行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※

※　　　　　　　　　　　　　　　　　　　　　　　※

※　利用事後機率來檢定家族相關聚集性是否存在　※

※　　　　　　　　　　　　　　　　　　　　　　　※

※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：☑個別型計畫　　□整合型計畫

計畫編號：NSC －89－2118－M－002－014

執行期間：88 年 8 月 1 日至 89 年 7 月 31 日 (1/2)
89 8 1 89 7 31 (2/2)

計畫主持人：蕭朱杏
共同主持人：
戴政

本成果報告包括以下應繳交之附件：

- □赴國外出差或研習心得報告一份
- □赴大陸地區出差或研習心得報告一份
- □出席國際學術會議心得報告及發表之論文各一份
- □國際合作研究計畫國外研究報告書一份

執行單位：台灣大學公衛學院流行病學研究所

中 華 民 國 90 年 10 月 5 日

Bayesian Marginal Inference via Candidate's Formula

Chuhsing Kate Hsiao*
Division of Biostatistics,
Institute of Epidemiology,
National Taiwan University

Su-Yun Huang
Institute of Statistical Science
Academia Sinica, Taipei

Ching-Wei Chang
Division of Biostatistics,
Institute of Epidemiology,
National Taiwan University

Running head: Candidate's estimate

July 3, 2001

*Correspondence: Dr. Chuhsing Kate Hsiao, Division of Biostatistics, Institute of Epidemiology, National Taiwan University, No.1, Jen-Ai Rd., Sec.1, Rm. 1542, Taipei 100, Taiwan, R.O.C. Fax: 011-886-2-2341-8562, email:ckhsiao@ha.mc.ntu.edu.tw

Abstract

In the context of Bayesian inference, a nonparametric kernel estimate via Candidate's formula is developed for computing the marginal density of the sample data. The estimate is computed based on Markov chains outputs. The deficiency of high dimensional density estimation, known as the curse of dimensionality, can be handled well in this Bayesian marginal inference problem. An approach is introduced to ease the tension caused by data sparseness. This nonparametric Candidate's estimate does not require knowledge of full conditional densities. We find it a convenient and comprehensible way to estimate the posterior density. The asymptotic behavior of the estimate is studied and the best point for evaluating the estimate is derived. It is found that the posterior mode may not be the best point. A simulation study shows that the nonparametric Candidate's estimate has better accuracy than the Laplace type estimates. Laplace method is based on analytic asymptotic approximation, which often requires a large size of observed data; while the Candidate's method asks for simply a large size of simulated Markov chains as the posterior samples. It is much easier to get more simulated posterior samples at a low cost rather than to obtain more observations.

Key words and phrases: Bayes factor, Gibbs sampler, kernel density estimation, Laplace approximation, Laplace-Metropolis method, Laplace volume correction, Markov chain Monte Carlo, Metropolis-Hasting algorithm.

1 Introduction

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

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

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

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

where $\eta = g(\theta)$ for some function g and where $\eta \in \Omega \subset R^d$ with $1 \leq d < p$.

Computing the marginal probability has long been an important issue in Bayesian inference. The computation is necessary when the model selection is of interest, or when the posterior distributions, moments, Bayes factors, or predictive densities are requested. The quantity $m(y)$, or $m(y|M_k)$, is sometimes referred to as a normalizing constant, especially when the integrand is taken to be the product of likelihood function and prior density. Much efforts have since been placed uopn 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. The approximation requires the evaluation of the mode and variance, which may not be straightforward in many complex models or applications. In addition, even the integrand in the above equation may involve further integration of other nuisance parameters and thus have no closed form. Furthermore, when parameters are close to the boundaries, such as that in a variance component model, the usual Laplace's method may fail. Recently, 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 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 MLE's.

The advent of the Markov chain Monte Carlo method has provided an easy means to obtain samples from the target distribution such as the posterior density (see, e.g., Gelfand and Smith 1990; Gilks, Richardson, and Spiegelhalter 1996). The MCMC method has helped the development of mainly two types of estimates for the normalizing constant. The first one combines Laplace approximation and posterior simulations. Based on the simulated posterior samples, mode and variance of the approximate normal probability density to $\pi(\theta|y)$ can be estimated. Lewis and Raftery (1997) estimated the mode and variance based on the simulation and utilized those with Laplace's method to approximate the Bayes factor. DiCiccio, Kass, Raftery, and Wasserman (1997) followed the same line and improved Laplace's method using a volume of high density points for correction to approximate $\pi(\theta|y)$. Huang, Hsiao and Chang (2000) further derived the best volume for correction.

