Title	Predictive marginallikelihood for the evaluation of Bayesian models
Sub Title	
Author	安道, 知寛(Ando, Tomohiro)
Publisher	慶應義塾経営管理学会
Publication year	2007
Jtitle	慶應義塾経営管理学会リサーチペーパー・シリーズ No.96 (2007. 4)
JaLC DOI	
Abstract	The problem of evaluating the goodness of the predictive distributions of Bayesian models is investigated. A predictive marginal likelihood is proposed as an estimator of the posterior mean of the expected likelihood of the predictive distribution. Under the model misspecification situation, the proposed criterion is developed by correcting the asymptotic bias of the posterior mean of the likelihood as an estimate of its expected likelihood. The use of the resampling approach in model evaluation is also discussed.
Notes	
Genre	Technical Report
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40003002-00000096-0001

Predictive marginal likelihood for the evaluation of Bayesian models

安道　知寛

慶應義塾大学大学院経営管理研究科

慶應義塾経営管理学会
リサーチペーパー・シリーズ
No.９６　（2007年4月）

Predictive marginal likelihood for the evaluation of Bayesian models

BY TOMOHIRO ANDO

Graduate School of Business Administration, Keio University, 2-1-1 Hiyoshi-Honcho, Kohoku-ku, Yokohama-shi, Kanagawa, 223-8523, Japan

andoh@hc.cc.keio.ac.jp

SUMMARY

The problem of evaluating the goodness of the predictive distributions of Bayesian models is investigated. A predictive marginal likelihood is proposed as an estimator of the posterior mean of the expected likelihood of the predictive distribution. Under the model misspecification situation, the proposed criterion is developed by correcting the asymptotic bias of the posterior mean of the likelihood as an estimate of its expected likelihood. The use of the resampling approach in model evaluation is also discussed.

Some key words: Bayesian models; Markov chain Monte Carlo; Model misspecification; Posterior Bayes factor.

1. Introduction

Suppose a set of n independent observations y_n are generated from an unknown probability density $g(y_n)$ and that a parametric family of distributions with densities $\{f(y|\theta); \theta \in \Theta \subset R^p\}$ is utilised as an approximation of the true model. In the Bayesian framework, an inference for θ is provided by its posterior distribution, $\pi(\theta|y_n) \propto L(y_n|\theta)\pi(\theta)$, where $L(y_n|\theta)$ is the likelihood function and $\pi(\theta)$ is a prior distribution. The predictive distribution for a future observation z_n from the true model is $q(z_n|y_n) = \int f(z_n|\theta)\pi(\theta|y_n)d\theta$. The remained problem is how to evaluate the goodness of the predictive distribution, known as model selection problem.

Model selection is a fundamental task in statistical modeling process. The Bayes factor (Kass & Raftery, 1995) has played a major role in the evaluation of the goodness of the Bayesian models. Despite its popularity, the Bayes factor has come under increasing criticism. The most serious difficulty in the use of the Bayes factor is its sensitivity to prior distributions. The use of non-informative prior leads to the severe situation; the Bayes factor is not well-defined. Many studies therefore have been conducted to evaluate the goodness of Bayesian models (Gelfand & Dey, 1994; Kass & Raftery, 1995; O'Hagan 1995; Berger & Pericchi, 1996; Perez & Berger, 2002; Ando 2006).

To overcome the difficulties of the Bayes factor, Aitkin (1991) proposed a posterior Bayes factor that chooses the model with the largest value of the posterior mean of the likelihood. However, note that, the same data are used both to construct the posterior distribution and to compute the posterior mean of the likelihood (Kadane & Lazar, 2004). It is obvious that the posterior mean of the likelihood has a positive bias comparing with the posterior mean of the expected likelihood.

From a predictive point of view, it is natural to evaluate the constructed model based on the posterior mean of the expected likelihood. The main aim of this paper is to propose a predictive marginal likelihood, an estimator of the posterior mean of the expected likelihood of the predictive distribution. Under the model misspecification, the proposed criterion is developed by using an asymptotic theory. It consists of two terms; the posterior mean of the likelihood and an asymptotic bias of the posterior mean of the likelihood as an estimate of its expected likelihood.

§2 gives a main result. Some observations of the proposed criterion are provided in §3. §4 $\sim$ §6 conduct Monte Carlo simulations to investigate the performance of the proposed criterion.

2. Main result; Predictive marginal likelihood

The best predictive distribution is determined by maximizing the logarithmic posterior mean of expected likelihood

$$\eta = E_{z_n}[\log m(z_n|y_n)] = \int \left[\log\left\{\int L(z_n|\theta)\pi(\theta|y_n)d\theta\right\}\right] g(z_n) \tag{1}$$

among different Bayesian models. It is obvious that the quantity η depends on the specified model, and further depends on the unknown true model $g(z)$. The problem therefore is how to estimate the logarithmic posterior mean of expected likelihood.

A natural estimator of (1) is the logarithmic posterior mean of the likelihood, proposed by Aitkin (1991):

$$\hat{\eta} = \log m(y_n|y_n) = \log \int L(y_n|\theta)\pi(\theta|y_n)d\theta. \tag{2}$$

It is obvious that the logarithmic posterior mean of the likelihood generally provides a positive bias as an estimator of η. Therefore, the bias correction of the logarithmic posterior mean of the likelihood is required.

Employing an information theoretic argument (Akaike, 1974), the bias b of $\hat{\eta}$ in estimating η is defined as

$$b = E_{y_n}(\hat{\eta} - \eta) = \int (\hat{\eta} - \eta) g(y_n), \tag{3}$$

where expectation is taken over the joint distribution of y_n. Under some regularity conditions, we evaluated the asymptotic bias (3) under the model misspecification. The result is given in the following theorem.

Theorem 1 *Let η and $\hat{\eta}$ be the logarithmic posterior mean of expected likelihood in (1) and the likelihood in (2), respectively. Suppose that the specified family of probability distributions does not necessarily contain the true model. Then, under some regularity conditions, the asymptotic bias (3) is given approximately by*

$$\hat{b} \approx \frac{1}{2} \mathrm{tr} \left[\hat{J}^{-1}_{y_n y_n} \{\hat{\theta}(y_n, y_n)\} I_{y_n y_n} \{\hat{\theta}(y_n, y_n)\} \right], \tag{4}$$

where the notation $\approx$ indicates that the difference between the two sides of the equation tends to zero as $n \to \infty$, $\hat{\theta}(y_n, y_n)$ is the mode of $\log\{L(y_n|\theta)L(y_n|\theta)\pi(\theta)\}$, and the $p \times p$ matrices $\hat{I}_{y_n y_n}(\theta)$ and $\hat{J}_{y_n y_n}(\theta)$ are given by

$$\hat{I}_{y_n y_n}(\theta) = \frac{1}{2n} \sum_{\alpha=1}^{n} \left\{ \frac{\partial \log \eta(y_\alpha, y_\alpha|\theta)}{\partial \theta} \frac{\partial \log \eta(y_\alpha, y_\alpha|\theta)}{\partial \theta^{\mathrm{T}}} \right\},$$

$$\hat{J}_{y_n y_n}(\theta) = -\frac{1}{2n} \sum_{\alpha=1}^{n} \left\{ \frac{\partial^2 \log \eta(y_\alpha, y_\alpha|\theta)}{\partial \theta \partial \theta^{\mathrm{T}}} \right\},$$

with $\log \eta(z_\alpha, y_\alpha|\theta) = \log f(z_\alpha|\theta) + \log f(y_\alpha|\theta) + \log \pi(\theta)/n$.

