Ann. Inst. Statist. Math.
Vol. 55, No. 4, 671–687 (2003)

SELECTION OF SMOOTHING PARAMETERS IN B-SPLINE NONPARAMETRIC REGRESSION MODELS USING INFORMATION CRITERIA

SEIYA IMOTO[1] AND SADANORI KONISHI[2]

[1] *Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan*
[2] *Graduate School of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan*

(Received October 4, 2001; revised November 15, 2002)

Abstract. We consider the use of B-spline nonparametric regression models estimated by the maximum penalized likelihood method for extracting information from data with complex nonlinear structure. Crucial points in B-spline smoothing are the choices of a smoothing parameter and the number of basis functions, for which several selectors have been proposed based on cross-validation and Akaike information criterion known as AIC. It might be however noticed that AIC is a criterion for evaluating models estimated by the maximum likelihood method, and it was derived under the assumption that the true distribution belongs to the specified parametric model. In this paper we derive information criteria for evaluating B-spline nonparametric regression models estimated by the maximum penalized likelihood method in the context of generalized linear models under model misspecification. We use Monte Carlo experiments and real data examples to examine the properties of our criteria including various selectors proposed previously.

Key words and phrases: B-spline smoothing, generalized linear model, information criteria, smoothing parameter selection.

1. Introduction

Smoothing methods in nonparametric regression have drawn a large amount of attention in recent years. Many different methods such as kernel and spline smoothing have been proposed for nonparametric curve fitting (see, e.g., Silverman (1986), Eubank (1988), Härdle (1990), Green and Silverman (1994), Kitagawa and Gersch (1996), Simonoff (1996)). In this paper we consider the problem of constructing B-spline nonparametric regression models estimated by the maximum penalized likelihood method in generalized linear models (McCullagh and Nelder (1989)).

Crucial points of model construction are the choices of a smoothing parameter and the number of basis functions (or knots), for which several attempts have been made based on cross-validation (Stone (1974)), generalized cross-validation (Craven and Wahba (1979)) and Akaike's (1973, 1974) information criterion AIC. Eilers and Marx (1996) replaced the number of free parameters in AIC with the trace of a hat matrix, and introduced an information criterion for evaluating B-spline nonparametric regression models with Gaussian noise. Recently Hurvich *et al.* (1998) proposed an improved version of the AIC for smoothing parameter selection in the context of nonparametric regression.

In the information criteria proposed in the literature, attention has been focused on the bias correction of log-likelihood for a model estimated by the maximum penalized likelihood method. AIC is however derived under the assumptions that the parametric model is estimated by the maximum likelihood and that the true distribution belongs to a parametric family of densities. Hence the problem still remains to be done in constructing an information-theoretic criterion for evaluating B-spline nonparametric regression models estimated by the maximum penalized likelihood method. We also noticed that in practice it is usually difficult to obtain precise information on distributional form and data structures from a finite number of observations. It is therefore of interest to construct a criterion under model misspecification.

The purpose of the present paper is to derive information criteria for evaluating B-spline nonparametric regression models estimated by the maximum penalized likelihood under model misspecification both for distributional and structural assumptions. Section 2 describes B-spline nonparametric regression models in the context of generalized linear models. Section 3 presents information criteria in model selection and evaluation.

The information criteria proposed are applied to choose the smoothing parameter and the number of basis functions in nonparametric curve fitting. We also consider the use of Akaike's (1980a, 1980b) Bayesian information criterion as a smoothing parameter selector. In Section 4 Monte Carlo experiments are conducted to examine the performance of the proposed criteria and to compare various types of procedures. We use real data examples to investigate the properties of the proposed procedure in practice.

2. B-spline nonparametric regression

2.1 *Model*

Suppose that we have n observations $\{(x_\alpha, y_\alpha); \alpha = 1, \ldots, n\}$ and that the responses y_α are generated from an unknown true distribution $G(y \mid x)$ having probability density $g(y \mid x)$. To draw information from the data, we use the exponential family of densities

$$f(y_\alpha \mid x_\alpha; \xi_\alpha, \phi) = \exp\left\{\frac{y_\alpha \xi_\alpha - u(\xi_\alpha)}{\phi} + v(y_\alpha, \phi)\right\}, \tag{2.1}$$

where $u(\cdot)$ and $v(\cdot, \cdot)$ are specific functions and ξ_α and ϕ are unknown parameters. Under the generalized linear model framework, the conditional expectation $\mathrm{E}[Y_\alpha \mid x_\alpha] = \mu_\alpha$ $(= u'(\xi_\alpha))$ is related to the predictor η_α by $h(\mu_\alpha) = \eta_\alpha$, where $h(\cdot)$ is a link function. It is assumed that the predictor is

$$h(u'(\xi_\alpha)) = \eta_\alpha = \sum_{j=1}^{m} \gamma_j B_j(x_\alpha), \qquad \alpha = 1, \ldots, n \tag{2.2}$$

where $\{B_j(x); j = 1, \ldots, m\}$ $(m < n)$ is a prescribed set of m basis functions. We consider basis functions as B-splines of degree 3, constructed from polynomial pieces. Figure 1 is an example of B-splines of degree 3 with equidistant knots $t_1, \ldots, t_{10}$. For B-splines we refer to de Boor (1978), Dierckx (1993) and Eilers and Marx (1996).

Combining the random component (2.1) and the systematic component (2.2), we have a B-spline nonparametric regression model

$$f(y_\alpha \mid x_\alpha; \boldsymbol{\gamma}, \phi) = \exp\left\{\frac{y_\alpha r(\boldsymbol{\gamma}^T \boldsymbol{b}(x_\alpha)) - s(\boldsymbol{\gamma}^T \boldsymbol{b}(x_\alpha))}{\phi} + v(y_\alpha, \phi)\right\}, \tag{2.3}$$

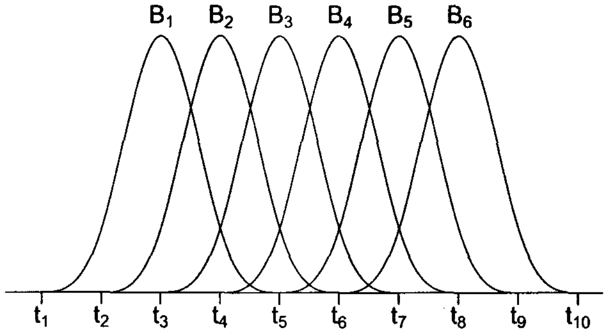


Fig. 1. B-splines of degree 3 with knots $t_1, \dots, t_{10}$.

where $\boldsymbol{b}(x_\alpha) = (B_1(x_\alpha), \dots, B_m(x_\alpha))^T$, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m)^T$, $r(\cdot) = u'^{-1} \circ h^{-1}(\cdot)$ and $s(\cdot) = u \circ u'^{-1} \circ h^{-1}(\cdot)$.

2.2 *Estimation*

The unknown parameters $\boldsymbol{\gamma}$ and ϕ in (2.3) are estimated by a suitable estimation procedure. If one uses the predictor η_α with small number of basis functions, then the parameters may be estimated by maximum likelihood. In practical situations, however, it often happens that a model with a small number of parameters cannot satisfactorily approximate the data, and we employ a model with more parameters.

One problem is that the maximum likelihood method then yields unstable parameter estimates and leads to overfitting. In such a case the adopted model is estimated by maximizing the penalized log-likelihood function

$$l_\lambda(\boldsymbol{\gamma}, \phi) = \sum_{\alpha=1}^{n} \log f(y_\alpha \mid x_\alpha; \boldsymbol{\gamma}, \phi) - \frac{\lambda}{2} n \text{ (roughness penalty)},$$

where λ is a smoothing parameter that controls the smoothness of a regression curve. The maximum penalized likelihood method was originally introduced by Good and Gaskins (1971) and has been investigated by Silverman (1985), Green (1987), Green and Silverman (1994) and references therein.

