Statistics and Computing **14**: 59–66, 2004

Bayesian marginal inference via candidate's formula

CHUHSING KATE HSIAO[1], SU-YUN HUANG[2] and CHING-WEI CHANG[3]

[1]*Division of Biostatistics, Institute of Epidemiology, National Taiwan University, Taipei, 100, Taiwan, R.O.C.*
ckhsiao@ha.mc.ntu.edu.tw
[2]*Institute of Statistical Science, Academia Sinica, Taipei, 115, Taiwan, R.O.C.*
syhuang@stat.sinica.edu.tw
[3]*Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, 115, Taiwan, R.O.C.*
ashely@nhri.org.tw

Received December 2001 and accepted June 2003

Computing marginal probabilities is an important and fundamental issue in Bayesian inference. We present a simple method which arises from a likelihood identity for computation. The likelihood identity, called Candidate's formula, sets the marginal probability as a ratio of the prior likelihood to the posterior density. Based on Markov chain Monte Carlo output simulated from the posterior distribution, a nonparametric kernel estimate is used to estimate the posterior density contained in that ratio. This derived nonparametric Candidate's estimate requires only one evaluation of the posterior density estimate at a point. The optimal point for such evaluation can be chosen to minimize the expected mean square relative error. The results show that the best point is not necessarily the posterior mode, but rather a point compromising between high density and low Hessian. For high dimensional problems, we introduce a variance reduction approach to ease the tension caused by data sparseness. A simulation study is presented.

Keywords: Bayes factor, Gibbs sampler, kernel density estimation, marginal likelihood, marginal likelihood identity, Markov chain Monte Carlo, Metropolis-Hasting algorithm

1. Introduction

In Bayesian inference, a joint posterior distribution is available through the formulation of the likelihood function and a prior distribution of the parameter of interest. Consider an $n \times 1$ vector of observations y with sampling probability density $f(y \mid \theta)$ given the $p \times 1$ vector of parameters $\theta = (\theta_1, \ldots, \theta_p)$. Assume that the parameter θ has a prior density $\pi_k(\theta)$ under model M_k $(k = 1, 2, \ldots, K)$. In Bayesian inference, including variable selection, model selection, or hypothesis testing, one may need to evaluate the marginal density of the sample data

$$m(y \mid M_k) = \int_{R^p} f(y \mid \theta)\pi_k(\theta)\, d\theta. \tag{1}$$

In other Bayesian analysis, one may need to evaluate the marginal posterior density of the form

$$\pi(\eta \mid y) = \frac{1}{m(y)} \int_{\{\theta : g(\theta)=\eta\}} f(y \mid \theta)\pi(\theta)\, d\theta_1 \cdots d\theta_{p-d}, \tag{2}$$

where the parameter of interest is $\eta = g(\theta)$ for some function g and where $\eta \in \Omega \subset R^d$ with $1 \leq d < p$. The computation of marginal probability has long been an important and challenging issue in Bayesian inference. The derivation of its value is necessary when model selection is of interest, or when the posterior distributions, moments, Bayes factors, or predictive densities are requested. The quantity $m(y)$, or $m(y \mid M_k)$, is sometimes referred to as a normalizing constant, especially when the intergrand is taken to be the product of likelihood function and prior density. Much efforts have since been placed upon the estimation of $m(y)$.

Several authors (e.g., Mosteller and Wallace 1964, Tierney and Kadane 1986) proposed an analytic approximation, Laplace's method, to approximate the integration when the data size is large. The basic idea is to use a normal probability density function to approximate the integrand. This approximation has been shown fairly accurate in many applications. Nevertheless, some difficulties may arise. For instance, the approximation

0960-3174

requires the evaluation of the mode and variance, which may not be straightforward in many complex models or applications. In addition, the integrand itself may involve further integration of other nuisance parameters and thus may not be easily derived analytically. Furthermore, when the parameters are close to the boundary, e.g., parameters in a constrained space such as the variance components models, the usual Laplace method may fail. Erkanli (1994) and Hsiao (1997) proposed a modification of Laplace approximation for boundary cases. Their methods applied to cases where the local maximum likelihood estimate lies at or near the boundary. Pauler, Wakefield and Kass (1999) proposed a more general Laplace approximation for boundary cases, but it requires the knowledge of both unrestricted and restricted MLEs.

The advent of the Markov chain Monte Carlo method has provided easy means to draw samples from the unnormalized target distribution (see, for instance, Hastings 1970, Geman and Geman 1984, Gelfand and Smith 1990, Gilks Richardson and Spiegelhalter 1996, among many others). Several MCMC simulation based methods for computing marginal probabilities have been developed, including the importance sampling (Geweke 1989, Chen and Shao 1997, for ratio importance sampling), the bridge sampling (Meng and Wong 1996), the path sampling (Gelman and Meng 1998), Chib's method using Gibbs output (Chib 1995), the harmonic mean (Newton and Raftery 1994, Gelfand and Dey 1994), and hybrid methods by combining simulation and Laplace approximation (Lewis and Raftery 1997, DiCiccio *et al*. 1997, Huang, Hsiao and Chang 2003).

