Title	Predictive Bayesian model selection
Sub Title	
Author	安道, 知寛(Ando, Tomohiro)
Publisher	慶應義塾経営管理学会
Publication year	2009
Jtitle	慶應義塾経営管理学会リサーチペーパー・シリーズ No.99 (2009. 7)
JaLC DOI	
Abstract	The problem of evaluating the goodness of the predictive distributions of Bayesian models is investigated. To evaluate the Bayesian model, deviance information criteria (DIC) has been extensively employed in various study areas, thanks to its simplicity of calculation from the posterior output. Unfortunately, it is also true that the DIC has been criticized due to the over fitting. Inheriting the simplicity form of DIC, we propose a new criterion that overcomes the over fitting problem. 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 proposed criteria are robust to any improper priors. Monte Carlo simulations are conducted to investigate the properties of the proposed criteria. We also show that the proposed criteria can avoid over fitting problem.
Notes	
Genre	Technical Report
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40003002-00000099-0001

Predictive Bayesian Model Selection

安道　知寛
Tomohiro Ando

慶應義塾大学経営管理研究科准教授

慶應義塾経営管理学会
リサーチペーパー・シリーズ
No.99 （2009年7月）

Predictive Bayesian Model Selection

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. To evaluate the Bayesian model, deviance information criteria (DIC) has been extensively employed in various study areas, thanks to its simplicity of calculation from the posterior output. Unfortunately, it is also true that the DIC has been criticized due to the over fitting. Inheriting the simplicity form of DIC, we propose a new criterion that overcomes the over fitting problem. 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 proposed criteria are robust to any improper priors. Monte Carlo simulations are conducted to investigate the properties of the proposed criteria. We also show that the proposed criteria can avoid over fitting problem.

Some key words: Effective number of parameters; Empirical Bayes; Markov chain Monte Carlo; Model misspecification; Hierarchical Bayesian model.

1. Introduction

Suppose a set of n observations y are generated from an unknown probability density $g(y)$ 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, $\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$, where $f(y|\theta)$ is the likelihood function and $\pi(\theta)$ is a prior distribution. The predictive distribution for a future observation z from the true model is $p(z|y) = \int f(z|\theta)\pi(\theta|y)d\theta$. 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. Although the Bayes factor (See e.g., Kass & Raftery, 1995) has been playing a major role in the evaluation of the goodness of the Bayesian models, the Bayes factor has come under increasing criticism due to its sensitivity to prior distributions. Under non-informative priors, the Bayes factor is frequently not well-defined. Many studies have been therefore conducted to evaluate the goodness of Bayesian models (Aitkin, 1991; Gelfand & Dey, 1994; Kass & Raftery, 1995; O'Hagan 1995; Berger & Pericchi, 1996; Perez & Berger, 2002; Ando 2007).

Aitkin (1991) proposed a posterior Bayes factor that chooses the model with the largest value of the posterior mean of the likelihood. Unfortunately, Akaike (1991) pointed out the possibility of the overfitting problem, which is caused by the repeated use of the observation y (Kadane & Lazar, 2004). To avoid over fitting problem, the following research was done by Spiegelhalter et al. (2002). A deviance information criterion, DIC, was proposed. DIC was defined as a posterior mean of the loglikelihood plus a penalty term, or equivalently, model complexity term. A penalty term was constructed on the notion of effective number of parameters. Unfortunately, many literatures (e.g., Robert & Titterington 2002, Ando 2007) still pointed out the overfitting problem since a penalty term does not reflect real model complexity.

However, the DIC is widely used thanks to its simplicity of calculation from posterior outputs. Solving the overfitting problem of DIC, but inheriting its ease

of calculation, we propose an information criterion from predictive point of view. The proposed criterion is applicable for evaluating the predictive distributions of hierarchical Bayesian and empirical Bayes models even when the specified family of probability distributions does not contain the true model $g(y)$.

§2 gives a main result. Some observations of the proposed criterion are provided in §3. §4 and §5 conduct Monte Carlo simulations to investigate the performance of the proposed criterion. Concluding and remarks are provided in §6.