For B-spline regression model, Eilers and Marx (1996) proposed a penalty based on finite differences of the coefficients of adjacent B-splines in the form

$$\lambda \sum_{j=k+1}^{m} (\Delta^k \gamma_j)^2 = \lambda \boldsymbol{\gamma}^T D_k^T D_k \boldsymbol{\gamma},$$

where Δ is the difference operator such as $\Delta\gamma_j = \gamma_j - \gamma_{j-1}$ and D_k is an $(m-k) \times m$ matrix representation given by

$$D_k = \begin{pmatrix} (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k & 0 & \cdots & 0 \\ 0 & (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k \end{pmatrix},$$

with ${}_nC_k = n!/\{k!(n-k)!\}$. We estimate the unknown parameters $\boldsymbol{\gamma}$ and ϕ by maximizing the penalized log-likelihood function

$$l_\lambda(\boldsymbol{\gamma},\phi) = \sum_{\alpha=1}^{n}\left\{\frac{y_\alpha r(\boldsymbol{\gamma}^T\boldsymbol{b}(x_\alpha)) - s(\boldsymbol{\gamma}^T\boldsymbol{b}(x_\alpha))}{\phi} + v(y_\alpha,\phi)\right\} - \frac{\lambda}{2}n\boldsymbol{\gamma}^T D_k^T D_k\boldsymbol{\gamma}. \tag{2.4}$$

The B-spline nonparametric regression model estimated by the penalized likelihood method was originally introduced by Eilers and Marx (1996) and they called it P-splines.

The maximum penalized likelihood estimate $\hat{\boldsymbol{\gamma}}$ is a solution of the penalized likelihood equation $\partial l_\lambda(\boldsymbol{\gamma},\phi)/\partial\boldsymbol{\gamma} = \mathbf{0}$. This equation is generally nonlinear in $\boldsymbol{\gamma}$, so we use Fisher's scoring algorithm (Nelder and Wedderburn (1972), Green and Silverman (1994)). For fixed values of ϕ, λ and the number of basis functions, the Fisher scoring iterations may be expressed as

$$\boldsymbol{\gamma}^{new} = (B^TWB + n\lambda D_k^T D_k)^{-1}B^TW\boldsymbol{\zeta}, \tag{2.5}$$

where $B = (\boldsymbol{b}(x_1),\ldots,\boldsymbol{b}(x_n))^T$, W is an $n\times n$ diagonal matrix with i-th diagonal element $w_{ii} = \{\phi u''(\xi_i)h'(\mu_i)^2\}^{-1}$ and $\boldsymbol{\zeta}$ an n dimensional vector with $\zeta_i = (y_i-\mu_i)h'(\mu_i) + \boldsymbol{\gamma}^T\boldsymbol{b}(x_i)$. In each Fisher scoring step $\boldsymbol{\gamma}$ is updated to $\boldsymbol{\gamma}^{new}$ by (2.5) until a suitable convergence criterion is satisfied. If $h(\cdot)$ is the canonical link, W and $\boldsymbol{\zeta}$ are simplified to $w_{ii} = u''(\xi_i)/\phi$ and $\zeta_i = (y_i-\mu_i)/u''(\boldsymbol{\gamma}^T\boldsymbol{b}(x_i)) + \boldsymbol{\gamma}^T\boldsymbol{b}(x_i)$.

Suppose that the observations y_α are independently and normally distributed with mean μ_α and variance σ^2. Then the B-spline nonparametric regression model with Gaussian noise is

$$f_N(y_\alpha \mid x_\alpha;\boldsymbol{\gamma},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\{y_\alpha - \boldsymbol{\gamma}^T\boldsymbol{b}(x_\alpha)\}^2}{2\sigma^2}\right], \tag{2.6}$$

and the maximum penalized likelihood estimates of $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}^2$ are

$$\hat{\boldsymbol{\gamma}} = (B^TB + n\beta D_k^T D_k)^{-1}B^T\boldsymbol{y}, \qquad \hat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{y} - B\hat{\boldsymbol{\gamma}}\|^2, \tag{2.7}$$

where $\beta = \hat{\sigma}^2\lambda$ for a given value of λ and $\boldsymbol{y} = (y_1,\ldots,y_n)^T$.

3. Information criteria for model evaluation

3.1 *Proposed criterion*

We recall that the independent responses $y_1,\ldots,y_n$ are generated from an unknown true distribution $G(y \mid x)$ having probability density $g(y \mid x)$, and that the statistical model $f(y \mid x;\hat{\boldsymbol{\gamma}},\hat{\phi})$ is constructed within the generalized linear model framework, using B-splines. We will assess the closeness of $f(y \mid x;\hat{\boldsymbol{\gamma}},\hat{\phi})$ and the true model $g(y \mid x)$ from a predictive point of view.

Suppose that $z_1,\ldots,z_n$ are future observations for the response variable Y drawn from $g(y \mid x)$. Let $f(\boldsymbol{z} \mid X;\hat{\boldsymbol{\theta}}) = \prod_{\alpha=1}^n f(z_\alpha \mid x_\alpha;\hat{\boldsymbol{\gamma}},\hat{\phi})$ and $g(\boldsymbol{z} \mid X) = \prod_{\alpha=1}^n g(z_\alpha \mid x_\alpha)$. Then we use as an overall measure of the divergence of $f(\boldsymbol{z} \mid X;\hat{\boldsymbol{\theta}})$ from $g(\boldsymbol{z} \mid X)$ the Kullback-Leibler information (Kullback and Leibler (1951))

$$\begin{aligned} I\{g,f\} &= \mathrm{E}_{G(\boldsymbol{z}|X)}\left[\log\frac{g(\boldsymbol{z} \mid X)}{f(\boldsymbol{z} \mid X;\hat{\boldsymbol{\theta}})}\right] \\ &= \mathrm{E}_{G(\boldsymbol{z}|X)}\left[\log g(\boldsymbol{z} \mid X)\right] - \mathrm{E}_{G(\boldsymbol{z}|X)}\left[\log f(\boldsymbol{z} \mid X;\hat{\boldsymbol{\theta}})\right] \end{aligned} \tag{3.1}$$

conditional on $\hat{\boldsymbol{\theta}}$. The first term of (3.1) depends only on the true model and does not relate to model evaluation. So it is clear that the second term of (3.1) is essential for model evaluation based on the Kullback-Leibler information. This implies that the minimization of $I\{g, f\}$ is equivalent to the maximization of the expected log-likelihood $\mathrm{E}_{G(\boldsymbol{z}|X)}[\log f(\boldsymbol{z} \mid X; \hat{\boldsymbol{\theta}})]$.

We estimate the expected log-likelihood $\mathrm{E}_{G(\boldsymbol{z}|X)}[\log f(\boldsymbol{z} \mid X; \hat{\boldsymbol{\theta}})]$ by the (average) log-likelihood $\log f(\boldsymbol{y} \mid X; \hat{\boldsymbol{\theta}})/n$. The log-likelihood generally provides an overestimation of the expected log-likelihood. We therefore consider the bias correction of the log-likelihood. By correcting a bias of the log-likelihood in the estimation of the expected log-likelihood, we have an information criterion

$$-2\sum_{\alpha=1}^{n} \log f(y_\alpha \mid x_\alpha; \hat{\boldsymbol{\theta}}) + 2\widehat{\mathrm{ASB}},$$

where $\widehat{\mathrm{ASB}}$ is in general an estimate of the asymptotic bias of

$$\mathrm{E}_{G(\boldsymbol{y}|X)}[\log f(\boldsymbol{y} \mid X; \hat{\boldsymbol{\theta}}) - \mathrm{E}_{G(\boldsymbol{z}|X)}[\log f(\boldsymbol{z} \mid X; \hat{\boldsymbol{\theta}})]].$$

Under the assumption that the specified family of probability distributions does not contain the true model generating the data, Konishi and Kitagawa ((1996), p. 877) derived the asymptotic bias as a function of the empirical influence function of the estimator and the score function of the parametric model (see also Konishi (1999)). The result is given by

$$\text{(3.2)} \qquad \mathrm{ASB} = \mathrm{tr}\left\{ \int \boldsymbol{T}^{(1)}(z \mid x; G) \frac{\partial \log f(z \mid x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} \mathrm{d}G \right\},$$

where $\boldsymbol{T}^{(1)}(z \mid x; G)$ is the influence function of the maximum penalized likelihood estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$ and $\hat{G}$ is the empirical distribution.

