Statistica Sinica **6**(1996), 831-860

SIMULATING RATIOS OF NORMALIZING CONSTANTS VIA A SIMPLE IDENTITY: A THEORETICAL EXPLORATION

Xiao-Li Meng and Wing Hung Wong

The University of Chicago and The Chinese University of Hong Kong

Abstract: Let $p_i(w), i = 1, 2$, be two densities with common support where each density is known up to a normalizing constant: $p_i(w) = q_i(w)/c_i$. We have draws from each density (e.g., via Markov chain Monte Carlo), and we want to use these draws to simulate the ratio of the normalizing constants, c_1/c_2. Such a computational problem is often encountered in likelihood and Bayesian inference, and arises in fields such as physics and genetics. Many methods proposed in statistical and other literature (e.g., computational physics) for dealing with this problem are based on various special cases of the following simple identity:

$$\frac{c_1}{c_2} = \frac{E_2[q_1(w)\alpha(w)]}{E_1[q_2(w)\alpha(w)]}.$$

Here E_i denotes the expectation with respect to p_i $(i = 1, 2)$, and α is an arbitrary function such that the denominator is non-zero. A main purpose of this paper is to provide a theoretical study of the usefulness of this identity, with focus on (asymptotically) optimal and practical choices of α. Using a simple but informative example, we demonstrate that with sensible (not necessarily optimal) choices of α, we can reduce the simulation error by orders of magnitude when compared to the conventional importance sampling method, which corresponds to $\alpha = 1/q_2$. We also introduce several generalizations of this identity for handling more complicated settings (e.g., estimating several ratios simultaneously) and pose several open problems that appear to have practical as well as theoretical value. Furthermore, we discuss related theoretical and empirical work.

Key words and phrases: Bridge sampling, Bayes factor, Hellinger distance, importance sampling, iterative simulation, likelihood ratio, free-energy difference, posterior odds, Markov chain Monte Carlo.

1. Motivation and Applications

A computational problem arising frequently in statistical and other analyses is the computation of normalizing constants for probability densities from which we have random draws. More generally, finding definite integrals of positive functions can be formulated as a problem of computing normalizing constants. Typically, we are interested in the ratios of such normalizing constants or generally the relative values of normalizing constants with respect to a reference value

(which can be chosen to be known). Mathematically, this problem can be formulated as follows. Let $p_i(w), i = 1, 2$, be two densities (with respect to a common measure, which will be implicit hereafter), from which we have (dependent or independent) draws. We know each density up to a *normalizing constant*,

$$p_i(w) = \frac{q_i(w)}{c_i}, \qquad w \in \Omega_i \subset R^d, \tag{1.1}$$

where Ω_i is the support of $p_i(w)$, and the unnormalized density $q_i(w)$ can be evaluated at any $w \in \Omega_i$, $i = 1, 2$. We are interested in calculating the ratio of the two normalizing constants: $r = c_1/c_2$.

As a direct application, consider the problem of computing likelihood ratios for hypothesis testing. Let w be the data and i be the index of the likelihood at parameter value $\theta = \theta_i$, that is, $p_i = p(w|\theta_i)$, $q_i(w) = q(w|\theta_i)$, $c_i = c(\theta_i)$, $i = 1, 2$. Then

$$\frac{L(\theta_2|w)}{L(\theta_1|w)} = \frac{p(w|\theta_2)}{p(w|\theta_1)} = \frac{q(w|\theta_2)}{q(w|\theta_1)} \times \frac{c(\theta_1)}{c(\theta_2)}.$$

Often, for a given observation w and a parameter θ, the density $p(w|\theta)$ is easy to evaluate up to a multiplicative constant, i.e., $q(w|\theta)$ is known. The calculation of the likelihood ratio then reduces to the calculation of the ratio of the normalizing constants.

Another use of this formulation in likelihood inference occurs in the computation of likelihood ratios in the presence of missing (or latent) data. Specifically, let $Y = (Y_{obs}, Y_{mis})$ be the complete data consisting of the observed part, Y_{obs}, and the missing (latent) part, Y_{mis}. Then

$$p(Y_{mis}|Y_{obs}, \theta) = \frac{p(Y|\theta)}{p(Y_{obs}|\theta)} = \frac{L(\theta|Y)}{L(\theta|Y_{obs})}. \tag{1.2}$$

In other words, the observed-data likelihood can be viewed as a normalizing constant of the conditional density: $p(Y_{mis}|Y_{obs}, \theta)$, with the complete-data likelihood, $L(\theta|Y)$ being the unnormalized density. This formulation has important applications, for instance, in genetic linkage analysis where direct calculation of the likelihood of θ (e.g., locations of disease genes relative to a set of known markers) based on the observed data (e.g., genotypes of each individual marker for some members of a pedigree) is typically prohibitive for a large pedigree with many loci. On the other hand, given the full information such as the allele types each person in the pedigree inherited from his/her parents, the computation of the complete-data likelihood is straightforward, and simulating Y_{mis} from $p(Y_{mis}|Y_{obs}, \theta)$ is feasible (e.g., Irwin, Cox and Kong (1994)). Another application of (1.2) is in the context of monitoring the convergence of the Monte Carlo

EM algorithm (e.g., Wei and Tanner (1990)), as detailed in Meng and Schilling (1996).

In Bayesian inference, a variety of evaluations of density ratios can be formulated as computations of ratios of normalizing constants. For example, the marginal density of the data, $p(Y)$, is a normalizing constant of the posterior density:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}, \tag{1.3}$$

where $p(\theta)$ is a prior density of θ. Calculating the ratio of $p(Y^{(i_1)})$ and $p(Y^{(i_2)})$, for instance, arises in congenial Bayesian inference with multiply-imputed data sets (Meng (1994)), where $Y^{(i_1)}$ and $Y^{(i_2)}$ represent two imputed data sets, and random draws from $p(\theta|Y^{(i)})$ are available as a by-product of complete-data Bayesian analysis performed on each completed-data set created by multiple imputation. The ratios are needed for calculating importance weights for applying the extended multiple-imputation combining rules discussed in Meng (1994), Sec. 5.

A slightly more complicated application arises in computing the ratio of marginal posterior densities of a parameter λ, which is a component of the model parameter $\theta = (\lambda, \phi)$. In other words, we want to compute the posterior odds

$$\frac{p(\lambda_1|Y)}{p(\lambda_2|Y)} \equiv \frac{p(Y|\lambda_1)}{p(Y|\lambda_2)} \times \frac{p(\lambda_1)}{p(\lambda_2)}.$$

Assuming the prior odds $p(\lambda_1)/p(\lambda_2)$ are known, the problem reduces to computing the Bayes factor $p(Y|\lambda_1)/p(Y|\lambda_2)$. The direct computation, however, is often difficult because of the integration:

$$p(Y|\lambda) = \int p(Y|\lambda, \phi)p(\phi|\lambda)d\phi.$$

We notice, however,

$$p(\phi|Y, \lambda) = \frac{p(\phi, Y|\lambda)}{p(Y|\lambda)} = \frac{p(Y|\lambda, \phi)p(\phi|\lambda)}{p(Y|\lambda)},$$

and thus $p(Y|\lambda_i)$ can be viewed as c_i of (1.1), with $p_i = p(\phi|Y, \lambda_i)$ and $q_i = p(Y|\lambda_i, \phi)p(\phi|\lambda_i)$, $i = 1, 2$. Simulation from $p(\phi|Y, \lambda)$ can often be facilitated by the Gibbs sampler (Geman and Geman (1984)), or more generally by iterative simulations (e.g, Gelfand and Smith (1990); Gelman and Rubin (1992)). For more details on Bayes factors, see the recent review article by Kass and Raftery (1995).

Finally, the problem of estimating a ratio of normalizing constants has been of great interest in computational physics, where the problem is known as estimating free energy differences (e.g., Bennett (1976), Torrie and Valleau (1977),

Voter (1985)). Since these papers are most related to our presentation, we will discuss them in more detail in Section 9, together with discussion of other related work. In other literature, importance sampling with simple identities, such as

$$\frac{c_1}{c_2} = E_2\Big[\frac{q_1(w)}{q_2(w)}\Big], \qquad (\text{when } \Omega_1 \subseteq \Omega_2), \tag{1.4}$$

where E_i denotes the expectation with respect to p_i $(i = 1, 2)$, has played a key role in simulating c_1/c_2 (e.g., Ott (1979), Geyer and Thompson (1992), Green (1992)). In particular, in deriving the method of "reweighting mixture", Geyer (1994) proposed the method of "reverse logistic regression" for computing several normalizing constants simultaneously, a method that can be derived by iteratively choosing q_2 in (1.4) as a mixture density, as we will detail in Section 7.

This paper provides a theoretical study of simulating c_1/c_2 via generalizations of (1.4) that permit efficient use of random draws from more than one density; drawing from several densities is a task that is typically only slightly more complicated than simulating from one of them when these densities are from the same parametric family. These generalizations (presented in Sections 2, 7, and 8) are most useful with the currently popular iterative simulation using Markov chains (e.g., Tanner and Wong (1987), Gelfand and Smith (1990), Gelman and Rubin (1992), Geyer and Thompson (1992), Smith and Roberts (1993), Besag and Green (1993)), as evidenced in some of their very successful applications in computational physics (e.g., Bennett (1976)). For theoretical tractability, we will first (in Section 3) assume independence among draws when deriving results regarding Monte Carlo errors. We then extend (in Section 6) our exploration, via the notion of "effective sample sizes", to more general settings involving dependent draws. Empirical studies, as reported in DiCiccio, Kass, Raftery and Wasserman (1996) and in Meng and Schilling (1996) (see Section 9), suggest that the optimal or near optimal procedures constructed under the independence assumption (see Sections 4 and 5) can work remarkably well in general, providing orders of magnitude improvement over other methods with similar computational efforts. Nevertheless, we hope our exploration under general settings will stimulate further research, which may provide additional reduction of Monte Carlo errors in situations where the dependence among draws is strong.

2. A Simple Identity

Following the notation of (1.1), let $\alpha(w)$ be an arbitrary function defined on $\Omega_1 \cap \Omega_2$, the common support of p_1 and p_2, such that

$$0 < \Big| \int_{\Omega_1 \cap \Omega_2} \alpha(w) p_1(w) p_2(w) dw \Big| < \infty. \tag{2.1}$$

The existence of such an α (the values of α outside $\Omega_1 \cap \Omega_2$ are irrelevant) is guaranteed if and only if

$$\int_{\Omega_1 \cap \Omega_2} p_1(w)p_2(w)dw > 0, \tag{2.2}$$

implying that the common support of p_1 and p_2 is non-trivial. The quantity in (2.2) is a measure of the "overlap" between p_1 and p_2, and in Section 9 we will discuss a method that can handle the "no-overlap" cases (i.e., when (2.2) does not hold).

