Semiparametric regression model selections

Article *in* Journal of Statistical Planning and Inference · February 1999

DOI: 10.1016/S0378-3758(98)00175-X

2 authors:

Peide Shi
Peking University
51 PUBLICATIONS 988 CITATIONS
SEE PROFILE

Chih-Ling Tsai
University of California, Davis
94 PUBLICATIONS 11,715 CITATIONS
SEE PROFILE

Journal of Statistical Planning and
Inference 77 (1999) 119–139

journal of
statistical planning
and inference

Semiparametric regression model selections

Peide Shi[a], Chih-Ling Tsai[b,*]

[a] Department of Probability and Statistics, Peking University,
Beijing 100871, People's Republic of China

[b] Graduate School of Management, University of California, Davis, CA 95616-8909, USA

Received 12 January 1997; accepted 5 August 1998

Abstract

In semiparametric regression models, we have developed a small-sample criterion, AICC, for the selection of explanatory variables in the parametric component as well as for choosing the number of spline knots to estimate the nonparametric component. In contrast to the Akaike Information Criterion (AIC), AICC provides a nearly unbiased estimator of the expectation of the Kullback–Leibler information. Monte Carlo results show that AICC outperforms AIC, *Cp* (Mallows, 1973), FPE (Akaike, 1970), and SIC (Schwartz, 1978) for small samples. In addition, we show that AICC, AIC, *Cp*, FPE, and GCV provide asymptotically efficient selections. The asymptotic optimalities of GIC (Nishii, 1984) and SIC are also obtained. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: AICC; Asymptotic efficiency; B-spline function; Kullback–Leibler information

1. Introduction

Regression analyses have always played an important role in studying the relationship between a response variable and explanatory variables. Usually, several competing (or candidate) parametric models are available; hence, selecting an appropriate model from all possible candidates is an important task. One method often used is the Akaike (1973) Information Criterion (AIC) which was designed to be an approximately unbiased estimator of the expected Kullback–Leibler information for the fitted model. Although this criterion has been widely used, it can have serious deficiencies. The AIC can be drastically biased when used for parametric regression models, and a corrected version of this statistic, the AICC, was presented by Hurvich and Tsai (1989). The AICC is more nearly unbiased and tends to provide better model choices than AIC for small samples.

[*] Corresponding author.

PII: S0378-3758(98)00175-X

For parametric regression models, Shibata (1981) has shown that AIC and other asymptotically equivalent methods provide an asymptotically efficient selection. Specifically, the AIC-selected estimator of the regression function is asymptotically as good, in terms of mean integrated squared error, as the estimator which uses the best approximating model in the class of candidates. Recently, Hurvich and Tsai (1995) provided the relative rate of convergence for the mean integrated squared error of the AIC-selected estimator. It can be shown in a straightforward manner that the same relative rate is also attained for the AICC-selected estimator. In contrast to efficient selection criteria, Schwarz (1978) obtained a consistent selection criterion, SIC. In other words, it selects the correct model with probability approaching 1 in large samples when the true model is finite-dimensional. Detailed discussions of efficient and consistent criteria can be found in Shao (1997).

For nonparametric regression models (e.g., nearest-neighbor nonparametric regression (Li, 1987), regression splines (Friedman and Silverman, 1989), and generalized Fourier series regression (Eubank, 1988), there are infinitely many parameters. In this case, there is no finite order (or number of parameters) to be estimated consistently. In addition, Shibata (1981) showed that consistent selectors do not produce efficient estimators. Hence, the focus in nonparametric regression model is to obtain an efficient estimation of the mean function (see Burman and Chen, 1989). Recently, Hurvich et al. (1998) obtained an efficient selector AICC for nonparametric regression models, and they showed that AICC outperforms the generalized cross validation, GCV, proposed by Craven and Wahba (1979) and AIC.

In many applications the parametric model itself is at best an approximation of the true one, and the search for an adequate model from the parametric family is not easy. When there are no persuasive parametric models available, the family of semiparametric regression models (or partial linear models) can be considered as a promising extension of the parametric family. Useful reference books on this topic are Eubank (1988), Hastie and Tibshirani (1990), Wahba (1990), Härdle (1990), and Green and Silverman (1994). Basically, semiparametric models contain both parametric and nonparametric components that provide a convenient way to include nonlinearities of an unspecified form in a regression model. Chen and Shiau (1994) have studied the asymptotic behavior of two efficient estimators of the parametric component of a partial linear model when the smoothing parameter is chosen by GCV, or by the Mallows *Cp* criterion (Mallows, 1973). In addition, Green and Silverman (1994) suggested using Allen's (1974) cross validation CV and GCV to choose an appropriate value for the smoothing parameter. However, both Chen and Shiau and Green and Silverman focused only on selecting the smoothing parameter from the nonparametric component. They did not consider selection of both the explanatory variables and smoothing parameters from the parametric and nonparametric model components, respectively.

In this paper we describe the development of the AICC for semiparametric regression model selection. We first focus on selecting variables and the smoothing parameter from candidate models which incorporate both parametric regression and B-splines (Section 2). There are finite number of parameters in the parametric component and many parameters (spline knots) in the nonparametric component. Our task is to obtain AICC to select a set of explanatory variables from the parametric component and the number of spline knots from the nonparametric component. We next demonstrate that, under reasonable conditions, the AIC, AICC, *Cp*, FPE, and GCV provide asymptotically efficient selections (Section 3). Hence, these selection criteria produce an efficient estimator of the semiparametric function. In addition, we obtain the asymptotic optimality of the selected estimators from the generalized Information Criterion (GIC) and SIC criteria. Monte Carlo studies demonstrate that AICC outperforms AIC, *Cp*, FPE (Akaike, 1970, Eq. (4.7)) GCV, and SIC for small samples. As the sample size increases, SIC performs better than AICC in the selection of the parametric component. We present these simulation results in Section 4. Finally, we make our concluding remarks in Section 5.

2. Model selection for semiparametric models

2.1. The true model

Here we assume that the data are generated from the (true) model,

$$Y=\mu+\varepsilon=X_0\beta_0+h_0+\varepsilon, \tag{2.1}$$

where $Y=(Y_1,\dots,Y_n)'$, $\mu=(\mu_1,\dots,\mu_n)'$, $X_0=(x_{01},\dots,x_{0n})'$, x_{0i} is a $p_0\times 1$ known vector for $i=1,\dots,n$, $h_0=(h_0(T_1),\dots,h_0(T_n))'$, $\varepsilon=(\varepsilon_1,\dots,\varepsilon_n)'$, β_0 is a $p_0\times 1$ unknown vector, $h_0(\cdot)$ is an unknown function defined on $[0,1]$ with $\int_0^1 h_0(t)\,\mathrm{d}t=0$, and $\varepsilon\sim \mathrm{N}(0,\sigma_0^2 I_{n\times n})$.

2.2. The B-spline approximation to the nonparametric component

Let $M=m+k_n$, where m is a given nonnegative integer and k_n is a positive integer, and let $t_1,\dots,t_{k_n}$ $(t_1<\cdots<t_{k_n}=1)$ be a D_0-quasi uniform sequence partitioned over $[0,1]$ (see Schumaker, 1981, p. 216). Further, let $0=s_1=\cdots=s_{m+1}$, $s_{m+2}=t_1,\dots,s_{m+k_n}=t_{k_n-1},s_{m+k_n+1}=\cdots=s_{2m+k_n+1}=1$, and let $B(t)=(B_1(t),\dots,B_M(t))'$ be a vector of normalized B-splines of order $m+1$ associated with the extended partition $\{s_1,\dots,s_M\}$ of $[0,1]$ (see Schumaker, 1981, p. 224). Then, from Corollary 6.21 in Schumaker (1981), we can approximate $h_0(t)$ by the B-spline function $\pi(t)'\alpha$, where $\pi(t)=B(t)-\int_0^1 B(t)\,\mathrm{d}t$, and α is an $M\times 1$ vector of unknown parameters.

