Biometrika (2007), **94**, 2, *pp.* 443–458
© 2007 Biometrika Trust
Printed in Great Britain

doi:10.1093/biomet/asm017
Advance Access publication 28 February 2007

Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes 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 hierarchical Bayesian and empirical Bayes models is investigated. A Bayesian predictive information criterion is proposed as an estimator of the posterior mean of the expected loglikelihood of the predictive distribution when the specified family of probability distributions does not contain the true distribution. The proposed criterion is developed by correcting the asymptotic bias of the posterior mean of the loglikelihood as an estimator of its expected loglikelihood. In the evaluation of hierarchical Bayesian models with random effects, regardless of our parametric focus, the proposed criterion considers the bias correction of the posterior mean of the marginal loglikelihood because it requires a consistent parameter estimator. The use of the bootstrap in model evaluation is also discussed.

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

1. INTRODUCTION

Model selection is a major issue in statistical science, and a number of studies have been conducted to evaluate the goodness of hierarchical Bayesian and empirical Bayes models (Gelfand & Dey, 1994; Kass & Raftery, 1995; Perez & Berger, 2002).

Spiegelhalter et al. (2002) considered asymptotic bias correction of a posterior mean of the loglikelihood as an estimator of a posterior mean of the expected loglikelihood and proposed a deviance information criterion, DIC. Unfortunately, there are at least two critical problems in their information-theoretic justification of DIC.

First, in the derivation of DIC, Spiegelhalter et al. (2002, p.604) assumed that the specified parametric family of probability distributions that generate future observations encompasses the true model; as shown in the Appendix, this assumption is not needed to derive DIC. This assumption does not always hold, and we have to consider model assessment procedures in that scenario (Stone, 2002). Secondly, the observed data are used both to construct the posterior distribution and to compute the posterior mean of the expected loglikelihood. The bias estimate of DIC tends to underestimate the true bias considerably.

The main aim of this paper is to propose a Bayesian predictive information criterion, BPIC, for evaluating the predictive distributions of hierarchical Bayesian and empirical Bayes models when the specified family of probability distributions does not contain the true model.

2. BAYESIAN PREDICTIVE INFORMATION CRITERION

2·1. *Preliminaries: empirical and hierarchical Bayesian models*

Suppose a set of n independent observations $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ are generated from an unknown distribution $G(y)$ with a probability density $g(y)$, and that a parametric family of distributions with densities $\{f(y|\theta); \theta \in \Theta \subset R^p\}$ is used to approximate the true model.

In the Bayesian approach, a prior density $\pi(\theta|\psi)$ is specified for the parameter vector θ, where ψ is a hyperparameter. The parametric model sometimes involves (θ, ψ) and has the form $f(y|\theta, \psi)$. In an empirical Bayes model, the hyperparameter is considered as nonstochastic and a likelihood function and prior density are specified as $L(y|\theta, \psi) = \prod_{\alpha=1}^{n} f(y_\alpha|\theta, \psi)$ and $\pi(\theta|\psi)$ for a given value of ψ. An alternative to the empirical Bayes approach is a fully Bayesian approach, which further includes a prior $\pi(\psi)$ for the unknown hyperparameters, thereby creating a hierarchical model. The full probability model generally factorizes as $f(y|\theta, \psi)\pi(\theta|\psi)\pi(\psi)$. Depending on the parameters of interest, the hierarchical model can compose various specifications of the likelihood and the prior distribution (Spiegelhalter et al., 2002, p.585).

In each case, inference for γ is provided by its posterior distribution $\pi(\gamma|y)$, where γ denotes the parameter vector specified to be the random variable in the prior distribution. In the empirical Bayes model, $\gamma = \theta, \pi(\gamma) = \pi(\theta|\psi)$ and $\pi(\gamma|y) = \pi(\theta|y, \psi)$, respectively. Various specifications of γ and $\pi(\gamma|y)$ are provided in §§3–6.

The predictive distribution for a future observation z generated from unknown distribution $g(z)$ is $q(z|y) = \int f(z|\gamma)\pi(\gamma|y)d\gamma$. Even if the integration problem cannot be solved analytically, the predictive distribution can be easily approximated by Markov chain Monte Carlo methods.

2·2. *Theoretical framework and main result*

The critical issue with Bayesian modelling is how to evaluate the goodness of the predictive distributions. Akaike (1974) proposed the information criterion AIC under the assumptions that (i) a specified parametric family of probability distributions encompasses the true model and (ii) a model is estimated by the maximum likelihood method. The divergence of the fitted model from the true model is measured by the Kullback-Leibler information, or equivalently by an expected loglikelihood $\int \log f(z|\hat{\theta})dG(z)$, where $\hat{\theta}$ is the maximum likelihood estimator. Subsequent generalizations of AIC include TIC (Takeuchi, 1976), which relaxed the assumption (i), and GIC (Konishi & Kitagawa, 1996), which relaxed both assumptions (i) and (ii).

Recently, Spiegelhalter et al. (2002) implicitly considered the maximization of the posterior mean of the expected loglikelihood,

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

to measure the deviation of $q(z|y)$ from the true model $g(z)$; the best predictive distribution maximizes (1) among different statistical models. If the unknown true model is assumed to be the predictive distribution, i.e. $g(z) = q(z|y)$, (1) reduces to a version of the predictive discrepancy measure proposed by Gelfand & Ghosh (1998). Thus the maximization of η is an extension of their concepts and constitutes a more general approach.

It is obvious that η in (1) depends on the model fitted, and on the unknown true model $g(z)$. A natural estimator of η is the posterior mean of the loglikelihood,

$$\hat{\eta} = \frac{1}{n} E_{\gamma|y} \{\log L(y|\gamma)\} = \frac{1}{n} \int \log L(y|\gamma)\pi(\gamma|y)d\gamma, \tag{2}$$

where $L(y|\gamma) = \prod_{\alpha=1}^{n} f(y_\alpha|\gamma)$. The quantity, $\hat{\eta}$, is generally positively biased as an estimator of η, because the same data y are used both to construct the posterior distribution $\pi(\gamma|y)$ and to evaluate η. Therefore, bias correction should be considered, where the bias b_γ is

$$b_\gamma = E_y(\hat{\eta} - \eta) = \int \left(\frac{1}{n} E_{\gamma|y} \{\log L(y|\gamma)\} - E_z \left[E_{\gamma|y}\{\log f(z|\gamma)\} \right] \right) dG(y). \tag{3}$$

If b_γ can be estimated, by $\hat{b}_\gamma$, the bias-corrected posterior mean of the loglikelihood is given by $n^{-1}E_{\gamma|y}\{\log L(y|\gamma)\} - \hat{b}_\gamma$, which is usually used in the form $\text{IC} = -2E_{\gamma|y}\{\log L(y|\gamma)\} + 2n\hat{b}_\gamma$.

Under this framework, Spiegelhalter et al. (2002) proposed DIC, where

$$\text{DIC} = -2E_{\gamma|y} \{\log L(y|\gamma)\} + P_D^\gamma. \tag{4}$$

The second term is an effective number of parameters, defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters: $P_D^\gamma = 2[\log L(y|\bar{\gamma}_n) - E_{\gamma|y}\{\log L(y|\gamma)\}]$, where $\bar{\gamma}_n$ is the posterior mean.

Although it is not clear that Spiegelhalter et al. (2002) tried to estimate the asymptotic bias (3), it is obviously estimated by $\hat{b}_\gamma^{\text{DIC}} = P_D^\gamma/(2n)$. The bias term of DIC is derived in Appendix 1. Spiegelhalter et al. (2002) gave an asymptotic justification for deriving the effective number of parameters. However, as pointed out by Robert & Titterington (2002), the same data were used twice in the construction of P_D^γ. As a result, the predictive distribution chosen by DIC is more complex than that with BPIC and overfits the observed data. In this paper, we evaluated the asymptotic bias (3) more accurately under model misspecification.

