Statistica Sinica **20** (2010), 537-560

ESTIMATION OF LARGE FAMILIES OF BAYES FACTORS FROM MARKOV CHAIN OUTPUT

Hani Doss

University of Florida

Abstract: We consider situations in Bayesian analysis where the prior is indexed by a hyperparameter taking on a continuum of values. We distinguish some arbitrary value of the hyperparameter, and consider the problem of estimating the Bayes factor for the model indexed by the hyperparameter vs. the model specified by the distinguished point, as the hyperparameter varies. We assume that we have Markov chain output from the posterior for a finite number of the priors, and develop a method for efficiently computing estimates of the entire family of Bayes factors. As an application of the ideas, we consider some commonly used hierarchical Bayesian models and show that the parametric assumptions in these models can be recast as assumptions regarding the prior. Therefore, our method can be used as a model selection criterion in a Bayesian framework. We illustrate our methodology through a detailed example involving Bayesian model selection.

Key words and phrases: Bayes factors, control variates, ergodicity, importance sampling, Markov chain Monte Carlo.

1. Introduction

Suppose we have a data vector Y whose distribution has density p_θ, for some unknown $\theta \in \Theta$. Let $\{\nu_h,\, h \in \mathcal{H}\}$ be a family of prior densities on θ that we are contemplating. The selection of a particular prior from the family is important in Bayesian data analysis, and when making this choice one will often want to consider the marginal likelihood of the data under the prior ν_h, given by $m_h(y) = \int \ell_y(\theta)\nu_h(\theta)\, d\theta$, as h varies over the hyperparameter space $\mathcal{H}$. Here, $\ell_y(\theta) = p_\theta(y)$ is the likelihood function. Values of h for which $m_h(y)$ is relatively low may be considered poor choices, and consideration of the family $\{m_h(y),\, h \in \mathcal{H}\}$ may be helpful in narrowing the search of priors to use. It is therefore useful to have a method for computing the family $\{m_h(y),\, h \in \mathcal{H}\}$. For the purpose of model selection, if c is a fixed constant, the information given by $\{m_h(y),\, h \in \mathcal{H}\}$ and $\{c\, m_h(y),\, h \in \mathcal{H}\}$ is the same. From a computational and statistical point of view however, it is usually easier to fix a particular hyperparameter value h_* and focus on $\{m_h(y)/m_{h_*}(y),\, h \in \mathcal{H}\}$. Given two hyperparameter values h and h_*, the quantity $B(h, h_*) = m_h/m_{h_*}$ is called the Bayes factor of the model indexed by h vs. the model indexed by h_* (we write m_h instead of $m_h(y)$ from now on).

In this paper we present a method for estimating the family $\{B(h, h_*),\ h \in \mathcal{H}\}$. We have in mind situations where $B(h, h_*)$ cannot be obtained analytically and, moreover, we need to calculate $B(h, h_*)$ for a large set of h's, so that computational efficiency is essential. Our approach requires that there are k hyperparameter values $h_1, \ldots, h_k$, and for $l = 1, \ldots, k$, we are able to get a sample $\theta_i^{(l)}$, $i = 1, \ldots, n_l$, from $\nu_{h_l,y}$, the posterior density of θ given $Y = y$, assuming that the prior is ν_{h_l}.

To set the framework, consider the trivial case where $k = 1$, and we have a sample from the posterior $\nu_{h_1,y}$ generated by an ergodic Markov chain. Our objective is to estimate $\{B(h, h_1),\ h \in \mathcal{H}\}$. For any h such that $\nu_h(\theta) = 0$ whenever $\nu_{h_1}(\theta) = 0$, we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\frac{\nu_h(\theta_i)}{\nu_{h_1}(\theta_i)} &\to \int \frac{\nu_h(\theta)}{\nu_{h_1}(\theta)}\nu_{h_1,y}(\theta)\,d\theta \qquad (1.1)\\ &= \frac{m_h}{m_{h_1}}\int \frac{\ell_y(\theta)\nu_h(\theta)/m_h}{\ell_y(\theta)\nu_{h_1}(\theta)/m_{h_1}}\nu_{h_1,y}(\theta)\,d\theta \\ &= \frac{m_h}{m_{h_1}}\int \frac{\nu_{h,y}(\theta)}{\nu_{h_1,y}(\theta)}\nu_{h_1,y}(\theta)\,d\theta \quad = \quad \frac{m_h}{m_{h_1}}. \end{aligned}$$

Therefore, the left side of (1.1) is a consistent estimate of the Bayes factor $B(h, h_1)$.

To fix ideas, consider as a simple example the following standard three-level hierarchical model:

$$\text{conditional on } \psi_j, \qquad Y_j \overset{\text{indep}}{\sim} \phi_{\psi_j,\sigma_j}, \quad j = 1, \ldots, m \qquad (1.2a)$$

$$\text{conditional on } \mu, \tau, \qquad \psi_j \overset{\text{i.i.d.}}{\sim} \phi_{\mu,\tau}, \quad j = 1, \ldots, m \qquad (1.2b)$$

$$(\mu, \tau) \sim \lambda_{c_1,c_2,c_3,c_4}, \qquad (1.2c)$$

where $\phi_{m,s}$ denotes the density of the normal distribution with mean m and standard deviation s. In (1.2a), the σ_i's are assumed known. In (1.2c), $\lambda_{c_1,c_2,c_3,c_4}$ is the normal / inverse gamma distribution indexed by four hyperparameters (see Section 3). This is a very commonly used model but, as we discuss later, in some situations it is preferable to replace (1.2b) with $\psi_j \overset{\text{i.i.d.}}{\sim} t_{v,\mu,\tau}$, where $t_{v,\mu,\tau}$ is the density of the t distribution with v degrees of freedom, location μ and scale τ. In this case, consider now the estimate in the left side of (1.1). The likelihood of (μ, τ) is

$$\ell_Y(\mu, \tau) = \int\!\cdots\!\int \prod_{j=1}^{m} \phi_{\psi_j,\sigma_j}(Y_j) \prod_{j=1}^{m} t_{v,\mu,\tau}(\psi_j)\,d\psi_1 \ldots d\psi_m.$$

This likelihood cannot be computed in closed form, and therefore its cancellation in (1.1) gives a non-trivial simplification: calculation of the estimate requires only the ratio of the densities of the *priors* and not the posteriors.

Consider (1.2) with $t_{v,\mu,\tau}$ instead of $\phi_{\mu,\tau}$ in the middle stage, and suppose now that we would like to select v, with the choice $v = \infty$ signifying the choice of the normal distribution $\phi_{\mu,\tau}$. The distribution of Y is determined by $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)$. A completely equivalent way of describing the model is therefore through the two-level hierarchy in which we let $\theta = (\boldsymbol{\psi}, \mu, \tau)$, and stipulate:

$$\begin{aligned} \text{conditional on } \theta, \qquad Y_j &\overset{\text{indep}}{\sim} \phi_{\psi_j,\sigma_j}, \quad j = 1, \ldots, m \\ (\boldsymbol{\psi}, \mu, \tau) &\sim \nu_h, \end{aligned}$$

where $\nu_h(\boldsymbol{\psi}, \mu, \tau) = \left(\prod_{j=1}^m t_{v,\mu,\tau}(\psi_j)\right) \lambda_{c_1,c_2,c_3,c_4}(\mu, \tau)$. Here, the hyperparameter is $h = (v, c_1, c_2, c_3, c_4)$, which includes the number of degrees of freedom. Estimation of the family of Bayes factors $\{B(h, h_1),\, h \in \mathcal{H}\}$ therefore enables a model selection step.

We now discuss briefly the accuracy of the estimate on the left side of (1.1). When ν_h is nearly singular with respect to ν_{h_1} over the region where the θ_i's are likely to be, the estimate will be unstable. (Formally, the estimate will satisfy a central limit theorem if the chain mixes fast enough and the random variable $\nu_h(\theta)/\nu_{h_1}(\theta)$ (where $\theta \sim \nu_{h_1,y}$) has a high enough moment. This is discussed in more detail in Section 2.3.) From a practical point of view, this means that there is effectively a "radius" around h_1 within which one can safely move.

In all but the very simplest models, the dimension of $\mathcal{H}$ is greater than 1, and therefore estimation of the Bayes factor as h ranges over $\mathcal{H}$ raises serious computational difficulties, and it is essential that for each h, the estimate of $B(h, h_1)$ is both accurate and can be computed quickly. Our approach is to select k hyperparameter points $h_1, \ldots, h_k$, and get Markov chain samples from $\nu_{h_l,y}$ for each $l = 1, \ldots, k$. The prior ν_{h_1} in the denominator of the left side of (1.1) is replaced by a mixture $w_1\nu_{h_1} + \cdots + w_k\nu_{h_k}$, with appropriately chosen weights. We show how judiciously chosen control variates can be used in conjunction with multiple Markov chain streams to produce accurate estimates even with small samples, so that the net result is a computationally feasible method for producing reliable estimates of the Bayes factors for a wide range of hyperparameter values. Our approach is motivated by and uses ideas developed in Kong, McCullagh, Meng, Nicolae and Tan (2003), which deals with the situation where we have independent samples from k unnormalized densities, and we wish to estimate all possible ratios of the k normalizing constants. Owen and Zhou (2000) and Tan (2004) also discuss the use of control variates to increase the accuracy of Monte Carlo estimates. In Section 4 we return to these three papers and discuss in detail how

our approach fits in the context of this work. The paper is organized as follows. Section 2 contains the main methodological development; there, we present our method for estimating the family of Bayes factors and state supporting theoretical results. Section 3 illustrates the methodology through a detailed example that involves a number of issues, including selection of the parametric family in the model. Section 4 gives a discussion of other possible approaches and related work, and the Appendix gives the proof of the main theoretical result of the paper.

