Statistica Sinica **15**(2005), 1033-1048

TWO-STEP CROSS-VALIDATION SELECTION METHOD FOR PARTIALLY LINEAR MODELS

Panagiotis Avramidis

London School of Economics

Abstract: In this article, we deal with the selection of the linear component and the nonparametric component in a partially linear model. Our method combines the leave-one-out cross-validation for the nonparametric component and the leave-n_v-out Monte Carlo Cross Validation (MCCV) for the parametric component. Under some mild regularity conditions, we show that the estimators are consistent. Although the results are presented for models involving the mean regression function, we extend them to include the variance function while bandwidth selection is discussed separately. Numerical examples demonstrate the gain in efficiency using the proposed selection procedure compared to a fully nonparametric procedure.

Key words and phrases: Bandwidth, cross validation, kernel, partially linear model, subset selection, Monte Carlo.

1. Introduction

Linear regressors selection has been the subject of extensive research. The Akaike information criterion (Akaike (1974), Shibata (1981)), cross-validation (CV) (Stone (1974), Shao (1993)) and the generalized-cross validation (GCV) (Craven and Wahba (1979)), are among the most frequently used procedures. See also Wei (1992) for an overall discussion on the problem of regressors selection in a linear model and Shao (1997) for an extensive study on the asymptotic behavior of the generalized-AIC, the CV and GCV criteria.

In parallel with work on parametric fitting, there have been substantial developments in variable selection in the context of nonparametric regression. These include cross validation CV (Cheng and Tong (1992, 1993) and Zhang (1991)) and multifold cross validation (Zhang (1993)). Vieu (1994) examines cross validation with respect to an error measuring function. Tjøstheim and Auestad (1994) proposed an analog of the Akaike criterion, the Final Prediction Error (FPE) criterion, while Yang (1999) introduced a penalty term to an Akaike-based criterion. See also Bickel and Zhang (1992) for an extension of CV and FPE to nonparametric selection of categorical covariates. Yao and Tong (1994) establish asymptotic results for the cross-validation criterion based on kernel estimation. Their approach includes time series while Tschernig and Yang (2000) extend these

results to include nonparametric autoregressive models with heteroskedasticity. See also Vieu (1995) for a similar idea.

Recently, a new class of models combining linear and nonparametric settings, has been introduced. The advantage of this new class, partially linear models, is that the parametric part can be estimated efficiently while the flexibility of the nonparametric model is retained. Härdle, Liang and Gao (2000) propose an estimation procedure for both the parametric and nonparametric component and establish asymptotic properties. Moreover, they address the issue of variable selection for the nonparametric component. Gao and Tong (2004) propose a simultaneous selection procedure for the parametric and nonparametric components.

In this paper, we study a two step selection procedure. In particular, we first use the leave-one-out cross validation to select the nonparametric regressors, while at the second step the linear regressors are selected using the leave-n_v-out cross validation. Under some mild conditions and assumption (A1) on the existence of the true model, we prove that the proposed model selection procedure is consistent. Moreover, we extend the results derived for the mean regression function to the case of the variance function. Simulation examples are presented to illustrate the theoretical findings. It appears that the proposed selection procedure outperforms a fully nonparametric selection procedure emphasizing the need of a more flexible method when dealing with partially linear models.

2. Model and Selection of the Nonparametric Component

Let $(Y_t, \mathbf{W}_t)$ be a strictly stationary process with a scalar Y_t and a vector of predictors $\mathbf{W}_t$. In a partially linear model the regression function is of the form $E(Y_t|\mathbf{W}_t) = \mathbf{X}_t^T\theta + g(\mathbf{Z}_t)$, where $\mathbf{X}_t = (\mathbf{W}_{t,i})^T$, $i \in \mathcal{P}$, $\mathbf{Z}_t = (\mathbf{W}_{t,i})^T$, $i \in \mathcal{Q}$, the linear and the nonlinear regressors. In the context of time series analysis, $\mathbf{X}_t$ and $\mathbf{Z}_t$ may contain some lagged variables of Y_t. We introduce the enlarged regression model

$$Y_t = E(Y_t|\mathbf{X}_t, \mathbf{Z}_t) + \epsilon_t = \mathbf{X}_t^T\theta + g(\mathbf{Z}_t) + \epsilon_t, \tag{1}$$

where $g : \mathbb{R}^Q \to \mathbb{R}$ is an unknown function, $\theta = (\theta_1, \ldots, \theta_P)^T$ is a vector of parameters and $\epsilon_t = Y_t - E(Y_t|\mathbf{X}_t, \mathbf{Z}_t)$ is an error term. It is easy to see that $E(\epsilon_t|\mathbf{X}_t, \mathbf{Z}_t) = 0$. If $U_t = Y_t - \mathbf{X}_t^T\theta$, then (1) yields $E(U_t|\mathbf{Z}_t) = g(\mathbf{Z}_t)$. We introduce conditions to ensure the existence of a reduced true model. Define the variance function from $\mathbb{R}^k \to \mathbb{R}$

$$\sigma^2(A) = E[U_t - E(U_t|\mathbf{Z}_t^A)]^2 \tag{2}$$

with $\mathbf{Z}_t^A = (Z_{t,i} : i \in A)^T$ for $A = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, Q\}$.

