

## OPTIMAL VOLUME-CORRECTED LAPLACE-METROPOLIS METHOD

SU-YUN HUANG<sup>1</sup>, CHUHSING KATE HSIAO<sup>2</sup> AND CHING-WEI CHANG<sup>3</sup>

<sup>1</sup>*Institute of Statistical Science, Academia Sinica, Taipei, 115, Taiwan, R.O.C.,*  
e-mail: syhuang@stat.sinica.edu.tw

<sup>2</sup>*Division of Biostatistics, Institute of Epidemiology, National Taiwan University, Taipei, 100,  
Taiwan, R.O.C., e-mail: ckhsiao@ha.mc.ntu.edu.tw*

<sup>3</sup>*Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, 115,  
Taiwan, R.O.C., e-mail: ashely@nhri.org.tw*

(Received January 11, 2002; revised October 31, 2002)

**Abstract.** The article provides a refinement for the volume-corrected Laplace-Metropolis estimator of the marginal likelihood of DiCiccio *et al.* The correction volume of probability  $\alpha$  in DiCiccio *et al.* is fixed and suggested to take the value  $\alpha = 0.05$ . In this article  $\alpha$  is selected based on an asymptotic analysis to minimize the mean square relative error (MSRE). This optimal choice of  $\alpha$  is shown to be invariant under linear transformations. The invariance property leads to easy implementation for multivariate problems. An implementation procedure is provided for practical use. A simulation study and a real data example are presented.

**Key words and phrases:** Bayes factor, Laplace approximation, marginal probability, Markov chain Monte Carlo.

### 1. Introduction

Computing marginal probabilities (or normalizing constants) is an important and fundamental issue in Bayesian inference. For simple cases such that the likelihood  $p(y | \theta)$  and the prior  $p(\theta)$  are analytically tractable, the integration  $\int p(y | \theta)p(\theta)d\theta$ , namely the normalizing constant, can be evaluated. For complex models, the integration may not be analytically feasible and various approximations can be applied. Similar difficulty in computation may also be present when evaluating Bayes factor is of interest. The computation can be even more complex when other integrations are required to derive the likelihood, such as the case of hierarchical models. One of the asymptotic approximations to the above integral is Laplace's method (see, for instance, Erkanli (1994); Kass *et al.* (1990); Kass and Vaidyanathan (1992); Hsiao (1997); Lindley (1980); Mosteller and Wallace (1964); Tierney and Kadane (1986)). Recently the enhancement of computational methods such as Markov chain Monte Carlo (MCMC) has helped greatly in obtaining posterior samples. With the simulated posterior sample, statistics can be derived to provide estimates such as the posterior mode, mean, and covariance. This result helps the applicability of Laplace's method in constructing a normal density approximation to the integrand (see, for instance, Chib (1995); Lewis and Raftery (1997); Tierney (1994)). This approach is referred to as Laplace-Metropolis method.

Based on the Laplace-Metropolis method, DiCiccio *et al.* (1997) took one step further to improve the method by the correction of a small volume around the posterior mode. They suggested using a volume of probability  $\alpha$  around the mode to correct

and to improve the point estimate obtained by the Laplace-Metropolis method. They suggested a default value of  $\alpha = 0.05$ , when the size of the posterior sample is between 10,000 and 100,000. This choice may, however, vary from case to case. For instance, when the posterior distribution is skewed, the accuracy of approximation depends highly on the precision of the estimated mode and thus a volume of 0.05 probability around that point may be as sensitive as is the point estimate of mode. In this article, we derive the optimal choice for  $\alpha$  based on an asymptotic analysis to minimize the MSRE.

The rest of the article is arranged as follows. In Section 2 we give a brief literature review of the progress from the Laplace approximation to the volume-corrected Laplace-Metropolis method. In Section 3 we derive the optimal choice of  $\alpha$  for volume correction. The optimal choice is shown to be invariant under linear transformations of prior parameters. Simulation studies including univariate and multivariate data and a real data example are presented in Section 4. Concluding remarks are given in Section 5. Proofs are in the Appendix.

## 2. Literature review

### 2.1 Second order Laplace approximation

Let  $\ell(\theta) = -n^{-1}\{\log p(y | \theta) + \log p(\theta)\}$ , where  $n$  is the sample size of observations  $y$ . Write the marginal probability, or the normalizing constant, as  $C = \int \exp\{-n\ell(\theta)\} d\theta$ . The idea of Laplace approximation is to approximate  $C$  by the integration of leading terms (up to the quadratic ones) of Taylor series of  $\ell(\theta)$ . The approximation is given by

$$(2.1) \quad C \approx C_{Lap} \stackrel{def}{=} \frac{e^{-n\ell(\theta^*)}}{\phi(\theta^*; \theta^*, \Sigma^*)} = (2\pi)^{d/2} |\Sigma^*|^{1/2} e^{-n\ell(\theta^*)},$$

where  $\phi(\cdot; \mu, A)$  denotes a normal density with mean  $\mu$  and covariance matrix  $A$ ,  $\theta^*$  is the mode of  $\ell(\theta)$  and

$$(2.2) \quad \Sigma^* = \left( \frac{n\partial^2\ell(\theta)}{\partial\theta^2} \Big|_{\theta=\theta^*} \right)^{-1}.$$

The approximation has error order  $C_{Lap} = C\{1 + O(n^{-1})\}$ . See Olver (1968), and McClure and Wong (1983).

For application of the Laplace approximation to marginal probabilities or Bayes factors, see Chow (1981), Jeffreys ((1961), Section 5.31), and Kass and Raftery (1995). Later, Hsiao (1997) and Paudler *et al.* (1999) investigated the asymptotic order of Laplace's method and Bayes factor in the case where the mode occurs at the boundary. All these approximations require the evaluation of the mode  $\theta^*$  and the inverse Hessian matrix  $\Sigma^*$ . To overcome the difficulty in estimating  $\theta^*$  and  $\Sigma^*$ , and to utilize the MCMC method in generating the posterior sample of  $\theta$ , hybrid methods of combining posterior simulation and Laplace approximation were proposed and studied by several authors. These are known as Laplace-Metropolis type methods.

### 2.2 Laplace-Metropolis method

Simulate a sample  $\theta^{(1)}, \dots, \theta^{(m)}$  from the posterior distribution. Based on the posterior sample, there are various estimators for the posterior mode  $\theta^*$  and the inverse Hessian matrix  $\Sigma^*$ . Let them be denoted as  $\hat{\theta}$  and  $\hat{\Sigma}$  respectively. For instance,  $\hat{\theta}$  can be the posterior sample mode or the posterior sample mean, and  $\hat{\Sigma}$  can be the inverse

Hessian of the log-posterior evaluated at  $\hat{\theta}$  or a sample covariance. Chib (1995) and Lewis and Raftery (1997) suggested to estimate the mode and covariance matrix based on the posterior sample. Hsiao *et al.* (2000) employed the method in hierarchical models for spatial data and termed it the double Laplace-Metropolis method.

