Posterior Distributions on Normalizing Constants

Valen E. Johnson

August 8, 1998
Revised June 10, 1999

Abstract

This article describes a procedure for defining a posterior distribution on the value of a normalizing constant or ratio of normalizing constants using output from Monte Carlo simulation experiments. The resulting posterior distribution provides a simple diagnostic for assessing the adequacy of a simulation experiment for estimating these quantities, and is particularly useful in cases for which standard estimators perform poorly, since in such situations asymptotic properties of standard diagnostics are unlikely to hold.

Keywords: Marginal likelihood, partition function, Markov chain Monte Carlo, Ising model, γ coupling.

1 Introduction

This article describes a simulation-based method for computing a posterior distribution on either a single normalizing constant or a ratio of normalizing constants. The method relies on a coupling argument to define two sequences of Bernoulli random variables whose success probabilities, given the true values of the normalizing constants, are equal. The posterior distribution defined by this procedure is exact when sample values are drawn independently from both target densities. When only dependent sequences of draws are available, as is the case when analyzing output from Markov chain Monte Carlo (MCMC) simulations, an approximate posterior distribution is obtained by modeling the derived sequences of Bernoulli random variables as two-state Markov chains.

The problem of estimating normalizing constants arises frequently in both physics and statistics, and one of two general approaches is usually taken towards estimating them. In one approach, ratios of normalizing constants are estimated using Monte Carlo approximations to expressions of the type

$$\frac{c_2}{c_1} = \frac{E_2[f_1(x)\alpha(x)]}{E_1[f_2(x)\alpha(x)]}, \tag{1}$$

where $c_i f_i(x)$ denote probability density or mass functions and $E_i(\cdot)$ denotes expectation with respect to density i. The function $\alpha(x)$ is an arbitrary function defined so that the denominator in (1) is nonzero. The generality of this equation was exposed in Meng and Wong (1996), where its relation to standard importance sampling estimates and to estimates proposed by Bennett (1976), Newton and Raftery (1994) and Gelfand and Dey (1994) were discussed in detail. There are also strong ties between (1) and the estimators advocated by Geyer and Thompson (1992) and Geyer (1994). Such Monte Carlo approximation schemes can be highly efficient when there is substantial overlap between the densities $c_1 f_1$ and $c_2 f_2$. Meng and Wong discuss an iterative scheme to determine the asymptotically optimal choice of α, which can be expressed as

$$\alpha(x) = \frac{1}{f_1(x) + (\hat{c}_2/\hat{c}_1) f_2(x)}, \tag{2}$$

where $(\hat{c}_2/\hat{c}_1)$ is obtained iteratively using this equation and (1).

For situations in which there is limited overlap between the target densities, Meng and Wong (1996) described a bridge sampling scheme that works by interjecting sampled values from densities "between" $c_1 f_1$ and $c_2 f_2$ into the Monte Carlo approximation, thus improving the overlap between each pair of adjacent densities. The connection between bridge sampling and path sampling was explored by Gelman and Meng (1998), who also described the relationship between between bridge sampling, path sampling and more standard statistical simulation-based methods as discussed in, for example, Meng and Wong and DiCiccio, Kass, Raftery and Wasserman (1997). Both path sampling and bridge sampling represent important extensions of Monte Carlo approximations to normalizing constants in situations in which the overlap between the densities used in the simulations is small. These techniques appear particularly important in statistical physics where natural paths between various equilibrium distributions can often be defined. However, in this article, attention is restricted to the case of "binary" bridges, or in other words, the case where only two densities are used in the simulation. The diagnostic obtained in this setting can be extended to bridge sampling schemes by considering adjacent densities in the bridges on a pairwise basis. Sampling from only two densities may be the most common situation encountered in statistical applications, and is necessary when comparing models defined on different parameter spaces because in such cases bridges between model spaces cannot be constructed.

The second approach commonly taken towards estimating normalizing constants relies on large sample properties of the target densities. Estimates of this type are based on Laplace approximations to the posterior distribution or sampling distribution (Tierney and Kadane (1986)) and are usually much more efficient from a computational perspective than are Monte Carlo estimators. Unfortunately, Laplace approximation-based estimators of normalizing constants are often not sufficiently accurate for statistical modeling, though methods for improving their accuracy using output from MCMC simulation experiments have recently been proposed by DiCiccio, Kass, Raftery and Wasserman (1997) and Lewis and Raftery (1997).

In this article, a Bayesian approach towards inference regarding a normalizing constant or ratio of normalizing constants is described. The resulting posterior distribution yields consistent estimates of normalizing constants or ratios of normalizing constants as simulation sample sizes become large. However, estimators based on this posterior distribution–like the posterior mean and posterior mode–are inefficient. But because the purpose of introducing the posterior is to assess the uncertainty in more efficient estimators, this inefficiency should not be considered problematic. The marginal posterior distribution on the normalizing constant, or ratio of normalizing constants, is a critical attribute of the method; as most statisticians will attest, knowing how well an estimator performs is as important as knowing its value. This is particularly true in simulation studies, since additional data can usually be generated if uncertainty is discovered to be unacceptably high.

The remainder of this paper is organized as follows. In the next section, posterior distributions on normalizing constants or ratios of normalizing constants are defined under the assumption of independent draws from the target distributions. The sampling properties of this posterior distribution are explored in a stylized example in Section 3, where the average posterior variance is found to track well the mean square error of the asymptotically optimal estimator described by Meng and Wong (1996). In Section 4, this methodology is extended to situations in which only dependent samples from unnormalized densities are available. The approximations developed in this section can be applied directly to estimating normalizing constants or ratios of normalizing constants based on output from MCMC algorithms. Examples of this approximation are provided in Section 5. The final section concludes with a summary of results and discussion.

2 Methodology

To fix notation, let $cf(\cdot)$ denote a probability density function or probability mass function. The normalizing constant to be estimated is denoted by c. For notational simplicity, the problem of estimating a single normalizing constant is described below; estimation of a ratio of normalizing constants follows along identical lines if c is replaced by c_1/c_2 in the discussion. In the context of estimating a single normalizing constant, let $g(\cdot)$ denote an approximation to cf defined on a common probability space. The density g need not provide an accurate approximation to cf as long as the two densities have non-negligible overlap in the L^1 sense. In the context of estimating a ratio of normalizing constants, g denotes the second unnormalized density.