Definition 1. If there is a subset $A_0 \equiv \{1, \ldots, q\}$ of $\{1, \ldots, Q\}$ with $|A_0| = q \leq Q$ for which

(a) $\sigma^2(A_0) = \sigma^2(1, \ldots, Q)$, and
(b) for any $A = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, Q\}$ with $k \leq q$ and $\{i_1, \ldots, i_k\} \neq \{1, \ldots, q\}$, it holds that $\sigma^2(A) > \sigma^2(A_0)$,

then the set $\mathbf{Z}_t^{A_0} = \{Z_{t,1}, \ldots, Z_{t,q}\}$ is called the optimal regression subset of the nonparametric component in (1).

Further, let $V_t = Y_t - g(\mathbf{Z}_t^{A_0})$ and define the variance function

$$\bar{\sigma}^2(M) = E[V_t - E(V_t|\mathbf{X}_t^M)]^2 \tag{3}$$

for any $M = \{j_1, \ldots, j_r\} \subset \{1, \ldots, P\}$.

Definition 2. If there is a subset $M_0 \equiv \{1, \ldots, p\}$ of $\{1, \ldots, P\}$ with $|M_0| = p \leq P$ for which
(a) $\bar{\sigma}^2(M_0) = \bar{\sigma}^2(1, \ldots, P)$, and
(b) for any $M = \{j_1, \ldots, j_r\} \subseteq \{1, \ldots, P\}$ with $r \leq p$ and $\{j_1, \ldots, j_r\} \neq \{1, \ldots, p\}$, it holds that $\bar{\sigma}^2(M) > \bar{\sigma}^2(M_0)$,

then the set $\mathbf{X}_t^{M_0} = \{X_{t,1}, \ldots, X_{t,p}\}$ is called the optimal regression subset of the parametric component in (1).

At this point, we are ready to impose the necessary condition that will ensure the existence and identifiability of the true model.

A1 We assume that the true model is the model with the optimal nonparametric $\mathbf{Z}_t^{A_0} = \{Z_{t,1}, \ldots, Z_{t,q}\}$ and parametric $\mathbf{X}_t^{M_0} = \{X_{t,1}, \ldots, X_{t,p}\}$ components. Further, we assume that there is $C > 0$ a constant such that $\min_{j \in \{1,\ldots,Q\} - A_0} \inf_{\alpha,\beta} E(E(g(\mathbf{Z}_t^{A_0})|Z_{t,j}) - \alpha - \beta Z_{t,j})^2 > C$.

It is easy to see that if A1 holds, $E(U_t|\mathbf{Z}_t) = E(U_t|\mathbf{Z}_t^{A_0})$ almost surely, i.e., the optimal subset contains almost all the information on U_t available from $\mathbf{Z}_t$. Further, from Definition 2, $E(V_t|\mathbf{X}_t) = E(V_t|\mathbf{X}_t^{M_0})$ almost surely, so we conclude that some of the linear predictors are insignificant and should be omitted. Note also that in A1, the nonparametric component of $g(\mathbf{Z}_t^{A_0})$ cannot be explained by any linear term. Chen and Chen (1991) and Gao and Tong (2004) impose similar conditions to ensure the identifiability of the model.

Given the existence of the true model as a reduced form of (1), we try to identify the optimal regressors for both linear and nonlinear components of the regression function. We propose a two-step selection procedure. The first step is the selection of the nonparametric component. We use the leave-one-out cross validation procedure on the residuals, after regressing over the full set of linear regressors, to estimate the optimal subset. Call $\hat{\theta}$ the parameter estimator calculated by regressing Y_t against all linear regressors, $X_{t,1}, \ldots, X_{t,P}$, and let $\hat{U}_t = Y_t - \mathbf{X}_t^T \hat{\theta}$ be the residuals. For any $A = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, Q\}$ let

$\mathbf{Z}_t^A = (Z_{t,i_1}, \dots, Z_{t,i_k})^T$. Take the standard Nadaraya-Watson estimator to be

$$g_n(\mathbf{z}) = \sum_{t=1}^{n} w_{t,A}(\mathbf{z})(Y_t - \mathbf{X}_t^T\theta), \tag{4}$$

with $w_{t,A} : \mathbb{R}^k \to \mathbb{R}$, $w_{t,A}(\mathbf{z}) = K_h(\mathbf{Z}_t^A - \mathbf{z})/\sum_{r=1}^n K_h(\mathbf{Z}_r^A - \mathbf{z})$ the weighting function, and $K_h : \mathbb{R}^k \to \mathbb{R}$ a k-dimensional kernel function. Similar to (4), we set

$$\hat{g}_n(\mathbf{z}) = \sum_{t=1}^{n} w_{t,A}(\mathbf{z})(Y_t - \mathbf{X}_t^T\hat{\theta}), \tag{5}$$

with θ replaced by the estimator $\hat{\theta}$. The leave-one-out estimators are $g_n^{(-s)}(\mathbf{z}) = \sum_{t=1,t\neq s}^n w_{t,A}^{(-s)}(\mathbf{z})(Y_t - \mathbf{X}_t^T\theta)$ and $\hat{g}_n^{(-s)}(\mathbf{z}) = \sum_{t=1,t\neq s}^n w_{t,A}^{(-s)}(\mathbf{z})(Y_t - \mathbf{X}_t^T\hat{\theta})$, with $w_{t,A}^{(-s)}(\mathbf{z}) = K_h(\mathbf{Z}_t^A - \mathbf{z})/\sum_{r=1,r\neq s}^n K_h(\mathbf{Z}_r^A - \mathbf{z})$. Then, for $A = \{i_1, \dots, i_k\} \subset \{1, \dots, Q\}$, the cross validation function is given by

$$CV(A) = \frac{1}{n}\sum_{s=1}^{n}\{\hat{U}_s - \hat{g}_n^{(-s)}(\mathbf{Z}_t^A)\}^2. \tag{6}$$

Definition 3. The estimator for the optimal regression subset of the nonparametric component is

$$\hat{A} = \arg \min_{A=\{i_1,\dots,i_k\}\subseteq\{1,\dots,Q\}} CV(A). \tag{7}$$

Next we state the assumptions and introduce the notation. Let $C > 0$ be a constant that can take different values in different places.

