SEMIPARAMETRIC MODEL SELECTION IN LARGE SAMPLES*

SHI Peide WANG Haiyan ZHENG Zhongguo
(*Department of Probability and Statistics, Peking University, Beijing* 100871, *China*)

Abstract. For semiparametric regression model selection, based on a model selection criterion there is no finite order (or number of parameters) of the nonparametric part to be estimated consistently, but there is a finite order (or number of predictor variables) of the linear part to be estimated consistently. The models selected by using AIC and $AICC$ are not consistent estimates of linear part of the true model. In this paper, we study the consistency in model selection by investigating the asymptotic properties of AIC^* and $AICC^*$, the modified versions of AIC and $AICC$ respectively, which were proposed by a referee of the reference Shi and Tsai. Under some regular conditions, we prove that the parametric models of the semiparametric regression selected with AIC^* and $AICC^*$ converge to the true model in probability. In addition, in terms of the mean integrated squared error plus a penalty, these two criteria can also provide an asymptotically efficient selection.

Key words. AIC, $AICC$, model selection, information criterion.

1 Introduction

There are obvious reasons for the popularity of linear regression among which are interpretation and simplicity in computation. But this does not mean that a linear relationship is always sufficient. In some applications, the mean is linearly related to some variables but the relation to additional variables are not easily parameterized. Partly linear models become the natural choices in such applications: the linear model is minimally altered to allow one or a few of the independent variables to have complicated effects (cf. [1] and [2]).

Suppose that the responses $Y_1, Y_2, \cdots, Y_n$ are generated from the true model

$$Y_i = X_{0i}^T \beta_0 + g_0(T_i) + \varepsilon_i, \quad 1 \le i \le n, \tag{1.1}$$

where the Y_i's are real-valued responses, $X_{0i} \in R^{p_0}$ and $T_i \in [0,1]$ are known explanatory variables and the ε_i's are i.i.d. random errors. This model consists of a p_0-dimensional parameter β_0 and an unspecified univariate function g_0.

The increasing recognition of partly linear models has attracted a number of authors to study the asymptotic behavior of both the parameter and function estimates (see [3]–[8], among others). From a practical point of view, however, model (1.1) is only of theoretical interest because the true model is generally unknown in practical applications. One may fit the approximating models

$$Y_i = X_i^T \beta + g_{n\theta}(T_i) + e_i, \quad 1 \le i \le n, \tag{1.2}$$

Received May 25, 1998.
Revised January 17, 2000.
*This research supported in part by Postdoctoral Science Foundation and NSF of China.

instead of the true one, where $X_i \in R^p$, $g_{n\theta}$ is a function with unknown parameter θ, and the e_i's are i.i.d random errors. For example, $g_{n\theta}$ may be a spline function (cf. [8]), or a kernel estimate of g_0 given β (cf. [7]). The use of an excessive number of predictor variables in applications usually reduces prediction accuracy (cf. [9]). One interesting question is how to identify the true model among all possible combinations of the linear predictor variables based on the observed data. Selecting the linear predictor variables based on data is a powerful tool for model choice.

For nonparametric regression models such as nearest neighbor nonparametric regression (cf. [10]), regression splines (cf. [11]), and generalized Fourier series regression (cf. [12]), there are infinitely many parameters. In this case, there is no finite order (or number of parameters) to be estimated consistently.

For linear regression models, in contrast to efficient selection criteria, Schwarz in [13] proposed a consistent selection criterion, BIC, defined by

$$BIC = n\log(\hat{\sigma}^2) + p\log(n),$$

where $\hat{\sigma}^2$ is the estimate of the variance of the associated approximating model and p is the dimension of the regression coefficients. In other words, when the true model is finite-dimensional, except for an event whose probability tends to zero with sample size, it always selects the correct model. The detailed discussions about efficient and consistent criteria can be found in [14].

Recently, Shi and Tsai in [15] developed $AICC$ for semiparametric regression model selection,

$$AICC = n\log(\hat{\sigma}^2) + 2(p+N) + \frac{2(p+N+1)(p+N+2)}{(n-p-N-2)}, \tag{1.3}$$

where $g_{n\theta}$ is a spline function with a vector of parameters θ, N is the dimension of θ, p is the number of parameters in the linear part of the approximating model and $\hat{\sigma}^2$ is the estimate of the variance of the random errors. They used $AICC$ to select variables and smoothing parameters from approximating models and showed that $AICC$ and other asymptotically equivalent criteria provide an asymptotically efficient selection. Specifically, the $AICC$-selected estimator of the regression function is asymptotically as good, in terms of the mean (integrated) squared error, as the estimator which uses the best approximating model in the class of candidates.

For a general model selection problem, Shibata in [16] showed that consistent selectors do not produce efficient estimators. In the semiparametric regression model selection case, although there is no finite order (or number of parameters) of the nonparametric part to be estimated consistently, there is a finite order (or number of predictor variables) of the linear part to be estimated consistently. Hence, the focus of semiparametric regression selection is on obtaining an efficient estimation of the nonparametric part (cf. [17]) and a consistent estimation of the linear part. To incorporate both of these two situations, one of the referee of [15] suggested that a sensible modification of $AICC$ defined in (1.3) is to replace its penalty with $p\log(n) + 2N + 2(N+1)(N+2)/(n-N-2)$, that is, we may use the following modified AIC and $AICC$ criteria

$$AIC^* = n\log(\hat{\sigma}^2) + p\log(n) + 2N$$

and

$$AICC^* = n\log(\hat{\sigma}^2) + p\log(n) + 2N + \frac{2(N+1)(N+2)}{(n-N-2)}$$

for partly linear model selection. This replacement results in selection criteria, each of which may not only produce an efficient estimator of the mean function but also be a consistent estimation of the order of the parametric part. However, they did not get the asymptotic results for these two criteria.