2. Main result

In this paper, the best predictive distribution is determined by maximizing the posterior mean of the expected loglikelihood:

$$\eta = E_z\left[E_{\theta|y}\left\{\log f(z|\theta)\right\}\right] = \int\left\{\int \log f(z|\theta)\pi(\theta|y)d\theta\right\} dG(z) \tag{1}$$

among different Bayesian models. This utility function was recently employed Ando (2007) and implicitly by Spiegelhalter et al. (2002). When the unknown true model is replaced with the predictive distribution, the η reduces to a specified version of the predictive discrepancy measure proposed by Gelfand & Ghosh (1998).

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 η is the posterior mean of the loglikelihood

$$\hat{\eta} = \frac{1}{n}E_{\theta|y}\left\{\log f(y|\theta)\right\} = \frac{1}{n}\int \log f(y|\theta)\pi(\theta|y)d\theta. \tag{2}$$

As pointed out in §1, it is obvious that the posterior mean of the loglikelihood $\hat{\eta}$ generally provides a positive bias as an estimator of η. Therefore, bias correction should be considered. Ando (2007) defined the bias b of $\hat{\eta}$ in estimating η as

$$b_\theta = E_{y_n}(\hat{\eta} - \eta) = \int(\hat{\eta} - \eta)dG(y), \tag{3}$$

where expectation is taken over the joint distribution of y. Once we have an estimator of the bias, $\hat{b}_\theta$, the bias-corrected posterior mean of the loglikelihood is given by

$n^{-1}E_{\theta|y}\{\log f(y|\theta)\}-\hat{b}_\theta$, which is usually used in the form IC $= -2E_{\theta|y}\{\log L(y|\theta)\}+2n\hat{b}_\theta$. Under a certain mild regularity condition, we obtained the bias term. The result is in the following theorem.

Theorem 1 *Let η and $\hat{\eta}$ be as defined in (1) and (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 of $\hat{\eta}$ is given approximately by*

$$nb_\theta \approx P_D = 2\log f(y|\hat{\theta}_n) - 2E_{\theta|y}\{\log f(y|\theta)\}, \tag{4}$$

where $\hat{\theta}_n$ is the posterior mean, and the notation $\approx$ indicates that the difference between the two sides of the equation tends to zero as $n \to \infty$. The term P_D is an effective number of parameters (Spiegelhalter et al. (2002)), defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters:

The regularity conditions and derivation are given in Appendix.

Correcting the asymptotic bias of $\hat{\eta}$, an information criterion is given as

$$\text{IC} = -2E_{\theta|y}\left\{\log f(y|\theta)\right\} + 2P_D, \tag{5}$$

where $\hat{b}_\theta$ is given by the right-hand side of equation (4). We choose the predictive distribution that minimises the IC score. Recall that the original DIC was defined as DIC $= -2E_{\theta|y}\left\{\log f(y|\theta)\right\} + P_D$. Comparing with (5), the penalty term of the proposed criterion is twice of of that of original DIC.

The criterion (5) is available under the situation that (a) the consistency of θ holds, and (b) the penalised likelihood has a single mode. To ensure the consistency, we integrated out the random effects from the likelihood function when we consider the Bayesian model with random effects. See Section 5. The assumption (b) 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 characterized by multimodality. We would like to point out that when the Bayesian predictive information criterion (BPIC; Ando, 2007) is applicable, then

the proposed criterion is also available. Since the scope of BPIC is less limited than other model selection criteria (Ando, 2007), the proposed criterion can be widely applied.

However, as pointed out by Spiegelhalter et al. (2002), the developed penalty term based on P_D is not invariant to reparameterization. Thus, the proposed criterion in (5) suffers the same problem.