2. Estimation of the Family of Bayes Factors

Suppose that for $l = 1, \ldots, k$, we have Markov chain Monte Carlo (MCMC) samples $\theta_i^{(l)}$, $i = 1, \ldots, n_l$ from the posterior density of θ given $Y = y$, assuming that the prior is ν_{h_l}, having the form

$$\nu_{h_l,y}(\theta) = \frac{\ell_y(\theta)\nu_{h_l}(\theta)}{m_{h_l}}.$$

We assume that the k sequences are independent of one another.

We will not assume we know any of the m_{h_l}'s. However, we now explain how knowledge of the Bayes factors m_{h_l}/m_{h_1}, for $l = 2, \ldots, k$ would result in two important benefits. If we knew these Bayes factors we could then form the estimate

$$\hat{B}(h, h_1) = \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} n_s \nu_{h_s}(\theta_i^{(l)}) m_{h_1}/m_{h_s}}. \tag{2.1}$$

Let $n = \sum_{s=1}^{k} n_s$, and assume that $n_s/n \to a_s$, $s = 1, \ldots, k$. We then have

$$\begin{aligned}
\hat{B}(h, h_1) &= \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{\ell_y(\theta_i^{(l)})\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} n_s \ell_y(\theta_i^{(l)})\nu_{h_s}(\theta_i^{(l)}) m_{h_1}/m_{h_s}} \\
&= \frac{m_h}{m_{h_1}} \sum_{l=1}^{k} \frac{1}{n_l} \sum_{i=1}^{n_l} \frac{\frac{n_l}{n}\nu_{h,y}(\theta_i^{(l)})}{\sum_{s=1}^{k} \frac{n_s}{n}\nu_{h_s,y}(\theta_i^{(l)})} \\
&\xrightarrow{\text{a.s.}} \frac{m_h}{m_{h_1}} \sum_{l=1}^{k} \int \frac{a_l \nu_{h,y}(\theta)}{\sum_{s=1}^{k} a_s \nu_{h_s,y}(\theta)} \nu_{h_l,y}(\theta)\, d\theta \;=\; \frac{m_h}{m_{h_1}}.
\end{aligned} \tag{2.2}$$

The almost sure convergence in (2.2) occurs under minimal conditions on the Markov chains $\theta_i^{(l)}$, $i = 1, \ldots, n_l$. Asymptotic normality requires more restrictive conditions, and is discussed in Section 2.3. To compute $\hat{B}(h, h_1)$, the n quantities $\sum_{s=1}^{k} n_s \nu_{h_s}(\theta_i^{(l)}) m_{h_1}/m_{h_s}$ are calculated once, and stored. Then, for every new

value of h, the computation of $\hat{B}(h, h_1)$ requires taking n ratios and a sum. Since this is to be done for a large number of h's, it is essential that for each l, the sequence $\theta_i^{(l)}$, $i = 1, \ldots, n_l$ be as independent as possible, so that the value of n be made as small as possible.

We now briefly recall the use of control variates in Monte Carlo sampling. Suppose we wish to estimate the expected value of a random variable Y, and we can find a random variable Z that is correlated with Y, and such that $E(Z)$ is known (without loss of generality, $E(Z) = 0$). Then for any β, the estimate $Y - \beta Z$ is an unbiased estimate of $E(Y)$, and the value of β minimizing the variance of $Y - \beta Z$ is $\beta = \mathrm{Cov}\,(Y, Z)/\mathrm{Var}\,(Z)$. The idea may be used when there are several variables $Z_1, \ldots, Z_r$ that are correlated with Y.

In the present context, we may consider the functions

$$Z_j(\theta) = \frac{\nu_{h_j}(\theta) m_{h_1}/m_{h_j} - \nu_{h_1}(\theta)}{\sum_{s=1}^{k} (n_s/n) \nu_{h_s}(\theta) m_{h_1}/m_{h_s}}, \qquad j = 2, \ldots, k,$$

whose expectations under $\sum_{s=1}^{k} (n_s/n) \nu_{h_s, y}$ are 0. The calculation of these functions requires knowledge of the Bayes factors m_{h_s}/m_{h_1}, $s = 2, \ldots, k$.

The method proposed in this paper can now be briefly summarized as follows.

1. For each $l = 1, \ldots, k$, get Markov chain samples $\theta_i^{(l)}$, $i = 1, \ldots, N_l$ from $\nu_{h_l, y}$. Based on these, the Bayes factors m_{h_s}/m_{h_1}, $s = 2, \ldots, k$ are estimated. The sample sizes N_l should be very large, so that these estimates are very accurate.
2. For each $l = 1, \ldots, k$, we obtain new samples $\theta_i^{(l)}$, $i = 1, \ldots, n_l$ from $\nu_{h_l, y}$. Using these, together with the Bayes factors computed in Step 1 we form the estimate $\hat{B}_{\mathrm{reg}}(h, h_1)$, which is similar to (2.1), except that we use the functions Z_j, $j = 2, \ldots, k$ as control variates.

The samples in the two steps are used for different purposes. Those in Step 1 are used solely to estimate m_{h_s}/m_{h_1}, $s = 2, \ldots, k$, and in fact, once these estimates are formed, the samples may be discarded. The samples in Step 2 are used to estimate the family $B(h, h_1)$. On occasion, special analytical structure enables the use of numerical methods to estimate m_{h_s}/m_{h_1}, $s = 2, \ldots, k$, as long as k is not too large—so Step 1 is bypassed. A review of the literature for this approach is given in Kass and Raftery (1995). Ideally, the samples in Step 2 should be independent or nearly so, which may be accomplished by subsampling a very long chain. If we have a Markov transition function that gives rise to a uniformly ergodic chain, it is possible to use this Markov transition function to obtain perfect samples (Hobert and Robert (2004)), although the time it takes to generate a perfect sample of length n_l may be much greater than the time to generate the Markov chain of length n_l.

One may ask what is the point of having two steps of sampling, i.e., why not just use the samples from Step 1 for both estimation of m_{h_s}/m_{h_1}, $s = 1, \ldots, k$, and for subsequent estimation of the family $B(h, h_1)$. The reason for having the two stages is that the estimate of $B(h, h_1)$ needs to be computed for a large number of h's, and for every h the amount of computation is linear in n, so this precludes a large value of n. Therefore, given that a relatively modest sample size must be used, we need to reduce the variance of the estimate as much as possible, and this is the reason for carrying out Step 1. The amount of computation to generate the Step 1 samples is typically one or two orders of magnitude less than the amount of computation needed to calculate the estimates of $B(h, h_1)$ from the Step 2 samples (see the discussion at the end of Section 3).

To summarize, the benefit of the two-step approach is a better tradeoff between statistical efficiency and computational time. To see this, it is helpful to consider a very simple example in which the variances of various estimators can actually be computed. Consider the unnormalized density $q_h = t^h I(t \in (0, 1))$, and let m_h be the normalizing constant. Now suppose we wish to estimate m_h/m_1 as h ranges over a grid of $4,000$ points in the interval $(1.5, 2.5)$ and that we are able to generate i.i.d. observations from q_1/m_1 and q_3/m_3. We may use the estimator in Kong et al. (2003) (discussed later in this paper), which estimates both m_h/m_1 and m_3/m_1 from the same sample. Given one minute of computer time, using the machine whose specifications are described in Section 3, the requirement that we calculate such a large number of ratios of normalizing constants limits the total sample size to $n = 2 \times 90$. A formula for the asymptotic variance $\rho^2(h)$ of the Kong et al. (2003) estimate is given in Tan (2004, equation (8)), and in this situation all quantities that are needed in the formula are available explicitly. Now if we take the minute and divide it into two parts, 3 seconds and 57 seconds, then with the 3 seconds we can estimate m_3/m_1 with essentially perfect accuracy, and with the remaining 57 seconds, if we use the estimate $\hat{B}(h, 1)$, we can handle a sample size of $n \times 57/60$. A formula for the asymptotic variance $\tau^2(h)$ of this estimator—which uses the value of m_3/m_1 calculated in the first stage—is given in Theorem 1 of the present paper, and can also be evaluated explicitly. The ratio $\tau^2(h)/\rho^2(h)$ is bounded above by 0.2 over the entire grid, and so with the same computer resources, the variance of the two-stage estimator is uniformly at most $0.2 \times 60/57 \approx .21$ that of the one-stage estimator. (The gains if we use $\hat{B}_{\text{reg}}$ instead of $\hat{B}$ can be far greater; see Section 3 for an illustration.)

In Section 2.1 we show how the MCMC approach to Step 1 may be implemented. In Section 2.2 we show how estimation in Step 2 may be implemented, and also discuss the benefits of using the control variates. In Section 2.3 we give a result regarding asymptotic normality of the estimates of the Bayes factors.

