w)$ is correctly weighted with respect to $\pi(\boldsymbol{\theta}|D)$, i.e., $\int w\pi_1(\boldsymbol{\theta}, w)\, dw \propto \pi(\boldsymbol{\theta}|D)$. It can be shown that after a one-step transition of the R-type move with $\alpha = \alpha(\boldsymbol{\theta}, w) > 0$ for all $(\boldsymbol{\theta}, w)$, the new joint state $(\boldsymbol{\theta}^*, w^*)$ has a joint distribution $\pi_2(\boldsymbol{\theta}^*, w^*)$ that is also correctly weighted with respect to $\pi(\boldsymbol{\theta}|D)$, i.e.,

$$\int w^*\pi_2(\boldsymbol{\theta}^*, w^*)\, dw^* \propto \pi(\boldsymbol{\theta}^*|D). \tag{2.7.4}$$

If $\alpha \to 0$, then the IWIW property holds for both the Q- and R-type moves. However, when $\alpha > 0$, the Q-type move only approximately satisfies the

IWIW property. A more detailed discussion of the properties of the Q-type move can be found in Liu, Liang, and Wong (1998b).

The dynamic reweighting method has been successfully applied to simulation and global optimization problems arising from multimodal sampling, neural network training, high-dimensional integration, the Ising models (Wong and Liang 1997; Liu, Liang, and Wong 1998b), and Bayesian model selection problems (Liu and Sabatti 1999). The applications of IWIW to the computation of posterior quantities of interest will be discussed further in Chapter 3.

2.8 Toward "Black-Box" Sampling

Chen and Schmeiser (1998) propose a random-direction interior-point (RDIP) Markov chain approach to black-box sampling. The purpose of such a black-box sampler is to free the analyst from computational details without paying a large computational penalty, in contrast to specialized samplers such as the Gibbs sampler or the Metropolis–Hastings algorithm.

Assume that the target posterior distribution is of the form

$$\pi(\boldsymbol{\theta}|D) = \frac{L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})}{c(D)}, \tag{2.8.1}$$

where $L(\boldsymbol{\theta}|D)$ is the likelihood function, $\pi(\boldsymbol{\theta})$ is a prior distribution, and $c(D)$ is the normalizing constant. We further assume that $L(\boldsymbol{\theta}|D)$ and $\pi(\boldsymbol{\theta})$ can be computed at any point $\boldsymbol{\theta}$. The key idea of RDIP is to introduce an auxiliary random variable δ so that the joint posterior distribution of $(\boldsymbol{\theta}, \delta)$ has the form

$$\pi(\boldsymbol{\theta}, \delta|D) = \begin{cases} 1/c(D), & \text{if } 0 \le \delta \le \pi(\boldsymbol{\theta}|D), \\ 0, & \text{otherwise.} \end{cases} \tag{2.8.2}$$

Integrating out δ from $\pi(\boldsymbol{\theta}, \delta|D)$ yields the marginal distribution of $\boldsymbol{\theta}$ as $\pi(\boldsymbol{\theta}|D)$. This result implies that if a Markov chain $\{(\boldsymbol{\theta}_i, \delta_i), i = 1, 2, ...\}$ has the unique uniform stationary distribution $\pi(\boldsymbol{\theta}, \delta|D)$, then the "marginal" Markov chain $\{\boldsymbol{\theta}_i, i = 1, 2, ...\}$ has a stationary distribution which is the target posterior $\pi(\boldsymbol{\theta}|D)$.

Let Ω be the interior of the $(p+1)$-dimensional region lying beneath $\pi(\boldsymbol{\theta}|D)$ and over the support of $\pi(\boldsymbol{\theta}|D)$. Then the RDIP sampler has three fundamental characteristics:

(i) Sampling generates points $(\boldsymbol{\theta}, \delta)$ from the interior of Ω.

(ii) The stationary distribution of $(\boldsymbol{\theta}, \delta)$ is uniform over Ω. Therefore, the stationary distribution of $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta}|D)$.

(iii) The Markov chain evolution from point to point is based on random directions.

Computationally, whether $\pi(\boldsymbol{\theta}|D)$ integrates to one or not is unimportant. Suppose that $\pi(\boldsymbol{\theta}|D) = L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})$. The advantage of this is that the normalizing constant $c(D)$ need not be computed. Based on this convention, the general version of the RDIP sampler, which essentially defines a class of samplers, can be stated as follows:

The RDIP Sampler

Step 1. (Random direction.) Generate a unit-length $(p+1)$-dimensional direction $\boldsymbol{d} \sim g_1(\boldsymbol{d}|\boldsymbol{\theta}, \delta)$.

Step 2. (Random distance.) Generate $\lambda \sim g_2(\lambda|\boldsymbol{\theta}, \delta, \boldsymbol{d})$.

Step 3. (Candidate point.) Set $(\boldsymbol{\theta}^*, \delta^*) = (\boldsymbol{\theta}, \delta) + \lambda \boldsymbol{d}$.

Step 4. (Candidate posterior density.) Compute $\pi^* = \pi(\boldsymbol{\theta}^*|D)$.

Step 5. (Inside Ω?) If $0 < \delta^* < \pi^*$ is false, go to Step 7.

Step 6. (Move?) Set $(\boldsymbol{\theta}, \delta) = (\boldsymbol{\theta}^*, \delta^*)$ with probability $a(\boldsymbol{\theta}^*, \delta^*|\boldsymbol{\theta}, \delta)$.

Step 7. (Done.) Return $(\boldsymbol{\theta}, \delta)$.

The densities g_1 and g_2 in Steps 1 and 2 and the jump probability a in Step 6 must be chosen to provide a valid sampler. In typical versions, g_1, g_2, and a can be chosen so that the transition kernel of the random sequence of points $\{(\boldsymbol{\theta}_i, \delta_i),\ \ i \geq 0\}$ is doubly stochastic in Ω, guaranteeing uniformity over Ω in the limit. In general, Steps 1 and 2 can be combined to generate a $(\boldsymbol{d}, \lambda)$ conditional on $(\boldsymbol{\theta}, \delta)$. From the above description, it can be seen that the RDIP sampler is a special case of the H&R sampler.