The definition of the posterior distribution on c is motivated by a simulation technique called γ coupling (see, for example, Lindvall (1992)). Gamma coupling is a numerical procedure that can be used to generate pairs of random variables, say x and y, from densities cf and g so that $x \sim cf$, $y \sim g$, and the probability that $x = y$ is maximized.

Practically speaking, γ coupling can be implemented in the following way. To begin, x is drawn from density cf and u is drawn from a uniform distribution on the unit interval ($\equiv U(0,1)$). If $ucf(x) \leq g(x)$, then x is also accepted as a draw from density g; that is, $y = x$. On the other hand, if $u\,cf(x) > g(x)$, x is not accepted as a draw from g, and y must instead be obtained by repeatedly drawing values $t \sim g$ and $v \sim U(0,1)$ until the drawn values satisfy $v\,g(t) > cf(t)$. The value of y is then taken to be t.

The same scheme can also be implemented by reversing the roles of cf and g. That is, y is first drawn from density g and v is drawn from a $U(0,1)$ distribution. If $v\,g(y) \leq cf(y)$, then $x = y$ is accepted as a draw from cf. Otherwise, values of $x \sim cf$ and $u \sim U(0,1)$ are drawn until $u\,cf(x) > g(x)$; the value of x for which this condition is first satisfied is accepted as the draw from cf.

This coupling scheme applies to both continuous probability density functions and discrete probability mass functions, and extends directly to random vectors of arbitrary dimension.

Because cf and g are densities, the probability that a draw taken from the first sampled density is rejected as a draw from the second is invariant to the labeling of the densities. In fact, the probability that the sampled values are not equal under this sampling scheme equals the total variation distance between densities cf and g, defined here as

$$r \equiv \frac{1}{2} \int_X |cf(x) - g(x)| dx.$$

Returning now to the problem of estimating c, suppose that $x_1, \ldots, x_m$ denotes a random sample of size m from cf, and $y_1, \ldots, y_n$ denotes an independent random sample of size n from g. Take $u_1, \ldots, u_m$ and $v_1, \ldots, v_n$ to be independent $U(0,1)$ random deviates, and define independent Bernoulli random variables $W_i(s)$ and $Z_j(s)$ according to

$$W_i(s) = \begin{cases} 1 & \text{if } g(x_i)/sf(x_i) < u_i \\ 0 & \text{otherwise} \end{cases} \qquad Z_j(s) = \begin{cases} 1 & \text{if } sf(y_j)/g(y_j) < v_j \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

Note that in these equations, the normalizing constant c has been replaced by an arbitrary scalar s. Ultimately, our primary interest lies in the case that $s = c$, but to avoid confusion in a conditioning argument that follows, for now consider the more general setting of arbitrary s.

Analytically, the probability that $W_i(s) = 1$ can be expressed

$$p_s \equiv Pr[W_i(s) = 1] = \int_{\{x : sf(x) > g(x)\}} cf(z)\, dz - \int_{\{x : sf(x) > g(x)\}} \frac{c}{s}\, g(z)\, dz.$$

In principle, the value of this function can be computed precisely, but in practice its value is unknown and so, like c, it must be estimated from simulation data. Similar comments apply to the probability that $Z_j(s) = 1$, which is henceforth denoted by q_s.

Treating the function $\mathbf{W}(s) = \{W_1(s), \ldots, W_m(s)\}$ as data from an experiment in which interest lies in estimating s, it follows that the sampling density of $\mathbf{W}(s)$, given s and p_s, may be written

$$f(\mathbf{W}(s)|s, p_s) = \prod_{i=1}^{m} p_s^{W_i(s)}(1-p_s)^{1-W_i(s)}.$$

Likewise, the sampling density of $\mathbf{Z}(s) = \{Z_1(s), \ldots, Z_n(s)\}$, given s and q_s, is

$$f(\mathbf{Z}(s)|s, q_s) = \prod_{j=1}^{n} q_s^{Z_j(s)}(1-q_s)^{1-Z_j(s)}.$$

Because the values of p_s and q_s are unknown, assume that prior information concerning the values of these parameters, conditionally on the value of s, is conveyed through prior densities $\pi(p_s|s)$ and $\pi(q_s|s)$. For computational convenience, these densities are assumed to have the form of beta densities with parameters (α, β) and (γ, δ), respectively. In general, p_s is a non-decreasing function of s, while q_s is a non-increasing function of s.

To obtain the posterior distribution on (s, p_s, q_s), it is important to account for the fact that the particular values of $(\mathbf{W}(s), \mathbf{Z}(s))$ observed depend on the value of s. The marginal likelihood of the data therefore depends on s, and must therefore be explicitly incorporated into the posterior distribution. Defining

$$a_s = \sum_1^m W_i(s) \qquad \text{and} \qquad b_s = \sum_1^n Z_j(s),$$

the marginal likelihood of the data for a particular value of s is given by

$$\begin{aligned} h(\mathbf{W}(s), \mathbf{Z}(s)) = & \\ & k \int_A p_s^{a_s+\alpha-1}(1-p_s)^{m-a_s+\beta-1} \\ & \times\ q_s^{b_s+\gamma-1}(1-q_s)^{(n-b_s+\delta-1)} \\ & \times\ \pi(s)\, dp_s\, dq_s\, ds \end{aligned} \tag{4}$$

where $A = [0,1] \times [0,1] \times K(a_s, b_s)$, and $K(a_s, b_s)$ denotes the interval over which the sum of the Bernoulli sequences $\mathbf{W}$ and $\mathbf{Z}$ equal a and b, respectively. The function $\pi(s)$ denotes the prior on s (discussed below), and k is a constant that does not depend on s, p_s, or q_s.

Simplifying (4) and substituting into the posterior leads to

$$\begin{aligned} f(s, p_s, q_s | \mathbf{W}, \mathbf{Z}) \propto & \qquad\qquad (5) \\ & B(a_s + \alpha, m - a_s + \beta) B(b_s + \gamma, n - b_s + \delta) \frac{1}{\kappa(a_s, b_s)} \\ \times & \quad p_s^{a_s+\alpha-1} (1 - p_s)^{(m-a_s+\beta-1)} q_s^{b_s+\gamma-1} (1 - q_s)^{n-b_s+\delta-1} \pi(s) \end{aligned}$$

where

$$\kappa(a_s, b_s) = \int_{K(a_s, b_s)} \pi(s).$$

For values of s for which the set $K(a_s, b_s)$ is empty, the posterior density is defined to be 0. The function $B(u, v)$ denotes the beta function $\Gamma(u + v)/(\Gamma(u)\Gamma(v))$.