The second type of approach essentially uses only Monte Carlo posterior simulation but without analytic model approximation. This kind of estimates utilize a large amount of simulated samples from the target distribution (say, posterior density) for estimation. For instance, Newton and Raftery (1994) applied the simulated samples in the harmonic mean of likelihood to estimate the marginal. However, this estimate is not stable due to the fact that the inverse likelihood does not have finite variance. Gelfand and Dey (1994) modified the harmonic mean estimate by introducing a thin-tail tuning function to stablize the estimate. The tuning function is required to have tails thinner than the posterior. However, it is quite difficult to determine a tuning function, especially in high-dimensional problem. Gelfand and Smith (1990) and Chib (1995) used Gibbs outputs to estimate the marginal density based on full conditional distributions. Their methods require knowledge of each conditional density.

Besag (1989) gave the Candidate's formula for Bayesian prediction

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

This equation holds for all θ values in the support of prior $\pi(\cdot)$. It is also known as the basic marginal likelihood identity (Chib 1995). Chib discussed this estimate based on the assumption that the full set of conditional densities are known. When the knowledge of full conditional densities is intractable, the posterior density $\pi(\theta|y)$ can be estimated by nonparametric method. We refer to it the nonparametric Candidate's estimate. Nonparametric density estimation has been a well developed research topic in recent decades. The techniques, such as histograms and kernel estimates, have been used in our daily practice of statistical analysis. Several books (see, for instance, Silverman 1986; Scott 1992; Simonoff 1996) have given a very good account of the theoretical background and usage for practical applications. In this article we first adopt the kernel method which deals with the regular posterior distributions with open support and interior mode. This kernel method can also be adapted to the boundary case by using a one-sided kernel (or any other types of boundary kernels). For example, a right-sided kernel may be used in the variance component models to avoid putting any weight exceeding the left boundary. If an i.i.d. random sample from the posterior distribution $\pi(\theta|y)$ is available, it is not difficult to compute its kernel density estimate at a certain point. This nonparametric Candidate's method is simple to implement. It requires the evaluation of the likelihood function, the prior density and the estimate of posterior. We intend to use the posterior samples generated via MCMC method to obtain the nonparametric estimate. Although the MCMC samples are usually not independent, there have been several approaches to reducing the correlation 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.

In Section 2 we demonstrate that the nonparametric Candidate's estimate in equation (3) can be easy, costless, and more accurate than the Laplace type estimates. Its asymptotic behavior is evaluated and the best point for estimation is derived. In Section 3 the method is compared with the Laplace type estimates in a simulation study using two groups of distributions. For the first group, observations are generated from various shapes of unimodal distributions. For the second group, we focus on distributions that are highly skewed and/or have mode close to the boundary of the support. The Laplace type estimates usually fail in this situation unless special treatments are applied; while the Candidate's estimate provides better solutions. We also discuss a modification for the case where the mode is close to the boundary. A real application about the germination of beans is analyzed in Section 4. A concluding discussion is given in Section 5. All proofs are in the Appendix.

2 Nonparametric Candidate's method

2.1. The approach. Suppress the model index from $m(y|M_k)$ and concentrate on the estimation of the marginal density, or the normalizing constant, $m(y)$. For simplicity, we use notation C to denote the normalizing constant

$$C \equiv \frac{f(y|\theta)\pi(\theta)}{\pi(\theta|y)},$$

where any value of θ in the support of prior $\pi(\cdot)$ provides the same answer. This normalizing constant can be estimated by

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

where $\hat{\pi}(\theta|y)$ is a kernel density estimate of $\pi(\theta|y)$ based on the posterior samples $\theta^{(1)},\ldots,\theta^{(m)}$ and is given by

$$\hat{\pi}(\theta|y) = \frac{1}{m|H|^{1/2}} \sum_{i=1}^{m} \mathcal{K}\left((\theta-\theta^{(i)})^T H^{-1}(\theta-\theta^{(i)})\right). \tag{5}$$

Here H is a $p \times p$ symmetric positive definite matrix, $|H|$ is the determinant of H, and $\mathcal{K}(\cdot)$ is a p-dimensional kernel function satisfying conditions

C1. $\mathcal{K}$ is non-negative and integrates to one: $\int_{R^p} \mathcal{K}(\theta^T\theta)d\theta = 1$.