Given any α satisfying (2.1), we have

$$\frac{\int_{\Omega_2} q_1(w)\alpha(w)p_2(w)dw}{\int_{\Omega_1} q_2(w)\alpha(w)p_1(w)dw} = \frac{c_1}{c_2} \times \frac{\int_{\Omega_1 \cap \Omega_2} \alpha(w)p_1(w)p_2(w)dw}{\int_{\Omega_1 \cap \Omega_2} \alpha(w)p_1(w)p_2(w)dw},$$

which yields the key identity

$$\frac{c_1}{c_2} = \frac{E_2[q_1(w)\alpha(w)]}{E_1[q_2(w)\alpha(w)]}. \tag{2.3}$$

This identity unifies many identities used in the literature for simulating normalizing constants or other similar computation. The most general one of them, to our best knowledge, was given by Bennett (1976), who proposed (2.3) in the context of simulating free-energy differences with $q_i = \exp(-U_i)$, where U_i is the temperature-scaled potential energy and $i = 1, 2$ indexes two canonical ensembles on the same configuration space. Taking $\alpha(w) = q_2^{-1}(w)$ leads to (1.4), assuming $\Omega_1 \subseteq \Omega_2$. When $\Omega_1 = \Omega_2$ and Ω_1 has a finite Lebesgue measure, taking $\alpha(w) = [q_1(w)q_2(w)]^{-1}$ leads to a generalization of the "harmonic rule" given in Newton and Raftery (1994) (also see Gelfand and Dey (1994)),

$$\frac{c_1}{c_2} = \frac{E_2[q_2^{-1}(w)]}{E_1[q_1^{-1}(w)]}. \tag{2.4}$$

Before discussing in detail the choices of α, we first define the Monte Carlo estimator of c_1/c_2 based on (2.3). Given random draws $w_{i1}, \ldots, w_{in_i}$ from $p_i(w), i = 1, 2$, and a choice of α, the corresponding estimator for $r = c_1/c_2$ is

$$\hat{r}_\alpha = \frac{n_2^{-1} \sum_{j=1}^{n_2} q_1(w_{2j})\alpha(w_{2j})}{n_1^{-1} \sum_{j=1}^{n_1} q_2(w_{1j})\alpha(w_{1j})}. \tag{2.5}$$

For any α satisfying (2.1), $\hat{r}_\alpha$ consistently estimates r as long as the sample averages in (2.5) converge to their corresponding population averages, a requirement

that is met by both independent Monte Carlo and Markov chain Monte Carlo simulations under standard regularity conditions (e.g., ergodicity).

3. Asymptotically Optimal Choice of α

Since the estimator $\hat{r}_\alpha$ depends on α, a natural question of interest is the optimal choice of α. A standard measure of accuracy in such a setting is the relative mean-square error:

$$RE^2(\hat{r}_\alpha) \equiv \frac{E(\hat{r}_\alpha - r)^2}{r^2}, \tag{3.1}$$

where the expectation is taken over all random draws. The exact calculation of (3.1) depends on how the simulation is conducted, and typically is intractable because $\hat{r}_\alpha$ is a ratio estimator. With large numbers of draws from each density, however, we can approximate (3.1) by its first-order term, which essentially ignores the (negligible) bias term; we will use " $\doteq$ " to denote a first-order equality. Under the assumption that $\{w_{i1}, \ldots, w_{in_i}\}$ are identical and independent draws from $p_i(w), i = 1, 2$, and that the two sets of draws are independent, we have (see Appendix for proof)

$$\begin{aligned} RE^2(\hat{r}_\alpha) &\doteq \frac{1}{ns_1s_2}\Big\{\frac{\int_{\Omega_1\cap\Omega_2} p_1p_2(s_1p_1 + s_2p_2)\alpha^2 dw}{(\int_{\Omega_1\cap\Omega_2} p_1p_2\alpha dw)^2} - 1\Big\} \\ &= \frac{1}{n}\frac{\int_{\Omega_1\cap\Omega_2} \tilde{p}_1\tilde{p}_2(\tilde{p}_1 + \tilde{p}_2)\alpha^2 dw}{\big(\int_{\Omega_1\cap\Omega_2} \tilde{p}_1\tilde{p}_2\alpha dw\big)^2} - \frac{1}{n_1} - \frac{1}{n_2}, \end{aligned} \tag{3.2}$$

where $n = n_1 + n_2, s_i = n_i/n$, $\tilde{p}_i = s_ip_i$, and $s_i (i = 1, 2)$ are assumed to be asymptotically between 0 and 1. Bennett (1976) gave the same expression as the asymptotic mean-squared error of $\log(\hat{r}_\alpha)$, which is asymptotically the same as the relative error in (3.1); he also correspondingly gave the following result without proof.

Theorem 1. *The right side of* (3.2)*, as a functional of* α*, is minimized at*

$$\alpha_O(w) \propto \frac{1}{s_1p_1(w) + s_2p_2(w)} \equiv \frac{1}{\tilde{p}_1(w) + \tilde{p}_2(w)}, \qquad w \in \Omega_1 \cap \Omega_2, \tag{3.3}$$

with the minimum value

$$\frac{1}{n}\Big[\int_{\Omega_1\cap\Omega_2} (\tilde{p}_1^{-1} + \tilde{p}_2^{-1})^{-1} dw\Big]^{-1} - \frac{1}{n_1} - \frac{1}{n_2}. \tag{3.4}$$

Proof. By the Cauchy-Schwartz inequality

$$\Big(\int_{\Omega_1\cap\Omega_2} \tilde{p}_1\tilde{p}_2\alpha dw\Big)^2 \le \Big\{\int_{\Omega_1\cap\Omega_2} \Big[\sqrt{\frac{\tilde{p}_1\tilde{p}_2}{\tilde{p}_1 + \tilde{p}_2}}\,\Big]\Big[\sqrt{\tilde{p}_1\tilde{p}_2(\tilde{p}_1 + \tilde{p}_2)}\,|\alpha|\Big] dw\Big\}^2$$

$$\leq \Big\{\int_{\Omega_1\cap\Omega_2} \frac{\tilde{p}_1\tilde{p}_2}{\tilde{p}_1+\tilde{p}_2}dw\Big\}\Big\{\int_{\Omega_1\cap\Omega_2} \tilde{p}_1\tilde{p}_2(\tilde{p}_1+\tilde{p}_2)\alpha^2 dw\Big\}.$$

Thus,

$$\frac{\int_{\Omega_1\cap\Omega_2} \tilde{p}_1\tilde{p}_2(\tilde{p}_1+\tilde{p}_2)\alpha^2 dw}{(\int_{\Omega_1\cap\Omega_2} \tilde{p}_1\tilde{p}_2\alpha dw)^2} \geq \Big[\int_{\Omega_1\cap\Omega_2} \frac{\tilde{p}_1\tilde{p}_2}{\tilde{p}_1+\tilde{p}_2}dw\Big]^{-1},$$

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

$$\sqrt{\tilde{p}_1\tilde{p}_2(\tilde{p}_1+\tilde{p}_2)}\alpha \propto \sqrt{\tilde{p}_1\tilde{p}_2/(\tilde{p}_1+\tilde{p}_2)}\ ,$$

which yields (3.3).

This (asymptotically) optimal choice is intuitively appealing. It represents (the inverse of) the mixture of p_1 and p_2 with mixture proportions determined by the sampling rates of the two distributions. It is, however, not of direct use because α_O depends on the unknown ratio $r = c_1/c_2$, since

$$\alpha_O \propto \frac{1}{s_1q_1 + s_2rq_2}. \tag{3.5}$$

Furthermore, it depends on the ratio of the two sample sizes, because $\alpha_O \propto 1/(q_1+(n_2/n_1)rq_2)$. When the draws are independent, we know the exact sample sizes, n_1 and n_2. With dependent draws, n_1 and n_2 are no longer the true sample sizes, since the dependence among draws typically reduces the "effective sample sizes" and thus using n_1/n_2 may lead to large simulation errors. We will discuss this issue in Section 5 and Section 6 after we deal with the issue of unknown r in (3.5).

4. Iterative Choice of α

The expression (3.5) immediately suggests an iterative method of choosing α. Starting with an initial guess of $r, \hat{r}_O^{(0)} > 0$, we calculate our estimate of r iteratively by using the optimal α based on the previous estimate of r. Specifically, at the $(t+1)st$ iteration, we compute

$$\hat{r}_O^{(t+1)} = \frac{\frac{1}{n_2}\sum_{j=1}^{n_2}\Big[\frac{q_1(w_{2j})}{s_1q_1(w_{2j})+s_2\hat{r}_O^{(t)}q_2(w_{2j})}\Big]}{\frac{1}{n_1}\sum_{j=1}^{n_1}\Big[\frac{q_2(w_{1j})}{s_1q_1(w_{1j})+s_2\hat{r}_O^{(t)}q_2(w_{1j})}\Big]} \equiv \frac{\frac{1}{n_2}\sum_{j=1}^{n_2}\Big[\frac{l_{2j}}{s_1l_{2j}+s_2\hat{r}_O^{(t)}}\Big]}{\frac{1}{n_1}\sum_{j=1}^{n_1}\Big[\frac{1}{s_1l_{1j}+s_2\hat{r}_O^{(t)}}\Big]}, \tag{4.1}$$

where $l_{ij} = q_1(w_{ij})/q_2(w_{ij})$ $(j = 1,\ldots,n_i, i = 1,2)$ need only be computed once at the beginning of the algorithm. (For simplicity, we assume $q_i(w_{ij}) > 0$ for all i and j.)

Since for any $\hat{r}_O^{(0)} > 0, \hat{r}_O^{(1)}$ is a consistent estimator of r, by induction it is easy to show that every iterate $\hat{r}_O^{(t)}$ is a consistent estimator of r as $n \to \infty$. Furthermore, we have the following result, the proof of which is given in the Appendix.

Theorem 2. *For any given set* $\{l_{ij} > 0 : 1 \le j \le n_i, i = 1, 2\}$, *the iterative sequence* $\{\hat{r}_O^{(t)}, t \ge 0\}$ *defined by* (4.1) *converges to a unique limit,* $\hat{r}_O$, *with the property*

$$|\hat{r}_O^{(t+1)} - \hat{r}_O| < |\hat{r}_O^{(t)} - \hat{r}_O|, \quad \text{if } \hat{r}^{(t)} \ne \hat{r}_O \quad (t \ge 0).$$

Furthermore, when all draws are independent,

$$RE^2(\hat{r}_O) \doteq \frac{1}{n}\Big[\int_{\Omega_1 \cap \Omega_2} (\tilde{p}_1^{-1} + \tilde{p}_2^{-1})^{-1} dw\Big]^{-1} - \frac{1}{n_1} - \frac{1}{n_2}.$$

In other words, $\hat{r}_O$ achieves the same asymptotic minimum relative mean-square error (hereafter, minimum error) as does $\hat{r}_{\alpha_O}$, the estimator based on the (infeasible) optimal choice α_O of (3.5). Bennett (1976) suggested adjusting a pair of equations until a "self-consistent" solution is reached to deal with the problem that the optimal α depends on r, a method that can be viewed as an implicit construction of the iterative scheme of (4.1), although Bennett (1976) did not discuss the issue of convergence. We note that the first-order approximation in Theorem 2 for $RE^2(\hat{r}_O)$ becomes exact in the trivial case $p_1 = p_2$, for which case $r_O^{(t)} \equiv r$, and thus $RE^2(\hat{r}_O) = 0$. We also note that, upon convergence, the numerator of (4.1) provides a consistent estimator of $\int_{\Omega_1 \cap \Omega_2} (p_1 p_2)/(\tilde{p}_1 + \tilde{p}_2)\, dw$, and thus a consistent estimator of $RE^2(\hat{r}_O)$ is obtained without additional computation. When the draws are not independent, the estimation of $RE^2(\hat{r}_O)$ is generally quite complicated (see Geyer (1994)). But if one is implementing multiple sequences, as Gelman and Rubin (1992) advocate, then constructing variance estimates is an easy task with the replications.

