PREDICTIVE INFORMATION CRITERIA FOR BAYESIAN NONLINEAR REGRESSION MODELS

Kim, Daeju
Graduate School of Mathematics, Kyushu University

Kawano, Shuichi
Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University

Konishi, Sadanori
Department of Mathematics, Faculty of Science and Engineering, Chuo University

https://doi.org/10.5109/1495408

出版情報：Bulletin of informatics and cybernetics. 44, pp.17-28, 2012-12. 統計科学研究会
バージョン：
権利関係：

PREDICTIVE INFORMATION CRITERIA FOR BAYESIAN NONLINEAR REGRESSION MODELS

by

Daeju Kim, **Shuichi Kawano** and **Sadanori Konishi**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.44

FUKUOKA, JAPAN
2012

Bulletin of Informatics and Cybernetics, Vol. 44, 2012

PREDICTIVE INFORMATION CRITERIA FOR BAYESIAN NONLINEAR REGRESSION MODELS

By

Daeju Kim[*], **Shuichi Kawano**[†] and **Sadanori Konishi**[‡]

Abstract

Bayesian nonlinear regression modeling based on basis expansions provides efficient methods for analyzing data with complicated structure. A crucial issue in the model building process is the choice of adjusted parameters including hyper-parameters for prior distribution and the number of basis functions. Choosing these parameters can be viewed as a model selection and evaluation problem. We present an information criterion for evaluating Bayesian nonlinear regression models. Our proposed modeling procedure enables us to select the appropriate values of hyper-parameters and the number of basis functions. We use a real data analysis and simulation studies to validate the performance of the proposed nonlinear regression modeling. Simulation studies show that our proposed modeling strategy performs well in various situations.

Key Words and Phrases: Basis expansions, Bayesian predictive distribution, Model selection, Nonlinear regression models.

1. Introduction

Nonlinear regression models based on basis expansions have emerged as useful tools to draw information from data with complex structure (see, e.g., Bishop, 2006; Figueiredo, 2003; Kohn *et al.*, 2001; Minka, 2000 and so on). The essential idea for basis expansions is to express a regression function as a linear combination of specified functions, called basis functions (Hastie *et al.*, 2009; Konishi and Kitagawa, 2008). In constructing a model, various basis functions are used to represent a regression function according to the structure of data or the purpose of analysis. For example, splines (Green and Silverman, 1994), B-splines (de Boor, 2001) and radial basis functions (Hastie *et al.*, 2009) have been widely used to construct nonlinear regression models. Nonlinear regression models are generally characterized by many parameters to be estimated. Since maximum likelihood methods yield unstable parameter estimates, the adopted model is usually estimated by the method of regularization or the Bayesian approach (Bishop, 2006; Denison *et al.*, 2002; Figueiredo, 2003).

In nonlinear regression models constructed by Bayesian approach, a crucial issue is to select hyper-parameters in prior distributions and the number of basis functions.

[*] Graduate School of Mathematics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan. t-kin@math.kyushu-u.ac.jp

[†] Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University, 1-1 Gakuen-cho, Sakai, Osaka 599-8531, Japan. skawano@ms.osakafu-u.ac.jp

[‡] Department of Mathematics, Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan. konishi@math.chuo-u.ac.jp

The selection of the tuning parameters can be viewed as a model evaluation problem. In order to overcome the problem, several information criteria have been proposed; e.g., mAIC (Hasite and Tibshirani, 1990), NIC (Murata *et al.*, 1994), RIC (Shibata, 1989), GIC (Konishi and Kitagawa, 1996). While these criteria are generally constructed based on the Kullback-Leibler information (Kullback and Leibler, 1951) for a plug-in predictive distribution, we take a Bayesian predictive distribution instead of the plug-in predictive distribution. Until now, only Kitagawa (1997) presented an information criterion for Bayesian predictive distributions, called the predictive information criterion (PIC). In regression models with known variance, PIC can be exactly derived since the Bayesian predictive distribution belongs to the normal distribution. However, in practice, it is difficult to consider the situation that the variance is known. Hence, we should assume regression models with unknown variance in practical situations. In such cases, PIC cannot be directly provided, since the Bayesian predictive distribution does not belong to the normal distribution.

In this paper, we derive an information criterion for evaluating Bayesian nonlinear regression models with unknown variance. The proposed criterion enables us to choose the values of hyper-parameters in the prior distribution and the number of basis functions simultaneously. Our Bayesian nonlinear regression modeling procedure is investigated through some numerical examples.

The remainder of this article is organized as follows. In Section 2, we describe a framework of nonlinear regression models based on basis expansions. Section 3 derives Bayesian predictive distributions for nonlinear regression models. In Section 4, we obtain a model selection criterion for evaluating Bayesian predictive distributions for nonlinear regression models. In Section 5, we investigate the performance of the proposed modeling procedure by real data analysis and Monte Carlo simulations. Some concluding remarks are given in Section 6.

2. Nonlinear regression models based on basis expansions

Suppose that $\{(y_i, \boldsymbol{x}_i); i = 1, ..., n\}$ are n sets of data obtained in terms of the response variable y and p-dimensional explanatory variables $\boldsymbol{x} = (x_1, ..., x_p)^{\mathrm{T}}$. In order to draw information from the data, we consider the Gaussian nonlinear regression model

$$y_i = u(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, ..., n, \tag{1}$$

where $u(\cdot)$ is a true smooth function and errors ε_i are independently, identically distributed according to $N(0, \sigma^2)$. The unknown function $u(\cdot)$ is approximated by a linear combination of basis functions

$$u(\boldsymbol{x}_i) = \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i), \tag{2}$$

where $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), ..., \phi_m(\boldsymbol{x}))^{\mathrm{T}}$ is a vector of basis functions and $\boldsymbol{w} = (w_1, ..., w_m)^{\mathrm{T}}$ is an unknown coefficient parameter vector. Often times, natural cubic splines (Green and Silverman, 1994), B-splines (de Boor, 2001) and radial basis functions (Hastie *et al.*, 2009) are used for basis functions.