Chen and Schmeiser (1998) discuss three variations of the general version of the RDIP sampler. One of these variations is called the state-dependent direction-and-radius sampler, which uses the location information of the current state as well as the relative height of the current location without requiring much extra computation. The detailed steps involved in this special case are given as follows:

The State-Dependent Direction-and-Radius Sampler

Step 1. (Random direction.)

- Generate a uniform unit-length p-dimensional direction $(d_1, d_2, \ldots, d_p)$;
- generate $\alpha \sim g_1^*(\alpha|r)$; and
- the $(p+1)$-dimensional random direction is $\boldsymbol{d} = (d_1 \cos\alpha, d_2 \cos\alpha, \ldots, d_p \cos\alpha, \sin\alpha)$.

Step 2. (Random distance.) Generate $\lambda \sim g_2^*(\lambda|\boldsymbol{\theta}, \delta, \boldsymbol{d})$.

Step 3. (Candidate point.) Set $(\boldsymbol{\theta}^*, \delta^*) = (\boldsymbol{\theta}, \delta) + \lambda \boldsymbol{d}$.

Step 4. (Candidate posterior density.) Compute $\pi^* = \pi(\boldsymbol{\theta}^*|D)$ and $r^* = \delta^*/\pi(\boldsymbol{\theta}^*|D)$.

Step 5. (Inside Ω?) If $0 < \delta^* < \pi^*$ is false, go to Step 7.

Step 6. (Move?) Let $\pi(\boldsymbol{\theta}^*, \delta^*|\boldsymbol{\theta}, \delta)$ be the conditional probability density function of the next candidate point $(\boldsymbol{\theta}^*, \delta^*)$ given the current point $(\boldsymbol{\theta}, \delta)$, and let $\pi(\boldsymbol{\theta}, \delta|\boldsymbol{\theta}^*, \delta^*)$ be the conditional probability density function by switching the positions of $(\boldsymbol{\theta}^*, \delta^*)$ and $(\boldsymbol{\theta}, \delta)$. Then set $(\boldsymbol{\theta}, \delta) = (\boldsymbol{\theta}^*, \delta^*)$ with probability $\min\{\pi(\boldsymbol{\theta}, \delta|\boldsymbol{\theta}^*, \delta^*)/\pi(\boldsymbol{\theta}^*, \delta^*|\boldsymbol{\theta}, \delta), 1\}$.

Step 7. (Done.) Return $(\boldsymbol{\theta}, \delta)$.

Next, we discuss some possible choices of the angle density $g_1^*(\alpha|r)$, the distance density $g_2^*(\lambda|\boldsymbol{\theta}, \delta, \boldsymbol{d})$, the conditional density $\pi(\boldsymbol{\theta}^*, \delta^*|\boldsymbol{\theta}, \delta)$, and the jump probability.

The angle α is with respect to the horizontal plane $\Delta = \delta$. The domain of α is $(-\pi/2, \pi/2)$, with $\alpha = -\pi/2$ corresponding to a move straight down and $\alpha = \pi/2$ corresponding to a move straight up in $p+1$ dimensions. A reasonable choice of $g_1^*(\alpha|r)$ might be a mixture of beta densities

$$\begin{aligned} g_1^*(\alpha|r) =& P(\text{neg})p(\alpha|\text{neg}) + P(\text{pos})p(\alpha|\text{pos}) \\ =& r\,\frac{(-\alpha)^{a_r-1}((\pi/2)+\alpha)^{b_r-1}}{B(a_r,b_r)(\pi/2)^{a_r+b_r-1}} 1\{-\pi/2 < \alpha < 0\} \\ &+ (1-r)\,\frac{(\alpha)^{a_r-1}((\pi/2)-\alpha)^{b_r-1}}{B(a_r,b_r)(\pi/2)^{a_r+b_r-1}} 1\{0 < \alpha < \pi/2\}, \end{aligned} \tag{2.8.3}$$

where $a_r > 0$, $b_r > 0$, a_r and b_r might depend on r, and $B(a_r, b_r)$ is the beta function. Note that $E(\alpha|r) = [a_r/(a_r+b_r)](\pi/2)(1-2r)$. Here we choose the probability of moving down to be $P(\text{neg}) = r$ and the probability of moving up to be $P(\text{pos}) = 1-r$. The reason for this choice is that if the point $(\boldsymbol{\theta}, \delta)$ is close to the surface $\delta = \pi(\boldsymbol{\theta}|D)$ (i.e., r is close to 1), more probability should be assigned to negative values of α (i.e., the next move should be down), and vice versa. In addition, when r is close to 1, more probability should be assigned to the large absolute value of α, and, therefore, a_r should be large while b_r should be small. Similarly, when r is close to zero, a_r should be small and b_r should be large. Chen and Schmeiser (1998) empirically show that despite choosing a_r and b_r to be constants, the sampler still performs reasonably well.

It is desirable to choose the distance density g_2^* so that it depends upon the current location and the angle α, without incurring expensive computation. One source of almost-free information is to compute the intersection of the line through the point $(\boldsymbol{\theta}, \delta)$ with the direction $\boldsymbol{d}$ and the horizontal plane $\delta = 0$. This intersection is

$$Pt(\boldsymbol{\theta}, \boldsymbol{d}, 0) = (\boldsymbol{\theta} - (\delta \cos\alpha/\sin\alpha)(d_1, d_2, \ldots, d_p), 0)\,.$$

Similarly, the intersection of the line through the point $(\boldsymbol{\theta}, \delta)$ with the direction $\boldsymbol{d}$ and the horizontal plane $\delta = \pi(\boldsymbol{\theta}|D)$ is

$$Pt(\boldsymbol{\theta}, \boldsymbol{d}, \pi(\boldsymbol{\theta}|D)) = (\boldsymbol{\theta} - ((\delta - \pi(\boldsymbol{\theta}|D))\cos\alpha/\sin\alpha)(d_1, d_2, \ldots, d_p), \pi(\boldsymbol{\theta}|D)).$$

