9

Bayesian Information Criteria

This chapter considers model selection and evaluation criteria from a Bayesian point of view. A general framework for constructing the Bayesian information criterion (BIC) is described. The BIC is also extended such that it can be applied to the evaluation of models estimated by regularization. Section 9.2 presents Akaike's Bayesian information criterion (ABIC) developed for the evaluation of Bayesian models having prior distributions with hyperparameters. In the latter half of this chapter, we consider information criteria for the evaluation of predictive distributions of Bayesian models. In particular, Section 9.3 gives examples of analytical evaluations of bias correction for linear Gaussian Bayes models. Section 9.4 describes, for general Bayesian models, how to estimate the asymptotic biases and how to perform the second-order bias correction by means of Laplace's method for integrals.

9.1 Bayesian Model Evaluation Criterion (BIC)

9.1.1 Definition of BIC

The Bayesian information criterion (BIC) or Schwarz's information criterion (SIC) proposed by Schwarz (1978) is an evaluation criterion for models defined in terms of their posterior probability [see also Akaike (1977)]. It is derived as follows.

Let $M_1, M_2, \ldots, M_r$ be r candidate models, and assume that each model M_i is characterized by a parametric distribution $f_i(x|\boldsymbol{\theta}_i)$ ($\boldsymbol{\theta}_i \in \Theta_i \subset R^{k_i}$) and the prior distribution $\pi_i(\boldsymbol{\theta}_i)$ of the k_i-dimensional parameter vector $\boldsymbol{\theta}_i$. When n observations $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$ are given, then, for the i^{th} model M_i, the marginal distribution or probability of $\boldsymbol{x}_n$ is given by

$$p_i(\boldsymbol{x}_n) = \int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i. \tag{9.1}$$

This quantity can be considered as the likelihood of the i^{th} model and is referred to as the *marginal likelihood* of the data.

According to Bayes' theorem, if we suppose that the prior probability of the i^{th} model is $P(M_i)$, the posterior probability of the i^{th} model is given by

$$P(M_i|\boldsymbol{x}_n) = \frac{p_i(\boldsymbol{x}_n)P(M_i)}{\sum_{j=1}^{r} p_j(\boldsymbol{x}_n)P(M_j)}, \qquad i = 1, 2, \ldots, r. \tag{9.2}$$

This posterior probability indicates the probability of the data being generated from the i^{th} model when data $\boldsymbol{x}_n$ are observed. Therefore, if one model is to be selected from r models, it would be natural to adopt the model that has the largest posterior probability. This principle means that the model that maximizes the numerator $p_i(\boldsymbol{x}_n)P(M_i)$ must be selected, since all models share the same denominator in (9.2).

If we further assume that the prior probabilities $P(M_i)$ are equal in all models, it follows that the model that maximizes the marginal likelihood $p_i(\boldsymbol{x}_n)$ of the data must be selected. Therefore, if an approximation to the marginal likelihood expressed in terms of an integral in (9.1) can readily be obtained, the need to compute the integral on a problem-by-problem basis will vanish, thus making the BIC suitable for use as a general model selection criterion.

The BIC is actually defined as the natural logarithm of the integral multiplied by -2, and we have

$$\begin{aligned} -2\log p_i(\boldsymbol{x}_n) &= -2\log\left\{\int f_i(\boldsymbol{x}_n|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i\right\} \\ &\approx -2\log f_i(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}_i) + k_i\log n, \end{aligned} \tag{9.3}$$

where $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimator of the k_i-dimensional parameter vector $\boldsymbol{\theta}_i$ of the model $f_i(x|\boldsymbol{\theta}_i)$. Consequently, from the r models that are to be evaluated using the maximum likelihood method, the model that minimizes the value of BIC can be selected as the optimal model for the data.

Thus, even under the assumption that all models have equal prior probabilities, the posterior probability obtained by using the information from the data serves to contrast the models and helps to identify the model that generated the data. We see in the next section that the BIC can be obtained by approximating the integral using Laplace's method.

Bayes factors. For simplicity, let us compare two models, say M_1 and M_2. When the data produce the posterior probabilities $P(M_i|\boldsymbol{x}_n)$ $(i = 1, 2)$, the posterior odds in favor of model M_1 against model M_2 are

$$\frac{P(M_1|\boldsymbol{x}_n)}{P(M_2|\boldsymbol{x}_n)} = \frac{p_1(\boldsymbol{x}_n)}{p_2(\boldsymbol{x}_n)}\frac{P(M_1)}{P(M_2)}. \tag{9.4}$$

Then the ratio

$$B_{12} = \frac{p_1(\boldsymbol{x}_n)}{p_2(\boldsymbol{x}_n)} = \frac{\int f_1(\boldsymbol{x}_n|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{\int f_2(\boldsymbol{x}_n|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2} \tag{9.5}$$

is defined as the *Bayes factor*.

Akaike (1983a) showed that model comparisons based on the AIC are asymptotically equivalent to those based on Bayes factors. Kass and Raftery (1995) commented that from a Bayesian viewpoint this is true only if the precision of the prior is comparable to that of the likelihood, but not in the more usual situation where prior information is limited relative to the information provided by the data. For Bayes factors, we refer to Kass and Raftery (1995), O'Hagan (1995), and Berger and Pericchi (2001) and references given therein.

9.1.2 Laplace Approximation for Integrals

In order to explain the Laplace approximation method [Tierney and Kadane (1986), Davison (1986), and Barndorff-Nielsen and Cox (1989, p. 169)], we consider the approximation of a simple integral given by

$$\int \exp\{nq(\boldsymbol{\theta})\}d\boldsymbol{\theta}, \tag{9.6}$$

where $\boldsymbol{\theta}$ is a p-dimensional parameter vector. Notice that in the Laplace approximation of an actual likelihood function, the form of $q(\boldsymbol{\theta})$ also changes as the number n of observations increases.

The basic concept underlying the Laplace approximation takes advantage of the fact that when the number n of observations is large, the integrand is concentrated in a neighborhood of the mode $\hat{\boldsymbol{\theta}}$ of $q(\boldsymbol{\theta})$, and consequently, the value of the integral depends solely on the behavior of the integrand in that neighborhood of $\hat{\boldsymbol{\theta}}$.

It follows from $\partial q(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}$ that the Taylor expansion of $q(\boldsymbol{\theta})$ around $\hat{\boldsymbol{\theta}}$ yields the following:

$$q(\boldsymbol{\theta}) = q(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T J_q(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \cdots, \tag{9.7}$$

where

$$J_q(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 q(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{9.8}$$

Substituting the Taylor expansion of $q(\boldsymbol{\theta})$ into (9.6) gives

$$\begin{aligned}\int \exp\left[n\left\{q(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T J_q(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \cdots\right\}\right] d\boldsymbol{\theta} \qquad (9.9)\\ \approx \exp\left\{nq(\hat{\boldsymbol{\theta}})\right\}\int \exp\left\{-\frac{n}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T J_q(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta}.\end{aligned}$$

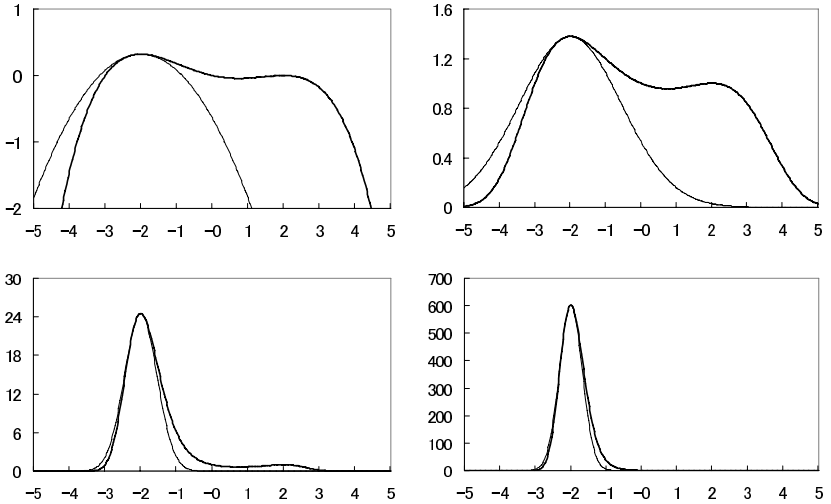

Fig. 9.1. Laplace approximation. Top left: $q(\theta)$ and its quadratic function approximation. Top right, bottom left, and bottom right: $\exp\{nq(\theta)\}$ and Laplace approximations with n=1, 10, and 20, respectively.

By noting the fact that the p-dimensional random vector $\boldsymbol{\theta}$ follows the p-variate normal distribution with mean vector $\hat{\boldsymbol{\theta}}$ and variance covariance matrix $n^{-1}J_q(\hat{\boldsymbol{\theta}})^{-1}$, calculation of the integral on the right-hand side of (9.9) yields

$$\int \exp\left\{-\frac{n}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T J_q(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} = \frac{(2\pi)^{p/2}}{n^{p/2}|J_q(\hat{\boldsymbol{\theta}})|^{1/2}}. \tag{9.10}$$

Therefore, we obtain the following Laplace approximation of the integral (9.6).

Laplace approximation of integrals. Let $q(\boldsymbol{\theta})$ be a real-valued function of a p-dimensional parameter vector $\boldsymbol{\theta}$, and let $\hat{\boldsymbol{\theta}}$ be the mode of $q(\boldsymbol{\theta})$. Then the Laplace approximation of the integral is given by

$$\int \exp\{nq(\boldsymbol{\theta})\}d\boldsymbol{\theta} \approx \frac{(2\pi)^{p/2}}{n^{p/2}|J_q(\hat{\boldsymbol{\theta}})|^{1/2}} \exp\left\{nq(\hat{\boldsymbol{\theta}})\right\}, \tag{9.11}$$

where $J_q(\hat{\boldsymbol{\theta}})$ is defined by (9.8).