A2 For the least squares estimator $\hat{\theta}$, $E \parallel \hat{\theta} - \theta \parallel^2 = O(n^{-1})$.

A3 The density functions of the random processes $\mathbf{Z}_t$ and $\mathbf{X}_t$, f and p, are Lipschitz functions, and the sets $B_1 = \{\mathbf{z} : f(\mathbf{z}) > 0\}$ and $B_2 = \{\mathbf{z} : p(\mathbf{z}) > 0\}$ are compact subsets of $\mathbb{R}^Q$ and $\mathbb{R}^P$, respectively.

A4 For the strictly stationary process $\{(Y_t, \mathbf{X}_t, \mathbf{Z}_t) : t = 1, 2, \dots\}$ let $\beta(n) = \sup_{k\geq 1} E\{\sup_{B\in\Im_{k+n}^\infty} |P(B|\Im_1^k) - P(B)|\}$ where $\Im_k^n$ the sigma-field generated by $\{(Y_t, \mathbf{X}_t, \mathbf{Z}_t) : k \leq t \leq n\}$. Then $\beta(n) = O(n^{-(2+\delta)/\delta})$ where $0 \leq \delta \leq 2/5$. In addition, there are positive integers m_n and $l_n = [n/(2m_n)]$ such that $\limsup_{n\to\infty}(1 + 6\sqrt{e}\beta(m_n)^{1/(1+l_n)})^{l_n} < \infty$.

A5 For $|A| = k$, $1 \leq k \leq Q$, denote with $K_h(\mathbf{u}) = K(\mathbf{u}/h)$ the kernel function where $K : \mathbb{R}^k \to \mathbb{R}$ is a symmetric density function with bounded support satisfying a Lipschitz condition. Further, for the bandwidth $h = n^{-\lambda(k)}$, it holds that $0 < k\lambda(k) < 1/2$ for $1 \leq k \leq Q$.

A6 For m_n defined in A4, $\limsup_{n\to\infty} l_n n^{-\lambda(k)} < \infty$ for all $1 \leq k \leq Q$.

A7 $E|Y_t|^6 < \infty$, $E||\mathbf{X}_t||^6 < \infty$, $E(Y_t|\mathbf{X}_t, \ldots, \mathbf{X}_1, \mathbf{Z}_t, \ldots, \mathbf{Z}_1) = E(Y_t|\mathbf{X}_t, \mathbf{Z}_t)$ for $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,P})^T$ and $\mathbf{Z}_t = (Z_{t,1}, \ldots, Z_{t,Q})^T$.

A8 It holds that $|g(\mathbf{z}_1) - g(\mathbf{z}_2)| \leq C \parallel \mathbf{z}_1 - \mathbf{z}_2 \parallel^{\gamma}$ where $g(\mathbf{z}) = E(U_t|\mathbf{Z}_t^A = \mathbf{z})$, with $|A| = k$, $1 \leq k \leq Q$, and γ a positive constant.

A9 $(k + \gamma)\lambda(k) > 1/2$ for all $1 \leq k \leq Q$ and for γ in A8.

A10 $k\lambda(k)$ is a strictly increasing function of k.

Remark 1. The above assumptions are not the weakest possible and may be altered at the cost of a lengthier proof. Assumption A1 of existence of the true model is a common assumption in the context of regressor selection, while A2 is a standard result of the linear regression theory and requires no further explanation. Further, in A3, we follow Yao and Tong (1994) who point out that in practice any reasonably stationary data could be considered as bounded and thus we assume a density with bounded support. This is a technical assumption which facilitates the proof and can be relaxed by introducing a weight function in the definition of the cross-validation function. Assumption A4 implies that we are dealing with absolutely regular processes, while the assumption on the rate of $\beta(n)$ and A6 allow us to use the results of Yoshihara (1976) and Roussas (1988). Assumption A5, A7−A8 need no further explanation, A9 is standard in nonparametric order determination, while A10 is essential in the proof of convergence in probability of the CV-estimator.

We now state the main theorem on the consistency of the proposed leave-one-out cross validation criterion.

Theorem 1. *Under assumptions* A1−A10, $\lim_{n\to\infty} P(\hat{A} = A_0) = 1$.

The proof of Theorem 1 is postponed to Appendix A.

Remark 2. The main terms in the decomposition of the CV-function are those derived in Yao and Tong (1994) for a fully nonparametric model.