The regularity conditions and derivation are given in Appendix 1.

Correcting the asymptotic bias of $\hat{\eta}$, we propose the predictive marginal likelihood (PML):

$$\begin{aligned} \mathrm{PML} &= -2 \times (\hat{\eta} - \hat{b}) \\ &= -2 \log \int L(y_n|\theta)\pi(\theta|y_n) d\theta + \mathrm{tr} \left[\hat{J}^{-1}_{y_n y_n} \{\hat{\theta}(y_n, y_n)\} I_{y_n y_n} \{\hat{\theta}(y_n, y_n)\} \right]. \end{aligned} \tag{5}$$

We choose the predictive distribution that minimizes PML score.

If we further impose some assumptions on the situation, the bias term (4) reduces to a simple form. The result is given in the following theorem.

Theorem 2 *Additionally to the regularity condition of Theorem 1, we assume that (a) the prior is assumed to be dominated by the likelihood as n increases, say,* $\log \pi(\gamma) = O(1)$, *and (b) the specified parametric models contain the true model, or are similar to the true model. Then the estimated bias term* $\hat{b}$ *in (4) reduceds to* $\hat{b} \approx p/2$, *where p is the dimension of* θ.

The derivation is given in Appendix 2.

In this situation, the predictive marginal likelihood (5) is

$$\text{PML} = -2\log \int L(y_n|\theta)\pi(\theta|y_n)d\theta + p. \tag{6}$$

3. Some observations about PML

3·1. Some regularity conditions

The proposed criterion is available under the situation that (a) the observations are independent, (b) the consistency of θ holds, and (c) the penalised likelihood has a single mode. From (a), the criterion cannot be applied to dependent data, since the derivation of the bias (4) utilizes an asymptotic theory. To ensure the consistency (b), we integrated out the random effects from the likelihood function when we consider the Bayesian model with random effects. The assumption (c) is needed so that the Laplace approximation to be reasonably accurate. From theoretical

perspective, the proposed criterion is not available when the penalised likelihood is characterised by multimodality. We would like to point out that when the Bayesian predictive information criterion (BPIC; Ando, 2006) is applicable, then the proposed criterion is also available. Since the scope of BPIC is less limited than other model selection criteria (Ando, 2006), the proposed criterion can be widely applied.

3·2. Other utility maximization

Under the general situation that the specified family of probability distributions does not contain the true distribution, several Bayesian model selection criteria have been proposed. Konishi & Kitagawa (1996) considered the maximization of the expected log-likelihood of the predictive distribution. Approximating the predictive distribution by using the maximum likelihood estimate, the generalised information criterion is proposed. Unfortunately, as pointed out by Konishi et al. (2004), when we develop various types of nonlinear models such as neural networks and splines, the maximum likelihood method often yields unstable parameter estimates and so leads to large errors in predicting future observations. Ando (2006) considered the maximization of the posterior mean of expected log-likelihood for the evaluation of hierarchical Bayesian and empirical Bayes models. Although these two criteria measure the predictive ability of the model, the utilities to be maximised are different from each other.

3·3. Resampling method

An estimator of the logarithmic posterior mean of expected likelihood, Gelfand et al. (1992) proposed the cross validation predictive density: $\log\{\int \prod_{\alpha=1}^{n}(y_\alpha|\theta)\pi(\theta|y_{-\alpha})d\theta$,

where $y_{-\alpha}$ is the all elements y_n except y_α. Berger & Pericchi (1996) proposed an instric Bayes factor, which was originally introduced to cover the weakness of the original Bayes factor.

Similar to Konishi & Kitagawa (1996) and Ando (2006), the bootstrap method (Efron & Tibshirani, 1993) allows us to perform the bias correction of $\hat{\eta}$. The bootstrap analogues of η and $\hat{\eta}$ are $\eta^{(b)} = \log \int L(y_n|\theta)\pi(\theta|y_n^*)d\theta$ and $\hat{\eta}^{(b)} = \log \int L(y_n^*|\theta)\pi(\theta|y_n^*)d\theta$, where y_n^* is the empirical distribution based on bootstrap samples. The bootstrap bias estimator of b in (3) is then $b^{(b)} = E_{y_n^*}(\eta^{(b)} - \hat{\eta}^{(b)})$. Estimating the asymptotic bias b_γ by the bootstrap bias estimator $b_\gamma^{(b)}$, we can also construct an estimator for η. In practice, the bootstrap bias estimate $b_\gamma^{(b)}$ is approximated by $\hat{b}_\gamma^{(b)}$, which is obtained by Monte Carlo simulation.

4 Normal example

To give insight into the proposed criterion, we first apply the proposed criterion to a simple normal model with known variance. Suppose that a set of n independent observations $y^1, ..., y^n$ are generated from a normal distribution with true mean μ_t and known variance σ^2, i.e. $g(z|\mu_t) = N(\mu_t, \sigma^2)$. We assume the data are generated from a normal distribution $f(z|\mu) = N(\mu, \sigma^2)$. The use of a normal prior $\mu \sim N(\mu_0, \tau_0^2)$ leads to the posterior distribution of μ being normal with mean $\hat{\mu}_n = (\mu_0/\tau_0^2 + \sum_{\alpha=1}^n y^\alpha/\sigma^2)/(1/\tau_0^2 + n/\sigma^2)$ and variance $\sigma_n^2 = 1/(1/\tau_0^2 + n/\sigma^2)$.