Example 1 (Laplace approximation)@ Figure 9.1 shows how Laplace's method for integrals works. The upper left graph illustrates a suitably defined function $q(\boldsymbol{\theta})$ and its approximation in terms of its Taylor expansion. The curve with two peaks shown in bold lines represents the function $q(\boldsymbol{\theta})$, and the thin line indicates its approximation by the Taylor series expansion up

Table 9.1. The integral of the function given in Figure 9.1 and its Laplace approximation.

n	1	10	20	50
Integral	398.05	1678.76	26378.39	240282578
Laplace approximation	244.51	1403.40	24344.96	240282578
Relative errors	0.386	0.164	0.077	0

to the second term. In this graph, only the left peak of the two peaks is approximated, and it can hardly be considered a good approximation. The other three graphs show the integrand $\exp\{nq(\boldsymbol{\theta})\}$ and approximations to it. The upper right, lower left, and lower right graphs represent the cases $n = 1$, 10, and 20, in the indicated order. The graph for $n = 1$ fails to describe the peak on the right side. However, as n increases to $n = 10$ and $n = 20$, the right peak vanishes rapidly, indicating that making use of the Taylor series expansion yields a good approximation. Therefore, it is clear that, when the value of n is large, this method provides a good approximation to the integral.

Table 9.1 shows the integral of the function $\exp\{nq(\boldsymbol{\theta})\}$ given in Figure 9.1, its Laplace approximation, and the relative error (= |true value − approximation|/|true value|). In this case, the relative error is as large as 0.386 when $n = 1$, but it diminishes as n increases, and the relative error becomes 0 when $n = 50$.

9.1.3 Derivation of the BIC

The marginal likelihood or the marginal distribution of data $\boldsymbol{x}_n$ can be approximated by using Laplace's method for integrals. In this section, we drop the notational dependence on the model M_i and represent the marginal likelihood of (9.1) as

$$p(\boldsymbol{x}_n) = \int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{9.12}$$

where $\boldsymbol{\theta}$ is a p-dimensional parameter vector. This equation may be rewritten as

$$\begin{aligned} p(\boldsymbol{x}_n) &= \int \exp\left\{\log f(\boldsymbol{x}_n|\boldsymbol{\theta})\right\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \int \exp\left\{\ell(\boldsymbol{\theta})\right\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}, \end{aligned} \tag{9.13}$$

where $\ell(\boldsymbol{\theta})$ is the log-likelihood function $\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{x}_n|\boldsymbol{\theta})$.

The Laplace approximation takes advantage of the fact that when the number n of observations is sufficiently large, the integrand is concentrated in

a neighborhood of the mode of $\ell(\boldsymbol{\theta})$ or, in this case, in a neighborhood of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, and that the value of the integral depends on the behavior of the function in this neighborhood. Since $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$ holds for the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, the Taylor expansion of the log-likelihood function $\ell(\boldsymbol{\theta})$ around $\hat{\boldsymbol{\theta}}$ yields

$$\ell(\boldsymbol{\theta}) = \ell(\hat{\boldsymbol{\theta}}) - \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \cdots, \tag{9.14}$$

where

$$J(\hat{\boldsymbol{\theta}}) = -\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\frac{1}{n}\frac{\partial^2 \log f(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{9.15}$$

Similarly, we can expand the prior distribution $\pi(\boldsymbol{\theta})$ in a Taylor series around the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ as

$$\pi(\boldsymbol{\theta}) = \pi(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \cdots. \tag{9.16}$$

Substituting (9.14) and (9.16) into (9.13) and simplifying the results lead to the approximation of the marginal likelihood as follows:

$$\begin{aligned} p(\boldsymbol{x}_n) &= \int \exp\left\{\ell(\hat{\boldsymbol{\theta}}) - \frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \cdots\right\} \\ &\quad \times \left\{\pi(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} + \cdots\right\} d\boldsymbol{\theta} \\ &\approx \exp\left\{\ell(\hat{\boldsymbol{\theta}})\right\} \pi(\hat{\boldsymbol{\theta}}) \int \exp\left\{-\frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta}. \end{aligned} \tag{9.17}$$

Here we used the fact that $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$ in probability with order $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{-1/2})$ and also that the following equation holds:

$$\int (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \exp\left\{-\frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} = \mathbf{0}. \tag{9.18}$$

In (9.17), integrating with respect to the parameter vector $\boldsymbol{\theta}$ yields

$$\int \exp\left\{-\frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} = (2\pi)^{p/2} n^{-p/2} |J(\hat{\boldsymbol{\theta}})|^{-1/2}, \tag{9.19}$$

since the integrand is the density function of the p-dimensional normal distribution with mean vector $\hat{\boldsymbol{\theta}}$ and variance covariance matrix $J^{-1}(\hat{\boldsymbol{\theta}})/n$. Consequently, when the sample size n becomes large, it is clear that the marginal likelihood can be approximated as

$$p(\boldsymbol{x}_n) \approx \exp\left\{\ell(\hat{\boldsymbol{\theta}})\right\} \pi(\hat{\boldsymbol{\theta}}) (2\pi)^{p/2} n^{-p/2} |J(\hat{\boldsymbol{\theta}})|^{-1/2}. \tag{9.20}$$

Taking the logarithm of this expression and multiplying it by -2, we obtain

$$-2\log p(\boldsymbol{x}_n) = -2\log\left\{\int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\right\} \tag{9.21}$$
$$\approx -2\ell(\hat{\boldsymbol{\theta}}) + p\log n + \log|J(\hat{\boldsymbol{\theta}})| - p\log(2\pi) - 2\log\pi(\hat{\boldsymbol{\theta}}).$$

Then the following model evaluation criterion BIC can be obtained by ignoring terms with order less than $O(1)$ with respect to the sample size n.

Bayesian information criterion (BIC). Let $f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}})$ be a statistical model estimated by the maximum likelihood method. Then the Bayesian information criterion BIC is given by

$$\text{BIC} = -2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) + p\log n. \tag{9.22}$$

From the above argument, it can be seen that, BIC is an evaluation criterion for models estimated by using the maximum likelihood method and that the criterion is obtained under the condition that the sample size n is made sufficiently large. We also see that it was obtained by approximating the marginal likelihood associated with the posterior probability of the model by Laplace's method for integrals and that it is not an information criterion, leading to an unbiased estimation of the K-L information.

We shall now consider how to extend the BIC to an evaluation criterion that permits the evaluation of models estimated by the regularization method described in Subsection 5.2.4. In the next section, we derive a model evaluation criterion that represents an extension of the BIC through the application of Laplace approximation.

Minimum description length (MDL). Rissanen (1978, 1989) proposed a model evaluation criterion (MDL) based on the concept of minimum description length in transmitting a set of data by coding using a family of probability models $\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset \boldsymbol{R}^p\}$.

Assume that the data $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$ are obtained from $f(x|\boldsymbol{\theta})$. Since the parameter vector $\boldsymbol{\theta}$ of the model is unknown, we first encode $\boldsymbol{\theta}$ and send it to the receiver, and then encode and send the data $\boldsymbol{x}_n$ by using the probability distribution $f(x|\boldsymbol{\theta})$ specified by $\boldsymbol{\theta}$. Then, given the parameter vector $\boldsymbol{\theta}$, the description length necessary for encoding the data is $-\log f(\boldsymbol{x}_n|\boldsymbol{\theta})$ and the total description length is defined by $-\log f(\boldsymbol{x}_n|\boldsymbol{\theta})$ plus the description length of the probability distribution model. The probability distribution model that minimizes this total description length is such a model that can encode the data $\boldsymbol{x}_n$ in minimum length.

If the parameter is a real number, an infinite description length is necessary for exact coding. Therefore, we consider encoding the parameter by discretizing through segmentation of the parameter space $\Theta \in \boldsymbol{R}^p$ into infinitesimal cubes of size δ. Then the total description length depends on the

value of δ, and its minimum can be approximated as

$$\ell(\boldsymbol{x}_n) = -\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) + \frac{p}{2}\log n - \frac{p}{2}\log 2\pi + \log \int \sqrt{|J(\boldsymbol{\theta})|}d\boldsymbol{\theta} + O(n^{-1/2}), \tag{9.23}$$

where $J(\boldsymbol{\theta})$ is Fisher's information matrix. By considering terms up to order $O(\log n)$, the minimum description length is defined as

$$\text{MDL} = -\log f(\boldsymbol{x}_n|\boldsymbol{\theta}) + \frac{p}{2}\log n. \tag{9.24}$$

The first term on the right-hand side is the description length in sending the data $\boldsymbol{x}_n$ by using the probability distribution $f(x|\hat{\boldsymbol{\theta}})$ specified by the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ as the encoding function, and the second term is the description length for encoding the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ with accuracy $\delta = O(n^{-1/2})$. In any case, it is interesting that the minimum description length MDL coincides with the BIC that was derived in terms of the posterior probability of the model within the Bayesian framework.

9.1.4 Extension of the BIC

Let $f(x|\hat{\boldsymbol{\theta}}_P)$ be a statistical model estimated by the regularization method for the parametric model $f(x|\boldsymbol{\theta})$ ($\boldsymbol{\theta} \in \Theta \subset R^p$), where $\hat{\boldsymbol{\theta}}_P$ is an estimator of dimension p obtained by maximizing the penalized log-likelihood function

$$\ell_\lambda(\boldsymbol{\theta}) = \log f(\boldsymbol{x}_n|\boldsymbol{\theta}) - \frac{n\lambda}{2}\boldsymbol{\theta}^T K\boldsymbol{\theta}, \tag{9.25}$$

and where K is a $p\times p$ specified matrix with rank $d = p-k$ [for the typical form of K, see (5.135)]. Our objective here is to obtain a criterion for evaluation and selection of a statistical model $f(x|\hat{\boldsymbol{\theta}}_P)$, from a Bayesian perspective.

The penalized log-likelihood function in (9.25) can be rewritten as

$$\begin{aligned}\ell_\lambda(\boldsymbol{\theta}) &= \log f(\boldsymbol{x}_n|\boldsymbol{\theta}) + \log\left\{\exp\left(-\frac{n\lambda}{2}\boldsymbol{\theta}^T K\boldsymbol{\theta}\right)\right\} \\ &= \log\left\{f(\boldsymbol{x}_n|\boldsymbol{\theta})\exp\left(-\frac{n\lambda}{2}\boldsymbol{\theta}^T K\boldsymbol{\theta}\right)\right\}.\end{aligned} \tag{9.26}$$

By considering the exponential term on the right-hand side as a p-dimensional degenerate normal distribution with mean vector $\mathbf{0}$ and singular variance covariance matrix $(n\lambda K)^-$ and adding a constant term to yield a density function, we obtain

$$\pi(\boldsymbol{\theta}|\lambda) = (2\pi)^{-d/2}(n\lambda)^{d/2}|K|_+^{1/2}\exp\left(-\frac{n\lambda}{2}\boldsymbol{\theta}^T K\boldsymbol{\theta}\right), \tag{9.27}$$

where $|K|_+$ denotes the product of nonzero eigenvalues of the specified matrix K with rank d. This distribution can be thought of as a prior distribution in which the smoothing parameter λ is a hyperparameter.