3. Selection of Parametric Component

The second step of the proposed procedure is the selection of the parametric regressors. For $M \subset \{1, \ldots, P\}$ we write

$$Y_t = (\mathbf{X}_t^M)^T \theta_M + g(\mathbf{Z}_t^{A_0}) + \epsilon_{t,M}, \tag{8}$$

where $\mathbf{X}_t^M = (X_{t,i} : i \in M)^T$ and $\epsilon_{t,M} = Y_t - E(Y_t|\mathbf{X}_t^M, \mathbf{Z}_t^{A_0})$. We denote this model as $\mathcal{M}_M$. To this end, we classify the models $\mathcal{M}_M$ into two groups:

Category I: at least one nonzero component of θ is not in θ_M;

Category II: θ_M contains all the nonzero components of θ.

In the first category, we have models $\mathcal{M}_M$ that are incorrect in the sense that they do not include all the significant regressors. Models in the second category

include all the significant regressors but may include regressors unrelated to the response variable. The optimal model M_0 is the one in category II with the smallest dimension. To estimate $\mathcal{M}_M$, we substitute $g(\cdot)$ with the nonparametric estimator $g_n(\mathbf{Z}_t^{A_0}) = \sum_{s=1}^n w_{s,A_0}(\mathbf{Z}_t^q)(Y_s - \mathbf{X}_s^T\theta)$, and we find that the least squares estimator of θ_M is

$$\hat{\theta}_M = (\tilde{\mathbf{X}}_M^T\tilde{\mathbf{X}}_M)^{-1}\tilde{\mathbf{X}}_M^T\tilde{\mathbf{Y}}, \tag{9}$$

where $\tilde{Y}_t = Y_t - \sum_{s=1}^n w_{s,A_0}(\mathbf{Z}_t^{A_0})Y_s$ and $\tilde{\mathbf{X}}_M = (\tilde{\mathbf{X}}_1^M, \ldots, \tilde{\mathbf{X}}_n^M)^T$ with $\tilde{\mathbf{X}}_t^M = \mathbf{X}_t^M - \sum_{s=1}^n w_{s,A_0}(\mathbf{Z}_t^{A_0})\mathbf{X}_s^M$. Then the mean square prediction error for model $\mathcal{M}_M$ is given by

$$MSE_n(M) = \frac{1}{n}\tilde{\epsilon}^T\tilde{\epsilon} - \frac{1}{n}\tilde{\epsilon}^T\mathbf{P}_M\tilde{\epsilon} + \frac{1}{n}\theta^T\tilde{\mathbf{X}}^T\mathbf{H}_M\tilde{\mathbf{X}}\theta + \frac{2}{n}\tilde{\epsilon}^T\mathbf{H}_M\tilde{\mathbf{X}}\theta, \tag{10}$$

where $\tilde{\epsilon} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n)$ with $\tilde{\epsilon}_t = \tilde{Y}_t - \tilde{\mathbf{X}}_t^T\theta$, $\mathbf{P}_M = \tilde{\mathbf{X}}_M(\tilde{\mathbf{X}}_M^T\tilde{\mathbf{X}}_M)^{-1}\tilde{\mathbf{X}}_M^T$, and $\mathbf{H}_M = I_n - \mathbf{P}_M$. Note that by definition, $\tilde{\epsilon}_t = \tilde{Y}_t - \tilde{\mathbf{X}}_t^T\theta = \epsilon_t - \sum_{s=1}^n w_{s,A_0}(\mathbf{Z}_t^{A_0})\epsilon_s$, where $\epsilon_t = Y_t - \mathbf{X}_t^T\theta$ and, using Proposition 1 below with $p = 1 - q\lambda(q)$ where $q = |A_0|$ and $r = 6$ (see A5, A7), we conclude that $\max_t |\sum_{s=1}^n w_{s,A_0}(\mathbf{Z}_t^{A_0})\epsilon_s| = o_p(n^{q\lambda(q)-1/2}) = o_p(1)$. Consequently, the nonparametric term does not affect the rate of convergence of the mean square prediction error under the assumptions imposed. Surprising as it may look, a similar conclusion was reached by Speckman (1988), who noticed that for a certain choice of bandwidth the parametric estimator $\hat{\theta}$ remains a $\sqrt{n}$-consistent estimator, see also Härdle, Liang and Gao (2000). It follows from (10) that the conditional expected mean square error is given by

$$EMSE_n(M) = \sigma_{\tilde{\epsilon}}^2 - \frac{m}{n}\sigma_{\tilde{\epsilon}}^2 + \Omega_{n,M} \text{ a.s.}, \tag{11}$$

where $\sigma_{\tilde{\epsilon}}^2 = n^{-1}E(\tilde{\epsilon}^T\tilde{\epsilon})$ and $\Omega_{n,M} = n^{-1}\theta^T\tilde{\mathbf{X}}^T\mathbf{H}_M\tilde{\mathbf{X}}\theta$. For every $M \subset \{1, \ldots, P\}$ when $\mathcal{M}_M$ is in category II, it follows that

$$MSE_n(M) = \frac{1}{n}\tilde{\epsilon}^T\tilde{\epsilon} - \frac{1}{n}\tilde{\epsilon}^T\mathbf{P}_M\tilde{\epsilon} + \frac{2}{n}\tilde{\epsilon}^T\mathbf{H}_M\tilde{\mathbf{X}}\theta + o_p(1)$$

and $EMSE_n(M) = (1 - m/n)\sigma_{\tilde{\epsilon}}^2$ (the latter from the fact that in category II $\tilde{\mathbf{X}}_M\theta_M = \tilde{\mathbf{X}}\theta$). Now, assume that

B1 For every $M \subset \{1, \ldots, P\}$,
- (i) $E(\tilde{\mathbf{X}}_M^T\tilde{\mathbf{X}}_M)$ is a positive definite matrix of order m;
- (ii) if $\mathcal{M}_M$ in category I, $\liminf_{n\to\infty} \Omega_{n,M} > 0$ in probability.