### 2.3 Volume-corrected Laplace-Metropolis method

DiCiccio *et al.* (1997) proposed a local volume correction for the hybrid Laplace-Metropolis method to improve the accuracy. Let

$$(2.3) \quad B_{\theta^*, \Sigma^*}(\delta) = \{\theta \in R^d : (\theta - \theta^*)^t (\Sigma^*)^{-1} (\theta - \theta^*) \leq \delta^2\},$$

$$(2.4) \quad \alpha = \int_{B_{\theta^*, \Sigma^*}} \phi(\theta; \theta^*, \Sigma^*) d\theta = \Phi(B_{\theta^*, \Sigma^*}; \theta^*, \Sigma^*).$$

Let  $P(B_{\theta^*, \Sigma^*} | y) = \int_{B_{\theta^*, \Sigma^*}} p(\theta | y) d\theta$ , which is the posterior probability of  $B_{\theta^*, \Sigma^*}$ , and let  $\hat{P}(B_{\theta^*, \Sigma^*} | y)$  be the proportion of the sampled posterior values inside  $B_{\theta^*, \Sigma^*}$ . For all  $\theta$  in the prior support, the following equalities hold:

$$C = \frac{e^{-n\ell(\theta)}}{p(\theta | y)} = \frac{e^{-n\ell(\theta)}}{\phi(\theta; \theta^*, \Sigma^*)} \times \frac{\phi(\theta; \theta^*, \Sigma^*)}{p(\theta | y)}.$$

Therefore, DiCiccio *et al.* proposed a volume-correction modification:

$$(2.5) \quad \hat{C}_{vc} = \frac{e^{-n\ell(\theta^*)}}{\phi(\theta^*; \theta^*, \Sigma^*)} \times \frac{\alpha}{\hat{P}(B_{\theta^*, \Sigma^*} | y)},$$

where  $\theta^*$  and  $\Sigma^*$  can be estimated based on simulated posterior sample as discussed in Subsection 2.2. It was reported in DiCiccio *et al.*, based on their simulation experience, that the mean square relative error  $MSRE = E(C/\hat{C}_{vc} - 1)^2$  was stable as a function of  $\alpha$ , except when the skewness was extreme. They suggested  $\alpha = 0.05$  as a default value. However, this correction factor can be sensitive to both the shape of the posterior density and the accuracy of the mode estimate  $\hat{\theta}$ , and it can also be affected by the size of posterior sample.

## 3. Optimal volume-corrected Laplace-Metropolis method

The optimal choice of  $\alpha$  is derived based on an asymptotic analysis to minimize MSRE. We assume the following conditions, which are the common conditions for the Laplace approximation.

**ASSUMPTIONS.** The function  $\exp\{-n\ell(\theta)\}$ , as a function of  $\theta$ , (i) has continuous second derivatives, (ii) is uni-modal with an interior mode  $\theta^*$ , and (iii) its tails decrease fairly rapidly so that the integral  $\int \exp\{-n\ell(\theta)\} d\theta$  converges for all sufficiently large  $n$ .

### 3.1 The univariate case

Let  $\sigma^* = (n\ell''(\theta) |_{\theta=\theta^*})^{-1/2}$ . Recall that  $B_{\theta^*, \sigma^*}(\delta) = \{\theta \in R : |\theta - \theta^*|/\sigma^* \leq \delta\}$ . Then, we have

$$(3.1) \quad E(C/\hat{C}_{vc} - 1)^2$$

$$\begin{aligned}
&= \left\{ \frac{e^{-n\ell(\theta^*)}}{\phi(\theta^*; \theta^*, \sigma^*)} \frac{\phi(\theta^*; \theta^*, \sigma^*)}{p(\theta^* | y)} \div \left( \frac{e^{-n\ell(\theta^*)}}{\phi(\theta^*; \theta^*, \sigma^*)} \frac{\alpha}{\hat{P}(B_{\theta^*, \sigma^*} | y)} \right) - 1 \right\}^2 \\
&= \frac{\phi^2(\theta^*; \theta^*, \sigma^*)}{p^2(\theta^* | y)} E \left( \frac{\hat{P}(B_{\theta^*, \sigma^*} | y)}{\alpha} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \right)^2.
\end{aligned}$$

To select the optimal  $\alpha$  minimizing the MSRE, we find, equivalently, that

$$\arg \min_{\delta > 0} E \left( \frac{\hat{P}(B_{\theta^*, \sigma^*} | y)}{\alpha} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \right)^2.$$

**THEOREM 3.1.** *Assume assumptions (i), (ii) and (iii) hold, then the argument minimizing  $\delta$  is given by*

$$(3.2) \quad \delta_{opt} = \left\{ \frac{4.5p(\theta^* | y)}{m\sigma^*[(\sigma^*)^2 p''(\theta^* | y) + p(\theta^* | y)]^2} \right\}^{1/5},$$

where the second derivative  $p''(\theta^* | y)$  is taken with respect to  $\theta$ .

When the posterior distribution is normal, the denominator in (3.2) is zero. Therefore, the optimal  $\delta$  is infinity, resulting in  $\alpha = \hat{P} = 1$ . When  $\delta$  is infinity, the correction is done over the entire real line resulting in correction factor one. The effect of correction factor being one is equivalent to no correction at all. Therefore, there is no need to do volume correction when the posterior distribution is truly normal. In fact, the Laplace type method is a normal approximation to the posterior. When the posterior is truly normal, it becomes a parametric problem and obviously there is no need to do correction from the parametric point of view, which agrees with the theoretical phenomenon  $\delta = \infty$ . As for the case of  $\delta \rightarrow 0$ , or equivalently  $\alpha \rightarrow 0$ , the correction volume shrinks to a degenerated point, then the correction factor becomes

$$\Phi(B_{\theta^*, \Sigma^*}(\delta)) / \hat{P}(B_{\theta^*, \Sigma^*}(\delta) | y) \xrightarrow{\delta \rightarrow 0} \phi(\theta^*; \theta^*, \Sigma^*) / p(\theta^* | y).$$

As a function of  $\theta$ , the posterior  $p(\theta | y)$  is proportional to  $p(y | \theta)p(\theta)$ . When  $p(y | \theta)p(\theta)$  is known, then quantities  $\theta^*$  and  $\sigma^*$  may be derived analytically. Otherwise, they can be estimated from the posterior sample. Expression in (3.2) also involves the unknown quantities  $p(\theta^* | y)$  and  $p''(\theta^* | y)$ . Again, they should be estimated based on the posterior sample. The following proposition provides an easy means for practical usage.

**PROPOSITION 3.1.** *The quantity  $\delta_{opt}$  in (3.2) is invariant under location and scale transformation of  $\theta$ . That is, one can scale or shift the posterior variable and still get the same  $\delta_{opt}$  using the scaled/shifted posterior.*

Proposition 3.1 allows us to use linearly transformed posterior sample in calculating  $\delta_{opt}$ . It does not seem much of use for the univariate case, as there is not much trouble in getting estimates of  $p(\theta^* | y)$  and  $p''(\theta^* | y)$  from the posterior sample. However, later in the multivariate case, it becomes tedious in computation when deriving the estimates

of  $p(\theta^* | y)$  and its second derivatives  $p^{(2)}(\theta^* | y)$ . It would be helpful that we first standardize the posterior sample.