In this article, we present a simple approach arising from the identify (3) below by employing a kernel estimate for $\pi(\theta \mid y)$. Besag (1989) gave the Candidate's formula for Bayesian prediction

$$m(y) = \frac{f(y \mid \theta)\pi(\theta)}{\pi(\theta \mid y)}. \tag{3}$$

This equation holds for all θ values in the support of prior density $\pi(\cdot)$. It is also known as the marginal likelihood identity (Chib 1995). Although this kernel estimate is easy to understand conceptually, its application has long been limited due to the estimation difficulty known as "the curse of dimensionality" arising from data sparseness in high dimensional problems. In this article, we introduce an averaging process of evaluations over various θ values to effectively utilize more of the posterior sample and to ease the tension of data sparseness. In Section 2 we demonstrate that the derivation of nonparametric Candidate's estimate can be easy and inexpensive. We also investigate its behavior when the posterior sample size is larger. The best point for the estimation is derived next to attain the minimum mean square relative error. To successfully implement the procedure, we also discuss the guidelines for obtaining the posterior sample, deriving the bandwidth for kernel estimate, and handling the multi-dimensional problems. In Section 3 a simulation study is carried out using two groups of probability distributions. The first group consists of various unimodal distributions, while the second group consists of those that are highly skewed and/or have the mode close to the boundary. A concluding discussion is given in Section 4. All proofs are shown in the Appendix.

2. Nonparametric Candidate's method

2.1. *The approach*

For simplicity, we use notation C to denote the normalizing constant, $C \equiv m(y) = (y \mid \theta)\pi(\theta)/\pi(\theta \mid y)$. Any value of θ in the support of prior density $\pi(\cdot)$ provides the same answer. Based on simulated posterior sample $\{\theta^{(1)}, \ldots, \theta^{(m)}\}$, the normalizing constant can be estimated by plugging in an estimate for the posterior density

$$\hat{C} \equiv \frac{f(y \mid \theta)\pi(\theta)}{\hat{\pi}(\theta \mid y)}. \tag{4}$$

Here we adopt a kernel estimate given by

$$\hat{\pi}(\theta \mid y) = \frac{1}{m \mid H \mid^{1/2}} \sum_{i=1}^{m} \mathcal{K}\big((\theta - \theta^{(i)})^T H^{-1}(\theta - \theta^{(i)})\big), \tag{5}$$

where H is a $p \times p$ symmetric positive definite matrix, $|H|$ is the determinant of H, and $\mathcal{K}(\cdot) : R \to R^+$ is a kernel function satisfying the following two conditions

C1. $\mathcal{K}$ integrates to one, $\int_{R^p} \mathcal{K}(\theta^T\theta)\, d\theta = 1$, and
C2. $\mathcal{K}$ is order 2 kernel in the sense that

$$\int_{R^p} \theta_j \mathcal{K}(\theta^T\theta)\, d\theta = 0, \ \forall j = 1, 2, \ldots, p,$$

$$k_2 = \int_{R^p} \theta_j^2 \mathcal{K}(\theta^T\theta)\, d\theta > 0, \ \forall j = 1, 2, \ldots, p,$$

$$v = \int_{R^p} \mathcal{K}^2(\theta^T\theta)\, d\theta < \infty.$$

Nonparametric density estimation has been a well developed research topic in recent decades. See, for references, Silverman (1986), Scott (1992) and Simonoff (1996) for general theory and applications of the techniques. Here we intend to use the posterior samples generated via MCMC methods to obtain the estimate of $\pi(\theta \mid y)$. Although MCMC samples are usually not independent, there have been several approaches to reducing the correlations and computing the covariance (see Geyer 1992, for more details and references). When given a sufficiently long burn-in of iterations and by taking MCMC outputs sufficiently long apart, the posterior sample is close to i.i.d.

2.2. *Theoretical results*

We state in the following theorems the asymptotic order of accuracy of the nonparametric Candidate's estimate and the best point for evaluating the estimate.

Theorem 1. *Assume that the likelihood function and the prior have continuous second derivatives in a neighborhood of a certain interior point* $\theta \in supp\{\pi(\theta)\}$. *Also assume that the kernel*

function $\mathcal{K}$ satisfies conditions C1 and C2, and that as $m \to \infty$, trace$(H) \to 0$ and $m|H|^{1/2} \to \infty$, then

$$E_{\theta^{(i)},\ldots,\theta^{(m)}|y}\left(\frac{C}{\hat{C}} - 1\right)^2 = O(\{\text{trace}(H)\}^2) + O(m^{-1}|H|^{-1/2}),$$

where the expectation $E_{\theta^{(1)},\ldots,\theta^{(m)}|y}$, is taken with respect to the joint posterior distribution of $\theta^{(1)}, \ldots, \theta^{(m)}$ given y. Furthermore, choose the bandwidth matrix H so that $\lambda_j(H) = O(m^{-2/(4+p)})$ for all $j = 1, \ldots, p$, then

$$E_{\theta^{(i)},\ldots,\theta^{(m)}|y}\left(\frac{C}{\hat{C}} - 1\right)^2 = O\big(m^{-4/(4+p)}\big),$$

where λ_j denotes the jth largest eigenvalue.

Theorem 1 states that as long as the size of posterior sample gets large, the order of Candidate's estimate can be guaranteed. Once Theorem 1 is established, the next question arising naturally is that at what θ value the posterior should be estimated. Because the Candidate's formula is valid for all θ in the prior support, any choice of θ, as long as the posterior density is smooth in a neighborhood of this θ, is theoretically valid. However, consideration for efficiency suggests that the estimator (4) should be evaluated at high density and low Hessian points. We therefore derive the best point below under the mean square relative error criterion.