Remark 3. Assumption (i) is necessary for the consistency of $\hat{\theta}_M$ (see Härdle, Liang and Gao (2000)). Assumption (ii) is an identifiability condition which is a

very minimal argument for asymptotic analysis. Gao and Tong (2004) show that assumption B1(ii) can be replaced by

$$\liminf_{n\to\infty} \frac{1}{n}(\mathbf{u}\theta)^T(I_n - \mathbf{u}_M(\mathbf{u}_M^T\mathbf{u}_M)^{-1}\mathbf{u}_M^T)(\mathbf{u}\theta) > 0,$$

where $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)^T$, $\mathbf{u}_M = (\mathbf{u}_{1,M}, \ldots, \mathbf{u}_{n,M})^T$, $\mathbf{u}_t = \mathbf{X}_t - E(\mathbf{X}_t|\mathbf{Z}_t^{A_0}) = (u_{t,1}, \ldots, u_{t,P})^T$ and $\mathbf{u}_{t,M} = \mathbf{X}_t^M - E(\mathbf{X}_t^M|\mathbf{Z}_t^{A_0}) = (u_{t,j} : j \in M)^T$. This is an extension of the partial linear context of the identifiability condition 2.5 in Shao (1993).

The proposed estimator is based on the leave-n_v-out cross validation function. We split the data into two parts: $\{(\tilde{Y}_t, \tilde{\mathbf{X}}_t) : t \in N\}$ and $\{(\tilde{Y}_t, \tilde{\mathbf{X}}_t) : t \in N^c\}$, where $N \subseteq \{1, \ldots, n\}$ and N^c is its complement. Hence, if we call n_v and n_c the size of N and N^c, respectively, then $n_v + n_c = n$. Suppose that the model $\mathcal{M}_M$ is fitted using the subsample N^c, called the construction data, while the prediction error is calculated using the subsample N, called the validation data. Then the leave-n_v-out cross validation function is defined by $CV(M, n_v) = 1/n_v \parallel \tilde{Y}_N - \tilde{\mathbf{X}}_{N,M}\hat{\theta}_{N^c,M} \parallel^2$. It is understood that the simplest case would be the leave-one-out cross validation. However, it has been shown that the leave-one-out cross validation criterion yields asymptotically inconsistent estimators, see Shao (1993). On the other hand, for n large there are too many possible subsamples and this is computationally inconvenient.

A good compromise is to use the Monte Carlo-CV(n_v). We randomly draw a collection $\mathcal{B}$ of b subsets of $\{1, \ldots, n\}$, each one with size n_v, and we choose the model that minimizes

$$MCCV(M, n_v) = \frac{1}{bn_v} \sum_{N\in\mathcal{B}} \parallel \tilde{Y}_N - \tilde{\mathbf{X}}_{N,M}\hat{\theta}_{N^c,M} \parallel^2 . \tag{12}$$

Definition 4. The estimator for the optimal regression subset of the linear component is

$$\hat{M} = \arg \min_{M\subset\{1,\ldots,P\}} MCCV(M, n_v). \tag{13}$$

Suppose that

B2 As $n \to \infty$, $n_v/n \to 1$, $n_c = n - n_v \to \infty$, and $n^2/n_c^2 b \to 0$.

Theorem 2. *Let* A1−A10 *of Theorem* 1 *and* B1−B2 *hold.*

(1) *If $\mathcal{M}_M$ in category I, then there exists $R_n \geq 0$ such that*

$$MCCV(M, n_v) = \frac{1}{b} \sum_{N\in\mathcal{B}} \tilde{\epsilon}_N^T\tilde{\epsilon}_N + \Omega_{n,M} + R_n + o_p(1),$$

where $\tilde{\epsilon}_N = \tilde{Y}_N - \tilde{\mathbf{X}}_N\theta$.

(2) *If $\mathcal{M}_M$ in category II, then*

$$MCCV(M, n_v) = \frac{1}{b}\sum_{N\in\mathcal{B}} \tilde{\epsilon}_N^T \tilde{\epsilon}_N + \frac{m}{n_c}\sigma_{\tilde{\epsilon}}^2 + o_p(n_c^{-1}).$$

(3) $\lim_{n\to\infty} P(\hat{\mathcal{M}}_M = \mathcal{M}_{M_0}) = 1$..

The proof of Theorem 2 is based on the following proposition, an extension of the results of Lemma A.3, Härdle, Liang and Gao (2000), for α-mixing processes. Proofs are postponed to Appendix A.