Combining Equations (1) and (2), we have the Gaussian nonlinear regression model

$$y_i = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, ..., n$$

with a probability density function

$$f(y_i|\boldsymbol{x}_i;\boldsymbol{w},\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left\{y_i-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i)\right\}^2}{2\sigma^2}\right],\quad i=1,...,n.$$

Hereafter, for simplicity, we omit the description of the explanatory variables and denote $f(y_i|\boldsymbol{x}_i;\boldsymbol{w},\sigma^2)$ as $f(y_i|\boldsymbol{w},\sigma^2)$. For a future data $\boldsymbol{z}$ generated independently from the observed data $\boldsymbol{y}$, the plug-in type predictive distribution is given by $f(\boldsymbol{z}|\hat{\boldsymbol{w}}_{\mathrm{MLE}},\hat{\sigma}^2_{\mathrm{MLE}})$, where $\hat{\boldsymbol{w}}_{\mathrm{MLE}}$ and $\hat{\sigma}^2_{\mathrm{MLE}}$ are the maximum likelihood estimators,

$$\hat{\boldsymbol{w}}_{\mathrm{MLE}}=(\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\boldsymbol{y},\quad \hat{\sigma}^2_{\mathrm{MLE}}=\frac{1}{n}(\boldsymbol{y}-\Phi\hat{\boldsymbol{w}}_{\mathrm{MLE}})^{\mathrm{T}}(\boldsymbol{y}-\Phi\hat{\boldsymbol{w}}_{\mathrm{MLE}})$$

with $\Phi=(\boldsymbol{\phi}(\boldsymbol{x}_1),\ldots,\boldsymbol{\phi}(\boldsymbol{x}_n))^{\mathrm{T}}$ and $\boldsymbol{y}=(y_1,\ldots,y_n)^{\mathrm{T}}$. It is known that the maximum likelihood methods in nonlinear regression modeling often yield unstable parameter estimates, and lead to overfitting (Konishi and Kitagawa, 2008). Then the adopted model is usually estimated by the regularization method or the Bayesian approach. In the next section we consider a predictive distribution from the Bayesian viewpoint.

3. Bayesian predictive distributions

Given data $\boldsymbol{y}$, it follows from Bayes rule that a joint posterior distribution of $\boldsymbol{w}$ and σ^2 is defined by

$$\pi(\boldsymbol{w},\sigma^2|\boldsymbol{y})=\frac{f(\boldsymbol{y}|\boldsymbol{w},\sigma^2)\pi(\boldsymbol{w},\sigma^2)}{\int f(\boldsymbol{y}|\boldsymbol{w},\sigma^2)\pi(\boldsymbol{w},\sigma^2)d\boldsymbol{w}d\sigma^2},\tag{3}$$

where $\pi(\boldsymbol{w},\sigma^2)$ is a joint prior distribution of $\boldsymbol{w}$ and σ^2, and $f(\boldsymbol{y}|\boldsymbol{w},\sigma^2)$ is a joint probability density function defined as $f(\boldsymbol{y}|\boldsymbol{w},\sigma^2)=\prod_{i=1}^n f(y_i|\boldsymbol{w},\sigma^2)$. For the joint prior distribution $\pi(\boldsymbol{w},\sigma^2)$, we use a normal inverse-gamma distribution (Denison *et al.*, 2002), which is defined by

$$\pi(\boldsymbol{w},\sigma^2)=\pi_1(\boldsymbol{w}|\sigma^2)\pi_2(\sigma^2),$$

where the density functions $\pi_1(\boldsymbol{w}|\sigma^2)$ and $\pi_2(\sigma^2)$ are given by

$$\boldsymbol{w}|\sigma^2\sim N_m\left(\boldsymbol{w}|\boldsymbol{0},\left(\sigma^2/n\lambda\right)\mathrm{I}_m\right),\quad \sigma^2\sim IG\left(\sigma^2|\nu_0/2,\eta_0/2\right),$$

respectively. Here $N_m(\boldsymbol{w}|\cdot,\cdot)$ is an m-dimensional normal distribution, $IG(\sigma^2|\cdot,\cdot)$ is an inverse gamma distribution and λ, ν_0 and η_0 are hyper-parameters with positive values. We set $\nu_0=\eta_0=10^{-10}$ so that the prior distribution does not affect estimators of the parameters.

Then the joint posterior distribution (3) can be expressed as

$$\pi(\boldsymbol{w},\sigma^2|\boldsymbol{y})=\pi_1(\boldsymbol{w}|\sigma^2,\boldsymbol{y})\pi_2(\sigma^2|\boldsymbol{y}),$$

where the density functions of posterior distributions are given by

$$\boldsymbol{w}|\sigma^2,\boldsymbol{y}\sim N_m\left(\boldsymbol{w}|\hat{\boldsymbol{w}}_n,\sigma^2A_n\right),\quad \sigma^2|\boldsymbol{y}\sim IG\left(\sigma^2|\nu_n/2,\hat{\eta}_n/2\right),\tag{4}$$

respectively. Here,

$$A_n = (\Phi^{\mathrm{T}}\Phi + n\lambda \mathrm{I}_m)^{-1}, \quad \hat{\boldsymbol{w}}_n = A_n\Phi^{\mathrm{T}}\boldsymbol{y},$$
$$\nu_n = n + \nu_0, \quad \hat{\eta}_n = \eta_0 + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \hat{\boldsymbol{w}}_n^{\mathrm{T}}A_n^{-1}\hat{\boldsymbol{w}}_n.$$

From (4), we obtain the Bayesian predictive distribution which is an n-dimensional Student t-distribution with ν_n degrees of freedom,

$$\begin{aligned} h(\boldsymbol{z}|\boldsymbol{y}) &= \int f(\boldsymbol{z}|\boldsymbol{w},\sigma^2)\pi(\boldsymbol{w},\sigma^2|\boldsymbol{y})d\boldsymbol{w}d\sigma^2 \\ &= \frac{\Gamma\left((n+\nu_n)/2\right)}{\Gamma\left(\nu_n/2\right)(\pi\nu_n)^{\frac{n}{2}}}\left|\hat{\Sigma}\right|^{-\frac{1}{2}}\left[1+(1/\nu_n)(\boldsymbol{z}-\Phi\hat{\boldsymbol{w}}_n)^{\mathrm{T}}\hat{\Sigma}^{-1}(\boldsymbol{z}-\Phi\hat{\boldsymbol{w}}_n)\right]^{-\left(\frac{n+\nu_n}{2}\right)} \end{aligned} \tag{5}$$

where $\hat{\Sigma} = (\hat{\eta}_n/\nu_n)(\Phi A_n\Phi^{\mathrm{T}} + \mathrm{I}_n)$ and $\Gamma(\cdot)$ is the Gamma function.