3. Some observations

3·1. Further simplification

If we impose assumptions, the bias term (4) reduces to a simple form. 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, $\log \pi(\theta) = O(1)$, 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) reduces to $\hat{b} \approx p$, where p is the dimension of θ (See Spiegelhalter et al. (2002)). In this situation, the criterion reduces to

$$\mathrm{IC} = -2E_{\theta|y}\left\{\log f(y|\theta)\right\} + 2p. \tag{6}$$

Thus, we can easily calculate the score. Also, note that, the penalty term has an advantage because it does not contain the simulation errors. This score is the simplified version of Bayesian predictive information criterion (BPIC; Ando, 2007),

3·2. Simple example

To illustrate 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 $\hat{b}_\mu$ are

$$b_\mu = E_y \left\{ \frac{1}{2} + \frac{(\mu_t - \hat{\mu}_n)^2}{2\sigma^2} - \frac{1}{n} \sum_{\alpha=1}^{n} \frac{(y_\alpha - \hat{\mu}_n)^2}{2\sigma^2} \right\},$$
$$\hat{b}_\mu = P_D^{\mu}/n = \sigma_n^2/\sigma^2.$$

==========

Insert Figure 1 around here.

==========

Figure 1 shows the true bias b_μ and the bias estimate $\hat{b}_\mu$ for various sample sizes n. The quantities are evaluated by a Monte Carlo simulation with 1,000,000 repetitions. The true mean, true variance and the prior mean are set to be $\mu_t = 0.2$, $\sigma = 0.5$ and $\mu_0 = 0.0$, so that the prior mean is slightly different from the true mean. In Figs 1 (a) and (b), the prior variances are set to be $\tau_0 = 0.1$ and $\tau_0 = 10$, 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 bias estimate is close to the true bias.

In this case, the true sampling density belongs to the specified parametric family of models. Also, the prior information becomes weak as sample size increases, the bias estimate thsu converges to the number of parameters $p = 1$.

4 Empirical Bayesian Modeling

We conduct Monte Carlo experiments to compare the proposed criterion with its competitors: Bayesian predictive information criterion (BPIC; Ando, 2007), deviance information criterion (DIC; Spiegelhalter et al., 2002), Bayes factor (Kass and Raftery, 1995), and the posterior Bayes factor (Aitkin, 1991), respectively. We also considered the frequentist's criteria network information criterion (NIC; Muarta et al., 1994), and modified AIC (AICM; Eilers & Marx, 1996), bias-corrected AIC

(AICC; Hurvich et al., 1998), respectively. However, note that, as pointed out by Ando et al, (2008), NIC tends to overfitting.

§4.1 considers P-spline generalised linear models. A tailor-made version of the proposed criterion is then derived. Numerical results are summarized in §4.2.

4·1. Generalised linear models with basis expansion predictors

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. Substituting the density function $f(y_\alpha|x_\alpha; \theta)$ into the equation (5), a tailor made version of our criterion can be derived.

4·2. Results

P-spline Gaussian regression model.

As an Gaussian example, datasets $\{(y_\alpha, x_\alpha); \alpha = 1, ..., n\}$ are repeatedly generated from the true regression model $y_\alpha = \sin(5\pi x_\alpha) + \varepsilon_\alpha$ for $x_\alpha = (2\alpha - 1)/(2n)$. The errors ε_α are independently and identically 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.

We 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. Note that the true model is mis-specified in both distributional and structural equations. Estimating the parameter vector by producing the posterior samples, the predictive distribution is obtained. In this case, taking

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

the IC in (5) is derived.

P-spline logistic regression model.

We generated a set of n observations for according to $\mathrm{pr}(Y_\alpha = 1|x_\alpha) = 1/[1 + \exp\{\sin(5\pi$ for $x_\alpha = (2\alpha - 1)/(2n)$. It is assumed that the probability $\pi(x_\alpha)$ is of the form: $\log[\pi(x_\alpha)/\{1 - \pi(x_\alpha)\}] = w^{\mathrm{T}} b(x_\alpha)$. Improved DIC in (5) for evaluating the predictive distribution can be obtained by taking

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

Results summary

In both examples, 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 Table 1 around here.

==========

Tables 1 compares the mean squared error between the true and estimated conditional expectations: MSE $= \sum_{\alpha=1}^{n} \{E(Y_\alpha|x_\alpha) - \hat{y}(x_\alpha)\}^2 / n$. 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 $\log_{10}(\lambda) = -5$. The number of basis functions ranges from 6 to 15. The simulation results were obtained from 200 repeated Monte Carlo trials.