THEOREM 1. *Let η and $\hat{\eta}$ be as defined in* (1) *and* (2)*, respectively, and suppose that the specified family of probability distributions does not necessarily contain the true model. Then, under certain regularity conditions, the asymptotic bias of $\hat{\eta}$ is given approximately by*

$$nb_\gamma \approx E_{\gamma|y}[\log\{L(y|\gamma)\pi(\gamma)\}] - \log\{L(y|\hat{\gamma}_n)\pi(\hat{\gamma}_n)\} + \text{tr}\left\{J_n^{-1}(\hat{\gamma}_n)I_n(\hat{\gamma}_n)\right\} + p/2, \tag{5}$$

where the notation $\approx$ indicates that the difference between the two sides of the equation tends to zero as $n \to \infty$, p is the dimension of γ, $\hat{\gamma}_n = argmax_\gamma \pi(\gamma|y)$ is the posterior mode and the matrices $I_n(\hat{\gamma}_n)$ and $J_n(\hat{\gamma}_n)$ are given by

$$I_n(\gamma) = \frac{1}{n} \sum_{\alpha=1}^{n} \left\{ \frac{\partial \eta_n(y_\alpha, \gamma)}{\partial \gamma} \frac{\partial \eta_n(y_\alpha, \gamma)}{\partial \gamma^{\mathrm{T}}} \right\}, \quad J_n(\gamma) = -\frac{1}{n} \sum_{\alpha=1}^{n} \left\{ \frac{\partial^2 \eta_n(y_\alpha, \gamma)}{\partial \gamma \partial \gamma^{\mathrm{T}}} \right\},$$

respectively. Here $\eta_n(y_\alpha, \gamma) = \log f(y_\alpha|\gamma) + \log \pi(\gamma)/n$ and the prior distribution $\pi(\gamma)$ may depend on n as long as $\lim_{n\to\infty} n^{-1} \log \pi(\gamma)$ *is finite.*

The regularity conditions and derivation are given in Appendix 2.

To correct the asymptotic bias of $\hat{\eta}$, we propose the following Bayesian predictive information criterion:

$$\text{BPIC} = -2E_{\gamma|y} \{\log L(y|\gamma)\} + 2n\hat{b}_\gamma, \tag{6}$$

where $\hat{b}_\gamma$ is given by the right-hand side of equation (5). We choose the predictive distribution that minimizes BPIC.

3. A SIMPLE EXAMPLE

To give insight into BPIC, we first apply it to a simple normal model with known variance. Suppose that we have n independent observations, $y_1, \ldots, y_n$, each from a normal distribution with true mean μ_{T} and known variance σ^2, i.e. $g(z|\mu_{\mathrm{T}})$ corresponds to $N(\mu_{\mathrm{T}}, \sigma^2)$. We assume the data are generated from a normal distribution $f(z|\mu)$, corresponding to $N(\mu, \sigma^2)$. 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 $\sigma_n^2 = 1/(1/\tau_0^2 + n/\sigma^2)$. Thus

$$\eta = \int \left\{ \int \log f(z|\mu)\pi(\mu|y)d\mu \right\} dG(z) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{\sigma^2 + (\mu_{\mathrm{T}} - \hat{\mu}_n)^2 + \sigma_n^2}{2\sigma^2}$$

$$\hat{\eta} = \frac{1}{n}\sum_{\alpha=1}^n \int \log f(y_\alpha|\mu)\pi(\mu|y)d\mu = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{n}\sum_{\alpha=1}^n \frac{(y_\alpha - \hat{\mu}_n)^2 + \sigma_n^2}{2\sigma^2},$$

and the true bias (3) is

$$b_\mu = E_y\left\{ \frac{1}{2} + \frac{(\mu_{\mathrm{T}} - \hat{\mu}_n)^2}{2\sigma^2} - \frac{1}{n}\sum_{\alpha=1}^n \frac{(y_\alpha - \hat{\mu}_n)^2}{2\sigma^2} \right\}.$$

Straightforward calculations give $\log\{L(y|\hat{\mu}_n)\} - E_{\mu|y}[\log\{L(y|\mu)\}] = n\sigma_n^2/(2\sigma^2)$, $\log\{\pi(\hat{\mu}_n)\} - E_{\mu|y}[\log\{\pi(\mu)\}] = \sigma_n^2/(2\tau_0^2)$,

$$I_n(\hat{\mu}_n) = \sum_{\alpha=1}^n \{(y_\alpha - \hat{\mu}_n)/\sigma^2 + (\mu_0 - \hat{\mu}_n)/(n\tau_0^2)\}^2/n \tag{7}$$

and $J_n(\hat{\mu}_n) = 1/(n\sigma_n^2)$. The asymptotic bias estimate (5) is then

$$n\hat{b}_\mu = -n\sigma_n^2/(2\sigma^2) - \sigma_n^2/(2\tau_0^2) + J_n^{-1}(\hat{\mu}_n)I_n(\hat{\mu}_n) + 1/2, \tag{8}$$

whereas the asymptotic bias estimate of DIC is $n\hat{b}_\mu^{\mathrm{DIC}} = nP_D^\mu/(2n) = n\sigma_n^2/(2\sigma^2)$.

Figure 1 plots the true bias b_μ and the bias estimates of BPIC and DIC for various sample sizes n, for which the quantities are evaluated by a Monte Carlo simulation with 100 000 repetitions. The true mean and variance are arbitrarily set to be $\mu_{\mathrm{T}} = 0$ and $\sigma^2 = (0{\cdot}5)^2$, respectively. The prior mean is set to be $\mu_0 = 0$. In Fig. 1(a) and (b), the prior variances are set to be $\tau_0^2 = (0{\cdot}1)^2$, corresponding to a rather informative prior and $\tau_0^2 = (100)^2$, corresponding to a flat noninformative prior, respectively. Figure 1 shows that $\hat{\eta}$ has a significant bias as an estimator of η. The estimated asymptotic bias of BPIC is close to the true bias, whereas the estimated asymptotic bias of DIC underestimates the true bias considerably. We found the results described above to be essentially unchanged when investigated under model misspecification.

When the prior information is weak, $n\hat{b}_\mu^{\mathrm{DIC}} = 1/2 + O(n^{-1})$. Since the specified parametric family of probability distributions encompasses the true model and the prior is dominated by the likelihood as n increases, we have that $J_n(\hat{\mu}_n) \doteq I_n(\hat{\mu}_n)$, and $n\hat{b}_\mu = 1 + O(n^{-1})$. In this case, therefore, the asymptotic bias estimate of BPIC is approximated by the dimension of γ, double that for DIC. A more general scenario is discussed in § 7·3.

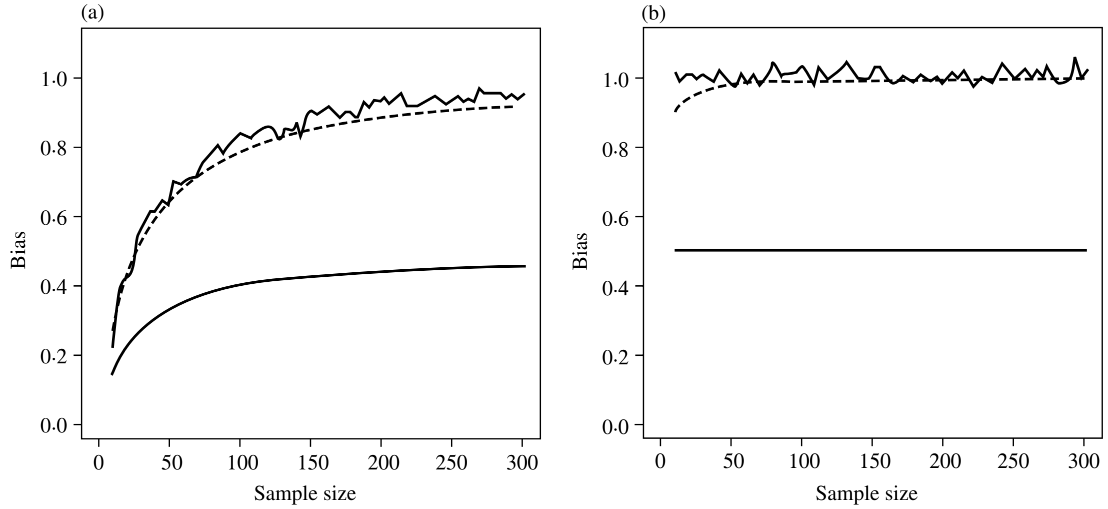

Fig. 1. Simple normal example. Comparison of the true bias nb_μ (—), the estimated asymptotic bias of BPIC (- - -) and the estimated asymptotic bias of DIC (—) for various sample sizes. (a) under the rather informative prior with $\tau_0 = 0{\cdot}1$ and (b) under a flat noninformative prior with $\tau_0 = 100$.