Note that the Bayesian predictive distribution (5) includes a hyper-parameter λ and the number of basis functions m. We choose the optimal value of hyper-parameter and the number of basis functions from given data, objectively. This problem will be discussed in the next section.

4. Model selection criterion

4.1. Proposed model selection criterion

In the Bayesian nonlinear regression model based on basis expansions, a crucial problem is the choices of the hyper-parameter λ for the prior distribution and the number of basis functions m.

In this paper, we derive a predictive information criterion to evaluate the Bayesian predictive distribution given in Equation (5). Kitagawa (1997) proposed information criterion PIC for evaluating the Bayesian predictive distribution given as a multivariate normal distribution. The PIC for the Bayesian predictive distribution is, in general, given by

$$\mathrm{PIC} = -2\log h(\boldsymbol{y}|\boldsymbol{y}) + 2B_p,$$

where B_p is a bias term defined as

$$B_p = E_{q(\boldsymbol{y})}\left[\log h(\boldsymbol{y}|\boldsymbol{y}) - E_{q(\boldsymbol{z})}\left[\log h(\boldsymbol{z}|\boldsymbol{y})\right]\right]$$

for the true distribution $q(\cdot)$. It is, however, difficult to derive PIC for the predictive distribution $h(\boldsymbol{z}|\boldsymbol{y})$ in Equation (5) analytically, because it does not belong to the normal distribution. To overcome this difficulty, we derive the PIC approximately by using the Laplace method (Davison, 1986; Tierney and Kadane, 1986). According to Konishi and Kitagawa (2008), the Laplace method yields the approximation of the Bayesian predictive distribution in Equation (5) as

$$h(\boldsymbol{z}|\boldsymbol{y}) = f(\boldsymbol{z}|\tilde{\boldsymbol{w}},\tilde{\sigma}^2)(1+O_p(n^{-1})),$$

where $\tilde{\boldsymbol{w}}$ and $\tilde{\sigma}^2$ are, respectively, given by

$$\tilde{\boldsymbol{w}} = (\Phi^{\mathrm{T}}\Phi + n\lambda\mathrm{I}_m)^{-1}\Phi^{\mathrm{T}}\boldsymbol{y}, \quad \tilde{\sigma}^2 = \frac{(\boldsymbol{y}-\Phi\tilde{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\Phi\tilde{\boldsymbol{w}}) + n\lambda\tilde{\boldsymbol{w}}^{\mathrm{T}}\tilde{\boldsymbol{w}} + \eta_0}{n+m+\nu_0+2}.$$

For the approximated Bayesian predictive distribution $f(\boldsymbol{z}|\tilde{\boldsymbol{w}}, \tilde{\sigma}^2)$, we define PIC* as follows:

$$\mathrm{PIC}^* = -2\log f(\boldsymbol{y}|\tilde{\boldsymbol{w}}, \tilde{\sigma}^2) + 2B'_p,$$

where B'_p is the approximated bias term of B_p, given by

$$B'_p = E_{q(\boldsymbol{y})}\left[\log f(\boldsymbol{y}|\tilde{\boldsymbol{w}}, \tilde{\sigma}^2) - E_{q(\boldsymbol{z})}\left[\log f(\boldsymbol{z}|\tilde{\boldsymbol{w}}, \tilde{\sigma}^2)\right]\right].$$

The bias term B'_p can be written as

$$B'_p = -\frac{1}{2\tilde{\sigma}^2}\mathrm{tr}\left[E_{q(\boldsymbol{y})}\left[(\boldsymbol{y}-\Phi\tilde{\boldsymbol{w}})(\boldsymbol{y}-\Phi\tilde{\boldsymbol{w}})^{\mathrm{T}} - E_{q(\boldsymbol{z})}\left[(\boldsymbol{z}-\Phi\tilde{\boldsymbol{w}})(\boldsymbol{z}-\Phi\tilde{\boldsymbol{w}})^{\mathrm{T}}\right]\right]\right]. \quad (6)$$

It is assumed here that the true distribution $q(\boldsymbol{z})$ is $f(\boldsymbol{z}|\boldsymbol{w}^*, \sigma^{2*})$, where $\boldsymbol{w}^* \in \mathbb{R}^m$ and $\sigma^{2*} \in \mathbb{R}$ are the true regression coefficients and variance, respectively. Then the second term in Equation (6) can be expressed as

$$E_{f(\boldsymbol{z}|\boldsymbol{w}^*,\sigma^{2*})}\left\{(\boldsymbol{z}-\Phi\boldsymbol{w}^*)(\boldsymbol{z}-\Phi\boldsymbol{w}^*)^{\mathrm{T}}\right\} + (\Phi\boldsymbol{w}^*-\Phi\tilde{\boldsymbol{w}})(\Phi\boldsymbol{w}^*-\Phi\tilde{\boldsymbol{w}})^{\mathrm{T}},$$

noting that the following equations hold

$$\begin{aligned}\Phi\boldsymbol{w}^* - \Phi\tilde{\boldsymbol{w}} &= \Phi(\Phi^{\mathrm{T}}\Phi + n\lambda \mathrm{I}_m)^{-1}\Phi^{\mathrm{T}}(\Phi\boldsymbol{w}^* - \boldsymbol{y}) + n\lambda\Phi(\Phi^{\mathrm{T}}\Phi + n\lambda\mathrm{I}_m)^{-1}\boldsymbol{w}^*,\\ \boldsymbol{y} - \Phi\tilde{\boldsymbol{w}} &= (\mathrm{I}_n - \Phi(\Phi^{\mathrm{T}}\Phi + n\lambda\mathrm{I}_m)^{-1}\Phi^{\mathrm{T}})(\boldsymbol{y} - \Phi\boldsymbol{w}^*) + n\lambda\Phi(\Phi^{\mathrm{T}}\Phi + n\lambda\mathrm{I}_m)^{-1}\boldsymbol{w}^*.\end{aligned}$$