2.1. Estimation of the Bayes Factors m_{h_s}/m_{h_1}

We now assume that for $l = 1, \ldots, k$, we have a sequence $\theta_i^{(l)}$, $i = 1, \ldots, N_l$ from a Markov chain corresponding to the posterior $\nu_{h_l,y}$. Also, these k sequences are independent of one another. Let $N = \sum_{l=1}^{k} N_l$, and $a_l = N_l/N$. We wish to estimate m_{h_l}/m_{h_1}, $l = 2, \ldots, k$.

Meng and Wong (1996) considered this problem and, to understand their method, it is helpful to consider first the case where $k = 2$ and we wish to estimate $d = m_{h_2}/m_{h_1}$. For any function α defined on the common support of $\nu_{h_1,y}$ and $\nu_{h_2,y}$ such that $\int \alpha(\theta)\nu_{h_1}(\theta)\ell_y(\theta)\nu_{h_2}(\theta)\,d\theta < \infty$, we have

$$\frac{\int \alpha(\theta)\nu_{h_2}(\theta)\nu_{h_1,y}(\theta)\,d\theta}{\int \alpha(\theta)\nu_{h_1}(\theta)\nu_{h_2,y}(\theta)\,d\theta} = \frac{(1/m_1)\int \alpha(\theta)\nu_{h_2}(\theta)\ell_y(\theta)\nu_{h_1}(\theta)\,d\theta}{(1/m_2)\int \alpha(\theta)\nu_{h_1}(\theta)\ell_y(\theta)\nu_{h_2}(\theta)\,d\theta} = \frac{m_{h_2}}{m_{h_1}}.$$

Therefore,

$$\hat{d} = \frac{\sum_{i=1}^{N_1} \alpha(\theta_i^{(1)})\nu_{h_2}(\theta_i^{(1)}) \Big/ N_1}{\sum_{i=1}^{N_2} \alpha(\theta_i^{(2)})\nu_{h_1}(\theta_i^{(2)}) \Big/ N_2} \tag{2.3}$$

is a consistent estimate of d, under the minimal assumption of ergodicity of the two chains.

Meng and Wong (1996) show that when $\{\theta_i^{(j)}\}_{i=1}^{N_j}$ are independent draws from $\nu_{h_j,y}$, the optimal α to use is

$$\alpha_{\text{opt}}(\theta) = \frac{1}{a_1\nu_{h_1}(\theta) + a_2\nu_{h_2}(\theta)/d}, \tag{2.4}$$

which involves the quantity we wish to estimate. This suggests the iterative scheme

$$\hat{d}^{(t+1)} = \frac{\sum_{i=1}^{N_1} [\nu_{h_2}(\theta_i^{(1)})/(a_1\nu_{h_1}(\theta_i^{(1)}) + a_2\nu_{h_2}(\theta_i^{(1)})/\hat{d}^{(t)})] \Big/ N_1}{\sum_{i=1}^{N_2} [\nu_{h_1}(\theta_i^{(2)})/(a_1\nu_{h_1}(\theta_i^{(2)}) + a_2\nu_{h_2}(\theta_i^{(2)})/\hat{d}^{(t)})] \Big/ N_2}, \tag{2.5}$$

for $t = 1, 2, \ldots$.

For the general case where $k \geq 2$, let $\boldsymbol{d} = (m_{h_2}/m_{h_1}, \ldots, m_{h_k}/m_{h_1})$, but it is more convenient to work with the vector of component-wise reciprocals of $\boldsymbol{d}$, call it $\boldsymbol{r}$. For $i = 2, \ldots, k$, and $j = 1, \ldots, k$, $j \neq i$, let α_{ij} be known functions defined

on the common support of ν_{h_i} and ν_{h_j} satisfying $\int \alpha_{ij}(\theta)\nu_{h_i}(\theta)\ell_y(\theta)\nu_{h_j}(\theta)\,d\theta < \infty$. Let

$$\begin{aligned} b_{ii} &= \textstyle\sum_{j\neq i} E_{\nu_{h_j},y}\big(\alpha_{ij}(\theta)\nu_{h_i}(\theta)\big) \qquad && 2 \le i \le k, \\ b_{ij} &= E_{\nu_{h_i},y}\big(\alpha_{ij}(\theta)\nu_{h_j}(\theta)\big) && i \neq j, \end{aligned} \tag{2.6}$$

and

$$\boldsymbol{B} = \begin{pmatrix} b_{22} & -b_{23} & \dots & -b_{2k} \\ -b_{32} & b_{33} & \dots & -b_{3k} \\ \vdots & \vdots & \ddots & \vdots \\ -b_{k2} & -b_{k3} & \dots & b_{kk} \end{pmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} b_{21} \\ b_{31} \\ \vdots \\ b_{k1} \end{pmatrix}.$$

Then assuming that $\boldsymbol{B}$ is nonsingular, we have $\boldsymbol{r} = \boldsymbol{B}^{-1}\boldsymbol{b}$. If $\hat{\boldsymbol{B}}_{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{b}}_{\boldsymbol{\alpha}}$ are the natural estimates of $\boldsymbol{B}$ and $\boldsymbol{b}$ based on the functions α_{ij} and the samples $\{\theta_i^{(j)}\}_{i=1}^{N_j}$, $j = 1, \dots, k$, then $\boldsymbol{r}$ may be estimated via

$$\hat{\boldsymbol{r}} = \hat{\boldsymbol{B}}_{\boldsymbol{\alpha}}^{-1}\hat{\boldsymbol{b}}_{\boldsymbol{\alpha}}. \tag{2.7}$$

Meng and Wong (1996) consider the functions

$$\alpha_{ij} = \frac{a_i a_j}{\sum_{s=1}^{k} a_s r_s \nu_{h_s}}, \tag{2.8}$$

which involve the unknown $\boldsymbol{r}$. The natural extension of (2.5) is $\hat{\boldsymbol{r}}^{(t+1)} = \hat{\boldsymbol{B}}_{\boldsymbol{\alpha_t}}^{-1}\hat{\boldsymbol{b}}_{\boldsymbol{\alpha_t}}$, with the vector of functions $\boldsymbol{\alpha}_t$ given by (2.8), where we use $\hat{\boldsymbol{r}}^{(t)}$ instead of $\boldsymbol{r}$.

2.2. Using control variates

The use of control variates has had many successes in Monte Carlo sampling, and a particularly important paper is Owen and Zhou (2000). This paper considers the use of control variates in conjunction with importance sampling, when the importance sampling density is a mixture, and the paper motivates some of the ideas below.

We now assume that we have samples $\theta_i^{(l)}$, $i = 1, \dots, n_l$, from $\nu_{h_l,y}$, $l = 1, \dots, k$, with independence across samples, and that we know the constants $d_2, \dots, d_k$. For unity of notation, we define $d_1 = 1$. As before $n = \sum_{l=1}^{k} n_l$ and $n_l/n = a_l$. The estimate $\hat{B}(h, h_1)$ in (2.1) is an average of n draws from the mixture distribution $p_{\boldsymbol{a}} = \sum_{s=1}^{k} a_s \nu_{h_s,y}$. However, these are not independent and identically distributed since they form a stratified sample: we have exactly n_s draws from $\nu_{h_s,y}$, $s = 1, \dots, k$, a fact which causes no problems.

We wish to estimate the integral

$$I_h = \int \frac{\ell_y(\theta)\nu_h(\theta)}{m_{h_1}}\,d\theta = B(h, h_1).$$

Define the functions

$$H_j(\theta) = \frac{\ell_y(\theta)\nu_{h_j}(\theta)}{m_{h_j}} - \frac{\ell_y(\theta)\nu_{h_1}(\theta)}{m_{h_1}}, \qquad j = 2, \ldots, k.$$

We have

$$\int H_j(\theta)\, d\theta = 0, \qquad \text{or equivalently} \qquad E_{p_{\boldsymbol{a}}}\Big(\frac{H_j(\theta)}{p_{\boldsymbol{a}}(\theta)}\Big) = 0,$$

where the subscript indicates that the expectation is taken with respect to the mixture distribution $p_{\boldsymbol{a}}$. Therefore, for every $\boldsymbol{\beta} = (\beta_2, \ldots, \beta_k)$ the estimate

$$\hat{I}_{h,\boldsymbol{\beta}} = \frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n_l} \frac{\ell_y(\theta_i^{(l)})\frac{\nu_h(\theta_i^{(l)})}{m_{h_1}} - \sum_{j=2}^{k}\beta_j\big[\ell_y(\theta_i^{(l)})\big(\frac{\nu_{h_j}(\theta_i^{(l)})}{m_{h_j}} - \frac{\nu_{h_1}(\theta_i^{(l)})}{m_{h_1}}\big)\big]}{\sum_{s=1}^{k} a_s \nu_{h_s,y}(\theta_i^{(l)})}$$

is unbiased. As written, this estimate is not computable, because it involves the normalizing constants m_{h_j}, which are unknown, and also the likelihood $\ell_y(\theta)$, which may not be available. We rewrite it in computable form as

$$\hat{I}_{h,\boldsymbol{\beta}} = \frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n_l} \frac{\nu_h(\theta_i^{(l)}) - \sum_{j=2}^{k}\beta_j\big[\nu_{h_j}(\theta_i^{(l)})/d_j - \nu_{h_1}(\theta_i^{(l)})\big]}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/d_s}. \tag{2.9}$$