Given the data distribution $f(\boldsymbol{x}_n|\boldsymbol{\theta})$ and the prior distribution $\pi(\boldsymbol{\theta}|\lambda)$ with hyperparameter λ, the marginal likelihood of the model is defined by

$$p(\boldsymbol{x}_n|\lambda) = \int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta}. \tag{9.28}$$

When the prior distribution of $\boldsymbol{\theta}$ is given by the p-dimensional normal distribution in (9.27), this marginal likelihood can be rewritten as

$$\begin{aligned} p(\boldsymbol{x}_n|\lambda) &= \int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta} \\ &= \int \exp\left[n \times \frac{1}{n}\log\left\{f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\lambda)\right\}\right] d\boldsymbol{\theta} \\ &= \int \exp\left\{nq(\boldsymbol{\theta}|\lambda)\right\} d\boldsymbol{\theta}, \end{aligned} \tag{9.29}$$

where

$$\begin{aligned} q(\boldsymbol{\theta}|\lambda) &= \frac{1}{n}\log\left\{f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\lambda)\right\} \\ &= \frac{1}{n}\left\{\log f(\boldsymbol{x}_n|\boldsymbol{\theta}) + \log\pi(\boldsymbol{\theta}|\lambda)\right\} \\ &= \frac{1}{n}\left\{\log f(\boldsymbol{x}_n|\boldsymbol{\theta}) - \frac{n\lambda}{2}\boldsymbol{\theta}^T K\boldsymbol{\theta}\right\} \\ &\quad - \frac{1}{2n}\left\{d\log(2\pi) - d\log(n\lambda) - \log|K|_+\right\}. \end{aligned} \tag{9.30}$$

We note here that the mode, $\hat{\boldsymbol{\theta}}_P$, of $q(\boldsymbol{\theta}|\lambda)$ in the above equation coincides with a solution obtained by maximizing the penalized log-likelihood function (9.25). By approximating it using Laplace's method for integrals in (9.11), we have

$$\int \exp\{nq(\boldsymbol{\theta})\}d\boldsymbol{\theta} \approx \frac{(2\pi)^{p/2}}{n^{p/2}|J_\lambda(\hat{\boldsymbol{\theta}}_P)|^{1/2}} \exp\left\{nq(\hat{\boldsymbol{\theta}}_P)\right\}. \tag{9.31}$$

Taking the logarithm of this expression and multiplying it by -2, we obtain the following model evaluation criterion [Konishi et al. (2004)]:

Generalized Bayesian information criterion (GBIC). Suppose that the model $f(\boldsymbol{x}_n|\boldsymbol{\theta}_P)$ is constructed by maximizing the penalized log-likelihood function (9.25). Then the model evaluation criterion based on a Bayesian approach is given by

$$\begin{aligned} \mathrm{GBIC} = &-2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}_P) + n\lambda\hat{\boldsymbol{\theta}}_P^T K\hat{\boldsymbol{\theta}}_P + (p-d)\log n \\ &+ \log|J_\lambda(\hat{\boldsymbol{\theta}}_P)| - d\log\lambda - \log|K|_+ - (p-d)\log(2\pi), \end{aligned} \tag{9.32}$$

where K is a $p \times p$ specified matrix of rank d, $|K|_+$ is the product of the d nonzero eigenvalues of K, and

$$J_\lambda(\hat{\boldsymbol{\theta}}_P) = -\frac{1}{n}\frac{\partial^2 \log f(\boldsymbol{x}_n|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\hat{\boldsymbol{\theta}}_P} + \lambda K. \tag{9.33}$$

Since the model evaluation criterion GBIC can be used for the selection of a smoothing parameter λ, we select λ that minimizes the GBIC as the optimal smoothing parameter. This results in the selection of an optimal model from a family of models characterized by smoothing parameters.

By interpreting the regularization method based on the above argument from a Bayesian point of view, it can be seen that the regularized estimator agrees with the estimate that is obtained through maximization (mode) of the following posterior probability, depending on the value of the smoothing parameter:

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_n;\lambda) = \frac{f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\lambda)}{\displaystyle\int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta}}, \tag{9.34}$$

where $\pi(\boldsymbol{\theta}|\lambda)$ is the density function resulting from (9.27) as a prior probability of the p-dimensional parameter $\boldsymbol{\theta}$ for the model $f(\boldsymbol{x}_n|\boldsymbol{\theta})$. For the Bayesian justification of the maximum penalized likelihood approach, we refer to Silverman (1985) and Wahba (1990, Chapter 1).

The use of Laplace's method for integrals has been extensively investigated as a useful tool for approximating Bayesian predictive distributions, Bayes factors, and Bayesian model selection criteria [Davison (1986), Clarke and Barron (1994), Kass and Wasserman (1995), Kass and Raftery (1995), O'Hagan (1995), Konishi and Kitagawa (1996), Neath and Cavanaugh (1997), Pauler (1998), Lanterman (2001), and Konishi et al. (2004)].

Example 2 (Nonlinear regression models) Suppose that n observations $\{(\boldsymbol{x}_\alpha, y_\alpha);\ \alpha = 1, 2, \ldots, n\}$ are obtained in terms of a p-dimensional vector of explanatory variables $\boldsymbol{x}$ and a response variable Y. We assume the regression model based on the basis expansion described in Section 6.1 as follows:

$$\begin{aligned} y_\alpha &= \sum_{i=1}^{m} w_i b_i(\boldsymbol{x}_\alpha) + \varepsilon_\alpha \\ &= \boldsymbol{w}^T\boldsymbol{b}(\boldsymbol{x}_\alpha) + \varepsilon_\alpha, \qquad\qquad \alpha = 1, 2, \ldots, n, \end{aligned} \tag{9.35}$$

where $\boldsymbol{b}(\boldsymbol{x}_\alpha) = (b_1(\boldsymbol{x}_\alpha), \ldots, b_m(\boldsymbol{x}_\alpha))^T$ and ε_α, $\alpha = 1, 2, \ldots, n$, are independently and normally distributed with mean zero and variance σ^2. Then the regression model based on the basis expansion can be expressed in terms of the probability density function

$$f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\{y_\alpha - \boldsymbol{w}^T\boldsymbol{b}(\boldsymbol{x}_\alpha)\}^2}{2\sigma^2}\right], \tag{9.36}$$

where $\boldsymbol{\theta} = (\boldsymbol{w}^T, \sigma^2)^T$.

If we estimate the parameter vector $\boldsymbol{\theta}$ of the model by maximizing the penalized log-likelihood function (9.25), the estimators for $\boldsymbol{w}$ and σ^2 are respectively given by

$$\hat{\boldsymbol{w}} = (B^TB + n\lambda\hat{\sigma}^2K)^{-1}B^T\boldsymbol{y}, \quad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - B\hat{\boldsymbol{w}})^T(\boldsymbol{y} - B\hat{\boldsymbol{w}}), \tag{9.37}$$

where B is an $n \times m$ basis function matrix given by $B = (\boldsymbol{b}(\boldsymbol{x}_1), \boldsymbol{b}(\boldsymbol{x}_2), \cdots, \boldsymbol{b}(\boldsymbol{x}_n))^T$ (see Section 6.1). Then the probability density function $f(y_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{\theta}}_P)$ in which the parameters $\boldsymbol{\theta} = (\boldsymbol{w}^T, \sigma^2)^T$ in (9.36) are replaced with their estimators $\hat{\boldsymbol{\theta}}_P = (\hat{\boldsymbol{w}}^T, \hat{\sigma}^2)^T$ is the resulting statistical model.

By applying the GBIC in (9.32), the model evaluation criterion for the statistical model $f(y_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{\theta}}_P)$ estimated by the regularization method is given by

$$\begin{aligned} \mathrm{GBIC} = {} & n\log\hat{\sigma}^2 + n\lambda\hat{\boldsymbol{w}}^TK\hat{\boldsymbol{w}} + n + n\log(2\pi) \\ & + (m+1-d)\log n + \log|J_\lambda(\hat{\boldsymbol{\theta}}_P)| - \log|K|_+ \\ & - d\log\lambda - (m+1-d)\log(2\pi), \end{aligned} \tag{9.38}$$

where the $(m+1) \times (m+1)$ matrix $J_\lambda(\hat{\boldsymbol{\theta}}_P)$ is

$$J_\lambda(\hat{\boldsymbol{\theta}}_P) = \frac{1}{n\hat{\sigma}^2}\begin{bmatrix} B^TB + n\lambda\hat{\sigma}^2K & \frac{1}{\hat{\sigma}^2}B^T\boldsymbol{e} \\ \frac{1}{\hat{\sigma}^2}\boldsymbol{e}^TB & \frac{n}{2\hat{\sigma}^2} \end{bmatrix} \tag{9.39}$$

with the n-dimensional residual vector

$$\boldsymbol{e} = \left(y_1 - \hat{\boldsymbol{w}}^T\boldsymbol{b}(\boldsymbol{x}_1), y_2 - \hat{\boldsymbol{w}}^T\boldsymbol{b}(\boldsymbol{x}_2), \cdots, y_n - \hat{\boldsymbol{w}}^T\boldsymbol{b}(\boldsymbol{x}_n)\right)^T, \tag{9.40}$$

and K is an $m \times m$ specified matrix of rank d and $|K|_+$ is the product of the d nonzero eigenvalues of K.