*Implementation.* Often the posterior sample is cheap to obtain, we can make the posterior sample size as large as we prefer. In calculating  $\delta_{opt}$ , we need estimates of  $p(\theta^* | y)$  and  $p''(\theta^* | y)$  for plug-in. Here we adopt kernel estimates  $\hat{p}$  and  $\hat{p}''$ , which are described below. Let  $\mathcal{G}$  be the standard Gaussian kernel and let  $\mathcal{W}(t) = (t^2 - 1)\mathcal{G}(t)$ . Take

$$(3.3) \quad \hat{p}(\theta^* | y) = \frac{1}{mh_1} \sum_{i=1}^m \mathcal{G}\left(\frac{\theta^* - \theta^{(i)}}{h_1}\right)$$

$$(3.4) \quad \hat{p}''(\theta^* | y) = \frac{1}{mh_2^3} \sum_{i=1}^m \mathcal{W}\left(\frac{\theta^* - \theta^{(i)}}{h_2}\right)$$

where  $h_1$  and  $h_2$  are bandwidths for kernels  $\mathcal{G}$  and  $\mathcal{W}$  respectively. Rules of thumb for bandwidths are given by

$$(3.5) \quad h_1 = \sigma^*(\sqrt{2m})^{-1/5} = .9330\sigma^*m^{-1/5},$$

$$(3.6) \quad h_2 = \sigma^* \left( \frac{12\sqrt{2m}}{5} \right)^{-1/9} = .8730\sigma^*m^{-1/9}.$$

The set of bandwidths (3.5) and (3.6) are derived based on criterion to minimize the mean square error for estimates of  $p(\theta^* | y)$  and  $p''(\theta^* | y)$ . See Appendix for details on theoretical background. Note that one may use other types of kernels for  $\mathcal{G}$  and  $\mathcal{W}$  as well.

### 3.2 The multivariate case

Suppose that  $\theta = (\theta_1, \dots, \theta_d)$  are  $d$ -dimensional parameters. Recall  $B_{\theta^*, \Sigma^*}(\delta)$  and  $\alpha$  defined respectively in (2.3) and (2.4). Similar to the univariate case, we have

$$(3.7) \quad E(C/\hat{C}_{vc} - 1)^2 = \frac{\phi^2(\theta^*; \theta^*, \Sigma^*)}{p^2(\theta^* | y)} E\left(\frac{\hat{P}(B_{\theta^*, \Sigma^*} | y)}{\alpha} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \Sigma^*)}\right)^2.$$

To select the optimal  $\alpha$  minimizing the MSRE, we find, equivalently, the minimizing  $\delta$  in the following theorem.

**THEOREM 3.2.** *Assume assumptions (i), (ii) and (iii) hold, then the argument minimizing  $\delta$  is given by*

$$(3.8) \quad \delta_{opt} = \left( \frac{d(d+2)^2 p(\theta^* | y) \Gamma(d/2 + 1)}{m \pi^{d/2} |\Sigma^*|^{1/2} \{ \text{tr}(\Sigma^* p^{(2)}(\theta^* | y)) + dp(\theta^* | y) \}^2} \right)^{1/(d+4)},$$

where  $p^{(2)}(\theta^* | y)$  is a  $d \times d$  matrix with  $(i, j)$ -th entry given by  $\partial^2 p(\theta^* | y) / \partial \theta_i \partial \theta_j$ . When  $\Sigma^*$  is diagonal, we have

$$(3.9) \quad \delta_{opt} = \left( \frac{d(d+2)^2 p(\theta^* | y) \Gamma(d/2 + 1)}{m \pi^{d/2} \prod_{i=1}^d \sigma_i^* [\sum_{i=1}^d (\sigma_i^*)^2 \partial^2 p(\theta^* | y) / \partial \theta_i^2 + dp(\theta^* | y)]^2} \right)^{1/(d+4)}.$$

**PROPOSITION 3.2.** *The quantity  $\delta_{opt}$  in (3.8) is invariant under invertible linear transformation of  $\theta$ . Then, one can use the standardized posterior sample, or any other linear transformed posterior sample, in calculating  $\delta_{opt}$ .*

*Implementation.* By Proposition 3.2, we first standardize the posterior sample and denote the standardized posterior sample by  $\eta^{(1)}, \dots, \eta^{(m)}$ . The optimal volume-corrected Laplace-Metropolis method can then be applied to  $\eta$ . Notice that  $C$  can be derived from

$$C = \int p(y | \theta) p_\theta(\theta) d\theta = \int p(y | \eta) p_\eta(\eta) d\eta.$$

The optimal  $\delta_{opt}$  can be now calculated based on the new variable  $\eta$  using equation (3.9) with  $\sigma_1 = \dots = \sigma_d = 1$ . The density function  $p(\eta^* | y)$  and its second derivatives can be estimated by the standardized posterior sample

$$(3.10) \quad \hat{p}(\eta^* | y) = \frac{1}{mh_1^d} \sum_{j=1}^m \prod_{i=1}^d \mathcal{G}\left(\frac{\eta_i^{(j)} - \eta_i^*}{h_1}\right),$$

and

$$(3.11) \quad \frac{\partial^2 \hat{p}(\eta^* | y)}{\partial \eta_i^2} = \frac{1}{mh_2^{2+d}} \sum_{j=1}^m \mathcal{W}\left(\frac{\eta_i^{(j)} - \eta_i^*}{h_2}\right) \prod_{\substack{\ell=1 \\ \ell \neq i}}^d \mathcal{G}\left(\frac{\eta_\ell^{(j)} - \eta_\ell^*}{h_1}\right),$$

where  $\eta^*$  stands for the mode (which is zero after standardization), and  $\mathcal{G}$  and  $\mathcal{W}$  are as defined before. Rules of thumb for bandwidths  $h_1$  and  $h_2$  are given by

$$(3.12) \quad h_1 = (2^{d/2} dm)^{-1/(4+d)}$$

and

$$(3.13) \quad h_2 = \left( \frac{.02351(4+d)(2\pi)^{d/2}}{dm} \right)^{1/(8+d)},$$

which are derived based on minimizing the mean square error for estimates of  $p(\eta^* | y)$  and  $p^{(2)}(\eta^* | y)$ .

In calculating the optimal volume, we plug in kernel estimates for posterior  $p(\theta | y)$  and its second derivatives  $p^{(2)}(\theta | y)$ . One may question the applicability of the density estimates when the dimensionality of  $\theta$  is large. In the original volume correction method, the correction ratio  $\alpha(\hat{P}(B | y))^{-1}$  is based on a uniform kernel density estimate (proportion of counts for posterior sample falling into a ball  $B$  is the same as a kernel density estimate using uniform kernel over the ball  $B$ ) at the mode with a default bandwidth corresponding to the 5% probability volume. When dimension is high, this default value becomes less reliable. The optimal choice based on an asymptotic analysis becomes more helpful, as it uses other information in assessing the right correction volume. As for the deficiency of high dimensional function estimation, the deficiency is due to the data sparseness. With low cost of generating posterior samples, the tension of data sparseness can be reduced by simulating sufficient posterior data. Moreover, the kernel estimates for  $p(\theta = 0 | y)$  and its second derivatives  $p^{(2)}(\theta = 0 | y)$  are simply two plug-in values among some other quantities used for selecting  $\alpha$ . We do not need to estimate the entire density function nor the entire second derivative function. We need

their estimates at one single point, the mode, which is a high density point with denser posterior sample data around. The data sparseness is not so critical for high density points. Furthermore, the variance of kernel density estimate for a  $d$ -dimensional  $p(\theta | y)$  is of order  $m^{-1}h^{-d}$ . This high variation is diminished quite a lot in the calculation of optimal correction volume, as we can see from Theorem 3.2 that  $\delta$  has  $p(\theta | y)$  to the power of  $1/(d + 4)$ . In all, the high dimensionality is not a critical issue here for computing the optimal volume. A step-by-step procedure is provided in Example 4.2 for illustration.