In this paper, we study the asymptotic properties of AIC^* and $AICC^*$. Under some regular conditions, we prove that the parametric models of the semiparametric regression selected with AIC^* and $AICC^*$ converge to the true model in probability. In addition, in terms of the mean integrated squared error plus a penalty (see [14]), these two criteria can provide an asymptotically efficient selection. The main results are introduced in Section 2 and their technical proofs are given in Section 3.

2 Main Results

We basically follow the method of [18] (see also [19]) by using B-spline function $g_{n\theta}(t)$ to approximate the smooth unknown function $g_0(t)$ in (1.1).

Let k_n be a positive integer. Given a partition $0 = s_0 < s_1, \cdots < s_{k_n} = 1$ of $[0,1]$, we denote by $\pi_i(t)$ $(i = 1, 2, \cdots, k_n + m)$ the normalized B-splines (of order $m+1$) associated with an extended partition of $[0,1]$ determined by $\{s_i\}$ (cf. [18]). The details can be seen in [19]. The spline knot s_i is placed on the i/k_nth quantile of $T_1, T_2, \cdots, T_n$ as in [11], where the class of the resulting knot sets is denoted by Λ_2. Let $\pi(t) = (\pi_1(t), \pi_2(t), \cdots, \pi_{k_n+m}(t))'$. Then, $g_{n\theta}(t) = \pi(t)'\theta$ and θ is an unknown projection parameter of $g_0(t)$.

Let $\Lambda_1 = \{\lambda_1 : \lambda_1 \text{ is a subset of } \{1, 2, \cdots, p\}\}$, $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)'$, $\boldsymbol{e} = (e_1, e_2, \cdots, e_n)'$, $\boldsymbol{X}_0 = (X_{01}, X_{02}, \cdots, X_{0n})'$, $\boldsymbol{X}(\lambda) = (X_1, X_2, \cdots, X_n)'$, $g_{n\theta\lambda} = (g_{n\theta}(T_1), g_{n\theta}(T_2), \cdots, g_{n\theta}(T_n))'$, and $\Lambda_1^0 = \{\lambda_1 \in \Lambda_1 : \boldsymbol{X}_0\beta_0 = \boldsymbol{X}(\lambda)\beta(\lambda), \lambda = (\lambda_1, \lambda_2)\}$, which is a subset of Λ_1.

From Assumption 3 below, there exist constants $C > m$ and $W_1 > 0$ depending only on m such that

$$g_0(T) = \pi(T)'\theta^*(g_0) + R_{nT} \quad \text{and} \quad \sup_{T\in[0,1]} |R_{nT}| \le W_1 k_n^{-C}, \tag{2.1}$$

where $\theta^*(g_0)$ is a vector depending on g_0. Then, the model in (1.2) can be rewritten in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}(\lambda)\beta(\lambda) + g_{n\theta\lambda} + \boldsymbol{e} \tag{2.2}$$

for $\lambda = (\lambda_1, \lambda_2)$, $\lambda_1 \in \Lambda_1$, and $\lambda_2 \in \Lambda_2$. Let λ_1^0 be the model in Λ_1^0 with the smallest dimension, and let $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2)$ denote the model selected from (2.2) by using AIC^*. By definition, we can see that λ_1^0 is the best model in the linear part of (2.2), $AIC^*(\hat{\lambda})$ is the smallest among those for all possible candidate models and the linear part of each model associated with $\lambda_1 \in \Lambda_1^0$ includes that of the true model.

Definition 1 The selection criterion AIC^* is said to be consistent if

$$p\{\hat{\lambda}_1 = \lambda_1^0\} \longrightarrow 1$$

as n tends to infinity.

Definition 2 A random variable ξ is sub-Gaussian if there exists a constant $\varphi > 0$ such that for all real d

$$E\{\exp(d\xi)\} \le \exp(\varphi d^2).$$

This definition is adopted from [20].

Assumption 1 Both $\boldsymbol{X}(\lambda)'Q(\lambda)\boldsymbol{X}(\lambda)$ and $\Pi(\lambda)'\Pi(\lambda)$ are of full rank for any $\lambda \in \Lambda$, where $Q(\lambda) = I - \Pi(\lambda)(\Pi(\lambda)'\Pi(\lambda))^{-1}\Pi(\lambda)'$ and $\Pi(\lambda) = (\pi(T_1), \pi(T_2), \cdots, \pi(T_n))'$.

Assumption 2 $k_n \to \infty$, $k_n/n \to 0$, $\#(\Lambda_2)k_n = o(n)$ as $n \to \infty$, and there exists a positive definite matrix Σ such that $\Sigma - \boldsymbol{X}_0'\boldsymbol{X}_0/n > 0$ for all n, where $\#(\Lambda_2)$ is the cardinal number of Λ_2.

Assumption 3 g_0 is $m(> 0)$ times differentiable.