C2. $\mathcal{K}$ is an order 2 kernel in the sense that

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

2.2. Theoretical results. We state in the following theorem the order of accuracy of this nonparametric Candidate's estimate.

Theorem 1 *Assume that the likelihood function and the prior have continuous second derivative 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 $trace(H) \to 0$ and $m|H|^{1/2} \to \infty$, as the posterior sample size $m \to \infty$. Then, with*

$$\hat{C} = \frac{f(y|\theta)\pi(\theta)}{\hat{\pi}(\theta|y)},$$

the following mean square error is of order

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

where the expectation $E_{\theta^{(1)},\ldots,\theta^{(m)}|y}$ is taken with respect to the conditional joint posterior distribution of $\theta^{(1)},\ldots,\theta^{(m)}$ given y. Furthermore, choose H so that $\lambda_j(H) = O(m^{-2/(4+p)})$ for $j = 1,\ldots,p$, where λ_j denotes the jth largest eigenvalue. Then we have

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

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 arisen naturally is at what value of θ the posterior should be estimated. The Candidate's formula is valid for all θ in the support of prior, and thus any choice of θ, such that the posterior function in the neighborhood of θ is smooth, is theoretically valid. However, consideration for efficiency may suggest the estimator (4) be evaluated at high density points. Under the criterion of minimum mean square error, we derive the best point in the following Theorem 2.

Theorem 2 *The estimator (4) has asymptotically minimum value for*

$$E_{\theta^{(1)},\ldots,\theta^{(m)}|y}\left(\frac{C}{\hat{C}}-1\right)^2$$

at

$$\theta^* = \arg\min_{\theta}\left|\frac{det(\nabla^2\pi(\theta|y))}{[\pi(\theta|y)]^{p+2}}\right|. \tag{6}$$

For univariate θ, i.e. $p=1$, (6) becomes

$$\theta^* = \arg\min_{\theta}\frac{\left|\frac{\partial^2}{\partial\theta^2}\pi(\theta|y)\right|}{[\pi(\theta|y)]^3}.$$

As the posterior is proportional to $f(y|\theta)\pi(\theta)$, the above minimization problem gives the same θ^* by using $f(y|\theta)\pi(\theta)$.

To attain the minimum mean square error, it is best to evaluate the Candidate's estimate at the value of θ which minimizes the right hand side of equation (6). The posterior mode may not be the most accurate point. For instance, for normal posteriors the Candidate's estimate is most accurate if evaluated at points of one standard deviation from the mode. See Table 1 for accuracies at mode, mean, and best point.

2.3. Some guidelines for computation. To make the nonparametric Candidate's estimate easy to implement, listed below are some guidelines.

1. The marginal p.d.f. $m(y)$ is invariant under changes of prior variables. Therefore, we may standardize the posterior sample. The matrix H in (5) can then be taken as hI_p, where $h > 0$ is a scalar serving as the kernel bandwidth and I_p is the $p \times p$ identity matrix. The standardization procedure makes posterior density estimate a lot easier.

2. For a high dimensional problem, the kernel method suffers from the curse of dimensionality and is not efficient. The phenomenon of the curse of dimensionality is due to data sparseness in a high dimensional space. It is not a problem occurring only in kernel estimates, but it exists in all kinds of nonparametric methods. Unless, a certain model structure is imposed, the phenomenon of the curse of dimensionality persists. Fortunately, the deficiency problem encountered in this particular Bayesian marginal inference problem can be easily handled on two accounts. (a) The posterior sample is cheap to obtain. (b) 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 $O((mh^p)^{-1})$. We see that the deficiency is caused by the high variation. Since the Candidate's formula for marginal density is valid for all θ values in the prior support, one may average over a set of $\hat{m}(y) = \pi(y|\theta)\pi(\theta)/\hat{\pi}(\theta|y)$ for various θ values. The averaging procedure effectively utilizes a lot more posterior sample points, which lessens the tension of data sparseness and tends to stablize the high variation.