It is informative to compare (4.1) with a similar iterative scheme based on the well-known scheme of importance sampling using a mixture (e.g., Geyer (1994)). If we treat the pooled sample $\{w_1, \ldots, w_n\} = \{w_{ij}, j = 1, \ldots, n_i, i = 1, 2\}$ as a sample of independent and identical draws from the mixture $s_1 p_1 + s_2 p_2$, we can construct, by analogy to (4.1), the following iterative scheme

$$\hat{r}_M^{(t+1)} = \frac{\frac{1}{n}\sum_{j=1}^{n} \Big[\frac{q_1(w_j)}{s_1 q_1(w_j) + s_2 \hat{r}_M^{(t)} q_2(w_j)}\Big]}{\frac{1}{n}\sum_{j=1}^{n} \Big[\frac{q_2(w_j)}{s_1 q_1(w_j) + s_2 \hat{r}_M^{(t)} q_2(w_j)}\Big]}, \qquad t = 0, 1, \ldots \tag{4.2}$$

The rationale behind (4.2) is clear: if $\hat{r}_M^{(t)} = r$, then the numerator and the denominator would be consistent estimates of 1 and c_2/c_1 respectively, and thus

the ratio would estimate c_1/c_2. Iteration is needed to update the approximation to the mixture density and thus to ensure consistency in the limit. Since the numerator of (4.2) converges to 1, a third iterative scheme is to replace the numerator of (4.2) by 1, which yields a sequence that converges to the "reverse logistic regression" estimator described in Geyer (1994) for the two-density cases (for more general cases, see Section 7).

When $s_i > 0, i = 1, 2$, the first part of Theorem 2 also applies to $\{\hat{r}_M^{(t)}, t \geq 0\}$, that is, it converges to a unique limit, $\hat{r}_M$, and $|\hat{r}_M^{(t+1)} - \hat{r}_M| < |\hat{r}_M^{(t)} - \hat{r}_M|$ if $\hat{r}_M^{(t)} \neq \hat{r}_M \ \ (t \geq 0)$. Furthermore, it is easy to show (see Appendix) that $\hat{r}_M = \hat{r}_O$, that is, the two iterative schemes yield the same limit. The same conclusions also apply to the third iterative scheme discussed earlier. The fundamental difference between scheme (4.1) and scheme (4.2) (or its modification, the third scheme), however, is that the former provides a consistent estimator at each iteration, whereas the latter does so only in the limit. It is thus reasonable to expect that (4.1) converges more rapidly than the other iterations. This has been confirmed in an empirical study by Meng and Schilling (1996) (see Section 9 for more detail).

When $s_1 = 0$, that is, when all samples are drawn from p_2, (4.1) reduces to the non-iterative importance-sampling estimator based on (1.4), that is,

$$\hat{r}_S = \frac{1}{n} \sum_{j=1}^{n} q_1(w_j)/q_2(w_j), \tag{4.3}$$

which has an exact relative mean-square error given by (under the assumption of independent draws)

$$RE^2(\hat{r}_S) = \frac{1}{n} \int_{\Omega_2} \frac{(p_1 - p_2)^2}{p_2} \, dw. \tag{4.4}$$

Similarly, when $s_2 = 0$, the inverse of the right hand side of (4.1) provides an unbiased estimate of c_2/c_1, with (4.3) and (4.4) adjusted accordingly by switching the subscripts 1 and 2. It is well-known that the right hand side of (4.4) can be infinite because p_1/p_2 may not be square integrable with respect to p_2. In contrast, the right hand side of (3.2) is finite with many choices of α, as we will illustrate in the next section. However, mathematically, it is possible for (4.4) to be less than (3.4), because the sample size in (4.4) is n, not n_1 or n_2.

5. Non-iterative Choice of α

Although the iterative choice of α given in (4.1) leads to an estimator that achieves the minimum error, it is also desirable in practice to have simple non-iterative procedures that have good, not necessarily optimal, properties. Such a non-iterative estimator, for example, can be used as a starting value of the

iteration defined in (4.1). Also, such estimators can be better than $\hat{r}_O$ when the draws are not independent. In this section, we discuss several simple choices of α that seem appealing and to have good potential. For simplicity of presentation, for each chosen α, we only list the corresponding identity as a special case of (2.3), with r being subscribed to identify its estimator.

(I) *The Geometric,* $\alpha = (q_1 q_2)^{-1/2}$:

$$r_G = \frac{E_2[(q_1/q_2)^{1/2}]}{E_1[(q_2/q_1)^{1/2}]}. \tag{5.1}$$

Compared to the original importance ratio, q_1/q_2 of (1.4), the square root in (5.1) not only stabilizes the magnitudes of the importance ratios, but also guarantees that both $(q_1/q_2)^{1/2}$ and $(q_2/q_1)^{1/2}$ are square integrable with respect to p_2 and p_1, respectively. Furthermore, the (asymptotic) relative error of $\hat{r}_G$ has a simple and appealing form (derived from (3.2) with $\alpha = (q_1 q_1)^{-1/2}$)

$$RE^2(\hat{r}_G) \doteq \frac{1}{n s_1 s_2}\Big\{\frac{b}{[\int_{\Omega_1\cap\Omega_2}(p_1 p_2)^{1/2}\,dw]^2} - 1\Big\}, \tag{5.2}$$

where $b = s_1 p_1(\Omega_1\cap\Omega_2) + s_2 p_2(\Omega_1\cap\Omega_2)$, with $p_i(\Omega_1\cap\Omega_2) = \int_{\Omega_1\cap\Omega_2} p_i(w)dw$, $i = 1, 2$. Note that $b \le 1$, with equality when $\Omega_1 = \Omega_2$, a condition that typically holds in practice. Rewriting (5.2), we obtain

$$RE^2(\hat{r}_G) \doteq \frac{1}{n s_1 s_2}\Big\{b\Big[1 - \frac{1}{2}H^2(p_1, p_2)\Big]^{-2} - 1\Big\}, \tag{5.3}$$

where

$$H(p_1, p_2) = \Big[\int_{\Omega_1\cup\Omega_2}(\sqrt{p_1} - \sqrt{p_2})^2\,dw\Big]^{\frac{1}{2}} \tag{5.4}$$

is the Hellinger distance between p_1 and p_2. This connection with the Hellinger distance will be further discussed in Section 8.

When $\Omega_1 = \Omega_2$ (and thus $b = 1$), we note from (5.2) that if $\hat{r}_G$ is used, then the optimal allocation of sample sizes, given $n_1 + n_2 = n$, is $n_1 = n_2 = n/2$. Equal-sample-size allocation, or more generally, equal-time allocation if sampling from the two densities requires different amount of time per draw, was also recommended by Bennett (1976) based on a more general study.

(II) *The Power Family,* $\alpha(k, A) = [q_1^{1/k} + (Aq_2)^{1/k}]^{-k}$, for pre-selected constants $A > 0$ and $k > 0$:

$$r_P(k, A) = \frac{E_2\Big[1 + (Aq_2/q_1)^{1/k}\Big]^{-k}}{E_1\Big[(q_1/q_2)^{1/k} + A^{1/k}\Big]^{-k}}. \tag{5.5}$$

This class of α's is motivated by both $\hat{r}_O$ and $\hat{r}_G$. When the draws are independent, the optimal α given by (3.5) corresponds to $\alpha(1, A_O)$ with $A_O = rn_2/n_1$; and thus a sensible choice of A can make $\hat{r}_P(1, A)$ close to the optimal estimator $\hat{r}_O$. On the other hand, bad choices of A can result in large errors for $\hat{r}_P(1, A)$, as we will illustrate in the next section. When draws are not independent, however, the matter is more complicated, because n_2/n_1 does not necessarily correspond to the ratio of the effective sample sizes; here we conjecture that the optimal choice of α when draws are dependent is still of the form $\alpha(1, A)$ but with A determined by r and the ratio of the effective sample sizes. Since it is generally difficult to decide upon the effective sample sizes with Markov chain Monte Carlo, it seems to be more relevant in practice to search for good choices of α's that are not too sensitive to the effective sample sizes.

This motivates us to consider $\alpha(k, A)$ for k other than 1. We note that $\lim_{k\to\infty} 2^k \alpha(k, A) = (Aq_1q_2)^{-1/2}$, which implies that when k approaches $+\infty$, $\hat{r}_P(k, A)$ approaches $\hat{r}_G$ based on (5.1) for any $A > 0$, because multiplying α by any constant factor dose not change the ratio estimator. This suggests that $\hat{r}_P(k, A)$ may become less sensitive to A, and thus to the effective sample sizes, for large k. However, an undesirable feature of $\hat{r}_G$ (corresponding to $k = +\infty$) is that the resulting integrands, $(q_1/q_2)^{1/2}$ and $(q_2/q_1)^{1/2}$, are not necessarily bounded, in contrast to the integrands in (5.5), which are bounded by $\max\{1, A^{-1}\}$. The unboundness is a main source of large variations of the resulting estimators. This suggests a compromise when choosing k in order to achieve small Monte Carlo error when the effective sample sizes are hard to determine: we want bounded integrands as well as robustness against the misspecification of the effective sample sizes and thus of A. We also note that when $k \to 0$, $\alpha(k, A) \to 1/\max(q_1, Aq_2)$, another interesting choice of α. We will investigate the choice of k in the next section with a specific example; how to choose k in general is an interesting open problem.

(III) *Constant,* $\alpha = 1$:

$$r_C = E_2(q_1)/E_1(q_2). \tag{5.6}$$

This choice of α was suggested by Andrew Gelman, and works remarkably well in the simple example of the next section. Its full potential remains to be explored. One disadvantage of (5.6) is that, unlike (5.1) or (5.5), the integrands in (5.6) are not constrained to be a constant when $q_1 = q_2$. As a consequence, $RE\,(\hat{r}_C) > 0$ even when $q_1 = q_2$ and thus r is known to be 1.

6. A Theoretical Illustration

To examine to what extent the choice of α can affect the error of $\hat{r}_\alpha$, we provide theoretical calculations of $RE^2\,(\hat{r}_\alpha)$ for α's discussed in previous sections

with a simple and informative example. Let $p_1 = N(0,1)$, $p_2 = N(\mu,1)$, where μ is an arbitrary constant. In this case, both normalizing constants are known and can be chosen arbitrarily. We choose $c_1 = c_2$ for simplicity.