The true bias (3) and its estimate $\hat{b}$ are

$$b = \frac{1}{2}E_{y_n}\left[w_n^{-1}\left\{\frac{(\sum_{\alpha=1}^n y^\alpha)^2 - n^2\mu_t^2 - n\sigma^2}{\sigma^4} + 2\hat{\mu}_n \times \frac{\sum_{\alpha=1}^n (y^\alpha - \mu_t)}{\sigma_n^2\sigma^2}\right\} - \sum_{\alpha=1}^n \frac{(y^\alpha)^2 - \mu_t^2 - \sigma^2}{\sigma^2}\right],$$

$$\hat{b} = \mathrm{tr}[\hat{J}^{-1}_{y_n y_n}\{\hat{\mu}(y_n, y_n)\} I_{y_n y_n}\{\hat{\mu}(y_n, y_n)\}]/2,$$

where $w_n^{-1} = n/\sigma^2 + 1/\sigma_n^2$, $\hat{\mu}(y_n, y_n)$ is the mode of $\log\{L(y_n|\mu)L(y_n|\mu)\pi(\mu)\}$, $\hat{I}_{y_n y_n}(\mu) = n^{-1}\sum_{\alpha=1}^{n}\{(y^\alpha - \mu)/\sigma^2 + (\mu_0 - \mu)/(2n\tau_0^2)\}^2$ and $\hat{J}_{y_n y_n}(\mu) = 1/\sigma^2 + 1/(2n\tau_0^2)$.

==========

Insert Figure 1 around here.

==========

Figure 1 shows the true bias and the bias estimate of PML for various sample sizes n. The quantities are evaluated by a Monte Carlo simulation with 100,000 repetitions. We arbitrarily set the true mean, true variance and the prior mean to be $\mu_t = 0$, $\sigma = 0.5$ and $\mu_0 = 0$, respectively. In Figs 1 (a) and (b), the prior variances are set to be $\tau_0 = 0.1$ and $\tau_0 = 100$, corresponding to a rather informative prior and a flat informative prior, respectively. Figure 1 shows that $\hat{\eta}$ has a significant bias as an estimator of η. It can be seen that the estimated asymptotic bias of PML is close to the true bias.

5 Nonlinear regression models

5·1. Preamble

We conduct Monte Carlo experiments to compare the proposed criterion with its competitors: Bayesian predictive information criterion (BPIC; Ando, 2006), deviance information criterion (DIC; Spiegelhalter et al., 2002), harmonic mean Bayes factor (HMBF; Newton & Raftery, 1994), GIC (Konishi & Kitagawa, 1996) and network information criterion (NIC; Muarta et al., 1994), respectively.

§5.2 considers P-spline generalised linear models. A tailor-made version of the proposed criterion is then derived. The resulting formulae are applied in §5.3 to special cases involving Gaussian, logistic and Poisson models. Numerical results are summarised in §5.4.

5·2. P-spline generalised linear models

Suppose that we have n independent observations y_α corresponding to design points x_α, for $\alpha = 1, ..., n$. In generalised linear models (McCullagh & Nelder, 1989), y_α are assumed to be drawn from the exponential family of distributions with density $f(y_\alpha|x_\alpha; \xi_\alpha, \phi) = \exp[\{y_\alpha\xi_\alpha - u(\xi_\alpha)\}/\phi + v(y_\alpha, \phi)]$, where $u(\cdot)$ and $v(\cdot, \cdot)$ are functions specific to each distribution, and ϕ is an unknown scale parameter. The conditional expectation $E(y_\alpha|x_\alpha) = \mu_\alpha = u'(\xi_\alpha)$ is linked to a predictor $\eta_\alpha = h(\mu_\alpha)$, where $h(\cdot)$ is a link function. In this paper, we use the B-spline function for the predictor $\eta_\alpha = \sum_{j=1}^{m} w_j b_j(x_\alpha)$ (Eilers & Marx, 1996).

Then it follows from the density and the predictor that the data are summarised by a model from a class of probability densities of the form $f(y_\alpha|x_\alpha; \theta) = \exp([y_\alpha r\{w^{\mathrm{T}}b(x_\alpha)\} - s\{w^{\mathrm{T}}b(x_\alpha)\}]/\phi + v(y_\alpha, \phi))$, where $\theta = (w^{\mathrm{T}}, \phi)^{\mathrm{T}}$, $w = (w_1, \cdots, w_m)^{\mathrm{T}}$ is the m-dimensional coefficient vector, $b(x) = (b_1(x), \cdots, b_m(x))^{\mathrm{T}}$ is the m-dimensional basis function vector, $r(\cdot) = u'^{-1} \circ h^{-1}(\cdot)$ and $s(\cdot) = u \circ u'^{-1} \circ h^{-1}(\cdot)$.

For posterior inference, we shall use a singular multivariate normal prior density (Konishi et al., 2004) $\pi(\theta) = \{n\lambda/(2\pi)\}^{(m-2)/2}|R|_+^{1/2} \exp\{-n\lambda\theta^{\mathrm{T}}R\theta/2\}$, where λ is a smoothing parameter, m is the number of basis functions, $R = \mathrm{diag}\{D, 0\}$ is a block diagonal matrix and $|R|_+$ is the product of $(m-2)$ nonzero eigenvalues of R.

The remaining problem is how to choose the smoothing parameter λ and the number of basis functions m. We use the proposed criterion (5) to choose appropriate

values for these parameters. The result is summarised in the following theorem.

Theorem 3 *Let $f(y_\alpha|x_\alpha;\theta)$ be P-spline generalised linear model, to be estimated by the Bayesian approach. Then the $(m+1)\times(m+1)$ matrices $I_{y_n y_n}\{\hat{\theta}(y_n,y_n)\}$ and $J_{y_n y_n}\{\hat{\theta}(y_n,y_n)\}$ in the predictive marginal likelihood (5) are given by*

$$\hat{I}_{y_n y_n}\{\hat{\theta}_n(y_n,y_n)\} = \frac{1}{n}\begin{pmatrix} B^{\mathrm{T}}\Lambda/\hat{\phi}_n - \lambda D\hat{w}_n 1_n^{\mathrm{T}}/2 \\ p^{\mathrm{T}} \end{pmatrix}\left(\Lambda B/\hat{\phi}_n - \lambda 1_n^{\mathrm{T}}\hat{w}_n D/2, p\right),$$

$$\hat{J}_{y_n y_n}\{\hat{\theta}_n(y_n,y_n)\} = \frac{1}{n}\begin{pmatrix} B^{\mathrm{T}}\Gamma/\hat{\phi}_n + n\lambda D/2 & B^{\mathrm{T}}\Lambda 1_n/\hat{\phi}_n^2 \\ 1_n^{\mathrm{T}}\Lambda B/\hat{\phi}_n^2 & -q^{\mathrm{T}}1_n \end{pmatrix}.$$