We would like to use the value of $\boldsymbol{\beta}$, call it $\boldsymbol{\beta}_{\text{opt}}$, that minimizes the variance of $\hat{I}_{h,\boldsymbol{\beta}}$, but this $\boldsymbol{\beta}_{\text{opt}}$ is generally unknown. As in Owen and Zhou (2000), we can do ordinary linear regression of $Y_{i,l}^{(h)}$ on predictors $Z_{i,l}^{(j)}$, where

$$Y_{i,l}^{(h)} = \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/d_s}, \qquad Z_{i,l}^{(j)} = \frac{\nu_{h_j}(\theta_i^{(l)})/d_j - \nu_{h_1}(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/d_s}, \quad j = 2, \ldots, k, \tag{2.10}$$

and all required quantities are available. We then use the least squares estimate $\hat{\boldsymbol{\beta}}$, i.e., the estimate of I_h is $\hat{I}_{h,\hat{\boldsymbol{\beta}}}$. It is easy to see that $\hat{I}_{h,\hat{\boldsymbol{\beta}}}$ is simply $\hat{\beta}_0$, the estimate of the intercept term in the bigger regression problem where we include the intercept term, i.e.,

$$\hat{I}_{h,\hat{\boldsymbol{\beta}}} = \hat{\beta}_0. \tag{2.11}$$

One can show that if the k sequences are all i.i.d. sequences, then $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}_{\text{opt}}$, and $\hat{I}_{h,\hat{\boldsymbol{\beta}}}$ is guaranteed to be at least as efficient as the naive estimator $\hat{B}(h, h_1)$. But when we have Markov chains this is not the case, especially if the chains mix at different rates. In Section 2.3 we consider the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{I}_{h,\hat{\boldsymbol{\beta}}}$ directly. In particular, we give a precise definition of the nonrandom value $\boldsymbol{\beta}^*$

that $\hat{\boldsymbol{\beta}}$ is estimating (it is $\boldsymbol{\beta}_{\lim}^{(h)}$ in equation (A.3)), and show that the effect of using $\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}^*$ is asymptotically negligible.

It is natural to consider the problem of estimating $\boldsymbol{\beta}_{\text{opt}}$ in the Markov chains setting. Actually, before thinking about minimizing the variance of (2.9) with respect to $\boldsymbol{\beta}$, one should first note the following. The constants $a_s = n_s/n$, $s = 1, \ldots, k$, used in forming the values $Y_{i,l}^{(h)}$ are sensible in the i.i.d. setting, but when dealing with Markov chains one would want to replace n_s with an "effective sample size," as discussed by Meng and Wong (1996). Therefore, the real problem is two-fold:

- How do we find optimal (or good) values to use in place of the a_s's in the $Y_{i,l}^{(h)}$'s?
- Using the $Y_{i,l}^{(h)}$'s based on these values, how do we estimate the value of $\boldsymbol{\beta}$ that minimizes the variance of (2.9)?

Both problems appear to be very difficult. Intuitively at least, the method described here should perform well if the mixing rates of the Markov chains are not very different. But in any case, the results in Section 2.3 show that, whether or not $\hat{I}_{h,\hat{\boldsymbol{\beta}}}$ is optimal, it is a consistent and asymptotically normal estimator whose variance can be estimated consistently.

Note that if we do not use control variates, our estimate is just

$$\frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n_l}\frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/d_s},$$

which is exactly (2.1).

Reduction in Variance from Using the Control Variates. Consider the linear combination of the responses $Y^{(h)}$ and predictors $Z^{(j)}$ given by

$$L_1 = \sum_{j=2}^{k} a_j Z^{(j)} + Y^{(h)}.$$

(We are dropping the subscripts i, l.) A calculation shows that if $h = h_1$ then $L_1 = 1$, meaning that we have an estimate with zero variance. Similarly, for $t = 2, \ldots, k$, let L_t be the linear combination given by

$$L_t = \sum_{j=2}^{k} a_j Z^{(j)} + \Big(\frac{1}{d_t}\Big) Y^{(h)} - Z^{(t)}.$$

If $h = h_t$, then $L_t = 1$. Thus *if $h \in \{h_1, \ldots, h_k\}$, our estimate of the Bayes factor $B(h, h_1)$ has zero variance.* This is not surprising since, after all, we are

assuming that we know $B(h_j, h_1)$, for $j = 1, \ldots, k$; however, this does indicate that if we use these control variates, our estimate will be very precise as long as h is close to at least one of the h_j's. This advantage does not exist if we use the plain estimate (2.1).

The intercept term in the regression of the $Y_{i,l}^{(h)}$'s on the $Z_{i,l}^{(j)}$'s is simply a linear combination of the form

$$\hat{\beta}_0 = \sum_{l=1}^{k} \sum_{i=1}^{n_l} w_{i,l} Y_{i,l}^{(h)}. \tag{2.12}$$

The $w_{i,l}$'s need to be computed just once, so for every new value of h the calculation of $\hat{B}_{\mathrm{reg}}(h, h_1)$ requires n operations, which is the same as the number of operations needed to compute $\hat{B}(h, h_1)$ given by (2.1). To summarize, using control variates can greatly improve the accuracy of the estimates, at no (or trivial) increase in computational cost.

2.3. Asymptotic normality and estimation of the variance

Here we state a result that says that under certain regularity conditions $\hat{B}_{\mathrm{reg}}(h, h_1)$ and $\hat{B}(h, h_1)$ are asymptotically normal, and we show how to estimate the variance. As discussed in Section 2.2, we typically prefer that $\theta_i^{(l)}$, $i = 1, \ldots, n_l$, be an i.i.d. sample for each l. Nevertheless, our results pertain to the more general case where these samples arise from Markov chains. (As before, we assume that $n_l/n \to a_l \in (0, 1)$ and, when dealing with the asymptotics, strictly speaking we need to make a distinction between n_l/n and its limit; however we write a_l for both as this makes the bookkeeping easier, and blurring the distinction never creates a problem.)

Recall that $Y_{i,l}^{(h)}$ and $Z_{i,l}^{(j)}$, $j = 2, \ldots, k$, are defined in (2.10) and, for economy of notation, we define $Z_{i,l}^{(1)}$ to be 1 for all i, l. Let $\boldsymbol{R}$ be the $k \times k$ matrix defined by

$$R_{jj'} = E\Big(\textstyle\sum_{l=1}^{k} a_l Z_{1,l}^{(j)} Z_{1,l}^{(j')}\Big), \qquad j, j' = 1, \ldots, k.$$

We assume that for the Markov chains a strong law of large numbers holds (sufficient conditions are given, for example, in Theorem 2 of Athreya, Doss and Sethuraman (1996)), and we refer to the following conditions.

A1 For each $l = 1, \ldots, k$, the chain $\{\theta_i^{(l)}\}_{i=1}^{\infty}$ is geometrically ergodic.

A2 For each $l = 1, \ldots, k$, there exists $\epsilon > 0$ such that $E\big(\big|Y_{1,l}^{(h)}\big|^{2+\epsilon}\big) < \infty$.

A3 The matrix $\boldsymbol{R}$ is nonsingular.

Theorem 1. *Under conditions* A1 *and* A2

$$n^{1/2}\big(\hat{B}(h, h_1) - B(h, h_1)\big) \xrightarrow{d} \mathcal{N}\big(0, \tau^2(h)\big),$$

and under conditions A1−A3

$$n^{1/2}\big(\hat{B}_{\text{reg}}(h,h_1) - B(h,h_1)\big) \xrightarrow{d} \mathcal{N}\big(0,\sigma^2(h)\big),$$

with $\tau^2(h)$ *and* $\sigma^2(h)$ *given by equations* (A.9) *and* (A.7) *below.*

The proof is given in the Appendix, which also explains how one can estimate the variances.

Theorem 1 assumes that the vector $\boldsymbol{d}$ is known—either because it can be computed analytically or because the sample sizes from Stage 1 sampling are so large that this is effectively true. Buta (2010) has obtained a version of Theorem 1 that takes into account the variability from the first stage. Very briefly, if N is the total sample size from the first stage, and if $N \to \infty$ and $n \to \infty$ in such a way that $n/N \to q \in [0,\infty)$, then $n^{1/2}\big(\hat{B}(h,h_1) - B(h,h_1)\big) \xrightarrow{d} \mathcal{N}\big(0, q\tau^2_{S1}(h) + \tau^2(h)\big)$, where $\tau^2_{S1}(h)$ is a correction term that inflates the variance when the sample sizes in Stage 1 are finite. Also, she has a similar result for the estimate that uses control variates.

The variances of $\hat{B}_{\text{reg}}(h,h_1)$ and $\hat{B}(h,h_1)$ depend on the choice of the points $h_1,\ldots,h_k$, and finding good values of k and $h_1,\ldots,h_k$ is in general a very difficult problem. In our experience, we have found that the following method works reasonably well. Having specified the range $\mathcal{H}$, we select trial values $h_1,\ldots,h_k$, and in pilot runs plot the variance function $\tau^2(h)$, or $\sigma^2(h)$; then if we find a region where this is unacceptably large, we "cover" this region by moving some h_l's closer to the region, or by simply adding new h_l's in that region, which increases k.