We start with $\hat{r}_S$ of (4.3), which is the conventional importance sampling estimator using all $n = n_1 + n_2$ draws from one density, say p_2. Straightforward calculation of (4.4) yields

$$RE^2_\mu(\hat{r}_S) = \frac{1}{n}[\exp(\mu^2) - 1]. \tag{6.1}$$

In comparison,

$$RE^2_\mu(\hat{r}_G) \doteq \frac{4}{n}\Big[\exp(\frac{\mu^2}{4}) - 1\Big] \tag{6.2}$$

and

$$RE^2_\mu(\hat{r}_C) \doteq \frac{4}{n}\Big[\frac{2}{\sqrt{3}}\exp(\frac{\mu^2}{6}) - 1\Big], \tag{6.3}$$

where $\hat{r}_G$ and $\hat{r}_C$ correspond respectively to $\alpha = (q_1 q_2)^{-1/2}$ and $\alpha = 1$, with $n_1 = n_2 = n/2$ (assuming n is even). The optimal error, (3.4), is not in closed form, but we have

$$RE^2_\mu(\hat{r}_O) \doteq \frac{4}{n}\Big[\frac{1}{\beta(\mu)\sqrt{2\pi}}|\mu|\exp(\frac{\mu^2}{8}) - 1\Big], \tag{6.4}$$

where

$$\beta(\mu) = \frac{1}{\pi}\int_0^\infty \frac{\exp(-y^2/(2\mu^2))}{\cosh(y/2)}\,dy \tag{6.5}$$

with the property that $\beta(\mu) \le 1 = \lim_{|\mu|\to+\infty}\beta(\mu)$. Thus,

$$RE^2_\mu(\hat{r}_O) \ge \frac{4}{n}\Big[\frac{1}{\sqrt{2\pi}}|\mu|\exp(\frac{\mu^2}{8}) - 1\Big], \tag{6.6}$$

and the right-hand side of (6.6) approximates $RE^2_\mu(\hat{r}_O)$ for large $|\mu|$.

It is striking to see that a good choice of α can reduce the coefficient of the n^{-1} term by orders of magnitude, especially for large $|\mu|$, that is, when the two densities are far apart. Table 1 gives $RE_\mu(\hat{r}_\alpha)$ (relative standard error) for $n_1 = n_2 = 50$, where $RE_\mu(\hat{r}_O)$ is obtained by numerical integration. It is seen that the improvement of using two densities over one density is dramatic, which is expected since sampling from the overlap of two densities is much more stable than sampling from one extreme tail.

To compare the $RE_\mu(\hat{r}_\alpha)$'s on a finer scale, we graph, in Figure 1, $RE_\mu(\hat{r}_\alpha)$ $/RE_\mu(\hat{r}_O)$ for $\mu \in [0,5]$; we will use the optimal standard error as the baseline in all the figures. It is seen that $\hat{r}_G$ outperforms $\hat{r}_S$ for all values of μ (it is easy

to show that $\exp(\mu^2) - 1 \geq 4[\exp(\mu^2/4) - 1]$ for all μ); $\hat{r}_C$ performs a little worse at the beginning due to the problem mentioned in Section 5, but soon catches up with an error that is almost identical to the optimal one (i.e., $\hat{r}_O$) before it eventually takes off.

Table 1. Comparison of four relative standard errors

μ	$RE_\mu(\hat{r}_S)$	$RE_\mu(\hat{r}_G)$	$RE_\mu(\hat{r}_C)$	$RE_\mu(\hat{r}_O)$
0	0	0	0.079	0
1	0.131	0.107	0.121	0.101
2	0.732	0.262	0.224	0.221
3	9.001	0.583	0.409	0.403
4	298.1	1.464	0.790	0.737
5	26834	4.548	1.714	1.439

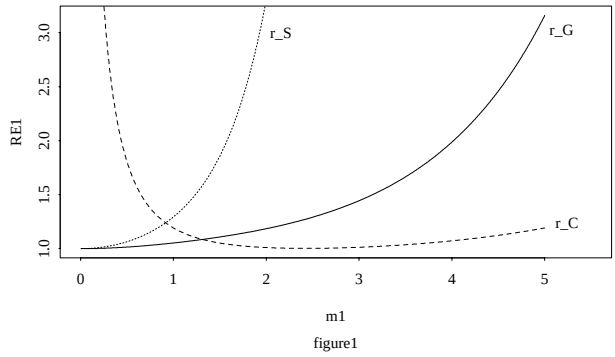

figure1

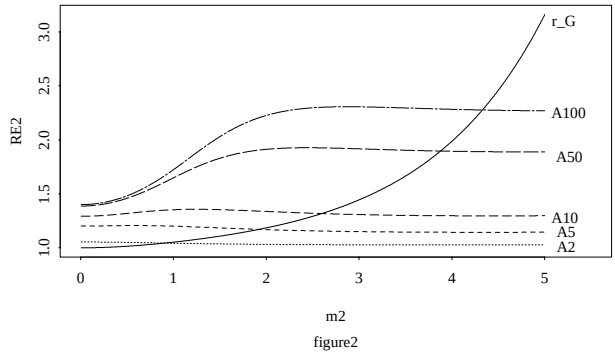

figure2

To show how the choices of k and of A can affect the performance of the power family defined by (5.5), Figure 2 displays $RE_\mu(\hat{r}_P(1, A))/RE_\mu(\hat{r}_O)$ as a function of μ with several choices of A (note $\hat{r}_O = \hat{r}_P(1, 1)$ in this example). This is to check the sensitivity of the optimal estimator $\hat{r}_P(1, A)$ to the misspecification of A. (Since the issue of unknown r can be handled by iteration, as detailed in Section 4, misspecification of A is equivalent to misspecifi-

cation of the ratio of the effective sample sizes.) For comparisons, we also plot $RE_\mu(\hat{r}_G)/RE_\mu(\hat{r}_O) \equiv RE_\mu(\hat{r}_P(+\infty, A))/RE_\mu(\hat{r}_O)$; we will include this comparison in all figures. From Figure 2, we see that the further the A departs from its optimal value $A = A_O$ (=1 in this example), the larger the Monte Carlo error for $\hat{r}_P(1, A)$, which thus becomes worse than $\hat{r}_G$ for $\mu \in [0, \mu_A]$, where μ_A increases with $A > 1$ (we focus on $A > 1$ because $RE_\mu(\hat{r}_P(1, A)) = RE_\mu(\hat{r}_P(1, A^{-1}))$ when $s_1 = s_2$). This confirms our intuition that if the correct A cannot be assessed within a reasonable range (e.g., $|\log(A/A_0)| \leq \log(5)$), it may be better to use a non-optimal estimator that is less sensitive to the choice of A, at least when the two densities are not too far apart. When the two densities are far apart (e.g., when μ is large in our example), the increase in $RE_\mu(\hat{r}_P(k, A))$, and thus $RE_\mu(\hat{r}_P(1, A))$, due to the misspecification of A becomes less and eventually negligible when compared to a $RE_\mu(\hat{r}_\alpha)$ that is a function of the distance between the two densities with a higher order (e.g., $RE(\hat{r}_G)$), because A can only affect $RE_\mu((\hat{r}_P(k, A))$ by a multiplicative factor when the distance become infinite.

More specifically, for the current problem, we have (see the Appendix)

$$\lim_{|\mu|\to\infty} \frac{RE_\mu(\hat{r}_P(k, A))}{RE_\mu(\hat{r}_P(1, A_O))} = \frac{\tau(k)}{\tau(1)} \left[\frac{(A/A_O)^{1/2} + (A_O/A)^{1/2}}{2}\right]^{1/2} \geq 1, \tag{6.7}$$

where

$$\tau(k) = \frac{\sqrt{B(\frac{3k}{2}, \frac{k}{2})}}{\sqrt{k} B(\frac{k}{2}, \frac{k}{2})}, \quad k > 0, \tag{6.8}$$

with $B(a, b)$ the standard Beta function. Thus the asymptotes in Figure 2 for large μ correspond to the multiplicative factor $[(\sqrt{A} + 1/\sqrt{A})/2]^{1/2}$ since $k = 1$ and $A_O = 1$ in the plots. We observe that the theoretical limits are well approximated even when μ is as small as 3.

We also observe that there are asymptotes when $\mu \to 0$ in Figure 2. The theoretical expression underlying these asymptotes is (see the Appendix for proof)

$$\lim_{\mu\to 0} \frac{RE_\mu(\hat{r}_P(k, A))}{RE_\mu(\hat{r}_P(1, A_O))} = \left\{1 + \frac{1}{s_1 s_2}\Big[\frac{(s_1/s_2)^{\frac{1}{k}}}{(s_1/s_2)^{\frac{1}{k}} + (A/A_O)^{\frac{1}{k}}} - s_1\Big]^2\right\}^{1/2} \geq 1, \tag{6.9}$$

where s_1 and s_2 are the proportions of the *effective* sample size for the two samples, respectively. It follows then that, when $\mu \to 0$, the optimal choice of k for given A is

$$k_O = 1 + \frac{\log(A/A_O)}{\log(s_2/s_1)}. \tag{6.10}$$

When $A = A_O$, $k_O = 1$ is optimal unless $s_1 = s_2$, in which case any k is optimal when $\mu \to 0$; this can be seen in Figure 3, which exhibits plots of

$RE_\mu(\hat{r}_P(k, A_O))/RE_\mu(\hat{r}_O)$ with several k. When A is misspecified, however, $k = 1$ is no longer optimal when $\mu \to 0$. In fact, when $s_1 = s_2$, $k = +\infty$ (corresponding to $\hat{r}_G$) is optimal when $\mu \to 0$. This is in contrast to the optimal choice of k when $|\mu| \to +\infty$, in which case $k = 1$ is optimal regardless of the value of A, as implied by (6.7). Figure 4, which exhibits plots of $RE_\mu(\hat{r}_P(k, 10))/RE_\mu(\hat{r}_O)$ with several choices of k, provides an numerical illustration of this conflict. However, despite the sharp conflict in choosing k (i.e., $k = 1$ for $\mu \to +\infty$ and $k = +\infty$ for $\mu \to 0$), it is possible to find a compromise (e.g., $k = 5$ in Figure 4) that works remarkably well when compared to the optimal estimator, as long as μ is not too large (when μ is large, even the optimal estimator is not usable). Of course, in real application, we will not know by how much we have misspecified the ratio of the effective sample sizes, but the misspecification of this *ratio* cannot be too extreme in reality (e.g., $A/A_O = 10$ which we used in this example should be quite extreme in practice). We believe it is possible in real applications to find such "compromise" estimators that will work well as long as the misspecification is not too extreme, and we hope our example here will serve as a stimulus for general search of such estimators.