Then it is easy to compute the distances from the point $(\boldsymbol{\theta}, \delta)$ to $Pt(\boldsymbol{\theta}, \boldsymbol{d}, 0)$ and $Pt(\boldsymbol{\theta}, \boldsymbol{d}, \pi(\boldsymbol{\theta}|D))$, say, λ_1 and λ_2, respectively:

$$\lambda_1 = ||(\boldsymbol{\theta}, \delta) - Pt(\boldsymbol{\theta}, \boldsymbol{d}, 0)|| = \frac{\delta}{|\sin\alpha|}, \tag{2.8.4}$$

$$\lambda_2 = ||(\boldsymbol{\theta}, \delta) - Pt(\boldsymbol{\theta}, \boldsymbol{d}, \pi(\boldsymbol{\theta}|D))|| = \frac{\pi(\boldsymbol{\theta}|D) - \delta}{|\sin\alpha|}. \tag{2.8.5}$$

The distance distribution is chosen based on this information. For example, a gamma distribution might be appropriate when α is positive and a uniform distribution over $(0, \lambda_1)$ might be appropriate when α is negative. More specifically, we can choose

$$g_2^*(\lambda|\theta, \delta, \boldsymbol{d}) = g_{2a}^*(\lambda|\boldsymbol{\theta}, \delta)1\{\alpha < 0\} + g_{2b}^*(\lambda|\boldsymbol{\theta}, \delta)1\{\alpha > 0\}, \tag{2.8.6}$$

where

$$g_{2a}^*(\lambda|\boldsymbol{\theta}, \delta) = \begin{cases} |\sin\alpha|/\delta & \text{for } 0 \le \lambda \le \delta/|\sin\alpha|, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$g_{2b}^*(\lambda|\boldsymbol{\theta}, \delta) = \begin{cases} \dfrac{\lambda^2 \exp\{-6|\sin\alpha|\lambda/(\pi(\boldsymbol{\theta}|D) - \delta)\}}{\Gamma(3)\left((\pi(\boldsymbol{\theta}|D) - \delta)/6|\sin\alpha|\right)^3} & \text{for } \lambda > 0, \\ 0 & \text{otherwise.} \end{cases}$$

That is, $g_2^*(\lambda|\boldsymbol{\theta}, \delta, \boldsymbol{d})$ in (2.8.6) is either the uniform distribution $U(0, \delta/|\sin\alpha|)$ or the gamma distribution with a shape parameter 3 and a scale parameter $(\pi(\boldsymbol{\theta}|D) - \delta)/(6|\sin\alpha|)$.

With the above choices of $g_1^*(\alpha|r)$ and $g_2^*(\lambda|\boldsymbol{\theta}, \delta, \boldsymbol{d})$, the density of the next candidate point $(\boldsymbol{\theta}^*, \delta^*)$ conditional on the current point $(\boldsymbol{\theta}, \delta)$ is

$$\pi(\boldsymbol{\theta}^*, \delta^*|\boldsymbol{\theta}, \delta) \propto \frac{g_1^*(\alpha|r)g_2^*(\lambda|\boldsymbol{\theta}, \delta, \boldsymbol{d})}{|\lambda|^p|\cos\alpha|^{p-1}}, \tag{2.8.7}$$

where $\lambda = ||(\boldsymbol{\theta}^* - \boldsymbol{\theta}, \delta^* - \delta)||$, $\boldsymbol{d} = (\boldsymbol{\theta}^* - \boldsymbol{\theta}, \delta^* - \delta)/\lambda$, $\alpha = \sin^{-1}((\delta^* - \delta)/\lambda)$, and $r = \delta/\pi(\boldsymbol{\theta}|D)$. Let α^*, $\boldsymbol{d}^*$, and λ^* denote the angle, the direction, and the distance from the point $(\boldsymbol{\theta}^*, \delta^*)$ back to the point $(\boldsymbol{\theta}, \delta)$. Then $\lambda^* = ||(\boldsymbol{\theta} - \boldsymbol{\theta}^*, \delta - \delta^*)|| = \lambda$, $\boldsymbol{d}^* = (\boldsymbol{\theta} - \boldsymbol{\theta}^*, \delta - \delta^*)/\lambda = -\boldsymbol{d}$, $\alpha^* = \sin^{-1}((\delta - \delta^*)/\lambda) = -\alpha$, and $r^* = \delta^*/\pi(\boldsymbol{\theta}^*|D)$. Therefore, the jump ratio is

$$\frac{\pi(\boldsymbol{\theta}, \delta|\boldsymbol{\theta}^*, \delta^*)}{\pi(\boldsymbol{\theta}^*, \delta^*|\boldsymbol{\theta}, \delta)} = \left[\frac{g_1^*(-\alpha|r^*)}{g_1^*(\alpha|r)}\right] \cdot \left[\frac{g_2^*(\lambda|\boldsymbol{\theta}^*, \delta^*, -\boldsymbol{d})}{g_2^*(\lambda|\boldsymbol{\theta}, \delta, \boldsymbol{d})}\right]. \tag{2.8.8}$$