The influence function of the estimator $\hat{\boldsymbol{\theta}} = (\hat{\gamma}^T, \hat{\phi})^T$ in our model $f(\boldsymbol{y} \mid X; \hat{\boldsymbol{\theta}})$ is given as follows: Let $\boldsymbol{T}(\cdot)$ be the p dimensional functional implicitly defined by

$$\text{(3.3)} \qquad \int \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f(y \mid x; \boldsymbol{\theta}) - \frac{\lambda}{2}\gamma^T D_k^T D_k \gamma \right\} \bigg|_{\boldsymbol{\theta}=\boldsymbol{T}(G)} \mathrm{d}G = 0,$$

where $\boldsymbol{\theta} = (\gamma^T, \phi)^T$ and G is the joint distribution of (y, x) constructed formally. By replacing G in (3.3) by the empirical distribution function $\hat{G}$ based on the observations, we have

$$\frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f(y_\alpha \mid x_\alpha; \boldsymbol{\theta}) - \frac{\lambda}{2}\gamma^T D_k^T D_k \gamma \right\} \bigg|_{\hat{\boldsymbol{\theta}}=\boldsymbol{T}(\hat{G})} = 0.$$

This implies that the maximum penalized likelihood estimators $\hat{\boldsymbol{\theta}}$ can be written as $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$ for the functional $\boldsymbol{T}(G)$ implicitly defined by (3.3).

Replacing G in (3.3) by $G_\varepsilon = (1-\varepsilon)G + \varepsilon\delta_{(y,x)}$ with $\delta_{(y,x)}$ being a point of mass at (y, x) and differentiating with respect to ε yields the influence function of the maximum penalized likelihood estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$ in the form

$$\text{(3.4)} \qquad \boldsymbol{T}^{(1)}(y \mid x; G) = J_\lambda(G)^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \log f(y \mid x; \boldsymbol{\theta}) - \frac{\lambda}{2}\gamma^T D_k^T D_k \gamma \right\} \bigg|_{\boldsymbol{T}(G)},$$

where

$$J_\lambda(G) = -\int \frac{\partial^2 \left\{ \log f(y \mid x; \boldsymbol{\theta}) - \frac{\lambda}{2}\gamma^T D_k^T D_k \gamma \right\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \mathrm{d}G.$$

The result may be obtained by an argument similar to that in Hampel *et al.* ((1986), p. 101) in which they derived the influence function of an M-estimator.

Then substituting (3.4) in the asymptotic bias (3.2) and using Theorem 2.1 given in Konishi and Kitagawa ((1996), p. 876), we have the following theorem.

THEOREM 3.1. *Let $f(y_\alpha \mid x_\alpha; \boldsymbol{\gamma}, \phi)$ be the B-spline nonparametric regression model defined by* (2.3), *and let $f(y_\alpha \mid x_\alpha; \hat{\boldsymbol{\gamma}}, \hat{\phi})$ be the statistical model fitted by the maximum penalized likelihood method in* (2.4). *Suppose that the exponential family with the linear predictor replaced by B-splines does not necessarily contain the true model generating the data. Then an information criterion for evaluating the statistical model $f(y_\alpha \mid x_\alpha; \hat{\boldsymbol{\gamma}}, \hat{\phi})$ is*

$$\begin{aligned}\mathrm{SPIC}(\lambda, m) = &-2\sum_{\alpha=1}^{n} \left\{ \frac{y_\alpha r(\hat{\boldsymbol{\gamma}}^T \boldsymbol{b}(x_\alpha)) - s(\hat{\boldsymbol{\gamma}}^T \boldsymbol{b}(x_\alpha))}{\hat{\phi}} + v(y_\alpha, \hat{\phi}) \right\} \\ &+ 2\,\mathrm{tr}\{I_\lambda(\hat{G}) J_\lambda(\hat{G})^{-1}\},\end{aligned}$$

where $I_\lambda(\hat{G})$ and $J_\lambda(\hat{G})$ are the $(m+1) \times (m+1)$ matrices

$$\begin{aligned} I_\lambda(\hat{G}) &= \frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial \left\{ \log f(y_\alpha \mid x_\alpha; \boldsymbol{\gamma}, \phi) - \frac{\lambda}{2}\gamma^T D_k^T D_k \gamma \right\}}{\partial \boldsymbol{\theta}} \\ &\qquad \cdot \left. \frac{\partial \log f(y_\alpha \mid x_\alpha; \boldsymbol{\gamma}, \phi)}{\partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= \frac{1}{n\hat{\phi}} \begin{pmatrix} B^T \Lambda / \hat{\phi} - \lambda D_k^T D_k \hat{\gamma} \mathbf{1}_n^T \\ \boldsymbol{p}^T \end{pmatrix} (\Lambda B, \quad \hat{\phi} \boldsymbol{p}), \end{aligned} \tag{3.5}$$

$$\begin{aligned} J_\lambda(\hat{G}) &= -\frac{1}{n}\sum_{\alpha=1}^{n} \left. \frac{\partial^2 \left\{ \log f(y_\alpha \mid x_\alpha; \boldsymbol{\gamma}, \phi) - \frac{\lambda}{2}\gamma^T D_k^T D_k \gamma \right\}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= \frac{1}{n\hat{\phi}} \begin{pmatrix} B^T \Gamma B + n\hat{\phi}\lambda D_k^T D_k, & B^T \Lambda \mathbf{1}_n / \hat{\phi} \\ \mathbf{1}_n^T \Lambda B / \hat{\phi}, & -\hat{\phi} \boldsymbol{q}^T \mathbf{1}_n \end{pmatrix}. \end{aligned} \tag{3.6}$$

Here Λ and Γ are $n \times n$ diagonal matrices with i-th diagonal elements

$$\Lambda_{ii} = \frac{y_i - \hat{\mu}_i}{u''(\hat{\xi}_i) h'(\hat{\mu}_i)},$$

$$\Gamma_{ii} = \frac{(y_i - \hat{\mu}_i)\{u'''(\hat{\xi}_i) h'(\hat{\mu}_i) + u''(\hat{\xi}_i)^2 h''(\hat{\mu}_i)\}}{\{u''(\hat{\xi}_i) h'(\hat{\mu}_i)\}^3} + \frac{1}{u''(\hat{\xi}_i) h'(\hat{\mu}_i)^2},$$

respectively, and $\mathbf{1}_n = (1, \ldots, 1)^T$, $\boldsymbol{p}$ and $\boldsymbol{q}$ are n dimensional vectors with i-th elements

$$p_i = -\frac{y_i r(\hat{\boldsymbol{\gamma}}^T \boldsymbol{b}(x_i)) - s(\hat{\boldsymbol{\gamma}}^T \boldsymbol{b}(x_i))}{\hat{\phi}^2} + \left. \frac{\partial}{\partial \phi} v(y_i, \phi) \right|_{\phi=\hat{\phi}}, \qquad q_i = \left. \frac{\partial p_i}{\partial \phi} \right|_{\phi=\hat{\phi}}.$$

Canonical link functions relate the parameter ξ_α in the exponential family (2.1) directly to the predictor $\eta_\alpha = \sum_{j=1}^m \gamma_j B_j(x_\alpha)$ in (2.2), and lead to

$$f_{cl}(y_\alpha \mid x_\alpha; \hat{\gamma}, \hat{\phi}) = \exp\left\{ \frac{y_\alpha \hat{\gamma}^T \boldsymbol{b}(x_\alpha) - u(\hat{\gamma}^T \boldsymbol{b}(x_\alpha))}{\hat{\phi}} + v(y_\alpha, \hat{\phi}) \right\}. \tag{3.7}$$

Then we have the following theorem.