In practical calculations, the selection criterion is usually used to choose spline knots for estimating the nonparametric component when the parametric component is given. For example, AIC can be used to choose spline knots from the order statistics $T_{(1)},\dots,T_{(n)}$ of the sample points $T_1,\dots,T_n$ based on the forward placement and backward deletion algorithms in He and Shi (1996) except that the spline knots are optimized at the deletion step (see Shi, 1995; Friedman and Silverman, 1989). The resulting set of the collection of candidate knots is denoted by Λ_2. As $\beta_0=0$ and

h_0 is a bivariate function, He and Shi (1996) compared the performance of their algorithm with that MARS (Friedman, 1991) and PIMPLE (Breiman, 1991) through Monte Carlo studies. They show that their algorithm is either comparable or superior to both of MARS and PIMPLE.

2.3. *The approximating family of semiparametric models*

Consider the approximating (or candidate) family of models

$$Y = X(\lambda)\beta(\lambda) + \Pi(\lambda)\alpha(\lambda) + e, \tag{2.2}$$

where $\beta(\lambda) \in R^p$, $\alpha(\lambda) \in R^N$, $e \sim \mathrm{N}(0, \sigma^2 I_{n\times n})$, $\lambda = \{\lambda_1, \lambda_2\}$, $\lambda_1 = (1, \ldots, p) \in \Lambda_1$ and $\lambda_2 \in \Lambda_2, \Lambda_1$ is a collection of all candidate models of the parametric part, $X(\lambda)$ and $\Pi(\lambda) = (\pi(T_1), \ldots, \pi(T_n))'$ are $n \times p$ and $n \times N$ matrices determined by indices λ_1 and λ_2, respectively, $\mathrm{N} = \mathrm{N}(\lambda) = m + l$, and l is the cardinal number of λ_2. Thus, the true model in Eq. (2.1) can be reexpressed in the form

$$Y = X_0\beta_0 + \Pi(\lambda)\alpha(\lambda) + R_n + \varepsilon \tag{2.3}$$

for the given λ, where R_n is an $n \times 1$ vector such that $n^{-1}R_n'R_n \leqslant C_1 N^{-2C_2}$, C_1 is a constant, and $C_2 \geqslant m$ is an indicator of the smoothness of h_0. A detailed discussion of R_n can be found in Shi and Li (1995).

2.4. *Kullback–Leibler information and AICC*

Let $G(\cdot)$ and $F(\cdot|\theta(\lambda))$ be the distributions of the true model and the approximating model, respectively, where $\theta(\lambda) = (\beta(\lambda)', \alpha(\lambda)', \sigma^2(\lambda))'$. A useful measure of the discrepancy between the true and approximating models is the Kullback–Leibler information criterion,

$$\begin{aligned} \Delta\{\theta(\lambda)\} &= 2E_G[\log\{g(Y)/f(Y|\theta(\lambda)\}] \\ &= -n + n\log(\sigma^2(\lambda)/\sigma_0^2) + E_0\|Y - X(\lambda)\beta(\lambda) - \Pi(\lambda)\alpha(\lambda)\|^2/\sigma^2(\lambda), \end{aligned}$$

where g and f are the density functions of G and F, respectively, and E_G denotes the expectation under the true model. A reasonable standard for judging the quality of the approximating family in the light of the data is $E_G[\Delta\{\hat{\theta}(\lambda)\}]$, where $\hat{\theta}(\lambda) = (\hat{\beta}(\lambda)', \hat{\alpha}(\lambda)', \hat{\sigma}^2(\lambda))'$, and $\hat{\beta}(\lambda)$, $\hat{\alpha}(\lambda)$, $\hat{\sigma}^2(\lambda)$ are the maximum likelihood estimates of β, α, and σ^2 in the approximating family. The detailed expression for $\hat{\theta}(\lambda)$ is given in Appendix A. Let $\theta_0(\lambda) = (\beta_0(\lambda)', \alpha_0(\lambda)', \sigma_0^2(\lambda))'$ be the estimator that minimizes $\Delta\{\theta(\lambda)\}$ for a given λ. Then, for a given λ, we have (ignoring the constant, $-n$)

$$\begin{aligned} \Delta\{\theta(\lambda)\} = {} & n[\log\{\sigma^2(\lambda)/\sigma_0^2\}] + n\sigma_0^2(\lambda)/\sigma^2(\lambda) \\ & + \|X(\lambda)(\beta(\lambda) - \beta_0(\lambda)) + \Pi(\lambda)(\alpha(\lambda) - \alpha_0(\lambda))\|^2/\sigma^2(\lambda). \end{aligned}$$

In addition, replacing Y by μ in the expression of $\hat{\beta}(\lambda)$ and $\hat{\alpha}(\lambda)$ gives $\beta_0(\lambda)$ and $\alpha_0(\lambda)$, respectively, and $\sigma_0^2(\lambda) = \frac{1}{n}\mu'(I - H)\mu + \sigma_0^2$.

We now assume that the parametric component of the approximating model includes the parametric component of the true model. Under this assumption, the mean function μ given in Eq. (2.3) can be expressed as

$$\mu = X(\lambda)\tilde{\beta}(\lambda) + \Pi(\lambda)\alpha(\lambda) + R_n,$$

where $X(\lambda) = (X_0|X_1)$, X_1 is an $n \times (p - p_0)$ matrix, $\tilde{\beta}(\lambda) = (\beta_0'|0')'$ and 0 is a $(p - p_0) \times 1$ vector of zeros. This leads to

$$\Delta\{\hat{\theta}(\lambda)\} = n[\log\{\hat{\sigma}^2(\lambda)/\sigma_0^2\}] + n\sigma_0^2(\lambda)/\hat{\sigma}^2(\lambda) + (Y - \mu)'H(Y - \mu)/\hat{\sigma}^2(\lambda), \quad (2.4)$$

where $H = Q(\lambda)X(\lambda)(X(\lambda)'Q(\lambda)X(\lambda))^{-1}X(\lambda)'Q(\lambda) + \Pi(\lambda)(\Pi'(\lambda)\Pi(\lambda))^{-1}\Pi'(\lambda)$ and $Q(\lambda) = I - \Pi(\lambda)(\Pi'(\lambda)\Pi(\lambda))^{-1}\Pi'(\lambda)$.

Given a collection of competing approximating models for Y, the one that minimizes $E_G[\Delta\{\hat{\theta}(\lambda)\}]$ has the smallest discrepancy and is therefore preferred. Using the distributional results given in Appendix B, we have

$$\begin{aligned} E_G\left[\frac{\sigma_0^2}{\hat{\sigma}(\lambda)^2}\right] &= E_G\left[\frac{1 + n^{-1}R_n'R_n/\sigma_0{}^2}{n^{-1}\chi^2_{n-N-p,\delta}}\right] \\ &= \sum_{j=0}^{\infty} \mathrm{e}^{-\delta/2}(\delta/2)^j \frac{n(1 + \mathrm{O}(N^{-2C_2}))}{j!(n - N - p + 2j - 2)}, \end{aligned}$$

and

$$\begin{aligned} E_G&\left[\frac{\|X(\lambda)(\hat{\beta}(\lambda) - \beta_0(\lambda)) + \Pi(\lambda)[\hat{\alpha}(\lambda) - \alpha_0(\lambda)]\|^2}{n\hat{\sigma}(\lambda)^2}\right] \\ &= \sum_{j=0}^{\infty} \frac{(\delta/2)^j \ (N + p)\mathrm{e}^{-\delta/2}}{j!(n - N - p + 2j - 2)}. \end{aligned}$$

Note that if $h_0(\cdot)$ is a spline function or a polynomial of degree no greater than m,

$$E_G\left[\frac{\sigma_0^2}{\hat{\sigma}(\lambda)^2}\right] = \frac{n(1 + \mathrm{O}(N^{-2C_2}))}{n - N - p - 2}$$

and

$$E_G\left[\frac{\|X(\lambda)(\hat{\beta}(\lambda) - \beta_0(\lambda)) + \Pi(\lambda)[\hat{\alpha}(\lambda) - \alpha_0(\lambda)]\|^2}{n\hat{\sigma}(\lambda)^2}\right] = \frac{N + p}{n - N - p - 2},$$

and that

$$\begin{aligned} 0 &\leqslant \frac{1}{n - N - p - 2} - \sum_{j=0}^{\infty} \mathrm{e}^{-\delta/2}(\delta/2)^j \frac{1}{j!(n - N - p + 2j - 2)} \\ &\leqslant \frac{\mathrm{O}(N^{-2C_2})}{(n - N - p - 2)}, \end{aligned}$$

where C_2 is defined in Eq. (2.3). Thus, as n tends to infinity, $Nn^{-1/(2C_2+1)}$ is bounded away from zero and infinity, and $C_2 > \frac{1}{2}$,

$$\begin{aligned} E_G[\Delta\{\hat{\theta}(\lambda)\}] &= E_G(\mathrm{AICC}(\lambda)) - n\log(\sigma_0^2) + r_{1n} \\ &= E_G(\mathrm{AIC}(\lambda)) - n\log(\sigma_0^2) + r_{1n} + r_{2n}, \end{aligned} \tag{2.5}$$

where

$$\mathrm{AICC}(\lambda) = n[\log\{\hat{\sigma}^2(\lambda)\} + 1] + 2n(N + p + 1)/(n - N - p - 2), \tag{2.6}$$

$$\mathrm{AIC}(\lambda) = n[\log\{\hat{\sigma}^2(\lambda)\} + 1] + 2(N + p + 1), \tag{2.7}$$