To obtain the posterior distribution on the normalizing constant c, recall that at $c = s$, the probabilities p_s and q_s are equal. Thus, the posterior distribution on c can be obtained by conditioning on the event $|p_s - q_s| \to 0$. Passing to the limit $p_s = q_s \equiv r$, the conditional posterior distribution on (c, r) becomes

$$\begin{aligned} & f(c, r | \mathbf{W}, \mathbf{Z}) \propto \qquad\qquad (6) \\ & B(a_c + \alpha, m - a_c + \beta) B(b_c + \gamma, n - b_c + \delta) \frac{1}{\kappa(a_c, b_c)} \times \\ & r^{a_c+b_c+\alpha+\gamma-2} (1 - r)^{(m+n-a_c-b_c+\beta+\delta-2)} \pi(c). \end{aligned}$$

Equation (6) defines a joint posterior distribution on c and r that can be analyzed numerically to obtain a marginal posterior distribution on c given $\mathbf{W}$ and $\mathbf{Z}$, which are functions of the simulated data $\{g(x_i)/(u_i f(x_i)\}$ and $\{v_j g(y_j)/f(y_j)\}$. If interest focuses exclusively on c, the parameter r can be integrated out of this joint posterior density to obtain the marginal posterior

$$\begin{aligned} & f(c | \mathbf{W}, \mathbf{Z}) \propto \qquad\qquad (7) \\ & \frac{B(a_c + \alpha, m - a_c + \beta) B(b_c + \gamma, n - b_c + \delta)}{B(a_c + b_c + \alpha + \gamma - 1, m + n - a_c - b_c + \beta + \delta - 1)} \frac{\pi(c)}{\kappa(a_c, b_c)} \end{aligned}$$

Several properties of this posterior distribution warrant comment. First, because the likelihood function is flat for sufficiently small or large values of c, the posterior distribution is not integrable unless a proper prior is specified on c. Also, the likelihood function is piecewise constant in c, changing value only at the points $\{g(x_i)/(u_i f(x_i))\}$ and $\{v_j g(y_j)/f(y_j)\}$. As a result, standard asymptotic limit theorems do not apply. However, it is possible to show that the posterior mode converges to values of c and r that satisfy $a_c/m = b_c/n$ and $r = (a_c + b_c)/(m + n)$ as a_c, $m - a_c$, b_c and $n - b_c$ tend to infinity. Also, the log-posterior is concave in a_c and b_c when a_c, $m - a_c$, b_c and $n - b_c$ are all positive. For $0 < r < 1$, consistency of the posterior mode follows from the facts that (i) a_c is a non-increasing function of c, (ii) b_c is a non-decreasing function of c, and (iii) by the strong law of large numbers, $a_c/m \to r$ and $b_c/n \to r$ as $m, n \to \infty$.

In the absence of specific prior information regarding the value of c, a vague, data-dependent prior may be specified by assuming that the prior for c is uniform on the interval

$$[\min\left(\min_i \frac{g(x_i)}{f(x_i)}, \min_j \frac{g(y_j)}{f(y_j)}\right), \max\left(\max_i \frac{g(x_i)}{f(x_i)}, \max_j \frac{g(y_j)}{f(y_j)}\right)]. \qquad (8)$$

This interval contains values of c for which both $a_c = 0$ and $b_c = 0$. For convenience, this prior is assumed for c in all of the examples that follow.

Heuristically, this method for defining a joint posterior distribution on (c, r) is based on finding values of c that lead to approximately the same proportion of rejections in hypothetical γ-coupling between cf and g. The posterior distribution is informative provided that $r < 1$. In the case that g is not a normalized density and the values $\{y_j\}$ are in fact drawn from dg, the posterior distribution on c described above instead represents the posterior distribution on the value of c/d.

As an aside, it is interesting to note that, at least in theory, multiple samples of the marginal posterior density on c specified in (7) can be obtained by resampling the uniform deviates $\{u_i\}$ and $\{v_i\}$ for fixed values of $\{g(x_i)/f(x_i)\}$ and $\{g(y_j)/f(y_j)\}$. (Note that these uniform deviates cannot be included as parameters in the posterior distribution without violating the assumption of independent Bernoulli sampling used in the definition of the likelihood.) The posterior mean of the density that would be obtained by averaging multiple realizations of (7) provides a more efficient estimator of the quantity c than does the posterior mean of a single realization of this density. However, the posterior variance of the averaged density is slightly larger than the average value of the variance of a single realization. Because the posterior uncertainty reflected in a single realization of this posterior density provides a conservative diagnostic for assessing the uncertainty in more efficient estimators, the additional computational effort required to average densities over draws of these uniform deviates is generally not worthwhile.

3 Estimating the normalizing constant of a Student t-density

This example explores the use of the posterior distribution on the normalizing constant of a Student t-density on 4 degrees of freedom in assessing the uncertainty of two more efficient estimators. The first estimator is the asymptotically optimal estimator (AOE) described in (2), and the second is described below. The data used in this simulation experiment consist of 100 values, x_i, $i = 1, \ldots, 100$, drawn from a standard Student t distribution on 4 degrees of freedom; 100 standard normal deviates, y_j, $j = 1, \ldots, 100$; and 200 $U(0,1)$ deviates, u_i, $i = 1, \ldots, 100$ and v_j, $j = 1, \ldots, 100$. The unnormalized t-density was assumed to have the form

$$cf(x) = \frac{c}{[1 + x^2/4]^{2.5}}$$

The marginal posterior density for c based on (7) for these data is depicted in Figure 1. Despite the irregular shape of this density, the posterior mass appears to concentrate near the correct value of $c = 0.375$. The posterior mean and standard deviation of this density are 0.372 and 0.0175, respectively. The 95% posterior probability interval extends from approximately 0.332 to 0.389. By comparison, the AOE was 0.370.