Theorem 2. *The nonparametric Candidate's estimator* (4) *achieves asymptotically the minimum value for $E_{\theta^{(1)},\ldots,\theta^{(m)}|y}$ $(C/\hat{C} - 1)^2$ at*

$$\theta^* = \arg\min_{\theta} \frac{\text{abs}\{|\pi^{(2)}(\theta|y)|\}}{[\pi(\theta \mid y)]^{p+2}}, \tag{6}$$

where $\pi^{(2)}(\theta \mid y)$ is the second derivative matrix, $|\cdot|$ stands for the matrix determinant and abs$\{\cdot\}$ *stands for the absolute value. Since $\pi(\theta \mid y)$ is proportional to $f(y \mid \theta)\pi(\theta)$, it can be replaced by $f(y \mid \theta)\pi(\theta)$ in both the numerator and denominator in Eq.* (6)*.*

To attain the minimum mean square relative error, it is most accurate to evaluate the Candidate's estimate at above θ^*. Note that the best point may not coincide with the posterior mode. For instance, the Candidate's estimate for a normal posterior is most accurate when evaluated at points of plus or minus one standard devation away from the mode. Later in Section 3 we demonstrate by simulation that the estimate evaluated at the best point does have better performance, particularly when the underlying posterior density is highly skewed.

2.3. *Guidelines for implementation*

To make the nonparametric Candidate's estimate easy to implement, listed below are some guidelines. These guidelines are useful in obtaining an appropriate sample $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}\}$, choosing an optimal bandwidth, constructing the procedures for high dimensional problems, and dealing with boundary cases.

1. *Standardization*. Because the marginal density $m(y)$ is invariant under changes of prior parameters by an arbitrary transformation (as long as such transformation is invertible to avoid degeneration), we may use the standardized posterior sample for derivation. Other types of transformation which reduce the degree of correlation and skewness in the sample may be considered as well. The matrix H in (5) can then be taken as $h^2 I_p$, where $h > 0$ is a scalar serving as the kernel bandwidth and I_p is the $p \times p$ identity matrix. The posterior standardization procedure will ease the estimation of $m(y)$. The transformation of parameters can even be carried out before running a Markov chain. This procedure helps to achieve a faster convergence rate of the chain and to obtain an easier kernel estimate later on. On the other hand, if the unnormalized posterior sample is available, it is possible to find an appropriate transformation to standardize, in a rough way, the parameters.
2. *Bandwidth Selection*. The choice of the bandwidth h is important. The theoretical value for the optimal bandwidth, in the asymptotic IMSE sense, involves the unknown parameters of the second derivatives of the posterior density $\pi^{(2)}(\theta \mid y)$ (see, for instance, Parzen 1962, Simonoff 1996). For convenience, the second derivatives can be estimated using a reference distribution. If we choose the standard normal distribution as our reference posterior (with the posterior having been standardized), some straightforward calculation leads to a rule of thumb bandwidth choice given by
$$h_{\text{opt}} = \left\{2^{p+2}\pi^{p/2} v k_2^{-2} \big/ (p+2)\right\}^{1/(p+4)} m^{-1/(p+4)},$$
when v and k_2 are defined in the kernel condition C2. Alternatively, Terrell (1990) suggested making a scale transformation on the data to get a better understanding of the second derivatives.
3. *Averaging over Points*. For a high dimensional problem, we suggest to average over a set of $\hat{C} = f(y \mid \theta)\pi(\theta)/\hat{\pi}(\theta \mid y)$ evaluated at various θ values. The averaging process effectively utilizes a lot more posterior sample points, which can lessen the tension of data sparseness and stabilize the high variation. Let M denote the number of θ points used in the averaging process. The corresponding rule of thumb bandwidth choice becomes
$$h_{\text{opt}} = \left\{2^{p+2}\pi^{p/2} v k_2^{-2}/(M(p+2))\right\}^{1/(p+4)} m^{-1/(p+4)}.$$
Again, this suggestion is based on the criterion of minimal asymptotic IMSE with the standard normal distribution as our reference posterior. Suggestion on the size M is discussed in the simulation study in Section 3.
4. *Uniform Kernel*. It helps to use a uniform kernel (either over a ball or a cube) for posterior density estimate to reduce the computational load, especially in high dimensional problems. When a uniform kernel is used to compute the posterior density estimate at a point θ, one simply counts the proportion of posterior samples falling into the ball (or the cube) centered at θ with radius h (or with edge width h). This

procedure will save time in computation with only a small loss in efficiency.

5. *Estimation at Boundary*. As long as the estimate is evaluated at point(s) chosen in the interior of the parameter space, the nonparametric Candidate's estimate discussed above is valid for posteriors with modes located either in the interior or at the boundary. The estimate can also be modified to adapt to the boundary by evaluating at boundary point(s) using a one-sided kernel or any other type of boundary kernels. For instance, a right-sided boundary kernel may be used in the variance component models to avoid putting any weight beyond the left boundary. For kernel estimation with boundary adjustment, see, e.g., Gasser and Müller (1979), Rice (1984), Müller (1991, 1993), Cowling and Hall (1996), and Müller and Stadtmüller (1999).