$r_{1n} = \mathrm{O}(n^{1-2C_2/(2C_2+1)})$, and $r_{2n} = 2(N + p + 1)(N + p + 2)/(n - N - p - 2)$. Eq. (2.5) shows that $E_G\{\mathrm{AICC}(\lambda)\}$ is closer to $E_G[\Delta\{\hat{\theta}(\lambda)\}]$ than $E_G\{\mathrm{AIC}(\lambda)\}$.

Other well-known criteria such as *Cp* (Mallows, 1973), GCV, SIC (Schwarz, 1978) and FPE (Akaike, 1970) can be analogously defined as: $Cp(\lambda) = \hat{\sigma}^2(\lambda) + 2(N + p)\tilde{\sigma}^2/n$, $\mathrm{GCV}(\lambda) = \hat{\sigma}^2(\lambda)/(1-(N+p)/n)^2$, $\mathrm{SIC}(\lambda) = n\log(\hat{\sigma}^2(\lambda)) + (N+p)\log(n)$ and $\mathrm{FPE}(\lambda) = (n + N + p)\hat{\sigma}^2(\lambda)/(n - N - p)$, where $\tilde{\sigma}$ is an estimate of σ_0.

2.5. Selection procedure

Here we describe the procedure for applying the AICC criterion to select the best of the candidate models. Let a family of approximating models be indexed by $\Lambda = \{\lambda : \lambda = (\lambda_1, \lambda_2) : \lambda_1 \in \Lambda_1, \lambda_2 \in \Lambda_2\}$, where Λ_2 and Λ_1 are defined in Sections 2.2 and 2.3, respectively. In general, for given p-variables, $x_i = (x_{i1}, \ldots, x_{ip})'$, in the parametric part, we choose spline basis functions (spline knots accordingly), $\pi_1(t), \ldots, \pi_{N_0}(t)$, such that $\mathrm{AICC}(\lambda_1, \lambda_2) = n[\log(\hat{\sigma}^2(\lambda)) + 1] + 2n(N_0 + p + 1)/(n - N_0 - p - 2)$ is minimized over all candidate spline bases (spline knot sets), where $\hat{\sigma}^2$ is the MLE of $\sigma^2(\lambda)$ under the model $y_i = x_i'\beta + \sum_{k=1}^{N_0} \pi_k(T_i)\alpha_i + e_i$ for $i = 1, \ldots, n$. That is, we minimize $\mathrm{AICC}(\lambda_1, \lambda_2)$ over Λ_2 for any given λ_1, where the minimizer is denoted by $\hat{\lambda}_2$, and then choose λ_1 over Λ_1 to minimize $\mathrm{AICC}(\lambda_1, \hat{\lambda}_2)$. The resulting value is denoted $\hat{\lambda}_1$. Thus $\mathrm{AICC}(\hat{\lambda})$ selects the best model from all possible candidate models, where $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2)$.

The algorithm for generating spline knot sets for the given λ_1 can be obtained from authors by request. This algorithm is also applicable to model selection using spline knots chosen with the AIC, *Cp*, FPE, GCV, and SIC criteria.

3. The optimality of the model selection

For variable selection in parametric regression models, Shibata (1981) has shown that AIC and other equivalent methods provide an asymptotically efficient selection. In this section we extend Shibata's results to semiparametric models subject to some regularity conditions. Throughout this section, we assume that data are generated from Eq. (2.1) and $\varepsilon_1, \ldots, \varepsilon_n$ are iid random errors with mean zero and variance σ_0^2.

Let $\lambda_1^{(0)}$ denote index of the true model of the parametric part and let

$$U_n(\lambda)=\|X(\lambda)\hat{\beta}(\lambda)+\Pi(\lambda)\hat{\alpha}(\lambda)-X_0\beta_0-h_0\|^2/n$$

for $\lambda\in\Lambda$. In addition, let $\hat{\lambda}$ and λ^* be the model selected from Λ by minimizing AIC(λ) and $L_n(\lambda)=E_G\{U_n(\lambda)\}$, respectively. Then, by the following theorem, the ratio of the mean integrated squared error of the AIC-selected estimator to that of the optimal estimator tends to one in probability. Before we state the theorem we must make the following assumptions – the first to assure the uniqueness of $(\hat{\beta}(\lambda)',\hat{\alpha}(\lambda)')'$, and the last two to assure the consistency of B-spline estimators:

Assumption 1. $X(\lambda)'Q(\lambda)X(\lambda)$ is full rank for any model $\lambda\in\Lambda$.

Assumption 2. $k_n\to\infty$ and $k_n/n\to 0$ as $n\to\infty$, and there exists a positive-definite matrix Σ such that $\Sigma - X_0'X_0/n>0$ for all n.

Assumption 3. h_0 is m (>0) times differentiable.

Theorem 1. *If Assumptions* 1–3 *are satisfied, and* $\sum_{\lambda_2\in\Lambda_2}1/N^3(\lambda_2)<\infty$, $E_G(\varepsilon_1^{12})<\infty$ *and* $\inf_{\lambda\in\Lambda} nL_n(\lambda)\to\infty$, *then, regardless of whether there exists a correct model for the parametric part or not,*

$$L_n(\hat{\lambda})/L_n(\lambda^*)-1=o_p(1).$$

Proof. See Appendix C.

If we rewrite AIC as the equivalent form of $\hat{\sigma}^2\exp(2(N+p)/n)$, then Theorem 1 implies that AIC(λ)$=L_n(\lambda)+n^{-1}\sum_{i=1}^n\varepsilon_i^2+o_p(L_n(\lambda))$. This implies that AIC($\lambda$) is a reasonable estimator of $\sigma_0^2+E_G(L_n(\lambda))$ regardless of whether there exists a correct model in the parametric part or not.

In order to further investigate the asymptotic properties of the estimator selected by AIC criterion, we first present the following lemma.

Lemma. *If Assumptions* 1–3 *are satisfied, and* $E_G(\varepsilon_1^2)<\infty$ *and* $\lambda_1^{(0)}\in\Lambda_1$, *then*

$$L_n(\lambda^*)=o(1).$$

Proof. See Appendix D.

Then, applying both Theorem 1 and the above lemma, we obtain the following asymptotic result.

Corollary 1. *Under the same conditions as given in Theorem* 1 *and the above Lemma,*

$$L_n(\hat{\lambda})=o_p(1).$$

Corollary 1 implies that the estimator of the true model selected by AIC criterion is L_n-consistent.

The asymptotic optimality discussed in Theorem 1 is based on the mean integrated squared error measure. The following corollary shows that this asymptotic optimality is also valid in terms of the mean squared error, which has been used by Li (1987) to study the asymptotic optimality of *Cp* in the nonparametric regression settings.

Corollary 2. *Under the same conditions as stated in Theorem* 1, *we have*

$$U_n(\hat{\lambda})/\inf_{\lambda\in\Lambda} U_n(\lambda) - 1 = \mathrm{o_p}(1).$$

Proof. Applying the same techniques used in the proof of Theorem 1, we have

$$\sup_{\lambda} U_n(\lambda)/L_n(\lambda) = 1.$$

This result in connection with Theorem 1 leads to the conclusion that Corollary 2 holds.