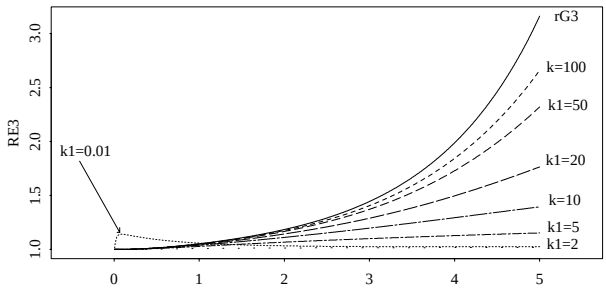

Figure3

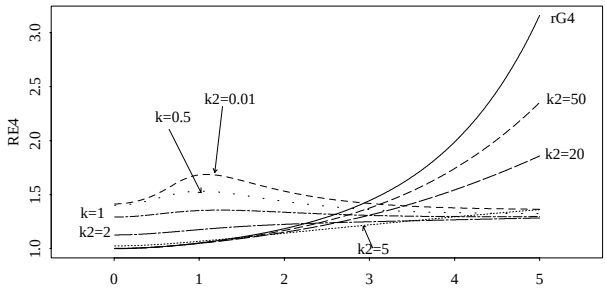

Figure4

7. Extensions – When Draws From All Densities Are Available

The utility of the identity (2.3) can be enhanced by considering its extensions to cases involving more than two densities. The theory underlying these extensions is more complex, and here we confine ourselves to the most basic formulation of these extensions. There are two types of multi-density extensions, one type regarding cases where draws from all densities are available, and the other one regarding cases where only draws from some densities are available. The latter extension will be discussed in the next section.

Consider a setting where we have draws $\mathcal{W} = \{w_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, m\}$ from $p_i(w) = q_i(w)/c_i, i = 1, \ldots, m$ and we are interested, say, in calculating $r_i = c_1/c_i, i = 2, \ldots, m$. An application of this, as in the genetic problem discussed in Section 1, is when we need to compute a likelihood at a variety of locations (relative to a reference value). Another application, which is the focus of Geyer (1994), is for creating a sensible importance-sampling density using a mixture of m densities, from which we have draws (e.g., via Gibbs sampler) and the unnormalized densities. The methods described before can obviously be applied to each ratio, c_1/c_i, by using draws from each pair of densities, p_1 and p_i. Such methods, however, can be made more efficient by using draws from all densities. There are at least two approaches for constructing a multi-density extension of the identity (2.3). The following approach is more intuitive, and thus we describe it first. For simplicity, consider a case with $m = 3$, as represented by the graph in Figure 5.

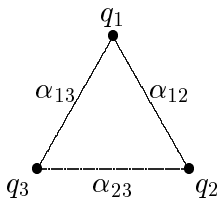

Figure 5. A complete graph for three densities

In Figure 5, each vertex represents a (unnormalized) density, and each edge represents a "bridge" from one density to another, with α_{ij} being a weighting function. As illustrated in the previous sections, the accuracy of the simulation based on (2.3) is determined by the "overlap" of the two densities, where the "overlap" is defined by a measure (based on the choice of α_{ij}) on the common support. When we have three densities, there are two paths from, say, q_1 to q_2. The direct path from q_1 to q_2 represents a direct estimation of c_1/c_2. The path from q_1 to q_3 and then to q_2 represents estimating c_1/c_2 via the product:

$(c_1/c_3)(c_3/c_2) = c_1/c_2$. Combining these two paths yields an extension of (2.3) with three densities:

$$\frac{c_1}{c_2} = \frac{E_2(q_1\alpha_{12}) + E_2(q_3\alpha_{23}) \cdot E_3(q_1\alpha_{13})}{E_1(q_2\alpha_{12}) + E_3(q_2\alpha_{23}) \cdot E_1(q_3\alpha_{13})} \equiv \frac{A_1 + A_2}{B_1 + B_2}, \tag{7.1}$$

which is easy to verify directly.

The advantage of (7.1) is that it uses the two paths in an efficient way. For instance, if there is no direct overlap between q_1 and q_2, implying $E_2(q_1\alpha_{12}) = E_1(q_2\alpha_{12}) = 0$, then (7.1) uses the indirect path automatically. We can also set some α's to zero to "cut off" some paths. For example, choosing $\alpha_{23} = 0$ or $\alpha_{13} = 0$ disconnects the indirect path, and reduces (7.1) to (2.3). Another advantage of (7.1) is that it suggests more general identities that may be useful for simulating c_1/c_2. For example, for any $\xi \neq 0$, using A_i and B_i $(i = 1, 2)$ defined in (7.1), we have

$$\frac{c_1}{c_2} = \left\{ \frac{A_1^\xi + A_2^\xi}{B_1^\xi + B_2^\xi} \right\}^{\frac{1}{\xi}}. \tag{7.2}$$

For $m = 2$, this extension is the same as (2.3) because of the cancellation of the powers. We note that when $\xi \to 0$, $+\infty$, $-\infty$, the right side of (7.2) converges to $\sqrt{A_1A_2}/\sqrt{B_1B_2}$, $\max\{A_1, A_2\}/\max\{B_1, B_2\}$, $\min\{A_1, A_2\}/\min\{B_1, B_2\}$, respectively. Thus, (7.2) can be defined for all $\xi \in [-\infty, +\infty]$. Besides the choices of α_{ij}, the choice of ξ seems to be a problem worthy of investigation with finite samples.

The extensions of (7.2) to $m \geq 4$ are straightforward by considering a complete graph connecting all m densities, with α_{ij} being a weighting function on the edge with q_i and q_j $(i \neq j)$ as its end points. There are $(m-2)![\sum_{l=0}^{m-2}(l!)^{-1}]$ possible paths from q_1 to q_2, and the corresponding extension of (7.2) is

$$\frac{c_1}{c_2} = \left\{ \frac{\sum_{l=0}^{m-2} \sum_{(2,i_1,\ldots,i_l,1)} \left[\prod_{k=0}^{l} E_{i_k}(q_{i_{k+1}}\alpha_{i_k i_{k+1}}) \right]^\xi}{\sum_{l=0}^{m-2} \sum_{(2,i_1,\ldots,i_l,1)} \left[\prod_{k=0}^{l} E_{i_{k+1}}(q_{i_k}\alpha_{i_k i_{k+1}}) \right]^\xi} \right\}^{\frac{1}{\xi}}, \quad \xi \in [-\infty, +\infty], \tag{7.3}$$

where $(2, i_1, \ldots, i_l, 1)$ represents a path from q_2 to q_1 crossing l distinct vertices (excluding q_1 and q_2), $i_0 = 2$, and $i_{l+1} = 1$ for all l as the largest index of k. Taking $\xi = 1$ in (7.3) provides an extension of (7.1) to m densities. The identity (7.3) provides a large class of estimators, and the search for (asymptotically) optimal $\alpha_{i_k i_{k+1}}$ (and ξ) is a challenging task, but the results can be quite useful for implementations in practice.

A second approach of construction, which is theoretically as well as computationally more attractive, proceeds by setting up a linear system for $\vec{r} =$

$(r_2, \ldots, r_m)^\top$. Let $\{\alpha_{ij}(w) : 2 \le i \le m; 1 \le j \le m; i \ne j\}$ be $(m-1)^2$ known functions, then the key identity (2.3) implies $B\vec{r} = \vec{b}$ where

$$B = \begin{pmatrix} b_{22} & -b_{23} & \ldots & -b_{2m} \\ -b_{32} & b_{33} & \ldots & -b_{3m} \\ \vdots & \vdots & \ddots & \vdots \\ -b_{m2} & -b_{m3} & \ldots & b_{mm} \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} b_{21} \\ b_{31} \\ \vdots \\ b_{m1} \end{pmatrix},$$

$$\text{with} \begin{cases} b_{ii} = \sum_{j \ne i} E_j[\alpha_{ij}(w) q_i(w)], & 2 \le i \le m, \\ b_{ij} = E_i[\alpha_{ij}(w) q_j(w)], & i \ne j. \end{cases}$$

If all α_{ij} are chosen such that B is well-defined and non-singular, then we have

$$\vec{r} = B^{-1}\vec{b}, \tag{7.4}$$

which is a matrix extension of (2.3). An estimator for $\vec{r}$ is then obtained by replacing B and $\vec{b}$ in (7.4) by their sample counterparts, B_n and $\vec{b}_n$, the sample averages constructed using the draws in $\mathcal{W}$.

Just as (2.3) leads to an iterative sequence converging to the "reverse logistic regression" estimator of Geyer (1994) when $m = 2$, (7.4) also yields an analogous sequence for $m > 2$. Specifically, Geyer's (1994) approach is to first construct a "profile-loglikelihood" that can be expressed in our notation as

$$L(\vec{r}\,|\mathcal{W}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log\Big[\frac{r_i s_i q_i(w_{ij})}{\sum_{k=1}^{m} r_k s_k q_k(w_{ij})}\Big], \tag{7.5}$$

where $s_i = n_i/n > 0$, $n = \sum_{i=1}^{m} n_i$, and $r_1 \equiv 1$, and then to compute the corresponding maximizer as an estimator for $\vec{r}$. An easier way of understanding this approach is to view it as an application of (1.4) using $q_{mix} = \sum_k r_k s_k q_k(w) \equiv c_1 p_{mix}$ as the denominator (i.e., $q_2(w)$) of the right side of (1.4). More specifically, if we apply (1.4) with $q_2 = q_{mix}$ and $q_1 = q_i$ (here q_1 and q_2 are viewed as generic notation), we obtain

$$r_i = (c_i/c_1)^{-1} = \left[E_{mix}\left(q_i(w)/q_{mix}(w)\right)\right]^{-1}, \qquad i = 2, \ldots, m, \tag{7.6}$$

where the expectation is taken with respect to the mixture density $p_{mix} \equiv q_{mix}/c_1$. Since q_{mix} depends on $\vec{r}$, (7.6) cannot be used directly to construct an estimator but it immediately suggests an iterative sequence if we view $\mathcal{W}$ as a set of draws from p_{mix}; for $m = 2$, this sequence was constructed and referred to as the "third iterative scheme" in Section 4. It is easy to verify that the fixed-point (or "self-consistent") equation defined by this iteration is identical to the "score" equation corresponding to (7.5), and thus Geyer's (1994) approach

is effectively an "iterative" application of Torrie and Valleau's (1977) "umbrella sampling" with p_{mix} as the umbrella density (see Section 9), though Geyer's (1994) derivation is more statistical.

For $m = 2$, we mentioned in Section 4 that Geyer's (1994) estimator is the same as our optimal estimator $\hat{r}_O$, and our iteration (4.1) provides a fast and stable algorithm for computing it. For $m > 2$, it is an open problem whether the maximizer of (7.5) is also optimal among the class of estimators constructed from (7.4) or more generally from (7.3). It is, however, easy to construct an iterative sequence from (7.4) that converges to the maximizer of (7.5). This is achieved by setting $\alpha_{ij}(w) = s_i s_j / q_{mix}(w)$ for all i, j, and then iterate via (7.4), that is,

$$\vec{r}_O^{(t+1)} = [B_n(\vec{r}_O^{(t)})]^{-1} \vec{b}_n(\vec{r}_O^{(t)}), \quad t = 0, 1, \ldots, \tag{7.7}$$

where we use the subscript "O" to indicate that (7.7) is a direct extension of the $\hat{r}_O^{(t)}$ sequence of (4.1), and use the argument in B_n and $\vec{b}_n$ to mark their dependence on the previous iterate $\vec{r}^{(t)}$ through the dependence of q_{mix} on r. Because the $\vec{r}_O^{(t)}$ of (7.7) provides a consistent estimator for $\vec{r}$ for any $t \geq 1$, just as (4.1) does when $m = 2$, we expect that (7.7) will provide a fast and stable algorithm for computing $\vec{r}_O$, an algorithm that might be more appreciated in practice than the Newton-Raphson or a slow successive algorithm discussed in Geyer (1994) for maximizing (7.5) (the matrix inversion in (7.7) can be avoided by directly solving the linear equation $B_n(\vec{r}_O^{(t)})\vec{r}_O^{(t+1)} = \vec{b}_n(\vec{r}_O^{(t)})$ for $\vec{r}_O^{(t+1)}$).