We shall explain the rationale behind the averaging process in item 3 above. The deficiency encountered in high dimensional estimation, known as the curse of dimensionality, is due to the data sparseness in a high dimensional space. This problem is not specific to kernel estimates; indeed, it exists in almost all kinds of nonparametric methods. This phenomenon persists, unless a certain model structure is imposed to confine either explicitly or implicitly the data to an effective space of lower dimension. Fortunately, the deficiency problem encountered in this particular marginal probability computation can be easily handled on two accounts: (1) the posterior sample is cheap to obtain, and (2) for the kernel estimate (of order 2) in a p-dimensional problem, it has pointwise bias of order $O(h^2)$ and pointwise variance of order $O((mh^p)^{-1})$. As the dimension p increases, the variance of kernel estimation gets higher and higher, while the bias order remains the same. That is, the deficiency is caused by the high variation of estimation in a high dimensional space and occurs for the estimation of the posterior density at all points. However, our interest is not in the posterior estimation itself, but rather in using it for computing marginal probability. Also, because the Candidate's formula for marginal density is valid for all θ values in the prior support, one may average over a set of M many θ values to stabilize the estimation variance to the order $O((Mmh^p)^{-1})$ utilizing more posterior sample points.

3. A simulation study

In this section, we compare the nonparametric Candidate's estimate with the volume-corrected Laplace estimate. Both methods are easy to implement for routine use. For other methods, see DiCiccio *et al.* (1997) for theoretical remarks and comparisons. Other methods including the importance sampling, the bridge sampling, and the Chib's method require evaluations of the posterior density at all posterior sample points. In contrast, the nonparametric Candidate's and the volume-corrected Laplace estimates require only one single evaluation of the posterior density. The computational load for Candidate's and the Laplace methods is not as heavy as that for the above mentioned methods. In other words, when the evaluation of the posterior is difficult or expensive, it is suggested to use the Laplace-type estimates or the nonparametric Candidate's method. Therefore, our simulation study is limited to the comparison of the volume-corrected Laplace estimate and the nonparametric Candidate's estimate.

Recall the Laplace approximation

$$C_{\text{Lap}} = \frac{f(y \mid \theta^*)\pi(\theta^*)}{\phi(\theta^*; \theta^*, \Sigma^*)},$$

where $\phi(\cdot; \theta^*, \Sigma^*)$ is the multivariate normal density function with mean θ^* and covariance matrix Σ^*. This approximation has $C = C_{\text{Lap}}\{1 + O(n^{-1})\}$, where n is the size of the observed data y. Lewis and Raftery (1997) derived the mode and variance estimates for ϕ based on simulated Markov chains and combined them with the Laplace approximation. The accuracy of the Laplace approximation C_{Lap} depends heavily on the shape of the posterior $\pi(\theta \mid y)$. In other words, it depends on the degree of resemblance between $\pi(\theta \mid y)$ and the density of a normal distribution. DiCiccio *et al.* (1997) proposed to improve the Laplace approximation. They used a volume of probability α around the mode to adjust the Laplace approximation:

$$C_{\text{vol-cor}} = \frac{f(y \mid \theta^*)\pi(\theta^*)}{\phi(\theta^*; \theta^*, \Sigma^*)} \cdot \frac{\alpha}{\hat{P}},$$

where α was recommended to be fixed at .05 and $\hat{P}$ is the proportion of posterior samples falling into the ball $B_r(\theta^*, \Sigma^*) = \{\theta : (\theta - \theta^*)'(\Sigma^*)^{-1}(\theta - \theta^*) \le r^2\}$ with radius r determined by $\int_{B_r(\theta^*, \Sigma^*)} \phi(\theta; \theta^*, \Sigma^*)\, d\theta = \alpha$. In a later work by Huang, Hsiao and Chang (2003), the choice of α is investigated.

3.1. *One dimensional case with interior mode*

Table 1 lists the results from simulations based on four different distributions as the nominal posteriors. The four distributions are standard normal, Student-t with degrees of freedom 5, Student-t with degrees of freedom 3, and gamma(2,1). These distributions are chosen to represent various shapes of posterior distributions such as symmetry with light tails, symmetry with heavy tails, and skewed distributions. In each replication, either $m = 1{,}000$, 10,000 or $m = 100{,}000$ posterior samples are drawn from each distribution. There are 100 replications under each setting. We use the mean square relative error $(C/\hat{C} - 1)^2$ as the measure of accuracy.

The Candidate's estimates are evaluated at three different points: the mode, the mean, and the best point. For the normal distribution, the best points are the mode plus or minus one standard deviation. Here we consider the one at the mode plus one standard deviation. The best point for the Student-t is its mode. For gamma(2,1), its best point locates at the mean. Table 1 lists the averages and standard errors of the mean square relative errors from 100 replications. Some numbers are omitted as they are the same as the numbers in the best point column. It can be seen that all three Candidate's estimates are more accurate than the volume-corrected Laplace estimates. Among the three Candidate's estimates evaluated at different points, the one at the best point does outperform the rest.

Table 1. *These are the averages and standard errors of mean square relative errors from 100 replications for Laplace volume correction and Candidate's estimates*

	vol. cor. Laplace (mode)	Candidate (mode)	Candidate (best point)	Candidate (mean)
Normal				
$m = 1{,}000$	16.1 (2.47)	3.07 (.44)	1.72 (.22)	3.05 (.43)
$m = 10{,}000$	2.2 (.31)	.50 (.07)	.25 (.03)	.49 (.07)
$m = 100{,}000$	.2 (.02)	.08 (.01)	.05 (.01)	.08 (.01)
$t(5)$				
$m = 1{,}000$	70.5 (5.21)	–	4.46 (.42)	4.23 (.39)
$m = 10{,}000$	48.9 (1.38)	–	.74 (.08)	.73 (.08)
$m = 100{,}000$	51.0 (.60)	–	.15 (.02)	.15 (.02)
$t(3)$				
$m = 1{,}000$	171. (8.13)	–	9.97 (.63)	9.88 (.63)
$m = 10{,}000$	176. (3.78)	–	2.13 (.14)	2.11 (.14)
$m = 100{,}000$	176. (1.60)	–	.37 (.02)	.37 (.02)
Gamma(2,1)				
$m = 1{,}000$	89.9 (6.11)	5.70 (.50)	1.66 (.21)	–
$m = 10{,}000$	86.3 (1.65)	.75 (.07)	.31 (.04)	–
$m = 100{,}000$	85.5 (.64)	.16 (.02)	.05 (.01)	–

All reported numbers here have been multiplied by 10^3.