THEOREM 3.2. *Let h be the canonical link function, so $h(\cdot) = u'^{-1}(\cdot)$. Then an information criterion for evaluating the statistical model $f_{cl}(y_\alpha \mid x_\alpha; \hat{\gamma}, \hat{\phi})$ given by* (3.7) *is*

$$\begin{aligned} \mathrm{SPIC}_{Cl} = & -2\sum_{\alpha=1}^{n} \left\{ \frac{y_\alpha \hat{\gamma}^T \boldsymbol{b}(x_\alpha) - u(\hat{\gamma}^T \boldsymbol{b}(x_\alpha))}{\hat{\phi}} + v(y_\alpha, \hat{\phi}) \right\} \\ & + 2\,\mathrm{tr}\{ I_\lambda^{(Cl)}(\hat{G}) J_\lambda^{(Cl)}(\hat{G})^{-1} \}, \end{aligned}$$

where $I_\lambda^{(Cl)}(\hat{G})$ and $J_\lambda^{(Cl)}(\hat{G})$ are defined by (3.5) *and* (3.6) *with*

$$\begin{aligned} \Lambda_{ii} &= y_i - u'(\hat{\gamma}^T \boldsymbol{b}(x_i)), \qquad \Gamma_{ii} = u''(\hat{\gamma}^T \boldsymbol{b}(x_i)), \\ p_i &= -\frac{y_i \hat{\gamma}^T \boldsymbol{b}(x_i) - u(\hat{\gamma}^T \boldsymbol{b}(x_i))}{\hat{\phi}^2} + \frac{\partial}{\partial \phi} v(y_i, \phi) \bigg|_{\phi=\hat{\phi}}, \\ q_i &= 2\frac{y_i \hat{\gamma}^T \boldsymbol{b}(x_i) - u(\hat{\gamma}^T \boldsymbol{b}(x_i))}{\hat{\phi}^3} + \frac{\partial^2}{\partial \phi^2} v(y_i, \phi) \bigg|_{\phi=\hat{\phi}}. \end{aligned}$$

We choose the value of a smoothing parameter λ and the number of basis functions m which minimize the information criterion SPIC.

Ordinarily, P-splines transfer the issue of the number and the position of knots into the choice of the smoothing parameter. Eilers and Marx (1996) employed a modest number of knots and concentrated on the choice of the smoothing parameter. In fact, P-spline procedure is a useful tool for fitting a curve to data with nonlinear structure. We consider the number of knots (or basis functions) as an unknown parameter, since it may relate to the stability of the estimated model. Also the information criteria are constructed as asymptotically unbiased estimators of the expected log-likelihood under model misspecification. Hence we consider the problem of choosing not only the smoothing parameter but also the number of basis functions. We illustrate the procedure in Section 4.

Example 1. Suppose that the observations y_α are independently and normally distributed with mean μ_α and variance σ^2. Then the B-spline nonparametric regression model with Gaussian noise (2.6) estimated by the maximum penalized likelihood method can be expressed as $f_N(y_\alpha \mid x_\alpha; \hat{\gamma}, \hat{\sigma}^2)$, where $\hat{\gamma}$ and $\hat{\sigma}^2$ are given by (2.7). Taking $u(\hat{\xi}_\alpha) = \hat{\xi}_\alpha^2/2$, $\hat{\phi} = \hat{\sigma}^2$ and $v(y_\alpha, \hat{\sigma}^2) = -(y_\alpha/\hat{\sigma})^2/2 - \log(\hat{\sigma}\sqrt{2\pi})$ in Theorem 3.2, we have the following information criterion for evaluating the statistical model $f_N(y_\alpha \mid x_\alpha; \hat{\gamma}, \hat{\sigma}^2)$,

$$\mathrm{SPIC}_N = n\log\hat{\sigma}^2 + n\log(2\pi) + n + 2\,\mathrm{tr}\{ I_\lambda^{(N)}(\hat{G}) J_\lambda^{(N)}(\hat{G})^{-1} \}, \tag{3.8}$$

where $I_\lambda^{(N)}(\hat{G})$ and $J_\lambda^{(N)}(\hat{G})$ are given by (3.5) and (3.6) with

$$\Lambda_{ii} = y_i - \hat{\gamma}^T \boldsymbol{b}(x_i), \quad \Gamma_{ii} = 1,$$
$$p_i = \{y_i - \hat{\gamma}^T \boldsymbol{b}(x_i)\}^2/(2\hat{\sigma}^4) - 1/(2\hat{\sigma}^2), \quad q_i = -\{y_i - \hat{\gamma}^T \boldsymbol{b}(x_i)\}^2/\hat{\sigma}^6 + 1/(2\hat{\sigma}^4).$$

Example 2. Suppose that we have n observations $\{(x_\alpha, y_\alpha), \alpha = 1, \ldots, n\}$, where x_α are explanatory variables and y_α are independent random variables coded as either 0 or 1. Consider the B-spline nonparametric logistic regression model

$$f_L(y_\alpha \mid x_\alpha; \boldsymbol{\gamma}) = \pi(x_\alpha)^{y_\alpha}\{1 - \pi(x_\alpha)\}^{1-y_\alpha},$$

where $\Pr(Y_\alpha = 1 \mid x_\alpha) = \pi(x_\alpha)$, $\Pr(Y_\alpha = 0 \mid x_\alpha) = 1 - \pi(x_\alpha)$ and $\pi(x_\alpha) = 1/\{1 + \exp(-\boldsymbol{\gamma}^T \boldsymbol{b}(x_\alpha))\}$. The m dimensional parameter vector $\boldsymbol{\gamma}$ is estimated by the maximum penalized likelihood method. Taking

$$u(\hat{\xi}_\alpha) = \log\{1 + \exp(\hat{\xi}_\alpha)\}, \quad v(y_\alpha, \phi) = 0, \quad h(\hat{\mu}_\alpha) = \log \frac{\hat{\mu}_\alpha}{1 - \hat{\mu}_\alpha} \quad \text{and} \quad \phi = 1$$

in Theorem 3.2, we have the following information criterion for evaluating the statistical model $f_L(y_\alpha \mid x_\alpha; \hat{\boldsymbol{\gamma}})$,

$$\text{(3.9)} \qquad \begin{aligned} \text{SPIC}_L = {} & 2\sum_{\alpha=1}^{n}[\log\{1 + \exp(\hat{\boldsymbol{\gamma}}^T \boldsymbol{b}(x_\alpha))\} - y_\alpha \hat{\boldsymbol{\gamma}}^T \boldsymbol{b}(x_\alpha)] \\ & + 2\,\text{tr}\{I_\lambda^{(L)}(\hat{G}) J_\lambda^{(L)}(\hat{G})^{-1}\}, \end{aligned}$$

where

$$I_\lambda^{(L)}(\hat{G}) = B^T \Lambda^2 B - \lambda D_k^T D_k \hat{\boldsymbol{\gamma}} \boldsymbol{1}_n^T \Lambda B, \qquad J_\lambda^{(L)}(\hat{G}) = B^T \Gamma B + n\lambda D_k^T D_k,$$

with $\Lambda_{ii} = y_i - 1/\{1 + \exp(-\hat{\boldsymbol{\gamma}}^T \boldsymbol{b}(x_\alpha))\}$ and $\Gamma_{ii} = \exp(\hat{\boldsymbol{\gamma}}^T \boldsymbol{b}(x_\alpha))/\{1 + \exp(\hat{\boldsymbol{\gamma}}^T \boldsymbol{b}(x_\alpha))\}^2$.