Here $B=(b(x_1),...,b(x_n))^{\mathrm{T}}$, $1_n=(1,...,1)^{\mathrm{T}}$, Λ and Γ are $n\times n$ diagonal matrices and p and q are n-dimensional vectors with α th diagonal elements and α th elements

$$\Lambda_{\alpha\alpha} = \frac{y_\alpha-\hat{\mu}_\alpha}{u''(\hat{\xi}_\alpha)h'(\hat{\mu}_\alpha)},\ p_\alpha = -\frac{y_\alpha r\{\hat{w}_n^{\mathrm{T}}b(x_\alpha)\}-s\{\hat{w}_n^{\mathrm{T}}b(x_\alpha)\}}{\hat{\phi}_n^2}+\frac{\partial}{\partial\phi}v(y_\alpha,\phi)\Big|_{\phi=\hat{\phi}_n},$$

$$\Gamma_{\alpha\alpha} = \frac{(y_\alpha-\hat{\mu}_\alpha)\{u'''(\hat{\xi}_\alpha)h'(\hat{\mu}_\alpha)+u''(\hat{\xi}_\alpha)^2h''(\hat{\mu}_\alpha)\}}{\{u''(\hat{\xi}_\alpha)h'(\hat{\mu}_\alpha)\}^3}+\frac{1}{u''(\hat{\xi}_\alpha)h'(\hat{\mu}_\alpha)^2},\ q_\alpha=\frac{\partial p_\alpha}{\partial\phi}\Big|_{\phi=\hat{\phi}_n}.$$

Substituting the density functions $f(y_\alpha|x_\alpha;\theta)$ and $\pi(\theta)$ into the equation (5), the bias term can be derived. The adjusted parameters λ and m are determined as the minimisers of the predictive marginal likelihood in (5).

5·3. Some special cases

Example 1. *P-spline Gaussian regression model.* Consider P-spline Gaussian regression model $y_\alpha = w^{\mathrm{T}}b(x_\alpha)+\varepsilon_\alpha$ $(\alpha=1,...,n)$, where the errors ε_α are independently and normally distributed with mean zero and variance σ^2. Estimating the parameter vector θ by producing the posterior samples, the predictive distribution is obtained. In this case, taking

$$u(\xi_\alpha)=\xi_\alpha^2/2,\ \phi=\sigma^2,\ v(y_\alpha,\phi)=-y_\alpha^2/(2\sigma^2),\ -\log\left(\sigma\sqrt{2\pi}\right)\ \text{ and }\ h(\mu_\alpha)=\mu_\alpha$$

in Theorem 3, the bias term of the predictive marginal likelihood is derived.

Example 2. *P-spline logistic regression model.* Suppose that we have n independent binary observations y_α, each from a logistic distribution with conditional expectation $\mathrm{pr}(Y_\alpha = 1|x_\alpha) = \pi(x_\alpha)$. It is assumed that the probability $\pi(x_\alpha)$ is of the form: $\log[\pi(x_\alpha)/\{1-\pi(x_\alpha)\}] = w^{\mathrm{T}}b(x_\alpha)$. Posterior inference can be done by producing the posterior samples. Then taking

$$u(\xi_\alpha) = \log\{1+\exp(\xi_\alpha)\},\ v(y_\alpha,\psi) = 0,\ h(\mu_\alpha) = \log\frac{\mu_\alpha}{1-\mu_\alpha} \text{ and } \phi = 1$$

in Theorem 3, we derive the predictive marginal likelihood for evaluating the predictive distribution.

Example 3. *P-spline Poisson regression model.* Let y_α, $\alpha = 1,...,n$ be independent observations, each from a Poisson distribution with conditional expectation $E(Y_\alpha|x_\alpha) = \gamma(x_\alpha)$. Then the conditional expectation is expressed as $\log\{\gamma(x_\alpha)\} = w^{\mathrm{T}}b(x_\alpha)$, $\alpha = 1,2,...,n$. We estimate the unknown parameter vector w by producing the posterior samples. The predictive marginal likelihood for eavluating the constructed model can be obtained by taking $u(\xi_\alpha) = \exp(\xi_\alpha)$, $\phi = 1$, $v(y_\alpha,\phi) = -\log(y_\alpha!)$ and $h(\mu_\alpha) = \log(\mu_\alpha)$ in Theorem 3.

5·4. Results

As an Gaussian example, datasets $\{(y_\alpha, x_\alpha); \alpha = 1,...,n\}$ are repeatedly generated from the true regression model $y_\alpha = \sin(4\pi x_\alpha) + \varepsilon_\alpha$ for $x_\alpha = (2\alpha-1)/(2n)$. The errors ε_α are independently and identicaly distributed according to a mixture of normal distributions $g(\varepsilon_\alpha) = \beta N(\varepsilon_\alpha|0,\sigma_1^2) + (1-\beta)N(\varepsilon_\alpha|0,\sigma_2^2)$, where β is a mixing proportion, and $N(\varepsilon|\mu,\sigma^2)$ is the normal density function with mean μ and variance σ^2. The values of the mixing proportion and sample variances are set to be $\beta = 0.8$,

$\sigma_1 = 0.25$ and $\sigma_2 = 0.5$, respectively. To fit the logistic model, we generated a set of n observations for according to $\mathrm{pr}(Y_\alpha = 1|x_\alpha) = 1/\left[1 + \exp\left\{3\sin(\pi x_\alpha)\right\}\right]$ for $x_\alpha = (2\alpha - 1)/(2n)$.

The total number of Markov chain Monte Carlo iterations is chosen to be 11,000. The first 1,000 iterations are discarded. To save computational time, the initial value of the parameter is chosen to be the posterior mode.

==========

Insert Tables 1 and 2 around here.

==========

Tables 1 and 2 compare the average squared error between the true and estimated conditional expectations: $\mathrm{ASE} = \sum_{\alpha=1}^{n} \left\{E(Y_\alpha|x_\alpha) - \hat{y}(x_\alpha)\right\}^2 /n$. The means and standard deviations of the selected smoothing parameter λ and the number of basis functions m are also given. The values in parentheses indicate standard deviations for the means. The value of sample sizes is set to be $n \in \{100,\ 200\}$. The candidates for the smoothing parameter were chosen on an evenly spaced grid of 10 values between $\log_{10}(\lambda) = 0$ and $\log_{10}(\lambda) = -9$. The number of basis functions ranges from 6 to 10. The simulation results were obtained from 100 repeated Monte Carlo trials.

It may be seen from the simulation results that BPIC achieved the best performance in almost all cases. However, Gaussian models evaluated by PML are superior to those based on other criteria in $n = 100$; they give smaller mean values with smaller variances for ASE. When $n = 200$, the performances of Gaussian model evaluated by PML and BPIC are almost the same. The mean values of the smoothing parameter chosen by DIC and PBF were smaller than those based on

other criteria. PML tends to choose fewer basis functions and larger values of λ than those based on DIC and PBF. It indicates that DIC and PBF are generally more variable and more likely to undersmooth than PML. ASE indicates that the model selected by DIC and PBF overfits to the observed data.

6 Stochastic volatility models

6·1. Preamble

As a hierarchical Bayes example, stochastic volatility model selection problem is considered. We fit six different stochastic volatility models to the simulated data including the true model from which the data are generated. An objective is to investigate whether the proposed criterion is capable of identifying the true model from which the data are generated.