3. Illustration

There are many classes of models to which the methodology developed in Section 2 applies. These include the usual parametric models, and also Bayesian nonparametric models involving mixtures of Dirichlet processes (Antoniak (1974)), in which one of the hyperparameters is the so-called total mass parameter—very briefly, this hyperparameter controls the extent to which the nonparametric model differs from a purely parametric model. Another application involves some problems in Bayesian variable selection, and this is described in Doss (2007). In this section we give an example involving the hierarchical Bayesian model described in Section 1. While models of much greater complexity can be considered, this relatively simple example has the advantage that the data can be visualized quickly, and the hyperparameters have a straightforward interpretation so that our analysis can be easily understood.

Table 1. Fifteen studies on aspirin and colon cancer. Here, PPW represents the dose (number of 325 mg pills per week), RR is the observed risk ratio for aspirin vs. no aspirin, LRR is its logarithm, and SE(LRR) is an estimate of the standard error of LRR.

Publication	PPW	RR	LRR	SE(LRR)
Coogan, 00	4	0.50	−0.69	0.172
Friedman, 98	3	0.70	−0.36	0.068
Garcia-Rod., 01	7	0.60	−0.51	0.207
Giovannucci, 94	2	0.68	−0.39	0.154
Giovannucci, 95	2	0.56	−0.58	0.242
LaVecchia, 97	4	0.70	−0.36	0.182
Muscat, 94	3	0.64	−0.45	0.212
Paganini-Hill, 89	7	1.50	0.41	0.195
Peleg, 94	7	0.25	−1.39	0.547
Reeves, 96	2	0.79	−0.24	0.277
Rosenberg, 91	4	0.50	−0.69	0.240
Rosenberg, 98	4	0.70	−0.36	0.128
Schr. & Ev., 94	1	0.74	−0.30	0.202
Suh, 93	7	0.24	−1.43	0.374
Thun, 91	4	0.48	−0.73	0.234

Meta-Analysis of Data on Non-Steroidal Anti-Inflammatory Drugs and Cancer Risk

Over the last decade, a large number of epidemiological studies have reported a link between intake of nonsteroidal anti-inflammatory drugs (NSAIDs) and cancer risk. The studies, which involve different cancers and different NSAIDs, strongly suggest that long-term intake of NSAIDs results in a significant reduction in cancer risk for all the major types: colon, breast, lung, and prostate cancer. In Harris, Beebe-Donk, Doss and Burr (2005) we carry out a comprehensive review of the published scientific literature on NSAIDs and cancer. Our review spans 90 papers, which investigate several NSAIDs and ten cancers, including the four major types. We have extracted data from these papers to make tables such as Table 1, which pertains to aspirin and colon cancer. The table gives, for each of 15 studies, the dose, reported risk ratio (for NSAID use vs. non-NSAID use), and the log reported risk ratio together with a standard error. (Harris et al. (2005) does not give these standard errors; it gives 95% confidence intervals for the risk ratios, which can be used to form 95% confidence intervals for the log risk ratios, which in turn can be used to determine the standard errors.) See Harris et al. (2005) for more information on this table and references for the 15 studies.

As can be seen from the table, there is some inconsistency in the studies, with some indicating a large reduction in cancer risk, while others indicate a smaller reduction, in spite of a large dose. This is not surprising, since there is heterogeneity in the patient and control pools (characteristics such as age, ethnicity, and health status vary greatly across the studies). It is therefore of interest to carry out a meta-analysis of these studies. Although there have been a few meta-analyses in the literature, these have been rather informal: all of them have used fixed effects models, and none have taken into account the dose information.

Assume temporarily that all studies involved the same dose. In a random-effects meta-analysis, for each study j there is a latent variable, say ψ_j, that gives the true log risk ratio that would be obtained if the sample sizes for that study were infinite. One is then led to a model such as (1.2), in which the distribution of the study-specific effect is the normal distribution in (1.2b). Two modelling issues now arise. The first is that whereas the first normality assumption (line (1.2a)) is supported by a theoretical result (the approximate normality of functions of binomial estimates), the second normality assumption (line (1.2b)) is not but is typically made for the sake of convenience. In fact, data for several of the other cancers include outliers (see Harris et al. (2005)), and therefore one may wish to use a t distribution instead, this decision being made prior to looking at the colon cancer data. An important modelling issue is then to decide on the number of degrees of freedom.

The second issue is to determine the parameters of the normal / inverse gamma prior $\lambda_{\boldsymbol{c}}$ in (1.2c). Here $\boldsymbol{c} = (c_1, c_2, c_3, c_4)$, where $c_1, c_2, c_4 > 0$ and $c_3 \in \mathbb{R}$ and, under this prior, the distribution of (μ, τ) is as follows: $\gamma = 1/\tau^2 \sim$ Gamma(c_1, c_2) and, conditional on τ, $\mu \sim \mathcal{N}(c_3, c_4\tau^2)$. This prior is commonly used because it is conjugate to the family $\mathcal{N}(\mu, \tau^2)$. With appropriate hyper-parameters, λ can be made to be a flat ("noninformative") prior, and common recommendations are to take c_1 and c_2 to be very small (so that the gamma distribution on γ is an approximation to $d\gamma/\gamma$, the improper Jeffrey's prior), and to take $c_3 = 0$ and c_4 to be very large. Indeed, this is the recommendation made in the examples in the `Bugs` documentation and tutorials. Nevertheless, such a set of hyperparameter values is now sometimes criticized because for small values of c_1 and c_2 the gamma distribution gives high probability to large values of γ (equivalently small values of τ), which greatly encourages the ψ_j's to all be equal to μ. In other words, this causes excessive shrinkage. See for example Gelman (2006).

We wish to address both these issues and now also would like to take into account the dose. Let L_j be the log of the observed risk ratio for study j. Let x_j be the dose, defined as number of pills per day (PPW/7), for study j. Consider the linear model

$$L_j = \alpha_j + \psi_j x_j + \varepsilon_j, \qquad j = 1, \ldots, m, \tag{3.1}$$

where α_j and ψ_j are parameters specific to study j, and ε_j is normally distributed with mean 0 and standard deviation σ_j (given in Column 5 of Table 1). Note that $\alpha_j = 0$, since $x_j = 0$ implies that the treatment and control groups are identical, so that L_j has mean 0. Thus, (3.1) is rewritten as $L_j = \psi_j x_j + \varepsilon_j$, from which we see that ψ_j has the interpretation as the true log risk ratio if the treatment group had taken 1 pill per day. Thus if we let $Y_j = L_j/x_j$, we

have $Y_j = \psi_j + \tilde{\varepsilon}_j$, $j = 1, \ldots, m$, where $\tilde{\varepsilon}_j$ is normal with mean 0 and standard deviation $\tilde{\sigma}_j = \sigma_j / x_j$.

We now consider the hierarchical model

$$Y_j \overset{\text{indep}}{\sim} \phi_{\psi_j, \tilde{\sigma}_j}, \qquad j = 1, \ldots, m, \tag{3.2}$$

with the distribution of $\boldsymbol{\psi}$ determined by the following:

$$\text{conditional on } \mu, \tau, \qquad \psi_j \overset{\text{i.i.d.}}{\sim} t_{v,\mu,\tau}, \quad j = 1, \ldots, m, \tag{3.3a}$$

$$(\mu, \tau) \sim \lambda_{\boldsymbol{c}}. \tag{3.3b}$$

Letting $\theta = (\boldsymbol{\psi}, \mu, \tau)$, the likelihood of $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ is given by (3.2), and the prior on θ is given by (3.3), which is indexed by $h = (v, \boldsymbol{c})$. Loosely speaking, the value of v determines the choice of the model, and the c's determines the prior. We may therefore fix some value h_1 and consider the family of Bayes factors $B(h, h_1)$ as h varies. We can estimate the family if for values h_j, $j = 1, \ldots, k$, of the hyperparameter h, we have samples from the posterior distributions $\nu_{h_j, \boldsymbol{Y}}$ of the entire vector θ.

We considered four different values of $\boldsymbol{c}$ in which $c_3 = 0$, $c_4 = 1,000$ were fixed (since there does not seem to be any controversy about these two parameters) and we took $c_1 = c_2$ and let the common value, denoted ϵ, start at 0.005 and increase by factors of 5 up to 0.625. We took the values of the degrees of freedom parameter to be $v = 1, 4, 12$, for a total of 12 values of the hyperparameter h. For each of these 12 values we ran a Markov chain of length about 1 million and used these to calculate the vector of ratios of normalizing constants, via the method of Meng and Wong (1996) reviewed in Section 2.1. We then ran new Markov chains to produce a sample of size 100 from each of the 12 posteriors. These samples, which were actually subsamples from longer chains (burn-in of $1,000$, then taking every 50^{th} value), can be considered i.i.d. for practical purposes, and were used to calculate the estimate $\hat{B}_{\text{reg}}(h, h_1)$ of Section 2.2. We took h_1 to be the specification corresponding to $v = 4$ and $\epsilon = 0.125$, since preliminary experiments indicated that this value of h gave a relatively high value of m_h. Figure 1 shows $\hat{B}_{\text{reg}}(h, h_1)$ as v and ϵ vary. The maximum standard error over the range of the graph was less than 0.01.