It may be seen from the simulation results that BPIC, BF and our proposed criteria performed well in almost all cases. The mean value of the smoothing parameter chosen by DIC is smaller than those based on other criteria. The proposed criterion tends to choose fewer basis functions and larger values of λ than those based on DIC. It indicates that DIC is generally more variable and more likely to undersmooth than the proposed criterion. MSE indicates that the model selected by DIC overfits to the observed data.

5 Hierarchical Bayesian Modeling

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, §5.1 describes observation and state equations, their distributional assumptions and the prior distributions for the unknown parameters. §5.2 summarises the results.

5·1. 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 of the basic stochastic volatility model 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).

5·2. Results

In the simulation design, Model 3 was employed for the true model. Datasets are generated from the true model with parameter values $\phi = 0.8$, $\mu = -8.0$, $\tau = 0.2$

and $\nu = 10$, respectively. We simulate 100 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. To compute the Bayes factor, we utilise Chib (1995)'s marginal likelihood method (Chib's BF). In this simulation study, we employed a simple form of BPIC because of its complicated form. Thus, the BPIC and the proposed criterion have the same form in (6). However, we also checked the accuracy of the proposed in (5). 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 3.

It may be seen from Table 3 that the proposed criterion is superior to DIC; it chooses the correct model frequently than DIC. Since DIC provide much less penalty for model complexity than those of other criteria, the best model chosen by DIC is relatively complex. On the other hand, the Bayes factor tends to choose the simpler models than those selected by other criteria. In fact the standard model was selected 6 times among 100 trials. In conclusion, the proposed criteria also perform well in the full Bayesian modeling.

==========

Insert Table 3 around here.

==========

6 Concluding and Remarks

The main aim of this paper was to improve the performance of deviance information criteria (Spiegelhalter *et al.* (2002)), while preserving the computational advantage of DIC, i.e, easy score calculation. Since theoretical framework of DIC was unclear, we employed the Ando (2007)'s framework that select the best model by maximizing the posterior mean of the expected log-likelihood. Under this framework, it was clarified that the more accurate penalty term is double of that of DIC.

We also conduct numerical experiments to compare the performance of the proposed criteria with other Bayesian model selection criteria as well as frequentist's model selection criteria. As demonstrated by various numerical experiments, the proposed information criterion performs fairly well in various situations.

Appendix 1

Proof of Theorem 1. First, we define $\theta_0 = \lim_{n\to\infty} E_{\theta|y}[\theta]$ and decompose the bias in (3) as $E_y(\hat{\eta} - \eta) = E_1 + E_2 + E_3$, where

$$\begin{aligned}
E_1 &= E_y\left[\frac{1}{n}E_{\theta|y}\left\{\log f(y|\theta) - \log f(y|\theta_0)\right\}\right], \\
E_2 &= E_y\left[\frac{1}{n}E_{\theta|y}\left\{\log f(y|\theta_0)\right\} - E_z\left[\log\{f(z|\theta_0)\}\right]\right], \\
E_3 &= E_y\left(E_z\left[\log\{f(z|\theta_0)\}\right] - E_z\left[E_{\theta|y}\{\log f(z|\theta)\}\right]\right).
\end{aligned}$$

We first evaluate E_1. Using the Taylor expansion of $\log\{f(y|\theta)\}$ around θ_0 and then taking expectations of the above equation with respect to the posterior distribution gives $E_{\theta|y}[\log\{f(y|\theta)\}] \approx \log\{f(y|\theta_0)\} + (\hat{\theta}_n - \theta_0)^{\mathrm{T}}\partial\log\{f(y|\theta_0)\}/\partial\theta - 0.5\mathrm{tr}\{L_n(\theta_0)E_{\theta|y}[(\theta-\theta_0)(\theta-\theta_0)^{\mathrm{T}}]\}$. Here

$$L_n(\theta_0) = -\left.\frac{\partial^2 \log f(y|\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\right|_{\theta=\theta_0}$$