Assumption 4 $n^{-1} \parallel \mu - H(\lambda)\mu \parallel^2$ is bounded away from zero and infinity over $\lambda_1 \in \Lambda_1 \backslash \Lambda_1^0$ and $\lambda_2 \in \Lambda_2$, where $H(\lambda) = Q(\lambda)\boldsymbol{X}(\lambda)[\boldsymbol{X}(\lambda)'Q(\lambda)\boldsymbol{X}(\lambda)]^{-1}\boldsymbol{X}(\lambda)'Q(\lambda) + \Pi(\lambda)(\Pi'(\lambda)\Pi(\lambda))^{-1} \times \Pi'(\lambda)$.

For fixed models $\lambda_1 \in \Lambda_1$ and $\lambda_1^0 \in \Lambda_1^0$, based on AIC^* criterion we can get the estimation of the spline knot sets $\hat{\lambda}_2(\lambda_1)$ and $\hat{\lambda}_2(\lambda_1^0)$ respectively with the forward/backward algorithm (cf. [11]).

Assumption 5 $N(\hat{\lambda}_2(\lambda_1)) - N(\hat{\lambda}_2(\lambda_1^0)) = o_P(\log n)$ for any $\lambda_1 \in \Lambda_1^0$.

Assumption 6 $n(\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) - \hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))) = o_p(\log n)$ for any $\lambda_1 \in \Lambda_1^0$.

Theorem 1 *If Assumptions 1 through 6 are satisfied and the ε_i's are sub-Gaussian, then AIC^* is consistent.*

Remark The sub-Gaussian condition was used by Zheng and Loh in [20] for proving the consistency of model selection for linear models; Assumptions 5 and 6 mean that the spline knot selection procedure always choose similar final spline knot sets when the linear part of the approximating models includes that of the true model. Our numerical experience indicates that it is true.

Note that AIC^* is equivalent to $\hat{\sigma}^2(\lambda)\exp\{(2N + p\log(n))/n\}$. We will use the latter form. Let $U_n(\lambda) = \parallel \boldsymbol{X}(\lambda)\hat{\beta}(\lambda) + \Pi(\lambda)\hat{\theta}(\lambda) - \boldsymbol{X}_0\beta_0 - g_0 \parallel^2 /n$ for $\lambda \in \Lambda$ and $L_n(\lambda) = E_0\{U_n(\lambda)\}$, where E_0 denotes the expectation under the true model. We will prove

$$AIC^*(\lambda) = n^{-1}\sum_{i=1}^{n}\varepsilon_i^2 + \overline{L}_n(\lambda) + o_p\{\overline{L}_n(\lambda)\},$$

where $\overline{L}_n(\lambda) = L_n(\lambda) + p(\log(n) - 2)\sigma_0^2/n$ (cf. [14]). Let $\hat{\lambda}$ be the model selected by AIC^* criterion. Then, we have the following results for asymptotically efficient selection.

Theorem 2 *If Assumptions 1–3 are satisfied,*

$$\sum_{\lambda_2 \in \Lambda_2} \frac{1}{N^3(\lambda_2)} < \infty, E_0(\varepsilon_1^{12}) < \infty \quad \textit{and} \quad \inf_{\lambda \in \Lambda} nL_n(\lambda) \to \infty,$$

then

(i) *Regardless of whether there exists a correct model in the parametric part or not*

$$\frac{\overline{L}_n(\hat{\lambda})}{\inf_\lambda \overline{L}_n(\lambda)} - 1 = o_p(1);$$

(ii) *If $\lambda_1 \in \Lambda_1^0$, then, $L_n(\hat{\lambda}) = o_p(1)$.*

The first part of Theorem 2 means that the AIC^*-selected estimator of the regression function is asymptotically as good, in terms of $\bar{L}_n(\cdot)$, as the estimator which uses the best approximating model in the class of candidates. The second part shows that the AIC^*-selected estimator of the regression function is consistent under $L_n(\cdot)$.

It is also important to study the asymptotically efficient selection under $\overline{U}_n(\lambda) = U_n(\lambda) + p(\log(n) - 2)\sigma_0^2/n$ (cf. [10] and [14]). We have the following results which are the direct consequence of Theorem 2.

Corollary *Under the conditions of Theorem 2, we have*

$$\frac{\overline{U}(\hat{\lambda})}{\inf_\lambda \bar{U}_n(\lambda)} - 1 = o_P(1)$$

and

$$U_n(\hat{\lambda}) = o_P(1) \qquad \textit{when} \quad \lambda_1 \in \Lambda_1^0.$$

Theorem 3 *The associated results of Theorems* 1 *and* 2 *remain true if the criterion* AIC^* *is substituted by* $AICC^*$.

3 Proof of Theorems

The proof of Theorem 3 is similar to that of Theorems 1 and 2 and the details are, therefore, omitted here. In order to show Theorems 1 and 2, we need the following results. The proofs of Lemmas 2–4 below are deferred to Appendices.

Lemma 1 *Suppose that* $w = (w_1, w_2, \cdots, w_n)'$ *is a sequence of random variables with i.i.d. components. Assume that* $E(w_1) = 0$, $\mathrm{Var}(w_1) < \infty$, *and* $E|\, w_1 \,|^{4s} < \infty$ *for some* $s > 1$. *Let* $\hat{a} = w'Bw$, *where* B *is a sequence of nonrandom symmetric matrices. Suppose that* $\lim_{n\to\infty} E(\hat{a}) = a$. *Then,*

(1) $E \mid \hat{a} - E(\hat{a}) \mid^{2s} \leq D(s)E(w_1^{4s})(\mathrm{tr}B^2)^s$, *where* $D(s)$ *is a constant depending only on* s.