The two plots in Figure 1 show different views of the same graph. From the left plot we see that a t distribution works better than does a normal, with the optimal number of degrees of freedom being about 3 or 4. The plot also shows clearly that a very small number of degrees of freedom is not appropriate. The right plot shows that as $\epsilon \to 0$, the Bayes factor converges to 0 rapidly (in particular, fixing $v = 4$, the recommendation in the `Bugs` literature to use $\epsilon = 0.001$ gives a Bayes factor of about 0.036, and for $\epsilon = 0.0001$ it is 0.0037), giving strong evidence that very small values of ϵ should not be used.

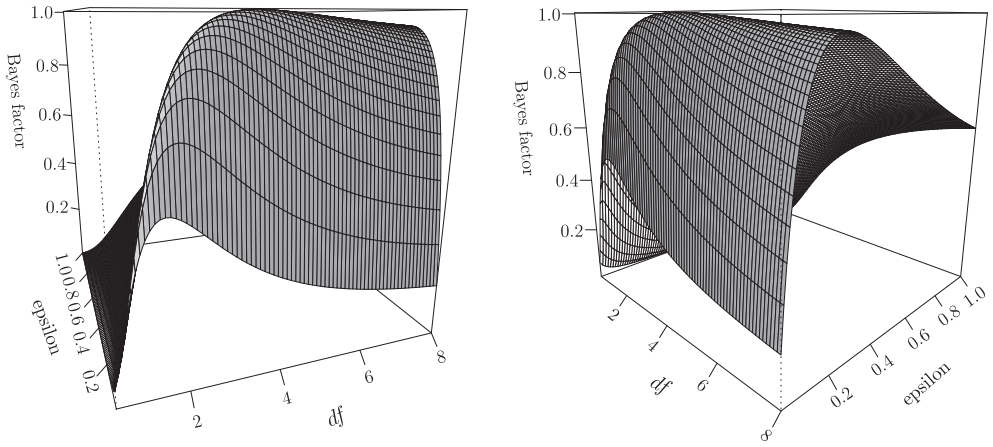

Figure 1. Model assessment for the aspirin and colon cancer data. The Bayes factor as a function of v, the number of degrees of freedom in (3.3a), and ϵ, the common value of c_1 and c_2 in the gamma prior in (3.3b), is shown from two different angles. Here the baseline value of the hyperparameter corresponds to $v = 4$ and $\epsilon = 0.125$.

For some models the improper prior $d\gamma/\gamma$ gives rise to a proper posterior, and for others, including model (3.3b), it is possible to prove that the posterior is improper (Berger (1985, p.187)), so that the pathological behavior resulting from $\epsilon \to 0$ should be expected. For some more complicated models, whether the posterior is proper or not is unknown (posterior propriety may even depend on the data values), and in these cases, plots such as those in Figure 1 may be useful because they may lead one to investigate a possible posterior impropriety.

The choice of hyperparameter h does have an influence on our inference. Let ψ_{new} denote the latent variable for a future study, a quantity of interest in meta-analysis. We considered two specifications of h: $(v = \infty, \epsilon = 0.001)$ and $(v = 4, \epsilon = 0.625)$. The first choice may be considered a "default choice," and the second a choice guided by consideration of the plot of Bayes factors. For the choice $(v = \infty, \epsilon = 0.001)$, we have $E(\psi_{\text{new}}) = -0.87$ and $P(\psi_{\text{new}} > 0) = 0.04$, whereas for $(v = 4, \epsilon = 0.625)$, we have $E(\psi_{\text{new}}) = -0.95$ and $P(\psi_{\text{new}} > 0) = 0.08$. In other words, the t model suggests a stronger aspirin effect, but the inference is more tentative.

Remarks on Computation and Accuracy

- We now give an idea of how the computational effort is distributed. The Stage 1 samples (12 chains, each of length 10^6) took 183 seconds to generate on a 3.8 GHz dual core P4 running Linux. By contrast, the plot in Figure 1, which involves a grid of $4,000$ points, took one hour to compute, in spite of

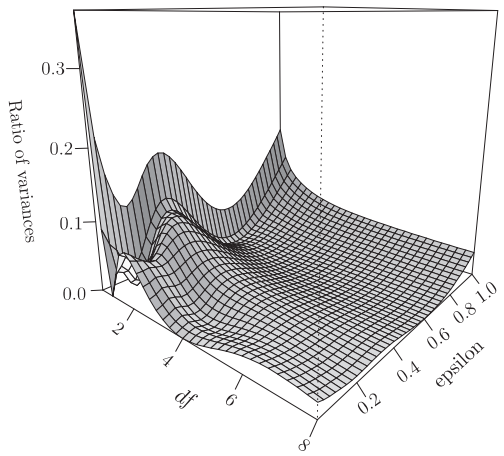

Figure 2. Improvement in accuracy from using control variates. Plot gives $\mathrm{Var}\left(\hat{B}_{\mathrm{reg}}(h, h_1)\right)/\mathrm{Var}\left(\hat{B}(h, h_1)\right)$ as h ranges over the same region as in Figure 1.

the fact that it is based on a total sample size n of only $1,200$, for what must be considered a rather simple model. Clearly using a very large value of n is not feasible, and this is why we need to run the preliminary chains in order to get a very accurate estimate of $\boldsymbol{d}$.

- We now illustrate the extent to which $\hat{B}_{\mathrm{reg}}(h,\ h_1)$ is more efficient than $\hat{B}(h, h_1)$. Figure 2 gives a plot of the ratio of the variances of the two estimates as h varies. Both $\hat{B}_{\mathrm{reg}}(h, h_1)$ and $\hat{B}(h, h_1)$ use the design discussed earlier, which involves a total sample size of $1,200$. This figure is obtained by generating 100 Monte Carlo replicates of $\hat{B}_{\mathrm{reg}}(h, h_1)$ and $\hat{B}(h, h_1)$ for each h in a grid somewhat more coarse than the one used in Figure 1. As can be seen from the figure, the ratio is about 0.01 over most of the grid, and is less than 0.1 over the entire grid, with the exception of the values of h for which $df = 0.5$ (for those values, the Bayes factor itself is very small, and the two estimates each have miniscule variances). We also note that the ratio is exactly 0 *at* the design points.

4. Discussion

When faced with uncertainty regarding the choice of hyperparameters, one approach is to put a prior on the hyperparameters, that is, add one layer to the hierarchical model. This approach, which goes under the general name of "Bayesian model averaging," can be very useful. On the other hand, there are several good reasons why one may want to avoid it. First, the choice of prior on the hyperparameters can have a great influence on the analysis. One is tempted

to use a "flat prior" but, as is well known, for certain parameters such a prior can in fact be very informative. In the illustration of Section 3, a flat prior on the degrees of freedom parameter in effect skews the results in favor of the normal distribution. Second, one may wish to do Bayesian model selection, as opposed to Bayesian model averaging, because the subsequent inference is then more parsimonious and interpretable. These points are discussed more fully in George and Foster (2000) and Robert (2001, Chap. 7).

There are a number of papers that deal with estimation of Bayes factors via MCMC. Chen, Shao and Ibrahim (2000, Chapter 5) and Han and Carlin (2001) give an overview of much of this work, and we mention also the more recent paper by Meng and Schilling (2002), which is directly relevant. Most of these papers deal with the case of a single Bayes factor, whereas the present paper is concerned with estimation of large families of Bayes factors. Nevertheless in principle, any of the methods in this literature can be applied to estimate the vector $\boldsymbol{d}$.

Especially important is Kong et al. (2003), whose work we describe in the notation of the present paper. The situation considered there has k known unnormalized densities $q_{h_1}, \ldots, q_{h_k}$, with unknown normalizing constants $m_{h_1}, \ldots, m_{h_k}$, respectively, and for $l = 1, \ldots, k$, there is an i.i.d. sample $\theta_1^{(l)}, \ldots, \theta_{n_l}^{(l)}$ from q_{h_l}/m_{h_l}. The problem is the simultaneous estimation of all ratios m_{h_l}/m_{h_s}, $l, s = 1, \ldots, k$, or equivalently, all ratios $d_l = m_{h_l}/m_{h_1}$, $l = 1, \ldots, k$. In a certain framework, they show that the maximum likelihood estimate (MLE) of $\boldsymbol{d}$ is obtained by solving the system of k equations

$$\hat{d}_r = \frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{q_{h_r}(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s q_{h_s}(\theta_i^{(l)})/\hat{d}_s}, \qquad r = 1, \ldots, k. \tag{4.1}$$

To put this in our context, let $q_{h_l}(\theta) = \ell_y(\theta)\nu_{h_l}(\theta)$, $l = 1, \ldots, k$, and suppose we have i.i.d. samples from the normalized q_{h_l}'s. We may imagine that we have $k + 1$ unnormalized densities $q_{h_1}, \ldots, q_{h_k}, q_h$, with a sample of size 0 from the normalized q_h. The estimate of m_h/m_{h_1} then becomes

$$\frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_l} \frac{\nu_h(\theta_i^{(l)})}{\sum_{s=1}^{k} a_s \nu_{h_s}(\theta_i^{(l)})/\hat{d}_s}.$$

We recognize this as precisely $\hat{B}(h, h_1)$ in (2.1), except that $\hat{d}_1, \ldots, \hat{d}_k$ are formed by solving (4.1), i.e., are estimated from the sequences $\theta_1^{(l)}, \ldots, \theta_{n_l}^{(l)}$, $l = 1, \ldots, k$. Thus, $\hat{B}(h, h_1)$ is the same as the estimate of Kong et al. (2003), except that the vector $\boldsymbol{d}$ is precomputed based on previously run very long chains. Therefore, it is perhaps natural to consider estimating $\boldsymbol{d}$ on the basis of these very long

Markov chains using the method of Kong et al. (2003) (as opposed to the method discussed in Section 2.1), and we now discuss this possibility.