3.2 *Other criteria*

The criteria proposed previously may be used as selectors in nonparametric curve fitting. This section describes the use of other criteria for the B-spline nonparametric regression model with Gaussian noise.

(1) Akaike's (1980*a*, 1980*b*) Bayesian information criterion

Akaike (1980*a*, 1980*b*) considered a smoothing problem in the Bayesian framework, and proposed the smoothness priors method based on the likelihood of a Bayesian model. Let $\pi(\boldsymbol{\gamma} \mid \lambda)$ be a prior distribution of the m dimensional parameter vector $\boldsymbol{\gamma}$ in the B-spline nonparametric regression model given by (2.3), where λ (> 0) is a hyperparameter. The hyperparameter corresponds to a smoothing parameter in the penalized log-likelihood function in (2.4).

When the observations $\{(x_\alpha, y_\alpha); \alpha = 1, \ldots, n\}$ are given, the posterior distribution is

$$\text{(3.10)} \quad \pi(\boldsymbol{\gamma} \mid \boldsymbol{y}; \lambda) = \prod_{\alpha=1}^{n} f(y_\alpha \mid x_\alpha; \boldsymbol{\gamma}, \phi)\pi(\boldsymbol{\gamma} \mid \lambda) \Big/ \int \prod_{\alpha=1}^{n} f(y_\alpha \mid x_\alpha; \boldsymbol{\gamma}, \phi)\pi(\boldsymbol{\gamma} \mid \lambda) d\boldsymbol{\gamma}.$$

The integral defining the denominator of equation (3.10)

$$L(\lambda, \phi) = \int \prod_{\alpha=1}^{n} f(y_\alpha \mid x_\alpha; \gamma, \phi)\pi(\gamma \mid \lambda)\mathrm{d}\gamma \tag{3.11}$$

is the likelihood for the unknown parameters λ and ϕ. In order to determine the value of λ, Akaike (1980a, 1980b) considered the maximization of the marginal likelihood (3.11) with respect to λ and ϕ (see also Good (1965)), or equivalently the minimization of

$$\mathrm{ABIC} = -2\log\left\{\int \prod_{\alpha=1}^{n} f(y_\alpha \mid x_\alpha; \gamma, \phi)\pi(\gamma \mid \lambda)\mathrm{d}\gamma\right\}. \tag{3.12}$$

Let $\hat{\lambda}$ and $\hat{\phi}$ be the minimizers of ABIC. Then the estimator of the parameter γ is chosen to maximize $\prod_{\alpha=1}^{n} f(y_\alpha \mid x_\alpha; \gamma, \hat{\phi})\pi(\gamma \mid \hat{\lambda})$ which corresponds to the maximizer of the posterior density (3.10) with respect to γ for fixed $\hat{\lambda}$ and $\hat{\phi}$. A number of successful applications of ABIC in statistical data analysis have been reported (see, e.g., Bozdogan (1994), Kitagawa and Gersch (1996)).

We now rewrite the penalized log-likelihood function (2.4) as

$$\begin{aligned} \log\left\{\prod_{\alpha=1}^{n} f(y_\alpha \mid x_\alpha; \gamma, \phi)(2\pi)^{-r/2}\left(\prod_{i=1}^{r} d_i\right)^{1/2} \exp\left(-\frac{n\lambda}{2}\gamma^T D_k^T D_k \gamma\right)\right\} \\ := \log\left\{\prod_{\alpha=1}^{n} f(y_\alpha \mid x_\alpha; \gamma, \phi)\pi(\gamma \mid \lambda)\right\}, \end{aligned} \tag{3.13}$$

where $r = m - k$ is the rank of the $m \times m$ matrix $D_k^T D_k$ and $d_1, \ldots, d_r$ are the nonzero eigenvalues of $n\lambda D_k^T D_k$. Hence the maximum penalized likelihood method is related to a Bayes model with improper prior distribution $\pi(\gamma \mid \lambda)$. For fixed values of λ and ϕ, the estimation problem of γ by maximizing the penalized log-likelihood function (2.4) is equivalent to obtain the mode of the posterior distribution (3.10) (Wahba (1978), Silverman (1985), Ishiguro and Arahata (1982), Tanabe and Tanaka (1983)).

Consider the B-spline nonparametric regression model with Gaussian noise given by (2.6). Then it follows from (3.12) and (3.13) that ABIC can be expressed as

$$\begin{aligned} \mathrm{ABIC}_N &= (n-k)\log(2\pi) + (n-k)\log\sigma^2 - (m-k)\log(n\beta) - \log\psi \\ &\quad + \log|B^T B + n\beta D_k^T D_k| + (\|\boldsymbol{y} - B\hat{\gamma}\|^2 + n\beta\hat{\gamma}^T D_k^T D_k \hat{\gamma})/\sigma^2, \end{aligned}$$

where $\hat{\gamma} = (B^T B + n\beta D_k^T D_k)^{-1} B^T \boldsymbol{y}$, $\beta = \sigma^2\lambda$ and ψ is a product of the nonzero eigenvalues of $D_k^T D_k$. For a given value of β, the value of σ^2 is chosen such that ABIC is minimal, and is given by $\hat{\sigma}_\beta^2 = (\|\boldsymbol{y} - B\hat{\gamma}\|^2 + n\beta\hat{\gamma}^T D_k^T D_k \hat{\gamma})/(n-k)$. The optimal value of β is obtained as the minimizer of $\mathrm{ABIC}_N(\beta, \hat{\sigma}_\beta^2)$.

(2) Modified AIC (Eilers and Marx (1996))

Under the assumptions that the model is estimated by maximum likelihood, and the true model belongs to the set of candidate models, Akaike's (1973) information criterion (AIC) is given by

-2(log-likelihood of the estimated model) $+\ 2$(the number of estimated parameters).

Eilers and Marx (1996) proposed to use AIC for the problem of choosing the optimal amount of smoothing, and gave criteria for Gaussian, Poisson and binomial models. For a Gaussian model, Eilers and Marx (1996) gave

$$\mathrm{AIC}_m = -2\sum_{\alpha=1}^{n} \log f_N(y_\alpha \mid x_\alpha; \hat{\gamma}, \hat{\sigma}_0^2) + 2\,\mathrm{tr}\, S,$$

where S is the hat matrix $B(B^TB + n\beta D_k^T D_k)^{-1}B^T$ and $\hat{\sigma}_0^2$ is the estimated error variance $\hat{\sigma}_0^2 = \|\boldsymbol{y} - B(B^TB)^{-1}B^T\boldsymbol{y}\|^2/n$. The variance was estimated by using the fitted value $\hat{\boldsymbol{y}}$ calculated at $\lambda = 0$.

When the number of basis functions is large compared with sample size, the inverse of the $m \times m$ matrix B^TB tends to be unstable and is often not computable. In our Monte Carlo simulation, we estimated σ^2 by $\hat{\sigma}^2 = \|\boldsymbol{y} - B\hat{\gamma}\|^2/n$, where $\hat{\gamma} = (B^TB + n\beta D_k^T D_k)^{-1}B^T\boldsymbol{y}$, and used instead the criterion

$$\mathrm{AIC}_m^* = -2\sum_{\alpha=1}^{n} \log f_N(y_\alpha \mid x_\alpha; \hat{\gamma}, \hat{\sigma}^2) + 2(\mathrm{tr}\, S + 1).$$

A problem may arise in theoretical justification for the use of the bias-correction term in AIC naturally, since AIC covers only models estimated by the maximum likelihood.