Hence we have

$$B'_p = \left(\frac{\sigma^{2*}}{\tilde{\sigma}^2}\right)\mathrm{tr}\left[\Phi(\Phi^{\mathrm{T}}\Phi + n\lambda\mathrm{I}_m)^{-1}\Phi^{\mathrm{T}}\right].$$

Consequently, we obtain the PIC* in the form

$$\begin{aligned}\mathrm{PIC}^* = n\log(2\pi) + n\log\tilde{\sigma}^2 + \frac{1}{\tilde{\sigma}^2}(\boldsymbol{y}-\Phi\tilde{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\Phi\tilde{\boldsymbol{w}})\\ +2\left(\frac{\sigma^{2*}}{\tilde{\sigma}^2}\right)\mathrm{tr}\left[\Phi(\Phi^{\mathrm{T}}\Phi + n\lambda\mathrm{I}_m)^{-1}\Phi^{\mathrm{T}}\right].\end{aligned}$$

Our proposed model selection criterion contains the unknown variance σ^{2*}. We consider two types of PIC*. We first take the maximum likelihood estimate for σ^{2*}. As a consequence, the PIC* can be expressed as follows:

$$\begin{aligned}\mathrm{PIC}_{\mathrm{MLE}} = n\log(2\pi) + n\log\tilde{\sigma}^2 + \frac{1}{\tilde{\sigma}^2}(\boldsymbol{y}-\Phi\tilde{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\Phi\tilde{\boldsymbol{w}})\\ +2\left(\frac{\hat{\sigma}^2_{\mathrm{MLE}}}{\tilde{\sigma}^2}\right)\mathrm{tr}\left[\Phi(\Phi^{\mathrm{T}}\Phi + n\lambda\mathrm{I}_m)^{-1}\Phi^{\mathrm{T}}\right]. \quad (7)\end{aligned}$$

Secondly, the variance is replaced with the mode of the posterior distribution, and hence we obtain

$$\begin{aligned}\mathrm{PIC}_{\mathrm{Mode}} = n\log(2\pi) + n\log\tilde{\sigma}^2 + \frac{1}{\tilde{\sigma}^2}(\boldsymbol{y}-\Phi\tilde{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\Phi\tilde{\boldsymbol{w}})\\ +2\mathrm{tr}\left[\Phi(\Phi^{\mathrm{T}}\Phi + n\lambda\mathrm{I}_m)^{-1}\Phi^{\mathrm{T}}\right]. \quad (8)\end{aligned}$$

We select the optimal values of the hyper-parameter and the number of basis functions that minimize either $\mathrm{PIC}_{\mathrm{MLE}}$ or $\mathrm{PIC}_{\mathrm{Mode}}$.

4.2. Other model selection criteria

Other model selection criteria include GIC (Konishi and Kitagawa, 1996) and mAIC (Hastie and Tibshirani, 1990). Konishi and Kitagawa (1996) introduced an evaluation criterion of statistical models estimated by various types of estimation procedures such as the robust and penalized likelihood procedures. In this section, we consider GIC for the nonlinear regression model based on basis expansions which are estimated by the penalized likelihood method. We estimate regression coefficients $\boldsymbol{w}$ and a variance σ^2 for maximizing the penalized likelihood function,

$$\ell_{\mathrm{P}}(\boldsymbol{w}, \sigma^2) = \log f(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) - \frac{n\gamma}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w},$$

where γ (> 0) is a smoothing parameter which controls the model complexity. The estimates of regression coefficients and the variance are respectively given by

$$\hat{\boldsymbol{w}}_{\mathrm{P}} = (\Phi^{\mathrm{T}}\Phi + n\gamma\hat{\sigma}_p^2\mathrm{I}_m)^{-1}\Phi^{\mathrm{T}}\boldsymbol{y}, \quad \hat{\sigma}_{\mathrm{P}}^2 = \frac{1}{n}(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}}_{\mathrm{P}})^{\mathrm{T}}(\boldsymbol{y} - \Phi\hat{\boldsymbol{w}}_{\mathrm{P}}).$$

Then, we derive the GIC for nonlinear regression models based on basis expansions as follows;

$$\mathrm{GIC} = n\left\{\log(2\pi) + 1\right\} + n\log\hat{\sigma}_{\mathrm{P}}^2 + 2\mathrm{tr}\left\{R^{-1}Q\right\}, \tag{9}$$

where R and Q are $(m+1)\times(m+1)$ matrices and are, respectively, given by

$$R = \frac{1}{n\hat{\sigma}_{\mathrm{P}}^2}\begin{bmatrix} \Phi^{\mathrm{T}}\Phi + n\gamma\hat{\sigma}_{\mathrm{P}}^2\mathrm{I}_m & \dfrac{1}{\hat{\sigma}_{\mathrm{P}}^2}\Phi^{\mathrm{T}}\Lambda\mathbf{1}_n \\ \dfrac{1}{\hat{\sigma}_{\mathrm{P}}^2}\mathbf{1}_n^{\mathrm{T}}\Lambda\Phi & \dfrac{n}{2\hat{\sigma}_{\mathrm{P}}^2} \end{bmatrix},$$

$$Q = \frac{1}{n\hat{\sigma}_{\mathrm{P}}^2}\begin{bmatrix} \dfrac{1}{\hat{\sigma}_{\mathrm{P}}^2}\Phi^{\mathrm{T}}\Lambda^2\Phi - \gamma\mathrm{I}_m\hat{\boldsymbol{w}}_{\mathrm{P}}\mathbf{1}_n^{\mathrm{T}}\Lambda\Phi & \dfrac{1}{2\hat{\sigma}_{\mathrm{P}}^4}\Phi^{\mathrm{T}}\Lambda^3\mathbf{1}_n - \dfrac{1}{2\hat{\sigma}_{\mathrm{P}}^2}\Phi^{\mathrm{T}}\Lambda\mathbf{1}_n \\ \dfrac{1}{2\hat{\sigma}_{\mathrm{P}}^4}\mathbf{1}_n^{\mathrm{T}}\Lambda^3\Phi - \dfrac{1}{2\hat{\sigma}_{\mathrm{P}}^2}\mathbf{1}_n^{\mathrm{T}}\Lambda\Phi & \dfrac{1}{4\hat{\sigma}_{\mathrm{P}}^6}\mathbf{1}_n^{\mathrm{T}}\Lambda^4\mathbf{1}_n - \dfrac{n}{4\hat{\sigma}_{\mathrm{P}}^2} \end{bmatrix}$$

with $\mathbf{1}_n = (1, ..., 1)^{\mathrm{T}}$ and $\Lambda = \mathrm{diag}[y_1 - \hat{\boldsymbol{w}}_{\mathrm{P}}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_1), ..., y_n - \hat{\boldsymbol{w}}_{\mathrm{P}}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_n)]$.