3. (Optional.) Use a uniform kernel for posterior density estimate. This suggestion is to reduce the computation load, especially in a high dimensional problem. 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 centered at θ with radius h.

3 A simulation study

In this section, we compare the nonparametric Candidate's estimate with other estimates, basically the volume correction Laplace method. Recall the Laplace approximation

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

where $\phi(\cdot;\theta^*,\Sigma^*)$ is the multivariate normal density function with mean θ^* and covariance matrix Σ^*. This approximation has $C = C_{Lap}\{1 + O(n^{-1})\}$, where n is the size of the observed data y. The asymptotic approximation applies when n is large, and it does not utilize any simulation. Lewis and Raftery (1997) derived the estimated mode and variance for ϕ based on simulated Markov chains and combined them with the Laplace's method. The accuracy of the Laplace approximation C_{Lap} depends heavily on the shape of the posterior $\pi(\theta|y)$, in other words, it depends on the degree of resemblance between $\pi(\theta|y)$ 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. It was called the volume correction Laplace method, denoted here $C_{vol-cor}$,

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

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

3.1 Regular case

Table 1 lists the results from simulations based on four different distributions as nominal posterior distributions. 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 in total. We use the mean square error $(C/\hat{C} - 1)^2$ as the measure of accuracy. Both the volume correction Laplace method and Candidate's estimate are considered in this comparison. The Candidate's estimates are evaluated at three different points: mode, mean, and the best point. For the normal distribution the best points are at mode plus or minus one standard deviation, and we take in this simulation study the point at mode plus one standard deviation. The best point for the Student-t is at mode. The best point for gamma(2,1) is at mean. Table 1 lists that the average and standard error of the mean square errors from 100 replications. It can be seen that both the average of MSEs of Candidate's estimates and their variability (standard error) are always smaller than those of the volume correction Laplace estimates. Among the three candidate's estimates evaluated at different points, the one at the best point does outperform the rest.

———— Place Table 1 here ————

3.2 Boundary case

In this section we consider gamma(1,1) as a 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 the boundary point. In this comparison, the volume correction Laplace and the Candidate's estimate are evaluated at one kernel bandwidth from the boundary. Table 2 lists the results for volume correction Laplace and Candidate's estimate for m =1,000, 10,000, and 100,000, respectively. Again, Candidate's estimates have better accuracy and attain much less standard error. 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.

———— Place Table 2 here ————-

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 hyper-parameter β from gamma(a, b). A practical application of this model, for example, is considering y as the number of eggs laid by certain species when conditioning on λ. The exponential prior for λ and the gamma for hyperparameter indicate that it is an endangered species. Suppose we are interested in the expected number of eggs λ after observing $y = 1$. In other words, we need to make inference based on the posterior distribution of λ given y, i.e., the integration of $f(y|\lambda)\pi(\lambda)$ over λ is required. Numerical integration can be used to derive the true normalizing constant $m(y)$. However, 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 correction Laplace, Candidate's, and the harmonic mean estimate (Newton and Raftery, 1994). The true value of $m(y)$ is .1917 when a and b are both assumed 1. The volume correction Laplace estimate is $C_{volm} = .2453$, which is evaluated at the maximum likelihood estimate of $\lambda = \sqrt{2} - 1$. The harmonic mean estimate is .2151. On the other hand, the Candidate's estimate is .1816, when evaluated at the estimated posterior mean. Again, the proposed estimate is better when dealing with the irregular shape of distributions.

———— Place Table 3 here ————-

3.3 High dimensional case

In this section we apply the nonparametric Candidate's estimate to high dimensional problem. The Candidate's formula $C = f(y|\theta)\pi(\theta)/\pi(\theta|y)$ is valid for all θ in the prior support. Therefore, we evaluate the Candidate's estimate at various θ-values and average over them. In this comparison study, we take the multivariate normal and the 'product of gammas' as the nominal posterior distributions, where the 'product of gammas' is set to be the product of coordinate-wise gamma(2,1) densities. We take 100 replications. In each replication, the posterior sample is standardized first, and then the 'final' Candidate's estimate (explained below) and its mean square error are computed. The average of these 100 mean square errors and its standard error are listed in Table 4.