## 4. Examples

### 4.1 Simulation study

A comparison study is carried out for the 5% volume-corrected Laplace-Metropolis estimates and the proposed optimal selection.

*Example 4.1.* Samples of various sizes are generated from four different distributions, standard normal (symmetric, light tails),  $t$  with degrees of freedom 3 (symmetric, heavy tails), gamma(1,1) (boundary mode) and gamma(2,1) (highly skewed) as nominal posteriors. The volume-corrected and the optimal volume-corrected Laplace estimates are computed. One hundred replication runs are carried out for each distribution with  $m$  posterior observations ( $m = 1000, 10000$  or  $100000$ ) simulated in each replication. The mean square relative error is used to evaluate the performance of the estimates. The mean and standard error of the MSREs from these 100 replications are tabulated in Table 1.

It is clear from the table that the optimal volume-corrected Laplace-Metropolis esti-

Table 1. The numbers are average MSRE and its standard error (se). There are 100 replications in each computation with simulated data of size  $m$ . The  $\theta^*$  here is the sample posterior mode for the first three distributions and the sample posterior mean for Gamma(1,1).

| Distribution      | MSRE (se)         |                   |
|-------------------|-------------------|-------------------|
|                   | optimal vol. cor. | vol. cor.         |
| <i>m</i> = 1000   |                   |                   |
| N(0,1)            | 9.79e-4 (1.29e-4) | 2.36e-2 (3.30e-3) |
| $t(3)$            | 5.35e-3 (4.43e-4) | 1.21e-2 (1.87e-3) |
| Gamma(2,1)        | 1.70e-3 (2.61e-4) | 1.63e-2 (2.31e-3) |
| Gamma(1,1)        | 2.51e-3 (2.98e-4) | 1.18e-2 (1.73e-3) |
| <i>m</i> = 10000  |                   |                   |
| N(0,1)            | 1.53e-4 (1.92e-5) | 2.01e-3 (2.49e-4) |
| $t(3)$            | 1.01e-3 (1.13e-4) | 1.05e-3 (1.19e-4) |
| Gamma(2,1)        | 4.25e-4 (7.04e-5) | 1.55e-3 (2.19e-4) |
| Gamma(1,1)        | 1.53e-4 (1.92e-5) | 2.01e-3 (2.49e-4) |
| <i>m</i> = 100000 |                   |                   |
| N(0,1)            | 3.04e-5 (5.98e-6) | 2.15e-4 (3.23e-5) |
| $t(3)$            | 3.56e-4 (2.61e-4) | 4.29e-4 (6.55e-4) |
| Gamma(2,1)        | 8.05e-5 (1.16e-5) | 1.47e-4 (2.09e-5) |
| Gamma(1,1)        | 1.46e-4 (1.76e-5) | 2.43e-4 (3.42e-5) |

mate achieves better accuracy than the 5% volume correction method. The advantage is apparent even when  $m$  is only 1000. The performance is also stable across distributions with light or heavy tails, boundary or interior mode, being symmetric or highly skewed.

*Example 4.2.* A multivariate normal distribution and a product-gamma distribution are taken as nominal posteriors. The normal distribution is a 10-dimensional one with zero mean and covariance matrix given by

$$\Sigma_{10} = \begin{bmatrix} 1 & 0.2 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.3 & 0 \\ 0.2 & 3 & 0.6 & 0 & 0 & 0 & 0 & 0.4 & 0 & 0.2 \\ 0 & 0.6 & 7 & 0 & 0 & 0.3 & 0 & 0 & 0.1 & 0.5 \\ 0 & 0 & 0 & 4 & 0.2 & 0 & 0 & 0 & 0.4 & 0.3 \\ 0 & 0 & 0 & 0.2 & 6 & 0 & 0.4 & 0.2 & 0.4 & 0 \\ 0 & 0 & 0.3 & 0 & 0 & 8 & 0 & 0.2 & 0.3 & 0.6 \\ 0.5 & 0 & 0 & 0 & 0.4 & 0 & 2 & 0 & 0.1 & 0.3 \\ 0 & 0.4 & 0 & 0 & 0.2 & 0.2 & 0 & 5 & 0.2 & 0.2 \\ 0.3 & 0 & 0.1 & 0.4 & 0.4 & 0.3 & 0.1 & 0.2 & 7 & 0 \\ 0 & 0.2 & 0.5 & 0.3 & 0 & 0.6 & 0.3 & 0.2 & 0 & 3 \end{bmatrix}.$$

The product-gamma distribution is referring to the product of coordinate-wise gamma(2,1) densities. Here we consider a 10-dimensional product-gamma as the nominal posterior. One hundred replication runs are carried out for each distribution with  $m$  posterior observations. The average MSRE and its standard error based on 100 replications are reported for each nominal posteriors. Here we use the 10-dimensional normal with posterior sample size  $m = 1000$  as example to provide a step-by-step procedure.

1. Compute  $h_1$  and  $h_2$  using (3.12) and (3.13) (for univariate case, using (3.5) and (3.6) instead) with  $d = 10$  and  $m = 1000$ .
2. Simulate a random sample of size 1000 from  $N(0, \Sigma_{10})$  and standardize the sample. In steps below, when posterior sample is called upon, refer to the standardized version.
3. Compute  $\hat{p}(0)$  and  $\hat{p}^{(2)}(0)$  using (3.10) and (3.11) (for univariate case, using (3.3) and (3.4) instead).
4. Compute the optimal  $\delta$  using (3.9) with  $\sigma_1^* = \sigma_2^* = \dots = \sigma_{10}^* = 1$ ,  $d = 10$ ,  $m = 1000$  and with  $\hat{p}(0)$  and  $\hat{p}^{(2)}(0)$  from step 3 above.
5. Compute  $\hat{C} = \hat{C}_{Lap} \times$  correction factor, which is

$$\hat{C} = (2\pi)^5 \times \frac{\Phi(B_{0,I}(\delta_{opt}))}{\text{counts of sampled posterior inside the ball } B_{0,I}(\delta_{opt})/1000},$$

where  $I$  is a 10-dimensional identity matrix.

6. Compute  $\text{MSRE} = (C/\hat{C} - 1)^2$  with true  $C = 1/\phi(0; 0, \Sigma_{10}) = (2\pi)^5 |\Sigma_{10}|^{1/2}$ .
7. Go to step 2 and repeat for 100 runs.
8. There are 100 MSREs from the 100 runs. Report the average and its standard error.

Results for the two nominal posterior distributions are listed in Table 2. The optimal method often achieves MSRE one decimal less than that of the 5% one. Its standard errors are smaller too. Note that this is a simulation example with 100 replication runs.