Proposition 1. *Let $\{X_i\}$ be a zero mean, strictly stationary, α-mixing random sequence with $t^6\alpha(t) \to 0$. Suppose $\sup_{1\le i\le n} E|X_i|^r < C < \infty$ for $r > 2$, and let $\alpha_{i,j}$, $i, j = 1, \ldots, n$, be a sequence of positive numbers such that $\sup_{1\le i,j\le n} |\alpha_{i,j}| \le n^{-p}$ for some $0 < p < 1$. Then $\max_{1\le j\le n} |\sum_{i=1}^n \alpha_{i,j} X_i| = o_p(n^{-p+1/3+1/r} \log n)$.*

4. Bandwidth Selection

The cross validation function defined above depends directly on the bandwidth. Let h_0 be the minimizer of the mean average square error: $MASE(h) = 1/n \sum_{t=1}^n E(\mathbf{X}_t^{M_0 T}\hat{\theta}_{M_0} - \mathbf{X}_t^T\theta + \hat{g}(\mathbf{Z}_t^{A_0}) - g(\mathbf{Z}_t^{A_0}))^2$. Then using standard results from the partial linear theory, it is easy to see that $h_0 = Cn^{-1/(4+q)}$. Further, if $(\hat{A}, \hat{h}) = \arg\min_{A\subseteq\{1,\ldots,Q\}} \min_{h\in H_n(k)} CV(A; h)$ then $\hat{h}/h_0 \xrightarrow{P} 1$ as $n \to \infty$, where $H_n(k) = [A_k n^{-1/(4+k)-c_k}, B_k n^{-1/(4+k)+c_k}]$ with constants $A_k, B_k > 0$, $0 < c_k < 1/(8 + 2k)$ (see Gao and Tong, (2004)). In other words, the CV function not only identifies the correct dimensionality of the nonparametric function but it automatically adjusts the bandwidth to have the optimal rate $h \sim n^{-1/(4+q)}$.

To prove consistency of the CV-estimator we assumed a bandwidth of rate $n^{-\lambda(k)}$ with $k\lambda(k) < 1/2$. For a bandwidth of this order, the nonparametric component does not affect the rate of convergence of the parametric estimator. The above condition may look strong, but in practice the rate of the bandwidth is as important as its constant, especially for large n. The numerical examples given in Section 6, with the bandwidth chosen as the minimizer of the CV-function, support this remark. Using a similar argument, Yao and Tong (1994) suggested that the more relevant data-driven bandwidths do not depart in principle from the bandwidth assumptions. They also allowed some minor modifications to ensure, for instance, the monotonicity of $k\lambda(k)$ when necessary. In practice, the cross validation function is calculated over a certain range of bandwidths, choosing for bandwidth estimate the value at which the function attains the minimum. A reasonable choice of the bandwidth range, suggested by Fan, Yao and Cai (2003), is $h = \sigma\, C\, 1.2^r$, for $r = 0, \ldots, 15$, where σ is an estimator of the standard deviation of the error term and C a constant depending on the selected

kernel (for example, $C = 0.2$ for the Epanichnikov kernel and $C = 1.2$ for the Gaussian kernel).

5. Further Discussion

The model considered so far is a mean regression model. However, with some modifications to the assumptions, the results can be extended to include modelling of the variance function. In particular, a second order partial linear model can be defined as

$$Y_t = \sigma_t \epsilon_t, \;\; \sigma_t^2 = \mathbf{X}_t^T \theta + g(\mathbf{Z}_t), \tag{14}$$

with Y_t scalar, $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,P})^T$, $X_{t,j} \geq 0$, and $\mathbf{Z}_t = (Z_{t,1}, \ldots, Z_{t,Q})^T$. Further, $g(\cdot) \geq 0$ and the error process ϵ_t is independent of $\{\mathbf{X}_s, \mathbf{Z}_s, s \leq t\}$, satisfying $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) = 1$. Note that if $X_{t,i} = Y_{t-i}^2$, (14) is a partial linear, in respect to Y_t^2, ARCH model, also called a semiparametric ARCH model. Re-arranging (14) yields $Y_t^2 = \sigma_t^2 \epsilon_t^2 = \sigma_t^2 + \sigma_t^2(\epsilon_t^2 - 1) \equiv \sigma_t^2 + \sigma_t^2 \xi_t$ with $\xi_t = \epsilon_t^2 - 1$. Obviously, $E(\xi_t) = 0$, so $E(\sigma_t^2 \xi_t | \mathbf{X}_t, \mathbf{Z}_t) = \sigma_t^2 E(\xi_t) = 0$. Hence, it follows that

$$E(Y_t^2 | \mathbf{X}_t, \mathbf{Z}_t) = \sigma_t^2 = \mathbf{X}_t^T \theta + g(\mathbf{Z}_t), \tag{15}$$

which is in the form of (1). Our main concern here is that the error term is heteroskedastic. However, Härdle, Liang and Gao (2000) have already shown that under some assumptions on initial estimates of σ_t^2, the weighted-LS estimator of θ is $\sqrt{n}$-consistent and asymptotically normally distributed. Based on that, we introduce weights in the proposed selection procedure. In particular, we first regress Y_t^2 on all the candidate parametric regressors $X_{t,j}$ and use the weighted leave-one-out CV criterion to find the optimal nonparametric regressors set. Then we apply the weighted leave-n_v-out cross validation to exclude, if necessary, the insignificant parametric regressors. The weights are based on some initial estimates of the variance. Consistency of the optimal subsets estimators will depend on the convergence rate of the initial variance estimates.

6. Numerical Applications

Two simulated examples involving regression models and one of variance modelling are presented. We use the multivariate kernel $K(\mathbf{u}) = \prod_{i=1}^{k} K(u_i)$, where $K(\cdot)$ is the Epanichnikov kernel. The selected kernel satisfies assumption A5 on the kernel function. Bandwidth is found by minimizing the cross validation criterion in grid points $h = 0.2, \; 1.2^a \sigma$ for $a = 1, \ldots, 15$, where σ is the sample standard deviation. It turns out that the resulting bandwidth estimator does not depart significantly from assumptions A5, A9−A10. Further, if we choose $b = n$ and $n_v = n - n_c$ with $n_c = [n^{3/4}]$ the largest integer part of $n^{3/4}$, then B2

holds. The simulations show that, using the above bandwidth estimator, the CV criterion identifies the correct model.