With $g_2^*(\lambda|\boldsymbol{\theta},\delta,\boldsymbol{d})$ given by (2.8.6), (2.8.8) can be further simplified as

$$\frac{\pi(\boldsymbol{\theta},\delta|\boldsymbol{\theta}^*,\delta^*)}{\pi(\boldsymbol{\theta}^*,\delta^*|\boldsymbol{\theta},\delta)} = \begin{cases} \dfrac{\pi(\boldsymbol{\theta}|D)}{3\pi(\boldsymbol{\theta}^*|D)} \cdot \left(\dfrac{\pi(\boldsymbol{\theta}|D)-\delta}{6|\sin\alpha|\lambda}\right)^2 \cdot \exp\left\{\dfrac{6|\sin\alpha|\lambda}{\pi(\boldsymbol{\theta}|D)-\delta}\right\} & \text{for } \alpha > 0, \\ \dfrac{3\pi(\boldsymbol{\theta}|D)}{\pi(\boldsymbol{\theta}^*|D)} \cdot \left(\dfrac{6|\sin\alpha|\lambda}{\pi(\boldsymbol{\theta}^*|D)-\delta^*}\right)^2 \cdot \exp\left\{-\dfrac{6|\sin\alpha|\lambda}{\pi(\boldsymbol{\theta}^*|D)-\delta^*}\right\} & \text{for } \alpha < 0. \end{cases} \tag{2.8.9}$$

In Step 1, the uniform distribution for a unit-length p- dimensional direction $(d_1, d_2, \ldots, d_p)$ can be extended to any continuous distribution over the surface of the p-dimensional unit sphere. Additional discussion can be found in Chen and Schmeiser (1996). Chen and Schmeiser (1998) show that with the above choices of $g_1^*(\alpha|r)$ and $g_2^*(\lambda|\boldsymbol{\theta},\delta,\boldsymbol{d})$, the Markov chain $\{(\boldsymbol{\theta}_i,\delta_i), i = 1,2,\ldots\}$ induced by the state-dependent direction-and-radius RDIP sampler is irreducible and doubly stochastic, and, therefore, has a unique stationary distribution $\pi(\boldsymbol{\theta},\delta|D)$. They further empirically study the performance of the RDIP sampler and find that the RDIP sampler works reasonably well for a bimodal distribution as well as for the ordinal response model in Section 2.5.3. The RDIP sampler is the first step toward black-box sampling. Further research in this direction needs to be done in the future.

2.9 Convergence Diagnostics

Convergence diagnostics are one of the most important components in MCMC sampling. For most practical problems, the MCMC sample generated from a user's selected MCMC sampling algorithm will ultimately be used for computing posterior quantities of interest. Thus, if a Markov chain induced by the MCMC algorithm fails to converge, the resulting posterior estimates will be biased and unreliable. As a consequence, an incorrect Bayesian data analysis will be performed and false conclusions may be drawn. Fortunately, many useful diagnostic tools along with their sound theoretical foundations have been developed during the last decade. Although no single diagnostic procedure can guarantee to diagnose convergence successfully, combining several diagnostic tools together may enable us to detect how fast or how slow a Markov chain converges and how well or how poorly a chain is mixing.

By now, the literature on convergence diagnostics is very rich. Excellent and comprehensive reviews are given by Cowles and Carlin (1996), Brooks and Roberts (1998), Mengersen, Robert, and Guihenneuc–Jouyaux (1998), and many other references therein. More recently, Robert (1998, Chap. 2) presents several useful methods on convergence control of MCMC

algorithms. In this section, we present several commonly used convergence diagnostic techniques, and we refer the reader to the above review articles for details of other available convergence diagnostic methods.

A simple but effective diagnostic tool is the trace plot. Two kinds of trace plots are useful, which are the trace plot of a single long-run sequence and the trace plots of several short-run sequences with overdispersed starting (initial) points. As discussed in Mengersen, Robert, and Guihenneuc–Jouyaux (1998), there is widespread debate about single run and multiple runs. A single sequence which has difficulty leaving the neighborhood of an attractive mode will exhibit acceptable behavior even though it has failed to explore the whole support of the target distribution $\pi(\boldsymbol{\theta}|D)$. Multiple sequences may have better exploratory power, but depend highly on the choice of starting points. On the other hand, a long-run single sequence may be advantageous in exploring potential coding bugs and the mixing behavior of the Markov chain, while multiple sequences suffer from a large increase in the number of wasted burn-in simulations for estimating posterior quantities. As a practical guideline, we suggest the use of both types of trace plots in exploring convergence and mixing behavior of the chain, and then generate a single long-run sequence with a large number of iterations (say 50,000) for estimation purposes.

For many practical problems, the dimension of the parameter space is high. Thus it may not be feasible to examine the trace plots for all parameters. In this case, we may construct trace plots for several selected parameters, which should include parameters, that are known to converge slowly, functions of parameters of interest, and some nuisance parameters. For example, for the ordinal response models in Section 2.5.3, we may need to monitor the trace plots for the regression coefficients $\boldsymbol{\beta}$, the cutpoints $\boldsymbol{\gamma}$, and some of the latent variables z_i's (nuisance parameters). If one is interested in estimating $\xi = h(\boldsymbol{\theta})$, it may be sufficient to monitor the trace plot for ξ only. However, we note that slow convergence of the nuisance parameters may seriously affect the convergence of parameters of interest as discussed in Section 2.5.3. Another related issue is that a simple time series plot for the sequence of ξ may not be effective if ξ is a discrete variable. In this regard, we propose the use of the cumulative sum (CUSUM) plot of Yu and Mykland (1998). The CUSUM plot can be constructed as follows. Given the output $\{\xi_1, \xi_2, \ldots, \xi_n\}$, we begin by discarding the initial n_0 iterations, which we believe to correspond to the burn-in period. Then, the following algorithm describes how to produce a CUSUM plot:

The CUSUM Plot

Step 1. Calculate $\bar{\xi} = (n - n_0)^{-1} \sum_{i=n_0+1}^{n} \xi_i$.

Step 2. Calculate the CUSUM

$$S_t = \sum_{i=n_0+1}^{t} (\xi_i - \bar{\xi}) \quad \text{for } t = n_0 + 1, \ldots, n.$$

Step 3. Plot s_t against t for $t = n_0 + 1, \ldots, n$, connecting successive points by line segments.

Yu and Mykland (1998) argue that the speed with which the chain is mixing is indicated by the smoothness of the resulting CUSUM plot, so that a smooth plot indicates slow mixing, while a "hairy" plot indicates a fast mixing rate for ξ.