is the observed Fisher information matrix evaluated at θ_0. Notice also that

$$E_y\left(E_{\theta|y}[(\theta-\theta_0)(\theta-\theta_0)^{\mathrm{T}}]\right) = E_y[V_n(\theta)] + V(\hat{\theta}_n),$$

where $V_n(\theta) = E_{\theta|y}[(\theta-\hat{\theta}_n)^{\mathrm{T}}(\theta-\hat{\theta}_n)]$ is the posterior covariance matrix of θ, and $V_n(\hat{\theta}_n) = E_y[(\hat{\theta}_n-\theta_0)^{\mathrm{T}}(\hat{\theta}_n-\theta_0)]$ is the covariance matrix of $\hat{\theta}_n$. Then, noting that noting that $\hat{\theta}_n \to \theta_0$, and $L_n(\hat{\theta}_n) \to L(\theta_0)$ in probability as $n \to \infty$, , we have the following approximation

$$E_y\left(E_{\theta|y}[\log\{f(y|\theta_0)\}]\right) \approx E_y\left(E_{\theta|y}[\log\{f(y|\theta)\}]\right) + 0.5\mathrm{tr}[L(\theta_0)\{E_y[V_n(\theta)] + V(\hat{\theta}_n)\}],$$

where $L(\theta_0)$ is the expected Fisher information matrix evaluated at θ_0.

Similarly, Taylor expansion of $\log\{f(y|\theta)\}$ around the posterior mean $\hat{\theta}_n$ and then taking expectations of the above equation with respect to the posterior distribution gives

$$E_y[E_{\theta|y}[\log\{f(y|\theta)\}]] = E_y[\log f(y|\hat{\theta}_n)] - 0.5\mathrm{tr}\{L(\theta_0)E_y[V_n(\theta)]\}.$$

Substituting these expression into E_1, we finally have

$$\begin{aligned} nE_1 &\simeq E_y[\log f(y|\hat{\theta}_n)] - E_y[E_{\theta|y}[\log\{f(y|\theta)\}]] - 0.5\mathrm{tr}\left\{L(\theta_0)V(\hat{\theta}_n)\right\} \\ &= 0.5E_y[P_D] - 0.5\mathrm{tr}\left\{L(\theta_0)V(\hat{\theta}_n)\right\} \end{aligned}$$

with P_D is the effective number of parameters.

The term E_2 can be regarded as zero, because

$$E_2 = E_y\left[\log\{f(y|\theta_0)\}\right] - E_z\left[\log\{f(z|\theta_0)\}\right] = 0.$$

Next, we evaluate the term E_3. We again use the Taylor expansion of $\log\{f(z|\theta)\}$ around θ_0 and then take the posterior expectation. This gives

$$\begin{aligned} E_{\theta|y}[\log\{f(z|\theta)\}] \approx \log\{f(z|\theta_0)\} + (\hat{\theta}_n - \theta_0)^{\mathrm{T}}\partial\log\{f(z|\theta_0)\}/\partial\theta \\ -0.5\mathrm{tr}[L_n(\theta_0)\{V_n(\theta) + V(\hat{\theta}_n)\}]. \end{aligned}$$

Using the same auguments used in the evaluation of E_1, we have the following approximation

$$nE_3 \simeq 0.5\mathrm{tr}[L(\theta_0)\{E_y[V_n(\theta)] + V(\hat{\theta}_n)\}].$$

When the above results are combined, the asymptotic bias is given by

$$\begin{aligned} nE_y(\hat{\eta} - \eta) &\simeq 0.5E_y[P_D] - 0.5\mathrm{tr}\left\{L(\theta_0)V(\hat{\theta}_n)\right\} + 0.5\mathrm{tr}[L(\theta_0)\{E_y[V_n(\theta)] + V(\hat{\theta}_n)\}] \\ &= 0.5E_y[P_D] + 0.5\mathrm{tr}[L(\theta_0)E_y[V_n(\theta)]]. \end{aligned}$$

Replaceing the expectation of y by the empirical distribution and the true parameter value θ_0 by $\hat{\theta}_n$, we finally have

$$E_y(\hat{\eta} - \eta) \simeq P_D.$$

Here we used $\mathrm{tr}\{L_n(\hat{\theta}_n)V_n(\theta)\} = 0.5P_D$ (Spiegelhalter *et al.* (2002)).