Example 1. We generate a time series data set from the model $Y_t = 0.5Y_{t-1} - 0.35Y_{t-2} - 0.75\exp(-Y_{t-3}^2) + 0.85(1 + Y_{t-4}^2)^{-1} + \epsilon_t$, with ϵ_t following a uniform distribution in $[-1, 1]$. Note that the error distribution has bounded support, hence there is no need for introducing a weighting function in the CV-function. Let the candidate linear components be $M_1 = \{1\}$, $M_2 = \{2\}$ and $M_0 = \{1, 2\}$ the true one. Let $\theta = (\theta_1, \theta_2)^T$, $u_{t,1} = Y_{t-1} - E(Y_{t-1}|Y_{t-3}, Y_{t-4})$ and $u_{t,2} = Y_{t-2} - E(Y_{t-2}|Y_{t-3}, Y_{t-4})$, $u_t = (u_{t,1}, u_{t,2})^T$, $u_{t,M_1} = u_{t,1}$, $u_{t,M_2} = u_{t,2}$, $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{u}_{M_i} = (u_{1,M_i}, \ldots, u_{n,M_i})$, for $i = 1, 2$. It holds that $\liminf_{n\to\infty} n^{-1}(\mathbf{u}\theta)^T(I_n - \mathbf{u}_{M_i}(\mathbf{u}_{M_i}^T\mathbf{u}_{M_i})^{-1}\mathbf{u}_{M_i}^T)(\mathbf{u}\theta) = (\theta_{3-i}\sum_{t=3}^n u_{t,1}^2 \sum_{t=3}^n u_{t,2}^2 - (\sum_{t=3}^n u_{t,1}u_{t,2})^2)/\sum_{t=3}^n u_{t,i}^2 > 0$ with probability one, because $P(u_{t,1} = u_{t,2}) = 0$ see Remark 3 for B1(ii). It is easy to see that A3 is met, while the generated process satisfies A4, A7. Note that A8 holds for $g(y, x) = -0.75e^{-y^2} + 0.85/(1 + x^2)$. Following the remark about bandwidth selection, we conclude that all of the assumptions are satisfied. We first regress Y_t against all Y_{t-j}, for $j = 1, 2, 3, 4$. Then using the residuals $\hat{U}_t$ we calculate the leave-one-out cross validation. The first three columns of Table 1 contain the probabilities of selection for each candidate component calculated from 80 iterations. Apparently, $\{Y_{t-3}, Y_{t-4}\}$ has the highest probability of selection even in a small sample size of $n = 50$ observations, with the omitted combinations having probability zero. Moreover, increasing the sample size yields even higher probability of selection, e.g., for a sample of size $n = 300$, the probability of selecting the correct regressors increased to 0.7625. In the second step, using $\{Y_{t-3}, Y_{t-4}\}$ as the nonparametric component, we calculate the leave-n_v-out CV. The results are presented in Table 2. The parametric component is identified from the MCCV and the probability of selecting the true regressors increases to 0.975 for $n = 300$.

Table 1. Probabilities of selection based on the leave-one-out CV calculated in 80 iterations for Example 1.

	Two-Step CV			Fully Nonparametric		
subset	$n = 50$	$n = 100$	$n = 300$	$n = 50$	$n = 100$	$n = 300$
$\{Y_{t-1}\}$	0.1375	0.1125	0.0875	0.075	0.0375	0.0125
$\{Y_{t-2}\}$	0.025	0.0125	0.0125	0.0625	0.025	0.00
$\{Y_{t-1}, Y_{t-2}\}$	0.1875	0.1375	0.0625	0.2125	0.2625	0.2375
$\{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$	0.075	0.0625	0.0	0.2375	0.25	0.2625
$\{Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}\}$	0.0875	0.05	0.05	0.3	0.375	0.475
$\{Y_{t-2}, Y_{t-3}\}$	0.0125	0.0	0.0	0.0125	0.0	0.0
$\{Y_{t-1}, Y_{t-3}\}$	0.025	0.025	0.025	0.025	0.0125	0.0125
$\{Y_{t-3}, Y_{t-4}\}$	0.45	0.6	0.7625	0.075	0.0375	0.0

Table 2. Probabilities of selection based on the MCCV calculated in 80 iterations with Y_{t-3}, Y_{t-4} for nonparametric regressors for Example 1.

subset	$n = 50$	$n = 100$	$n = 300$
$\{Y_{t-1}\}$	0.2875	0.2375	0.025
$\{Y_{t-2}\}$	0.1375	0.05	0.0
$\{Y_{t-1}, Y_{t-2}\}$	0.575	0.7125	0.975

Furthermore, in the last three columns of Table 1, we present the results using a fully nonparametric cross validation selection procedure. It is clear that the fully parametric selection method fails to distinguish the linear term from the nonparametric component, while the convergence rate appears significantly slower. It seems that the linear term dominates the nonparametric component and this is why $\{Y_{t-1}, Y_{t-2}\}$ and $\{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$ have high probabilities of selection even when the sample size is increased. The latter illustrates the importance of choosing a combined selection method instead of a fully nonparametric one when working with a semiparametric model.