The autocorrelation plots are the easiest tool for quantitatively assessing the mixing behavior of a Markov chain. It is important to check not only the within-sequence autocorrelations but also the intraclass (between-sequence) autocorrelations. However, care must be taken in computing autocorrelations. Without discarding iterations corresponding to the burn-in period, the autocorrelations may be under- or over-estimated, which may reflect a false mixing behavior of the Markov chain. For purposes of autocorrelation checking, a long-run single sequence may be more beneficial compared to multiple short-run sequences, since the long-run sequence will lead to more accurate estimates of autocorrelations. The slow decay in the autocorrelation plots indicates a slow mixing. On the other hand, the autocorrelations are also useful in estimating "effective sample size" for studying the convergence rates of the estimates of posterior quantities.

One of the most popular quantitative convergence diagnostics is the variance ratio method of Gelman and Rubin (1992). Gelman and Rubin's method consists of analyzing m independent sequences to form a distributional estimate for what is known about some random variable, given the observations simulated so far. Assume that we independently simulate $m \geq 2$ sequences of length $2n$, each beginning at different starting points from an overdispersed distribution with respect to the target distribution $\pi(\boldsymbol{\theta}|D)$. We discard the first n iterations and retain only the last n. Then, for any scalar function $\xi = h(\boldsymbol{\theta})$ of interest, we calculate the variance between the m sequence means defined by

$$\frac{B}{n} = \frac{1}{m-1} \sum_{i=1}^{m} (\bar{\xi}_{i\cdot} - \bar{\xi}_{\cdot\cdot})^2,$$

where

$$\bar{\xi}_{i\cdot} = \frac{1}{n} \sum_{t=n+1}^{2n} \xi_{it}, \quad \bar{\xi}_{\cdot\cdot} = \frac{1}{m} \sum_{i=1}^{m} \bar{\xi}_{i\cdot},$$

and $\xi_{it} = h(\boldsymbol{\theta}_{it})$ is the t^{th} observation of ξ from sequence i. Then we calculate the mean of the m within-sequence variances, s_i^2, each of which

has $n-1$ degrees of freedom, given by

$$W = \frac{1}{m}\sum_{i=1}^{m} s_i^2,$$

where $s_i^2 = (n-1)^{-1}\sum_{t=n+1}^{2n}(\xi_{it} - \bar{\xi}_{i.})^2$. An estimator of the posterior variance of ξ is

$$\hat{V} = \frac{n-1}{n}W + \Big(1+\frac{1}{m}\Big)\frac{B}{n},$$

which is asymptotically equivalent to W. Gelman and Rubin (1992) suggest the use of a t test, deduced from the approximation $B/W \sim \mathcal{F}(m-1, df)$, where $\mathcal{F}(m-1, df)$ denotes the F-distribution with degrees of freedom $(m-1, df)$, $df = 2V^2/\widehat{\mathrm{Var}}(V)$, and

$$\begin{aligned}\widehat{\mathrm{Var}}(V) &= \Big(\frac{n-1}{n}\Big)^2\frac{1}{m}\widehat{\mathrm{Var}}(s_i^2) + \Big(\frac{m+1}{mn}\Big)^2\frac{2}{m-1}B^2 \\ &\quad + 2\frac{(m+1)(n-1)}{mn^2}\frac{n}{m}[\widehat{\mathrm{Cov}}(s_i^2, \bar{\xi}_{i.}^2) - 2\bar{\xi}_{..}\widehat{\mathrm{Cov}}(s_i^2, \bar{\xi}_{i.})].\end{aligned}$$

Then, we monitor convergence by a *potential scale reduction* (PSR) factor, which is calculated by

$$\hat{R} = (V/W)df/(df-2).$$

A large value of R suggests that either V can be further decreased by more draws, or that further draws will increase W. A value of R close to 1 indicates that each of the m sets of n simulated observations is close to the target distribution, that is, convergence is achieved. The multivariate version of the PSR can be found in Brooks and Gelman (1998). The other quantitative convergence diagnostic methods include the spectral density diagnostic of Geweke (1992), the L^2 convergence diagnostics of Liu, Liu, and Rubin (1992), and Roberts (1994), geometric convergence bounds of Rosenthal (1995a,b) and Cowles and Rosenthal (1998), the convergence rate estimator of Garren and Smith (1995) and Raftery and Lewis (1992), and many others.

As with all statistical procedures, any convergence diagnostic technique can falsely indicate convergence when in fact it has not occurred. In particular, for slowly mixing Markov chains, convergence diagnostics are likely to be unreliable, since their conclusions will be based on output from only a small region of the state space. Therefore, it is important to emphasize that any convergence diagnostic procedure should not be unilaterally relied upon. As in Cowles and Carlin (1996), we recommend using a variety of diagnostic tools rather than any single plot or statistic, and learning as much as possible about the target posterior distribution before applying an MCMC algorithm. In addition, a careful study of the propriety of the posterior distribution is important, since an improper posterior makes Bayesian inference meaningless. Also, we recommend using the acceleration

tools described in Sections 2.5 and 2.6 as much as possible, since they can dramatically speed up an MCMC algorithm.

Exercises

2.1 Using the New Zealand apple data in Example 1.1, compute posterior estimates for β and σ^2 for the constrained multiple linear regression model with the prior specification for the model parameters given in Example 2.2, using the Gibbs sampler. Compare the results to those obtained by the classical-order restricted inference in Exercise 1.2.

2.2 Simulation Study

Construct a simulation study to examine the performance of the algorithm for sampling the correlation ρ given in Example 2.3. More specifically:

(i) generate a data set $D = \{\boldsymbol{y}_i = (y_{1i}, y_{2i})',\ i = 1, 2, \ldots, n\}$ from a bivariate normal distribution $N_2(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ with different values of n and ρ;

(ii) implement a Metropolis–Hastings algorithm to obtain a Markov chain of ρ; and

(iii) study the trace plot and autocorrelation of the chain as well as the acceptance probability of the algorithm.