(2) $\lim_{n\to\infty} \hat{a} = a$ *almost surely (a.s.), provided that there exist* $c > 0, \eta > 1/s$ *for which* $n^{\eta}\mathrm{tr}(B^2) < c$ *when* n *is large enough.*

The above Lemma is a special version of Equation 2.3.10 and Theorem 2.4.2 in [21].

Lemma 2 *If* $E(\varepsilon_1^{4s}) < \infty$ *for some* $s > 0$ *and Assumptions* 1 *and* 2 *hold, then,*

$$\varepsilon'(I - H(\lambda_1, \lambda_2))\mu = o_p(\log n) \quad \textit{uniformly for all } \lambda_1 \in \Lambda_1 \textit{ and } \lambda_2 \in \Lambda_2.$$

Lemma 3 *Under the conditions of Theorem* 1,

$$\frac{1}{n}\varepsilon' H(\lambda_1, \lambda_2))\varepsilon = o_p(1) \quad \textit{for all} \quad \lambda_1 \in \Lambda_1, \lambda_2 \in \Lambda_2.$$

Lemma 4 *Under the conditions of Theorem* 1,

(i) $n^{-1}\mu'(I - H(\lambda_1, \hat{\lambda}_2(\lambda_1)))\mu = o_p(1)$ *for all* $\lambda_1 \in \Lambda_1^0$,

(ii) $\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) = \sigma_0^2 + o_p(1)$ *for all* $\lambda_1 \in \Lambda_1^0$.

Proof of Theorem 1 Note that

$$AIC^*(\lambda_1, \hat{\lambda}_2(\lambda_1)) = n\log[\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1))] + p(\lambda_1)\log(n) + 2N(\hat{\lambda}_2(\lambda_1))$$

and

$$AIC^*(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)) = n\log(\hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))) + p(\lambda_1)\log(n) + 2N(\hat{\lambda}_2(\lambda_1^0))$$

are the AIC^*-values for models $\lambda_1 \in \Lambda_1$ and $\lambda_1^0 \in \Lambda_1^0$ respectively, where $\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) = \frac{1}{n}\boldsymbol{Y}'[I - H(\lambda_1, \hat{\lambda}_2(\lambda_1))]\boldsymbol{Y}$ and $\hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)) = \frac{1}{n}\boldsymbol{Y}'[I - H(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))]\boldsymbol{Y}$, and the hat-matrix $H(\lambda_1, \lambda_2)$ is as in Assumption 4.

We will divide our proof into three steps.

Step 1 Here we show that in probability

$$\frac{AIC^*(\lambda_1, \hat{\lambda}_2(\lambda_1)) - AIC^*(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))}{\log(n)} = p(\lambda_1) - p(\lambda_1^0) + o_P(1) > 0 \tag{3.1}$$

for all $\lambda_1 \in \Lambda_1^0$ and $\lambda_1 \neq \lambda_1^0$. When $\lambda_1 \in \Lambda_1^0$ and $\lambda_1 \neq \lambda_1^0$,

$$\begin{aligned}
&\log(\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1))) - \log(\hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))) \\
&= \log\left(1 + \hat{\sigma}^{-2}(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))[\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) - \hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))]\right) \\
&\leq O\left(\hat{\sigma}^{-2}(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))[\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) - \hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))]\right)
\end{aligned}$$

in probability whenever $|\hat{\sigma}^{-2}(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))[\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) - \hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))]| \leq 1$. Hence, from Assumption 6 and Lemma 4(ii), we have

$$\frac{n(\log \hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) - \log \hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)))}{\log n} = o_p(1).$$

From the last equality and Assumption 5, we obtain

$$\begin{aligned}
&\frac{AIC^*(\lambda_1, \hat{\lambda}_2(\lambda_1)) - AIC^*(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))}{\log(n)} \\
&= n\frac{\log(\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1))) - \log(\hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)))}{\log n} + p(\lambda_1) - p(\lambda_1^0) + o_P(1) \\
&= p(\lambda_1) - p(\lambda_1^0) + o_p(1) > 0
\end{aligned}$$

for all $\lambda_1 \in \Lambda_1^0$ and $\lambda_1 \neq \lambda_1^0$, which is (3.1).

Step 2 We show that

$$AIC^*(\lambda_1, \hat{\lambda}_2(\lambda_1)) - AIC^*(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)) \geq n\left\{\log\left(1 + \frac{c_0}{\sigma_0^2}\right) + o_P(1)\right\} \tag{3.2}$$

for all $\lambda_1 \in \Lambda_1 \setminus \Lambda_1^0$. According to Assumption 4, there exists a constant $c_0 > 0$ such that $\lim_{n\to\infty} n^{-1}\mu'[I - H(\lambda_1, \hat{\lambda}_2(\lambda_1))]\mu \geq c_0$ for all $\lambda_1 \in \Lambda_1 \setminus \Lambda_1^0$.