It can be shown straightforwardly that Theorem 1 and Corollaries 1 and 2 remain valid if AIC is replaced by AICC, CP, GCV or FPE. Furthermore, we can show that those asymptotic optimalities hold as long as the penalty term, a_n of the selection criterion satisfies

$$\exp(a_n/n) - 1 - 2(N+p)/n = \mathrm{o}((N+p)/n). \tag{3.1}$$

Note that the penalty term of SIC is $a_n = (N+p)\ln(n)$ and it does not satisfy the above equation. Hence, SIC has slightly different asymptotic properties in comparison to a criterion whose penalty term satisfies Eq. (3.1). In the rest of this section, we will discuss the asymptotic optimalities of SIC and its generalized form, generalized information criterion (GIC).

We first extend Nishii (1984) GIC criterion from parametric model settings to semiparametric model settings:

$$\mathrm{GIC}(\lambda) = n\log(\hat{\sigma}^2(\lambda)) + (N+p)\gamma_n,$$

where $\{\gamma_n\}$ is some given positive sequence. If $\gamma_n = \log(n)$, then GIC is SIC. In contrast, if $\gamma = 2$ then GIC is AIC. Furthermore, applying the fact that $\mathrm{GIC}(\lambda) = \hat{\sigma}^2(\lambda)$ $\exp\{(N+p)\gamma_n/n\}$, we can show that

$$\mathrm{GIC}(\lambda) = n^{-1}\sum_{i=1}^{n}\varepsilon_i^2 + \bar{L}_n(\lambda) + \mathrm{o_p}\{\bar{L}_n(\lambda)\},$$

where $\bar{L}_n(\lambda) = L_n(\lambda) + (N+p)\sigma_0^2(\gamma_n - 2)/n.$ Hence, we can use $\bar{L}_n(\lambda)$ as the risk to study the asymptotic optimality of GIC. Let $\bar{\lambda}$ be the model selected by GIC criterion. Then, we obtain the following asymptotic results which are analogous to Theorem 1.

Theorem 2. (1) *Under the conditions of Theorem* 1, *and if* $\gamma_n \to \infty$, $k_n\gamma_n^2/n \to 0$ *as* $n \to \infty$, *then*

$$\bar{L}_n(\bar{\lambda})/\bar{L}_n(\bar{\lambda}^*) - 1 = \mathrm{o_p}(1),$$

where $\bar{\lambda}^*$ *satisfies* $\bar{L}_n(\bar{\lambda}^*) = \inf_\lambda \bar{L}_n(\lambda)$;

(2) *In addition to the conditions of the lemma, if the assumptions in Part* (1) *hold, then*

$$\bar{L}_n(\bar{\lambda}) = \mathrm{o_p}(1) \quad \textit{and} \quad L_n(\bar{\lambda}) = \mathrm{o_p}(1).$$

Proof. See Appendix E.

Theorem 2 shows that the GIC-selected estimator is not only L_n-consistent but also $\bar{L}_n$-asymptotically efficient. In addition, the following corollary shows that SIC has the same asymptotic optimality as GIC.

Corollary 3. *Suppose that* $\bar{\lambda}$ *is the model selected by using SIC. If the conditions of Theorem* 2 *for* $\gamma_n = \log(n)$ *are all satisfied, then*

$$\bar{L}_n(\bar{\lambda})/\bar{L}_n(\bar{\lambda}^*) - 1 = \mathrm{o_p}(1) \quad \textit{and} \quad L_n(\bar{\lambda}) = \mathrm{o_p}(1).$$

Proof. Apply the same technique used in the proof of Theorem 2, except that γ_n is replaced by $\log(n)$.

To establish the connection between λ^* and $\bar{\lambda}^*$, we obtain the following theorem. The proof of this theorem is similar to the proof of Lemma 2.1 in Shao (1997), and the details are therefore omitted here.

Theorem 3. *If* h_0 *is not a spline function or a polynomial of degree less than* $m+1$, $\lambda_1^{(0)} \notin \Lambda_1$ *and*

$$\frac{(\gamma_n - 2)\sigma_0^2(N(\lambda^*) + p(\lambda^*))}{nL_n(\lambda^*)} = \mathrm{o}(1),$$

then

$$L_n(\lambda^*)/L_n(\bar{\lambda}^*) - 1 = \mathrm{o_p}(1) \quad \textit{and} \quad \bar{L}_n(\lambda^*)/\bar{L}_n(\bar{\lambda}^*) - 1 = \mathrm{o_p}(1).$$

The relationship between $\bar{L}_n$ and L_n is given in Corollary 4.

Corollary 4. *Under the same assumptions stated in Theorem* 3, $\bar{L}_n$ *and* L_n *are asymptotically equivalent.*

Proof. See Appendix F.

4. Simulation study

Monte Carlo studies were conducted to evaluate the performance of AICC, AIC, C_p, FPE, GCV, and SIC. In order to assess the impact of the dependency and the nonorthogonal structure of the predictors in model selection, we divide our studies into three subsections: simulation studies with independent predictors, simulation studies with dependent predictors, and simulation studies with nonorthogonal designs. In Sections 4.1 and 4.3, the selection criteria are used to choose the explanatory variables of the parametric component as well as the number of spline knots for estimating the nonparametric component. Since the spline knots chosen by the selection criteria cannot be measured by the selected model order, we only present the probabilities of the correct order of the true parametric component chosen by selection criteria. In Section 4.2, we assume that the parametric component of the true semiparametric model is known. Then we evaluate the performance of selection criteria by using the estimated mean squared errors calculated from the linear parametric estimates and the nonparametric function estimates, respectively.

4.1. Simulation studies with independent predictors

Five hundred realizations of size $n=15$, 20, 30, 40, and 50 were generated from the model in Eq. (2.1) with standard normal errors. For each specified interval of t, $(T_1,\dots,T_n)$ were chosen from the uniform distribution to be independent and identically distributed. Let p_0 denote the true model order of the parametric part. Seven random variables for the parametric component were chosen as candidates, and these were stored in $\boldsymbol{X}$, an $n\times 7$ matrix. Seven possible models were developed by the sequential inclusion of each of these variables. The first p_0 columns of $\boldsymbol{X}$ contained the variables which comprise the true model of the parametric component.

Assume that the rows of $\boldsymbol{X}$ are iid normal $\mathrm{N}(0,I_7)$, $\beta_0=(3,2,1)'$, $p_0=3$ and h_0 is one of three functions:

(1) $h_0=t^2/2-2/3$ where $t\in[-2,2]$,

(2) $h_0=2\sin(2\pi t)$, or

(3) $h_0=0.1\exp(4t)-(\exp(4)-1)/40$ where $t\in[0,1]$.

These three h_0 represent the following important and well-known classes of functions: polynomials, triangular polynomials and exponential functions.

Following the algorithm in Section 2.4, for each sample all seven models were evaluated by each criterion, and a candidate model with the smallest value for that criterion was selected. Table 1 lists the estimated probability that the correct order of the true parametric component was selected from 500 realizations when $n=15$, 20, 30, 40, and 50 for each of the three specified functions of h_0. For $n=15$, 20, 30, and 40, AICC outperforms its competitors in every case, as AIC, C_p, FPE, GCV and SIC tend to overfit the model. As the sample size increases to $n=50$, SIC performs better than AICC. A plausible explanation is that only a few spline knots are usually needed for fitting the nonparametric component in the semiparametric regression

Table 1
The estimated probabilities that the correct order of the true parametric component of model (2.1) was chosen by the criteria AICC, AIC, C_p, FPE, GCV and SIC from 500 realizations, where the nonparametric component is fitted with quadratic regression splines

h_0	n	AICC	AIC	C_p	FPE	GCV	SIC
$\frac{t^2}{2} - \frac{3}{2}$	15	0.674	0.028	0.224	0.082	0.238	0.034
	20	0.854	0.218	0.374	0.318	0.510	0.368
	30	0.846	0.444	0.516	0.518	0.618	0.716
	40	0.824	0.526	0.588	0.558	0.680	0.840
	50	0.808	0.544	0.580	0.562	0.674	0.894
$2\sin(2\pi t)$	15	0.534	0.018	0.222	0.080	0.238	0.030
	20	0.846	0.218	0.348	0.332	0.486	0.360
	30	0.826	0.464	0.504	0.498	0.622	0.696
	40	0.820	0.530	0.588	0.572	0.688	0.834
	50	0.810	0.552	0.592	0.578	0.700	0.900
$0.1\exp(4t) - \frac{\exp(4)-1}{40}$	15	0.660	0.020	0.214	0.078	0.246	0.032
	20	0.842	0.180	0.320	0.280	0.496	0.368
	30	0.838	0.444	0.518	0.508	0.620	0.708
	40	0.830	0.550	0.596	0.580	0.660	0.848
	50	0.800	0.546	0.574	0.554	0.672	0.890

model (see He and Shi, 1996). This component can be estimated well when the assumptions of Part (2) in Theorem 2 hold. Hence, under these assumptions, a variable selection criterion will have similar properties when used either to select the parametric component for the nonparametric regression model, or to select variables for the parametric regression model (e.g., AICC is an efficient criterion and SIC is a consistent criterion).