(3) Improved AIC (Hurvich *et al.* (1998))

In parametric linear regression and autoregressive time series models, Hurvich and Tsai (1989) proposed an improved version of AIC given by

$$-2(\text{log-likelihood of the estimated model}) + \frac{2n(p+1)}{n-p-2},$$

where p is the number of regression parameters in the model (see also Sugiura (1978) for a Gaussian linear regression model). Hurvich *et al.* (1998) replaced the number of parameters by the trace of the hat matrix S and introduced the criterion

$$\mathrm{AIC}_C = -2\sum_{\alpha=1}^{n} \log f_N(y_\alpha \mid x_\alpha; \hat{\gamma}, \hat{\sigma}^2) + \frac{2n(\mathrm{tr}\, S + 1)}{n - \mathrm{tr}\, S - 2},$$

being easy to apply in practical situations.

(4) Cross-validation

In cross-validation, the predictor for each observation is constructed based on the remaining data. Let $\hat{w}^{(-\alpha)}$ be a regression curve estimated by the observed data except (x_α, y_α). The cross-validation criterion is then

$$\mathrm{CV} = \frac{1}{n}\sum_{\alpha=1}^{n}(y_\alpha - \hat{w}^{(-\alpha)}(x_\alpha))^2 = \frac{1}{n}\sum_{\alpha=1}^{n}\left(\frac{y_\alpha - \hat{w}(x_\alpha)}{1 - s_{\alpha\alpha}}\right)^2, \tag{3.14}$$

where $s_{\alpha\alpha}$ is an α-th diagonal element of the hat matrix S and $\hat{w}^{(-\alpha)}(x_\alpha)$ is a predictive value of $\mathrm{E}[Y_\alpha \mid x_\alpha] = \mu_\alpha$.

Generalized cross-validation introduced by Craven and Wahba (1979) replaces $s_{\alpha\alpha}$ in (3.14) by the average $\sum_{\alpha=1}^{n} s_{\alpha\alpha}/n = \mathrm{tr}\, S/n$ and is

$$\mathrm{GCV} = \frac{1}{n}\sum_{\alpha=1}^{n}\left(\frac{y_\alpha - \hat{w}(x_\alpha)}{1 - \mathrm{tr}\, S/n}\right)^2.$$

4. Numerical results

4.1 *Analysis of real data*

- *The motorcycle impact data*

We illustrate the proposed procedure to choose the smoothing parameter and the number of basis functions through the analysis of the motorcycle impact data (Silverman (1985), Härdle (1990), Eilers and Marx (1996)). The motorcycle impact data were simulated to investigate the efficacy of crash helmets and comprise a series of measurements of head acceleration in units of gravity and times in milliseconds after impact.

We fit the B-spline nonparametric regression model with Gaussian noise (2.6) to the motorcycle impact data. The maximum penalized likelihood estimates $\hat{\gamma}$ and $\hat{\sigma}^2$ are given by equation (2.7). Then we choose the number of basis functions m and the smoothing parameter β that minimize the information criterion SPIC_N given by equation (3.8). For the analysis of the motorcycle impact data, we set the candidate values of m and β to $\{10,\ldots,30\}$ and $\{10^{10(i-100)/99}; i=1,\ldots,100\}$, respectively and optimal values of β and m could be chosen such that the criterion $\mathrm{SPIC}_N(\beta, m)$ is minimized. The roughness penalty in the penalized likelihood function (2.4) is taken as the second-order penalty defined by $\gamma^T D_2^T D_2 \gamma$. We choose the optimal values $\hat{m} = 16$ and $\hat{\beta} = 3.59 \times 10^{-4}$, and then $\mathrm{SPIC}_N = 1214.28$. The corresponding fitted curve is shown in Fig. 2 (a) (solid curve).

We implement our procedure against various types of criteria which introduced in Section 3.2. Table 1 gives the values of the number of basis functions and the smoothing parameter chosen by each criterion. We observe that, except for ABIC_N, the criteria yield similar values for $\hat{\beta}$ and $\hat{m}$, and are not directly comparable. The agreement in the variance estimates $\hat{\sigma}^2$ is close for all of the criteria.

We selected the optimal number of basis functions by SPIC_N. But we could not visually find difference among the fitted curves corresponding with $m = 15,\ldots,30$. One possible interpretation is that regarding the modest number of basis functions, the smoothing parameter can adjust the smoothness of B-spline curve fitting. Further research is needed for the effect of the number of basis function upon the B-spline smoothed estimate, making inference about its stability and reliability.

- *Kyphosis in laminectomy patients*

As our second example, we analyze the kyphosis data (Hastie and Tibshirani (1990)) by using B-spline nonparametric logistic regression model illustrated in Example 2. The data were collected from 83 patients undergoing corrective spinal surgery. The response y_α represents kyphosis after the operation and coded as either 0 (absence) or 1 (presence). We examined the relation between kyphosis and age in months at time of surgery.

The parameter vector γ is estimated by maximizing the penalized log-likelihood

Table 1. B-spline smoothed estimate for the motorcycle impact data.

	$SPIC_N$	CV	GCV	$ABIC_N$	AIC_m^*	AIC_C
$\hat{m}$	16	16	16	30	15	16
$\hat{\beta} \times 10^4$	3.59	3.68	5.86	59.9	3.68	5.86
$\hat{\sigma}^{2\dagger}$	464.0	464.2	468.0	461.9	470.8	468.0

$^\dagger \hat{\sigma}^2 = \sum_{\alpha=1}^n (y_\alpha - \hat{y}_\alpha)^2/n$.