4. An empirical bayes example

We consider a univariate normal mixture model, with $f(y|\theta) = \sum_{j=1}^{K} w_j f(y|\mu_j, \sigma_j^2)$, for $\theta = (w_1, \mu_1, \sigma_1^2, \ldots, w_K, \mu_K, \sigma_K^2)^{\mathrm{T}}$, $\sum_{j=1}^{K} w_j = 1$, and use the proposed criterion to assess the number of components in the model.

A set of $n = 200$ observations are generated from a skewed bimodal mixture model $0{\cdot}75N(0, 1) + 0{\cdot}25N(1{\cdot}5, 0{\cdot}33)$. Richardson (2002) has already pointed out that DIC tends to select overfitted models. Using the well known conjugate priors $w \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$, $\mu_j|\sigma_j^2 \sim N(\mu_{0j}, \sigma_j^2/k_{0j})$ and $\sigma_j^2 \sim \mathrm{IG}(a_j, b_j)$, we generated posterior samples by using a Gibbs sampling algorithm with the identifiability constraint $\mu_1 < \cdots < \mu_K$. The hyperparameters are set to be $\alpha_j = 5$, $a_j = 1$, $b_j = 5$ and $k_{0j} = 3$, respectively. A set of 2000 posterior samples are generated after 1000 burn-in realizations. We also evaluated AIC and BIC (Schwarz, 1978).

Table 1 shows that BPIC achieved the minimum score at $K = 2$ and identified the correct number of mixture components. The asymptotic bias estimate of BPIC becomes larger as the number of mixture components increases. On the other hand, the bias estimate of DIC gives almost the same values between two and five components. Consequently, DIC tends to select overfitted models with more than three components.

Table 1. *Galaxy data. Model selection scores for mixture models with different numbers of components. The bias estimates of* BPIC *and* DIC *are also shown*

K	p	BPIC	DIC	AIC	BIC	$\bar{L}$	$P_D^\gamma/2$	$n\hat{b}_\gamma$
2	5	622·14	617·30	622·16	638·65	−306·38	2·27	4·69
3	8	624·60	615·32	622·50	648·88	−305·00	2·66	7·30
4	11	631·40	615·20	621·84	658·12	−304·81	2·96	10·89
5	14	631·64	615·30	624·21	670·37	−304·64	3·01	11·20

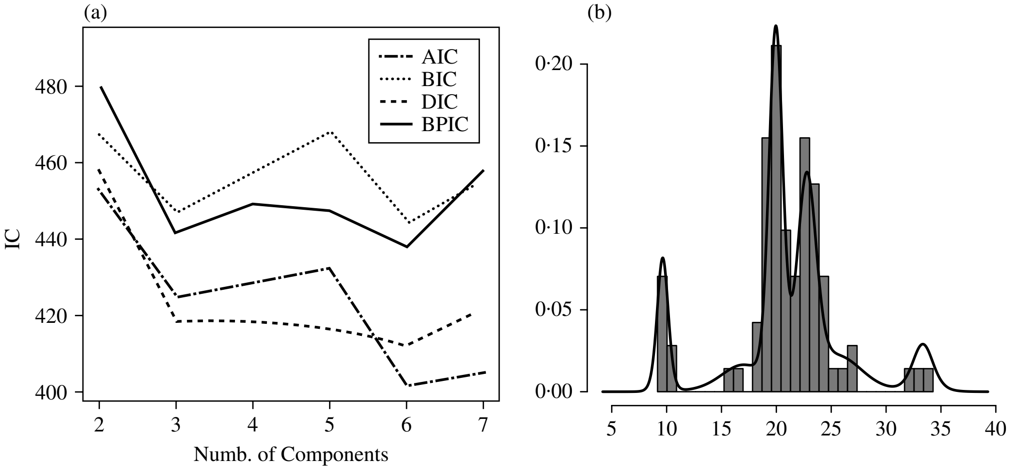

Fig. 2. Galaxy data. (a) Model evaluation scores for various number of components K and (b) the estimated predictive density.

The normal mixture model was also fitted to the galaxy data (Richardson & Green, 1997), which consists of 82 observations. Figure 2 (a) shows that all criteria selected the six-components mixture model. The reversible jump algorithm of Richardson & Green (1997) also identified $K = 6$ as the posterior mode. Figure 2 (b) shows the estimated predictive density with $K = 6$.

5. A HIERARCHICAL BAYES EXAMPLE

This section evaluates the default probabilities of companies listed on the Tokyo Exchange by using the functional logistic regression model with random effects:

$$\log\left(\frac{\mathrm{pr}\{y_\alpha = 1|x_\alpha(\cdot)\}}{\mathrm{pr}\{y_\alpha = 0|x_\alpha(\cdot)\}}\right) = \sum_{j=1}^{r}\int \beta_j(z)x_{\alpha j}(z)dz + c_\alpha, \quad \alpha = 1,\ldots,n,$$

where $\mathrm{pr}\{y_\alpha = 1|x_\alpha(\cdot)\}$ represents the default probability, $\beta(t) = (\beta_1(t),\ldots,\beta_r(t))^{\mathrm{T}}$ is the coefficient vector, $x_\alpha(t) = (x_{\alpha 1}(t),\ldots,x_{\alpha r}(t))^{\mathrm{T}}$ is the functional predictor (Ramsay & Silverman, 1997), which represents the transitions of financial ratios, and $c_\alpha \sim N(0,\sigma_c^2)$ is the random effect. The set of 7 financial ratios in Table 2 are considered as predictors. With the coefficients expressed as a linear combination of B-spline basis functions, $\beta_k(t) = \sum_{j=1}^{m} w_{kj}b_{kj}(t) = w_k^{\mathrm{T}}b_k(t)$ (Ramsay & Silverman, 1997), the default risk evaluation model, conditional on w and c_α is $f\{y_\alpha|x_\alpha(\cdot),c_\alpha,w\} = \pi\{x_\alpha(\cdot),c_\alpha,w\}^{y_\alpha}[1-\pi\{x_\alpha(\cdot),c_\alpha,w\}]^{1-y_\alpha}$, where $\pi\{x_\alpha(\cdot),c_\alpha,w\} = [1+\exp\{-\sum_{j=1}^{r}\int\beta_j(z)x_{\alpha j}(z)dz - c_\alpha\}]^{-1}$ is the default probability of the αth company and $w = (w_1^{\mathrm{T}},\ldots,w_r^{\mathrm{T}})^{\mathrm{T}}$.

Using the $m \times r$ dimensional normal prior $w \sim N(0,\lambda D_r)$ and an inverse gamma prior $\sigma^2 \sim \mathrm{IG}(a,b)$, we have the hierarchical Bayesian model

$$\prod_{\alpha=1}^{n} f\{y_\alpha|x_\alpha(\cdot);c_\alpha,w\}f(c_\alpha|\sigma_c^2)\pi(w)\pi(\sigma_c^2). \tag{9}$$

Here $D_r = \mathrm{diag}\{D,\ldots,D\}$ is an $m \times r$ block diagonal matrix and D is a matrix representation of the difference operator such that $\sum_{j=2}^{m}(\Delta^2 w_{kj})^2 = w_k^{\mathrm{T}}Dw_k$ with $\Delta w_{kj} = w_{kj} - w_{k,j-1}$.