Example 3 (Nonlinear logistic regression models) Let $y_1, \ldots, y_n$ be independent binary random variables with

$$\Pr(Y_\alpha = 1|\boldsymbol{x}_\alpha) = \pi(\boldsymbol{x}_\alpha) \quad \text{and} \quad \Pr(Y_\alpha = 0|\boldsymbol{x}_\alpha) = 1 - \pi(\boldsymbol{x}_\alpha), \tag{9.41}$$

where $\boldsymbol{x}_\alpha$ are p-dimensional explanatory variables. We model $\pi(\boldsymbol{x}_\alpha)$ by

$$\log\left\{\frac{\pi(\boldsymbol{x}_\alpha)}{1 - \pi(\boldsymbol{x}_\alpha)}\right\} = w_0 + \sum_{i=1}^{m} w_i b_i(\boldsymbol{x}_\alpha), \tag{9.42}$$

where $\{b_1(\boldsymbol{x}_\alpha), \ldots, b_m(\boldsymbol{x}_\alpha)\}$ are basis functions. Estimating the $(m+1)$-dimensional parameter vector $\boldsymbol{w} = (w_0, w_1, \ldots, w_m)^T$ by maximization of the penalized log-likelihood function (9.25) yields the model

$$f(y_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}) = \hat{\pi}(\boldsymbol{x}_\alpha)^{y_\alpha}\{1 - \hat{\pi}(\boldsymbol{x}_\alpha)\}^{1-y_\alpha}, \qquad \alpha = 1, \ldots, n, \tag{9.43}$$

where $\hat{\pi}(\boldsymbol{x}_\alpha)$ is the estimated conditional probability given by

$$\hat{\pi}(\boldsymbol{x}_\alpha) = \frac{\exp\left\{\hat{\boldsymbol{w}}^T \boldsymbol{b}(\boldsymbol{x}_\alpha)\right\}}{1 + \exp\left\{\hat{\boldsymbol{w}}^T \boldsymbol{b}(\boldsymbol{x}_\alpha)\right\}}. \tag{9.44}$$

By using the GBIC in (9.32), we obtain the model evaluation criterion for the model $f(y_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}})$ estimated by the regularization method as follows:

$$\begin{aligned}\text{GBIC} = 2\sum_{\alpha=1}^{n}\left[\log\left\{1 + \exp\left(\hat{\boldsymbol{w}}^T b(\boldsymbol{x}_\alpha)\right)\right\} - y_\alpha \hat{\boldsymbol{w}}^T b(\boldsymbol{x}_\alpha)\right] + n\lambda \hat{\boldsymbol{w}}^T K \hat{\boldsymbol{w}} \\ - (m+1-d)\log(2\pi/n) + \log|Q_\lambda^{(L)}(\hat{\boldsymbol{w}})| - \log|K|_+ - d\log\lambda,\end{aligned} \tag{9.45}$$

where $Q_\lambda^{(\mathrm{L})}(\hat{\boldsymbol{w}}) = B^{\mathrm{T}} \Gamma^{(\mathrm{L})} B/n + \lambda K$ with

$$\Gamma_{\alpha\alpha}^{(\mathrm{L})} = \frac{\exp\{\hat{\boldsymbol{w}}^T \boldsymbol{b}(\boldsymbol{x}_\alpha)\}}{[1 + \exp\{\hat{\boldsymbol{w}}^T \boldsymbol{b}(\boldsymbol{x}_\alpha)\}]^2} \tag{9.46}$$

as the α^{th} diagonal element of $\Gamma^{(\mathrm{L})}$.

Example 4 (Numerical results) For illustration, binary observations $y_1, \ldots, y_{100}$ were generated from the true models

$$\begin{aligned}&(1)\quad \Pr(Y = 1|x) = \frac{1}{1 + \exp\{-\cos(1.5\pi x)\}}, \\ &(2)\quad \Pr(Y = 1|x) = \frac{1}{1 + \exp\{-\exp(-3x)\cos(3\pi x)\}},\end{aligned} \tag{9.47}$$

where the design points are uniformly distributed in $[0, 1]$. We fitted the nonlinear logistic regression model based on B-splines discussed in Subsection 6.2.1 to the simulated data. The number of basis functions and the value of a smoothing parameter were selected as $m = 17$ and $\lambda = 0.251$ for case (1), and $m = 6$ and $\lambda = 6.31 \times 10^{-5}$ for case (2). Figure 9.2 shows the true and estimated conditional probability functions; the circles indicate the data.

9.2 Akaike's Bayesian Information Criterion (ABIC)

Let $f(\boldsymbol{x}_n|\boldsymbol{\theta})$ be the data distribution of $\boldsymbol{x}_n$ with respect to a parametric model $\{f(x|\boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta \subset \mathrm{R}^p\}$, and let $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ be the prior distribution of the p-dimensional parameter vector $\boldsymbol{\theta}$ with q-dimensional hyperparameter vector $\boldsymbol{\lambda}$

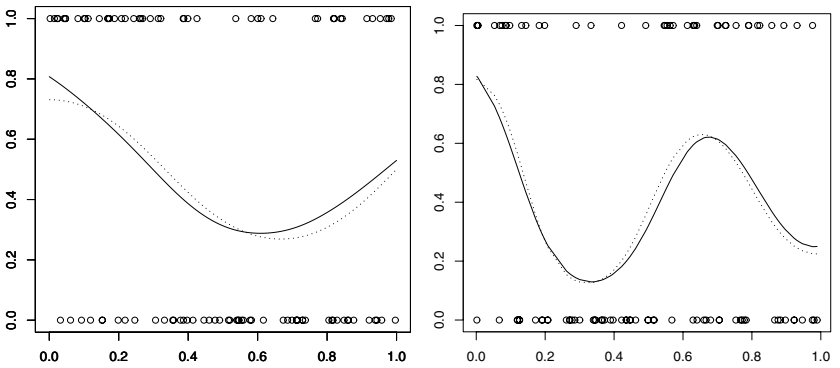

Fig. 9.2. *B*-spline logistic regression; the true (dashed line) and estimated (solid line) conditional probability functions. Case (1): left, case (2): right.

($\in \Lambda \subset \mathrm{R}^q$). Then the marginal distribution or marginal likelihood of the data $\boldsymbol{x}_n$ is given by

$$p(\boldsymbol{x}_n|\boldsymbol{\lambda}) = \int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})d\boldsymbol{\theta}. \tag{9.48}$$

If the marginal distribution $p(\boldsymbol{x}_n|\boldsymbol{\lambda})$ of the Bayes model is considered to be a parametric model with hyperparameter $\boldsymbol{\lambda}$, then evaluation of the model can be considered within the framework of the AIC, and the criterion is given by

$$\begin{aligned} \mathrm{ABIC} &= -2\log\left\{\max_{\boldsymbol{\lambda}} p(\boldsymbol{x}_n|\boldsymbol{\lambda})\right\} + 2q \\ &= -2\max_{\boldsymbol{\lambda}} \log\left\{\int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})d\boldsymbol{\theta}\right\} + 2q. \end{aligned} \tag{9.49}$$

This criterion for model evaluation, originally proposed by Akaike (1980b), is referred to as *Akaike's Bayesian information criterion* (ABIC).

According to the Bayesian approach based on the ABIC, the value of the hyperparameter $\boldsymbol{\lambda}$ of a Bayes model can be estimated by maximizing either the marginal likelihood $p(\boldsymbol{x}_n|\boldsymbol{\lambda})$ or the marginal log-likelihood $\log p(\boldsymbol{x}|\boldsymbol{\lambda})$. In other words, the hyperparameter $\boldsymbol{\lambda}$ can be regarded as being estimated using the maximum likelihood method in terms of $p(\boldsymbol{x}_n|\boldsymbol{\lambda})$. If there are two or more Bayes models characterized by a hyperparameter and if it is necessary to compare their goodness of fit, it suffices to select the model that minimizes the ABIC.

If the hyperparameter estimated in this way is denoted by $\hat{\boldsymbol{\lambda}}$, then we can determine the posterior distribution of the parameter $\boldsymbol{\theta}$ in terms of the prior distribution $\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$ as

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_n;\hat{\boldsymbol{\lambda}}) = \frac{f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})}{\int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})d\boldsymbol{\theta}}. \tag{9.50}$$

In general, the mode of the posterior distribution (9.50) is used in practical applications, i.e., the value $\hat{\boldsymbol{\theta}}$ that maximizes $\pi(\boldsymbol{\theta}|\boldsymbol{x}_n;\hat{\boldsymbol{\lambda}}) \propto f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\lambda}})$.

The ultimate objective of modeling using the information criterion ABIC is not to estimate the hyperparameter $\boldsymbol{\lambda}$. Rather, the objective is to estimate the parameter $\boldsymbol{\theta}$ or the distribution of data $\boldsymbol{x}_n$ specified by the parameters. Inferences performed through the minimization of the ABIC can be thought of as a two-step estimation process consisting first of the estimation of a hyperparameter and the selection of a model using the maximum likelihood method on the data distribution $p(\boldsymbol{x}_n|\boldsymbol{\lambda})$, which is given as a marginal distribution, and second, the determination of an estimate of $\boldsymbol{\theta}$ by maximizing the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x}_n;\hat{\boldsymbol{\lambda}})$ of the parameter $\boldsymbol{\theta}$.

The ABIC minimization method was originally used for the development of seasonal adjustments of econometric data [Akaike (1980b, 1980c) and Akaike and Ishiguro (1980a, 1980b, 1980c)]. Subsequently, it has been used for the development of a variety of new models, including cohort analyses [Nakamura (1986)], binary regression models [Sakamoto and Ishiguro (1988)], and earth tide analyses [Ishiguro and Sakamoto (1984)].

Akaike (1987) showed the relationship between, AIC and ABIC by introducing the Bayesian approach to control the occurrence of improper solutions in normal theory maximum likelihood factor analysis [see also Martin and McDonald (1975)].

9.3 Bayesian Predictive Distributions

Predictive distributions based on a Bayesian approach are constructed using a parametric model $\{f(x|\boldsymbol{\theta});\boldsymbol{\theta} \in \Theta \subset \mathrm{R}^p\}$ that defines the data distribution and a prior distribution $\pi(\boldsymbol{\theta})$ for the parameter vector $\boldsymbol{\theta}$. If the prior distribution, in turn, has a hyperparameter $\boldsymbol{\lambda}$, its distribution is denoted by $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ ($\boldsymbol{\lambda} \in \Theta_\lambda \subset \mathrm{R}^q$; $q < p$).

9.3.1 Predictive Distributions and Predictive Likelihood

Let $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$ be n observations that are generated from an unknown probability distribution $G(x)$ having density function $g(x)$. Let $f(x|\boldsymbol{\theta})$ denote a parametric model having a p-dimensional parameter $\boldsymbol{\theta}$, and let us consider a Bayes model for which the prior distribution of the parameter $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta})$.

Given data $\boldsymbol{x}_n$ and the distribution $f(\boldsymbol{x}_n|\boldsymbol{\theta})$, it follows from Bayes' theorem that the *posterior distribution* of $\boldsymbol{\theta}$ is defined by

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_n) = \frac{f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\displaystyle\int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{9.51}$$

Let $\boldsymbol{z} = \{z_1, \cdots, z_n\}$ be future data generated independently of the observed data $\boldsymbol{x}_n$. Using the posterior distribution (9.51), we approximate the distribution $g(\boldsymbol{z})$ of the future data by

$$\begin{aligned} h(\boldsymbol{z}|\boldsymbol{x}_n) &= \int f(\boldsymbol{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)d\boldsymbol{\theta} \\ &= \frac{\displaystyle\int f(\boldsymbol{z}|\boldsymbol{\theta})f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\displaystyle\int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \end{aligned} \tag{9.52}$$

The $h(\boldsymbol{z}|\boldsymbol{x}_n)$ is called a *predictive distribution*.