Table 2. The numbers are average MSRE and its standard error (se). There are 100 replications in each computation with simulated data of size  $m$ . The  $\theta^*$  here are the sample posterior modes.

| Distribution        | MSRE (se)         |                   |
|---------------------|-------------------|-------------------|
|                     | optimal           | vol. cor.         |
| $d = 10, m = 1000$  |                   |                   |
| $N(0, \Sigma_{10})$ | 2.84e-3 (2.85e-4) | 4.43e-2 (1.59e-3) |
| product-Gamma(2,1)  | 1.75e-1 (7.50e-3) | 4.31e-1 (1.11e-2) |
| $d = 10, m = 10000$ |                   |                   |
| $N(0, \Sigma_{10})$ | 3.21e-4 (1.21e-4) | 3.37e-3 (4.38e-4) |
| product-Gamma(2,1)  | 9.35e-2 (2.20e-3) | 4.15e-1 (4.41e-3) |

When calculating  $C$  in a real data example, perform steps 1 to 5 only, and then report  $\hat{C}$ .

#### 4.2 Toxicological experiment data

*Example 4.3.* We consider another example for the purpose of illustration. This was a toxicological experiment involved 16 female rats treated with a chemical diet during their pregnancy and lactation period. For each rat in its own litter, the number of pups  $n_i$  and the number of survived pups  $y_i$  at the 21st day were recorded. Listed below are the data given as fractions  $y_i/n_i$ .

$$12/12, 11/11, 10/10, 9/9, 10/11, 9/10, 9/10, 8/9, \\ 8/9, 4/5, 7/9, 4/7, 5/10, 3/6, 3/10, 0/7.$$

The data were analyzed by Williams (1975) and several others such as the BUGS group (Spiegelhalter *et al.* (1995)). For this treatment group, the observed  $y_i$  conditioning on  $q_i$ , the survival rate for  $i$ -th litter, follows a binomial distribution, and  $q_i$  is the probability of survival for the  $i$ -th litter. Williams had shown that there existed significant variability among the 16  $q_i$ 's, and therefore a distribution was considered to describe this heterogeneity. Here we choose a beta( $\alpha, \beta$ ) prior for the  $q_i$  and next assume a uniform(0, 1000) for both hyperparameters  $\alpha$  and  $\beta$  to indicate the unknown status of these two parameters. An alternative parameterization is to use parameters  $(\xi, \omega)$  given by  $\xi = \alpha/(\alpha + \beta)$  and  $\omega = 1/(\alpha + \beta)$ . The parameter  $\xi$  has a natural interpretation of being the mean while  $\omega$  can be viewed as a precision parameter. Consequently, the magnitude of  $\omega$  implies the deviation of the data from constant survival probability. The maximum likelihood estimate of  $(\xi, \omega)$  is (0.74, 0.47) indicating a high average surviving probability across all litters but great heterogeneity of surviving chances of these pups from different female rats.

To calculate further the posterior probability, we focus on the estimation of the marginal probability,

$$C = \int \left[ \int \prod_{i=1}^{16} p(y_i | q_i) p(q_i | \alpha, \beta) dq_i \right] p(\alpha, \beta) d\alpha d\beta,$$

where  $p(\cdot | \alpha, \beta)$  is the prior probability density function for  $q_i$  conditional on given  $\alpha$  and  $\beta$ , and  $p(\cdot, \cdot)$  is the joint prior probability density function for  $\alpha$  and  $\beta$ . Note that the

likelihood and prior can be written in terms of  $(\xi, \omega)$ , if it is preferred. When choosing the beta prior for  $q_i$ 's, the integration over the 16  $q_i$ 's are analytically feasible and a numerical integration over  $(\alpha, \beta)$  can be carried out. The  $C$  value obtained by numerical integration is reported as ‘true value’ and is compared with Laplace type approximations below.

Based on 5000 posterior sample points of  $\alpha$  and  $\beta$  using BUGS program with 1000 burn-in, the usual Laplace, volume corrected and optimal volume corrected Laplace-Metropolis approximations to  $\log C$  are listed below.

| true  | Laplace | vol. cor. | optimal vol. cor. |
|-------|---------|-----------|-------------------|
| -42.6 | -41.5   | -41.6     | -42.4             |

Although the magnitude of these numbers look similar enough, their mean square relative errors in the original scale are much different, as listed below.

|      | Laplace | vol. cor. | optimal vol. cor. |
|------|---------|-----------|-------------------|
| MSRE | 0.4451  | 0.3996    | 0.0329            |

The optimal volume correction gives the closest value to the true one obtained from numerical integration. If the estimated normalizing constant is used further to calculate the posterior probability of quantity of interest, it may have great influence as well. For instance, since the ratio of estimated posterior probability ( $\hat{P}$ ) to true posterior probability ( $P$ ) equals  $C/\hat{C}$ , it is possible that the ratio may become as small as 0.33 for the usual Laplace approximation, or 0.37 for the volume corrected approximation, while it remains 0.82, much closer to one, using the optimal volume correction.

## 5. Concluding remarks

We have demonstrated in this article how to select the appropriate correction volume for the Laplace-Metropolis method. We have also outlined implementation procedures for the method. The simulation study and the real data example both show that the proposed correction volume has better accuracy than the default choice of volume  $\alpha = 0.05$ .

There are several advantages of the proposed method over the 5% version. Firstly, the optimal choice takes into account the shape of the target distribution. Secondly, it adapts to the size of the simulated posterior sample and adjusts the correction volume accordingly. Thirdly, when the distribution has boundary mode, the optimal volume correction version has better adjustability to lead to the right correction factor. Also, when multivariate parameters are of interest, the method can still be easily implemented and effectively calculated by first standardizing the posterior sample.

## Appendix: Theoretical details

PROOF FOR THEOREM 3.1. Let  $\hat{p}(\theta^* | y) = \hat{P}(B_{\theta^*, \sigma^*} | y) / (2\delta\sigma^*)$ . Notice that  $\hat{p}(\theta^* | y)$  is actually a kernel density estimate using the uniform kernel  $\mathcal{K}(\theta) = 1/2$  over  $[-1, 1]$ , and the corresponding bandwidth is  $\delta\sigma^*$ .

$$\arg \min_{\delta > 0} E(C/\hat{C}_{vc} - 1)^2$$

$$\begin{aligned}
&= \arg \min_{\delta > 0} E \left( \frac{\hat{P}(B_{\theta^*} | y)}{\alpha} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \right)^2 \\
&= \arg \min_{\delta > 0} E \left( \frac{\hat{p}(\theta^* | y)}{(2\delta\sigma^*)^{-1} \int_{B_{\theta^*, \sigma^*}} \phi(\theta; \theta^*, \sigma^*) d\theta} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \right)^2 \\
&= \arg \min_{\delta > 0} \left\{ \left( \frac{E\hat{p}(\theta^* | y)}{(2\delta\sigma^*)^{-1} \int_{B_{\theta^*, \sigma^*}} \phi(\theta; \theta^*, \sigma^*) d\theta} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \right)^2 \right. \\
&\quad \left. + \text{Var} \left( \frac{\hat{p}(\theta^* | y)}{(2\delta\sigma^*)^{-1} \int_{B_{\theta^*, \sigma^*}} \phi(\theta; \theta^*, \sigma^*) d\theta} \right) \right\}.
\end{aligned}$$