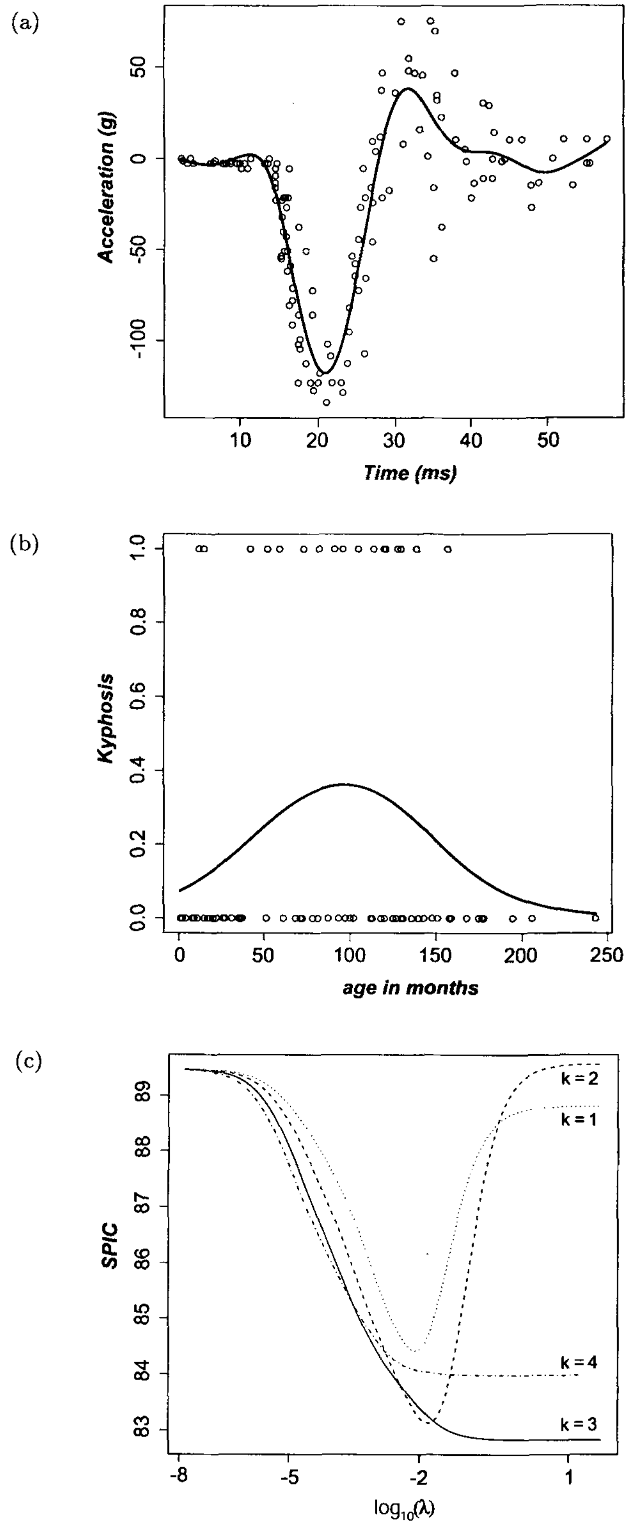
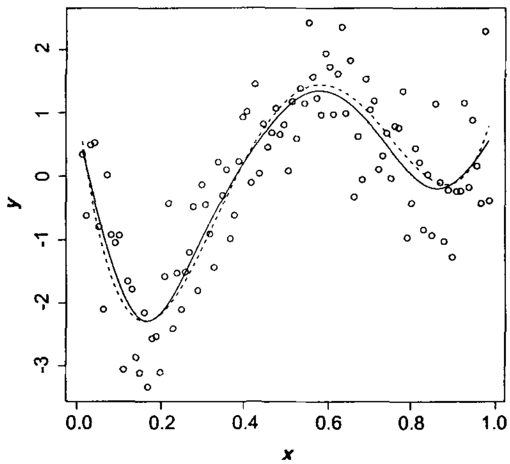
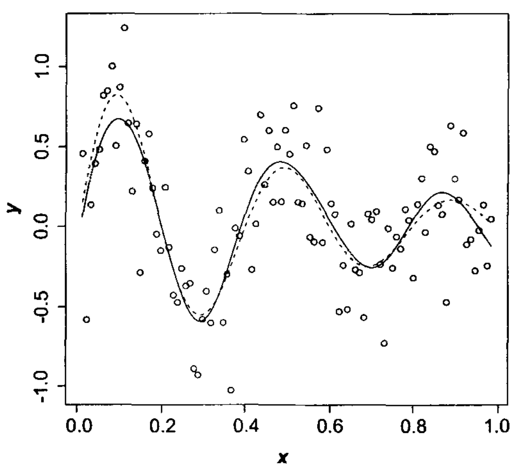


Fig. 2. Real data examples: (a) The motorcycle impact data. (b) and (c) The kyphosis data.

(a)



(b)



(c)

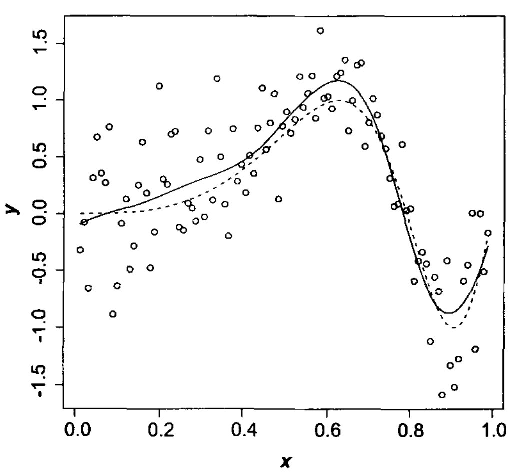


Fig. 3. Examples of simulated data: The dashed curve is the true regression curve, while the solid curve is B-spline smoothed estimate based on SPIC_N. $w(x)$ = (a) $1 - 48x + 218x^2 - 315x^3 + 145x^4$, (b) $\exp(-2x)\sin(5\pi x)$, (c) $\sin(2\pi x^3)$.

function with second order penalty $\gamma^T D_2^T D_2 \gamma$. The optimal values of λ and m are selected by SPIC$_L$ (3.10). Figure 2 (b) shows the fit for $\hat{\lambda} = 0.0132$ and $\hat{m} = 10$, and then SPIC$_L$ = 83.125. We can infer from the results on Fig. 2 (b) that the operation risk has a peak around 100 months after birth.

In a further research, we use the first, third and fourth order penalties and investigate the behaviors of SPIC$_L$. Figure 2 (c) represents the behaviors of SPIC$_L$ with the differences order $k = 1$, 2, 3 and 4. We can find the optimal value of λ which minimizes SPIC$_L$ in the first and second order penalty. However, in the third and fourth order penalty, SPIC$_L$ is a monotonous decreasing function and we cannot find the optimal value of λ. Within our research, when we use the third order penalty and a very large value of λ, SPIC$_L$ achieves the minimum (82.835). This implies that effectively the fitted curve for η is a second order polynomial (see Tanabe and Tanaka (1983), Eilers and Marx (1996)).

4.2 *Numerical comparisons*

In a Monte Carlo simulation repeated random samples $\{(x_\alpha, y_\alpha); \alpha = 1, \ldots, n\}$ were generated from the true regression model $y_\alpha = w(x_\alpha) + \varepsilon_\alpha$ for $x_\alpha = (2\alpha - 1)/(2n)$. The errors ε_α are assumed to be independently distributed according to a mixture of two normal distributions $\varepsilon_\alpha \sim \varepsilon N(0, \sigma^2) + (1 - \varepsilon) N(0, 3\sigma^2)$, where the standard deviation is taken as $\sigma = 0.05R_y$ or $0.1R_y$ with R_y being the range of $w(x)$ over $x \in [0, 1]$. The true curve $w(x)$ is assumed to be the following regression functions (see, e.g., Hurvich *et al.* (1998)).

$$w(x) = \begin{cases} 1 - 48x + 218x^2 - 315x^3 + 145x^4, \\ \sin(2\pi x^3), \\ \exp(-2x)\sin(5\pi x). \end{cases}$$

We fit B-spline nonparametric regression model with Gaussian noise defined by (2.6) to the simulated data. The model is estimated by the maximization of the penalized likelihood function (2.4) with the second-order penalty and 10 basis functions, since Monte Carlo simulations require a considerable amount of computation. Figure 3 shows examples of simulated data with B-spline smoothed estimates based on SPIC$_N$. In order to examine the properties of various types of criteria, we use the average squared error (ASE) and predictive average squared error (PASE) defined by ASE $= \sum_{\alpha=1}^{n} \{w(x_\alpha) - \hat{y}_\alpha\}^2/n$ and PASE $= \sum_{\alpha=1}^{n} (y_\alpha^* - \hat{y}_\alpha)^2/n$, where $y_1^*, \ldots, y_n^*$ are future observations generated from the true model. The simulation results were obtained by averaging over 300 repeated Monte Carlo trials. Table 2 summarizes the simulation results for each true regression curve, in which the notation MEAN and SD refer to the average value of $\hat{\beta}$ chosen by each criteria and its standard deviation, respectively.

Simulation results may be summarized as follows: Our proposed information criterion, SPIC$_N$, generally gives good estimates in the sense of ASE and PASE, and yields stable smoothing parameter estimates. The performance of ABIC$_N$ depends on complexity of the regression function and the error variance. For the regression function (b), ABIC$_N$ is the best selector and SPIC$_N$ is an alternative. But for the regression functions (a) and (c), ABIC$_N$ chooses unstable smoothing parameter estimate and gives larger ASE and PASE, whereas SPIC$_N$ gives good performance. The smoothing parameters chosen by CV have high variability, and lead to larger ASE and PASE compared with SPIC$_N$ in most situations.

Table 2. Monte Carlo results ($n = 100$).