Another method for assessing the performance is to plot the average values of each criterion and $\Delta\{\hat{\theta}(\lambda^*)\}$, DELTA, as $n=20$ and $h_0=t^2/2-2/3$ (Fig. 1), where λ^* is selected with $U_n(\lambda)$. It is apparent from this plot that the AICC curve mirrors that of $\Delta\{\hat{\theta}(\lambda^*)\}$ as $p \geqslant p_0$, whereas the rest of the criteria are strongly negatively biased estimators of $E_G[\Delta\{\hat{\theta}(\lambda^*)\}]$ as $p \geqslant p_0$ and therefore tend to favor large model order.

4.2. Simulation studies with dependent predictors

In this subsection, 500 realizations were generated from the true model

$$y_i = 2(x_{1i} + z_i/2) + x_{2i} + h_0(z_i) + \varepsilon_i, \tag{4.1}$$

where ε_i, x_{1i} and x_{2i} are iid standard normal random variables, $h_0(z_i) = z_i^2/2 - 2/3$ and z_i is iid $U(-2,2)$ random variable for $i=1,\dots,n$. Hence, $\beta_0=(2,1)'$ and $p_0=2$. Five sample sizes were used: $n=15$, 20, 30, 40 and 50, and the candidate models were constructed as follows: the first column of X is composed of $X_1 + Z/2$ with $X_1=(x_{11},\dots,x_{1n})'$ and $Z=(z_1,\dots,z_n)'$, the second column is $X_2=(x_{21},\dots,x_{2n})'$, and the rows from column 3 to column 7 of X are independent and identically drawn

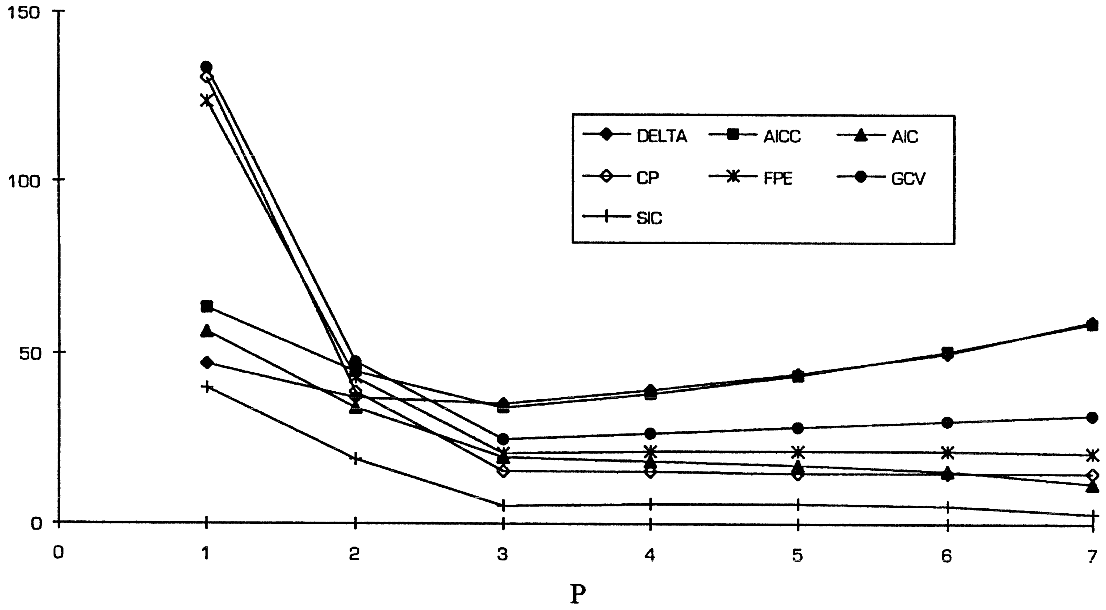

Fig. 1. Average criterion functions and Kullback–Leibler discrepancy in 500 realizations from a semiparametric regression model with $p_0 = 3$, $n = 20$, and $h_0 = t^2/2 - \frac{2}{3}$.

Table 2
The estimated probabilities that the correct order of the true parametric component of model (4.1) was chosen by the criteria AICC, AIC, C_p, FPE, GCV and SIC from 500 realizations, where the nonparametric component is fitted with quadratic regression splines

n	AICC	AIC	C_p	FPE	GCV	SIC
15	0.748	0.008	0.198	0.062	0.204	0.026
20	0.834	0.216	0.378	0.324	0.530	0.398
30	0.818	0.436	0.484	0.484	0.618	0.720
40	0.801	0.505	0.508	0.540	0.650	0.840
50	0.776	0.570	0.592	0.580	0.670	0.884

from $\mathrm{N}(0, I_5)$. Table 2 lists the estimated probabilities that the correct order of the true parametric component of model (4.1) is chosen by the criteria AICC, AIC, C_p, FPE, GCV and SIC from 500 realizations. For $n = 15$, 20, and 30, AICC outperforms AIC, C_p, FPE, GCV and SIC, whereas SIC performs better than AICC as $n = 40$ and 50. These findings basically are the same as those in Section 4.1. In other words, these studies indicate that the dependency of predictors does not affect the performance of AIC, C_p, AICC, FPE, GCV, and SIC. This suggests that AICC should be used routinely in semiparametric regression model selections when the sample size is small, and that SIC should be considered for large sample sizes.

In the rest of this subsection, we investigate the performance of model selection criteria on the nonparametric component of model (4.1). We first assume that the parametric component of true model is known and does not need to be selected. Then, we use quadratic B-splines to estimate the nonparametric component of model (4.1).

In addition, we assess how well the spline knots, λ_2, are selected by the criteria AICC, AIC, C_p, FPE, GCV, and SIC.

Since the spline knots selected by various criteria cannot be measured by the selected model order, we use the estimated mean squared errors computed from the linear parametric estimates and the nonparametric function estimates, respectively, to explore the quality of the models selection criteria. These two measures are defined as follows:

$$\mathrm{ERR}_L = \mathrm{average}\{(\hat{\beta}-\beta_0)'\Omega^{-1}(\hat{\beta}-\beta_0)\} \tag{4.2}$$

and

$$\mathrm{ERR}_N = \mathrm{average}\left\{n^{-1}\sum_{i=1}^{n}(\pi(T_i)'\hat{\alpha}-h_0(T_i))^2\right\}, \tag{4.3}$$

where Ω is the covariance matrix of x_{0i} and x_{0i} is the ith row of $X_0=(X_1',X_2')'$.

Table 3 presents the estimated mean squared errors, ERR_L and ERR_N, and their associated standard errors for the criteria AICC, AIC, FPE, GCV and SIC when $n=15$, 20, 30, 40, and 60 computed through 1000 realizations. For $n=15$, 20, and 30, AICC outperforms AIC, C_p, FPE, GCV and SIC, but SIC is slightly better than AICC when $n=40$ and 60. It is also worth noting that the model selected by AIC produces the largest estimated mean squared error and the associated standard error. This result shows that AIC should not be used for practical applications.