The second estimator considered for the normalizing constant c was defined as the value c_A that minimized

$$\left[\frac{1}{m} \sum_{i=1}^{m} \min\left(\frac{g(x_i)}{cf(x_i)}, 1 \right) - \frac{1}{n} \sum_{j=1}^{n} \min\left(\frac{cf(y_j)}{g(y_j)}, 1 \right) \right]^2 . \qquad (9)$$

Like the posterior density on c, this expression is motivated by the γ-coupling argument. Essentially, the value c_A that minimizes this sum of squares equates the expected values of a_c and b_c, where expectations are taken with respect to the uniform deviates $\{u_i\}$ and $\{v_j\}$. This estimator is consistent as both n and m become large, provided that $r < 1$. Because the influence of a single ratio of densities is bounded in (9), this estimator often provides a more robust estimate of the normalizing constant for small and moderate values of m and n than does the AOE. This point is explored further in the examples of the next section. The value of c that minimized (9) for these data was $c_A = 0.365$.

The value of the marginal posterior density displayed in Figure 1 derives from the information it provides in diagnosing the failure of a simulation experiment to adequately estimate a normalizing constant, or ratio of normalizing constants. As mentioned previously, the posterior distribution obtained through the analysis of the Bernoulli sequences $\mathbf{W}$ and $\mathbf{Z}$ cannot be efficient from a sampling perspective because of the additional randomization introduced by the uniform deviates $\{u_i\}$ and $\{v_j\}$. For this reason, it is important to investigate the degree to which the posterior distribution is inefficient, because extreme inefficiency might lead to a diagnostic procedure that was unjustifiably conservative. To this end, the experiment described above was expanded to study the sampling properties of the posterior distribution in relation to the mean square error of the AOE and the ad hoc estimator c_A.

In the expanded simulation, the mean of the normal distribution used in the definition of g was varied between 0 and 4 in increments of 0.5. For each value of the mean, samples of size 100 were drawn from both a Student t_4 density and from a normal distribution with specified mean and variance one. For each

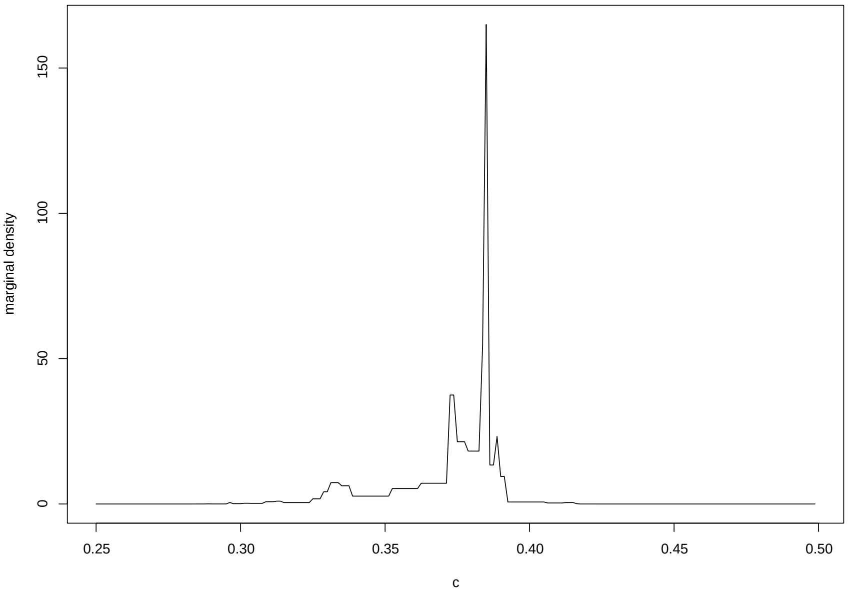

Figure 1: The marginal posterior distribution of the normalizing c of a Student t_4 density based on samples of size 100 from both a t_4 density and a standard normal density.

value of μ, 10,000 such sample were drawn, and the posterior mean, the posterior variance, the AOE, and c_A were calculated. Results from this experiment are summarized in Table 1.

The data in Table 1 suggest that the root mean square error (RMSE) of the posterior mean was approximately 50% larger than that of the AOE, and that the square root of the posterior variance tracks the RMSE of the posterior mean relatively well. This suggests that the posterior distribution provides a reasonably accurate, though slightly conservative, estimate of the uncertainty associated with both the posterior mean and the AOE. It is also interesting to note that the RMSE of the point estimator c_A nearly matched the efficiency of the AOE for all values of μ considered.

To evaluate the performance of the posterior distribution in assessing the uncertainty of individual estimates of the value of c, 90% and 95% posterior probability intervals were estimated for each simulation sample. Perhaps not surprisingly, the empirical probability that these credible regions contained the true value of the parameter corresponded nearly exactly to their specified levels. That is, the 90% posterior probability intervals contained the true value of c

	Estimator		
μ	c_A	Post. Mean	AOE
0.0	0.0087	0.0147 (0.017)	0.0083
0.5	0.0161	0.0260 (0.029)	0.0153
1.0	0.0279	0.0396 (0.044)	0.0265
1.5	0.0426	0.0573 (0.059)	0.0415
2.0	0.0569	0.0738 (0.078)	0.0556
2.5	0.0768	0.1022 (0.103)	0.0746
3.0	0.1030	0.1373 (0.138)	0.1002
3.5	0.1295	0.1716 (0.183)	0.1241
4.0	0.1601	0.2282 (0.242)	0.1526

Table 1: Comparison of the RMSE of three estimators of the Student t_4 density's normalizing constant. The numbers in parentheses in the second column represent the sample mean of the square root of the posterior variance obtained for each value of μ.

in 90.2% of the samples, while the 95% probability intervals contained the true value of the parameter in 95.1% of the samples. These proportions were stable across all values of μ.

4 Approximate posterior distributions for dependent samples

In many applications, independent draws from unnormalized densities are not available. Such is the situation when analyzing output from MCMC algorithms (e.g., Gelfand and Smith (1990) or Tierney (1994)) in which successive iterates in a chain are known to be correlated.

One strategy for dealing with MCMC output is to subsample the output chain to obtain what is, with high probability, an independent sample from the stationary distribution. For most MCMC algorithms, this can be done efficiently using the coupling-regeneration scheme proposed in Johnson (1998). Alternatively, the method of batch means (e.g., Bratley, Fox and Schrage (1987)) or methods based on analyzing autocorrelation functions might be used on a component-wise basis to determine the approximate subsampling interval required to obtain nearly independent draws from the target distribution.