8. Extensions – When Only Draws From Some Densities Are Available

In what has been described so far, we have assumed that random draws from all densities involved are available. Practically, this requirement may be undesirable as making draws from every density can be expensive; it is perhaps also unnecessary in some cases as explained below. Often in applications, the densities $p_i(w)$'s are related to each other as they all arise from a common parametric family:

$$p_i(w) = p(w|\theta_i) = q(w|\theta_i)/c(\theta_i), \qquad i = 1, 2, \ldots \tag{8.1}$$

We have seen in previous sections that the simulation error of using (2.3) is determined by a distance between two densities, the forms of which depend on the choice of α. When densities arise from (8.1), the distance between $p(w|\theta_1)$ and $p(w|\theta_2)$ is often a smooth function of a distance between θ_1 and θ_2. In such cases, the closeness of $p(w|\theta_1)$ and $p(w|\theta_2)$ tends to ensure the closeness of $p(w|\theta)$ to both $p(w|\theta_1)$ and $p(w|\theta_2)$ for any θ "between" θ_1 and θ_2. For such a θ, we can expect good accuracy of simulating $c(\theta)/c(\theta_i), i = 1, 2$, using only draws from $p(w|\theta_1)$ and $p(w|\theta_2)$. The following identity provides a basis for such simulations.

Let $p_i(w) = q_i(w)/c_i, w \in \Omega_i \subset R^d, i = 1, 2, 3$, be three densities such that $\Omega_3 \subseteq \Omega_1 \cup \Omega_2$. Also let $\alpha_i(w), i = 1, 2, 3$, be three arbitrary functions defined on $\Omega_1 \cup \Omega_2$ subject to the constraint

$$\alpha_1(w)q_1(w) + \alpha_2(w)q_2(w) = 1, \qquad w \in \Omega_1 \cup \Omega_2, \tag{8.2}$$

and $0 < |\int_{\Omega_1 \cap \Omega_2} \alpha_3(w)p_1(w)p_2(w)dw| < \infty$. Then

$$\frac{c_3}{c_1} = \frac{E_1(q_3\alpha_1)E_2(q_1\alpha_3) + E_2(q_3\alpha_2)E_1(q_2\alpha_3)}{E_2(q_1\alpha_3)}. \tag{8.3}$$

The right side of (8.3) only involves expectations with respect to the first two densities, where random draws are available. A class of choices of α_1 and α_2 that satisfies (8.2) is given by $\alpha_1 = 1/(q_1 + Aq_2)$ and $\alpha_2 = A/(q_1 + Aq_2)$ for any $A > 0$. Good properties of this class are expected because it is closely related to the optimal procedures discussed in Sections 4 and 5.

Practically, in the case of (8.1), if we need to compute $c(\theta)/c(\theta_1)$ for many values of θ and at the same time want to make draws from $p(w|\theta)$ at as few values of θ as possible, we can start with θ_1 and use draws from $p(w|\theta_1)$ to estimate a distance between $p(w|\theta_1)$ and $p(w|\theta_2)$ to determine the next density, $p(w|\theta_2)$, from which draws will be made. Once the draws from $p(w|\theta_2)$ are made, simulation of $c(\theta)/c(\theta_1)$ for any θ "between" θ_1 and θ_2 can be performed using (8.3) with $c_3 = c(\theta)$. We can then proceed from $p(w|\theta_2)$ as the new starting density. The key step in such a procedure is the search of the next density (when θ is multi-dimensional, one may need to search for several densities to construct a "convex region" in order to achieve better simulation efficiency), which is determined by the distance (from the previous density) that is acceptable given the required accuracy of simulation. A reasonable choice of distance is the Hellinger distance, $H(p_1, p_2)$ of (5.4), as explained below (where, again, we assume all draws are independent).

Consider a case where all densities have the same support (this is typically true under (8.1)), and equal number of draws will be made from all selected densities. In such a case, the (asymptotic) optimal relative error of simulating c_1/c_2 using (2.3), as given in Theorem 2, is determined by the "harmonic" distance between p_1 and p_2:

$$RE^2(\hat{r}_O) \doteq \frac{4}{n}\Big\{\Big[\int 2[p_1^{-1} + p_2^{-1}]^{-1}dw\Big]^{-1} - 1\Big\}.$$

It follows then, to the first order, that

$$\frac{4}{n}\Big\{\Big[\int \sqrt{p_1p_2}\,dw\Big]^{-1} - 1\Big\} \le RE^2(\hat{r}_O) \le \frac{4}{n}\Big\{\Big[\int \sqrt{p_1p_2}\,dw\Big]^{-2} - 1\Big\}. \tag{8.4}$$

The left inequality follows from the fact that a harmonic mean cannot exceed the corresponding geometric mean, and the right inequality holds because the right-most side of (8.4) is the relative error of a non-optimal estimator $\hat{r}_G$, as in (5.2) with $b = 1$. Thus, the optimal relative error is bounded both below and above by a simple function of $\int \sqrt{p_1 p_2}\, dw = 1 - \frac{1}{2}H^2(p_1, p_2)$.

Estimating $\int \sqrt{p_1 p_2}\, dw$ using draws from only one density is easy because of the following identity:

$$h \equiv \int \sqrt{p_1 p_2}\, dw = E_2\Big[\sqrt{q_1/q_2}\Big] \Big/ \sqrt{E_2(q_1/q_2)}. \tag{8.5}$$

Given draws from p_2, we can estimate h by

$$\hat{h} = \frac{\frac{1}{n_2}\sum_{j=1}^{n_2} \sqrt{\xi_j}}{\sqrt{\frac{1}{n_2}\sum_{j=1}^{n_2} \xi_j}}, \tag{8.6}$$

where $\xi_i = q_1(w_{2i})/q_2(w_{2i})$, with $w_{21}, \ldots, w_{2n_2}$ being n_2 draws from p_2. The asymptotic error of $\hat{h}$ under independent sampling is

$$E_1(\hat{h} - h)^2 \doteq \frac{1}{n_2}\Big\{1 - h \int \frac{p_1^{3/2}}{p_2^{1/2}}\, dw + \frac{h^2}{4}\int \frac{(p_1 - p_2)^2}{p_2}\, dw\Big\}. \tag{8.7}$$

Due to the square roots, estimating h is more stable than estimating c_1/c_2 directly using draws from p_2, as can be seen from the cancellation of the negative and positive parts in (8.7) when p_1 and p_2 are not too far apart. We also expected that when two densities are far apart, $\hat{h}$ would underestimate h due to large values of q_1/q_2 in the denominator. This would imply overestimation of the Hellinger distance when the distance is large, which is acceptable because it provides a conservative procedure. Of course, empirical studies are needed to investigate the practical performance of $\hat{h}$, as well as any procedure presented in this paper.

9. Epilogue – Related and Subsequent Work

The work presented in this paper was initially motivated by our observation that methods for simulating normalizing constants in statistical literature had mainly focused on the importance sampling scheme (1.4) using draws from one density, and we felt that it would be more efficient to use draws from both densities when simulating the corresponding ratio. To do that, we needed an identity that generalizes (1.4), and the key identity (2.3) was the result of our search. Because (2.3) is so simple and obviously powerful, we doubted that we were the first one who discovered it (at least in some forms), but it appeared to

be new to statisticians. Indeed, the comments from audiences during our several presentations of this work were quite enthusiastic, as many of them immediately saw the potential of (2.3), and so did the reviewers of *Statistica Sinica*. On the other hand, some anonymous comments we received earlier indicated that using draws from a single density had been a standard approach in statistical literature for so long that it might take sometime before the utility of (2.3) or its generalizations (e.g., Sections 7 and 8) can be fully appreciated.

The work presented here has stimulated much subsequent work and has helped to bring to our attention more related work, especially in computational physics (e.g., Bennett (1976)). Using the full-information item factor (FIIF) model (Bock and Aitken (1981)) as a working model, Meng and Schilling (1996) provide a detailed empirical investigation of most theoretical constructions presented in this paper (e.g., Sections 4 and 5). In particular, they illustrate how to use $\hat{r}_O$ or $\hat{r}_G$ to simulate ratios of likelihoods needed for monitoring the convergence of Monte Carlo EM. Their empirical evidence strongly supports the theoretical predictions here. For example, they find that compared to the importance-sampling estimator $\hat{r}_S$ of (4.3), $\hat{r}_O$ and $\hat{r}_G$ exhibit anywhere from 5 to 30 times lower mean-squared error in their FIIF applications. As another example, the sequence (4.1) converges much faster, as predicted, to $\hat{r}_O$ than (4.2) does, especially when p_1 and p_2 are far apart (e.g., (4.1) was about 7 times faster on average with a Hellinger distance $H(p_1, p_2) = 1.11$ between p_1 and p_2). A discrepancy with the prediction here is with respect to the underestimation of (8.6) for (8.5). In Meng and Schilling (1996), it was found that (8.6) can substantially overestimate (8.5) when $H(p_1, p_2)$ is large (e.g., > 1), an issue that needs further investigation.

DiCiccio, Kass, Raftery and Wasserman (1996) provide more empirical evidence of the superiority of the optimal estimator discussed in Section 4. They use $\hat{r}_O$ to compute a single normalizing constant (i.e., the $p(Y)$ of (1.3)) by coupling the unnormalized density (i.e., $p(Y|\theta)p(\theta)$) with its normal approximation, which is trivial to draw from. The empirical results they provide show that $\hat{r}_O$ dominates all other methods they have considered, including analytic approximations (i.e., Laplace approximation with or without Bartlett corrections) and simulation methods (e.g., the importance sampling method and the "reciprocal" method; both are special cases of (2.3)). Furthermore, $\hat{r}_O$ is often an order of magnitude better than other methods in terms of mean absolute deviation of $\log(\hat{r})$ (which is different from the error we considered in Section 3). They also report that using $\hat{r}_P(1, A)$ (see (5.5)) with A determined by their modified Laplace method works almost as well as $\hat{r}_O$. This is expected because, as suggested by Figure 2, $\hat{r}_P(1, A)$ is quite close to $\hat{r}_P(1, A_O)$ when A is reasonably close to A_O (e.g., within a factor of 2 to 5, which is expected because their modified Laplace method is

itself a reasonably accurate method). In fact, similar results will hold if we use other methods to determine A, or better, we can use the iterative approach given in (4.1), and thus avoid using any other methods. The simulation study from Meng and Schilling (1996) suggest that two to three iterations are often enough to produce an estimator that is very close to $\hat{r}_O$; they purposely choose an extremely variable estimator based on (2.4) to illustrate the remakable robustness of (4.1) to the starting value.