Table 2. *A set of financial ratios. A circle indicates that the model selection criterion employed that financial ratio for the predictor. The scores of area under the* ROC *curve,* AUC*, for training and test data are also calculated for each model, selected by* BPIC, DIC^m, DIC^c, GIC *and* HMBF*, respectively*

Financial ratio	BPIC	DIC^m	DIC^c	GIC	HMBF
Net income per employee	○	○	○	○	○
Net income per share	○	○	○	○	○
Net income		○	○	○	
Cash flow		○	○	○	
Cash flow to dept ratio		○	○	○	
Capital adequacy ratio		○	○	○	
Return on asset		○	○	○	
AUC for training data	0·9873	0·9943	0·9943	0·9943	0·9688
AUC for test data	0·9098	0·8981	0·8981	0·8981	0·8945

A subset of predictors is selected so as to minimize BPIC. With $m = 5, \lambda = 0{\cdot}01, a = 0{\cdot}0001$ and $b = 0{\cdot}0001$, a dataset of $n = 200$ training samples, of which 50 companies defaulted, was analyzed. Using the Metropolis-Hastings algorithm, we obtained a set of 5000 posterior samples $\{c_1, \ldots, c_n, w, \sigma_c^2\}$ from the posterior distribution. To ensure the consistency of γ in (6), we first integrated out the random effects c_α and then computed BPIC by using the following specification: $\gamma = (w^{\mathrm{T}}, \sigma_c^2)^{\mathrm{T}}, L(y|\gamma) = \prod_{\alpha=1}^{n} \int f\{y_\alpha|x_\alpha(\cdot); c_\alpha, w\} f(c_\alpha|\sigma_c^2) dc_\alpha$, and $\eta_n(y_\alpha, \gamma) = \log \int f\{y_\alpha|x_\alpha(\cdot); c_\alpha, w\} f(c_\alpha|\sigma_c^2) dc_\alpha + \log\{\pi(w)\pi(\sigma_c^2)\}/n$. The details of this specification will be discussed in § 7·1.

Table 2 shows the variable selection results based on BPIC, DIC, GIC and the harmonic mean estimated Bayes factor, HMBF (Newton & Raftery, 1994). Although DIC is designed for use in a situation where the random effects are of explicit interest, we have calculated the two types of DIC score, DIC^m, based on the marginal loglikelihood, and DIC^c, based on the conditional loglikelihood. A set of $r = 3$ financial ratios were selected by BPIC. The area under the ROC curve, AUC, for test data is calculated to evaluate the predictive ability of the selected model. The test data consist of 603 samples, in which 51 companies defaulted, and the default probabilities for the test data are calculated by $\int[\int \pi\{x_\alpha(\cdot), c_\alpha, w\} f(c_\alpha|\sigma_c^2) dc_\alpha]\pi(w, \sigma_c^2|y) dw d\sigma_c^2$. As shown in Table 2, the model selected by BPIC achieved the best AUC score. On the other hand, DIC^m, DIC^c and GIC selected an overfitted model with $r = 7$ financial ratios.

6. Numerical comparisons

6·1. *Preamble*

Monte Carlo experiments were conducted to compare BPIC with its competitors: DIC (Spiegelhalter et al., 2002), HMBF (Newton & Raftery, 1994), the crossvalidation predictive density approach (CVPD; Gelfand et al., 1992), GIC (Konishi & Kitagawa, 1996), the network information criterion (NIC; Murata et al., 1994), modified AIC (AIC$_M$; Eilers & Marx, 1996), bias-corrected AIC (AIC$_C$; Hurvich et al., 1998), the crossvalidation (CV; Stone, 1974) and generalized crossvalidation (GCV; Craven & Wahba, 1979), respectively.

In § 6·2, we consider a generalized linear model with basis-expansion predictor and derive a tailor-made version of BPIC. In § 6·3, we conduct Monte Carlo experiments and the resulting formula is applied.

6·2. *Generalized linear model with basis-expansion predictor*

Suppose we have n independent observations $\{(y_\alpha, x_\alpha); \alpha = 1, 2, \ldots, n\}$, where y_α are random response variables and x_α are q-dimensional explanatory variables. In generalized linear models (McCullagh & Nelder, 1989), y_α are assumed to be drawn from the exponential family of distributions with densities $f(y_\alpha|x_\alpha; \xi_\alpha, \phi) = \exp[\{y_\alpha\xi_\alpha - u(\xi_\alpha)\}/\phi + v(y_\alpha, \phi)]$, 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 the basis-expansion approach (Eilers & Marx, 1996), the unknown predictors η_α are approximated by a linear combination of basis functions $\eta_\alpha = \sum_{j=1}^m w_j b_j(x_\alpha) = w^{\mathrm{T}} b(x_\alpha)$, where $w = (w_1, \cdots, w_m)^{\mathrm{T}}$ is the m-dimensional coefficient vector and $b(x) = (b_1(x), \cdots, b_m(x))^{\mathrm{T}}$ is the m-dimensional basis function vector. Then one obtains generalized linear models with the basis-expansion predictor as a probability density $f(y_\alpha|x_\alpha; \theta) = \exp([y_\alpha r\{w^{\mathrm{T}} b(x_\alpha)\} - s\{w^{\mathrm{T}} b(x_\alpha)\}]/\phi + v(y_\alpha, \phi))$, where $\theta = (w^{\mathrm{T}}, \phi)^{\mathrm{T}}$, $r(\cdot) = u'^{-1} \circ h^{-1}(\cdot)$ and $s(\cdot) = u \circ u'^{-1} \circ h^{-1}(\cdot)$.

Posterior inference for θ can be achieved by simulating posterior samples. In this paper, we shall use a singular multivariate normal prior density (Konishi et al., 2004) $\pi(\theta|\psi) = \{n\lambda/(2\pi)\}^{(m-2)/2}|R|_+^{1/2}\exp\{-n\lambda\theta^{\mathrm{T}} R\theta/2\}$, where λ is a smoothing parameter, $\psi = (m, \lambda)^{\mathrm{T}}$ is a hyperparameter vector, $R = \mathrm{diag}\{D, 0\}$ 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 value of the hyperparameter ψ, or equivalently the smoothing parameter λ and the number of basis functions m. Substituting the density functions $f(y_\alpha|x_\alpha; \theta)$ and $\pi(\theta|\psi)$ into equation (6), we can derive an explicit version of BPIC. The $(m + 1) \times (m + 1)$ matrices $I_n(\hat{\theta}_n)$ and $J_n(\hat{\theta}_n)$ are given by

$$I_n(\hat{\theta}_n) = \frac{1}{n}\begin{pmatrix} B^{\mathrm{T}}\Lambda/\hat{\phi}_n - \lambda D\hat{w}_n 1_n^{\mathrm{T}} \\ p^{\mathrm{T}} \end{pmatrix}\begin{pmatrix} \Lambda B/\hat{\phi}_n - \lambda 1_n^{\mathrm{T}}\hat{w}_n D, & p \end{pmatrix},$$

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

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

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

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

We choose the hyperparameter ψ so as to minimize BPIC.

6·3. *Results*

Datasets $\{(y_\alpha, x_\alpha); \alpha = 1, \ldots, n\}$ are repeatedly generated from the true regression model $y_\alpha = \sin(3\pi x_\alpha) + \varepsilon_\alpha$ for $x_\alpha = (2\alpha - 1)/(2n)$. The errors ε_α are assumed to be independently distributed according to a mixture of normal distributions $g(\varepsilon_\alpha) = \beta N(\varepsilon_\alpha|0, \sigma_1^2) + (1 - \beta)N(\varepsilon_\alpha|0, \sigma_2^2)$, Here β is a mixing proportion, and $N(\varepsilon|\mu, \sigma^2)$ denotes the normal density function with mean μ and variance σ^2.

We estimate the true function by using a P-spline Gaussian regression model. In the Metropolis-Hastings algorithm, the Gaussian independence sampler is used. To save computational time, the initial value of the parameter is chosen to be the posterior mode $\hat{\theta}_n$. In our application, the total number of Markov chain Monte Carlo iterations is chosen to be 11 000, of which the first 1000 iterations are discarded. There was no evidence of lack of convergence based on an examination of trace plots.

With the AIC-type criteria, the model is estimated by the penalized maximum likelihood method, in which θ is estimated by maximizing $\ell(\theta|y, \psi) := \sum_{\alpha=1}^{n} \log f(y_\alpha|x_\alpha; \theta) - n\lambda\theta^{\mathrm{T}} R\theta$. Konishi et al. (2004) discussed the relationship between empirical Bayes estimation and the penalized maximum likelihood method. It is shown that the maximizer of $\ell(\theta|y, \psi)$ is equivalent to the posterior mode $\hat{\theta}_n$.