For each model, §6.2 describes observation and state equations, their distributional assumptions and the prior distributions for the unknown parameters. §6.3 summarises the results.

6·2. Models

Model 1 is the basic stochastic volatility model: $y_t = \exp(h_t/2)u_t$, $h_t = \mu + \phi(h_{t-1} - \mu) + \tau v_t$, where $\theta = (\mu, \phi, \tau^2)^{\mathrm{T}}$, h_t is an unobserved log-volatility of y_t and $u_t \sim N(0,1)$ and $v_t \sim N(0,1)$ are uncorrelated Gaussian white noise sequences. Following Kim et al. (1998), we assume that each parameter is a priori independent $\pi(\theta) = \pi(\mu)\pi(\phi)\pi(\tau^2)$ and use the same prior specifications of Kim et al. (1998).

For the prior densities of $(\phi+1)/2$, τ^2 and μ, a beta distribution $Be(20, 1.5)$, an inverse-gamma distribution $IG(2.5, 0.025)$ and a normal distribution $N(-5, 5^2)$ are utilised.

Model 2 utilises AR(2) structure for the state transitions: $h_t = \mu + \phi(h_{t-1} - \mu) + \psi(\theta_{t-2} - \mu) + \tau v_t,\ v_t \sim N(0,1)$ The observation equations are equal to the basic model. We use the same prior for ϕ, μ, τ^2 as for the basic stochastic volatility model and center the prior for ψ around zero using a uniform distribution $U[-1, 1]$.

Model 3 is equivalent to the basic stochastic volatility model including a leverage or asymmetric effect by allowing for correlation ρ between u_t and v_{t+1}. Following Berg et al. (2004), we specify a uniform prior distribution $U[-1, 1]$ for ρ.

In Model 4, the normal distribution of u_t in the observation equation is replaced by independent central Student-t distributions with ν degrees of freedom $St(\nu)$: $y_t = \exp(\theta_t/2)u_t,\ u_t \sim St(\nu)$. We use the same prior for ϕ, μ, τ^2 as for Model 1 and use the uniform prior distribution $U[2, 100]$ for ν.

Model 5 is equivalent to Model 4 including a leverage effect by allowing for correlation between u_t and v_{t+1}. We specify a uniform prior distribution $U[-1, 1]$ for ρ.

Model 6 is similar to the basic stochastic volatility model except that it contains a jump component in the observation equation to allow for large movements: $y_t = s_t q_t + \exp(\theta_t/2)u_t,\ u_t \sim N(0,1)$, where q_t follows a Bernoulli distribution which takes the value one with unknown probability κ and the time-varying variable s_t represents the size of the jump when a jump occurs. For the parameters, we follow the prior specifications of Chib et al. (2002).

6·3. Results

In the simulation design, Model 4 and Mode 5 are employed for the true model. Datasets are generated from the true model with parameter values $\phi = 0.8$, $\mu = -8.0$, $\tau = 0.2$, $\rho = -0.4$ and $\nu = 10$, respectively. We simulate 50 data series of $n = 800$ observations. In our application, the total number of MCMC iterations is chosen to be 1,000,000 in which the first 100,000 iterations are discarded as a burn-in period sample. After a burn-in period, we stored every 1,000th posterior sample.

Table 3 reports the model selection results obtained from 50 repeated Monte Carlo trials. Since the asymptotic bias of PML in (5) is not be available in closed form and requires numerical integration and derivatives, the PML in (6) is utilised. To compute the Bayes factor, we utilise Chib (1995)'s marginal likelihood method (Chib's BF).

As shown in Table 3, the proposed criterion selects the correct model 90% of the times against other models when the data is generated from Model 4. It may be seen from Table 3 that the proposed criterion is superior to PBF and DIC; it chooses the correct model frequently than DIC and PBF. Since DIC and PBF provide much less penalty for model complexity than that of BPIC and PML, the best models chosen by DIC and PBF are relatively complex than those of BPIC and PML. On the other hand, the Bayes factor tends to choose the simpler models than those selected by other criteria. In conclusion, PML relatively performs well in the full Bayesian approach.

==========

Insert Table 3 around here.

==========

Acknowledgement

This study was supported in part by a Grant-in-Aid for Young Scientists (B), 18700273.

Appendix 1

Proof of Theorem 1

We first describe some asymptotic properties of the parameter estimators. Let θ_0 and $\hat{\theta}(z_n, y_n)$ be the modes of $E_{z_n,y_n}\{L(z_n, y_n|\theta)\pi(\theta)\}$ and $L(y_n, z_n|\theta)\pi(\theta)$, respectively. For a simplicity of notation, we used $L(z_n, y_n|\theta) = L(z_n|\theta)L(y_n|\theta)$. Since $\log L(y_n, z_n|\theta)$ is the sum of the independently and identically distributed random variables, it follows from the law of large numbers that $\log\{L(y_n, z_n|\theta)\pi(\theta)\} \to E_{z_n,y_n}[\log\{L(z_n, y_n|\theta)\pi(\theta)\}]$ as n tends to infinity. Then $\hat{\theta}(z_n, y_n) \to \theta_0$ in probability as n tends to infinity. Hereafter, we restrict our attention to a proper situation in which the Hessian of $E_{z_n,y_n}[\log\{L(z_n, y_n|\theta)\pi(\theta)\}]$ is nonsingular at θ_0, which is uniquely determined and interior to Θ.

Lemma A1. *Assume regularity conditions similar to those of Ando (2006); i.e., the model is sufficiently smooth and the Hessian of* $E_{z_n,y_n}[\log\{L(z_n, y_n|\theta)\pi(\theta)\}]$ *is non-singular at* θ_0. *Then* $\sqrt{2n}\{\hat{\theta}(z_n, y_n) - \theta_0\}$ *is asymptotically normally distributed as* $N\{0, J^{-1}_{z_n y_n}(\theta_0) I_{z_n y_n}(\theta_0) J^{-1}_{z_n y_n}(\theta_0)\}$. *Here* $\hat{\theta}(z_n, y_n)$ *is the mode of* $L(z_n, y_n|\theta)\pi(\theta)$ *and* $I_{z_n y_n}(\theta)$ *and* $J_{z_n y_n}(\theta)$ *are the* $p \times p$ *matrices respectively defined by*

$$I_{z_n y_n}(\theta) = \frac{1}{2n} E_{z_n,y_n}\left[\sum_{\alpha=1}^{n} \frac{\partial \log\{\eta(z_\alpha, y_\alpha|\theta)\}}{\partial\theta} \frac{\partial \log\{\eta(z_\alpha, y_\alpha|\theta)\}}{\partial\theta^{\mathrm{T}}}\right],$$

$$J_{z_n y_n}(\theta) = -\frac{1}{2n} E_{z_n,y_n}\left[\sum_{\alpha=1}^{n} \frac{\partial^2 \log\{\eta(z_\alpha, y_\alpha|\theta)\}}{\partial\theta\partial\theta^{\mathrm{T}}}\right].$$