Some other diagnostic tools discussed in Section 2.9 may also be applied here.

2.3 Repeat Problem 2.1 using the Hit-and-Run algorithm described in Section 2.3. Compare the performance of the Gibbs sampler and the H&R algorithm.

2.4 Prove (2.5.2).

2.5 Bayesian Analysis for Sensory Data

(a) Construct a Bayesian probit model using an improper prior for the model parameters to analyze the MRE sensory data given in Table 1.2 with storage temperature, time, and their interaction as possible covariates.

(b) Derive the posterior distribution.

(c) Implement the Albert–Chib, Cowles, Nandram–Chen, and Chen–Dey algorithms described in Section 2.5.3, and compare their performance.

(d) Compute the posterior estimates of the cutpoints and the regression coefficients.

(e) Are the Bayesian results comparable to those obtained from Exercise 1.3?

2.6 Prove that the posterior $\pi(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\epsilon}|D)$ given in (2.5.26) is proper if $\delta_0 > 0$, $\gamma_0 > 0$, and X^* is of full rank, where X^* is a matrix induced by X and $\boldsymbol{y}$ with its t^{th} row equal to $1\{y_t > 0\}\boldsymbol{x}_t'$.

2.7 Show that the conditional posterior density $\pi(\boldsymbol{\eta}|\boldsymbol{\beta}, \sigma^2, \rho, D)$ given in (2.5.29) for the reparameterized random effects is log-concave in each component of $\boldsymbol{\eta}$.

2.8 Perform a fully Bayesian analysis for the 1994 pollen count data in Example 1.3.

(i) Implement the Gibbs sampler with hierarchical centering described in Section 2.5.4 for sampling from the reparameterized posterior $\pi(\boldsymbol{\beta}, \sigma^2, \rho, \boldsymbol{\eta}|D)$ given in (2.5.28). You may choose $\delta_0 = 0.01$ and $\gamma_0 = 0.01$.
(ii) Obtain the posterior estimates for $\boldsymbol{\beta}$, σ^2, and ρ.
(iii) Compare the Bayesian estimates with those obtained from Exercise 1.5 using the GEE approach.

2.9 A Counterexample (Liu and Sabatti 1998)
If the transition function T_θ does not satisfy (2.6.2), the target distribution π may not be preserved. Let θ take values in $\{0, 1, 2, 3, 4\}$ and suppose that the target distribution is uniform, i.e., $\pi(\theta|D) = \frac{1}{5}$. The group operation is the translation: $i * j = i + j \pmod 5$. The transition functions T_θ are, respectively, $T_0(i, j) = \frac{1}{3}$ for $|i - j| \leq 1$, $i = 1, 2, 3, 4$, $T_0(0, j) = \frac{1}{3}$ for $j = 0, 1, 4$, and $T_0(4, j) = \frac{1}{3}$ for $j = 3, 4, 0$; and $T_k(i, j) = \frac{1}{5}$ for all i, j and $k > 0$.

(a) Show that π is invariant under all T_θ with θ fixed.
(b) Show that the invariant distribution of $T_i(i, j)$ is proportional to $(3, 3, 2, 2, 3)$ instead of π.

2.10 Linear Regression Models with Censored Data
Consider the experiment of improving the lifetime of fluorescent lights (Hamada and Wu 1995). Carried out by a 2^{5-2} fractional factorial design, the experiment was conducted over a time period of 20 days, with inspection every two days. The design and the lifetime data are tabulated in Table 2.2.
Let $\boldsymbol{x}_i$ be the $p \times 1$ column vector of the factor levels, including the intercepts, A, B, C, D, E, AB, and BD, and y_i be the logarithm of the corresponding lifetime for $i = 1, 2, \ldots, n$, where $p = 8$ and $n = 2 \times 2^{5-2} = 16$. Also let $(Y_i^{(l)}, Y_i^{(r)})$ denote the observed censoring interval for y_i, i.e., $y_i \in (Y_i^{(l)}, Y_i^{(r)})$. Hamada and Wu (1995) consider the following model:

$$y_i \overset{\text{i.i.d.}}{\sim} N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2), \quad i = 1, 2, \ldots, n,$$

TABLE 2.2. Design and Lifetime Data for Light Experiment.

Run	Design A	B	C	D	E	Data (no. of days)	
1	+	+	+	+	+	(14, 16)	(20, ∞)
2	+	+	−	−	−	(18, 20)	(20, ∞)
3	+	−	+	+	−	(08, 10)	(10, 12)
4	+	−	−	−	+	(18, 20)	(20, ∞)
5	−	+	+	−	+	(20, ∞)	(20, ∞)
6	−	+	−	+	−	(12, 14)	(20, ∞)
7	−	−	+	−	−	(16, 18)	(20, ∞)
8	−	−	−	−	−	(12, 14)	(14, 16)

Source: Hamada and Wu (1995).

with the prior distribution specified by

$$\sigma^2 \sim \mathcal{IG}(\nu_0, \nu_0 s_0/2) \text{ and } \beta|\sigma^2 \sim N(\beta_0, \sigma^2 I_p/\tau_0)$$

for $(\boldsymbol{\beta}, \sigma^2)$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_7)'$ is the vector of regression coefficients with β_0 corresponding to the intercept, $\mathcal{IG}$ denotes the inverse gamma distribution, i.e., $\pi(\sigma^2) \propto (\sigma^2)^{-(\nu_0/2+1)} \exp\{-\nu_0 s_0/(2\sigma^2)\}$, $\nu_0 = 1$, $s_0 = 0.01$, $\boldsymbol{\beta}_0 = (3, 0, \ldots, 0)$, I_p is the $p \times p$ identity matrix, and $\tau_0 = 0.0001$.