Let  $k_2$  denote the second moment of the uniform kernel, then  $k_2 = 1/3$ . Start with the first term above (without the square), we have

$$\begin{aligned}
(A.1) \quad \text{bias} &= \frac{E\hat{p}(\theta^* | y)}{(2\delta\sigma^*)^{-1} \int_{B_{\theta^*, \sigma^*}} \phi(\theta; \theta^*, \sigma^*) d\theta} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \\
&= \frac{p(\theta^* | y) + k_2 p''(\theta^* | y)(\sigma^*\delta)^2/2 + o(\delta^2)}{\phi(\theta^*; \theta^*, \sigma^*) + \phi''(\theta^*; \theta^*, \sigma^*)(\sigma^*\delta)^2/6 + o(\delta^2)} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \\
&= \frac{\{p''(\theta^* | y)\phi(\theta^*; \theta^*, \sigma^*) - p(\theta^* | y)\phi''(\theta^*; \theta^*, \sigma^*)\}(\sigma^*\delta)^2/6}{\phi^2(\theta^*; \theta^*, \sigma^*)} + o(\delta^2) \\
&= \frac{\sqrt{2\pi}\sigma^*\delta^2\{(\sigma^*)^2p''(\theta^* | y) + p(\theta^* | y)\}}{6} + o(\delta^2),
\end{aligned}$$

where  $\phi(\theta^*; \theta^*, \sigma^*) = (\sqrt{2\pi}\sigma^*)^{-1}$  and  $\phi''(\theta^*; \theta^*, \sigma^*) = -(2\pi)^{-1/2}(\sigma^*)^{-3}$ .

The variance term above is given by

$$\begin{aligned}
(A.2) \quad \text{Var} &\left( \frac{\hat{p}(\theta^* | y)}{(2\delta\sigma^*)^{-1} \int_{B_{\theta^*}} \phi(\theta; \theta^*, \sigma^*) d\theta} \right) \\
&= \frac{p(\theta^* | y)\nu/(m\sigma^*\delta) + o((m\delta)^{-1})}{\{(2\delta\sigma^*)^{-1} \int_{B_{\theta^*}} \phi(\theta; \theta^*, \sigma^*) d\theta\}^2} \\
&= \frac{p(\theta^* | y)\nu}{\phi^2(\theta^*; \theta^*, \sigma^*)m\sigma^*\delta} + o\left(\frac{1}{m\delta}\right) \\
&= \frac{\pi\sigma^*p(\theta^* | y)}{m\delta} + o\left(\frac{1}{m\delta}\right),
\end{aligned}$$

where  $\nu = \int_{-1}^1 \mathcal{K}^2(\theta) d\theta$  ( $\nu = 1/2$  for uniform kernel).

From (A.1) and (A.2), the mean square relative error is given by

$$\begin{aligned}
(A.3) \quad \text{MSRE} &\approx \frac{\phi^2(\theta^*; \theta^*, \sigma^*)}{p^2(\theta^* | y)} \\
&\times \left( \frac{\pi(\sigma^*)^2\delta^4\{(\sigma^*)^2p''(\theta^* | y) + p(\theta^* | y)\}^2}{18} + \frac{\pi\sigma^*p(\theta^* | y)}{m\delta} \right).
\end{aligned}$$

Take the derivative of (A.3) with respect to  $\delta$ , set it to zero and solve for  $\delta$ . Then we get the stated optimal  $\delta$ .  $\square$

PROOF FOR THEOREM 3.2. Let  $v = \delta^d \pi^{d/2} |\Sigma^*|^{1/2} \{\Gamma(d/2 + 1)\}^{-1}$  be the volume of the ellipsoid  $B_{\theta^*, \Sigma^*}$ . Notice that  $\hat{p}(\theta^* | y) = \hat{P}(B_{\theta^*, \Sigma^*} | y)/v$  is an estimate for the posterior density  $p(\theta^* | y)$  with the following uniform kernel over an  $d$ -dimensional ellipsoid

$$(A.4) \quad \mathcal{K}(\theta) = \frac{\Gamma(d/2 + 1)}{\pi^{d/2} |\Sigma^*|^{1/2}}, \quad \text{for such } \theta \text{ that } \theta^t \Sigma^{*-1} \theta \leq 1,$$

and the corresponding kernel bandwidth is  $\delta$ . The optimal correction volume  $\alpha$  is determined by the argument minimizing  $\delta$  from

$$\arg \min_{\delta > 0} E \left( \frac{\hat{P}(B_{\theta^*, \Sigma^*} | y)}{\alpha} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \Sigma^*)} \right)^2.$$

The bias of  $\hat{p}(\theta^* | y)$  is given by

$$(A.5) \quad E\hat{p}(\theta^* | y) - p(\theta^* | y) = \frac{k_2 \delta^2 \text{tr}(\Sigma^* p^{(2)}(\theta^* | y))}{2} + o(\delta^2),$$

where  $k_2$  is the second moment of the uniform kernel (A.4) and its value is computed below:

$$\begin{aligned} k_2 &= \frac{\Gamma(d/2 + 1)}{\pi^{d/2}} \int_{\theta_1^2 + \dots + \theta_d^2 \leq 1} \theta_1^2 d\theta_1 \cdots d\theta_d \\ &= \frac{\Gamma(d/2 + 1)}{\pi^{d/2}} \int_{\theta_1^2 \leq 1} \theta_1^2 \int_{\theta_2^2 + \dots + \theta_d^2 \leq 1 - \theta_1^2} d\theta_2 \cdots d\theta_d d\theta_1 \\ &= \frac{\Gamma(d/2 + 1)}{\pi^{d/2}} \frac{\pi^{(d-1)/2}}{\Gamma((d+1)/2)} \int_{\theta_1^2 \leq 1} \theta_1^2 (1 - \theta_1^2)^{(d-1)/2} d\theta_1 \\ &= \frac{\Gamma(d/2 + 1)}{\pi^{1/2} \Gamma((d+1)/2)} \int_0^1 t^{1/2} (1-t)^{(d-1)/2} dt \\ &= \frac{\Gamma(d/2 + 1)}{\pi^{1/2} \Gamma((d+1)/2)} \frac{\Gamma(3/2) \Gamma((d+1)/2)}{\Gamma(d/2 + 2)} \\ &= \frac{1/2}{d/2 + 1} = \frac{1}{d+2}. \end{aligned}$$

The variance of  $\hat{p}(\theta^* | y)$  is given by

$$(A.6) \quad \text{Var } \hat{p}(\theta^* | y) = \frac{p(\theta^* | y)}{m \delta^d} \cdot \frac{\Gamma(d/2 + 1)}{\pi^{d/2} |\Sigma^*|^{1/2}} + o\left(\frac{1}{m \delta^d}\right).$$

Since that

$$E(C/\hat{C}_{vc} - 1)^2 = \frac{\phi^2(\theta^*; \theta^*, \sigma^*)}{p^2(\theta^* | y)} E \left( \frac{\hat{p}(\theta^* | y)}{v^{-1} \int_{B_{\theta^*, \Sigma^*}} \phi(\theta; \theta^*, \sigma^*) d\theta} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \right)^2.$$