In the following, we evaluate how well the predictive distribution approximates the distribution $g(\boldsymbol{z})$ that generates the data by using the expected log-likelihood

$$E_{G(\boldsymbol{z})}\left[\log h(\boldsymbol{Z}|\boldsymbol{x}_n)\right] = \int g(\boldsymbol{z}) \log h(\boldsymbol{z}|\boldsymbol{x}_n)d\boldsymbol{z}. \tag{9.53}$$

In actual modeling, the prior distribution $\pi(\boldsymbol{\theta})$ is rarely completely specified. In this section, we assume that the prior distribution of $\boldsymbol{\theta}$ is defined by a small number of parameters $\boldsymbol{\lambda} \in \Theta_\lambda \subset \mathrm{R}^q$ called *hyperparameters* and that they are expressed as $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$. In this situation, we denote the posterior distribution of $\boldsymbol{\theta}$, the predictive distribution of $\boldsymbol{z}$, and the marginal distribution of the data $\boldsymbol{x}_n$ by $\pi(\boldsymbol{\theta}|\boldsymbol{x}_n;\boldsymbol{\lambda})$, $h(\boldsymbol{z}|\boldsymbol{x}_n;\boldsymbol{\lambda})$, and $p(\boldsymbol{x}_n|\boldsymbol{\lambda})$, respectively.

For an ordinary parametric model $f(x|\boldsymbol{\theta})$, it is easy to see that

$$E_{G(\boldsymbol{x}_n)}\left[\log f(\boldsymbol{X}_n|\boldsymbol{\theta}) - E_{G(\boldsymbol{z})}\left[\log f(\boldsymbol{Z}|\boldsymbol{\theta})\right]\right] = 0, \tag{9.54}$$

as was shown in Chapter 3. Here, $E_{G(\boldsymbol{x}_n)}$ and $E_{G(\boldsymbol{z})}$ denote the expectations with respect to the data $\boldsymbol{x}_n$ and the future observations $\boldsymbol{z}$ obtained from the distribution G, respectively. Hence, in this case, the log-likelihood, $\log f(\boldsymbol{x}_n|\boldsymbol{\theta})$, is an unbiased estimator of the expected log-likelihood, and it provides a natural estimate of the expected log-likelihood. In the case of Bayesian models also, similar results can be derived with respect to the marginal distribution

$$p(\boldsymbol{z}) = \int f(\boldsymbol{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{9.55}$$

This implies that the log-likelihood provides a natural criterion for estimation of parameters.

In contrast, the Bayesian predictive distribution $h(\boldsymbol{z}|\boldsymbol{x}_n;\boldsymbol{\lambda})$ constructed by a prior distribution with hyperparameters $\boldsymbol{\lambda}$ generally takes the form

$$b_P(G,\boldsymbol{\lambda}) \equiv E_{G(\boldsymbol{x}_n)}\left[\log h(\boldsymbol{X}_n|\boldsymbol{X}_n;\boldsymbol{\lambda}) - E_{G(\boldsymbol{z})}\left[\log h(\boldsymbol{Z}|\boldsymbol{X}_n;\boldsymbol{\lambda})\right]\right] \neq 0. \tag{9.56}$$

Consequently, the log-likelihood $\log h(\boldsymbol{x}_n|\boldsymbol{x}_n;\boldsymbol{\lambda})$ is not an unbiased estimator of the expected log-likelihood $E_{G(\boldsymbol{z})}[\log h(\boldsymbol{Z}|\boldsymbol{x}_n;\boldsymbol{\lambda})]$. Therefore, in the estimation of the hyperparameters $\boldsymbol{\lambda}$, maximizing the expression $\log h(\boldsymbol{x}_n|\boldsymbol{x}_n;\boldsymbol{\lambda})$ does not result in maximizing the expected log-likelihood, even approximately.

The reason for this difficulty lies in the fact that, as in the case of previous information criteria, the same data $\boldsymbol{x}_n$ are used twice in the expression $\log h(\boldsymbol{x}_n|\boldsymbol{x}_n;\boldsymbol{\lambda})$. Therefore, when evaluating the predictive distribution for the estimation of hyperparameters in a Bayesian model, it is more natural to use the bias-corrected log-likelihood

$$\log h(\boldsymbol{x}_n|\boldsymbol{x}_n,\boldsymbol{\lambda}) - b_P(G,\boldsymbol{\lambda}) \tag{9.57}$$

as an estimate of the expected log-likelihood [Akaike (1980a) and Kitagawa (1984)].

In this section, in a similar way as the information criteria that have been presented thus far, we define the *predictive information criterion* (PIC) for Bayesian models as

$$\mathrm{PIC} = -2\log h(\boldsymbol{x}_n|\boldsymbol{x}_n;\boldsymbol{\lambda}) + 2b_P(G,\boldsymbol{\lambda}) \tag{9.58}$$

[Kitagawa (1997)]. If the hyperparameters $\boldsymbol{\lambda}$ are unknown, then the values of $\boldsymbol{\lambda}$ can be estimated by minimizing the PIC, in a manner similar to the maximum likelihood method described in Chapter 3. Given a predictive distribution of general Bayesian models, however, it is difficult to determine this bias analytically.

In the next section, we show that the bias correction term $b_P(G,\boldsymbol{\lambda})$ in (9.58) can be determined directly for a Bayesian normal linear model, and in Section 9.4, we describe how to use the Laplace integral approximation to determine it in the case of general Bayesian models.

9.3.2 Information Criterion for Bayesian Normal Linear Models

In this section, we consider a normal linear model in the Bayesian framework and determine the specific value of the bias term $b_P(G,\boldsymbol{\lambda})$.

Suppose that the n-dimensional observation vector $\boldsymbol{x}$ and the p-dimensional parameter vector $\boldsymbol{\theta}$ are both from multivariate normal distributions as follows:

$$\boldsymbol{X} \sim f(\boldsymbol{x}|\boldsymbol{\theta}) = N_n(A\boldsymbol{\theta},R), \qquad \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|\boldsymbol{\lambda}) = N_p(\boldsymbol{\theta}_0,Q), \tag{9.59}$$

where A is an $n\times p$ matrix, and R and Q are $n\times n$ and $p\times p$ nonsingular matrices, respectively. It is further assumed that the matrices A and R and the hyperparameters $\boldsymbol{\lambda}=(\boldsymbol{\theta}_0,Q)$ are all known.

The bias term $b_P(G, \boldsymbol{\lambda})$ for the Bayesian model given by (9.56) varies depending on the nature of the true distribution. For simplicity in what follows, we assume that the true distribution may be expressed as $g(x) = f(x|\boldsymbol{\theta})$ and $G(x) = F(x|\boldsymbol{\theta})$. In addition, we consider the case in which we evaluate the goodness of fit of the parameters $\boldsymbol{\theta}$, but not that of the hyperparameters $\boldsymbol{\lambda}$. We also assume that the observed data $\boldsymbol{x}$ and the future data $\boldsymbol{z}$ follow distributions having the same parameter $\boldsymbol{\theta}$. Then the bias can be determined exactly by calculating

$$\begin{aligned} b_P(F, \boldsymbol{\lambda}) &= E_{\Pi(\boldsymbol{\theta}|\boldsymbol{\lambda})} E_{F(\boldsymbol{x}|\boldsymbol{\theta})} \left[\log h(\boldsymbol{X}|\boldsymbol{X}; \boldsymbol{\lambda}) - E_{F(\boldsymbol{z}|\boldsymbol{\theta})} [\log h(\boldsymbol{Z}|\boldsymbol{X}; \boldsymbol{\lambda})] \right] \\ &= \int \left[\int \left\{ \log h(\boldsymbol{x}|\boldsymbol{x}, \boldsymbol{\lambda}) - \int f(\boldsymbol{z}|\boldsymbol{\theta}) \log h(\boldsymbol{z}|\boldsymbol{x}; \boldsymbol{\lambda}) d\boldsymbol{z} \right\} \right. \\ &\quad \left. \times f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \right] \pi(\boldsymbol{\theta}|\boldsymbol{\lambda}) d\boldsymbol{\theta}, \end{aligned} \tag{9.60}$$

where $\Pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ and $F(\boldsymbol{x}|\boldsymbol{\theta})$ are the distribution functions of $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ and $f(\boldsymbol{x}|\boldsymbol{\theta})$, respectively.

In the case of the Bayesian normal linear model, as will be shown in Subsection 9.3.3, we have the bias correction term

$$b_P(G, \boldsymbol{\lambda}) = \operatorname{tr} \left\{ (2W + R)^{-1} W \right\}, \tag{9.61}$$

where $W = AQA'$. Therefore, the PIC in this case is given by

$$\mathrm{PIC} = -2 \log f(\boldsymbol{x}|\boldsymbol{x}, \boldsymbol{\lambda}) + 2\mathrm{tr}\{(2W + R)^{-1} W\}. \tag{9.62}$$

Similarly, the bias correction term can also be determined when the parameters for the model $f(\boldsymbol{x}|\boldsymbol{\theta})$ depend on the MAP (maximum posterior estimate) defined by

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\boldsymbol{x}), \tag{9.63}$$

and in this case we have

$$\tilde{b}_P(G, \boldsymbol{\lambda}) = \operatorname{tr} \left\{ (W + R)^{-1} W \right\}. \tag{9.64}$$

9.3.3 Derivation of the PIC

To derive the information criterion PIC for the Bayesian normal linear model in (9.59), we use the following lemma [Lindley and Smith (1972)]:

Lemma (Marginal and posterior distributions for normal models) Assume that the distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$ of the n-dimensional vector $\boldsymbol{x}$ of random variables is an n-dimensional normal distribution $N_n(A\boldsymbol{\theta}, R)$ and that the distribution $\pi(\boldsymbol{\theta})$ of the p-dimensional parameter vector $\boldsymbol{\theta}$ is a p-dimensional normal distribution $N_p(\boldsymbol{\theta}_0, Q)$. Then we obtain the following results:

(i) The marginal distribution of $\boldsymbol{x}$ defined by

$$p(\boldsymbol{x}) = \int f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{9.65}$$

is distributed normally as $N_n(A\boldsymbol{\theta}_0, W + R)$, where $W = AQA^T$.