On the methodological side, Gelman and Meng (1994) studied a continuous extension of the key identity (2.3). They started with a re-expression of (2.3) that is operationally less convenient but intuitively more appealing. They first re-express $\alpha = q_0/(q_1 q_2)$ in terms of a new function q_0. Now, if we assume q_0 to be a non-negative function that can be normalized into a density $p_0 = q_0/c_0$, we can rewrite (2.3) as

$$\frac{c_1}{c_2} = \frac{c_0/c_2}{c_0/c_1} = \frac{E_2\left[q_0(w)/q_2(w)\right]}{E_1\left[q_0(w)/q_1(w)\right]}. \tag{9.1}$$

Comparing (9.1) to (1.4) we see that with (1.4) we have to use draws from p_2 to go all the way to "reach" p_1, whereas with (9.1) we can use draws from both p_1 and p_2 with q_0 as a connecting "bridge", and thus effectively shorten the distances between the densities, distances that are responsible for the magnitude of the errors of the estimators. This intuition obviously leads to extensions using multiple bridges, that is, by applying (9.1) in a "chain" fashion, which is a special case of (7.3) and was also briefly discussed in Bennett (1976). Gelman and Meng (1994) showed that the limit from using an infinitely many bridges leads to another identity that allows one to estimate the log of the ratio unbiasedly, an identity that underlies the method of Ogata (1989, 1990) for simulating high dimensional integration via Monte Carlo. Gelman and Meng (1994) thus showed that the method studied in this paper, which can be termed "bridge sampling", is a natural extension of importance sampling via (9.1), and that its further extension by using infinitely many bridges leads to the construction of "path sampling".

Another useful method of shortening the distance between p_1 and p_2 is to apply random-variable transformations and thus "physically" move the two densities closer before using the bridge sampling. The general idea is described in Meng and Schilling (1996a) and was inspired by Voter (1985), who suggested applying a location shift before applying the method of Bennett (1976), although Voter's choice of the "bridge" was more restrictive than Bennett's (1976). Applying a location shift does not change the normalizing constant or increase the computational effort for making the required draws, but can substantially reduce the distance between p_1 and p_2 with an appropriate choice of the amount of shift. Voter (1985) proposed to shift by an amount such that the two densities

will have the same mode if both densities have only one mode, and also proposed a random-shift scheme if there is more than one mode. Clearly, the idea of location shift can be generalized to other types of transformations, for example, rotation and scaling, that can further reduce the (Hellinger) distance between the targeted densities. We note that the method of transformation can handle cases where the original densities have no or little common support(s), a problem that motivated Voter (1985). The potential of handling multiple modes is also of importance in practice, as discussed in DiCiccio, Kass, Raftery and Wasserman (1996). Details of these developments are given in Meng and Schilling (1996a).

In the same spirit as reducing the distance between two densities, Torrie and Valleau (1977) proposed a related but different method. Briefly, their method is based on an "opposite" identity to (9.1), namely

$$\frac{c_1}{c_2} = \frac{c_1/c_0}{c_2/c_0} = \frac{E_0\left[q_1(w)/q_0(w)\right]}{E_0\left[q_2(w)/q_0(w)\right]},$$

where E_0 denotes the expectation of the "middle" density p_0. In other words, we now *sample* from the middle density, which is constructed to have, hopefully, large overlaps with both $p_i, i = 1, 2$. They termed this method "umbrella sampling", conveying the intention of constructing a middle density that "covers" both ends. Notice that with this method the draws from p_1 or p_2 are not used, at least not directly (these draws can be used to form a set of draws from p_0, if p_0 is taken as a mixture of p_1 and p_2; the iteration (4.2) is such an example although, as we have discussed there, this iteration converges much more slowly than the iteration (4.1)). Chen and Shao (1994) also consider this method, which they call the ratio importance sampling method. They also consider simple extensions of (2.3) to allow p_1 and p_2 to have different dimensions.

Researchers in computational physics have been pioneers in many advanced computational methods, especially in the field of Monte Carlo simulation. However, they typically pay more attention to their complicated problems, and their published work involves heavy specialized details so that sometimes it is difficult for researchers from other fields to see the generality of their methods. Once in the hands of statisticians, the applicability of their methods can often be enhanced, especially with generalizations and modifications guided by more statistical considerations. The current use of the Metropolis algorithm (Metropolis et al. (1953)), or more generally Markov chain Monte Carlo in statistics is a good example. We hope our work not only provides a general theoretical framework for bridge sampling but also brings to the attention of statisticians the success of such a method in computational physics. We conclude with a quote from the abstract of Bennett (1976), who, as we mentioned in Section 2, applied (2.3) for computing free-energy differences between two ensembles (i.e., distributions), to

emphasize that using (2.3) can be orders of magnitude more efficient than using (1.4) with little increase in computational effort:

> "The best estimate of the free energy difference is usually obtained by dividing the available computer time approximately equally between the two ensembles; its efficiency (variance $\times$ computer time)$^{-1}$ is never less, and may be several orders of magnitude greater, than that obtained by sampling only one ensemble ....."

Acknowledgements

An early version of this paper was presented as part of an invited talk delivered at the 1993 Joint Statistical Conference held in Taiwan, and at the 1994 IMS/WNAR meeting held in UCLA; comments from the audiences are acknowledged. We thank A. Gelman, A. Kong, K. Lange, S. Schilling, F. Vaida and D. Wallace for helpful conversations, and J. Barnard and S. Pedlow for computational assistance. We also thank Jeff Wu and the reviewers for encouraging comments that lead to a better presentation. This manuscript was prepared using computer facilities supported in part by several NSF grants awarded to the Department of Statistics at The University of Chicago, by the Fairchild Foundation, and by The University of Chicago Block Fund. The research was supported in part by NSF grant DMS 92-04504. Meng's research was also supported in part by NSA grant MDA904-96-1-0007.

Appendix

Proof of (3.2). Let $\bar{\eta}_1$ and $\bar{\eta}_2$ be respectively the numerator and denominator of $\hat{r}_\alpha$ of (2.5). Then under our assumptions, $\bar{\eta}_1$ and $\bar{\eta}_2$ are independent and

$$\eta_i \equiv E(\bar{\eta}_i) = c_i \int_{\Omega_1 \cap \Omega_2} \alpha p_1 p_2 \, dw, \qquad i = 1, 2, \tag{A.1}$$

$$V(\bar{\eta}_1) = \frac{c_1^2}{n_2}\Big\{ \int_{\Omega_1 \cap \Omega_2} p_1^2 p_2 \alpha^2 \, dw - \Big[\int_{\Omega_1 \cap \Omega_2} \alpha p_1 p_2 \, dw \Big]^2 \Big\}, \tag{A.2}$$

and

$$V(\bar{\eta}_2) = \frac{c_2^2}{n_1}\Big\{ \int_{\Omega_1 \cap \Omega_2} p_2^2 p_1 \alpha^2 \, dw - \Big[\int_{\Omega_1 \cap \Omega_2} \alpha p_1 p_2 \, dw \Big]^2 \Big\}. \tag{A.3}$$

By the δ-method, we have

$$RE^2(\hat{r}_\alpha) = \frac{E\left(\frac{\bar{\eta}_1}{\bar{\eta}_2} - \frac{\eta_1}{\eta_2}\right)^2}{\left(\frac{\eta_1}{\eta_2}\right)^2} = \frac{V(\bar{\eta}_1)}{\eta_1^2} + \frac{V(\bar{\eta}_2)}{\eta_2^2} + O\left(\frac{1}{n^2}\right). \tag{A.4}$$

Substituting (A.1) – (A.3) into (A.4) yields (3.2).

Proof of Theorem 2. By the construction of $\hat{r}_O^{(t+1)}$ of (4.1), its limit, if exists, must be a root of the following "score" function

$$S(r|w) = -\sum_{j=1}^{n_1} \frac{s_2 r q_2(w_{1j})}{s_1 q_1(w_{1j}) + s_2 r q_2(w_{1j})} + \sum_{j=1}^{n_2} \frac{s_1 q_1(w_{2j})}{s_1 q_1(w_{2j}) + s_2 r q_2(w_{2j})}. \quad \text{(A.5)}$$

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

$$\begin{aligned} S'(r|w) &\equiv \frac{dS(r|w)}{dr} \\ &= -\sum_{j=1}^{n_1} \frac{s_1 s_2 q_1(w_{1j}) q_2(w_{1j})}{[s_1 q_1(w_{1j}) + s_2 r q_2(w_{1j})]^2} - \sum_{j=1}^{n_2} \frac{s_1 s_2 q_1(w_{2j}) q_2(w_{2j})}{[s_1 q_1(w_{2j}) + s_2 r q_2(w_{2j})]^2} < 0 \end{aligned}$$

for all r, $S(r|w) = 0$ has a unique root. To check that this unique root, denoted by r^*, is the limit $\hat{r}_O$ of Theorem 2, we let $M(r)$ be the mapping defined by the iteration (4.1): $\hat{r}_O^{(t+1)} = M(\hat{r}_O^{(t)})$, and thus $r^* = M(r^*)$. It is easy to verify that $rM(r)$ is a strictly increasing function of r, and $M(r)/r$ is a strictly decreasing function of r. It follows that, if $\hat{r}_O^{(t)} > r^*$, then

$$\frac{\hat{r}_O^{(t+1)}}{\hat{r}_O^{(t)}} = \frac{M(\hat{r}_O^{(t)})}{\hat{r}_O^{(t)}} < \frac{M(r^*)}{r^*} = 1,$$

which implies

$$\hat{r}_O^{(t+1)} - r^* < \hat{r}_O^{(t)} - r^*. \quad \text{(A.6)}$$

On the other hand,

$$\hat{r}_O^{(t+1)} \hat{r}_O^{(t)} = M(\hat{r}_O^{(t)}) \hat{r}_O^{(t)} > M(r^*) r^* = (r^*)^2,$$

implying that

$$\hat{r}_O^{(t+1)} - r^* > \frac{r^*}{\hat{r}_O^{(t)}} (r^* - \hat{r}_O^{(t)}) > r^* - \hat{r}_O^{(t)}. \quad \text{(A.7)}$$

Combining (A.6) and (A.7) leads to

$$|\hat{r}_O^{(t+1)} - r^*| < |\hat{r}_O^{(t)} - r^*|. \quad \text{(A.8)}$$

Analogous arguments apply when $\hat{r}_O^{(t)} < r^*$. The convergence of $\{\hat{r}_O^{(t)}, t \geq 0\}$ to the unique limit r^* then follows from (A.8) and the Global Convergence Theorem (e.g., Wu (1983)) by choosing $|r - r^*|$ as the objective function. The same arguments apply to the sequence $\{\hat{r}_M^{(t)}, t \geq 0\}$ defined by (4.2).