3.2. *One dimensional case with boundary mode*

In this section we first consider gamma(1,1) as the nominal posterior distribution. It is noted that the shape of the observations is skewed and the mode locates at the boundary. The best point to evaluate the Candidate's estimate is at the mode, which is also the boundary point. In this comparison, the volume-corrected Laplace and the Candidate's estimate are evaluated at one kernel bandwidth from the boundary. Table 2 lists the results for the volume-corrected Laplace estimate and the Candidate's estimate for $m = 1{,}000$, 10,000, and 100,000, respectively. Again, the Candidate's estimates attain better accuracy and have much smaller standard errors. This procedure can be utilized when the target distribution is not normal or fairly skewed such as those seen in the random effects, or variance component models with constrained parameters.

Another illustration for the boundary case is a hierarchical model with y given λ from a Poisson(λ) distribution, where the parameter λ is from an exponential distribution with the hyper-

Table 2. *These are the averages and standard errors of mean square relative errors from 100 replications for Laplace volume correction and Candidate's estimates*

Gamma(1,1)	vol. cor. Laplace (h)	Candidate (best point, h)	Candidate (mean)
$m = 1{,}000$	11.6 (1.52)	1.3 (.16)	3.6 (.51)
$m = 10{,}000$	1.1 (.16)	.4 (.05)	.5 (.06)
$m = 100{,}000$	.1 (.02)	.4 (.02)	.1 (.01)

All reported numbers here have been multiplied by 10^3.

Table 3. *Estimation of the normalizing constant in a Poisson hierarchical model*

$m = 1000$	True value	vol. cor. Laplace	harmonic mean	Candidate
C or $\hat{C}$	.19	.25	.22	.18
MSRE	–	.048	.012	.003

parameter β from gamma(a, b). Suppose we are interested in the posterior mean of λ after observing $y = 1$, we then need to make inference based on the posterior distribution of λ. In other words, we need to compute the integration of $f(y \mid \lambda)\pi(\lambda)$ over λ. Numerical integration can be used to derive the true normalizing constant $m(y)$. Here, for the purpose of illustration, we first compute the Laplace estimate and then generate Gibbs samples with $m = 1{,}000$ from the full set of conditional distributions to derive the volume-corrected Laplace, Candidate's, and the harmonic mean estimate (Newton and Raftery 1994). The true value of $m(y)$ is .19 when a and b are both assumed 1. The volume-corrected Laplace estimate is $C_{\text{vol-cor}} = .25$, which is evaluated at the maximum likelihood estimate of $\lambda = \sqrt{2} - 1$. As seen in Table 3, the harmonic mean estimate is .22, and the Candidate's estimate is .18, evaluated at the estimated posterior mean. The mean square relative error for our proposal is only .003. The MSREs of the volume-corrected Laplace and harmonic mean are 16 and 4 times larger than that of the Candidate's estimate. Again, the Candidate's estimate is much better when dealing with distributions of irregular shape.

3.3. *Multi-dimensional case*

In this section we apply the nonparametric Candidate's estimate to multi-dimensional problems. The Candidate's formula $C = f(y \mid \theta)\pi(\theta)/\pi(\theta \mid y)$ is valid for all θ in the prior support. Therefore, we evaluate the estimate at various θ-values and average over them. In this comparison study, we choose the multivariate normal and the 'product of gammas' as the nominal posterior distributions, where the 'product of gammas' is referring to the product of coordinate-wise gamma(2,1) densities. There are 100 replications under each setting. Here is a reminder that the bandwidth should be adjusted to the number of evaluation points according to item 3 in the guidelines. Table 4 lists the results for the volume-corrected Laplace estimate and the Candidate's estimate for $m = 1{,}000$ and 10,000.

For the case of 4-dimensional normal and product-gammas, the Candidate's estimate is evaluated at points $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ with each θ_i set to be either the mode, mode minus one standard deviation or mode plus one standard deviation. There are $M = 3^4$ many of such θ points. We evaluate the Candidate's estimates at these 81 points and take their average. This average is the reported Candidate's estimate.

For the case of 10-dimensional normal and product-gammas, the Candidate's estimate is evaluated at points $\theta = (\theta_1, \ldots, \theta_{10})$ with each θ_i set to be either the mode or mode plus one standard

Table 4. *Multi-dimensional case with 100 replications*

	vol. cor. Laplace	Candidate
Dimension = 4		
Normal		
$m = 1{,}000$	2.43 (.35)	.89 (.14)
$m = 10{,}000$	.17 (.03)	.24 (.03)
Product-gammas		
$m = 1{,}000$	10.5 (.63)	.83 (.15)
$m = 10{,}000$	10.7 (.24)	.43 (.05)
Dimension = 10		
Normal		
$m = 1{,}000$	1.83 (.23)	9.41 (.69)
$m = 10{,}000$	.21 (.03)	4.78 (.36)
Product-gammas		
$m = 1{,}000$	39.0 (1.42)	12.5 (.32)
$m = 10{,}000$	39.4 (.43)	6.12 (.17)

All reported numbers here have been multiplied by 10^2.

deviation. There are $M = 2^{10}$ many of such θ points. Again, we evaluate the Candidate's estimates at these 1024 points and take their average. The results are reported in Table 4. Certainly one may increase, decrease, or customize the evaluation points. For instance, instead of all 10 coordinates, one may pick d many coordinates and evaluate at these $M = 2^d$ points only. Alternatively, instead of evaluation at all points of mode plus one standard deviation, one may pick points of mode plus one standard deviation in some coordinates and mode minus one standard deviation in other coordinates.