(ii) The posterior distribution of $\boldsymbol{\theta}$ defined by

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\displaystyle\int f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{9.66}$$

is distributed normally as $N_p(\boldsymbol{\xi}, V)$, where the mean vector $\boldsymbol{\xi}$ and the variance covariance matrix V are given by

$$\begin{aligned} \boldsymbol{\xi} &= \boldsymbol{\theta}_0 + QA^T(W + R)^{-1}(\boldsymbol{x} - A\boldsymbol{\theta}_0), \\ V &= Q - QA^T(W + R)^{-1}AQ \\ &= (A^T R^{-1} A + Q^{-1})^{-1}. \end{aligned} \tag{9.67}$$

For the prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ in (9.59), we derive specific forms of the marginal and posterior distributions by using the above lemma. In this case, ξ, V, and W in (9.67) depend on the hyperparameters $\boldsymbol{\lambda}$ and should be written as $\xi(\boldsymbol{\lambda})$, $V(\boldsymbol{\lambda})$, and $W(\boldsymbol{\lambda})$. For the sake of simplicity, in the following we shall denote them simply as ξ, V, and W.

By applying the results (i) and (ii) in the lemma to the Bayesian normal linear model of (9.59), the marginal distribution $p(\boldsymbol{x}|\boldsymbol{\lambda})$ and the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x};\boldsymbol{\lambda})$ are

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) \sim N_n(A\boldsymbol{\theta}_0, W + R), \quad \pi(\boldsymbol{\theta}|\boldsymbol{x};\boldsymbol{\lambda}) \sim N_p(\boldsymbol{\xi}, V), \tag{9.68}$$

where $\boldsymbol{\xi}$ and V are respectively the mean vector and the variance-covariance matrix of the posterior distribution given in (9.67). Then the predictive distribution defined by (9.52) in terms of the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x};\boldsymbol{\lambda})$ is an n-dimensional normal distribution, that is,

$$h(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\lambda}) = \int f(\boldsymbol{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x};\boldsymbol{\lambda})d\boldsymbol{\theta} \ \sim \ N_n(\boldsymbol{\mu}, \Sigma), \tag{9.69}$$

where the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix Σ are given by

$$\begin{aligned} \boldsymbol{\mu} &= A\boldsymbol{\xi} \\ &= W(W + R)^{-1}\boldsymbol{x} + R(W + R)^{-1}A\boldsymbol{\theta}_0, \end{aligned} \tag{9.70}$$

$$\begin{aligned} \Sigma &= AVA^T + R \\ &= W(W + R)^{-1}R + R \\ &= (2W + R)(W + R)^{-1}R. \end{aligned} \tag{9.71}$$

Consequently, using the log-likelihood of the predictive distribution written as

$$\log h(\boldsymbol{z}|\boldsymbol{x};\boldsymbol{\lambda}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu}), \quad (9.72)$$

the expectation of the difference between the log-likelihood and the expected log-likelihood may be evaluated as follows:

$$\begin{aligned} & E_{G(\boldsymbol{x})}\left[\log h(\boldsymbol{X}|\boldsymbol{X};\boldsymbol{\lambda}) - E_{G(\boldsymbol{z})}[\log h(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\lambda})]\right] \quad (9.73)\\ &= -\frac{1}{2}E_{G(\boldsymbol{x})}\left[(\boldsymbol{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) - E_{G(\boldsymbol{z})}[(\boldsymbol{Z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{Z}-\boldsymbol{\mu})]\right]\\ &= -\frac{1}{2}\mathrm{tr}\left\{\Sigma^{-1}E_{G(\boldsymbol{x})}\left[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T - E_{G(\boldsymbol{z})}[(\boldsymbol{Z}-\boldsymbol{\mu})(\boldsymbol{Z}-\boldsymbol{\mu})^T]\right]\right\}. \end{aligned}$$

We note that $\boldsymbol{\mu}$ in (9.70) depends on $\boldsymbol{X}$.

In the particular situation that the true distribution $g(\boldsymbol{z})$ is given by $f(\boldsymbol{z}|\boldsymbol{\theta}_0) \sim N_n(A\boldsymbol{\theta}_0, R)$, we have

$$\begin{aligned} & E_{F(\boldsymbol{z}|\boldsymbol{\theta})}\left[(\boldsymbol{Z}-\boldsymbol{\mu})(\boldsymbol{Z}-\boldsymbol{\mu})^T\right]\\ &\quad = E_{F(\boldsymbol{z}|\boldsymbol{\theta})}\left[(\boldsymbol{Z}-A\boldsymbol{\theta}_0)(\boldsymbol{Z}-A\boldsymbol{\theta}_0)^T\right] + (A\boldsymbol{\theta}_0-\boldsymbol{\mu})(A\boldsymbol{\theta}_0-\boldsymbol{\mu})^T\\ &\quad = R + (A\boldsymbol{\theta}_0-\boldsymbol{\mu})(A\boldsymbol{\theta}_0-\boldsymbol{\mu})^T. \quad (9.74) \end{aligned}$$

Writing $\Delta\boldsymbol{\theta} \equiv \boldsymbol{\theta} - \boldsymbol{\theta}_0$, we can see that

$$\begin{aligned} A\boldsymbol{\theta}_0 - \boldsymbol{\mu} &= W(W+R)^{-1}(A\boldsymbol{\theta}_0 - \boldsymbol{x}) + R(W+R)^{-1}A\Delta\boldsymbol{\theta},\\ \boldsymbol{x} - \boldsymbol{\mu} &= R(W+R)^{-1}\{(\boldsymbol{x} - A\theta_0) + A\Delta\boldsymbol{\theta}\}. \quad (9.75) \end{aligned}$$

Hence, by using $R = R(W+R)^{-1}W + R(W+R)^{-1}R$ and $\Sigma = R(W+R)^{-1}(2W+R)$, it follows from (9.74) and (9.75) that

$$\begin{aligned} & E_{F(\boldsymbol{x}|\boldsymbol{\theta})}\left[E_{F(\boldsymbol{z}|\boldsymbol{\theta})}[(\boldsymbol{Z}-\boldsymbol{\mu})(\boldsymbol{Z}-\boldsymbol{\mu})^T] - (\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T\right]\\ &\quad = R + W(W+R)^{-1}R(W+R)^{-1}W - R(W+R)^{-1}R(W+R)^{-1}R\\ &\quad = W(W+R)^{-1}R + R(W+R)^{-1}W\\ &\quad = \Sigma - R(W+R)^{-1}R. \quad (9.76) \end{aligned}$$

In this case, the bias correction term in (9.73) can be calculated exactly as

$$\begin{aligned} b_P(F,\boldsymbol{\lambda}) &= E_{\Pi(\boldsymbol{\theta})}E_{F(\boldsymbol{x}|\boldsymbol{\theta})}\left[\log h(\boldsymbol{X}|\boldsymbol{X};\boldsymbol{\lambda}) - E_{F(\boldsymbol{z}|\boldsymbol{\theta})}[\log h(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\lambda})]\right]\\ &= \frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\{\Sigma - R(W+R)^{-1}R\}\right]\\ &= \frac{1}{2}\mathrm{tr}\left\{I_n - (2W+R)^{-1}R\right\}\\ &= \mathrm{tr}\left\{(2W+R)^{-1}W\right\}. \quad (9.77) \end{aligned}$$

Since the expectation with respect to $F(\boldsymbol{x}|\boldsymbol{\theta})$ is constant and does not depend on the value of $\boldsymbol{\theta}$, integration with respect to $\boldsymbol{\theta}$ is not required. In

addition, the bias term does not depend on the individual observations $\boldsymbol{x}$ and is determined solely by the true variance covariance matrices R and Q.

By correcting the bias (9.77) for the log-likelihood of the predictive distribution in (9.72) and multiplying it by -2, we have the PIC for the Bayesian normal linear model in the form

$$\mathrm{PIC} = n\log(2\pi) + \log|\Sigma| + (\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + 2\mathrm{tr}\{(2W+R)^{-1}W\}, \tag{9.78}$$

where $\boldsymbol{\mu}$ and Σ are respectively given by (9.70) and (9.71).

9.3.4 Numerical Example

Suppose that we have n observations $\{x_\alpha; \alpha = 1, \ldots, n\}$ from a normal distribution model

$$x_\alpha = \mu_\alpha + w_\alpha, \qquad w_\alpha \sim N(0, \sigma^2), \tag{9.79}$$

where μ_α is the true mean and the variance σ^2 of the noise w_α is known. In order to estimate the mean-value function μ_α, we consider the trend model

$$x_\alpha = t_\alpha + w_\alpha, \qquad w_\alpha \sim N(0, \sigma^2). \tag{9.80}$$

For the trend component t_α, we assume a constraint model

$$t_\alpha = t_{\alpha-1} + v_\alpha, \qquad v_\alpha \sim N(0, \tau^2). \tag{9.81}$$

Then eqs. (9.80) and (9.81) can be formulated as the Bayesian model

$$\boldsymbol{x} = \boldsymbol{\theta} + \boldsymbol{w}, \qquad B\boldsymbol{\theta} = \boldsymbol{\theta}_* + \boldsymbol{v}, \tag{9.82}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, $\boldsymbol{\theta} = (t_1, \ldots, t_n)^T$, $\boldsymbol{w} = (w_1, \ldots, w_n)^T$, $\boldsymbol{v} = (v_1, \ldots, v_n)^T$, and B and $\boldsymbol{\theta}_*$ are, respectively, an $n \times n$ matrix and an n-dimensional vector given by

$$B = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}, \qquad \boldsymbol{\theta}_* = \begin{bmatrix} t_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{9.83}$$

In addition, for simplicity, we assume that $t_0 = \varepsilon_0$ ($\varepsilon_0 \sim N(0,1)$) and that the random variables $\boldsymbol{\theta}$ and $\boldsymbol{w}$ and $\boldsymbol{\theta}_*$ and $\boldsymbol{v}$ are mutually independent.