A second strategy that might be used to deal with dependent draws from cf and g is to model the dependence in the sequences $\mathbf{W}$ and $\mathbf{Z}$ as two-state Markov chains. (A similar idea has been used by Raftery and Lewis (1992) to define a convergence diagnostic for MCMC output.) In this approach, posterior uncertainty in the value of c is modeled directly, although the validity of the posterior distribution depends on the efficacy of the two-state Markov chain assumption.

In order to generalize the results of the previous sections to the case in which

$\mathbf{W}$ and $\mathbf{Z}$ are each modeled as two-state Markov chains, let $p_{k,l}$ and $q_{k,l}$ denote the conditional probabilities

$$p_{k,l} = \mathbf{Pr}[W_i(c) = l | W_{i-1}(c) = k], \qquad q_{k,l} = \mathbf{Pr}[Z_j(c) = l | Z_{j-1}(c) = k],$$
$$l, k \in \{0,1\}, \qquad i = 1, \ldots, m \qquad j = 1, \ldots, n,$$

where $\mathbf{W}(c)$ and $\mathbf{Z}(c)$ are defined as before. For simplicity, the values of $W_1(c)$ and $Z_1(c)$ are regarded as constant, and beta priors with parameters $\alpha_{k,l}$ and $\gamma_{k,l}$ are assumed for the probabilities $p_{k,l}$ and $q_{k,l}$, respectively (for example, $p_{k,0} \sim B(\alpha_{k,0}, \alpha_{k,1})$). Also, let $A_{k,l}(c)$ and $B_{k,l}(c)$ denote the number of transitions observed from k to l in chains $\mathbf{W}(c)$ and $\mathbf{Z}(c)$, respectively.

With this notation and the assumption that $\mathbf{W}(c)$ and $\mathbf{Z}(c)$ can be modeled approximately as two-state Markov chains, the sampling density of $\mathbf{W}(c)$ and $\mathbf{Z}(c)$ may be written

$$\begin{aligned} f(\mathbf{W}(c), \mathbf{Z}(c) | \{p_{k,l}, q_{k,l}\}, c) \quad = \quad & p_{00}^{A_{00}}(1-p_{00})^{A_{01}} p_{10}^{A_{10}}(1-p_{10})^{A_{11}} \\ & \times q_{00}^{B_{00}}(1-q_{00})^{B_{01}} q_{10}^{B_{10}}(1-q_{10})^{B_{11}} \end{aligned} \tag{10}$$

Multiplying (10) by the joint prior density on the model parameters and dividing by the normalizing constant specific to the observed value of $\{A_{k,l}(c)\}$ and $\{B_{k,l}(c)\}$ leads to a posterior distribution proportional to

$$\begin{aligned} & f(\{p_{k,l}, q_{k,l}\}, c | A_{k,l}(c), B_{k,l}(c)) \propto \\ & \quad \frac{1}{K(A,B)} B(A_{00}+\alpha_{00}, A_{01}+\alpha_{01}) B(A_{10}+\alpha_{10}, A_{11}+\alpha_{11}) \\ & \quad \times\ B(B_{00}+\gamma_{00}, B_{01}+\gamma_{01}) B(B_{10}+\gamma_{10}, B_{11}+\gamma_{11}) \\ & \quad \times\ p_{00}^{A_{00}}(1-p_{00})^{A_{01}} p_{10}^{A_{10}}(1-p_{10})^{A_{11}} \\ & \quad \times\ q_{00}^{B_{00}}(1-q_{00})^{B_{01}} q_{10}^{B_{10}}(1-q_{10})^{B_{11}} \pi(c) \end{aligned} \tag{11}$$

where $K(A,B)$ is the integral of $\pi(c)$ over values of c leading to the observed values of $\{A_{k,l}\}$ and $\{B_{k,l}\}$.

As in the independence case, interest again focuses on the posterior distribution of c given that the marginal probabilities $Pr[W_i = 1]$ and $Pr[Z_j = 1]$ are equal. This condition is satisfied when

$$\frac{p_{01} q_{10}}{q_{01} p_{10}} = 1.$$

Because this event occurs with probability 0, some care is needed in specifying the conditional distribution of c given this event. To uniquely define this conditional posterior distribution, condition on the limiting event

$$\xi \equiv p_{01} q_{10} - q_{01} p_{10} \to 0.$$

Changing variables according to the transformation $\lambda = p_{01}$, $\rho = q_{01}$, $\nu = q_{10}$ and ξ yields a conditional distribution of (λ, ρ, ν, c) at $\xi = 0$ proportional to

$$
\begin{aligned}
&f(\lambda,\rho,\nu,c|\xi=0,A_{k,l}(c),B_{k,l}(c)) \propto \qquad (12)\\
&\quad \frac{1}{K(A,B)}B(A_{00}+\alpha_{00},A_{01}+\alpha_{01})B(A_{10}+\alpha_{10},A_{11}+\alpha_{11})\\
&\quad \times\ B(B_{00}+\gamma_{00},B_{01}+\gamma_{01})B(B_{10}+\gamma_{10},B_{11}+\gamma_{11})\\
&\quad \times\ \lambda^{A_{01}+\alpha_{01}}(1-\lambda)^{A_{00}+\alpha_{00}}\rho^{B_{01}+\gamma_{01}}(1-\rho)^{B_{00}+\gamma_{00}}\nu^{B_{10}+\gamma_{10}}(1-\nu)^{B_{11}+\gamma_{11}}\\
&\quad \times \left(\frac{\lambda\nu}{\rho}\right)^{A_{10}+\alpha_{10}}\left[1-\left(\frac{\lambda\nu}{\rho}\right)\right]^{A_{11}+\alpha_{11}}\frac{1}{\rho}\pi(c)
\end{aligned}
$$

The conditional posterior distribution (12) is not of a standard form, but establishing a MCMC chain to sample from it is not difficult. One possibility is to define a Metropolis-Hastings algorithm according to the following outline:

1. Given current values of (λ,ρ,ν,c), generate proposals for each of λ, ρ and ν using the beta density components of the conditional distribution in (12). For example, a candidate point for λ might be drawn from a beta density with parameters $(A_{01}+\alpha_{01},A_{00}+\alpha_{00})$. For each candidate point so generated, accept or reject with probability determined by the ratio of the value of
$$\left(\frac{\lambda\nu}{\rho}\right)^{A_{10}+\alpha_{10}}\left[1-\left(\frac{\lambda\nu}{\rho}\right)\right]^{A_{11}+\alpha_{11}}\frac{1}{\rho}$$
at the candidate and current value of the transition probability.