When the posterior is normally distributed, the volume-corrected Laplace estimate is better than the Candidate's. When the posterior is product-gammas, Candidate's estimate outperforms the volume-corrected Laplace estimate.

4. Concluding remarks

This article attempts to draw the attention of readers to this simple and straightforward method for calculating marginal probabilities based on Markov chain output. This procedure can be applied for routine use by output from Metropolis algorithms or Gibbs sampler. Either the usual kernels for interior points or boundary kernels can be considered in the proposed method. Under the boundary case, an easy and direct application is to estimate the marginal probability at the point which is one bandwidth (h) away from the boundary. Unlike the methods requiring the knowledge of all conditional densities or evaluating the posterior density at all posterior sample points, the nonparametric Candidate's estimate does not require specific knowledge of the full set of conditional densities and it requires only the evaluation at one single point (or a couple more for the high dimensional problem). We have also shown that there exists a best point for evaluation and laid out its derivation. In cases of multi-dimensional problems, we suggest a procedure of averaging over some high density points to overcome the data spareness. The simulation study indicates that it performs reasonably well in overall cases. The Candidate's estimate is comprehensible, reliable and easy to compute.

Other simulation studies based on higher dimensionality, say 100 or even more, for comparison may be interesting. However, based on our experience, unless further structure of the joint posterior distribution can be available, the computational load, including the computation of the Hessian matrix, would be so large that the current computing facility fails to handle the problem. We are currently investigating solutions to ease the load when the complexity of the posterior distribution can be relaxed under various conditions.

Some issues concerning standardization are worth mentioning here. If the target distribution is not close to being coordinate-wise uncorrelated and symmetric, a common approach is to transform the parameter variables. This can be done via reparameterization before generating the Markov chains in order to achieve greater efficiency. Transformation will result in not only faster convergence but also an easier choice of the bandwidth matrix. As indicated in the implementation, though we have naively chosen the normal as the reference posterior to get a rule of thumb bandwidth h, the scale family suggested by Terrell (1991) may be an alternative. Nevertheless, even if the transformation is not carried out and the posterior distribution is highly skewed, the nonparametric Candidate's estimate can still work reasonably well.

The proposed averaged estimate over several points may be useful for high dimensional problems to reduce the curse of dimensionality. We further suggested to select some, say d, coordinates and average over points in those dimensions. This approach would be particularly useful when the dimensionality is so large that the Hessian matrix cannot be computed. The criteria for choosing the location and number of points for averaging process will be important and worth further investigation.

Although we have used the nonparametric Candidate's estimate to derive the marginal probability, we have no intention to abolish other types of estimates. Each has its merits and position to suit certain applications. For instance, the Laplace type methods are very accurate for well behaved $\pi(\theta \mid y)$ satisfying certain regularity conditions, the harmonic mean estimate is good when the samples are within reasonable range of the likelihood, the bridge sampling and the path sampling approaches are still quite accurate when other methods fail for extremely high dimensional problems. Nevertheless, the nonparametric Candidate's estimate is easy to apply with reasonable computational load.

Appendix of proofs

Proof for Theorem 1: By the Taylor expansion and properties of the kernel in C2, one can derive the bias of the posterior density estimate as

$$E_{\theta^{(1)},\ldots,\theta^{(m)}\mid y}\hat{\pi}(\theta \mid y) - \pi(\theta \mid y) = \frac{k_2}{2} trace\left\{H\pi^{(2)}(\theta \mid y)\right\} + o(trace(H)),$$

where $\pi^{(2)}(\theta \mid y)$ is a $p \times p$ matrix with the (i, j)th entry given by $\partial^2\pi(\theta \mid y)/(\partial\theta_i \partial\theta_j)$. The variance of the posterior density estimate is given by

$$var_{\theta^{(1)},\ldots,\theta^{(m)}|y}(\hat{\pi}(\theta \mid y)) = \frac{v\pi(\theta \mid y)}{m|H|^{1/2}} + o(m^{-1}|H|^{-1/2}),$$

where v is given in condition C2. Thus, we have

$$E_{\theta^{(1)},\ldots,\theta^{(m)}|y}\left(\frac{C}{\hat{C}} - 1\right)^2 = O(\{trace(H)\}^2) + O(m^{-1} \mid H \mid^{-1/2}). \tag{7}$$

By choosing the bandwidth matrix H satisfying $\lambda_j(H) = O(m^{-2/(4+p)})$, then we have

$$E_{\theta^{(1)},\ldots,\theta^{(m)}|y}(C/\hat{C} - 1)^2 = O(m^{-4/(4+p)}).$$

□

The following lemma is necessary for the proof of Theorem 2.