Therefore, from Lemma 4(i), Lemmas 2 and 3, we have that in probability

$$\begin{aligned}
&\lim_{n\to\infty} (\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) - \hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))) \\
= &\lim_{n\to\infty} \Big\{n^{-1}\mu'(I - H(\lambda_1, \hat{\lambda}_2(\lambda_1)))\mu) - (n^{-1}\mu'(I - H(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)) \\
&\quad + \Big(\frac{2}{n}\varepsilon'(H(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)) - H(\lambda_1, \hat{\lambda}_2(\lambda_1)))\mu\Big) \\
&\quad + (n^{-1}\varepsilon'(H(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)) - H(\lambda_1, \hat{\lambda}_2(\lambda_1)))\varepsilon)\Big\} \\
= &\lim_{n\to\infty} n^{-1}\mu'[I - H(\lambda_1, \hat{\lambda}_2(\lambda_1))]\mu + o_p(1) \geq c_0.
\end{aligned}$$

This, together with Lemma 4(ii), leads to

$$\lim_{n\to\infty} \hat{\sigma}^{-2}(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))(\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) - \hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)) \geq \frac{c_0}{\sigma_0^2} > 0$$

in probability, which implies

$$\begin{aligned}
&\lim_{n\to\infty} |n[\log(\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1))) - \log(\hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)))]| \\
= &\lim_{n\to\infty} |n\log(1 + \hat{\sigma}^{-2}(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))[\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) - \hat{\sigma}^2(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))])| \\
\geq &\lim_{n\to\infty} n\left(\log\left(1 + \frac{c_0}{\sigma_0^2 + o_p(1)}\right)\right) > 0.
\end{aligned}$$

This fact and Assumption 5 imply (3.2).

Step 3 From Step 1 and Step 2, we obtain that except for an event whose probability tends to zero with n,

$$AIC^*(\lambda_1, \hat{\lambda}_2(\lambda_1)) > AIC^*(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0)) \tag{3.3}$$

for all $\lambda_1 \in \Lambda_1$ and $\lambda_1 \neq \lambda_1^0$. On the other hand, since $\hat{\lambda}_1$ is the model satisfying $AIC^*(\hat{\lambda_1}, \hat{\lambda_2}) = \min_{\lambda_1 \in \Lambda_1} AIC^*(\lambda_1, \hat{\lambda}_2(\lambda_1))$. Therefore, we obtain that $AIC^*(\hat{\lambda}_1, \hat{\lambda}_2) \leq AIC^*(\lambda_1^0, \hat{\lambda}_2(\lambda_1^0))$. This fact together with (3.3) implies the conclusion of Theorem 1.

Proof of Theorem 2 Note that the maximum likelihood estimators of $\beta(\lambda)$, $\sigma^2(\lambda)$ and $\theta(\lambda)$ are respectively defined by

$$\hat{\beta}(\lambda) = [\boldsymbol{X}(\lambda)'Q(\lambda)\boldsymbol{X}(\lambda)]^{-1}\boldsymbol{X}(\lambda)'Q(\lambda)\boldsymbol{Y}, \quad \hat{\sigma}^2(\lambda) = \tfrac{1}{n}\boldsymbol{Y}'(I-H)\boldsymbol{Y}, \text{ and}$$
$$\hat{\theta}(\lambda) = (\Pi'(\lambda)\Pi(\lambda))^{-1}\Pi'(\lambda)\{I - \boldsymbol{X}(\lambda)[\boldsymbol{X}(\lambda)'Q(\lambda)\boldsymbol{X}(\lambda)]^{-1}\boldsymbol{X}(\lambda)'Q(\lambda)\}\boldsymbol{Y},$$

where $Q(\lambda)$ and H are defined as in Assumption 4. Let $R_n = (R_{n1}, \cdots, R_{nn})'$ and $R_{ni} = R_{nT_i}$. Then $\hat{\sigma}^2(\lambda) = \frac{1}{n} \parallel \boldsymbol{Y} - \Pi(\lambda)\hat{\theta}(\lambda) - \boldsymbol{X}(\lambda)\hat{\beta}(\lambda) \parallel^2 = L_n(\lambda) + \delta_n(\lambda) + \frac{1}{n}\varepsilon'\varepsilon + \frac{2}{n}\varepsilon'(I-H)R_n^* - \frac{2}{n}\varepsilon' H\varepsilon$, where $R_n^* = \tilde{R}_n + R_n$, $\delta_n(\lambda) = U_n(\lambda) - L_n(\lambda)$, $\tilde{R}_n = \boldsymbol{X}(\lambda)\beta^*(\lambda) - \boldsymbol{X}_0\beta_0$, $\beta^*(\lambda)$ is a vector which minimizes $\parallel \boldsymbol{X}(\lambda)\beta(\lambda) - \boldsymbol{X}_0\beta_0 \parallel^2$ over $\beta(\lambda) \in R_p$, and p is the number of columns in $\boldsymbol{X}(\lambda)$. Therefore,

$$\begin{aligned} AIC^*(\lambda) = \overline{L}_n(\lambda) + \delta_n(\lambda) + \frac{\varepsilon'\varepsilon}{n} + \frac{2\varepsilon'(I-H)R_n^*}{n} + \frac{2\mathrm{tr}(H)\sigma_0^2}{n} - \frac{2\varepsilon' H\varepsilon}{n} \\ + \frac{2N + p\log(n)}{n}(\hat{\sigma}^2(\lambda) - \sigma_0^2) + Z_n, \end{aligned}$$

where $Z_n = \hat{\sigma}^2(\lambda)[\exp(\frac{2N+p\log(n)}{n}) - 1 - \frac{2N+p\log(n)}{n}]$. Since $\sup_{\lambda\in\Lambda} \overline{L}_n(\lambda) \leq 2\sup_{\lambda\in\Lambda}(\frac{R_n'R_n}{n} + \beta_0'\Sigma\beta_0) + p(\log(n)-2)\sigma_0^2/n \leq D_1 < \infty$ for some constant D_1 and $\varepsilon'\varepsilon/n$ is independent of λ, the first part of Theorem 2 follows if we can show that in probability