Proof of Lemma A1. Since $\hat{\theta}(z_n, y_n)$ is the mode of $L(z_n, y_n|\theta)\pi(\theta)$, it satisfies the score equation $\partial[\log\{L(z_n, y_n|\theta)\pi(\theta)\}]/\partial\theta|_{\theta=\hat{\theta}(z_n,y_n)} = 0$. Taylor expansion leads to

$$-\frac{1}{2n}\frac{\partial^2 \log\{L(z_n, y_n|\theta)\pi(\theta)\}}{\partial\theta\partial\theta^{\mathrm{T}}}\bigg|_{\theta=\theta_0}\sqrt{2n}\{\hat{\theta}(z_n, y_n) - \theta_0\}$$
$$= \frac{1}{\sqrt{2n}}\frac{\partial \log\{L(z_n, y_n|\theta)\pi(\theta)\}}{\partial\theta}\bigg|_{\theta=\theta_0} + O_p\left(\frac{1}{\sqrt{2n}}\right).$$

It follows from the central limit theorem that the right-hand side is asymptotically distributed as $N\{0, I_{z_n y_n}(\theta_0)\}$, while the left-hand side converges to $J_{z_n y_n}(\theta_0)\sqrt{2n}\{\hat{\theta}(z_n, y_n) - \theta_0\}$. Thus we obtained the desired result. In the same way, we can proof that $\sqrt{n}\{\hat{\theta}(y_n, y_n) - \theta_0\}$ is asymptotically normally distributed as $N\{0, J^{-1}_{y_n y_n}(\theta_0)I_{y_n y_n}(\theta_0)J^{-1}_{y_n y_n}(\theta_0)\}$. Here $\hat{\theta}(y_n, y_n)$ is the mode of $L(y_n, y_n|\theta)\pi(\theta)$.

Proof of Theorem 1. We assume the regularity conditions of Lemma A1 and the Laplace approximation. For the regularity conditions of the Laplace approximation, we refer to Barndorff-Nielsen & Cox (1989). Using the basic Laplace approximation and a ratio of integrals (Gelfand & Day, 1994), we have

$$m(z_n|y_n) = \int L(z_n|\theta)\pi(\theta|y_n)d\theta = \frac{\int L(z_n, y_n|\theta)\pi(\theta)d\theta}{\int L(y_n|\theta)\pi(\theta)d\theta}$$
$$= \frac{L\{z_n, y_n|\hat{\theta}(z_n, y_n)\}\pi\{\hat{\theta}(z_n, y_n)\}}{L\{y_n|\hat{\theta}(y_n)\}\pi\{\hat{\theta}(y_n)\}}\left[\frac{|\hat{J}^{-1}_{z_n y_n}\{\hat{\theta}(z_n, y_n)\}|}{|\hat{J}^{-1}_{y_n}(\hat{\theta}(y_n))|}\right]^{1/2} + O(n^{-2}),$$
$$m(y_n|y_n) = \int L(y_n|\theta)\pi(\theta|y_n)d\theta = \frac{\int L(y_n, y_n|\theta)\pi(\theta)d\theta}{\int L(y_n|\theta)\pi(\theta)d\theta}$$
$$= \frac{L\{y_n, y_n|\hat{\theta}(y_n, y_n)\}\pi\{\hat{\theta}(y_n, y_n)\}}{L\{y_n|\hat{\theta}(y_n)\}\pi\{\hat{\theta}(y_n)\}}\left[\frac{|\hat{J}^{-1}_{y_n y_n}\{\hat{\theta}(y_n, y_n)\}|}{|\hat{J}^{-1}_{y_n}(\hat{\theta}(y_n))|}\right]^{1/2} + O(n^{-2}),$$

where $\hat{\theta}(z_n, y_n)$, $\hat{\theta}(y_n, y_n)$ and $\hat{\theta}(y_n)$ are the modes of $L(z_n, y_n|\theta)\pi(\theta)$, $L(y_n, y_n|\theta)\pi(\theta)$ and $L(y_n|\theta)\pi(\theta)$, respectively. Here $J_{y_n}(\theta) = -n^{-1}E_{y_n}[\partial^2\{\log L(y_n|\theta)\pi(\theta)\}/\partial\theta\partial\theta^{\mathrm{T}}]$, and $\hat{J}_{y_n y_n}$, $\hat{J}_{z_n y_n}$ and $\hat{J}_{y_n}$ are $p \times p$ matrices obtained by replacing the expectation operator by the empirical distribution.

Using the above expression, we decompose the bias in (3) as

$$E_{y_n}(\hat{\eta} - \eta) = B_1 + B_2 + B_3 + B_4 + O(n^{-2}),$$

where

$$\begin{aligned}
B_1 &= E_{y_n}(\log[L\{y_n, y_n|\hat{\theta}(y_n, y_n)\}\pi\{\hat{\theta}(y_n, y_n)\}]) - E_{y_n}[\log\{L(y_n, y_n|\theta_0)\pi(\theta_0)\}], \\
B_2 &= E_{y_n}[\log\{L(y_n, y_n|\theta_0)\pi(\theta_0)\}] - E_{z_n,y_n}[\log\{L(z_n, y_n|\theta_0)\pi(\theta_0)\}], \\
B_3 &= E_{z_n,y_n}[\log\{L(z_n, y_n|\theta_0)\pi(\theta_0)\}] - E_{z_n,y_n}(\log[L\{z_n, y_n|\hat{\theta}(z_n, y_n)\}\pi\{\hat{\theta}(z_n, y_n)\}]), \\
B_4 &= E_{y_n}\left[\frac{1}{2}\log\left|\hat{J}^{-1}_{y_n,y_n}\{\hat{\theta}(y_n, y_n)\}\right|\right] - E_{z_n,y_n}\left[\frac{1}{2}\log\left|\hat{J}^{-1}_{z_n,y_n}\{\hat{\theta}(z_n, y_n)\}\right|\right].
\end{aligned}$$

For the evaluation of B_1, considering $\partial \log[L\{y_n, y_n|\hat{\theta}(y_n, y_n)\}\pi\{\hat{\theta}(y_n, y_n)\}]/\partial\theta = 0$, the Taylor expansion of $\log\{L(y_n, y_n|\theta_0)\pi(\theta_0)\}$ around the posterior mode $\hat{\theta}(y_n, y_n)$ gives

$$\begin{aligned}
\log\{L(y_n, y_n|\theta_0)\pi(\theta_0)\} &= \log[L\{y_n, y_n|\hat{\theta}(y_n, y_n)\}\pi\{\hat{\theta}(y_n, y_n)\}] \\
&\quad -2n\{\theta_0 - \hat{\theta}(y_n, y_n)\}^{\mathrm{T}}\hat{J}_{y_n,y_n}\{\hat{\theta}(y_n, y_n)\}\{\theta_0 - \hat{\theta}(y_n, y_n)\}/2 + O_p(1/\sqrt{n}).
\end{aligned}$$