Table 3 compares the average squared error ASE $= \sum_{\alpha=1}^{n} \left\{\sin(3\pi x_\alpha) - \hat{y}(x_\alpha)\right\}^2 / n$ between the true and estimated functions. The values of the mixing proportion, sample size and sample variances are set to be $\beta = 0{\cdot}8$, $n \in \{100, 200\}$, $\sigma_1 = 0{\cdot}5$ and $\sigma_2 = 1{\cdot}0$, respectively. The candidates for the smoothing parameter were chosen on an evenly-spaced grid of 20 values between $\log_{10}(\lambda) = 0$ and $\log_{10}(\lambda) = -9$. The number of basis functions ranges from 6 to 15, and the simulation results were obtained from 100 repeated Monte

Table 3. *Comparison of the average squared errors based on various criteria. Averages are given and figures in parentheses give estimated standard deviations*

		$n = 100$			$n = 200$	
	m	$\log_{10}(\lambda)$	ASE	m	$\log_{10}(\lambda)$	ASE
BPIC	6·82	−3·964	0·0633	7·11	−3·948	0·0279
	(0·71)	(0·610)	(0·0333)	(1·64)	(0·534)	(0·0161)
DIC	8·70	−4·864	0·0741	8·62	−4·797	0·0339
	(1·52)	(0·629)	(0·0371)	(1·58)	(0·685)	(0·0161)
HMBF	8·70	−4·797	0·0734	8·67	−4·741	0·0334
	(1·52)	(0·571)	(0·0370)	(1·55)	(0·518)	(0·0159)
CVPD	9·00	−3·964	0·0743	8·96	−4·210	0·0341
	(1·51)	(0·820)	(0·0373)	(1·51)	(0·720)	(0·0161)
GIC	9·70	−3·446	0·0804	9·14	−3·578	0·0319
	(2·20)	(0·988)	(0·0568)	(1·85)	(0·639)	(0·0201)
NIC	9·35	−5·888	0·0943	8·86	−5·920	0·0389
	(2·12)	(2·449)	(0·0640)	(1·57)	(2·276)	(0·0229)
AIC$_M$	9·47	−3·221	0·0742	9·02	−3·524	0·0309
	(2·04)	(0·637)	(0·0472)	(1·74)	(0·540)	(0·0193)
AIC$_C$	9·08	−3·081	0·0676	8·83	−3·454	0·0290
	(1·58)	(0·573)	(0·0382)	(1·56)	(0·502)	(0·0179)
CV	9·24	−3·329	0·0706	8·95	−3·501	0·0304
	(1·72)	(1·150)	(0·0454)	(1·63)	(0·608)	(0·0191)
GCV	9·31	−3·151	0·0715	8·89	−3·493	0·0300
	(1·87)	(0·589)	(0·0434)	(1·61)	(0·549)	(0·0186)

ASE, average squared error.

Carlo trials. The means and standard deviations of the selected smoothing parameter λ and the number of basis functions m are also given.

The model evaluated by BPIC is superior to those based on other criteria, in terms of the value of ASE. The mean value of the smoothing parameter chosen by DIC was smaller than those based on other criteria, which implies that DIC overfits. The standard deviations of λ determined by BPIC were smaller than those for DIC, indicating that BPIC is more stable than DIC. Criteria like BIC tend to choose fewer basis functions.

7. SOME OBSERVATIONS ABOUT BPIC

7·1. *Some regularity conditions*

Remark 1. *Dependent data.* The proposed criterion is based on asymptotic theory that relies on the observations being independent. Otherwise, (5) does not hold, so that BPIC cannot be applied to dependent data.

Remark 2. *Consistency.* The proposed criterion offers the consistency of γ. Consider the hierarchical Bayesian model $f(y|\theta, \psi)\pi(\theta|\psi)\pi(\psi)$, where θ includes random effects. It is known that consistency of $\gamma = (\theta^{\mathrm{T}}, \psi^{\mathrm{T}})^{\mathrm{T}}$ in (6) does not hold. To ensure the consistency of γ, we shall use the marginal loglikelihood $L(y|\psi) = \int L(y|\theta, \psi)\pi(\theta|\psi)d\theta$ for $L(y|\gamma)$ in (6), with the random effects integrated out. Note that BPIC only requests the use of marginal loglikelihood to ensure the consistency of γ, but does not restrict the inference about θ. In fact, the posterior samples of the random effects θ can be easily drawn by using the Markov chain Monte Carlo algorithm.

Consider, for example, Bayesian inference for the stochastic volatility model, which is used in modelling the time-varying volatility of financial time series. It specifies the observation and system equations as $y_t = \exp(\theta_t/2)u_t$ and $\theta_t = \mu + \phi(\theta_{t-1} - \mu) + \tau v_t$ $(t = 1, \ldots, n)$, where θ_t is the unobserved log-volatility of y_t and the independent errors are distributed as $u_t \sim N(0, 1)$ and $v_t \sim N(0, 1)$. Posterior inference for $\theta = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$ and $\psi = (\tau, \mu, \phi)^{\mathrm{T}}$ can be performed through Markov chain Monte Carlo sampling. The asymptotic bias of BPIC is obtained by specifying $\gamma = \psi$, $L(y|\gamma) = \prod_{t=1}^{n} \int f(y_t|\theta_t, \psi) f(\theta_t|\theta_{t-1}, \psi)d\theta_t$ in equation (5). Here $f(y_t|\theta_t, \psi)$ and $f(\theta_t|\theta_{t-1}, \psi)$ are the normal densities specified by the observation equation and the system equation, respectively. The asymptotic normality and the consistency of γ holds (Jensen & Petersen, 1999).

The particle filtering method (Kitagawa, 1996) is generally used to approximate the integration. The first and second derivatives of the log posterior density $\eta_n(y_t, \psi) = \log \int f(y_t|\theta_t) f(\theta_t|\theta_{t-1}, \psi)d\theta_t + \log \pi(\psi)/n$ in I_n and J_n are evaluated by numerical differentiation. Therefore, the use of BPIC for evaluating the hierarchical Bayesian model with random effects clearly loses the advantage of the full Bayesian approach with the Markov chain Monte Carlo method, which allows us to estimate the model without computing the high-dimensional integrals. This problem also occurs in the evaluation of the Bayes factors. Furthermore, when we regard the unobserved log-volatilities θ as unknown parameters to be estimated, the asymptotic theory used in AIC-type criteria then breaks down since the number of parameters exceeds the sample size. Even though the marginal loglikelihood for a hierarchical model will generally not be available in closed form and requires numerical integration, it is not a computationally intensive task thanks to the progress in computer technology.

Recently, Vaida & Blanchard (2005) proposed conditional AIC to evaluate the goodness of the mixed-effects model. They made a distinction between the model evaluation problems with a focus on population and on clusters and discussed conditional versus marginal versions of AIC for a hierarchical model.

Remark 3. *Unimodality*. The proposed criterion requires a single mode $\hat{\gamma}_n$ so that the Laplace approximation can be reasonably accurate. Therefore, the use of BPIC for the evaluation of the normal mixture model discussed in § 4·1 is not theoretically justified since the model is characterized by multimodality. However, unimodality is also needed for BIC. Furthermore, there is no theoretical justification for the use of AIC-type criteria, such as AIC, TIC, GIC and NIC, since the usual asymptotic theory breaks down. Consequently, the scope of BPIC is less limited than other model selection criteria.

7·2. *Invariance*

It is well known that DIC suffers from a lack of invariance to parameterization (Smith, 2002; Spiegelhalter et al., 2002). When considering a one-parameter exponential family of distribution $f(y_\alpha|\xi,\phi) = \exp[\{y_\alpha\xi - u(\xi)\}/\phi + v(y_\alpha,\phi)]$ with a conjugate prior, Spiegelhalter et al. (2002) showed that P_D^γ in (4) has different values depending on whether one considers the mean parameterization $\mu = E(y_\alpha|\xi,\phi) = u'(\xi)$ or the canonical parameterization ξ. We investigate this problem in terms of BPIC.