2. Given current values of (λ,ρ,ν), update c using a random-walk Metropolis-Hastings step. The proposal density for c is best specified on the log scale since posteriors for c are often approximately normal on that scale. Note that the dependence of $\{A_{k,l}\}$ and $\{B_{k,l}\}$ on c has been suppressed in (12), and values for these beta functions must be computed separately for the candidate and current values of c when calculating the acceptance probabilities for the sampler.

5 Examples with dependent data

Two experiments were conducted to study the sampling properties of the posterior distribution on c defined for dependent data in (12). The first experiment was similar to the experiment described at the end of Section 3 for independent data, but instead of using independent samples of t_4 and standard normal deviates, dependent draws were generated from these densities. These dependent sequences were sampled using a random-walk Metropolis-Hastings algorithm with $N(0,1)$ increments, initialized with an exact draw from each of the target densities. As before, sample sizes of 100 from both dependent sequences were drawn for each simulated dataset. Based on these sampled values, the three

	Estimator		
μ	c_A	Post. mean	AOE
0.0	0.0329	0.0350 (0.028)	0.0320
0.5	0.0470	0.0559 (0.043)	0.0457
1.0	0.0770	0.0819 (0.062)	0.0770
1.5	0.1174	0.1257 (0.096)	0.1207
2.0	0.1688	0.1829 (0.142)	0.1720
2.5	0.2377	0.2752 (0.238)	0.2408
3.0	0.3592	0.4183 (0.364)	0.3726
3.5	0.4725	0.5951 (0.599)	0.4916
4.0	0.6690	0.8480 (0.933)	0.6817

Table 2: Comparison of three estimators of the Student t_4 density's normalizing constant for dependent data. The number in parentheses in third column is the square root of the sample estimate of the posterior variance of the simulated values for the normalizing constant obtained from the MCMC algorithm.

estimates of the normalizing constant presented in Section 3 were again calculated, and the RMSE of each estimate was estimated by repeatedly generating 2,000 such samples. For each of these samples, 25,000 MCMC deviates were generated from the posterior distribution to estimate the posterior mean and variance. Results from this simulation experiment are displayed in Table 2.

In the simulation results for the dependent data, the square root of the mean posterior variance (indicated in parentheses in column 3) did not mimic the RMSE as accurately as it did in the previous experiment using independent data. However, the posterior variance does provide a reasonably good, though conservative, approximation to the sampling variability of both the AOE and c_A, and the 95% posterior probability intervals had approximately correct coverage probabilities in repeated sampling. For $\mu = 0$, the empirical probability that the 95% posterior probability interval on c contained the true value was .85; for $\mu = 4.0$ the coverage probability for this same credible interval was .70. Differences between the empirical and posterior probability intervals are due, in part, to the fact that the sequences $\mathbf{W}$ and $\mathbf{Z}$ used to define the posterior distribution were not exactly distributed as two-state Markov chains. Despite these differences, the posterior distribution provided a simple mechanism for assessing the magnitude of the uncertainty for all three estimates displayed in the table.

It is also worth noting that the estimator c_A had the lowest RMSE for values of $\mu > 1$.

As a final example of this methodology, posterior distributions were next computed in the more challenging setting of estimating the ratio of normalizing constants for Ising models with different correlation parameters.

The Ising models considered here were defined on 32×32 square lattices, and were parameterized so that the range of the random variable x_s associated with site s in each lattice was ± 1. Letting $s \sim t$ denote the fact that s and t

are nearest neighbors (either horizontal or vertical), the density on the vector $\mathbf{x} = (x_{1,1}, \ldots, x_{32,32})$ may be written

$$p(\mathbf{x}) = \frac{1}{c(\beta)} \exp\left(\beta \sum_{s\sim t} x_s x_t\right). \tag{13}$$

Toroidal constraints were imposed on the boundary of the lattice so that, for example, $s = (1,1) \sim t = (32,1)$. Interest in this example focuses on estimating ratios of $c(\beta_1)/c(\beta_2)$ for values $(\beta_1, \beta_2) \in (0.2, 0.5)$.

Estimating all ratios in this range might be accomplished with a bridge sampling scheme (Meng and Wong, 1996) using sampled values from only a few Ising models defined for values of β selected from the interval $(0.2, 0.5)$. To implement such a strategy, a finite number of β values from this interval would be selected, and samples from the corresponding Ising models would subsequently be used to estimate the ratio of normalizing constants for all other models in this range. However, an obvious requirement for this strategy to succeed would be accurate evaluation of the ratios of normalizing constants at the neighboring values of selected β's.

Suppose then that we are interested in investigating the uncertainty in estimated values of the ratio $(c_\beta/c_{\beta-0.1})$, $\beta = 0.3, 0.4, 0.5$, using blocks of 5000 sampled lattices obtained from a simulation experiment comprised of the following steps.

First, for each value of β, a 32×32 lattice was randomly initialized using independent Bernoulli random variables with success probabilities equal to 0.5. Following initialization, 50,000 "burn-in" iterations were performed using a Gibbs sampler that updated each site in the lattice according to its full conditional distribution. After burn-in, 500,000 additional Gibbs sampling updates were made of each site in each of the lattices. Finally, these sampled lattices were split into 100 consecutive blocks of 5,000 dependent draws. Pairs of these blocks were then used to calculate the posterior mean, the AOE, and the estimator c_A for each of the three ratios of normalizing constants. Simultaneously, the sampling properties of these estimators, along with the posterior variance, were evaluated by computing these estimates for each of the 100 pairs of blocks obtained in the simulation study.

Table 3 displays the results from this simulation experiment. The second column of this table lists the exact values of the ratios of normalizing constants as determined by analytical methods described originally by Onsager (1944), and later more transparently by Kaufmann (1949). As this table demonstrates, all three estimators perform poorly in this experiment. Of course, the high relative errors of the estimators are caused by the relatively small sample sizes used in computing each estimate and the high dependence between successive draws from the Gibbs samplers. From this table, it is clear that the only ratio for which reasonably accurate estimates were obtained was $c(0.3)/c(0.2)$. In this case, the relative error of the AOE averaged about 19%, which was slightly

	Estimator			
	True value	c_A	Post. mean	AOE
$c(0.3)/c(0.2)$	8.3×10^{24}	1.7×10^{24}	2.1×10^{24} (2.3×10^{24})	1.6×10^{24}
$c(0.4)/c(0.3)$	3.1×10^{39}	1.2×10^{39}	3.7×10^{47} (1.5×10^{48})	9.7×10^{38}
$c(0.5)/c(0.4)$	1.3×10^{65}	9.1×10^{66}	1.3×10^{75} (8.6×10^{75})	1.9×10^{67}

Table 3: Estimated ratios of normalizing constants for Ising model. The first column indicates the ratio being estimated, the second the true value of this ratio based on the formulas given in Kaufmann (1949). The third column provides the RMSE for the 100 sampled values of c_A. The fourth column provides the RMSE for the posterior mean based on 50,000 MCMC iterations for each of the 100 blocks of 5,000 sampled Ising lattices. The fourth column is the RMSE for the asymptotically optimal estimate.

better than both the posterior mean and c_A.