Thus $B_1 = \mathrm{tr}(E_{y_n}[\hat{J}_{y_n,y_n}\{\hat{\theta}(y_n, y_n)\}\sqrt{n}\{\theta_0 - \hat{\theta}(y_n, y_n)\}\sqrt{n}\{\theta_0 - \hat{\theta}(y_n, y_n)\}^{\mathrm{T}}]) + O(1/\sqrt{n})$. From Lemma, the variance matrix of $\sqrt{n}\{\theta_0 - \hat{\theta}(y_n, y_n)\}$ is asymptotically given by $J^{-1}_{y_n y_n}(\theta_0)I_{y_n y_n}(\theta_0)J^{-1}_{y_n y_n}(\theta_0)$. With this result and since $\hat{J}_{y_n y_n}\{\hat{\theta}(y_n, y_n)\} \to J_{y_n y_n}(\theta_0)$ in probability as $n \to \infty$, we have $B_1 = \mathrm{tr}\{J^{-1}_{y_n y_n}(\theta_0)I_{y_n y_n}(\theta_0)\}$.

Using the Taylor expansion of $\log\{L(z_n, y_n|\theta_0)\pi(\theta_0)\}$ around $\hat{\theta}(z_n, y_n)$, we have

$$\begin{aligned}
\log\{L(z_n, y_n|\theta_0)\pi(\theta_0)\} &= \log[L\{z_n, y_n|\hat{\theta}(z_n, y_n)\}\pi\{\hat{\theta}(z_n, y_n)\}] \\
&\quad -2n\{\theta_0 - \hat{\theta}(z_n, y_n)\}^{\mathrm{T}}\hat{J}_{z_n,y_n}\{\hat{\theta}(z_n, y_n)\}\{\theta_0 - \hat{\theta}(z_n, y_n)\}/2 + O_p(1/\sqrt{2n}).
\end{aligned}$$

The term B_3 is then $B_3 = -E_{z_n,y_n}[\sqrt{2n}\{\hat{\theta}(z_n, y_n) - \theta_0\}^{\mathrm{T}}\hat{J}_{z_n,y_n}(\theta_0)\sqrt{2n}\{\hat{\theta}(z_n, y_n) - \theta_0)\}]/2 + O(1/\sqrt{2n})$. Considering Lemma A1, we have $B_3 = -\mathrm{tr}\{J^{-1}_{z_n y_n}(\theta_0)I_{z_n y_n}(\theta_0)\}/2 +$

$O_p(1/\sqrt{2n})$. The terms B_2 and B_4 are zero, because

$$E_{y_n}[\log\{L(y_n, y_n|\theta_0)\pi(\theta_0)\}] = E_{z_n,y_n}[\log\{L(z_n, y_n|\theta_0)\pi(\theta_0)\}],$$

$$E_{y_n}\left[\log|\hat{J}^{-1}_{y_n y_n}\{\hat{\theta}(y_n, y_n)\}|\right] = E_{z_n,y_n}\left[\log|\hat{J}^{-1}_{z_n y_n}\{\hat{\theta}(z_n, y_n)\}|\right].$$

When the above results are combined, the asymptotic bias is given by $E_{y_n}(\hat{\eta} - \eta) = \mathrm{tr}\{J^{-1}_{y_n y_n}(\theta_0) I_{y_n y_n}(\theta_0)\} - \mathrm{tr}\{J^{-1}_{z_n y_n}(\theta_0) I_{z_n y_n}(\theta_0)\}/2$. Estimating the matrices $I_{z_n y_n}(\theta_0)$ and $I_{y_n y_n}(\theta_0)$ by $\hat{I}_{y_n y_n}\{\hat{\theta}(y_n, y_n)\}$ and $J_{z_n y_n}(\theta_0)$ and $J_{y_n y_n}(\theta_0)$ by $\hat{J}_{y_n y_n}\{\hat{\theta}(y_n, y_n)\}$, we obtain the required result.

Appendix 2

Proof of Theorem 2

Under the situation $\log \pi(\gamma) = O(1)$, the quantity $\log \eta(z_\alpha, y_\alpha|\theta)$ in the $p \times p$ matrices $I_{z_n y_n}(\theta)$ and $J_{z_n y_n}(\theta)$ reduces to $\log \eta(z_\alpha, y_\alpha|\theta) = \log f(z_\alpha|\theta) + \log f(y_\alpha|\theta)$ as $n \to \infty$. Since the specified parametric models contain the true model, or are similar to the true model, it can be shown that $I_{z_n y_n}(\theta) \simeq J_{z_n y_n}(\theta)$. Therefore, the bias $\hat{b}$ in (4) reduces to a half dimension of θ.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Auto. Contr.* **19**, 716–23.

Ando, T. (2006). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, in press.

Aitkin, M. (1991). Posterior Bayes Factor (with Discussion). *J. R. Statist. Soc.* B **53**, 111–42.

BARNDORFF-NIELSEN, O. E. & COX, D. R. (1989). *Asymptotic Techniques for Use in Statistics.* London: Chapman and Hall.

BERG, A., MEYER, R. & YU, J. (2004). Deviance information criterion comparing stochastic volatility models. *J. Bus. Econom. Statist.* **22**, 107–20.

BERGER, J. O. & PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Am. Statist. Assoc.* **91**, 109–22

CHIB, S. (1995). Marginal Likelihood from the Gibbs Output. *J. Am. Statist. Assoc.* **90**, 1313-21.

CHIB, S., NARDARI, F. & SHEPHARD, N. (2002). Markov Chain Monte Carlo Methods for Stochastic Volatility Models. *J. Econometrics* **108**, 281–316.

EFRON, B. & TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

EILERS, P. H. C. & MARX, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statist. Sci.* **11**, 89–121.

GELFAND, A. E. & DEY, D. K. (1994). Bayesian model choice: asymptotic and exact calculations. *J. R. Statist. Soc.* B **56**, 501–14.

GELFAND, A. E., DEY, D. K. & CHANG, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with Discussion). In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 147–67. Oxford: Oxford University Press.

KADANE, J. B. & LAZAR, N. A. (2004). Methods and Criteria for Model Selection. *J. Am. Statist. Assoc.* **99**, 279–90.

KASS, R. & RAFTERY, A. (1995). Bayes factors and model uncertainty. *J.*

Am. Statist. Assoc. **90**, 773–95.