For the case of 4-dimensional normal, 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 mode, mode minus one standard deviation or mode plus one standard deviation. There are 3^4 many of such θ-values. Evaluate the Candidate's estimates at these 3^4 points and take average. This average is the final Candidate's estimate. The whole procedure is repeated 100 times. For the case of 10-dimensional normal, the Candidate's estimate is evaluated at points $\theta = (\theta_1, \ldots, \theta_{10})$ with each θ_i set to be either mode or mode plus one standard deviation. There are 2^{10} many of such θ-values. Again, evaluate the Candidate's estimates at these 2^{10} points and take average. Of course one may evaluate Candidate's estimates at all the 3^{10} points of mode and mode plus/minus one standard deviation. In this simulation, we only use 2^{10} many θ-values to reduce the computational load to about 1/60 compared to using all the 3^{10}

θ-values. For the case of product-gamma, the Candidate's estimate is evaluated at points $\theta = (\theta_1, \ldots, \theta_d)$ with each θ_i set to be either mode or mode plus one standard deviation. There are 2^d many of such θ-values.

The above high-dimensional Candidate's estimate effectively uses more posterior data points to stablize the high variation due to data sparseness in a high dimensional space. When the posterior is normally distributed, volume correction Laplace is slightly better than Candidate's. When data are from product Gamma, Candidate's estimate outperforms.

——————— Place Table 4 here ————————

4 Discussion

We have discussed the general use of the nonparametric Candidate's estimate for calculating the marginal probability based on Markov chain outputs. This procedure can be applied widely to outputs from Metropolis algorithms and Gibbs sampler. Either the usual kernel for interior points or the boundary kernel can be considered in the proposed method. Under the boundary case, however, an easier and direct application is to estimate the marginal probability at an interior point which is at least one bandwidth (h) away from the endpoint. Unlike methods requiring the knowledge of all conditional densities (Chib 1995; DiCiccio et al. 1997), the nonparametric Candidate's estimate does not require specific knowledge of full conditional densities and the simulation study indicates that it performs reasonably well in overall cases including the high dimensional problem.

Moreover, the Candidate's estimate is comprehensible, reliable and easy to compute.

Two issues about the non-normality and multi-dimensionality are worth mentioned here. First, if the target distribution is not close to be normally distributed, a common approach is to transform the variables. This can be done via parameterization before generating the Markov chains to achieve greater efficiency. Nevertheless, even if the transformation is not carried out and the resulting simulations are skewed, the kernel density estimate can still work. As the recommendation of which kernel to use, the optimal is the Epanechinikov kernel based on mean integrated square error (MISE) but the Gaussian kernel is robust enough to work well. Second, when the multivariate kernel estimate is to be utilized, the window width can be set equal for all dimensions if the multivariate data spread out roughly the same in each direction. On the other hand, it may be better to use a smoothing vector if some dimension of the target distribution varies greater than other directions. Transforming the samples to be independent among different dimensions is also an alternative. The transformation of parameters even before running a Markov chain may help to achieve faster convergence rate of the chain and obtain an easier kernel estimate later on.

Although we have used the density estimate here to derive the marginal probability, we have no intention to abolish the other types of estimates. Each has its merits and position to be the best in certain applications. For instance, Laplace's method is very accurate under regularity conditions for well behaved $\pi(\theta|y)$ and harmonic mean estimate is good when the samples are within reasonable range of the likelihood. Nevertheless, the

Candidate's estimate is very easy to apply when the size of the simulated samples can be enlarged to a great number at a very low cost of computing, and when all conditional distributions are not known.

Appendix: Proofs.

Proof for Theorem 1.

By Taylor expansion one can derive the bias as

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

where $\nabla^2\pi(\theta|y)$ is a $p \times p$ matrix with the (i,j)th entry given by $\partial^2\pi(\theta|y)/(\partial\theta_i\partial\theta_j)$ and k_2 is given in condition C2. The variance is given by

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

where v is given in condition C2. Thus,

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

By choosing H so that $\lambda_j(H) = O(m^{-2/(4+p)})$ and plugging it into (7), we have

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

□

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

Lemma 1 *For a symmetric positive semi-definite matrix A and constants c_1 and c_2, the solution for the minimization problem*

$$\arg\min_H \; c_1 \left(trace\{HA\} \right)^2 + c_2 |H|^{-1/2},$$

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

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

where α_{opt} is a function of c_i, λ_j, and p. Moreover, the minimal value is given below:

$$\begin{aligned} & \inf_H \; c_1 \left(trace\{HA\} \right)^2 + c_2 |H|^{-1/2} \\ = \; & c_1^{p/(p+4)} c_2^{4/(p+4)} \left(p^{(2p+4)/(p+4)} + p^{-p/(p+4)} \right) |A|^{2/(p+4)}. \end{aligned}$$