In their approach, Kong et al. (2003) assume that the q_{h_l}'s are densities with respect to a dominating measure μ, and they obtain the MLE $\hat{\mu}$ of μ ($\hat{\mu}$ is given up to a multiplicative constant). They can then estimate the ratios m_{h_l}/m_{h_s} since the normalizing constants are known functions of μ. Their approach works if for each l, $\theta_1^{(l)}, \ldots, \theta_{n_l}^{(l)}$ is an i.i.d. sample. Although they extend it to the case where these are a Markov chain, in the extension q_{h_l} is replaced by the Markov transition functions $P_{h_l}(\cdot, \theta_i^{(l)})$, $i = 0, \ldots, n_l - 1$, assumed absolutely continuous with respect to a sigma-finite measure μ (precluding Metropolis-Hastings chains), and if each of these is known only up to a normalizing constant—as is typically the case—then the system (4.1) becomes a system of $n \times k$ equations. This is prohibitively difficult to solve.

Tan (2004) shows how control variates can be incorporated in the likelihood framework of Kong et al. (2003). When there are r functions H_j, $j = 1, \ldots, r$, for which we know that $\int H_j \, d\mu = 0$, the parameter space is restricted to the set of all sigma-finite measures satisfying these r constraints. For the case where $\theta_i^{(l)}$, $i = 1, \ldots, n_l$, are i.i.d. for each $l = 1, \ldots, k$, he obtains the MLE of μ in this reduced parameter space, and therefore a corresponding estimate of m_h/m_{h_1}, and shows that this approach gives estimates that are asymptotically equivalent to estimates that use control variates via regression. His estimate can still be used when we have Markov chain draws, but is no longer optimal—for the same reason that the estimate in the present paper is not optimal (see the discussion in the middle of Section 2.2). The optimal estimator is obtained by using the likelihood that arises from the Markov chain structure, and in the case of general Markov chains its calculation is computationally very demanding. See Tan (2006, 2008) for advances in this direction. Tan (2004) also obtains results on asymptotic normality of his estimators that are valid when we have the i.i.d. structure, but it should be possible to obtain versions for Markov chain draws, under regularity conditions such as those of the present paper.

Owen and Zhou (2000) use control variates in conjunction with importance sampling. In the notation above, they assume that the q_{h_l}'s are *normalized* densities, and that for every l, they have an i.i.d. sample of size n_l from q_{h_l}. As before, let $a_l = n_l / \sum_{s=1}^k n_s$. Because these are normalized densities, each of the k variables $q_{h_l}(\theta)/\big(a_s \sum_{s=1}^k q_{h_s}(\theta)\big)$ has expectation 1 under the distribution $\sum_{s=1}^k a_s q_{h_s}$, and so can be used as control variates. Their method does not work directly in our situation because the $q_{h_l} = \ell_y(\theta)\nu_{h_l}(\theta)$ are unnormalized densities. It is therefore natural to consider estimating the normalizing constants of q_{h_l}, $l = 1, \ldots, k$, from the Stage 1 runs. Indeed, there are methods for doing this

from Markov chain output (Chib (1995), Chib and Jeliazkov (2001)). However, estimation of ratios of normalizing constants tends to be far more stable than estimation of the normalizing constants themselves. For example, if we wish to estimate m_h/m_{h_1}, then a procedure that involves estimating m_h and m_{h_1} separately and then taking the ratio is not guaranteed to provide accurate estimates even when $h = h_1$, whereas in this case the simple estimate (1.1) gives an unbiased estimate with zero variance. Moreover, if we run Markov chains for models indexed by $h_1, \ldots, h_k$, the estimate of a single ratio m_{h_s}/m_{h_1} using the method of Section 2.1 makes use of all the chains, providing greater stability. The control variates that we use are essentially equivalent to those used by Owen and Zhou (2000), but their computation requires only knowledge of the vector $\boldsymbol{d}$.

R functions for producing the estimates $\hat{B}(h, h_1)$ and $\hat{B}_{\text{reg}}(h, h_1)$, and plots such as those in Figure 1 for the hierarchical model (3.2)–(3.3) and relatives, are available from the author upon request.

Acknowledgements

I thank two referees for their careful reading, and Eugenia Buta for helpful comments. I am especially grateful to an associate editor for a very insightful and thorough report, and for suggestions that led to several improvements in the paper.

Appendix: Proof of Theorem 1

Under Conditions A1 and A2 we have a central limit theorem for the averages $n_l^{-1}\sum_{i=1}^{n_l} Y_{i,l}^{(h)}$ and $n_l^{-1}\sum_{i=1}^{n_l} Z_{i,l}^{(j)} Y_{i,l}^{(h)}$ for $l = 1, \ldots, k$, and $j = 2, \ldots, k$ (corollary to Theorem 18.5.3 of Ibragimov and Linnik (1971)); however, there are other sets of conditions that could be used. For example, the $\epsilon > 0$ is not needed, i.e., a finite second moment suffices if the chain is reversible (Roberts and Rosenthal (1997))—for instance if the chain is a Metropolis algorithm, or if it is a two-cycle Gibbs sampler—or if it is uniformly ergodic (Cogburn (1972)). These are the most commonly used assumptions, but for a fuller discussion of central limit theorems for Markov chains see Chan and Geyer (1994).

We first prove the assertion regarding $\hat{B}_{\text{reg}}(h, h_1)$. Let $\boldsymbol{Z}$ be the $n \times k$ matrix whose transpose is

$$\boldsymbol{Z}' = \begin{pmatrix} 1 & \ldots & 1 & 1 & \ldots & 1 & \ldots & 1 & \ldots & 1 \\ Z_{1,1}^{(2)} & \ldots & Z_{n_1,1}^{(2)} & Z_{1,2}^{(2)} & \ldots & Z_{n_2,2}^{(2)} & \ldots & Z_{1,k}^{(2)} & \ldots & Z_{n_k,k}^{(2)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Z_{1,1}^{(k)} & \ldots & Z_{n_1,1}^{(k)} & Z_{1,2}^{(k)} & \ldots & Z_{n_2,2}^{(k)} & \ldots & Z_{1,k}^{(k)} & \ldots & Z_{n_k,k}^{(k)} \end{pmatrix},$$

and let $\boldsymbol{Y} = \boldsymbol{Y}^{(h)} = \boldsymbol{Y} = \big(Y_{1,1}^{(h)}, \ldots, Y_{n_1,1}^{(h)}, Y_{1,2}^{(h)}, \ldots, Y_{n_2,2}^{(h)}, \ldots, Y_{1,k}^{(h)}, \ldots, Y_{n_k,k}^{(h)}\big)'$. Note: we sometimes suppress the superscript h in order to lighten the notation. The least squares estimate is $\big(\hat\beta_0^{(h)}, \hat{\boldsymbol{\beta}}^{(h)}\big) = n(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\,\boldsymbol{Z}'\boldsymbol{Y}/n$, assuming that $\boldsymbol{Z}'\boldsymbol{Z}$ is nonsingular. (Here, $\hat{\boldsymbol{\beta}}^{(h)} = (\hat\beta_2^{(h)}, \ldots, \hat\beta_k^{(h)})$).

Note that

$$\frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n_l} Z_{i,l}^{(j)} Z_{i,l}^{(j')} = \sum_{l=1}^{k}\frac{n_l}{n}\frac{1}{n_l}\sum_{i=1}^{n_l} Z_{i,l}^{(j)} Z_{i,l}^{(j')} \xrightarrow{\text{a.s.}} R_{j,j'}$$

by the strong law of large numbers (clearly $Z_{i,l}^{(j)}$ are bounded random variables). Therefore $\boldsymbol{Z}'\boldsymbol{Z}/n \xrightarrow{\text{a.s.}} \boldsymbol{R}$, so by A3 we have

$$n(\boldsymbol{Z}'\boldsymbol{Z})^{-1} \xrightarrow{\text{a.s.}} \boldsymbol{R}^{-1} \tag{A.1}$$

and, in particular, with probability one, $\boldsymbol{Z}'\boldsymbol{Z}$ is nonsingular for large n. We have

$$\frac{\boldsymbol{Z}'\boldsymbol{Y}}{n} = \begin{pmatrix} \frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n_l} Z_{i,l}^{(1)} Y_{i,l} \\ \vdots \\ \frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n_l} Z_{i,l}^{(k)} Y_{i,l} \end{pmatrix} \xrightarrow{\text{a.s.}} \begin{pmatrix} \sum_{l=1}^{k} a_l E\big(Z_{1,l}^{(1)} Y_{1,l}\big) \\ \vdots \\ \sum_{l=1}^{k} a_l E\big(Z_{1,l}^{(k)} Y_{1,l}\big) \end{pmatrix}. \tag{A.2}$$