First, consider a general situation in which the p-dimensional parameter γ is transformed into another parameter $\zeta = (\zeta_1,\ldots,\zeta_p)^{\mathrm{T}} = (s_1(\gamma),\ldots,s_p(\gamma))^{\mathrm{T}} = s(\gamma)$, where each $s_j(\cdot)$ has the inverse transformation $\gamma = (q_1(\zeta),\ldots,q_p(\zeta))^{\mathrm{T}} = q(\zeta)$. Additionally, we assume that the posterior modes satisfy $\hat{\zeta}_n = s(\hat{\gamma}_n)$. Under the regularity conditions, the $p \times p$ matrices $I_n^\zeta(\hat{\zeta}_n)$ and $J_n^\zeta(\hat{\zeta}_n)$ for the new parameterization, obtained using the chain rule, are $I_n^\zeta(\hat{\zeta}_n) = Q(\hat{\zeta}_n)^{\mathrm{T}} I_n\{q(\hat{\zeta}_n)\}Q(\hat{\zeta}_n)$ and $J_n^\zeta(\hat{\zeta}_n) = Q(\hat{\zeta}_n)^{\mathrm{T}} J_n\{q(\hat{\zeta}_n)\}Q(\hat{\zeta}_n)$, where $Q(\zeta)$ is the $p \times p$ Jacobian matrix with (i,j) element $Q_{ij}(\zeta) = \partial q_i(\zeta)/\partial\zeta_j$. Although the matrices $I_n(\cdot)$ and $J_n(\cdot)$ are dependent on the choice of parameterization, under the assumption that $Q^{-1}(\hat{\zeta}_n)$ exists, we have that $\mathrm{tr}\{J_n^{-1}(\hat{\gamma}_n)I_n(\hat{\gamma}_n)\} = \mathrm{tr}\{J_n^{\zeta^{-1}}(\hat{\zeta}_n)I_n^\zeta(\hat{\zeta}_n)\}$. Thus, when the quantity $E_{\gamma|y}[\log\{L(y|\gamma)\pi(\gamma)\}]$ in (5) is invariant to reparameterization, BPIC does not depend on the choice of parameterization.

Unfortunately, if $s(\hat{\gamma}_n) = \hat{\zeta}_n$ does not hold, BPIC, like DIC, is not invariant to parameterization. Suppose we have a set of n independent observations $y = (y_1,\ldots,y_n)^{\mathrm{T}}$ generated from a Poisson distribution with $f(y_\alpha|\mu) = \exp(-\mu)\mu^{y_\alpha}/y_\alpha!$. A conjugate prior, $\mu = \exp(\xi) \sim \mathrm{Ga}(a,b)$ leads to a posterior distribution $\mu = \exp(\xi) \sim \mathrm{Ga}(a + n\bar{y}_n, b+n)$, where $\bar{y}_n = \sum_{\alpha=1}^n y_\alpha/n$. The posterior modes are $\hat{\mu}_n = (a + n\bar{y}_n - 1)/(b+n)$ and $\hat{\xi}_n = \log\{(a+n\bar{y}_n)/(b+n)\}$. Therefore, under the reparameterization $\mu = \exp(\xi)$, the asymptotic bias estimate $\hat{b}_\mu$ is not equal to $\hat{b}_\xi$. Additionally, in the context of hierarchical modelling, the likelihood is not uniquely determined, in that it becomes $L(y|\theta,\psi) = \prod_{\alpha=1}^n f(y_\alpha|\theta,\psi)$ or $L(y|\psi) = \prod_{\alpha=1}^n \int f(y_\alpha|\theta,\psi)\pi(\theta|\psi)d\theta$. Thus, BPIC can also produce a lack of invariance in the sense that it depends on the likelihood specification.

An equivalence of the empirical Bayes estimator and the penalized maximum likelihood estimator is shown by Konishi et al. (2004). Therefore, the lack of invariance problem also occurs in other model selection criteria such as the Laplace-approximated Bayes factor, GIC, NIC, AIC$_M$ and AIC$_C$, since these scores are implicitly calculated by using the posterior mode.

7·3. *Dimension of the model*

Consider a model $f(y|\gamma)$ with $\log \pi(\gamma) = O(1)$, where the prior is assumed to be dominated by the likelihood as n increases. In this case, the asymptotic bias estimate of DIC can be approximated by $n\hat{b}_\gamma^{\text{DIC}} = nP_D^\gamma/(2n) \doteq p/2$ (Spiegelhalter et al., 2002, p.591).

If we assume further that the specified parametric models contain the true model, or are similar to the true model, then $I_n(\hat{\gamma}_n) \doteq J_n(\hat{\gamma}_n)$, i.e. $\text{tr}\{J_n^{-1}(\hat{\gamma}_n)I_n(\hat{\gamma}_n)\} \doteq p$. In this special case, $n\hat{b}_\gamma \doteq -P_D/2 + p + p/2 \doteq p$, double the value for DIC; see also § 3.

7·4. *Bootstrap bias correction*

The bias correction of $\hat{\eta}$ can be also performed by using the bootstrap method (Efron & Tibshirani, 1993). Since the bootstrap analogues of η and $\hat{\eta}$ are $\eta^{(b)} = n^{-1}E_{\gamma|y^*}\{\log L(y|\gamma)\}$ and $\hat{\eta}^{(b)} = n^{-1}E_{\gamma|y^*}\{\log L(y^*|\gamma)\}$, the bootstrap bias estimator, an estimator of b_γ in (3), is given by $b_\gamma^{(b)} = E_{y^*}(\eta^{(b)} - \hat{\eta}^{(b)})$. Here y^* is the empirical distribution based on bootstrap samples that has the probability n^{-1} at each data point $y^* = (y_1^*, \ldots, y_n^*)^{\text{T}}$. 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. This approach therefore provides a direct computational way of assessing the constructed model when the analytical expression of BPIC is unavailable. The numerical approach to constructing an original information criterion has been examined by Konishi & Kitagawa (1996).

In contrast to our approach that focuses on the bias correction of $\hat{\eta}$ as an estimator of η, Steck & Jaakkola (2003) considered the bias correction of the bootstrap analogue of a scoring function, which corresponds to the bootstrap analogue $\hat{\eta}^{(b)} = n^{-1}E_{\gamma|y^*}\{\log L(y^*|\gamma)\}$ of $\hat{\eta}$ to explore model uncertainty. Since the models estimated from bootstrap samples can be significantly biased towards over-complexity, the importance of considering this bias correction is pointed out. Their result would be useful in the investigation of model uncertainty.

ACKNOWLEDGEMENT

The author would like to thank Professor A. E. Gelfand and Dr. D. J. Spiegelhalter for constructive and helpful comments. The author is grateful to Professor S. Konishi for his comments and advices. The author gratefully acknowledges the constructive suggestions of the editor, the associate editor and two anonymous referees, which have greatly improved the paper. The author is also grateful to them for pointing out several references. This study was supported in part by a Grant-in-Aid for Young Scientists.

APPENDIX 1

Calculation of the bias term of DIC

From the definition of P_D^γ, we have $E_{\gamma|y}\{\log L(y|\gamma)\} = \log L(y|\bar{\gamma}_n) - P_D^\gamma/2$. It is also known that $E_{\gamma|y}[n^{-1}\log L(y|\bar{\gamma}_n) - E_{z|\gamma}\{\log f(z|\bar{\gamma}_n)\}] = P_D^\gamma/n$ (Spiegelhalter et al., 2002, p.604). Note that the future observation z is from $f(z|\gamma)$ with $\gamma \sim \pi(\gamma|y)$. If we use this assumption and replace the expectation of y by the empirical distribution, the bias (3) is then $\hat{b}_\gamma \doteq E_{\gamma|y}[n^{-1}L(y|\bar{\gamma}_n) - E_{z|\gamma}\{\log f(z|\gamma)\}] - P_D^\gamma/(2n)$, where the definition of P_D^γ is used. This indicates that the observed data were used twice (Robert & Titterington, 2002). Taylor expansion of $\log f(z|\gamma)$ around the posterior mean $\bar{\gamma}_n$ gives $E_{\gamma|y}[E_{z|\gamma}\{\log f(z|\gamma)\}] \simeq E_{\gamma|y}[E_{z|\gamma}\{\log f(z|\bar{\gamma}_n)\}] + E_{\gamma|y}[E_{z|\gamma}\{\partial \log f(z|\bar{\gamma}_n)/\partial\gamma^{\mathrm{T}}(\gamma - \bar{\gamma}_n)\}]$. Noting that $E_{\gamma|y}(\gamma - \bar{\gamma}_n) = 0$ and substituting the resulting expression into the bias, we have $\hat{b}_\gamma \simeq E_{\gamma|y}[n^{-1}L(y|\bar{\gamma}_n) - E_{z|\gamma}\{\log f(z|\bar{\gamma}_n)\}] - P_D^\gamma/(2n) = P_D^\gamma/n - P_D^\gamma/(2n) = P_D^\gamma/(2n)$.