Hastie and Tibshirani (1990) proposed to use the trace of the smoother matrix given by $H = \Phi(\Phi^{\mathrm{T}}\Phi + n\gamma\hat{\sigma}_{\mathrm{P}}^2\mathrm{I}_m)^{-1}\Phi^{\mathrm{T}}$, as an approximation of the effective degrees of freedom. By replacing the number of parameters in AIC (Akaike, 1974) with trace of the smoother matrix, we obtain

$$\mathrm{mAIC} = n\left\{\log(2\pi) + 1\right\} + n\log\hat{\sigma}_{\mathrm{P}}^2 + 2\mathrm{tr}\left\{\Phi(\Phi^{\mathrm{T}}\Phi + n\gamma\hat{\sigma}_{\mathrm{P}}^2\mathrm{I}_m)^{-1}\Phi^{\mathrm{T}} + 1\right\}. \tag{10}$$

We select the optimal values of m and γ that minimize either GIC or mAIC.

5. Numerical examples

5.1. Analysis of real data

We illustrate the proposed procedure for choosing the hyper-parameter in prior distribution and the number of basis functions through the analysis of the motorcycle

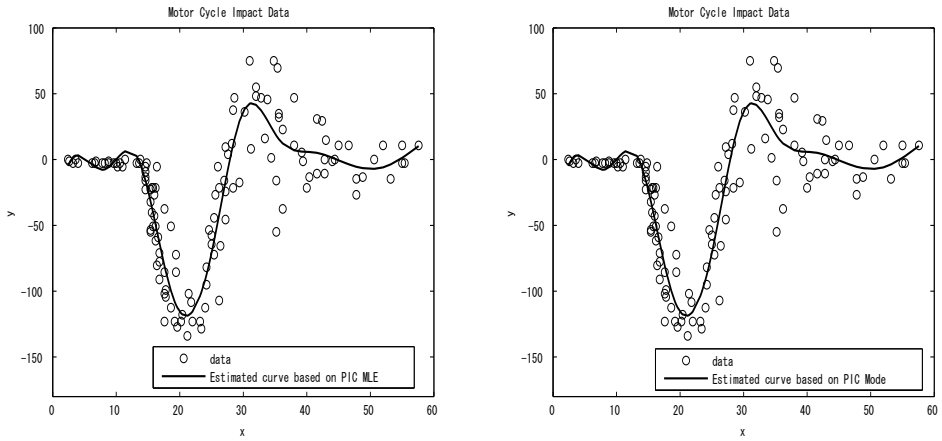

Figure 1: The motorcycle impact data. The left solid curve is the estimated curve based on $\mathrm{PIC}_{\mathrm{MLE}}$ while the right solid curve is the estimated curve based on $\mathrm{PIC}_{\mathrm{Mode}}$.

impact data (Eilers and Marx, 1996; Härdle, 1990; Silverman, 1985). The motorcycle impact data were simulated to investigate the efficacy of crash helmets and it comprised a series of measurements of head acceleration in units of gravity and times in milliseconds after impact.

We fitted our proposed nonlinear regression model based on cubic B-spline to the motorcycle impact data. Then we chose the number of basis functions m and the hyper-parameter λ that minimize the information criteria $\mathrm{PIC}_{\mathrm{MLE}}$ and $\mathrm{PIC}_{\mathrm{Mode}}$ given by Equation (7) and (8). For the analysis of the motorcycle impact data, we set the candidate values of m and λ to $\{4, ..., 25\}$ and $\left\{10^{10(i-100)/99}; i = 1, ..., 100\right\}$, respectively. The criterion $\mathrm{PIC}_{\mathrm{MLE}}$ selected $m = 13$ and $\lambda = 1.62 \times 10^{-9}$, while $\mathrm{PIC}_{\mathrm{Mode}}$ selected $m = 13$ and $\lambda = 2.59 \times 10^{-9}$. The corresponding fitted curve is shown in Figure 1 (solid curve). We compared our proposed procedure with two types of criteria which were introduced in Section 4.2. Table 1 gives the number of basis functions m, the smoothing parameter γ and the hyper-parameter λ chosen by each model selection criterion. The result shows that the variance estimators $\hat{\sigma}^2$ of all model selection criteria are almost equal to each other. Note that $\hat{y}_i$ is a predictive value which is estimated by each model selection criterion. Table 1 suggests that the estimated curves constructed by $\mathrm{PIC}_{\mathrm{MLE}}$ and $\mathrm{PIC}_{\mathrm{Mode}}$ are almost equal to those selected by other model selection criteria. In the next section, we show that our proposed nonlinear regression modeling is useful in some situations by simulation studies.

5.2. Monte Carlo simulations

We applied our proposed nonlinear regression modeling based on cubic B-spline to the simulated data. Repeated random samples $\{(x_i, y_i); i = 1, ..., n\}$ with $n = 50,\ 100$ and 300 were generated from a true regression model $y_i = u(x_i) + \varepsilon_i$. The design points x_i were uniformly distributed in $[0, 1]$ and the errors ε_i were independently, normally distributed with mean zero and variance τ^2, where the standard deviation is taken as

Table 1: The result of each nonlinear regression modeling for the motorcycle impact data.