Let $\boldsymbol{v} = (v_1, \ldots, v_k)$ be the vector on the right side of (A.2). From (A.1) and (A.2) we have

$$\big(\hat\beta_0^{(h)}, \hat{\boldsymbol{\beta}}^{(h)}\big) \xrightarrow{\text{a.s.}} \big(\beta_{0,\lim}^{(h)}, \boldsymbol{\beta}_{\lim}^{(h)}\big) = \boldsymbol{R}^{-1}\boldsymbol{v}. \tag{A.3}$$

Consider (2.9), using $\boldsymbol{\beta}_{\lim}^{(h)}$ for $\boldsymbol{\beta}$. We have

$$\hat I_{h,\boldsymbol{\beta}_{\lim}^{(h)}} = \frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n_l}\Big(Y_{i,l} - \textstyle\sum_{j=2}^{k}\beta_{j,\lim}^{(h)} Z_{i,l}^{(j)}\Big) = \displaystyle\sum_{l=1}^{k} a_l\left(\frac{1}{n_l}\sum_{i=1}^{n_l} U_{i,l}\right), \tag{A.4}$$

where $U_{i,l} = Y_{i,l} - \sum_{j=2}^{k}\beta_{j,\lim}^{(h)} Z_{i,l}^{(j)}$. Let $\mu_l(h) = E(U_{1,l})$. By A2, $E(|U_{1,l}|^{2+\epsilon}) < \infty$ and therefore, by A1 we have

$$n_l^{1/2}\left(\frac{\sum_{i=1}^{n_l} U_{i,l}}{n_l} - \mu_l(h)\right) \xrightarrow{d} \mathcal{N}\big(0, \sigma_l^2(h)\big),$$

where

$$\sigma_l^2(h) = \operatorname{Var}(U_{1,l}) + 2\textstyle\sum_{g=1}^{\infty}\operatorname{Cov}(U_{1,l}, U_{1+g,l}). \tag{A.5}$$

Since the Markov chains are independent, this implies that

$$n^{1/2}\Big(\hat I_{h,\boldsymbol{\beta}_{\lim}^{(h)}} - \textstyle\sum_{l=1}^{k} a_l\mu_l(h)\Big) \xrightarrow{d} \mathcal{N}\big(0, \sigma^2(h)\big), \tag{A.6}$$

where

$$\sigma^2(h) = \textstyle\sum_{l=1}^k a_l \sigma_l^2(h). \tag{A.7}$$

Note that $(1/n)\sum_{l=1}^k \sum_{i=1}^{n_l} Y_{i,l} \xrightarrow{\text{a.s.}} B(h,h_1)$ and $(1/n)\sum_{l=1}^k \sum_{i=1}^{n_l} Z_{i,l}^{(j)} \xrightarrow{\text{a.s.}} 0$, $j = 2,\ldots,k$. Therefore, from the first equation in (A.4), $\hat{I}_{h,\boldsymbol{\beta}_{\lim}^{(h)}} \xrightarrow{\text{a.s.}} B(h,h_1)$ which, together with (A.6), proves that $\sum_{l=1}^k a_l \mu_l(h) = B(h,h_1)$.

To conclude the proof, we consider the difference between $\hat{I}_{h,\hat{\boldsymbol{\beta}}^{(h)}}$ and $\hat{I}_{h,\boldsymbol{\beta}_{\lim}^{(h)}}$. Let $e(j,l) = E\big(Z_{1,l}^{(j)}\big)$. We have

$$\begin{aligned} n^{1/2}\Big(\hat{I}_{h,\hat{\boldsymbol{\beta}}^{(h)}} - \hat{I}_{h,\boldsymbol{\beta}_{\lim}^{(h)}}\Big) &= n^{1/2}\sum_{j=2}^k (\beta_{j,\lim} - \hat{\beta}_j)\left(\frac{1}{n}\sum_{l=1}^k\sum_{i=1}^{n_l} Z_{i,l}^{(j)}\right) \\ &= \sum_{j=2}^k (\beta_{j,\lim} - \hat{\beta}_j)\left(\sum_{l=1}^k a_l\, n^{1/2}\sum_{i=1}^{n_l}\left[\frac{Z_{i,l}^{(j)} - e(j,l)}{n_l}\right]\right), \end{aligned} \tag{A.8}$$

where the second equality in (A.8) follows from the fact that $\sum_{l=1}^k a_l e(j,l) = 0$. Now, for each $l = 1,\ldots,k$, and $j = 2,\ldots,k$, by A1, $n^{1/2}\sum_{i=1}^{n_l}\big[(Z_{i,l}^{(j)} - e(j,l))/n_l\big]$ is asymptotically normal, so in particular is bounded in probability. Together with (A.3), this implies that the right side of (A.8) converges in probability to 0. We conclude that $n^{1/2}\big(\hat{B}_{\text{reg}}(h,h_1) - B(h,h_1)\big) \xrightarrow{d} \mathcal{N}\big(0,\sigma^2(h)\big)$.

The proof for $\hat{B}(h,h_1)$ is simpler. Let $f_l = E(Y_{1,l})$, and note that $\sum_{l=1}^k a_l f_l = B(h,h_1)$. We have

$$\begin{aligned} n^{1/2}\big(\hat{B}(h,h_1) - B(h,h_1)\big) = n^{1/2}\left(\frac{1}{n}\sum_{l=1}^k\sum_{i=1}^{n_l} Y_{i,l} - f_l\right) &= \sum_{l=1}^k a_l^{1/2}\frac{\sum_{i=1}^{n_l}(Y_{i,l} - f_l)}{n_l^{1/2}} \\ &\xrightarrow{d} \mathcal{N}\big(0,\tau^2(h)\big), \end{aligned}$$

in which

$$\tau^2(h) = \textstyle\sum_{l=1}^k a_l \tau_l^2(h), \quad \text{where} \quad \tau_l^2(h) = \text{Var}\,(Y_{1,l}) + 2\sum_{g=1}^\infty \text{Cov}\,(Y_{1,l}, Y_{1+g,l}). \tag{A.9}$$

The variance term $\sigma_l^2(h)$ in (A.5) is the asymptotic variance of the standardized version of the average $\sum_{i=1}^{n_l} U_{i,l}$. If we knew the $U_{i,l}$'s, we could estimate $\sigma_l^2(h)$ by estimating the initial segment of the series in (A.5) using standard methods from time series (see Geyer (1992)) or via batching. Now the $U_{i,l}$'s involve $\boldsymbol{\beta}_{\lim}^{(h)}$, which is unknown, but our proof indicates that the effect of using $\hat{\boldsymbol{\beta}}^{(h)}$ instead of $\boldsymbol{\beta}_{\lim}^{(h)}$ in the expression for $U_{i,l}$ is asymptotically negligible.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152-1174.

Athreya, K. B., Doss, H. and Sethuraman, J. (1996). On the convergence of the Markov chain simulation method. *Ann. Statist.* **24**, 69-100.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* . Second edition. Springer-Verlag, New York.

Buta, E. (2010). Computational methods in Bayesian sensitivity analysis. Ph.D. thesis, University of Florida.

Chan, K. S. and Geyer, C. J. (1994). Comment on "Markov chains for exploring posterior distributions". *Ann. Statist.* **22**, 1747-1758.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.

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

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.* **96**, 270-281.

Cogburn, R. (1972). The central limit theorem for Markov processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, **2**. University of California Press, Berkeley.

Doss, H. (2007). Bayesian model selection: Some thoughts on future directions. *Statist. Sinica* **17**, 413-421.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515-534.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731-747.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (Disc: p483-503). *Statist. Sci.* **7**, 473-483.

Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *J. Amer. Statist. Assoc.* **96**, 1122-1132.

Harris, R., Beebe-Donk, J., Doss, H. and Burr, D. (2005). Aspirin, Ibuprofen and other non-steroidal anti-inflammatory drugs in cancer prevention: A critical review of non-selective COX-2 blockade. *Oncology Reports* **13**, 559-584.

Hobert, J. P. and Robert, C. P. (2004). A mixture representation of π with applications in Markov chain Monte Carlo and perfect sampling. *Ann. Appl. Probab.* **14**, 1295-1305.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

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

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D. and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussion). *J. Roy. Statist. Soc. Ser. B* **65**, 585-618.

Meng, X.-L. and Schilling, S. (2002). Warp bridge sampling. *J. Comput. Graph. Statist.* **11**, 552-586.

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

Owen, A. and Zhou, Y. (2000). Safe and effective importance sampling. *J. Amer. Statist. Assoc.* **95**, 135-143.

Robert, C. P. (2001). *The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation*. Springer-Verlag, New York.

Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability* **2**, 13-25.

Tan, Z. (2004). On a likelihood approach for Monte Carlo integration. *J. Amer. Statist. Assoc.* **99**, 1027-1036.

Tan, Z. (2006). Monte Carlo integration with acceptance-rejection. *J. Comput. Graph. Statist.* **15**, 735-752.

Tan, Z. (2008). Monte Carlo integration with Markov chain. *J. Statist. Plann. Inference* **138**, 1967-1980.

Department of Statistics, University of Florida, P.O. Box 118545, Gainesville, FL 32611-8545, U.S.A.

E-mail: doss@stat.ufl.edu

(Received December 2007; accepted January 2009)