Appendix 2

Proof of Theorem 1

We first describe some asymptotic properties of the parameter estimators and then give two Lemmas. Let $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{0p})'$ and $\hat{\gamma}_n = (\hat{\gamma}_{n1}, \ldots, \hat{\gamma}_{np})'$ be the modes of $E_z[\log\{f(z|\gamma)\pi_0(\gamma)\}]$ and $n^{-1}\log\{L(y|\gamma)\pi(\gamma)\}$, respectively. Here $\log\pi_0(\gamma) = \lim_{n\to\infty} n^{-1}\log\pi(\gamma)$. Since $\log L(y|\gamma)$ is the sum of the independently and identically distributed random variables $\log f(y_\alpha|\gamma)$, $\alpha = 1, \ldots, n$, it follows from the law of large numbers that $n^{-1}\log\{L(y|\gamma)\pi(\gamma)\} \to E_z[\log\{f(z|\gamma)\pi_0(\gamma)\}]$ as n tends to infinity. Then $\hat{\gamma}_n \to \gamma_0$ in probability as n tends to infinity.

Consider the case $\log\pi(\gamma) = O(n)$, i.e. the prior information grows with the sample size. Then $\log\pi_0(\gamma) = O(1)$, and the prior information cannot be ignored even when the sample size n is large. Next, we consider the case where $\log\pi(\gamma) = O(1)$. Then $n^{-1}\log\pi(\gamma) \to 0$ as $n \to \infty$ and the prior information can be ignored for a sufficiently large n. In this case, the mode γ_0 is the pseudo parameter value, which minimizes the Kullback-Leibler distance from the true model $g(z)$. In each case, $\log\pi_0(\gamma)$ can be approximated by $n^{-1}\log\pi(\gamma)$ for a moderate sample size. As shown by Konishi et al. (2004), the order of the prior distribution has a large influence on the calculation of the Bayes factor. Hereafter, we restrict our attention to a proper situation in which the Hessian of $E_z[\log\{f(z|\gamma)\pi_0(\gamma)\}]$ is nonsingular at γ_0, which is uniquely determined and interior to Θ.

Lemma A1. *Assume regularity conditions similar to those of White* (1982)*; i.e., the model is sufficiently smooth and the Hessian of* $E_z[\log\{f(z|\gamma)\pi_0(\gamma)\}]$ *is nonsingular at* $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{0p})^{\mathrm{T}}$. *Then* $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$ *is asymptotically normally distributed as* $N\{0, J^{-1}(\gamma_0)I(\gamma_0)J^{-1}(\gamma_0)\}$. *Here* $\hat{\gamma}_n$ *is the mode of* $\log\{L(y|\gamma)\pi(\gamma)\}$ *and* $I(\gamma)$ *and* $J(\gamma)$ *are the* $p \times p$ *matrices defined respectively by*

$$I(\gamma) = E_z\left(\frac{\partial\eta(z,\gamma)}{\partial\gamma}\frac{\partial\eta(z,\gamma)}{\partial\gamma^{\mathrm{T}}}\right), \qquad J(\gamma) = -E_z\left(\frac{\partial^2\eta(z,\gamma)}{\partial\gamma\partial\gamma^{\mathrm{T}}}\right),$$

with $\eta(z,\gamma) = \log f(z|\gamma) + \log\pi_0(\gamma)$.

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

$$-\frac{1}{n}\frac{\partial^2\log\{L(y|\gamma)\pi(\gamma)\}}{\partial\gamma\partial\gamma^{\mathrm{T}}}\bigg|_{\gamma=\gamma_0}\sqrt{n}(\hat{\gamma}_n - \gamma_0) = \frac{1}{\sqrt{n}}\frac{\partial\log\{L(y|\gamma)\pi(\gamma)\}}{\partial\gamma}\bigg|_{\gamma=\gamma_0} + O_p\left(\frac{1}{\sqrt{n}}\right).$$

It follows from the central limit theorem that the right-hand side is asymptotically distributed as $N\{0, I(\gamma_0)\}$, while the left-hand side converges to $J(\gamma_0)\sqrt{n}(\hat{\gamma}_n - \gamma_0)$. Thus we obtain the desired result. □

Lemma A2. *Additionally to the regularity conditions of Lemma* A1, *we further assume regularity conditions for the Laplace approximation of the posterior distribution, i.e. that the posterior distribution* $\pi(\gamma|y)$ *can be approximated by the normal distribution with mean* $\hat{\gamma}_n$ *and variance* $n^{-1}J_n(\hat{\gamma}_n)$. *Then*

$$E_y\left[E_{\gamma|y}\left\{(\gamma - \gamma_0)(\gamma - \gamma_0)^{\mathrm{T}}\right\}\right] \simeq \frac{1}{n}J^{-1}(\gamma_0) + \frac{1}{n}J^{-1}(\gamma_0)I(\gamma_0)J^{-1}(\gamma_0).$$

Proof. A simple modification leads to

$$\begin{aligned}E_y\left[E_{\gamma|y}\left\{(\gamma-\gamma_0)(\gamma-\gamma_0)^{\mathrm{T}}\right\}\right] &= E_y\left[E_{\gamma|y}\left\{(\gamma-\hat{\gamma}_n+\hat{\gamma}_n-\gamma_0)(\gamma-\hat{\gamma}_n+\hat{\gamma}_n-\gamma_0)^{\mathrm{T}}\right\}\right]\\ &= E_y\left[E_{\gamma|y}\left\{(\gamma-\hat{\gamma}_n)(\gamma-\hat{\gamma}_n)^{\mathrm{T}}\right\}\right]+\frac{1}{n}E_y\left\{\sqrt{n}(\hat{\gamma}_n-\gamma_0)\sqrt{n}(\hat{\gamma}_n-\gamma_0)^{\mathrm{T}}\right\}\\ &\quad+\frac{1}{n\sqrt{n}}E_y\left\{n(\bar{\gamma}_n-\hat{\gamma}_n)\sqrt{n}(\hat{\gamma}_n-\gamma_0)^{\mathrm{T}}+\sqrt{n}(\hat{\gamma}_n-\gamma_0)n(\bar{\gamma}_n-\hat{\gamma}_n)^{\mathrm{T}}\right\},\end{aligned}$$

where $\bar{\gamma}_n$ and $\hat{\gamma}_n$ are the posterior mean and the posterior mode, respectively. When we apply the Laplace approximation to the posterior distribution, the first term can be approximated by $n^{-1}J^{-1}(\gamma_0)$. From Lemma A1, the second term is asymptotically evaluated as $n^{-1}J^{-1}(\gamma_0)I(\gamma_0)J^{-1}(\gamma_0)$. From $\partial\log\{L(y|\hat{\gamma}_n)\pi(\hat{\gamma}_n)\}/\partial\gamma=0$, the posterior mode $\hat{\gamma}_n$ can be expanded as $\hat{\gamma}_n=\bar{\gamma}_n+n^{-1}J_n^{-1}(\bar{\gamma}_n)\partial\log\{L(y|\bar{\gamma}_n)\pi(\bar{\gamma}_n)\}/\partial\gamma+O_p(n^{-2})$. Since $\hat{\gamma}_{nj}-\gamma_{0j}=O_p(n^{-1/2})$ and $\bar{\gamma}_{nj}-\hat{\gamma}_{nj}=O_p(n^{-1})$ for $j=1,\cdots,p$, the third term can be ignored given a moderate sample size. Combination of these results verifies the Lemma. When the posterior mean $\bar{\gamma}_n$ and the posterior mode $\hat{\gamma}_n$ are identical, the third term drops out completely. For the regularity conditions of the Laplace approximation, we refer to Barndorff-Nielsen & Cox (1989). □