4.3. Simulation example with nonorthogonal designs

In this subsection, we study the model selection for semiparametric models with nonorthogonal designs. Five hundred simulated realizations were generated from

$$y_i=x_{1i}+x_{2i}+x_{3i}+x_{4i}+h_0(z_i)+\varepsilon_i,\quad (i=1,\dots,n) \tag{4.4}$$

where the ε_i are iid standard normal, $(x_{1i},x_{2i},x_{3i},x_{4i})'$, are the nonorthogonal D-optimal designs of the associated quadratic polynomial surface in four variables, and each variable takes only the values -1, 0 or 1. These design points were obtained by using Miller and Nguyen (1994) program which is available in STATLIB. Furthermore, $h_0(z_i)=z_i^2/2-2/3$ and $z_i=-2+4(i-0.5)/n$. Then, eight nonorthogonal variables were chosen, $(x_{1i},x_{2i},x_{3i},x_{4i},x_{1i}^2,x_{2i}^2,x_{3i}^2,x_{4i}^2)$, for the parametric component of candidates, and these were stored in $\boldsymbol{X}$, an $n\times 8$ matrix. Eight possible candidate models were developed by the sequential inclusion of each of these variables. The first four columns of $\boldsymbol{X}$ contained the variables which comprise the true model of the parametric component, and their corresponding coefficients are $\beta_0=(1,1,1,1)'$.

Table 4 lists the estimated probabilities that the correct order of the true parametric component of model (4.4) was chosen by the criteria AICC, AIC, Cp, FPE, GCV and SIC from 500 simulations of the sample size $n=20$, 30, 40, and 50. When $n=20$, 30, AICC outperforms AIC, Cp, FPE, GCV and SIC, but SIC performs better than AICC when $n=40$ and 50. This finding is consistent with the results in Sections 4.1 and 4.2.

Table 3
The errors of the quadratic spline fits for model (4.1) with AICC, AIC, FPE, GCV, and SIC criteria in 1000 realizations, where the numbers in the parentheses are the associated standard deviations of individual errors from simulated samples

n	Criteria	ERR_L	ERR_N
15	AICC	0.2216(0.26)	0.3216(0.31)
	AIC	0.4014(0.68)	0.7287(0.66)
	C_p	0.3517(0.59)	0.6446(0.68)
	FPE	0.3618(0.58)	0.6648(0.67)
	GCV	0.3026(0.41)	0.5383(0.49)
	SIC	0.3527(0.54)	0.6425(0.62)
20	AICC	0.1521(0.17)	0.2643(0.24)
	AIC	0.2000(0.26)	0.4433(0.34)
	C_p	0.1932(0.27)	0.4056(0.32)
	FPE	0.1934(0.25)	0.4247(0.34)
	GCV	0.1790(0.23)	0.3737(0.32)
	SIC	0.1842(0.24)	0.3798(0.33)
30	AICC	0.0913(0.10)	0.1942(0.17)
	AIC	0.1001(0.11)	0.2737(0.21)
	C_p	0.0971(0.11)	0.2554(0.20)
	FPE	0.0995(0.10)	0.2685(0.21)
	GCV	0.0960(0.10)	0.2396(0.20)
	SIC	0.0919(0.10)	0.2036(0.18)
40	AICC	0.0613(0.06)	0.1468(0.12)
	AIC	0.0660(0.06)	0.1926(0.16)
	C_p	0.0645(0.07)	0.1792(0.15)
	FPE	0.0657(0.06)	0.1887(0.15)
	GCV	0.0641(0.06)	0.1709(0.14)
	SIC	0.0603(0.06)	0.1341(0.12)
60	AICC	0.0398(0.04)	0.0993(0.08)
	AIC	0.0404(0.04)	0.1148(0.09)
	C_p	0.0401(0.04)	0.1107(0.09)
	FPE	0.0403(0.04)	0.1142(0.09)
	GCV	0.0402(0.04)	0.1083(0.08)
	SIC	0.0392(0.04)	0.0789(0.07)

Table 4
The estimated probabilities that the correct order of the true parametric component of model (4.4) was chosen by the criteria AICC, AIC, C_p, FPE, GCV and SIC from 500 realizations, where the nonparametric component is fitted with quadratic regression splines

n	AICC	AIC	C_p	FPE	GCV	SIC
20	0.904	0.168	0.366	0.288	0.510	0.294
30	0.862	0.360	0.444	0.430	0.616	0.678
40	0.830	0.506	0.588	0.564	0.684	0.830
50	0.828	0.566	0.606	0.592	0.704	0.894

5. Discussion

For semiparametric regression models, the model selection criteria AICC, AIC, Cp, GCV and FPE all make asymptotically efficient selections. Use of AICC in small sample Monte Carlo studies results in a better model choice, primarily because AICC gives less biased estimates of the expected Kullback–Leibler information. When the sample size is large, the error term r_{1n} in the approximation of AICC to the expected Kullback-Leibler information (see Eq. (2.5)) cannot be ignored. Since the nonparametric component can be estimated well with a few spline knots under the assumptions of Part (2) in Theorem 2, a variable selection criterion will have similar properties when used either to select the parametric component for the semiparametric regression model, or to select variables for the parametric regression model. Therefore, SIC may be preferable for semiparametric regression model selection when the sample size is large and the nonparametric part can be estimated well.

As suggested by a referee, a sensible alternative to modifying AIC (2.7) is to replace its penalty with $2N + \log(n)p$. This replacement results in a selection criterion that not only produces an efficient estimator of the mean function, but also is a consistent estimator of the order of the parametric part. This criterion needs to be further explored, and is under investigation.

Acknowledgements

The authors gratefully acknowledge the associate editor and the referee for their helpful comments and suggestions. The research of Peide Shi was supported by NSF of China, and the research of Chih-Ling Tsai was supported by NSF grant DMS 95-10511.

Appendix A. The derivation of $\hat{\theta}(\lambda)$

In the approximating family, the log-likelihood function of the semiparametric model is

$$\log(f(y|\theta(\lambda))) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) \\ - \|\boldsymbol{Y} - \boldsymbol{X}(\lambda)\beta(\lambda) - \Pi(\lambda)\alpha(\lambda)\|^2/2\sigma^2.$$

For the given λ, the maximum likelihood estimator of $\theta(\lambda)$ can be obtained by solving the equation $[\partial \log f(y|\theta(\lambda))]/\partial\theta = 0$. The resulting estimator is $\hat{\theta}(\lambda) = (\hat{\beta}(\lambda)', \hat{\alpha}(\lambda)', \hat{\sigma}^2(\lambda))'$, and

$$\hat{\beta}(\lambda) = (\boldsymbol{X}(\lambda)'Q(\lambda)\boldsymbol{X}(\lambda))^{-1}\boldsymbol{X}(\lambda)'Q(\lambda)\boldsymbol{Y},$$

$$\hat{\alpha}(\lambda) = (\Pi(\lambda)'\Pi(\lambda))^{-1}\Pi(\lambda)'(I - \boldsymbol{X}(\lambda)(\boldsymbol{X}(\lambda)'Q(\lambda)\boldsymbol{X}(\lambda))^{-1}\boldsymbol{X}(\lambda)'Q(\lambda))\boldsymbol{Y},$$

$$\hat{\sigma}^2(\lambda) = \frac{1}{n}\boldsymbol{Y}'(I - H)\boldsymbol{Y},$$

where $Q(\lambda)$ and H are defined in Eq. (2.4).

Appendix B. The distributions of $(Y-\mu)'H(Y-\mu)/\sigma_0^2$ and $Y'(I-H)Y/\sigma_0^2$

Since the true model can be reexpressed as in Eq. (2.3) and H is an idempotent matrix, the distribution of $Y'(I-H)Y/\sigma_0^2$ is noncentral chi-square, $\chi^2_{n-N-p,\delta}$, with noncentrality parameter $\delta=R_n'(I-H)R_n$. In addition, $(Y-\mu)'H(Y-\mu)/\sigma_0^2$ is distributed as χ^2_{N+p}, and is independent of the $Y'(I-H)Y/\sigma_0^2$ distribution.