The optimal  $\delta$  is given by solving the problem:

$$\begin{aligned} \arg \min_{\delta > 0} & \left\{ \left( \frac{E\hat{p}(\theta^* | y)}{v^{-1} \int_{B_{\theta^*, \Sigma^*}} \phi(\theta; \theta^*, \sigma^*) d\theta} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \right)^2 \right. \\ & \left. + \text{Var} \left( \frac{\hat{p}(\theta^* | y)}{v^{-1} \int_{B_{\theta^*, \Sigma^*}} \phi(\theta; \theta^*, \sigma^*) d\theta} \right) \right\}. \end{aligned}$$

By (A.5) and (A.6) and similar technique used in the univariate case, we have

$$(A.7) \quad \text{Var} \left( \frac{\hat{p}(\theta^* | y)}{v^{-1} \int_{B_{\theta^*, \Sigma^*}} \phi(\theta; \theta^*, \sigma^*) d\theta} \right) \approx \frac{p(\theta^* | y) \Gamma(d/2 + 1)}{m\pi^{d/2} \delta^d |\Sigma^*|^{1/2} \phi^2(\theta^*; \theta^*, \sigma^*)}$$

and

$$(A.8) \quad \begin{aligned} & \left( \frac{E\hat{p}(\theta^* | y)}{v^{-1} \int_{B_{\theta^*, \Sigma^*}} \phi(\theta; \theta^*, \sigma^*) d\theta} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \right)^2 \\ & \approx \left( \frac{p(\theta^* | y) + k_2 \delta^2 \text{tr}(\Sigma^* p^{(2)}(\theta^* | y)) / 2}{\phi(\theta^*; \theta^*, \sigma^*) + w(\theta^*, \Sigma^*)} - \frac{p(\theta^* | y)}{\phi(\theta^*; \theta^*, \sigma^*)} \right)^2 \\ & \approx \left( \frac{\phi(\theta^*; \theta^*, \sigma^*) k_2 \delta^2 \text{tr}(\Sigma^* p^{(2)}(\theta^* | y)) - 2w(\theta^*, \Sigma^*) p(\theta^* | y)}{2\phi^2(\theta^*; \theta^*, \sigma^*)} \right)^2, \end{aligned}$$

where  $w(\theta^*, \Sigma^*)$  is given by

$$\begin{aligned} w(\theta^*, \Sigma^*) &= \frac{1}{2v} \int_{B_{\theta^*, \Sigma^*}} (\theta - \theta^*)' \phi^{(2)}(\theta^*; \theta^*, \sigma^*) (\theta - \theta^*) d\theta \\ &= -\frac{\phi(\theta^*; \theta^*, \sigma^*)}{2v} \int_{B_{\theta^*, \Sigma^*}} (\theta - \theta^*)' (\Sigma^*)^{-1} (\theta - \theta^*) d\theta \\ &= -\frac{\phi(\theta^*; \theta^*, \sigma^*)}{2v} \int_{\theta_1^2 + \dots + \theta_d^2 \leq \delta^2} (\theta_1^2 + \dots + \theta_d^2) |\Sigma^*|^{1/2} d\theta \\ &= -\frac{\phi(\theta^*; \theta^*, \sigma^*) \delta^{d+2} |\Sigma^*|^{1/2}}{2v} \int_{\theta_1^2 + \dots + \theta_d^2 \leq 1} (\theta_1^2 + \dots + \theta_d^2) d\theta \\ &= \frac{-dk_2 \delta^2}{2(2\pi)^{d/2} |\Sigma^*|^{1/2}}. \end{aligned}$$

It is left to check that  $\delta_{opt}$  in (3.8) is the minimizing argument for the sum of (A.7) and (A.8), which is straightforward and is omitted here.  $\square$

**PROOF FOR PROPOSITION 3.2.** Take transformation  $\eta = C_1 \theta + C_2$ , where  $C_1$  is an arbitrary  $d \times d$  full rank matrix and  $C_2$  is a  $d$ -vector. Then the posterior mode of  $\eta$  is  $\eta^* = C_1 \theta^* + C_2$ , the conditional pdf of  $\eta$  given  $y$  is  $p(y | \eta) = p(y | \theta) |C_1|^{-1}$  and the inverse Hessian at  $\eta^*$  is

$$\Sigma_\eta^* = \left( -\frac{\partial^2 \log(p(y | \eta)p(\eta))}{\partial \eta^2} \Big|_{\eta=\eta^*} \right)^{-1} = C_1 \Sigma^* C_1^t.$$

The optimal  $\delta$  derived in Theorem 3.2 based on  $\eta$  is then given by

$$\begin{aligned} \delta_{opt, \eta} &= \left( \frac{d(d+2)^2 p(\eta^* | y) \Gamma(d/2 + 1)}{m\pi^{d/2} |\Sigma_\eta^*|^{1/2} \{ \text{tr}(\Sigma_\eta^* p^{(2)}(\eta^* | y)) + dp(\eta^* | y) \}^2} \right)^{1/(d+4)} \\ &= \left( \frac{d(d+2)^2 p(\theta^* | y) |C_1|^{-1} \Gamma(d/2 + 1)}{m\pi^{d/2} |\Sigma^*|^{1/2} |C_1| \{ \text{tr}(\Sigma^* p^{(2)}(\theta^* | y)) |C_1|^{-1} + dp(\theta^* | y) |C_1|^{-1} \}^2} \right)^{1/(d+4)} \\ &= \left( \frac{d(d+2)^2 p(\theta^* | y) \Gamma(d/2 + 1)}{m\pi^{d/2} |\Sigma^*|^{1/2} \{ \text{tr}(\Sigma^* p^{(2)}(\theta^* | y)) + dp(\theta^* | y) \}^2} \right)^{1/(d+4)} = \delta_{opt}. \end{aligned}$$

That is, the optimal  $\delta$  derived based on  $\eta$ , denoted by  $\delta_{opt,\eta}$ , is the same as the optimal  $\delta$  derived based on  $\theta$ , denoted by  $\delta_{opt}$ .  $\square$

*Theoretical background for rules of thumb (3.5) and (3.6).* The asymptotic mean square error for  $\hat{p}(\theta^* | y)$  is

$$E(\hat{p}(\theta^* | y) - p(\theta^* | y))^2 \approx \left( \frac{p''(\theta^* | y)}{2} \right)^2 h_1^4 + \frac{p(\theta^* | y)}{2\sqrt{\pi}mh_1}.$$

Thus, the optimal bandwidth minimizing the above asymptotic mean square error is given by

$$h_1 = \left( \frac{p(\theta^* | y)}{2\sqrt{\pi}(p''(\theta^* | y))^2 m} \right)^{1/5}.$$

The rule of thumb (3.5) is obtained by using a normal reference density with mean  $\theta^*$  and variance  $\sigma^*$  for  $p(\theta | y)$ . That is, plug in  $1/\sqrt{2\pi}$  for  $p(\theta^* | y)$  and  $-1/\sqrt{2\pi}$  for  $p''(\theta^* | y)$  above.