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

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

We shall now solve for the following minimization problem:

$$\arg\min_D \; c_1 \left(trace\{D\Lambda\} \right)^2 + c_2 |D|^{-1/2}. \tag{8}$$

Consider the set of matrices with fixed diagonal elements

$$\mathcal{D} = \{D : \text{symmetric positive definite with diagonal elements } d_1, \ldots, d_p\}.$$

Then

$$\arg\max_{D \in \mathcal{D}} |D| = diag(d_1, \ldots, d_p),$$

where $diag(d_1, \ldots, d_p)$ denotes a diagonal matrix. Therefore, the minimization problem (8) can be restricted to diagonal D. Notice that

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

we have

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

Setting expression (9) to zero, we get

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

Plugging α into the minimization problem (8) and solving for the minimizer α, the solution is

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

Thus, the solution for (8) is

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

Therefore,

$$H_{opt} = G' D_{opt} G.$$

□

Proof for Theorem 2.

Begin with

$$\begin{aligned} & E_{\theta^{(1)},\ldots,\theta^{(m)}|y}\left(\frac{C}{\hat{C}} - 1\right)^2 = E_{\theta^{(1)},\ldots,\theta^{(m)}|y}\left(\frac{\hat{\pi}(\theta|y)}{\pi(\theta|y)} - 1\right)^2 \qquad (10) \\ = & \frac{E_{\theta^{(1)},\ldots,\theta^{(m)}|y}\left(\hat{\pi}(\theta|y) - \pi^2(\theta|y)\right)^2}{\pi^2(\theta|y)} \\ = & \frac{k_2^2}{4\pi^2(\theta|y)}\left(trace\left\{H\,\nabla^2\pi(\theta|y)\right\}\right)^2 + \frac{v\,\pi(\theta|y)}{m|H|^{1/2}\pi^2(\theta|y)} \\ & + o\left(trace(H)\right) + o\left(m^{-1}|H|^{-1/2}\right). \end{aligned}$$

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

$$\arg\min_H \frac{k_2^2}{4}\left(trace\left\{H\,\nabla^2\pi(\theta|y)\right\}\right)^2 + \frac{v\,\pi(\theta|y)}{m|H|^{1/2}}.$$

Recall that

$$E_{\theta^{(1)},\ldots,\theta^{(m)}|y}\left(\frac{C}{\hat{C}} - 1\right)^2 \approx \frac{k_2^2\left(trace\left\{H\,\nabla^2\pi(\theta|y)\right\}\right)^2}{4\pi^2(\theta|y)} + \frac{v}{m\pi(\theta|y)|H|^{1/2}}.$$

By Lemma 1,

$$\begin{aligned} & \inf_H \left\{\frac{k_2^2\left(trace\left\{H\,\nabla^2\pi(\theta|y)\right\}\right)^2}{4\pi^2(\theta|y)} + \frac{v}{m\pi(\theta|y)|H|^{1/2}}\right\} \\ = & c_3\left(\pi(\theta|y)\right)^{-(2p+4)/(p+4)}\left|\nabla^2\pi(\theta|y)\right|^{2/(p+4)}, \end{aligned}$$

where the constant c_3 is given by

$$c_3 = \left(\frac{k_2^2}{4}\right)^{p/(p+4)}\left(\frac{v}{m}\right)^{4/(p+4)}\left(p^{(2p+4)/(p+4)} + p^{-p/(p+4)}\right).$$

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

$$\arg\min_{\theta} \frac{|\nabla^2 \pi(\theta|y)|^2}{(\pi(\theta|y))^{2p+4}}.$$

□

References

Besag, J. E. (1989), "A Candidate's Formula: a Curious Result in Bayesian Prediction," *Biometrika*, 76, 183.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313-1321.

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.

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.

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*, London, UK: Chapman & Hall.

Hsiao, C. K. (1997), "Approximate Bayes Factors When a Mode Occurs on the Boundary," *Journal of the American Statistical Association*, 92, 656-663.