Lemma 1. *Let $\mathcal{H}$ be the set of all $p \times p$ positive definite matrices. For a symmetric nonnegative definite matrix A and constants c_1 and c_2, the solution for the minimization problem*

$$\arg\min_{H\in\mathcal{H}} c_1(trace\{HA\})^2 + c_2|H|^{-1/2}$$

is given by $H_{opt} = G' D_{opt} G$, where G is an orthogonal matrix which diagonalizes A, i.e., $A = G\Lambda G'$ with $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$, and

$$D_{opt} = diag\left(\frac{\alpha_{opt}}{\lambda_1}, \ldots, \frac{\alpha_{opt}}{\lambda_p}\right)$$

with α_{opt} given later in (11)*. Moreover, the minimal value is given by*

$$\inf_H c_1(trace\{HA\})^2 + c_2|H|^{-1/2} = (c_1 p)^{p/(p+4)}(c_2/4)^{4/(p+4)}(p+4)|A|^{2/(p+4)}.$$

(*If any of the $\lambda_k, k = 1, \ldots, p$, is zero, then pass the limit $\lambda_k \to 0$ into D_{opt} and get $\lim_{\lambda_k\to 0} \alpha_{opt}/\lambda_k = \infty$. The lemma is still valid.*)

Proof: For a symmetric nonnegative definite matrix A, there exists an orthogonal matrix G such that $A = G\Lambda G'$, where Λ is a diagonal matrix with nonnegative diagonal entries. Let $D = G'HG$, then

$$\begin{aligned} &c_1(trace\{HA\})^2 + c_2 \mid H \mid^{-1/2} \\ &\quad = c_1(trace\{G'HG\Lambda\})^2 + c_2 \mid G'HG \mid^{-1/2} \\ &\quad = c_1(trace\{D\Lambda\})^2 + c_2 \mid D \mid^{-1/2}. \end{aligned}$$

We shall now solve for the following minimization problem:

$$\arg\min_{D\in\mathcal{H}} c_1(trace\{D\Lambda\})^2 + c_2 \mid D \mid^{-1/2}. \tag{8}$$

Let $\mathcal{D}$ be the set of all $p \times p$ diagonal matrices in $\mathcal{H}$. Note that

$$\begin{aligned} &\min_{D\in\mathcal{H}} c_1(trace\{D\Lambda\})^2 + c_2 \mid D \mid^{-1/2} \\ &\quad = \min_{D\in\mathcal{D}} c_1(trace\{D\Lambda\})^2 + c_2 \mid D \mid^{-1/2}. \end{aligned}$$

That is, the minimization problem (8) can be restricted to $\mathcal{D}$. We have

$$c_1(trace\{D\Lambda\})^2 + c_2|D|^{-1/2} = c_1\left(\sum_{i=1}^p d_i\lambda_i\right)^2 + c_2\left(\prod_{i=1}^p d_i\right)^{-1/2},$$

and

$$\frac{\partial}{\partial d_k}\{c_1(trace\{D\Lambda\})^2 + c_2 \mid D \mid^{-1/2}\} = 2\lambda_k c_1\left(\sum_{i=1}^p d_i\lambda_i\right) - \frac{c_2}{2d_k}\left(\prod_{i=1}^p d_i\right)^{-1/2}. \tag{9}$$

By setting expression (9) to zero, we get

$$\lambda_k d_k = \frac{c_2}{4c_1}\left(\prod_{i=1}^p d_i\right)^{-1/2}\left(\sum_{i=1}^p d_i\lambda_i\right)^{-1}. \tag{10}$$

Note that the right hand side of (10) is a constant. Define $\alpha = \lambda_k d_k$ and plug it into the minimization problem (8). The minimizing α is

$$\alpha_{opt} = \left(\frac{c_2\sqrt{\lambda_1\lambda_2\cdots\lambda_p}}{4c_1 p}\right)^{2/(p+4)} = \left(\frac{c_2\sqrt{|A|}}{4c_1 p}\right)^{2/(p+4)}. \tag{11}$$

Therefore, the solution for (8) is $D_{opt} = diag(\alpha_{opt}/\lambda_1, \ldots, \alpha_{opt}/\lambda_p)$ and the corresponding optimal matrix is $H_{opt} = G' D_{opt} G$. It is then easy to check that the minimal value is $c_1(trace\{H_{opt}A\})^2 + c_2|H_{opt}|^{-1/2} = (c_1 p)^{p/(p+4)}(c_2/4)^{4/(p+4)}(p+4)|A|^{2/(p+4)}$. □

Proof for Theorem 2: Begin with

$$\begin{aligned} E_{\theta^{(1)},\ldots,\theta^{(m)}|y}\left(\frac{C}{\hat{C}} - 1\right)^2 &= E_{\theta^{(1)},\ldots,\theta^{(m)}|y}\left(\frac{\hat{\pi}(\theta \mid y)}{\pi(\theta \mid y)} - 1\right)^2 \\ &= \frac{E_{\theta^{(1)},\ldots,\theta^{(m)}|y}(\hat{\pi}(\theta \mid y) - \pi(\theta \mid y)))^2}{\pi^2(\theta \mid y)} \\ &= \frac{k_2^2}{4\pi^2(\theta \mid y)}\left(trace\left\{H\pi^{(2)}(\theta \mid y)\right\}\right)^2 + \frac{v\pi(\theta \mid y)}{m|H|^{1/2}\pi^2(\theta \mid y)} \\ &\quad + o(trace(H)) + o(m^{-1} \mid H \mid^{-1/2}). \end{aligned} \tag{12}$$