The asymptotic mean square error for  $\hat{p}''(\theta^* | y)$  is

$$E(\hat{p}''(\theta^* | y) - p''(\theta^* | y))^2 \approx \left( \frac{12p^{(4)}(\theta^* | y)}{4!} \right)^2 h_2^4 + \frac{3p(\theta^* | y)}{8\sqrt{\pi}mh_2^5}.$$

Thus, the optimal bandwidth minimizing the above asymptotic mean square error is given by

$$h_2 = \left( \frac{15p(\theta^* | y)}{8\sqrt{\pi}(p^{(4)}(\theta^* | y))^2 m} \right)^{1/9}.$$

Again, by approximating  $p(\theta | y)$  by a normal reference density, we can get the rule of thumb (3.6).  $\square$

*Theoretical background for rules of thumb (3.12) and (3.13).* The asymptotic mean square error for

$$\hat{p}(\eta^* | y) = \frac{1}{mh_1^d} \sum_{j=1}^m \prod_{i=1}^d \mathcal{G} \left( \frac{\eta_i^{(j)} - \eta_i^*}{h_1} \right),$$

is given by

$$(A.9) \quad E(\hat{p}(\eta^* | y) - p(\eta^* | y))^2 \approx \left( \sum_{i=1}^d \frac{\partial^2 p(\eta^* | y)}{2\partial\eta_i^2} \right)^2 h_1^4 + \frac{p(\eta^* | y)}{(2\sqrt{\pi})^d mh_1^d}.$$

The optimal  $h_1$  is given by

$$h_1^{d+4} = d(2\sqrt{\pi})^{-d} p(\eta^* | y) \left( \sum_{i=1}^d \frac{\partial^2 p(\eta^* | y)}{\partial\eta_i^2} \right)^{-2} m^{-1}.$$

By using the standard normal density as a reference density for  $p(\eta | y)$ , then  $p(\eta^* | y)$  is replaced by  $(2\pi)^{-d/2}$  and  $\sum_{i=1}^d \partial^2 p(\eta^* | y)/\partial\eta_i^2$  is replaced by  $-d(2\pi)^{-d/2}$ . Thus, we have the rule of thumb (3.12).

The asymptotic mean square error for

$$\frac{\partial^2 \hat{p}(\eta^* | y)}{\partial \eta_i^2} = \frac{1}{mh_2^{2+d}} \sum_{j=1}^m \mathcal{W} \left( \frac{\eta_i^{(j)} - \eta_i^*}{h_2} \right) \prod_{\substack{\ell=1 \\ \ell \neq i}}^d \mathcal{G} \left( \frac{\eta_\ell^{(j)} - \eta_\ell^*}{h_1} \right),$$

is given below. For  $i = 1$ ,

$$\begin{aligned} (A.10) \quad E & \left( \frac{\partial^2 \hat{p}(\eta^* | y)}{\partial \eta_1^2} - \frac{\partial^2 p(\eta^* | y)}{\partial \eta_1^2} \right)^2 \\ & = \text{bias}^2 + \text{Var} \\ & \approx \left( \frac{12\partial^4 p(\eta^* | y)/\partial \eta_1^4}{4!p_1(\eta_1^* | y)} + \sum_{i=2}^d \frac{\partial^2 p(\eta^* | y)/\partial \eta_i^2}{2p_i(\eta_i^* | y)} \right)^2 \\ & \quad \times p^2(\eta^* | y)h_2^4 + \frac{3p(\eta^* | y)}{2^{d+2}\pi^{d/2}mh_2^{d+4}}, \end{aligned}$$

where  $p_i(\eta_i | y) = \int p(\eta | y)d\eta_1 \cdots d\eta_{i-1}d\eta_{i+1} \cdots d\eta_d$ . Then the optimal  $h_2$  is given by

$$h_2^{d+8} = \left( \frac{\partial^4 p(\eta^* | y)/\partial \eta_1^4}{p_1(\eta_1^* | y)} + \sum_{i=2}^d \frac{\partial^2 p(\eta^* | y)/\partial \eta_i^2}{p_i(\eta_i^* | y)} \right)^{-2} \times \frac{3(d+4)p^{-1}(\eta^* | y)}{2^{d+2}\pi^{d/2}m}.$$

Use the standard normal density as a reference density for  $p(\eta | y)$ . Then we can get the rule of thumb (3.13).

## REFERENCES

- Chib, S. (1995). Marginal likelihood from the Gibbs output, *J. Amer. Statist. Assoc.*, **90**, 1313–1321.
- Chow, G. C. (1981). A comparison of the information and posterior probability criteria for model selection, *J. Econometrics*, **16**, 21–33.
- DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximation, *J. Amer. Statist. Assoc.*, **92**, 903–915.
- Erkanli, A. (1994). Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space, *J. Amer. Statist. Assoc.*, **89**, 250–258.
- Hsiao, C. K. (1997). Approximation Bayes factor when a mode occurs on the boundary, *J. Amer. Statist. Assoc.*, **92**, 656–663.
- Hsiao, C. K., Tzeng, J. Y. and Wang, C. H. (2000). Comparing the performance of two indices for spatial model selection: Application to two mortality data, *Statistics in Medicine*, **19**, 1915–1930.
- Jeffreys, H. (1961). *Theory of Probability*, 3rd ed., Oxford University Press, Oxford, U.K.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *J. Amer. Statist. Assoc.*, **90**, 773–795.
- Kass, R. E. and Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions, *J. Roy. Statist. Soc. Ser. B*, **54**, 129–144.
- Kass, R. E., Tierney, L. and Kadane, J. B. (1990). The validity of posterior asymptotic expansions based on Laplace's method, *Bayesian and Likelihood Methods in Statistics and Econometrics* (eds. S. Geisser, J. S. Hodges, S. J. Press and A. Zellner), North-Holland, New York.
- Lewis, S. M. and Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator, *J. Amer. Statist. Assoc.*, **92**, 648–655.
- Lindley, D. V. (1980). Approximate Bayesian methods. With discussion, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. M. F. Smith), 223–245, University Press, Valencia, Spain.

- McClure, J. P. and Wong, R. (1983). Error bounds for multidimensional Laplace approximation, *J. Approx. Theory*, **37**, 372–390.
- Mosteller, F. and Wallace, D. L. (1964). *Applied Bayesian and Classical Inference*, first ed., reprinted in 1984 by Springer, New York.
- Olver, F. W. J. (1968). Error bounds for the Laplace approximation for definite integrals, *J. Approx. Theory*, **1**, 293–313.
- Pauler, D. K., Wakefield, J. C. and Kass, R. E. (1999). Bayes factors and approximations for variance component models, *J. Amer. Statist. Assoc.*, **94**, 1242–1253.
- Spiegelhalter, D. J., Thomas, A., Best, N. G. and Gilks, W. R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling*, Version 0.50, MRC Biostatistics Unit, Cambridge.
- Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Ann. Statist.*, **22**, 1701–1762.
- Tierney, L. and Kadane, J. B. (1986). Accurate approximation for posterior moments and marginal densities, *J. Amer. Statist. Assoc.*, **81**, 82–86.
- Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity, *Biometrics*, **31**, 949–952.