Proof of Theorem 1. We assume the regularity conditions of Lemmas A1 and A2. We decompose the bias b_γ in (3) as $E_y(\hat{\eta}-\eta)=E_1+E_2+E_3$, where

$$\begin{aligned}E_1&=E_y\left[\frac{1}{n}E_{\gamma|y}\{\log L(y|\gamma)\}-\frac{1}{n}\log\{L(y|\gamma_0)\pi(\gamma_0)\}\right],\\ E_2&=E_y\left(\frac{1}{n}\log\{L(y|\gamma_0)\pi(\gamma_0)\}-E_z\left[\log\{f(z|\gamma_0)\pi_0(\gamma_0)\}\right]\right),\\ E_3&=E_y\left(E_z\left[\log\{f(z|\gamma_0)\pi_0(\gamma_0)\}\right]-E_z\left[E_{\gamma|y}\{\log f(z|\gamma)\}\right]\right).\end{aligned}$$

For the evaluation of E_1, consider $\partial\log\{L(y|\hat{\gamma}_n)\pi(\hat{\gamma}_n)\}/\partial\gamma=0$. Then the Taylor expansion of $\log\{L(y|\gamma_0)\pi(\gamma_0)\}$ around the posterior mode $\hat{\gamma}_n$ gives

$$\log\{L(y|\gamma_0)\pi(\gamma_0)\}=\log\{L(y|\hat{\gamma}_n)\pi(\hat{\gamma}_n)\}+n(\gamma_0-\hat{\gamma}_n)^{\mathrm{T}}J_n(\hat{\gamma}_n)(\gamma_0-\hat{\gamma}_n)/2+O_p(n^{-1/2}).$$

Thus

$$\begin{aligned}E_1&=n^{-1}E_y[E_{\gamma|y}\{\log L(y|\gamma)\}-\log\{L(y|\hat{\gamma}_n)\pi(\hat{\gamma}_n)\}]\\ &\quad+(2n)^{-1}\mathrm{tr}[E_y\{J_n(\hat{\gamma}_n)\sqrt{n}(\gamma_0-\hat{\gamma}_n)\sqrt{n}(\gamma_0-\hat{\gamma}_n)^{\mathrm{T}}\}]+O_p(n^{-3/2}).\end{aligned}$$

From Lemma A1, the variance matrix of $\sqrt{n}(\hat{\gamma}_n-\gamma_0)$ is asymptotically given by $J^{-1}(\gamma_0)I(\gamma_0)J^{-1}(\gamma_0)$. With this result and since $J_n(\hat{\gamma}_n)\to J(\gamma_0)$ in probability as $n\to\infty$, E_1 can be approximated by

$$E_1\simeq\frac{1}{n}E_y\left[E_{\gamma|y}\{\log L(y|\gamma)\}-\log\{L(y|\hat{\gamma}_n)\pi(\hat{\gamma}_n)\}\right]+\frac{1}{2n}\mathrm{tr}\left\{J^{-1}(\gamma_0)I(\gamma_0)\right\}.$$

The term E_2 can be regarded as zero, because

$$\begin{aligned}E_2&=E_y\left[\log\{f(y|\gamma_0)\pi_0(\gamma_0)\}\right]-E_z\left[\log\{f(z|\gamma_0)\pi_0(\gamma_0)\}\right]-\log\pi_0(\gamma_0)+n^{-1}\log\pi(\gamma_0)\\ &=n^{-1}\log\pi(\gamma_0)-\log\pi_0(\gamma_0)\simeq 0.\end{aligned}$$

The term E_3 can be modified as follows:

$$\begin{aligned}E_3&=E_y(E_z[\log\{f(z|\gamma_0)\pi_0(\gamma_0)\}])-E_y\{E_{\gamma|y}(E_z[\log\{f(z|\gamma)\pi_0(\gamma)\}])\}\\ &\quad+E_y[E_{\gamma|y}\{\log\pi_0(\gamma)\}].\end{aligned}$$

By expanding $\log\{f(z|\gamma)\pi_0(\gamma)\}$ around γ_0 and using $\log\pi_0(\gamma) \doteq n^{-1}\log\pi(\gamma)$, we obtain

$$E_3 \doteq \frac{1}{2}\mathrm{tr}\left(J(\gamma_0)E_y\left[E_{\gamma|y}\left\{(\gamma-\gamma_0)(\gamma-\gamma_0)^{\mathrm{T}}\right\}\right]\right)+\frac{1}{n}E_y[E_{\gamma|y}\{\log\pi(\gamma)\}].$$

Considering Lemma A2, we finally have $E_3 \doteq (2n)^{-1}\mathrm{tr}\left\{J^{-1}(\gamma_0)I(\gamma_0)\right\}+p/(2n)+E_y[E_{\gamma|y}\{\log\pi(\gamma)\}]/n$.

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

$$\begin{aligned}E_y(\hat{\eta}-\eta) &\doteq n^{-1}E_y(E_{\gamma|y}[\log\{L(y|\gamma)\pi(\gamma)\}])-n^{-1}E_y[\log\{L(y|\hat{\gamma}_n)\pi(\hat{\gamma}_n)\}]\\ &\quad+n^{-1}\mathrm{tr}\left\{J^{-1}(\gamma_0)I(\gamma_0)\right\}+p/(2n).\end{aligned}$$

Replacing the expectation of y by the empirical distribution and estimating the matrices $I(\gamma_0)$ and $J(\gamma_0)$ by $I_n(\hat{\gamma}_n)$ and $J_n(\hat{\gamma}_n)$ given in (5), we obtain the required result. □

References

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

Barndorff-nielsen, O. E. & Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. London: Chapman and Hall.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377–403.

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,.

Gelfand, A. & Ghosh, S. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

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.

Jensen, J. L. & Petersen, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist.* **27**, 514–35.

Kass, R. & Raftery, A. (1995). Bayes factors and model uncertainty. *J. Am. Statist. Assoc.* **90**, 773–95.

Kitagawa, G. (1996). Monte Carlo filter and smoother for Gaussian nonlinear state space models. *J. Comp. Graph. Statist.* **5**, 1–25.

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.

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

Ramsay, J. O. & Silverman, B. W. (1997). *Functional Data Analysis.* New York: Springer.

Richardson, S. (2002). Discussion of a paper by D. J. Spiegelhalter et al. *J. R. Statist. Soc.* B **64**, 626–7.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with Discussion). *J. R. Statist. Soc.* B **59**, 731–92.

Robert, C. P. & Titterington, D. M. (2002). Discussion of a paper by D. J. Spiegelhalter et al. *J. R. Statist. Soc.* B **64**, 621–2.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.

Smith, J. (2002). Discussion of a paper by D. J. Spiegelhalter et al. *J. R. Statist. Soc.* B **64**, 619–20.

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.

STECK, H. & JAAKKOLA, T. (2003). Bias–corrected bootstrap and model uncertainty. In *Advances in Neural Information Processing Systems 16*, Ed. S. Thrun, L. K. Saul and B. Scholkopf, pp. 521–8, Cambridge, MA: MIT Press.

STONE, M. (1974). Cross-validatory choice and assessment of statistical prediction (with Discussion). *J. R. Statist. Soc.* B **36**, 111–47.

STONE, M. (2002). Discussion of a paper by D. J. Spiegelhalter et al. *J. R. Statist. Soc.* B **64**, 621.

TAKEUCHI, K. (1976). Distributions of information statistics and criteria for adequacy of models (in Japanese). *Math. Sci.* **153**, 12–8.

VAIDA, F. & BLANCHARD, S. (2005). Conditional Akaike information for mixed–effects models. *Biometrika* **92**, 351–70.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–26.

[*Received September* 2005. *Revised July* 2006]