Setting $Q_0 = \mathrm{diag}\{\tau^2 + 1, \tau^2, \ldots, \tau^2\}$ and $\boldsymbol{\theta}_0 = B^{-1}\boldsymbol{\theta}_*$, we have

$$\boldsymbol{\theta} \sim N_n(\boldsymbol{\theta}_0, B^{-1}Q_0(B^{-1})^T). \tag{9.84}$$

Therefore, by taking $A = I_n$, $Q = B^{-1}Q_0(B^{-1})^T$, and $R = \sigma^2 I_n$, where I_n is the n-dimensional identity matrix, this model turns out to be the Bayesian normal linear model of (9.59).

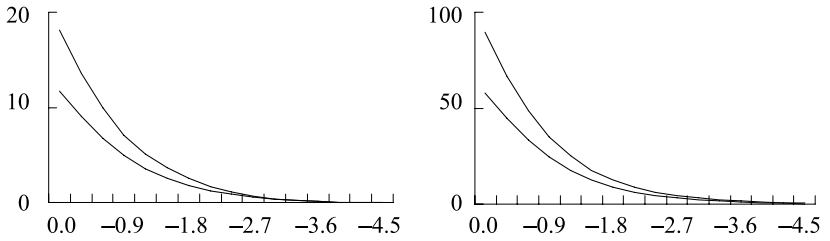

Fig. 9.3. Bias correction terms $2b_P(G, \lambda)$ and $2\tilde{b}_p(G, \lambda)$ for the Bayesian information criterion. The horizontal axis is λ, and the vertical axis shows the bias correction term. For the left graph, $n = 20$, and for the right graph, $n = 100$.

Figure 9.3 shows changes in the bias $2b_P(G, \lambda)$ and $2\tilde{b}_P(G, \lambda)$ as $n = 20$ and $n = 100$ for the values of $\lambda = \tau^2/\sigma^2 = 2^{-\ell}$ $(\ell = 0, 1, \ldots, 15)$, where $b_P(G, \lambda)$ and $\tilde{b}_P(G, \lambda)$ were obtained from (9.60) and (9.64), respectively. We note that, for a given value of n, the value of the bias depends solely on the variance ratio λ. As λ increases, the bias also increases significantly. In addition, the bias also increases as the number of observations increases, suggesting that the order is $O(n)$. From these results, we observe that the predictive likelihood without bias correction overestimates the goodness of fit when compared with the true predictive distribution, especially when the value of λ is large. Smoother estimates can be obtained by using a small λ that maximizes the predictive likelihood with a bias correction.

9.4 Bayesian Predictive Distributions by Laplace Approximation

This section considers a Bayesian model constructed from a parametric model $f(x|\boldsymbol{\theta})$ $(\boldsymbol{\theta} \in \Theta \subset R^p)$ and a prior distribution $\pi(\boldsymbol{\theta})$ for n observations $\boldsymbol{x}_n = \{x_1, \ldots, x_n\}$ that are generated from an unknown probability distribution $G(x)$ with density function $g(x)$.

For a future observation z that is randomly extracted independent of the data $\boldsymbol{x}_n$, we approximate the distribution $g(z)$ by the Bayesian predictive distribution

$$h(z|\boldsymbol{x}_n) = \int f(z|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)d\boldsymbol{\theta}, \tag{9.85}$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)$ is the posterior distribution of $\boldsymbol{\theta}$ given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_n) = \frac{f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{9.86}$$

By substituting this expression into (9.85), we can express the predictive distribution as

$$h(z|\boldsymbol{x}_n) = \frac{\int f(z|\boldsymbol{\theta}) f(\boldsymbol{x}_n|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int f(\boldsymbol{x}_n|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

$$= \frac{\int \exp\left[n\left\{n^{-1}\log f(\boldsymbol{x}_n|\boldsymbol{\theta}) + n^{-1}\log \pi(\boldsymbol{\theta}) + n^{-1}\log f(z|\boldsymbol{\theta})\right\}\right] d\boldsymbol{\theta}}{\int \exp\left[n\left\{n^{-1}\log f(\boldsymbol{x}_n|\boldsymbol{\theta}) + n^{-1}\log \pi(\boldsymbol{\theta})\right\}\right] d\boldsymbol{\theta}}$$

$$= \frac{\int \exp\left[n\left\{q(\boldsymbol{\theta}|\boldsymbol{x}_n) + n^{-1}\log f(z|\boldsymbol{\theta})\right\}\right] d\boldsymbol{\theta}}{\int \exp\left\{nq(\boldsymbol{\theta}|\boldsymbol{x}_n)\right\} d\boldsymbol{\theta}}, \tag{9.87}$$

where

$$q(\boldsymbol{\theta}|\boldsymbol{x}_n) = \frac{1}{n}\log f(\boldsymbol{x}_n|\boldsymbol{\theta}) + \frac{1}{n}\log \pi(\boldsymbol{\theta}). \tag{9.88}$$

We will now show that we can apply the information criterion GIC_M in (5.114) to the evaluation of a Bayesian predictive distribution, using Laplace's method for integrals described in Subsection 9.1.2 to approximate the predictive distribution in (9.87).

Let $\hat{\boldsymbol{\theta}}_q$ be a mode of $q(\boldsymbol{\theta}|\boldsymbol{x}_n)$ in (9.88). By applying the Laplace approximation to the denominator of (9.87), we obtain

$$\int \exp\left\{nq(\boldsymbol{\theta}|\boldsymbol{x}_n)\right\} d\boldsymbol{\theta}$$

$$= \frac{(2\pi)^{p/2}}{n^{p/2}\left|J_q(\hat{\boldsymbol{\theta}}_q)\right|^{1/2}} \exp\left\{nq(\hat{\boldsymbol{\theta}}_q|\boldsymbol{x}_n)\right\}\left\{1 + O_p(n^{-1})\right\}, \tag{9.89}$$

where $J_q(\hat{\boldsymbol{\theta}}_q) = -\partial^2\{q(\hat{\boldsymbol{\theta}}_q|\boldsymbol{x}_n)\}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$. Similarly, by letting $\hat{\boldsymbol{\theta}}_q(z)$ be a mode of $q(\boldsymbol{\theta}|\boldsymbol{x}_n) + n^{-1}\log f(z|\boldsymbol{\theta})$, we obtain the following Laplace approximation to the integral in the numerator:

$$\int \exp\left[n\left\{q(\boldsymbol{\theta}|\boldsymbol{x}_n) + \frac{1}{n}\log f(z|\boldsymbol{\theta})\right\}\right] d\boldsymbol{\theta}$$

$$= \frac{(2\pi)^{p/2}}{n^{p/2}|J_{q(z)}(\hat{\theta}_q(z))|^{1/2}} \exp\left[n\left\{q(\hat{\boldsymbol{\theta}}_q(z)|\boldsymbol{x}_n) + \frac{1}{n}\log f(z|\hat{\boldsymbol{\theta}}_q(z))\right\}\right]$$

$$\times \{1 + O_p(n^{-1})\}, \tag{9.90}$$

where $J_{q(z)}(\hat{\boldsymbol{\theta}}_q(z)) = -\partial^2\{q(\hat{\boldsymbol{\theta}}_q(z)|\boldsymbol{x}_n) + n^{-1}\log f(z|\hat{\boldsymbol{\theta}}_q(z))\}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$.

It follows from (9.89) and (9.90) that the predictive distribution $h(z|\boldsymbol{x}_n)$ can be approximated as follows:

$$h(z|\boldsymbol{x}_n) = \left(\frac{|J_q(\hat{\boldsymbol{\theta}}_q)|}{|J_{q(z)}(\hat{\boldsymbol{\theta}}_q(z))|} \right)^{\frac{1}{2}} \exp\left[n \left\{ q(\hat{\boldsymbol{\theta}}_q(z)|\boldsymbol{x}_n) - q(\hat{\boldsymbol{\theta}}_q|\boldsymbol{x}_n) + \frac{1}{n} \log f(z|\hat{\theta}_q(z)) \right\} \right] \times \{1 + O_p(n^{-2})\}. \tag{9.91}$$

Substituting functional Taylor series expansions for the modes $\hat{\boldsymbol{\theta}}_q$ and $\hat{\boldsymbol{\theta}}_q(z)$ into the resulting approximation and then simplifying the Laplace approximation (9.91) yield the Bayesian predictive distribution in the form

$$h(z|\boldsymbol{x}_n) = f(z|\hat{\boldsymbol{\theta}})\{1 + O_p(n^{-1})\}. \tag{9.92}$$

The form of the functional that defines the estimator $\hat{\boldsymbol{\theta}}$ is related to whether or not the prior distribution $\pi(\boldsymbol{\theta})$ depends upon the sample size n. Given a prior distribution, let us now consider two cases: (i) $\log \pi(\boldsymbol{\theta}) = O(1)$, (ii) $\log \pi(\boldsymbol{\theta}) = O(n)$. As can be seen from (9.88), in case (i), the estimator $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, and in case (ii), it becomes the mode $\hat{\boldsymbol{\theta}}_B$ of a posterior distribution. Functionals that define these estimators are solutions of

$$\int \left. \frac{\partial \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{T}_{\mathrm{ML}}(G)} dG(x) = \mathbf{0},$$

$$\int \left. \frac{\partial \log \{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{T}_B(G)} dG(x) = \mathbf{0}, \tag{9.93}$$

respectively.

In the information criterion GIC_M given by (5.114) in Subsection 5.2.3, by taking

$$\boldsymbol{\psi}(x, \hat{\boldsymbol{\theta}}) = \left. \frac{\partial \log f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{T}_{ML}(\hat{G})}, \tag{9.94}$$

$$\boldsymbol{\psi}(x, \hat{\boldsymbol{\theta}}) = \left. \frac{\partial \{\log f(\boldsymbol{x}|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{T}_B(\hat{G})}, \tag{9.95}$$

we obtain the information criterion for the Bayesian predictive distribution model $h(z|\boldsymbol{x}_n)$. It has the general form

$$\mathrm{GIC}_B = -2 \log h(\boldsymbol{x}_n|\boldsymbol{x}_n) + 2\mathrm{tr}\left\{ R(\boldsymbol{\psi}, \hat{G})^{-1} Q(\boldsymbol{\psi}, \hat{G}) \right\}. \tag{9.96}$$

In the case that $\log \pi(\boldsymbol{\theta}) = O(n)$, the asymptotic bias in (9.96) depends on the prior distribution through the partial derivatives of $\log \pi(\boldsymbol{\theta})$, while in the

case that $\log \pi(\boldsymbol{\theta}) = O(1)$, the asymptotic bias does not depend on the prior distribution and has the same form as that of TIC in (3.99). In the latter case, a more refined result is required in the context of smooth functional estimators.