(a) Write the posterior distribution for $(\boldsymbol{\beta}, \sigma^2)$ based on the observed data $D_{\text{obs}} = (\{(Y_i^{(l)}, Y_i^{(r)}), \boldsymbol{x}_i\},\ i = 1, 2, \ldots, n)$.
(b) Derive an expression for the posterior distribution based on the complete-observed data $D = ((y_i, \boldsymbol{x}_i),\ i = 1, 2, \ldots, n)$.
(c) Develop an efficient MCMC algorithm for sampling from the posterior distribution.
(*Hint:* Slow convergence of the original Gibbs sampler may occur; so an improved MCMC algorithm such as the GM-MGMC or CA-MCMC algorithm may be required.)
(d) Perform a fully Bayesian analysis for the lifetime data.

2.11 To sample from the posterior distribution in (2.6.9) for the rating data given in Table 2.1, consider the GM-MGMC algorithm with the following additive group transformations:

$$g(\boldsymbol{\beta}, \gamma_2, \boldsymbol{z}) = (\beta_0, \beta_1 + g, \gamma_2, \{z_i : x_i = 0\}, \{z_i + g : x_i = 1\}).$$

(a) Write the GM step.
(b) Study the performance of this version of the GM-MGMC algorithm.

2.12 Prove (2.7.4).

2.13 (i) Explain why the value of the normalizing constant $c(D)$ is not required in the RDIP sampler.

(ii) Derive (2.8.7), (2.8.8), and (2.8.9) for the conditional density and the jump ratio for the state-dependent and direction-and-radius RDIP sampler.

2.14 In Exercise 2.5, compute Gelman and Rubin's PSR factors for all four algorithms with five ($m = 5$) independent sequences of length $2n$ for $n = 500$ and $n = 1000$, and discuss which algorithm converges faster.

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

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

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*

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

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

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

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

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

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

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

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 $\boldsymbol{\lambda}(1) = \boldsymbol{\lambda}_2$:

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

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

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

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 $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n\}$ 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* $a > 0$,

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

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

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

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*

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

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*

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

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.,

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

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*

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

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

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

Then

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

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*

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

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,

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

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}	$2\,(2\Phi(\delta/2) - 1)$

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

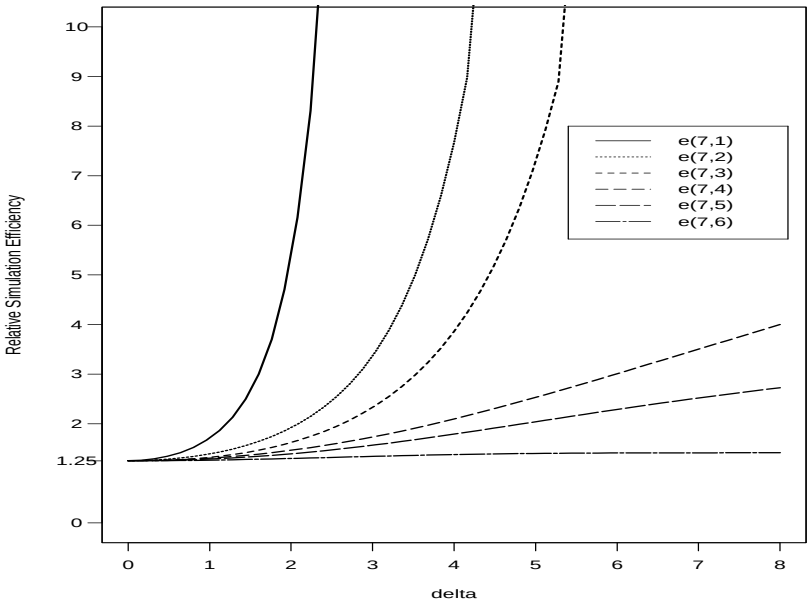

FIGURE 5.1. Relative simulation efficiency plot (I).

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:

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

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 $\alpha=(q_1q_2)^{-1/2}$	$\sqrt{2}(\Delta-1)/\sqrt{\Delta}$
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

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

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)

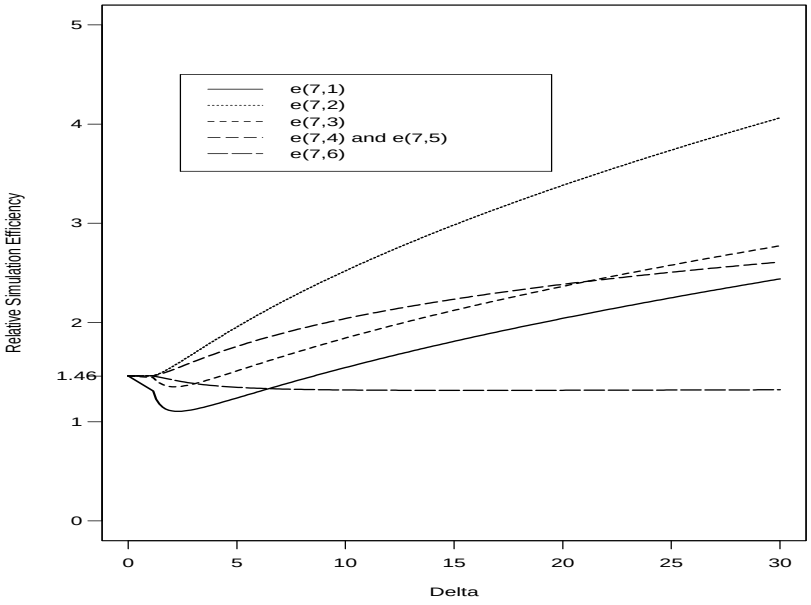

FIGURE 5.2. Relative simulation efficiency plot (II).