References

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

Akaike, H. (1991). Discussion on "Posterior Bayes factors" *J. R. Statist. Soc.* B **53**, B53, 135.

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

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94, 443–458.

Ando, T., Konishi, S. and Imoto, S. (2008). Nonlinear regression modeling via regularized radial basis function networks. *J. Stat. Plann. Inference*, 138, 3616–3633.

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.

HURVICH, C. M., SIMONOFF, J. S. & TSAI, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* B **60**, 271–93.

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.

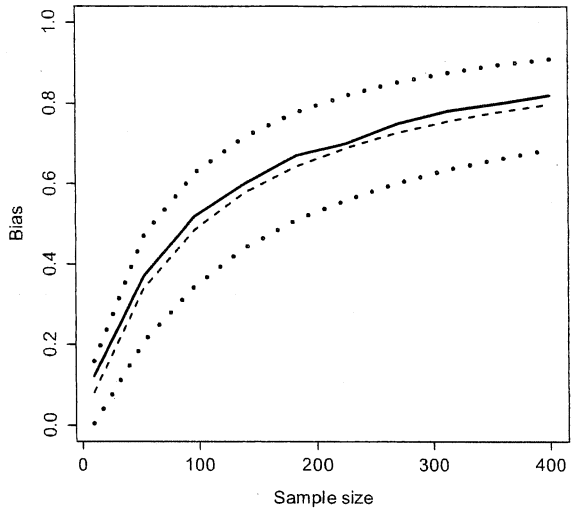
O'HAGAN, A. (1995). Fractional Bayes factors for model comparison (with Discussion). *J. R. Statist. Soc.* B **57**, 99–138.

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)

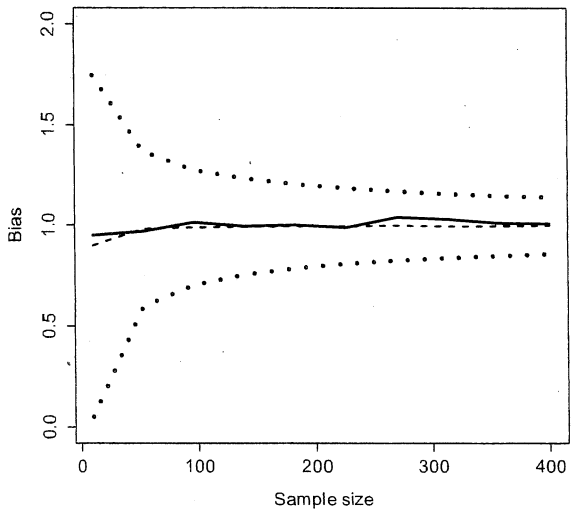


Figure 1: Simple normal example. Comparison of the true bias (—), the estimated bias (- - -) for various sample sizes. (a) under the rather informative prior with $\tau_0 = 0.1$ and (b) under a flat informative prior with $\tau_0 = 100$. The dotted lines in each panel indicates ± 2 times the standard deviation of the estimated bias.

Table 1: Results for the Gaussian regression model. Comparison of the mean squared errors based on various criteria. The mean squared error is defined as the difference between the true and estimated conditional expectations: MSE $= \sum_{\alpha=1}^{n} \{E(Y_\alpha|x_\alpha) - \hat{y}(x_\alpha)\}^2 / n$. We used the predictive mean of $y_\alpha|x_\alpha$ for $\hat{y}(x_\alpha)$. The simulation results were obtained from 100 repeated Monte Carlo trials. Averages are given in each of the first rows and figures in the second row give estimated standard deviations. The means and standard deviations of the selected smoothing parameter λ and the number of basis functions m are also given. NIC; network information criterion (Muarta et al., 1994), AICC; bias-corrected AIC (Hurvich et al., 1998), AICM; modified AIC (Eilers & Marx, 1996), BPIC; Bayesian predictive information criterion (Ando, 2007), DIC; deviance information criterion (Spiegelhalter et al., 2002), and BF; Bayes factor (Kass and Raftery, 1995).