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda} \left[U_n(\lambda) + \frac{p(\log(n)-2)}{n}\sigma_0^2\right] / \overline{L}_n(\lambda) = 1, \tag{3.4}$$

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda} \frac{2N + p\log(n)}{n} |\hat{\sigma}^2(\lambda) - \sigma_0^2| / \overline{L}_n(\lambda) = 0, \tag{3.5}$$

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda} \frac{2}{n} |\varepsilon'(I-H)R_n^*| / \overline{L}_n(\lambda) = 0, \tag{3.6}$$

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda} \frac{2}{n} |\mathrm{tr}(H)\sigma_0^2 - \varepsilon' H\varepsilon| / \overline{L}_n(\lambda) = 0, \tag{3.7}$$

and

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda} |Z_n| / \overline{L}_n(\lambda) = 0. \tag{3.8}$$

Since $U_n(\lambda) = (\varepsilon' H\varepsilon + R_n^*(I-H)R_n^*)/n$ and $\overline{L}_n(\lambda) = [\mathrm{tr}(H)\sigma_0^2 + R_n^*(I-H)R_n^* + p(\log(n) - 2)\sigma_0^2]/n$, we have that (3.4) follows from (3.7) and

$$\left| \frac{U_n(\lambda) + p(\log(n)-2)\sigma_0^2/n}{\overline{L}_n(\lambda)} - 1 \right| = \frac{n^{-1}|\mathrm{tr}(H)\sigma_0^2 - \varepsilon' H\varepsilon|}{\overline{L}_n(\lambda)}. \tag{3.9}$$

The fact that $n\bar{L}_n(\lambda) \geq \sigma_0^2[N + p(\log(n) - 2)]$ and the law of large numbers yield

$$\frac{2N + p\log(n)}{n}\left(\frac{\varepsilon'(I-2H)\varepsilon}{n} - \sigma_0^2\right) = o_P(\bar{L}_n(\lambda)). \tag{3.10}$$

Since $n\bar{L}_n(\lambda) \geq \sigma_0^2[N + p(\log(n) - 2)]$, we observe that

$$\begin{aligned}
&\frac{2N + p\log(n)}{n}(\hat{\sigma}^2(\lambda) - \sigma_0^2) \\
&= \frac{2N + p\log(n)}{n}\Big(\frac{\varepsilon'(I - 2H)\varepsilon}{n} - \sigma_0^2\Big) + \frac{2N + p\log(n)}{n}\Big[U_n(\lambda) + \frac{p(\log(n) - 2)\sigma_0^2}{n}\Big] \\
&\quad - \frac{(2N + p\log(n))p(\log(n) - 2)\sigma_0^2}{n^2} + \frac{2[2N + p\log(n)]}{n^2}\varepsilon'(I - H)R_n^* \\
&= \frac{2N + p\log(n)}{n}\Big(\frac{\varepsilon'(I - 2H)\varepsilon}{n} - \sigma_0^2\Big) + \frac{2N + p\log(n)}{n}\Big[U_n(\lambda) + \frac{p(\log(n) - 2)\sigma_0^2}{n}\Big] \\
&\quad + o_P(\bar{L}_n(\lambda)).
\end{aligned}$$

This fact, (3.4), and (3.10) imply (3.5). Similarly, the simple facts $\sup_{\lambda \in \Lambda} \hat{\sigma}^2(\lambda) = O_p(1)$ and $n\bar{L}_n(\lambda) \geq \sigma_0^2[N + p(\log(n) - 2)]$ yield (3.8).

From now on, we will show (3.6) and (3.7). Given any $\delta > 0$, from Chebychev's Inequality and Lemma 1, there exists a constant $D_2 > 0$ such that

$$\begin{aligned}
&P\Big\{\sup_{\lambda \in \Lambda} \frac{2}{n}\big|\mathrm{tr}(H)\sigma_0^2 - \varepsilon' H\varepsilon\big|/\overline{L}_n(\lambda) > \delta\Big\} \\
&\leq P\Big\{\sup_{\lambda \in \Lambda} \frac{2}{n}|\mathrm{tr}(H)\sigma_0^2 - \varepsilon' H\varepsilon|/L_n(\lambda) > \delta\Big\} \\
&\leq \sum_{\lambda \in \Lambda} \frac{2^6 E_0|\mathrm{tr}(H)\sigma_0^2 - \varepsilon' H\varepsilon|^6}{(nL_n(\lambda)\delta)^6} \leq \sum_{\lambda \in \Lambda} \frac{D_2(\mathrm{tr}(H^2))^3}{(nL_n(\lambda)\delta)^6} \\
&\leq D_2 \sum_{\lambda \in \Lambda} \frac{1}{(nL_n(\lambda))^3\delta^6}.
\end{aligned} \tag{3.11}$$