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 $\{\boldsymbol{\theta}_{2,1}, \boldsymbol{\theta}_{2,2}, \ldots, \boldsymbol{\theta}_{2,n}\}$, 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 $\{\boldsymbol{\theta}_{2,1}, \boldsymbol{\theta}_{2,2}, \ldots, \boldsymbol{\theta}_{2,n_2}\}$ 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 $\{\boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{1,2}, \ldots, \boldsymbol{\theta}_{1,n_1}\}$ 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

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

where $\hat{r}_{\text{RIS}}$ 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 $\hat{r}_{\text{RIS}}$ as follows:

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

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

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

and

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

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*

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

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

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

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 $\{\psi_1^*, \ldots, \psi_M^*\}$ that includes all sample points ψ_1, $\ldots$, 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:

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

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 $\hat{r}_{\mathrm{GOBS}}$ is given by

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

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 $\hat{r}_{\mathrm{OIS}}$, $\hat{r}_{\mathrm{GOBS}}$, and $\hat{r}_{\mathrm{GORIS}}$, 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 $\mathrm{RE}(\hat{r})$; 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

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

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

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

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

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$(\hat{r}(a_{\text{0pt}}))$
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 $\hat{r}(a_{\text{0pt}})$ 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

$$\text{Var}(\hat{r}(a_{\text{0pt}})) = \frac{1}{n}[\exp(\delta^2) 4\Phi(\delta)(1 - \Phi(\delta)) - 1],$$

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$(\hat{r}(a_{\text{0pt}}))$ 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 $\hat{r}(a_{\text{0pt}})$ dramatically improves the simulation efficiency compared to the importance sampling estimator $\hat{r}_{\text{IS}_2}$. For example, when $\delta = 3$ and $\kappa = 20$,

$$\text{Var}(\hat{r}_{\text{IS}_2})/\text{Var}(\hat{r}(a_{\text{0pt}})) = 117,421.51,$$

i.e., $\hat{r}(a_{\text{0pt}})$ 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:

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

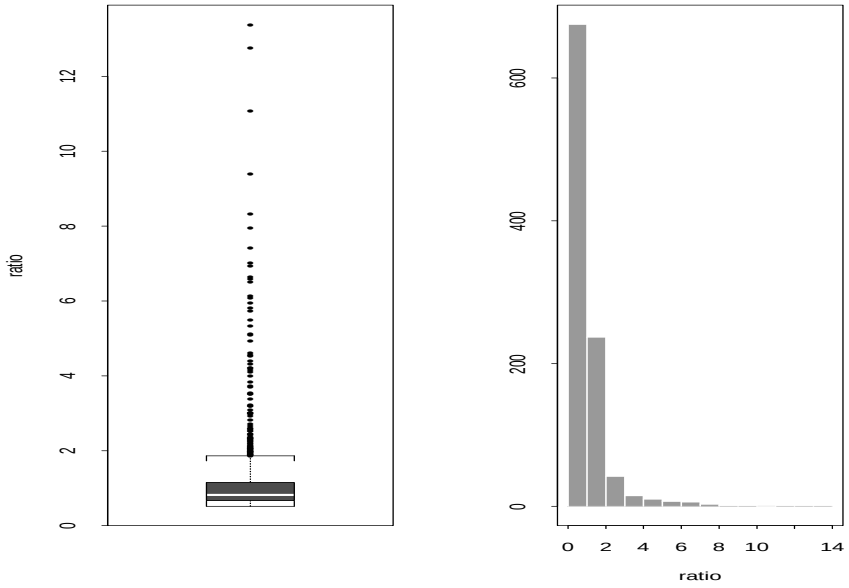

FIGURE 5.3. The box plot and histogram of the ratio $h(\boldsymbol{\theta}_i)$.

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

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

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

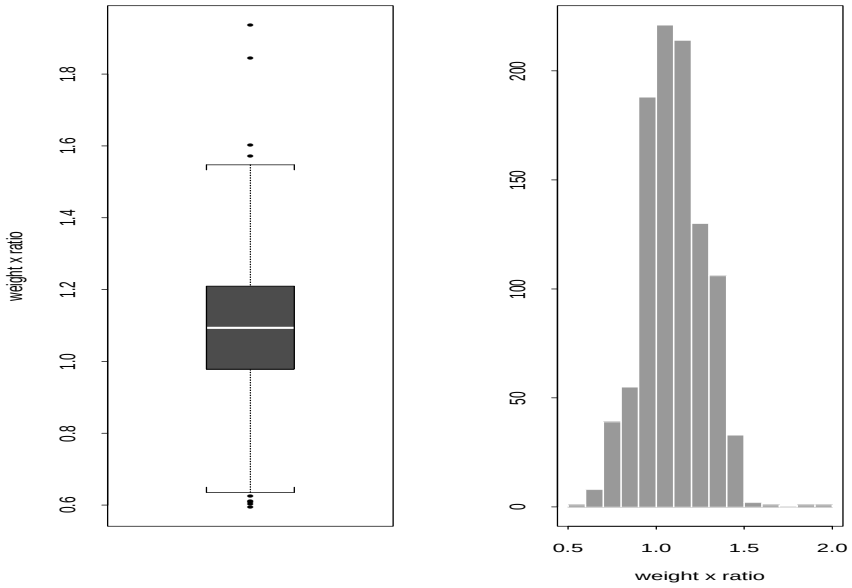

FIGURE 5.4. The box plot and histogram of the weight-times-ratio.

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

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

Thus,

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

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

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

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 $A \geq 2$,

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

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

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

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

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

Putting together the above inequalities yields

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

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

Next, we prove (5.5.6). Observe that

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

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

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

It is easy to see that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof of Theorem 5.5.6. Let

$$\zeta(x,t) = \frac{q_1(x)}{\psi q_1(x) + (1-\psi)tq_2(x)}.$$

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

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

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

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

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

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

and

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

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,

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

where

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

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). □

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 $\alpha(\boldsymbol{\theta}, \psi)$,

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

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 $g''(x) = -2 s_1 s_2/(s_1 x + s_2)^3 < 0$ 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. □

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

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

[*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, $\text{Var}(\hat{r}_{\text{IS}_2})$, of $\hat{r}_{\text{IS}_2}$ 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 $\hat{r}_{\text{BS}}$ given in (5.3.3) is called a geometric bridge sampling (GBS) estimator of r. Show that $\text{RE}^2(\hat{r}_{\text{BS}})$ given in (5.3.4) reduces to

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

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