Appendix C. The Proof of Theorem 1

Let $\mathrm{ASR}(\lambda)=\|Y-X(\lambda)\hat\beta(\lambda)-\Pi(\lambda)\hat\alpha(\lambda)\|^2/n$. Note that AIC has the equivalent form

$$\mathrm{AIC}(\lambda)=\mathrm{ASR}(\lambda)\exp\left(\frac{2(p+N)}{n}\right)$$

and $\mathrm{ASR}(\lambda)=L_n(\lambda)+\delta_n(\lambda)+(1/n)\varepsilon'\varepsilon+(2/n)\varepsilon'(I-H)R_n^*-(2/n)\varepsilon'H\varepsilon$, where $R_n^*=\widetilde{R}_n+R_n$, $\delta_n(\lambda)=U_n(\lambda)-L_n(\lambda)$, R_n is defined in Eq. (2.3), $\widetilde{R}_n=X(\lambda)\beta^*(\lambda)-X_0\beta_0$, $\beta^*(\lambda)$ is a vector which minimizes $\|X(\lambda)\beta(\lambda)-X_0\beta_0\|^2$ over $\beta(\lambda)\in R^p$, and p is the number of columns in $X(\lambda)$. Therefore,

$$\mathrm{AIC}(\lambda)=L_n(\lambda)+\delta_n(\lambda)+\frac{\varepsilon'\varepsilon}{n}+\frac{2}{n}\varepsilon'(I-H)R_n^*+\frac{2\mathrm{tr}(H)\sigma_0^2}{n}$$
$$-\frac{2\varepsilon'H\varepsilon}{n}+\frac{2(p+N)}{n}(ASR(\lambda)-\sigma_0^2)+V_n,$$

where

$$V_n=\mathrm{ASR}(\lambda)\left[\exp\left(\frac{2(p+N)}{n}\right)-1-\frac{2(p+N)}{n}\right].$$

Since $\sup_{\lambda\in\Lambda}L_n(\lambda)\leqslant 2\sup_{\lambda\in\Lambda}(R_n'R_n/n+\beta_0'\Sigma\beta_0)\leqslant D_1<\infty$ for some constant D_1 and $\varepsilon'\varepsilon/n$ is independent of λ, our Theorem 1 follows if we can show that in probability

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}U_n(\lambda)/L_n(\lambda)=1,$$

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}\frac{2(p+N)}{n}|ASR(\lambda)-\sigma_0^2|/L_n(\lambda)=0,$$

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}\frac{2}{n}|\varepsilon'(I-H)R_n^*|/L_n(\lambda)=0, \tag{C.1}$$

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}\frac{2}{n}|\mathrm{tr}(H)\sigma_0^2-\varepsilon'H\varepsilon|/L_n(\lambda)=0, \tag{C.2}$$

and

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}|V_n|/L_n(\lambda)=0. \tag{C.3}$$

A simple calculation yields

$$U_n(\lambda)=(\varepsilon' H\varepsilon+R_n^{*\prime}(I-H)R_n^*)/n,\quad |U_n(\lambda)/L_n(\lambda)-1|=\frac{n^{-1}|\varepsilon' H\varepsilon-\mathrm{tr}(H)\sigma_0^2|}{L_n(\lambda)}$$

and $L_n(\lambda)=(\mathrm{tr}(H)\sigma_0^2+R_n^{*\prime}(I-H)R_n^*)/n\geqslant\mathrm{tr}(H)\sigma_0^2/n$. Observe that

$$\frac{2(p+N)}{n}(\mathrm{ASR}(\lambda)-\sigma_0^2)=\frac{2(p+N)}{n}(\varepsilon'(I-H)\varepsilon/n-\sigma_0^2)+\frac{2(p+N)}{n}U_n(\lambda)$$
$$+\frac{2(p+N)}{n^2}\varepsilon'(I-H)R_n^*.$$

From these results, and given the law of large numbers, it suffices for the derivation to show Eqs. (C.1)–(C.3).

Noticing that $\sup_{\lambda\in\Lambda}(p(\lambda)+N(\lambda))/n\to 0$ $(n\to\infty)$, $\sup_{\lambda\in\Lambda}\mathrm{ASR}(\lambda)=\mathrm{O_P}(1)$ and $L_n(\lambda)\geqslant\sigma_0^2\mathrm{tr}(H)/n$, we can conclude Eq. (C.3).

Given any $\delta>0$, from Chebychev's Inequality we have

$$\mathscr{P}\left\{\sup_{\lambda\in\Lambda}\frac{1}{n}|2\mathrm{tr}(H)\sigma_0^2-2\varepsilon' H\varepsilon|/L_n(\lambda)>\delta\right\}\leqslant\sum_{\lambda\in\Lambda}\frac{n^{-6}E_G|2\mathrm{tr}(H)\sigma_0^2-2\varepsilon' H\varepsilon|^6}{L_n^6(\lambda)\delta^6}. \tag{C.4}$$

Also, from Theorem 2 of Whittle (1960) we obtain that

$$E_G|2\mathrm{tr}(H)\sigma_0^2-2\varepsilon' H\varepsilon|^6\leqslant D_2(\mathrm{tr}(HH')^3)$$

for some constant $D_2>0$, and $\sigma_0^2 n^{-1}\mathrm{tr}(HH')\leqslant L_n(\lambda)$. As a result,

$$\sum_{\lambda\in\Lambda}\frac{n^{-6}E_G|2\mathrm{tr}(H)\sigma_0^2-2\varepsilon' H\varepsilon|^6}{L_n^6(\lambda)\delta^6}\leqslant D_2\sum_{\lambda\in\Lambda}\frac{1}{(nL_n(\lambda))^3\delta^6}. \tag{C.5}$$

If we let D_3 be the cardinal number of Λ_1, then $D_3<\infty$. Note that $nL_n(\lambda)\geqslant\sigma^2\mathrm{tr}(H)$ and $\sum_{\lambda_2\in\Lambda_2}1/N^3(\lambda_2)<\infty$. For any given $\delta>0$, there is a subset of Λ_2, Λ_2^*, such that it contains only a finite number of knot sets and $\sum_{\lambda_2\in\Lambda_2\setminus\Lambda_2^*}D_3/\sigma_0^6 N^3(\lambda_2)<\delta/2$. Consequently,

$$\sum_{\lambda\in\Lambda}\frac{1}{(nL_n(\lambda))^3}\leqslant D_3 D\sup_{\lambda_1\in\Lambda_1,\lambda_2\in\Lambda_2^*}\frac{1}{(nL_n(\lambda_1,\lambda_2))^3}+D_3\sum_{\lambda_2\in\Lambda_2\setminus\Lambda_2^*}(N(\lambda_2)+1)^{-3}/\sigma_0^6$$
$$\leqslant D_3 D\sup_{\lambda\in\Lambda}\frac{1}{(nL_n(\lambda))^3}+\delta/2,$$

where D is the cardinal number of Λ_2^*. By assumption, $\inf_{\lambda\in\Lambda}nL_n(\lambda)\to\infty$ as $n\to\infty$. Hence, for any given D, $\sup_{\lambda\in\Lambda}D_3 D/(nL_n(\lambda))^3$ tends to zero with n. These assertions, along with Eqs. (C.4) and (C.5), imply (C.2).

For Eq. (C.1), we can find a constant $D_4 > 0$ such that

$$n^{-6}E_G|\varepsilon'(I-H)R_n^*|^6 \leqslant D_4(R_n^{*\prime}(I-H)R_n^*/n)^3 n^{-3} \leqslant D_4(L_n(\lambda)/n)^3.$$

From the last inequality and an argument similar to that used in the proof of Eq. (C.2), we can easily obtain (C.1).

Appendix D. The Proof of the Lemma

Since $\lambda_1^{(0)} \in A_1$, $X_0\beta_0 + h = X(\lambda)\tilde{\beta}(\lambda) + \Pi\alpha_0^* + R_n$ and $X(\lambda)\hat{\beta}(\lambda) + \Pi\hat{\alpha}(\lambda) = HY = X(\lambda)\tilde{\beta}(\lambda) + \Pi\alpha_0^* + HR_n + H\varepsilon$, we have $nL_n(\lambda) = E_0\|H\varepsilon - (I-H)R_n\|^2 = \mathrm{tr}(H)\sigma_0^2 + R_n'(I-H)R_n$. Let $\tilde{\lambda}_n = (\tilde{\lambda}_1, \tilde{\lambda}_2)$ be a sequence of models such that $\lambda_1^{(0)} \subseteq \tilde{\lambda}_1$ and $N(\tilde{\lambda}_2) = k_n$. Then,

$$\begin{aligned} 0 \leqslant L_n(\lambda^*) &\leqslant L_n(\tilde{\lambda}_n) = \sigma_0^2(k_n + p)/n + R_n'(I-H)R_n/n \\ &\leqslant \sigma_0^2(k_n + p_{\text{full}})/n + \mathrm{O}(k_n^{-m}) \to 0 \end{aligned}$$

as $n \to \infty$, where p_{full} is the number of columns of $X(\lambda)$ under the full model. This result, in connection with Assumption 2, proves the lemma.