Let D_3 be the cardinal number of Λ_1. Then, $D_3 < \infty$. Since $\sum_{\lambda_2 \in \Lambda_2} \frac{1}{N^3(\lambda_2)} < \infty$, for any given $\zeta > 0$, there is a subset Λ_2^* of Λ_2, such that it contains only a finite number of knot sets and $\sum_{\lambda_2 \in \Lambda_2 \setminus \Lambda_2^*} \frac{D_3}{\sigma_0^3 N^3(\lambda_2)} < \zeta/2$. Consequently,

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

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, $\sup_{\lambda \in \Lambda} D_3 D \frac{1}{(nL_n(\lambda))^3\delta^6} \to 0$ with n. These results and (3.11) imply (3.7).

For (3.6), we can find a constant $D_4 > 0$ such that

$$n^{-6}E_0|\varepsilon'(I - H)R_n^*|^6 \leq D_4(R_n^*(I - H)R_n^*)^3 n^{-6} \leq D_4\Big(\frac{\overline{L}_n(\lambda)}{n}\Big)^3.$$

From the above inequality and an argument similar to that used in the proof of (3.7), we can easily obtain (3.6).

Next, we will show the second part of Theorem 2. Since $\lambda_1 \in \Lambda_1^0$,

$$\boldsymbol{X}_0\beta_0 + g_0 = \boldsymbol{X}(\lambda)\tilde{\beta}(\lambda) + \Pi\theta_0^* + R_n$$

and

$$\boldsymbol{X}(\lambda)\hat{\beta}(\lambda) + \Pi\hat{\theta}(\lambda) = H\boldsymbol{Y} = \boldsymbol{X}(\lambda)\tilde{\beta}(\lambda) + \Pi\theta_0^* + HR_n + H\varepsilon,$$

we have

$$nL_n(\lambda) = \mathrm{tr}(H)\sigma_0^2 + R_n'(I-H)R_n.$$

Let $\overline{\lambda}_n(\lambda) = (\overline{\lambda}_1, \overline{\lambda}_2)$ be a sequence of models such that $\overline{\lambda}_1 \in \Lambda_1^0$ and $N(\overline{\lambda}_2) = k_n$. Then, from Assumption 2 and (2.1) we have

$$\begin{aligned} 0 \le \overline{L}_n(\hat{\lambda}) \le \overline{L}_n(\overline{\lambda}_n) &= \frac{(k_n+p)\sigma_0^2}{n} + \frac{R_n'(I-H)R_n}{n} + \frac{p(\log(n)-2)\sigma_0^2}{n} \\ &\le \frac{(k_n+p_{full})\sigma_0^2}{n} + \frac{p_{full}(\log(n)-2)\sigma_0^2}{n} + O(k_n^{-2C}) \to 0 \end{aligned}$$

as $n \to \infty$, where p_{full} is the number of columns of $\boldsymbol{X}(\lambda)$ under the full model (i.e., the model includes all explanatory variables of the linear part). This fact, the first part of Theorem 2, and $L_n(\lambda) \le \bar{L}_n(\lambda)$ for $n \ge \exp(2)$ imply the second part of Theorem 2.

4 Appendices

Appendix A

Proof of Lemma 2 Let $\xi_n = (\xi_{1n}, \xi_{2n}, \cdots, \xi_{nn})' = \frac{(I-H(\lambda_1,\lambda_2))\mu}{c_n\sqrt{n}}$, where $c_n = (n^{-1}\mu'(I - H(\lambda_1,\lambda_2))\mu)^{1/2}$. It is easy to see that $\frac{1}{\log n}\varepsilon'(I-H(\lambda_1,\lambda_2))\mu = \frac{c_n}{\log n}\sum_{i=1}^n \varepsilon_i\xi_{in}$ and $\xi'\xi = 1$. From Assumption 1, there exists a constant $c > 0$ such that $c_n \le c$ for all n. Let $d = \delta \log n/(8c\varphi)$ for any given $\delta > 0$. Since ε_1 is sub-Gaussian, it can be seen that

$$\begin{aligned} &P\left\{\max_{\lambda_1\in\Lambda_1,\lambda_2\in\Lambda_2}\left|\frac{c_n}{\log n}\sum_{i=1}^n \varepsilon_i\xi_{in}\right| > \delta\right\} \\ &\le 2^p\#(\Lambda_2)\max_{\lambda_1\in\Lambda 1,\lambda_2\in\Lambda_2} P\left\{\left|d\sum_{i=1}^n \varepsilon_i\xi_{in}\right| > \frac{\delta\log nd}{c}\right\} \\ &= 2^p\#(\Lambda_2)P\left\{\left|d\sum_{i=1}^n \varepsilon_i\xi_{in}\right| > 8\varphi d^2\right\} \\ &\le 2^p\#(\Lambda_2)\exp(-8\varphi d^2)E\Big(\exp\Big(\Big|d\sum_{i=1}^n \varepsilon_i\xi_{in}\Big|\Big)\Big) \le 2^{p+1}\#(\Lambda_2)\exp(-4\varphi d^2) \\ &\le 2^{p+1}\exp\Big(\frac{-4\delta^2(\log n)^2}{64c_1^2\varphi} + \log n\Big) \longrightarrow 0 \quad \text{as } n\to\infty. \end{aligned}$$