	$SPIC_N$	CV	GCV	$ABIC_N$	AIC_m^*	AIC_C
$w(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$, $\sigma/R_y = 0.05$						
$\varepsilon = 1.0$						
MEAN$\times 10^6$	9.102	13.66	15.78	30.70	14.01	18.80
SD$\times 10^5$	1.263	1.856	1.847	1.107	1.689	2.064
ASE$\times 10^3$	3.523	3.577	3.581	3.721	3.559	3.614
PASE$\times 10^2$	3.819	3.826	3.824	3.838	3.821	3.826
$\varepsilon = 0.9$						
MEAN$\times 10^5$	1.095	1.558	1.964	3.807	1.746	2.439
SD$\times 10^5$	1.268	1.838	2.175	1.415	1.929	2.730
ASE$\times 10^3$	4.085	4.136	4.138	4.297	4.117	4.184
PASE$\times 10^2$	4.498	4.503	4.503	4.523	4.500	4.508
$w(x) = \exp(-2x)\sin(5\pi x)$, $\sigma/R_y = 0.1$						
$\varepsilon = 1.0$						
MEAN$\times 10^5$	1.681	2.335	2.272	2.573	2.087	2.664
SD$\times 10^5$	1.399	1.852	1.629	0.718	1.558	1.830
ASE$\times 10^3$	2.363	2.387	2.383	2.339	2.377	2.403
PASE$\times 10^2$	2.120	2.122	2.122	2.117	2.121	2.123
$\varepsilon = 0.9$						
MEAN$\times 10^5$	1.794	2.601	2.424	3.014	2.182	2.852
SD$\times 10^5$	1.546	2.217	1.641	0.869	1.540	1.899
ASE$\times 10^3$	2.535	2.568	2.536	2.492	2.533	2.554
PASE$\times 10^2$	2.530	2.533	2.529	2.524	2.530	2.531
$w(x) = \sin(2\pi x^3)$, $\sigma/R_y = 0.2$						
$\varepsilon = 1.0$						
MEAN$\times 10^4$	1.309	1.682	1.961	5.774	1.772	2.363
SD$\times 10^4$	1.641	1.842	1.949	3.390	1.854	2.145
ASE$\times 10^2$	1.491	1.497	1.505	1.605	1.502	1.513
PASE$\times 10$	1.753	1.753	1.759	1.762	1.759	1.754
$\varepsilon = 0.9$						
MEAN$\times 10^4$	1.968	2.591	2.910	7.992	2.495	3.497
SD$\times 10^4$	2.771	3.225	3.141	5.652	2.729	3.461
ASE$\times 10^2$	1.872	1.899	1.876	2.005	1.878	1.891
PASE$\times 10$	2.133	2.139	2.133	2.150	2.133	2.145

MEAN, ASE and PASE are averages.

We observed that the smoothness of an estimated curve is mainly controlled by the smoothing parameter. Hence, in practice, we may employ a modest number of basis functions and then determine the smoothing parameter as the minimizer of the criterion.

ABIC$_N$ and AIC$_C$ work well when the error variances are relatively large (i.e. large smoothing parameter is appropriate) and have a tendency toward oversmoothing. GCV is better than CV in many situations. SPIC$_N$, GCV and AIC$_m^*$ work well in the cases where the error variances are relatively small. But SPIC_N is better than GCV and AIC_m^* in most cases. When the error variances are relatively large, SPIC_N still gives good performance. For large sample size of $n = 200$, all of the criteria stably choose the smoothing parameter and yield sufficiently small ASE and PASE.

Similar comparisons were made for other combinations of sample sizes, the number of basis functions and mixing proportions. We found the results described above to be essentially unchanged. We conclude from the results of Monte Carlo simulations that SPIC_N generally works well in practical situations.

We agree that AIC_m^* and AIC_C are easy to apply in practice. Despite the simplicity of these criteria, the problem still remains in theoretical aspect. We derived SPIC as an estimator of the Kullback-Leibler information under model misspecification and it has a sounder theoretical basis than AIC_m^* and AIC_C.

5. Discussion

In this article we derived information criteria for evaluating B-spline nonparametric regression models estimated by the maximum penalized likelihood method under model misspecification. We observed through Monte Carlo experiments and real data examples that the proposed criteria generally perform well for B-spline smoothing. The criteria were given as estimators of the Kullback-Leibler measure of discriminatory information between two probability distributions. An advantage of the information-theoretic approach is that it is not restricted to linear estimators of regression functions, but may be applied to construct a criterion for evaluating other nonparametric models like neural networks.

The bootstrap methods introduced by Efron (1979) offer an alternative approach to statistical model evaluation problems (Konishi and Kitagawa (1996), Ishiguro *et al.* (1997)). By bootstrapping the bias of a log-likelihood of estimated nonparametric model, we may construct a model evaluation criterion. However the bias estimate obtained numerically includes both the randomness of the observed data and simulation error which decreases as the number of bootstrap replication increases. Also Monte Carlo algorithm requires considerable amount of computations. Further work remains to be done in constructing a bootstrapping criterion.

Acknowledgements

The authors would like to thank three referees for their helpful comments and suggestions.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory* (eds. B. N. Petrov and F. Csaki), 267–281, Akademiai Kiado, Budapest (Reproduced in *Breakthroughs in Statistics*, Volume 1 (eds. S. Kotz and N. L. Johnson), Springer Verlag, New York (1992)).

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

Akaike, H. (1980*a*). On the use of predictive likelihood of a Gaussian model, *Ann. Inst. Statist. Math.*, **32**, 311–324.

Akaike, H. (1980*b*). Likelihood and the Bayes procedure, *Bayesian Statistic* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), University Press, Valencia, Spain.

Bozdogan, H. (ed.) (1994). *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informatial Approach*, Kluwer, Dordrecht.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Numer. Math.*, **31**, 377–403.

De Boor, C. (1978). *A Practical Guide to Splines*, Springer, Berlin.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*, Oxford University Press, Oxford.

Efron, B. (1979). Bootstrap methods: Another look at the jacknife, *Ann. Statist.*, **7**, 1–26.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with *B*-splines and penalties (with discussion), *Statist. Sci.*, **11**, 89–121.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

Good, I. J. (1965). *The Estimation of Probabilities*, M. I. T. Press, Cambridge, Massachusetts.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities, *Biometrika*, **58**, 255–277.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models, *International Statistical Review*, **55**, 245–259.

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

Hampel, F. R., Rousseeuw, P. J., Ronchetti, E. M. and Stahel, W. A. (1986). *Robust Statistics. The Approach on Influence*, Wiley, New York.

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

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

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples, *Biometrika*, **76**, 297–307.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *J. Roy. Statist. Soc. Ser. B*, **60**, 271–293.

Ishiguro, M. and Arahata, E. (1982). A Bayesian spline regression (in Japanese), *Proc. Inst. Statist. Math.*, **30**, 30–36.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC, *Ann. Inst. Statist. Math.*, **49**, 411–434.

Kitagawa, G. and Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*, Springer, New York.

Konishi, S. (1999). Statistical model evaluation and information criteria, *Multivariate Analysis, Design of Experiments and Surrey Sampling* (ed. S. Ghosh), 369–399, Marcel Dekker, New York.

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

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Ann. Math. Statist.*, **22**, 79–86.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman and Hall, London.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *J. Roy. Statist. Soc. Ser. A*, **135**, 370–384.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion), *J. Roy. Statist. Soc. Ser. B*, **47**, 1–52.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer, New York.

Stone, C. J. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion), *J. Roy. Statist. Soc. Ser. B*, **36**, 111–147.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections, *Comm. Statist. Theory Methods*, **A7**, 13–26.

Tanabe, K. and Tanaka, T. (1983). Fitting curves and surfaces to data by using Bayes model (in Japanese), *Chikyu*, **5**, 179–186.

Wahba, G. (1978). Improper priors, spline smoothing and the probrem of guarding against model errors in regression, *J. Roy. Statist. Soc. Ser. B*, **40**, 364–372.