	$\mathrm{PIC}_{\mathrm{MLE}}$	$\mathrm{PIC}_{\mathrm{Mode}}$	mAIC	GIC
m	13	13	13	13
γ	-	-	3.51×10^{-6}	2.21×10^{-6}
λ	1.62×10^{-9}	2.59×10^{-9}	-	-
$\hat{\sigma}^{2\dagger}$	464.3025	464.3022	467.3327	469.4488

$^{\dagger}\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n, \quad \hat{y}_i = \hat{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{\phi}(x_i).$

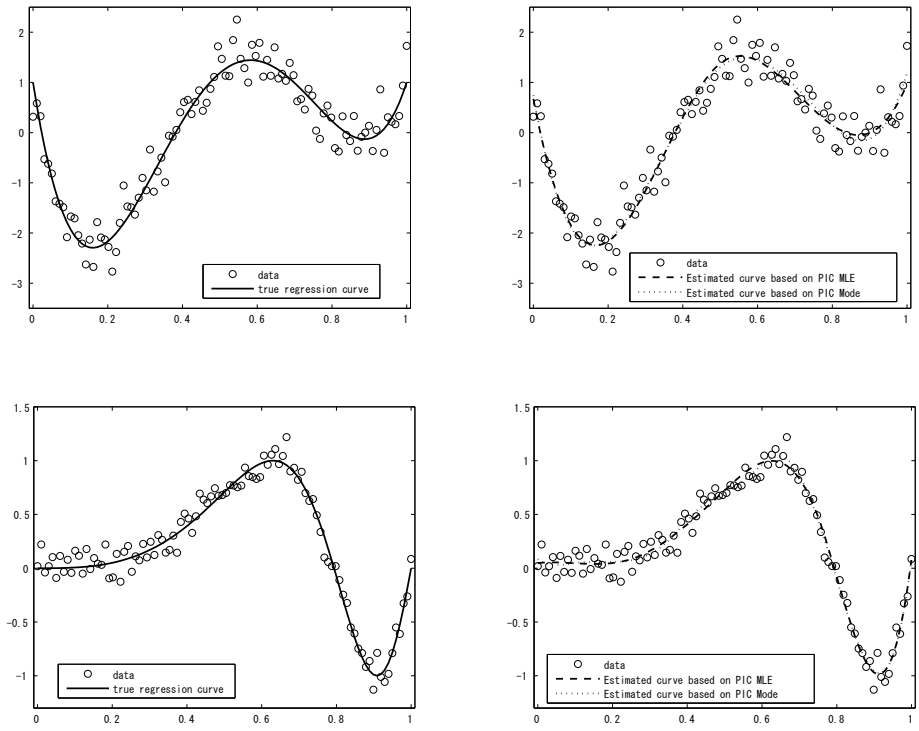

Figure 2: Examples of simulated data for $n = 100$. The solid curve is the true regression curve ($u_1(x)$; left top, $u_2(x)$; left bottom), while the chain curve and dashed curve (right top and bottom) are estimated curve based on $\mathrm{PIC}_{\mathrm{MLE}}$ and $\mathrm{PIC}_{\mathrm{Mode}}$, respectively.

Table 2: Comparison of results for curve fitting: function (a) for $\tau = 0.15R_y$.

Criterion	m mean (SD)	γ mean (SD)	λ mean (SD)	AMSE mean (SD)	APSE mean (SD)
($n = 50$)					
GIC	8.37	3.03×10^{-4}	-	2.20×10^{-2}	1.19×10^{-1}
	(3.06)	(1.70×10^{-2})	-	(1.28×10^{-2})	(2.55×10^{-2})
mAIC	7.61	6.73×10^{-4}	-	2.04×10^{-2}	1.18×10^{-1}
	(2.79)	(2.70×10^{-3})	-	(1.22×10^{-1})	(2.48×10^{-2})
PIC_{MLE}	6.57	-	4.04×10^{-4}	1.79×10^{-2}	1.16×10^{-1}
	(2.12)	-	(2.00×10^{-3})	(1.11×10^{-2})	(2.45×10^{-2})
PIC_{Mode}	7.14	-	1.00×10^{-8}	1.93×10^{-2}	1.17×10^{-1}
	(2.38)	-	(6.29×10^{-23})	(1.19×10^{-2})	(2.45×10^{-2})
($n = 100$)					
GIC	7.88	3.03×10^{-4}	-	1.03×10^{-2}	1.07×10^{-1}
	(2.60)	(1.70×10^{-3})	-	(0.62×10^{-2})	(1.64×10^{-2})
mAIC	7.45	4.71×10^{-4}	-	1.00×10^{-2}	1.07×10^{-1}
	(2.47)	(2.10×10^{-1})	-	(6.30×10^{-3})	(1.67×10^{-2})
PIC_{MLE}	6.60	-	1.00×10^{-8}	9.20×10^{-3}	1.06×10^{-1}
	(1.68)	-	(6.29×10^{-23})	(5.50×10^{-3})	(1.66×10^{-2})
PIC_{Mode}	7.16	-	1.00×10^{-8}	9.60×10^{-3}	1.07×10^{-1}
	(2.14)	-	(6.29×10^{-23})	(5.90×10^{-3})	(1.69×10^{-2})
($n = 300$)					
GIC	7.53	3.36×10^{-5}	-	3.20×10^{-3}	1.01×10^{-1}
	(1.86)	(5.83×10^{-4})	-	(1.90×10^{-3})	(9.00×10^{-3})
mAIC	7.47	3.36×10^{-5}	-	3.20×10^{-3}	1.01×10^{-1}
	(1.76)	(5.83×10^{-4})	-	(1.90×10^{-3})	(9.00×10^{-3})
PIC_{MLE}	7.24	-	1.00×10^{-8}	3.10×10^{-3}	1.01×10^{-1}
	(1.52)	-	(6.29×10^{-23})	(1.80×10^{-3})	(8.90×10^{-3})
PIC_{Mode}	7.43	-	1.00×10^{-8}	3.20×10^{-3}	1.01×10^{-1}
	(1.71)	-	(6.29×10^{-23})	(1.90×10^{-3})	(9.00×10^{-3})

Table 3: Comparison of results for curve fitting: function (a) for $\tau = 0.3R_y$.