KIM, S., SHEPHARD, N. & CHIB, S. (1998). Stochastic volatility: likelihood inference comparison with ARCH models. *Rev. Econom. Stud.* **65**, 361–93.

KONISHI, S. & KITAGAWA, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–90.

KONISHI, S., ANDO, T. & IMOTO, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91**, 27–43.

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

MURATA, N., YOSHIZAWA, S. & AMARI, S. (1994). Network information criterion determining the number of hidden units for an artificial neural network model. *IEEE Trans. Neural Networks* **5**, 865–72.

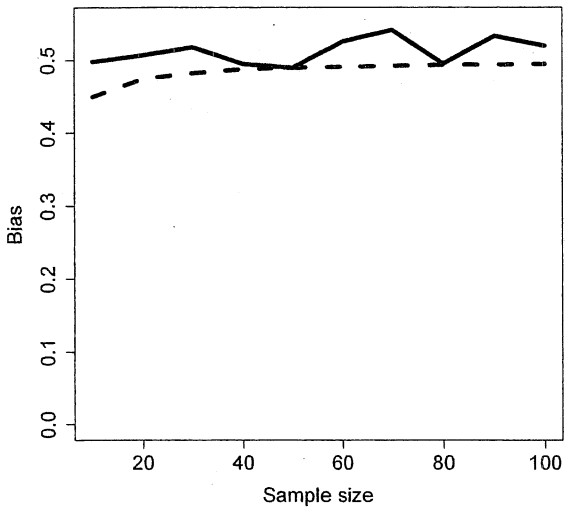
NEWTON, M. A. & RAFTERY, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with Discussion). *J. R. Statist. Soc.* B **56**, 3–48.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison (with Discussion). *J. R. Statist. Soc.* B **57**, 99–138.

PEREZ, J. M. & BERGER, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika* **89**, 491–512.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with Discussion). *J. R. Statist. Soc.* B **64**, 583–639.

(a)



(b)

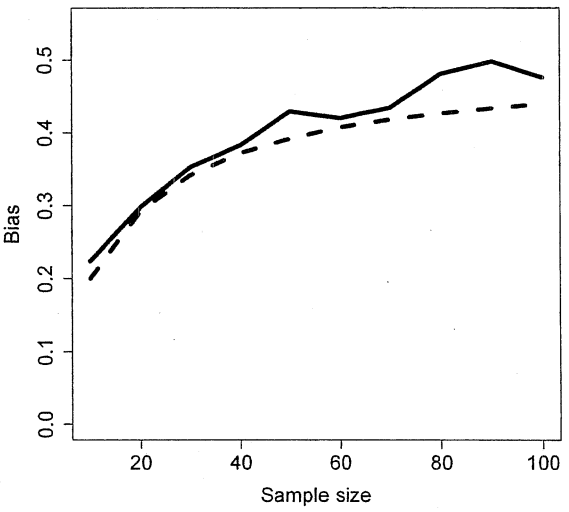


Figure 1: Simple normal example. Comparison of the true bias b (—), the estimated asymptotic bias BPIC (- - -) for various sample sizes. (a) under the rather informative prior with $\tau_0 = 0.1$ and (b) under a flat informative prior with $\tau_0 = 100$.

Table 1: Comparison of the average squared errors based on various criteria. Averages are given and figures in parentheses give estimated standard deviations. The results are for Gaussian example.

	$n = 100$			$n = 200$		
	m	$\log_{10}(\lambda)$	ASE	m	$\log_{10}(\lambda)$	ASE
PML	8.98	-4.15	0.01971	9.06	-4.26	0.0094
	(0.94)	(0.70)	(0.0092)	(0.82)	(0.68)	(0.0043)
PBF	9.64	-4.91	0.02057	9.64	-4.84	0.0100
	(0.64)	(0.28)	(0.0094)	(0.60)	(0.36)	(0.0045)
BPIC	8.07	-3.93	0.01989	8.36	-3.92	0.0093
	(1.04)	(0.68)	(0.0095)	(0.90)	(0.55)	(0.0045)
DIC	9.65	-4.81	0.02054	9.63	-4.75	0.0098
	(0.62)	(0.39)	(0.0094)	(0.60)	(0.43)	(0.0048)
HMBF	9.63	-4.86	0.02056	9.64	-4.80	0.0099
	(0.64)	(0.34)	(0.0094)	(0.78)	(0.39)	(0.0045)
GIC	8.36	—	0.02066	8.53	—	0.0102
	(1.00)	(—)	(0.0090)	(0.85)	(—)	(0.0046)
NIC	8.60	-4.91	0.02001	8.66	-5.00	0.0098
	(8.60)	(0.29)	(0.0094)	(0.29)	(2.276)	(0.0043)

ASE, average squared error

Table 2: Comparison of the average squared errors based on various criteria. Averages are given and figures in parentheses give estimated standard deviations. The fitting results are for logistic model.

	$n = 100$			$n = 200$		
	m	$\log_{10}(\lambda)$	ASE	m	$\log_{10}(\lambda)$	ASE
PML	8.65	-5.39	0.01423	8.09	-5.04	0.0060
	(1.39)	(0.96)	(0.0068)	(1.69)	(1.41)	(0.0045)
PBF	9.67	-6.13	0.01630	9.77	-8.52	0.0079
	(0.59)	(0.61)	(0.0068)	(0.44)	(0.93)	(0.0039)
BPIC	6.97	-4.86	0.01198	6.27	-4.09	0.0039
	(1.34)	(1.20)	(0.0070)	(0.84)	(0.85)	(0.0035)
DIC	9.67	-6.11	0.01629	9.78	-8.13	0.0079
	(0.59)	(0.55)	(0.0068)	(0.44)	(1.02)	(0.0040)
HMBF	9.67	-6.13	0.01630	9.77	-8.53	0.0079
	(1.23)	(0.61)	(0.0068)	(0.45)	(0.92)	(0.0039)
GIC	8.12	—	0.01668	6.89	—	0.0066
	(1.52)	(—)	(0.0079)	(1.38)	(—)	(0.0043)
NIC	8.43	-6.13	0.01507	7.81	-8.10	0.0063
	(1.51)	(0.62)	(0.0072)	(1.35)	(1.06)	(0.0042)

ASE, average squared error

Table 3: Frequency distribution of selected models across 50 simulated replications. The datasets are generated from Model 4 and Model 5.

True model; Model 4

Models	1	2	3	4	5	6
PML	3	2	0	44	1	0
PBF	2	1	0	43	4	0
BPIC	4	1	0	44	1	0
DIC	1	2	0	42	3	2
Chib's BF	5	0	0	45	0	0

True model; Model 5

Models	1	2	3	4	5	6
PML	4	0	3	3	40	0
PBF	2	1	3	4	40	0
BPIC	6	1	0	4	39	0
DIC	0	0	2	5	40	3
Chib's BF	5	0	0	8	37	0