To prove the second part of Theorem 2, we adopt the standard argument for establishing asymptotic variance of a root of the score function, which yields

$$E(\hat{r}_O - r)^2 = \frac{V[S(r|w)]}{E^2[S'(r|w)]} + O\left(\frac{1}{n^2}\right). \tag{A.9}$$

Direct calculations yield

$$V[S(r|w)] = ns_1s_2D(1-D), \quad \text{and} \quad E[S'(r|w)] = -\frac{ns_1s_2}{r}D, \tag{A.10}$$

where

$$D \equiv D(p_1, p_2) = \int_{\Omega_1 \cap \Omega_2} \frac{p_1 p_2}{s_1 p_1 + s_2 p_2}\, dw. \tag{A.11}$$

Substituting (A.10) – (A.11) into (A.9) completes our proof.

Proof of $\hat{r}_M = \hat{r}_O$ when $s_i > 0,\ i = 1, 2$. By analogy to $\hat{r}_O, \hat{r}_M$, the limit of $\hat{r}_M^{(t+1)}$ of (4.2), is a root of the following function

$$S_M(r|w) = \sum_{j=1}^{n_1} \frac{q_1(w_{1j}) - rq_2(w_{1j})}{s_1 q_1(w_{1j}) + s_2 r q_2(w_{1j})} + \sum_{j=1}^{n_2} \frac{q_1(w_{2j}) - rq_2(w_{2j})}{s_1 q_1(w_{2j}) + s_2 r q_2(w_{2j})}. \tag{A.12}$$

When $s_i > 0, i = 1, 2$, it is easy to see that

$$\sum_{j=1}^{n_1} \frac{q_1(w_{1j})}{s_1 q_1(w_{1j}) + s_2 r q_2(w_{1j})} = n - \frac{1}{s_1} \sum_{j=1}^{n_1} \frac{s_2 r q_2(w_{1j})}{s_1 q_1(w_{1j}) + s_2 r q_2(w_{1j})} \tag{A.13}$$

and

$$\sum_{j=1}^{n_2} \frac{r q_2(w_{2j})}{s_1 q_1(w_{2j}) + s_2 r q_2(w_{2j})} = n - \frac{1}{s_2} \sum_{j=1}^{n_2} \frac{s_1 q_1(w_{2j})}{s_1 q_1(w_{2j}) + s_2 r q_2(w_{2j})}. \tag{A.14}$$

Combining (A.12) – (A.14) with (A.5) gives $S_M(r|w) = S(r|w)/(s_1 s_2)$, which implies $\hat{r}_M = \hat{r}_O$ because $\hat{r}_O$ is the unique root of $S(r|w)$.

Proof (6.7) - (6.8). Given $p_1 = N(0, 1), p_2 = N(\mu, 1)$ $(\mu \neq 0)$ and s_1, with the variable transformation $y = \mu(w - \mu/2)$ we have from (3.2)

$$RE^2_\mu(\hat{r}_P(k, A)) = \frac{1}{s_1 s_2}\Big\{|\mu| \exp(\frac{\mu^2}{8})\sqrt{2\pi} \frac{\int_{-\infty}^{+\infty} f(y|k, A) \exp(\frac{-y^2}{2\mu^2})dy}{\Big[\int_{-\infty}^{+\infty} g(y|k, A) \exp(\frac{-y^2}{2\mu^2})dy\Big]^2} - 1\Big\}, \tag{A.15}$$

where

$$f(y|k, A) = \frac{\exp(\frac{y}{2})[s_1 + s_2 \exp(y)]}{\Big[1 + \rho^{\frac{1}{k}}(A) \exp(\frac{y}{k})\Big]^{2k}} \quad \text{and} \quad g(y|k, A) = \frac{\exp(\frac{y}{2})}{\Big[1 + \rho^{\frac{1}{k}}(A) \exp(\frac{y}{k})\Big]^{k}} \tag{A.16}$$

with $\rho(A) = (As_2)/(A_O s_1)$. Letting $x = \rho^{\frac{1}{k}}(A)\exp(\frac{y}{k})$, we have

$$\begin{aligned}
&\lim_{|\mu|\to+\infty}\int_{-\infty}^{+\infty} f(y|k,A)\exp(\frac{-y^2}{2\mu^2})dy = \int_{-\infty}^{+\infty} f(y|k,A)dy \\
&= \frac{s_1 k}{\rho^{1/2}(A)}\int_0^{+\infty}\frac{x^{\frac{k}{2}-1}}{(1+x)^{2k}}dx + \frac{s_2 k}{\rho^{3/2}(A)}\int_0^{+\infty}\frac{x^{\frac{3k}{2}-1}}{(1+x)^{2k}}dx \\
&= \frac{k}{\rho^{1/2}(A)}(s_1 + \frac{s_2}{\rho(A)})B(\frac{3k}{2},\frac{k}{2}),
\end{aligned} \tag{A.17}$$

and similarly

$$\lim_{|\mu|\to+\infty}\int_{-\infty}^{+\infty} g(y|k,A)\exp(\frac{-y^2}{2\mu^2})dy = \int_{-\infty}^{+\infty} g(y|k,A)dy = \frac{k}{\rho^{1/2}(A)}B(\frac{k}{2},\frac{k}{2}). \tag{A.18}$$

Substituting (A.17) and (A.18) into

$$\lim_{|\mu|\to+\infty}\frac{RE_\mu(\hat{r}_P(k,A))}{RE_\mu(\hat{r}_P(1,A_O))} = \Big[\frac{\int_{-\infty}^{+\infty} f(y|k,A)dy}{\int_{-\infty}^{+\infty} f(y|1,A_O)dy}\Big]^{\frac{1}{2}}\frac{\int_{-\infty}^{+\infty} g(y|1,A_O)dy}{\int_{-\infty}^{+\infty} g(y|k,A)dy}$$

yields (6.7) - (6.8).

Proof of (6.9). Letting $z = y/\mu$, we can rewrite (A.15) as

$$RE_\mu^2(\hat{r}_P(k,A)) = \frac{1}{s_1 s_2}\Big\{\exp(\frac{\mu^2}{8})\frac{E[f(\mu z|k,A)]}{E^2[g(\mu z|k,A)]} - 1\Big\}, \tag{A.19}$$

where all the expectations are with respect to $z \sim N(0,1)$. It then follows that

$$\begin{aligned}
\lim_{\mu\to 0}\frac{RE_\mu(\hat{r}_P(k,A))}{RE_\mu(\hat{r}_P(1,A_O))} &= \Big\{\lim_{\mu\to 0}\frac{\exp(\frac{\mu^2}{8})E[f(\mu z|k,A)] - E^2[g(\mu z|k,A)]}{\exp(\frac{\mu^2}{8})E[f(\mu z|1,A_O)] - E^2[g(\mu z|1,A_O)]}\Big\}^{\frac{1}{2}} \\
&\quad\times\lim_{\mu\to 0}\frac{E[g(\mu z|1,A_O)]}{E[g(\mu z|k,A)]}.
\end{aligned} \tag{A.20}$$

It is easy to check that

$$\lim_{\mu\to 0}\frac{E[g(\mu z|1,A_O)}{E[g(\mu z|k,A)]} = \frac{g(0|1,A_O)}{g(0|k,A)} = \frac{1}{s_1}(1+\rho^{\frac{1}{k}}(A))^k \equiv \frac{1}{s_1}p_{k,A}^{-k}. \tag{A.21}$$

Using L'Hôpital's rule twice, we obtain

$$\begin{aligned}
&\lim_{\mu\to 0}\frac{\exp(\frac{\mu^2}{8})E[f(\mu z|k,A)] - E^2[g(\mu z|k,A)]}{\exp(\frac{\mu^2}{8})E[f(\mu z|1,A_O)] - E^2[g(\mu z|1,A_O)]} \\
&= \frac{f(0|k,A) + 4f''(0|k,A) - 8g(0|k,A)g''(0|k,A)}{f(0|1,A_O) + 4f''(0|1,A_O) - 8g(0|1,A_O)g''(0|1,A_O)}.
\end{aligned} \tag{A.22}$$

By taking derivatives of $\log[f(y|k,A)]$ and $\log[g(y|k,A)]$ with respect to y and then evaluating them at $y=0$, we obtain easily

$$f(0|k,A)=p_{k,A}^{2k}, \quad f''(0|k,A)=p_{k,A}^{2k}[(2p_{k,A}-1.5+s_2)^2+s_1s_2-2p_{k,A}(1-p_{k,A})/k], \tag{A.23}$$

and

$$g(0|k,A)=p_{k,A}^{k}, \quad g''(0|k,A)=p_{k,A}^{k}[(p_{k,A}-0.5)^2-p_{k,A}(1-p_{k,A})/k], \tag{A.24}$$

where $p_{k,A}$ is defined in (A.21). Now substituting (A.23)-(A.24) into (A.22) and then substituting (A.22) together with (A.21) into (A.20) yields (6.9).

References

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **22**, 245-268.

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

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46**, 443-459.

Chen, M.-H. and Shao, Q. M. (1994). On Monte Carlo methods for estimating ratios of normalizing constants. Research Report 627, Department of Mathematics, National University of Singapore.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1996). Computing Bayes factors by combining simulation and asymptotic approximations. *J. Amer. Statist. Assoc.*, to appear.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser.B* **56**, 501-514.

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

Gelman, A. and Meng, X. L. (1994). Path sampling for computing normalizing constants: identities and theory. Technical Report 376, Department of Statistics, University of Chicago.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7**, 457-511.

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

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

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser.B* **54**, 657-699.

Green, P. J. (1992). Discussion of "Constrained Monte Carlo maximum likelihood for dependent data" by Geyer and Thompson *J. Roy. Statist. Soc. Ser.B* **54**, 683-684.

Irwin, M., Cox, N. and Kong, A. (1994). Sequential imputation for Multilocus linkage analysis. *Proc. Nat. Acad. Sci. U.S.A.* **91**, 11684-11688.

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

Meng, X. L. (1994). Multiple imputation inference under uncongenial sources of input (with discussion). *Statist. Science* **9**, 538-573.

Meng, X. L. and Schilling, S. (1996). Fitting full-information factor models and an empirical investigation of bridge sampling. *J. Amer. Statist. Assoc.* to appear.

Meng, X. L. and Schilling, S. (1996a). Bridge sampling after transformation. Technical Report 432, Department of Statistics, University of Chicago.

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

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

Ogata, Y. (1989). A Monte Carlo method for high dimensional integration. *Numer. Math.* **55**, 137-157.

Ogata, Y. (1990). A Monte Carlo method for an objective Bayesian procedure. *Ann. Inst. Statist. Math.* **42**, 403-433.

Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *Amer. J. Hum. Genet* **31**, 161-175.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs Sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Statist. Soc. Ser.B* **55**, 3-23.

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

Torrie, G. M. and Valleau, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J. Comput. Phys.* **23**, 187-199.

Voter, A. F. (1985). A Monte Carlo method for determining free-energy differences and transition state theory rate constants. *J. Chem. Phys.* **82**, 1890-1899.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85**, 699-704.

Wu, C. F. J (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11,** 95-103.

Department of Statistics, University of Chicago, Chicago, IL 60637, U.S.A.

Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong.

(Received March 1994; accepted February 1996)