Criterion	m mean(SD)	γ mean(SD)	λ mean(SD)	AMSE mean(SD)	APSE mean(SD)
($n = 50$)					
GIC	8.46	1.60×10^{-3}	-	3.28×10^{-1}	1.8932
	(3.57)	(3.70×10^{-3})	-	(2.15×10^{-1})	(4.14×10^{-1})
mAIC	8.14	2.90×10^{-3}	-	2.85×10^{-1}	1.8539
	(3.53)	(4.60×10^{-3})	-	(1.87×10^{-1})	(3.88×10^{-1})
PIC_{MLE}	9.75	-	7.80×10^{-3}	2.47×10^{-1}	1.8197
	(4.08)	-	(6.80×10^{-3})	(1.28×10^{-1})	(3.73×10^{-1})
PIC_{Mode}	7.94	-	3.10×10^{-3}	2.69×10^{-1}	1.8397
	(3.54)	-	(4.70×10^{-3})	(1.80×10^{-1})	(3.86×10^{-1})
($n = 100$)					
GIC	7.13	7.40×10^{-4}	-	1.42×10^{-1}	1.6976
	(3.04)	(2.60×10^{-3})	-	(1.00×10^{-1})	(2.64×10^{-1})
mAIC	6.77	9.09×10^{-4}	-	1.32×10^{-1}	1.6876
	(2.86)	(2.90×10^{-23})	-	(9.59×10^{-2})	(2.61×10^{-1})
PIC_{MLE}	6.93	-	2.00×10^{-3}	1.21×10^{-1}	1.6794
	(3.15)	-	(4.00×10^{-3})	(7.97×10^{-2})	(2.60×10^{-1})
PIC_{Mode}	6.55	-	7.40×10^{-4}	1.27×10^{-1}	1.6850
	(2.69)	-	(2.60×10^{-3})	(9.22×10^{-2})	(2.63×10^{-1})
($n = 300$)					
GIC	6.98	1.50×10^{-4}	-	4.49×10^{-2}	1.6200
	(2.56)	(4.11×10^{-4})	-	(3.03×10^{-2})	(0.14)
mAIC	6.78	1.40×10^{-4}	-	4.37×10^{-2}	1.6199
	(2.46)	(3.90×10^{-4})	-	(2.88×10^{-2})	(0.14)
PIC_{MLE}	6.50	-	1.51×10^{-4}	4.14×10^{-2}	1.6178
	(2.20)	-	(4.61×10^{-4})	(2.77×10^{-2})	(0.14)
PIC_{Mode}	6.59	-	1.13×10^{-4}	4.26×10^{-2}	1.6188
	(2.2904)	-	(3.50×10^{-4})	(2.88×10^{-2})	(0.14)

Table 4: Comparison of results for curve fitting: function (b) for $\tau = 0.15R_y$.

Criterion	m mean(SD)	γ mean(SD)	λ mean(SD)	AMSE mean(SD)	APSE mean(SD)
($n = 50$)					
GIC	10.45	7.90×10^{-3}	-	2.10×10^{-3}	9.90×10^{-3}
	(3.08)	(2.38×10^{-2})	-	(9.78×10^{-4})	(2.10×10^{-3})
mAIC	9.68	1.57×10^{-2}	-	2.00×10^{-3}	9.80×10^{-3}
	(3.06)	(3.47×10^{-2})	-	(9.86×10^{-4})	(2.10×10^{-3})
PIC_{MLE}	7.89	-	1.00×10^{-8}	1.70×10^{-3}	9.50×10^{-3}
	(1.76)	-	(6.29×10^{-23})	(8.99×10^{-4})	(2.00×10^{-3})
PIC_{Mode}	9.04	-	1.00×10^{-8}	1.90×10^{-3}	9.80×10^{-3}
	(2.70)	-	(6.29×10^{-23})	(8.59×10^{-4})	(2.10×10^{-3})
($n = 100$)					
GIC	9.70	4.10×10^{-3}	-	1.00×10^{-3}	9.00×10^{-3}
	(2.96)	(1.27×10^{-2})	-	(4.87×10^{-4})	(1.30×10^{-3})
mAIC	9.05	5.70×10^{-3}	-	1.00×10^{-3}	9.00×10^{-3}
	(2.82)	(1.64×10^{-2})	-	(4.90×10^{-4})	(1.30×10^{-3})
PIC_{MLE}	8.04	-	1.00×10^{-8}	9.11×10^{-4}	8.90×10^{-3}
	(1.99)	-	(6.29×10^{-23})	(4.6×10^{-4})	(1.30×10^{-3})
PIC_{Mode}	8.60	-	1.00×10^{-8}	9.56×10^{-4}	8.90×10^{-3}
	(2.49)	-	(6.29×10^{-23})	(4.73×10^{-4})	(1.40×10^{-3})
($n = 300$)					
GIC	10.31	8.75×10^{-4}	-	3.87×10^{-4}	8.40×10^{-3}
	(2.96)	(3.60×10^{-3})	-	(1.54×10^{-4})	(7.49×10^{-4})
mAIC	9.86	9.76×10^{-4}	-	3.85×10^{-4}	8.40×10^{-3}
	(2.92)	(4.00×10^{-3})	-	(1.50×10^{-4})	(7.49×10^{-4})
PIC_{MLE}	9.41	-	1.00×10^{-8}	3.74×10^{-4}	8.40×10^{-3}
	(2.79)	-	(6.29×10^{-23})	(1.43×10^{-4})	(7.50×10^{-4})
PIC_{Mode}	9.70	-	1.00×10^{-8}	3.77×10^{-4}	8.40×10^{-3}
	(2.84)	-	(6.29×10^{-23})	(1.48×10^{-4})	(7.48×10^{-4})

$\tau = 0.15R_y$ or $0.3R_y$ with R_y being the range of $u(x)$ over $x \in [0, 1]$. For the analysis of the simulated data, we partitioned the interval $[10^{-8}, 10^0]$ into 99 equal subintervals, to obtain candidate values $\lambda_1 = \gamma_1 = 10^{-8}$, $\lambda_2 = \gamma_2 = 10^{-8} + h$, ..., $\lambda_{100} = \gamma_{100} = 1$, where $h = (1 - 10^{-8})/99$. We also set the candidate values of m to $\{4, \ldots, 15\}$. We considered the following two cases for the true regression model:

$$(a) \quad u_1(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4,$$
$$(b) \quad u_2(x) = \sin(2\pi x^3).$$