	$n = 100$			$n = 200$		
	m	λ	MSE	m	λ	MSE
NIC	9.28	0.01408	0.00738	8.94	0.00449	0.00352
	1.90	0.02914	0.00469	1.54	0.00472	0.00217
AICC	9.56	0.35207	0.00530	9.44	0.11867	0.00274
	1.61	0.44831	0.00329	1.56	0.24710	0.00160
AICM	9.66	0.25258	0.00578	9.48	0.10318	0.00287
	1.68	0.40092	0.00380	1.68	0.23201	0.00173
BPIC	10.32	0.87183	0.00457	10.11	0.38025	0.00230
	1.89	1.67932	0.00295	1.86	0.44860	0.00116
DIC	10.54	0.35215	0.00604	10.66	0.21529	0.00296
	2.02	0.44835	0.00422	2.09	0.35953	0.00177
BF	13.38	0.88300	0.00463	11.73	0.35110	0.00236
	1.16	0.30419	0.00285	1.65	0.40679	0.00120
Improved DIC	10.49	0.77221	0.00471	10.44	0.42794	0.00246
	1.97	1.40230	0.00297	2.05	0.45123	0.00124

Table 2: Results for the Logistic regression model. Comparison of the mean squared errors based on various criteria. The mean squared error is defined as the difference between the true and estimated conditional expectations: MSE = $\sum_{\alpha=1}^{n} \{E(Y_\alpha|x_\alpha) - \hat{y}(x_\alpha)\}^2 / n$. We used the predictive mean of $P(y_\alpha = 1|x_\alpha)$ for $\hat{y}(x_\alpha)$. The simulation results were obtained from 100 repeated Monte Carlo trials. Averages are given in each of the first rows and figures in the second row give estimated standard deviations. The means and standard deviations of the selected smoothing parameter λ and the number of basis functions m are also given. NIC; network information criterion (Muarta et al., 1994), AICC; bias-corrected AIC (Hurvich et al., 1998), AICM; modified AIC (Eilers & Marx, 1996), BPIC; Bayesian predictive information criterion (Ando, 2007), DIC; deviance information criterion (Spiegelhalter et al., 2002), and BF; Bayes factor (Kass and Raftery, 1995).

	$n = 100$			$n = 200$		
	m	λ	MSE	m	λ	MSE
NIC	9.44	0.10102	0.02304	9.36	0.00002	0.01012
	1.73	0.30127	0.01114	1.12	0.00005	0.00499
AICC	8.61	0.12338	0.02148	8.90	0.00003	0.01015
	1.72	0.32653	0.01081	1.29	0.00005	0.00506
AICM	9.15	0.08333	0.02123	9.40	0.00006	0.01016
	1.79	0.27310	0.01108	1.49	0.00008	0.00496
BPIC	9.11	0.23447	0.01719	9.86	0.14607	0.00894
	2.03	0.34268	0.00849	2.19	0.27058	0.00351
DIC	11.02	0.00016	0.01977	11.13	0.00008	0.00989
	1.75	0.00034	0.00944	1.50	0.00013	0.00417
BF	11.98	0.35409	0.01877	12.13	0.07235	0.00948
	1.12	0.43461	0.00877	1.02	0.19429	0.00410
Improved DIC	10.81	0.00028	0.01730	10.83	0.00017	0.00901
	1.83	0.00070	0.00885	1.55	0.00032	0.00406

Table 3: Frequency distribution of selected models across 100 simulated replications. The datasets are generated from Model 3. BPIC (Bayesian predictive information criterion: Ando (2007)), DIC (Deviance information criterion: Spiegelhalter et al. (2002)), BF (Bayes factor: Kass and Raftery (1995)), Improved DIC (DIC1) in (5) and Improved DIC (DIC2) in (6), respectively. Note that the scores of BPIC and DIC2 are the same and thus give the same model selection results.

Models	1	2	3	4	5	6
BPIC	0	0	75	0	25	0
DIC	0	0	44	0	56	0
BF	6	0	70	0	24	0
DIC1	0	0	73	0	27	0
DIC2	0	0	75	0	25	0