The posterior distribution for $c(0.3)/c(0.2)$ displayed reasonably good sampling properties as well. Of the 100 95% probability intervals generated from the posterior distributions, 98 contained the true value of this ratio. Similarly, 91 of the 90% probability intervals contained the true parameter value.

Estimates of the ratios of the normalizing constants involving larger values of β had substantially higher relative errors. In the case of $c(0.5)/c(0.4)$, the RMSE of the AOE tended to be approximately two orders of magnitude larger than the value of the ratio itself, averaging about 150. This fact is, perhaps, not surprising since the estimator is based on only 5,000 highly-correlated draws from each of the target distributions. However, diagnosing this poor performance without prior knowledge of the true value of this ratio and only a single sample of 5000 draws could be problematic using standard sampling-based diagnostics.

To highlight the difficulties inherent in diagnosing the failure of estimators in this setting, a non-parametric bootstrap procedure was used to estimate the sampling distribution of the AOE in each of the 100 blocks of sampled lattices. The specific bootstrap procedure employed consisted of repeatedly resampling 5,000 lattices from the 5,000 sampled values obtained in each block. Each set of resampled values was then used to estimate the ratio $c(0.5)/c(0.4)$. Five thousand resamples in each block were used to assess the sample variability of the estimator.

The results of this experiment were not encouraging. The average ratio of the AOE to the standard deviation of the bootstrapped values of the AOE was 3.94. Taken at face value, this result would seem to imply that the ratio $c(0.5)/c(0.4)$ had been determined at least to the within the correct order of magnitude in many repetitions of the experiment. In fact, however, none of the estimated values of the AOE were correct to within a factor of 3, and 90% of the estimates were off by a factor of more than 40.

In contrast, each of the 100 posterior distributions obtained for $c(0.5)/c(0.4)$ provided a clear indication that the experimental data was not sufficient for estimating this ratio. Ninety-nine of the 100 posteriors had standard deviations that were larger than the posterior mean, with the average ratio of the posterior mean to posterior standard deviation being 0.30. This statistic alone suggests that the posterior mean, and by extension the other estimators, were not well determined by the blocks of 5,000 sampled Ising lattices. However, the strongest evidence that either additional data or a different design was needed to estimate this ratio is provided through an examination of the histogram summaries of the posterior distributions. Although space constraints prohibit the display of all of these histograms, the histogram summary of the 50,000 sampled values from the posterior distribution obtained from the first pair of sampled blocks of Ising lattices is depicted in Figure 2. (To exemplify the differences between the uncertainty reflected by the posterior distribution on this ratio and the uncertainty captured by the bootstrap procedure, a histogram of the bootstrapped values for the same data is also provided in this figure). The range of values assigned non-negligible mass under the posterior distribution is extremely wide, and indicates that the posterior uncertainty in the ratio $c(0.5)/c(0.4)$ spans several orders of magnitude. For example, the 0.975th quantile is approximately 20,000 times larger than the 0.025th quantile. In other words, the 95% credible interval spans a range of values that differ by a factor of 20,000. The geometric mean of this same factor across all 100 posterior distributions was 169,000. All of these factors, taken individually or together, provide ample evidence that this experimental design was simply not viable for estimating this ratio of normalizing constants.

To improve our hypothetical bridge sampling scheme, a revised experiment was next performed in an attempt to improve the estimation of the ratio of Ising normalizing constants on the interval $(0.2, 0.5)$. To this end, the spacing of β was reduced to 0.05 and the MCMC scheme described above repeated. Results obtained from the revised experiment are displayed in Table 4.

As might be expected, the estimates summarized in Table 4 were significantly better than those summarized in Table 3. In addition, this table reveals two interesting trends. First, the square root of the average posterior variance tracks the RMSE of the posterior mean and the other two estimators relatively well for these data, which in turn suggests that the posterior variance again provides a reasonable summary of the uncertainty of these estimators. This assertion is supported by an examination of the proportion of say, the 99%, probability intervals that covered the true parameter value. For these data, this proportion ranged from 99% coverage for the 100 99% probability intervals obtained for the ratio $c(.20)/c(.25)$, to 91% coverage for the 100 99% posterior probability intervals obtained for $c(.50)/c(.45)$. Although these intervals should not be taken too literally since the Bernoulli sequences $\mathbf{W}$ and $\mathbf{Z}$ are not exactly two-state Markov chains, the facts that the posterior standard deviation provides an approximate tracking of the RMSE, and that the empirical coverage probabilities and posterior probability intervals approximately agree, support the use of the posterior distribution as a diagnostic for determining when ratios

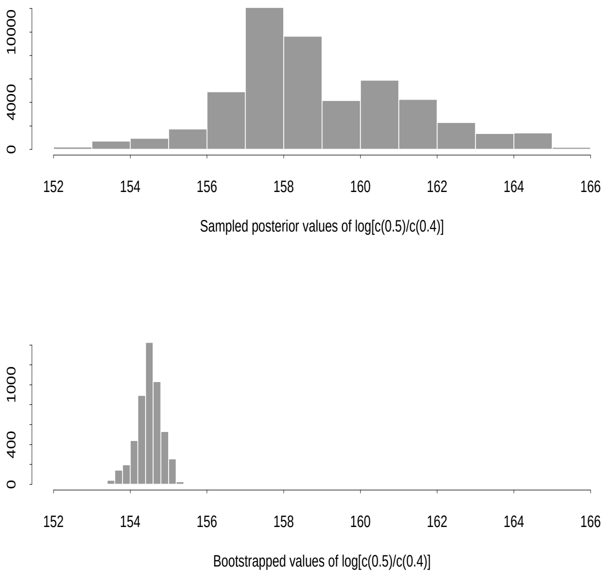

Figure 2: Histogram estimates of the estimated uncertainty in the ratio $c(0.5)/c(0.4)$. The top figure represents a histogram estimate of the posterior distribution on the natural logarithm of the value $c(0.5)/c(0.4)$. The lower figure depicts the distribution of 5,000 bootstrap samples obtained from the same data. The true value of the $\log(c(0.5)/c(0.4))$ is 149.