We performed 300 repetitions, and then calculated average mean squared errors (AMSE) defined by $\text{AMSE} = \sum_{i=1}^{n} \{u(x_i) - \hat{y}_i\}^2 / n$, average predictive squared errors (APSE) defined by $\text{APSE}=\sum_{i=1}^{n} \{z_i - \hat{y}_i\}^2 / n$ and deviations to assess the goodness of fit, respectively. Here, z is a future observation that is generated from true regression models. We compared the performance of nonlinear regression models based on PIC_{MLE} and PIC_{Mode} with that of GIC in (9) and mAIC in (10). Table 2, 3, 4 and 5 display simulation results with AMSE, APSE, the number of basis function m, smoothing parameter γ and hyper-parameter λ. Figure 2 shows the fitted curves by PIC_{MLE} and PIC_{Mode}.

Simulation results may be summarized as follows. For the regression function (a), our proposed modeling procedure performs well in terms of minimizing the AMSE and the APSE. In the regression function (b) in small and middle sample sizes (i.e., $n = 50$ and $n = 100$), the performance of our methods is competitive to other methods, while, in large sample size (i.e., $n = 300$), the proposed methods provide smaller values of AMSE and APSE than other methods (mAIC and GIC). From these descriptions, we conclude

Table 5: Comparison of results for curve fitting: function (b) for $\tau = 0.3R_y$.

Criterion	m mean(SD)	γ mean(SD)	λ mean(SD)	AMSE mean(SD)	APSE mean(SD)
($n = 50$)					
GIC	9.85	8.10×10^{-3}	-	2.89×10^{-2}	1.54×10^{-1}
	(2.95)	(1.48×10^{-2})	-	(1.55×10^{-2})	(3.28×10^{-2})
mAIC	9.80	1.41×10^{-2}	-	2.72×10^{-2}	1.53×10^{-1}
	(2.97)	(1.89×10^{-2})	-	(1.40×10^{-2})	(3.21×10^{-2})
$\mathrm{PIC_{MLE}}$	13.95	-	1.32×10^{-2}	2.89×10^{-2}	1.54×10^{-1}
	(1.35)	-	(5.00×10^{-3})	(1.14×10^{-2})	(3.18×10^{-2})
$\mathrm{PIC_{Mode}}$	11.84	-	8.10×10^{-3}	2.94×10^{-2}	1.55×10^{-1}
	(2.86)	-	(5.40×10^{-3})	(1.32×10^{-4})	(3.21×10^{-2})
($n = 100$)					
GIC	8.96	4.70×10^{-3}	-	1.38×10^{-2}	1.41×10^{-1}
	(2.68)	(9.70×10^{-3})	-	(7.90×10^{-3})	(2.18×10^{-2})
mAIC	8.88	6.20×10^{-3}	-	1.35×10^{-2}	1.40×10^{-1}
	(2.64)	(1.07×10^{-2})	-	(7.70×10^{-3})	(2.19×10^{-2})
$\mathrm{PIC_{MLE}}$	13.21	-	9.60×10^{-3}	1.48×10^{-2}	1.41×10^{-1}
	(2.09)	-	(2.40×10^{-3})	(6.20×10^{-3})	(2.13×10^{-2})
$\mathrm{PIC_{Mode}}$	10.22	-	4.90×10^{-3}	1.42×10^{-2}	1.41×10^{-1}
	(3.13)	-	(5.10×10^{-3})	(7.60×10^{-3})	(2.18×10^{-2})
($n = 300$)					
GIC	8.58	1.70×10^{-3}	-	2.20×10^{-3}	6.45×10^{-2}
	(2.45)	(4.40×10^{-3})	-	(1.20×10^{-3})	(5.80×10^{-3})
mAIC	8.50	2.00×10^{-3}	-	2.20×10^{-3}	6.44×10^{-2}
	(2.45)	(4.70×10^{-3})	-	(1.30×10^{-3})	(5.80×10^{-3})
$\mathrm{PIC_{MLE}}$	7.74	-	1.00×10^{-8}	2.00×10^{-3}	6.43×10^{-2}
	(1.59)	-	(6.29×10^{-23})	(1.30×10^{-3})	(5.80×10^{-3})
$\mathrm{PIC_{Mode}}$	7.84	-	1.00×10^{-8}	2.00×10^{-3}	6.43×10^{-2}
	(1.70)	-	(6.29×10^{-23})	(1.30×10^{-3})	(5.80×10^{-3})

that our proposed modeling procedure may be more useful than previously developed procedures in practical situations.

6. Concluding Remarks

We considered the problem of evaluating the predictive distributions for nonlinear regression models based on basis expansions. In order to select the optimal values of the hyper-parameters included in the prior distribution and the number of basis functions, we obtained information criteria for the evaluation of the Bayesian predictive distributions. The simulation results suggest that our proposed procedure provides improvements from the viewpoint of the mean squared errors and the predictive squared errors, and yields stable prediction results. A further research is to identify the shape of the prior distribution for the parameter σ^2 based on observed data; i.e., selecting the hyper-parameters ν_0 and η_0 in the prior distribution objectively.

Acknowledgement

The authors would like to thank Professor Yoshiyuki Ninomiya of Kyushu University for his helpful and constructive comments and suggestions. We are grateful to the anonymous reviewer for careful reading of the manuscript and his helpful comments.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

de Boor, C. (2001). *A Practical Guide to Splines. Revised Edition*. Springer.

Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika*, **73**, 323–332.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.

Figueiredo, M. A. T. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**, 1150–1159.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Communications in Statistics-Theory and Methods*, **26**, 2223–2246.

Kohn, R., Smith, M. and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, **11**, 313–322.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.

Minka, T. (2000). Bayesian linear regression. *Technical report, MIT*, 2000.

Murata, N., Yoshizawa, S. and Amari, S. (1994). Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, **5**, 865–872.

Shibata, R. (1989). An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society Series B*, **36**, 1–52.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Received March 8, 2012
Revised August 28, 2012