We shall solve the following minimization problem, which gives the optimal H minimizing the asymptotic mean square relative error:

$$\arg\min_{H\in\mathcal{H}} \frac{k_2^2}{4}\left(trace\left\{H\pi^{(2)}(\theta \mid y)\right\}\right)^2 + \frac{v\,\pi(\theta \mid y)}{m \mid H \mid^{1/2}}.$$

By Lemma 1,

$$\inf_{H\in\mathcal{H}}\left\{\frac{k_2^2\left(trace\left\{H\pi^{(2)}(\theta\mid y)\right\}\right)^2}{4\pi^2(\theta\mid y)}+\frac{v}{m\pi(\theta\mid y)\mid H\mid^{1/2}}\right\}$$
$$=c_3(\pi(\theta\mid y))^{-(2p+4)/(p+4)}\left\{\left|\pi^{(2)}(\theta\mid y)\right|^2\right\}^{1/(p+4)},$$

where the constant c_3 is given by

$$c_3=0.25\left(pk_2^2\right)^{p/(p+4)}\left(\frac{v}{m}\right)^{4/(p+4)}(p+4).$$

Therefore, the estimator (4) has asymptotically minimum value at

$$\arg\min_{\theta}\left|\pi^{(2)}(\theta\mid y)\right|^2(\pi(\theta\mid y))^{-(2p+4)},$$

or equivalently, at argument minimizer for abs$\{|\pi^{(2)}(\theta\mid y)|\}$ $(\pi(\theta|y))^{-(p+2)}$. □

References

Besag J.E. 1989. A candidate's formula: A curious result in Bayesian prediction. Biometrika 76: 183.

Chen M.H. and Shao Q.M. 1997. On Monte Carlo methods for estimating ratios of normalizing constants. The Annals of Statistics 25: 1563–1594.

Chib S. 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90: 1313–1321.

Cowling A. and Hall P. 1996. On pseudodata methods for removing boundary effects in kernel density estimation. Journal of the Royal Statistical Society, Ser., B 58: 551–563.

DiCiccio T.J., Kass R.E., Raftery A., and Wasserman L. 1997. Computing Bayes factors by combining simulation and asymptotic approximations. Journal of the American Statistical Association 92: 903–915.

Erkanli A. 1994. Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. Journal of the American Statistical Association 89: 250–258.

Gasser T. and Müller H.G. 1979. Kernel estimation of regression functions. In: Gasser and Rosenblatt (Ed.), Smoothing Techniques for Curve Estimation. Lecture Notes in Mathematics, 757, Springer-Verlag.

Gelfand A.E. and Dey D.K. 1994. Bayesian model choice: Asymptotics and exact calculations. Journal of the Royal Statistical Society, Ser. B 56: 501–514.

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.

Gelman A. and Meng X.L. 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Statistical Science 13: 163–185.

Geman S. and Geman D. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions of Pattern Analysis and Machine Intelligence 6: 721–741.

Geweke J. 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57: 1317–1340.

Geyer C.J. 1992. Practical Markov chain Monte Carlo" (with comments). Statistical Science 7: 473–451.

Gilks W.R., Richardson S., and Spiegelhalter D.J. 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall, London, UK.

Hastings W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57: 97–109.

Hsiao C.K. 1997. Approximate Bayes factors when a mode occurs on the boundary. Journal of the American Statistical Association 92: 656–663.

Huang S.Y., Hsiao C.K., and Chang C.W. 2003. Optimal volume-corrected Laplace-Metropolis method. Annals of the Institute of Statistical Mathematics 55: 655–670.

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.

Meng X.L. and Wong W.H. 1996. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. Statistica Sinica 6: 831–860.

Mosteller F. and Wallace D.L. 1964. Applied Bayesian and Classical Inference, 1st ed. Reprinted in 1984 by Springer-Verlag, New York.

Müller H.G. 1991. Smooth optimum kernel estimators near endpoints. Biometrika 78: 521–530.

Müller H.G. 1993. On the boundary kernel method for nonparametric curve estimation near endpoints. Scandinavian Journal of Statistics 20: 313–328.

Müller H.G. and Stadtmüller U. 1999. Multivariate boundary kernels and a continuous least squares principle. Journal of the Royal Statistical Society, Ser. B 61: 439–458.

Newton M.A. and Raftery A.E. 1994. Approximate Bayesian inference by the weighted likelihood bootstrap" (with discussion). Journal of the Royal Statistical Society, Ser. B 56: 3–48.

Parzen E. 1962. On the estimation of probability density function and mode. The Annals of Mathematical Statistics 33: 1065–1076.

Pauler D.K., Wakefield J.C., and Kass R.E. 1999. Bayes factors and approximations for variance component models. Journal of the American Statistical Association 94: 1242–1253.

Rice J. 1984. Boundary modification for kernel regression. Communications in Statistics, Part A 13: 893–900.

Scott D.W. 1992. Multivariate Density Estimation. John Wiley & Sons, Inc, New York, NY.

Simonoff J.S. 1996. Smoothing Methods in Statistics. Springer, New York, NY.

Silverman B.W. 1986. Density Estimation. Chapman & Hall, London.

Terrell G.R. 1990. The maximal smoothing principle in density estimation. Journal of the American Statistical Association 85: 470–477.

Tierney L. and Kadane J.B. 1986. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81: 82–86.

West M. 1993. Approximating posterior distributions by mixtures. Journal of the Royal Statistical Society, Ser. B 55: 409–422.