The strength of the influence exerted by the prior distribution $\pi(\boldsymbol{\theta})$ is principally captured by its first- and second-order derivatives, with the result that if the prior distribution is $\log \pi(\boldsymbol{\theta}) = O(1)$, it does not contribute its effect solely on the basis of the first-order bias correction term. In such a situation, by taking the higher-order bias correction terms into account, we obtain a more accurate result.

The second-order (asymptotic) bias correction term $b_{(2)}(\hat{G})$ is defined as an estimator of $b_{(2)}(G)$, which is generally given by

$$E_{G(\boldsymbol{x})}\left[\log h(\boldsymbol{X}_n|\boldsymbol{X}_n) - \mathrm{tr}\left\{R(\boldsymbol{\psi},\hat{G})^{-1}Q(\boldsymbol{\psi},\hat{G})\right\} - nE_{G(z)}[h(Z|\boldsymbol{X}_n)]\right]$$
$$= \frac{1}{n}b_{(2)}(G) + O(n^{-2}). \tag{9.97}$$

Then we have the second-order bias-corrected log-likelihood of the predictive distribution in the form

$$\mathrm{GIC}_{\mathrm{BS}} = -2\log h(\boldsymbol{x}_n|\boldsymbol{x}_n) + 2\mathrm{tr}\left\{R(\boldsymbol{\psi},\hat{G})^{-1}Q(\boldsymbol{\psi},\hat{G})\right\} + \frac{2}{n}b_{(2)}(\hat{G}). \tag{9.98}$$

In fact, $b_{(2)}(G)$ is given by subtracting the asymptotic bias of the first-order correction term $\mathrm{tr}\{R(\boldsymbol{\psi},\hat{G})^{-1}Q(\boldsymbol{\psi},\hat{G})\}$ from the second-order asymptotic bias term of the log-likelihood of the model (see Subsection 7.2.2). Derivation of the second-order bias correction term includes log-likelihood, a high-order differentiation of the prior distribution, and a higher-order, compact differentiation of the estimator, and analytically it can be extremely complex. In such cases, bootstrap methods offer an alternative numerical approach to estimate the bias.

Example 5 (Bayesian predictive distribution) We use a normal distribution model

$$f(x|\mu,\tau^2) = \left(\frac{\tau^2}{2\pi}\right)^{\frac{1}{2}}\exp\left\{-\frac{\tau^2}{2}(x-\mu)^2\right\} \tag{9.99}$$

that approximates the true distribution as a prior distribution of parameters μ and τ^2, we assume

$$\pi(\mu,\tau^2) = N(\mu_0,\tau_0^{-2}\tau^{-2})G_a(\tau^2|\lambda,\beta) \tag{9.100}$$
$$= \left(\frac{\tau_0^2\tau^2}{2\pi}\right)^{\frac{1}{2}}\exp\left\{-\frac{\tau_0^2\tau^2}{2}(\mu-\mu_0)^2\right\}\frac{\beta^\lambda}{\Gamma(\lambda)}\tau^{2(\lambda-1)}e^{-\beta\tau^2}.$$

Then the predictive distribution is given by

$$h(z|\boldsymbol{x}) = \frac{\Gamma\left(\dfrac{b+1}{2}\right)}{\Gamma\left(\dfrac{b}{2}\right)} \left(\frac{a}{b\pi}\right)^{\frac{1}{2}} \left\{1 + \frac{a}{b}(z-c)^2\right\}^{-(a+1)/2}, \tag{9.101}$$

where $\overline{x} = \frac{1}{n}\sum_{\alpha=1}^{n} x_\alpha$C $s^2 = \frac{1}{n}\sum_{\alpha=1}^{n}(x_\alpha - \overline{x})^2$, and a, b, and c are defined as

$$a = \frac{(n+\tau_0^2)(\lambda + \frac{1}{2}n)}{(n+\tau_0+1)\left\{\beta + \dfrac{1}{2}ns^2 + \dfrac{\tau^2 n}{2(\tau_0^2+n)}(\mu_0 - \overline{x})^2\right\}},$$
$$b = 2\lambda + n, \quad c = \frac{\tau_0^2\mu_0 + n\overline{x}}{\tau_0^2 + n}, \tag{9.102}$$

respectively.

From (9.96), the information criterion for the evaluation of the predictive distribution is then given by

$$\mathrm{GIC}_B = -2\sum_{\alpha=1}^{n} \log h(x_\alpha|\boldsymbol{x}_n) + 2\left\{\frac{1}{2} + \frac{\hat{\mu}_4}{2(s^2)^2}\right\} \tag{9.103}$$

with

$$\hat{\mu}_4 = \frac{1}{n}\sum_{\alpha=1}^{n}(x_\alpha - \overline{x})^4. \tag{9.104}$$

It can be seen that GIC_B, which is an information criterion for the predictive distribution of a Bayesian model, takes a form similar to the TIC. In addition, the second-order bias correction term is given by

$$E_{G(\boldsymbol{x}_n)}\left[\log h(\boldsymbol{X}_n|\boldsymbol{X}_n) - \left\{\frac{1}{2} + \frac{\hat{\mu}_4}{2(s^2)^2}\right\} - n\int g(z)\log h(z|\boldsymbol{X}_n)dz\right]. \tag{9.105}$$

Example 6 (Numerical result) We compare the asymptotic bias estimate $(\mathrm{tr}\,\hat{I}\hat{J}^{-1})$ in (9.103), the bootstrap bias estimate (EIC), and the second-order corrected bias ($\mathrm{GIC}_{\mathrm{BS}}$) with the bootstrap bias estimate in (9.105). In the simulation study, data $\{x_\alpha;\ \alpha = 1,\ldots,n)$ were generated from a mixture of normal distributions

$$g(x) = (1-\varepsilon)N(0,1) + \varepsilon N(0,d^2). \tag{9.106}$$

Table 9.2 shows changes in the values of the true bias $b(G)$, $\mathrm{tr}\{\hat{I}\hat{J}^{-1}\}$, and the biases for EIC and $\mathrm{GIC}_{\mathrm{BS}}$ for various values of the mixture ratio ε. For

Table 9.2. Changes of true bias $b(G)$, $\mathrm{tr}\{\hat{I}\hat{J}^{-1}\}$, and the biases for EIC and $\mathrm{GIC}_{\mathrm{BS}}$ for various values of the mixture ratio ε.

ε	$b(G)$	$\mathrm{tr}\{\hat{I}\hat{J}^{-1}\}$	EIC	$\mathrm{GIC}_{\mathrm{BS}}$
0.00	2.07	1.89	1.97	2.01
0.04	2.96	2.41	2.52	2.76
0.08	3.50	2.73	2.89	3.24
0.12	3.79	2.90	3.13	3.52
0.16	3.95	2.99	3.28	3.68
0.20	4.02	3.01	3.35	3.73
0.24	3.96	2.99	3.39	3.73
0.28	3.92	2.95	3.38	3.69
0.32	3.77	2.89	3.40	3.69
0.36	3.72	2.82	3.31	3.56
0.40	3.60	2.74	3.29	3.51

model parameters, we set $d^2 = 10$, $\mu_0 = 1$, $\tau_0^2 = 1$, $\alpha = 4$, and $\beta = 1$ and ran Monte Carlo trials with 100,000 repetitions. In the bias estimation for EIC, we used $B = 10$ for the bootstrap replications.

It can be seen from the table that the bootstrap bias estimate of EIC is closer to the true bias than the bias correction term $\mathrm{tr}\{\hat{I}\hat{J}^{-1}\}$ for TIC or GIC_B. It can also be seen that the second-order correction term of $\mathrm{GIC}_{\mathrm{BS}}$ is even more accurate than these other two correction terms.

9.5 Deviance Information Criterion (DIC)

Spiegelhalter et al. (2002) developed a deviance information criterion (DIC) from a Bayesian perspective, using an information-theoretic argument to motivate a complexity measure for the effective number of parameters in a model. Let $f(\boldsymbol{x}_n|\boldsymbol{\theta})$ ($\boldsymbol{\theta} \in \Theta \subset R^p$) and $\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)$ be, respectively, a probability model and a posterior distribution for the observed data $\boldsymbol{x}_n$. Spiegelhalter et al. (2002) proposed the effective number of parameters with respect to a model in the form

$$p_D = -2E_{\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)}\big[\log f(\boldsymbol{x}_n|\boldsymbol{\theta})\big] + 2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}), \tag{9.107}$$

where $\hat{\boldsymbol{\theta}}$ is an estimator of the parameter vector $\boldsymbol{\theta}$. Using the Bayesian deviance defined by

$$D(\boldsymbol{\theta}) = -2\log f(\boldsymbol{x}_n|\boldsymbol{\theta}) + 2\log h(\boldsymbol{x}_n), \tag{9.108}$$

where $h(\boldsymbol{x}_n)$ is some fully specified standardizing term that is a function of the data alone, eq. (9.107) can be written as

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}}), \tag{9.109}$$

where $\overline{\boldsymbol{\theta}}$ $(= \hat{\boldsymbol{\theta}})$ is the posterior mean defined by $\overline{\boldsymbol{\theta}} = E_{\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)}[\boldsymbol{\theta}]$ and $\overline{D(\boldsymbol{\theta})}$ is the posterior mean of the deviance defined by $\overline{D(\boldsymbol{\theta})} = E_{\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)}[D(\boldsymbol{\theta})]$.

This shows that a measure for the effective number of parameters in a model can be considered as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest. Note that when models are compared, the second term in the Bayesian deviance cancels out.

Spiegelhalter et al. (2002) defined DIC as

$$\begin{aligned} \text{DIC} &= \overline{D(\boldsymbol{\theta})} + p_D \\ &= -2E_{\pi(\boldsymbol{\theta}|\boldsymbol{x}_n)}\big[\log f(\boldsymbol{x}_n|\boldsymbol{\theta})\big] + p_D. \end{aligned} \tag{9.110}$$

It follows from (9.109) that the DIC can also be expressed as

$$\begin{aligned} \text{DIC} &= D(\overline{\boldsymbol{\theta}}) + 2p_D \\ &= -2\log f(\boldsymbol{x}_n|\overline{\boldsymbol{\theta}}) + 2p_D. \end{aligned} \tag{9.111}$$

The optimal model among a set of competing models is chosen by selecting one that minimizes the value of DIC. The DIC can be considered as a Bayesian measure of fit or adequacy, penalized by an additional complexity term p_D [Spiegelhalter et al. (2002)].