	Estimator			
	True value	c_A	Post. mean	AOE
$c(0.25)/c(0.20)$	8.5×10^{10}	5.2×10^{9}	5.5×10^{9} (3.9×10^{9})	5.1×10^{9}
$c(0.30)/c(0.25)$	9.7×10^{13}	1.9×10^{12}	3.0×10^{12} (4.8×10^{12})	1.8×10^{12}
$c(0.35)/c(0.30)$	3.5×10^{17}	1.1×10^{16}	1.5×10^{16} (2.1×10^{16})	1.0×10^{16}
$c(0.40)/c(0.35)$	8.8×10^{21}	3.2×10^{20}	4.7×10^{20} (7.6×10^{20})	3.3×10^{20}
$c(0.45)/c(0.40)$	2.6×10^{28}	1.6×10^{28}	1.8×10^{28} (9.7×10^{27})	1.7×10^{28}
$c(0.50)/c(0.45)$	5.1×10^{36}	7.0×10^{35}	7.1×10^{35} (4.0×10^{35})	7.5×10^{35}

Table 4: Estimated ratios of normalizing constants for Ising model. The first column indicates the ratio being estimated, the second the true value of this ratio based on the formulas given in Kaufmann (1949). The third column provides the RMSE for the 100 sampled values of c_A. The third column provides the RMSE for the posterior mean based on 50,000 MCMC iterations for each of the 100 blocks of 5,000 sampled Ising lattices; the square root of the average posterior variance appears in parentheses. The fourth column provides the RMSE for the AOE.

of normalizing constants have been accurately estimated.

Finally, it is interesting to note that the ad hoc estimator c_A outperformed the AOE for the larger values of β. As in the previous example involving dependent data, the superior performance of the estimator c_A for well-separated target densities can be attributed to its more robust formulation and the failure of the large-sample properties of the AOE estimator to apply in this setting. Asymptotics fail in these cases because of the high dependence between the iterates of the Gibbs sampler, and because of the large total variation distance between the target densities.

6 Summary

Simulation-based estimates of normalizing constants, or ratios of normalizing constants, are known to perform poorly in many complex statistical models. For this reason, assessing the uncertainty of these estimators is essential, though doing so can be difficult because of the highly non-Gaussian distributions of quantities used in their definition.

The posterior distributions proposed in this article provide a workable solution to this problem. Although the theoretical justification of these posterior distributions is somewhat involved, in practice they are easy to apply. Using this methodology, samples from a posterior distribution on a normalizing constant or ratio of normalizing constants can be generated automatically using software available from the author's homepage or from STATLIB. Because the data upon which these posteriors are based consists only of the values of (unnormalized) densities, and because similar values are also required to compute other standard estimators like the AOE, this methodology should prove extensible to a wide range of applications in which normalizing constants must be estimated.

Acknowledgments

The author would like to thank Michael Lavine, Peter Mueller, an anonymous referee and an anonymous associate editor for helpful comments and observations.

References

[1] Bennett, C.H. (1976), "Efficient Estimation of Free Energy Differences from Monte Carlo Data," *Journal of Computational Physics*, 22, 245-268.

[2] Bratley, P., Fox, B. and Schrage, L. (1987), *A Guide to Simulation*, New York: Springer-Verlag.

[3] DiCiccio, T.J., Kass, R.E., Raftery, A.E. and Wasserman, L. (1997), "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of the American Statistical Association*, (92), 903-915.

[4] Gelfand, A.E. and Dey, D.K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society*, series B, 56, 501-514.

[5] Gelfand, A.E. and Smith, A.F.M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.

[6] Gelman, A. and Meng, X.L. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163-185.

[7] Geyer, C.J. and Thompson, E.A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data," (with discussion) *Journal of the Royal Statistical Society*, series B, 54, 657-699.

[8] Geyer, C.J. (1994), "Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo," Technical Report 568, School of Statistics, University of Minnesota.

[9] Johnson, V.E. (1998), "A Coupling-Regeneration Scheme for Assessing Convergence of Markov Chain Monte Carlo Algorithms," *Journal of the American Statistical Association*, 238-248.

[10] Kass, R.E. and Raftery, A.E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773-795.

[11] Kaufman, B. (1949), "Crystal Statistics. II. Partition Function Evaluated by Spinor Analysis," *Physical Review*, **76**, 1232-1243.

[12] Lewis, S.M. and Raftery, A.E. (1997), "Estimating Bayes Factors Via Posterior Simulation with the Laplace-Metropolis Estimator," *Journal of the American Statistical Association*, 92, 648-655.

[13] Lindvall, T. (1992), *Lectures on the Coupling Method*, New York:John Wiley and Sons.

[14] Meng, X.L. and Wong, W.H. (1996), "Simulating Ratios of Normalizing Constants Via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 4, 831-860.

[15] Newton, M.A. and Raftery, A.E. (1994), "Approximate Bayesian Inference with Weighted Likelihood Bootstrap," (with discussion) *Journal of the Royal Statistical Society*, series B, 56, 3-48.

[16] Onsager, L. (1944), "Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition," *Physical Review*, **65**, 117-149.

[17] Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1995), *Numerical Recipes in C*, 2nd edition, Cambridge:Cambridge University Press.

[18] Raftery, A.E. (1995), "Hypothesis Testing and Model Selection via Posterior Simulation," in *Practical Markov Chain Monte Carlo,* eds. W. Gilks, S. Richardson, D.J. Spiegelhalter, London: Chapman and Hall, 163-188.

[19] Raftery, A.E. and Lewis, S.M. (1992), "How Many Iterations in the Gibbs SamplerΓ," in *Bayesian Statistics 4*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, Oxford: Oxford University Press, 765-776.

[20] Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley and Sons, Inc.

[21] Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," (with discussion) *Annals of Statistics*, **22**, 1701-1762.

[22] Tierney, L. and Kadane, J.B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82-86.