Appendix B

Proof of Lemma 3 For any given $\delta > 0$, from Chebychev's Inequality and Assumption 2 we have

$$\begin{aligned} &P\left\{\sup_{\lambda_1\in\Lambda_1}\sup_{\lambda_2\in\Lambda_2}\varepsilon'H(\lambda_1,\lambda_2)\varepsilon > \delta n\right\} \le 2^p\#(\Lambda_2)\max_{\lambda_1\in\Lambda_1,\lambda_2\in\Lambda_2} P\{(\varepsilon'H(\lambda_1,\lambda_2))\varepsilon > \delta n\} \\ &\le 2^p\#(\Lambda_2)\max_{\lambda_1\in\Lambda_1,\lambda_2\in\Lambda_2}\frac{E(\varepsilon'H(\lambda_1,\lambda_2)\varepsilon)}{\delta n} \le 2^p\#(\Lambda_2)\max_{\lambda_1\in\Lambda_1}\frac{k_n+m+p(\lambda_1)}{\delta n}\sigma_0^2 \to 0 \end{aligned}$$

as n tends to infinity. Therefore, Lemma 3 holds.

Appendix C

Proof of Lemma 4 We first prove (ii). From the definition of $\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1))$, for any given

$\lambda_1 \in \Lambda_1^0$, we have

$$\begin{aligned}\hat{\sigma}^2(\lambda_1, \hat{\lambda}_2(\lambda_1)) &= \frac{1}{n}\boldsymbol{Y}'[I - H(\lambda_1, \hat{\lambda}_2(\lambda_1))]\boldsymbol{Y} \\ &= n^{-1}\varepsilon'[I - H(\lambda_1, \hat{\lambda}_2(\lambda_1))]\varepsilon + \frac{2}{n}\varepsilon'[I - H(\lambda_1, \hat{\lambda}_2(\lambda_1))]\mu \\ &\quad + n^{-1}\mu'[I - H(\lambda_1, \hat{\lambda}_2(\lambda_1))]\mu.\end{aligned}$$

Thus, (ii) is the direct consequence of (i), Lemma 2 and Lemma 3.

Since $\lambda_1 \in \Lambda_1^0$, $H(\lambda)\boldsymbol{X}(\lambda) = 0$, $H(\lambda)\Pi = 0$ and

$$\boldsymbol{X}(\lambda)\hat{\beta}(\lambda) + \Pi\hat{\theta}(\lambda) = H(\lambda)\boldsymbol{Y} = H(\lambda)(\boldsymbol{X}(\lambda)\beta^* + \Pi\theta(\lambda) + R_n)$$

for some vector β^*, (i) follows from (2.1).

References

[1] R. J. Shiller, Smoothness priors and nonlinear regression, *J. Amer. Statist. Assoc.*, 1984, **79**: 609–615.

[2] R. F. Engle, C. W. J. Granger, J. Rice, & A. Weiss, Semiparametric estimates of the relation between weather and electricity sales, *J. Amer. Statist. Assoc.*, 1986, **81**: 310–320.

[3] G. Wahba, Partial spline models for the semiparametric estimation of functions of several variables, *Statistical Analysis of Time Series: Proceedings of the Japan U.S. Joint Seminar*, Tokyo, 1984, pp.319–329.

[4] N. E. Heckman, Spline smoothing in a partly linear model, *J. Roy. Statist. Soc. B*, 1986, **48**: 244–248.

[5] J. Rice, Convergence rates for partially splined models, *Statist. Probab. Lett.*, 1986, **4**: 203–208.

[6] H. Chen, Convergence rates for parametric components in a partly linear model, *Ann. Statist.*, 1988, **16**: 136–146.

[7] P. Speckman, Kernel smoothing in partly linear models, *J. R. Statist. Soc. B*, 1988, **50**: 413–436.

[8] X. He, & P. D. Shi, Bivariate tensor-product B-splines in a partly linear model, *Journal of Multivariate Analysis*, 1996, **58**: 162–181.

[9] L. Guttman, A new approach to factor analysis: the radex, In Lazarsfeld, ed., *Mathematical Thinking in the Social Sciences*, Free Press, Glencoe, 1954.

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

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

[12] R. L. Eubank, *Spline Smoothing and Nonparametric Regression,* New York: Marcel Dekker, 1988.

[13] G. Schwarz, Estimating the dimension of a model, *Ann. Statist.*, 1978, **6**: 461–464.

[14] J. Shao, An asymptotic theory for linear model selection, *Statistica Sinica*, 1997, **7**: 221–242.

[15] P. D. Shi, & C. L. Tsai, Semiparametric regression model selections, *Journal of Statistical Planning and Inference*, 1999, **77**: 119–139.

[16] R. Shibata, An optimal selection of regression variables, *Biometrika*, 1981, **68**: 45–54.

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

[18] L. L. Schumaker, *Spline Functions,* Wiley, New York, 1981.

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

[20] X. D. Zheng, & W. Y. Loh, Consistency variable selection in linear models, *J. Amer. Statist. Assoc.*, 1995, **90**: 151–156.

[21] C. R. Rao, & J. Kleffe, *Estimation of Variance Components and Applications*, Amsterdam: North-Holland, 1988.