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.

Mosteller, F. and Wallace, D. L. (1964), *Applied Bayesian and Classical Inference*, first ed., reprinted in 1984 by New York: Springer-Verlag.

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.

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.

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

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, New York, NY: Springer.

Silverman, B. W. (1986), *Density Estimation*, London: Chapman & Hall.

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

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

	vol. cor. Laplace	Candidate	Candidate	Candidate
	(mode)	(mode)	(best point)	(mean)
normal				
m=1,000	1.61e-2 (2.47e-3)	3.07e-3 (4.42e-4)	1.72e-3 (2.23e-4)	3.05e-3 (4.34e-4)
m=10,000	2.20e-3 (3.12e-4)	4.99e-4 (6.74e-5)	2.53e-4 (3.28e-5)	4.91e-4 (6.66e-5)
m=100,000	1.70e-4 (2.07e-5)	7.89e-5 (1.15e-5)	4.64e-5 (6.65e-6)	7.89e-5 (1.16e-5)
t(5)				
m=1,000	7.05e-2 (5.21e-3)	4.46e-3 (4.18e-4)	4.46e-3 (4.18e-4)	4.23e-3 (3.85e-4)
m=10,000	4.89e-2 (1.38e-3)	7.37e-4 (7.68e-5)	7.37e-4 (7.68e-5)	7.30e-4 (7.62e-5)
m=100,000	5.10e-2 (6.04e-4)	1.50e-4 (1.59e-5)	1.50e-4 (1.59e-5)	1.50e-4 (1.59e-5)
t(3)				
m=1,000	1.71e-1 (8.13e-3)	9.97e-3 (6.27e-4)	9.97e-3 (6.27e-4)	9.88e-3 (6.31e-4)
m=10,000	1.76e-1 (3.78e-3)	2.13e-3 (1.42e-4)	2.13e-3 (1.42e-4)	2.11e-3 (1.41e-4)
m=100,000	1.76e-1 (1.60e-3)	3.73e-4 (2.31e-5)	3.73e-4 (2.31e-5)	3.73e-4 (2.31e-5)
gamma(2,1)				
m=1,000	8.99e-2 (6.11e-3)	5.70e-3 (4.99e-4)	1.66e-3 (2.11e-4)	1.66e-3 (2.11e-4)
m=10,000	8.63e-2 (1.65e-3)	7.53e-4 (7.47e-5)	3.11e-4 (4.29e-5)	3.11e-4 (4.29e-5)
m=100,000	8.55e-2 (6.39e-4)	1.56e-4 (1.58e-5)	5.21e-5 (5.86e-6)	5.21e-5 (5.86e-6)

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

	vol. cor. Laplace	Candidate	Candidate
gamma(1,1)	(h)	(best point, h)	(mean)
m=1,000	1.16e-2 (1.52e-3)	1.29e-3 (1.64e-4)	3.56e-3 (5.07e-4)
m=10,000	1.08e-3 (1.55e-4)	3.85e-4 (5.02e-5)	5.48e-4 (6.27e-5)
m=100,000	1.27e-4 (2.26e-5)	3.53e-4 (2.14e-5)	1.01e-4 (1.40e-5)

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

	true value	vol. cor. Laplace	harmonic mean	Candidate
m = 1000	.1917	.2453	.2151	.1816

Table 4: High-dimensional case with 100 replications.

dimension=4	vol. cor. Laplace	Candidate
normal		
m=1,000	2.43e-2 (3.54e-3)	8.94e-3 (1.38e-3)
m=10,000	1.72e-3 (2.66e-4)	2.40e-3 (2.98e-4)
product-gamma		
m=1,000	1.05e-1 (6.27e-3)	8.26e-3 (1.49e-3)
m=10,000	1.07e-1 (2.42e-3)	4.27e-3 (5.28e-4)

dimension=10	vol. cor. Laplace	Candidate
normal		
m=1,000	1.83e-2 (2.31e-3)	9.41e-2 (6.93e-3)
m=10,000	2.07e-3 (3.06e-4)	4.78e-2 (3.57e-3)
product-gamma		
m=1,000	3.90e-1 (1.42e-2)	1.25e-1 (3.21e-3)
m=10,000	3.94e-1 (4.27e-3)	6.12e-2 (1.70e-3)