Appendix E. The Proof of Theorem 2

In order to prove the first part of Theorem 2, we first reexpress ASR(λ) and AIC(λ) as follows:

$$\mathrm{ASR}(\lambda) = L_n(\lambda) + \delta_n(\lambda) + \frac{1}{n}\varepsilon'\varepsilon + \frac{2}{n}\varepsilon'(I-H)R_n^* - \frac{2}{n}\varepsilon'H\varepsilon$$

and

$$\begin{aligned} \mathrm{AIC}(\lambda) = \bar{L}_n(\lambda) + \delta_n(\lambda) + \frac{\varepsilon'\varepsilon}{n} + \frac{2}{n}\varepsilon'(I-H)R_n^* + \frac{2\mathrm{tr}(H)\sigma_0^2}{n} - \frac{2\varepsilon'H\varepsilon}{n} \\ + \frac{(p+N)\gamma_n}{n}(ASR(\lambda) - \sigma_0^2) + W_n, \end{aligned} \tag{E.1}$$

where

$$W_n = ASR(\lambda)\left[\exp\left(\frac{(p+N)\gamma_n}{n}\right) - 1 - \frac{(p+N)\gamma_n}{n}\right]$$

and $\delta_n(\lambda)$ and R_n^* are defined in the proof of Theorem 1. From Eq. (E.1), the first part of Theorem 2 follows if we can show that in probability

$$\lim_{n\to\infty}\sup_{\lambda\in A} U_n(\lambda)/L_n(\lambda) = 1,$$

$$\lim_{n\to\infty}\sup_{\lambda\in A} \frac{2}{n}|\varepsilon'(I-H)R_n^*|/L_n(\lambda) = 0,$$

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}\frac{2}{n}|\mathrm{tr}(H)\sigma_0^2-\varepsilon' H\varepsilon|/L_n(\lambda)=0,$$

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}\frac{(p+N)\gamma_n}{n}|ASR(\lambda)-\sigma_0^2|/L_n(\lambda)=0, \tag{E.2}$$

and

$$\lim_{n\to\infty}\sup_{\lambda\in\Lambda}|W_n|/L_n(\lambda)=0. \tag{E.3}$$

By assumptions and arguments similar to those as given in the proof of Theorem 1, it suffices to show Eqs. (E.2) and (E.3) which are given below.

Note that

$$\frac{(p+N)\gamma_n}{n}(\mathrm{ASR}(\lambda)-\sigma_0^2)=\frac{(p+N)\gamma_n}{n}(\varepsilon'(I-H)\varepsilon/n-\sigma_0^2)+\frac{(p+N)\gamma_n}{n}U_n(\lambda)$$
$$+\frac{(p+N)\gamma_n}{n^2}\varepsilon'(I-H)R_n^*,$$

$$\frac{(p+N)\gamma_n}{n}(\varepsilon'(I-H)\varepsilon/n-\sigma_0^2)=\frac{(p+N)}{n}\mathrm{O_P}(\gamma_n n^{-1/2})+\frac{(p+N)^2\gamma_n\sigma_0^2}{n^2}$$

and

$$L_n(\lambda)\geqslant(N+p)\sigma_0^2/n.$$

Hence, Eq. (E.2) follows. Next, assertion (E.3) follows from the fact that

$$W_n=\mathrm{ASR}(\lambda)\mathrm{O}\left(\frac{(N+p)\gamma_n^2}{n}\right)\frac{N+p}{n},\quad \mathrm{ASR}(\lambda)=\mathrm{O_p}(1),$$

$$L_n(\lambda)\geqslant(N+p)\sigma_0^2/n \text{ and } \frac{k_n\gamma_n^2}{n}\to 0.$$

Finally, applying techniques similar to those as given in the proof of the lemma, together with the result from the first part of theorem 2, we can easily prove the second part of this theorem.

Appendix F. The Proof of Corollary 4

Let $H=H(\lambda)$ be the hat matrix in Eq. (2.4) determined by model λ. Using the fact that

$$nL_n(\lambda)=\mathrm{tr}(H(\lambda))\sigma_0^2+\mu'(I-H(\lambda))\mu, \tag{F.1}$$

and

$$n\bar{L}_n(\lambda)=\mathrm{tr}(H(\lambda))\sigma_0^2+\mu'(I-H(\lambda))\mu+\mathrm{tr}(H(\lambda))\sigma_0^2(\gamma_n-2), \tag{F.2}$$

we have

$$\frac{\bar{L}_n(\bar{\lambda}^*)}{L_n(\lambda^*)} = \frac{L_n(\bar{\lambda}^*)}{L_n(\lambda^*)} + \frac{\mathrm{tr}(H(\bar{\lambda}^*))\sigma_0^2(\gamma_n - 2)}{nL_n(\lambda^*)}. \tag{F.3}$$

Since $\mathrm{tr}(H(\lambda))\sigma_0^2(\gamma_n - 2)$ is an increasing function of $\mathrm{tr}(H(\lambda))$, the term $\mathrm{tr}(H(\bar{\lambda}^*)$ in the minimizer of Eq. (F.2) is smaller than the term $\mathrm{tr}(H(\lambda^*)$ in the minimizer of Eq. (F.1) when n is large enough. This result together with Eq. (F.3) and assumptions stated in Theorem 3 yield that $\bar{L}_n$ and L_n are asymptotically equivalent.

References

Akaike, H., 1970. Statistical predictor identification. Ann. Inst. Statist. Math. 22, 203–217.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: 2nd Internat. Symp. on Information Theory, Petrov, B.N., Csaki, F., Akademia Kiado, Budapest, pp. 267–281.

Allen, D.M., 1974. The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16, 125–127.

Breiman, L., 1991. The Π method for estimating multivariate functions from noisy data (with discussion). Technometrics 33, 125–143.

Burman, P., Chen, K.W., 1989. Nonparametric estimation of a regression function. Ann. Statist. 17, 1567–1596.

Chen, H., Shiau, J.J.H., 1994. Data-driven efficient estimators for a partially linear model. Ann. Statist. 22, 211–237.

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

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

Friedman, J.H., 1991. Multivariate adoptive regression splines (with discussion). Ann. Statist. 19, 1–67.

Friedman, J.H., Silverman, B.W., 1989. Flexible parsimonious smoothing and additive modeling (with discussion). Technometrics 31, 3–21.

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

Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.

He, X., Shi, P., 1996. Bivariate tensor-product B-splines in a partly linear model. J. Multivariate Anal. 58, 162–181.

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

Hurvich, C.M., Tsai, C.L., 1995. Relative rates of convergence for efficient model selection criteria in linear regression. Biometrika 82, 418–425.

Hurvich, C.M., Simonoff, J.S., Tsai, C.L., 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. Roy. Statist. Soc., to appear.

Li, K.C., 1987. Asymptotic optimality for C_P, C_L, cross-validation and generalized cross-validation: discrete index set. Ann. Statist. 15, 958–975.

Mallows, C.L., 1973. Some comments on Cp. Technometrics 15, 661–675.

Miller, A.J., Nguyen, N.K., 1994. A Fedorov exchange algorithm for D-optimal Design. Appl. Statist. 43, 669–678.

Nishii, R., 1984. Asymptotic properties of criteria for selection of variables in multiple regression. Ann. Statist. 12, 758–765.

Schumaker, L.L., 1981. Spline Functions. Wiley, New York.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Shao, J., 1997. An asymptotic theory for linear model selection. Statistica Sinica 7, 221–242.

Shi, P.D., 1995. An efficient algorithm for fitting additive models with regression splines. Technical Report, Peking University, China.

Shi, P.D., Li, G.Y., 1995. Global rates of convergence of B-spline M-estimates for nonparametric regression. Statistica Sinica 5, 303–318.

Shibata, R., 1981. An optimal selection of regression variables. Biometrika 68, 45–54.

Wahba, G., 1990. Spline Models for Observational Data. Monograph: SIAM, CBMS-NSF regional Conf. Series in Applied Mathematics, vol. 59.

Whittle, P., 1960. Bounds for the moments of linear and quadratic forms in independent variables. Theory Probab. Appl. 5, 302–305.