Example 2. We generate data from the model $Y_t = 0.35Y_{t-1} - 0.15Y_{t-2} + 0.5X_t/(1 + X_t^2) + e_t$, $X_t = 0.3X_{t-1} + 0.2X_{t-2} + \epsilon_t$, with $e_t \sim U[-0.25, 0.25]$ and $\epsilon_t \sim U[-0.5, 0.5]$. We proceed by regressing Y_t against the candidate linear regressors $Y_{t-i}, i = 1, 2, 3$, to calculate the residuals U_t. Note here that we include Y_{t-3} as a regressor, although it does not appear in the true model. This is to show that the procedure works even when insignificant regressors are used in the calculation of the residuals. The results of the leave-one-out CV are reported in Table 3. Table 4 contains the results for the MCCV, using X_t the nonparametric

Table 3. Probabilities of selection based on the leave-one-out CV calculated in 80 iterations for Example 2.

Regressors subset	$n = 50$	$n = 120$	$n = 300$
$\{X_t\}$	0.4625	0.575	0.7875
$\{X_{t-1}\}$	0.35	0.225	0.15
$\{X_t, X_{t-1}\}$	0.1875	0.2	0.0625

Table 4. Probabilities of selection based on the MCCV calculated in 80 iterations with X_t for nonparametric regressor for Example 2.

Parametric Regressors subset	$n = 50$	$n = 120$	$n = 300$
$\{Y_{t-1}\}$	0.25	0.225	0.125
$\{Y_{t-2}\}$	0.0625	0.025	0.0
$\{Y_{t-3}\}$	0.025	0.0	0.0
$\{Y_{t-1}, Y_{t-2}\}$	0.4375	0.575	0.8375
$\{Y_{t-1}, Y_{t-3}\}$	0.15	0.1125	0.025
$\{Y_{t-2}, Y_{t-3}\}$	0.0	0.0	0.0
$\{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$	0.075	0.0625	0.0125

regressor. The nonparametric component is identified with probability 0.7875 for $n = 300$, while the MCCV distinguishes the insignificant linear regressors which are excluded from the model.

Table 5. Probabilities of selection calculated in 80 iterations for Example 3.

	Two-step CV			Fully Nonparametric		
subset	$n = 50$	$n = 100$	$n = 300$	$n = 50$	$n = 100$	$n = 300$
$\{Y_{t-1}\}$	0.125	0.1375	0.075	0.1375	0.125	0.1375
$\{Y_{t-2}\}$	0.3375	0.4375	0.65	0.1875	0.2	0.1375
$\{Y_{t-3}\}$	0.0875	0.075	0.0625	0.0625	0.0375	0.0
$\{Y_{t-1}, Y_{t-2}\}$	0.15	0.1125	0.075	0.2625	0.325	0.4125
$\{Y_{t-1}, Y_{t-3}\}$	0.0875	0.0875	0.0625	0.1375	0.0875	0.0875
$\{Y_{t-2}, Y_{t-3}\}$	0.1	0.0625	0.0125	0.0375	0.0375	0.0125
$\{Y_{t-1}, Y_{t-2}, Y_{t-3}\}$	0.1125	0.0875	0.0625	0.175	0.1875	0.2125

Table 6. Probabilities of selection based on the MCCV calculated in 80 iterations with X_t for nonparametric regressor for Example 3.

Parametric Regressors subset	$n = 50$	$n = 100$	$n = 300$
$\{Y_{t-1}\}$	0.5375	0.7125	0.7875
$\{Y_{t-3}\}$	0.3	0.1875	0.1375
$\{Y_{t-1}, Y_{t-3}\}$	0.1625	0.1	0.075

Example 3. The data is a heteroscedastic time series generated from $Y_t = \sigma_t \epsilon_t$, with $\sigma_t^2 = (0.45Y_{t-1}^2 + 1.1 \sin(Y_{t-2}^2) \exp(-0.85Y_{t-2}))_+$, where ϵ_t is the sum of 35 independent random variables each uniformly distributed on $[-0.05, 0.05]$. According to the Central Limit Theorem, the resulting process is well approximated by a normal one when in fact the support of the error density is bounded, $[-1.75, 1.75]$. Note also that the assumption of an error process variance equal to 1 is satisfied. Since we deal with heteroskedastic data, it is suggested that we use the weighted least squares to calculate the residuals U_t. We calculate the predicted values and their standard errors from a simple linear regression of Y_t^2 on Y_{t-j}^2 for $j = 1, 2, 3$. Using the weighted least squares, we calculate the residuals along with their standard deviation. Then U_t are the standardized residuals. This is equivalent to introducing weights in the leave-one-out cross validation function as required by the heteroscedasticity of the error term.

The probabilities of selection for all the possible combinations of nonparametric regressors calculated with 80 iteration are presented in Table 5 while Table 6 contains the probabilities of selection for the linear regressors assuming that Y_{t-2}^2 has been identified and included as a nonlinear regressor. Both the nonparametric and parametric regressors are identified successfully with probability

0.65 for the CV and 0.7875 for MCCV, for sample size $n = 300$. In Table 5, we present the probabilities using a fully nonparametric method. Although the optimal set is correctly identified, the convergence rate is much slower compared to the rate achieved by using a combined procedure indicating the need to employ a more flexible selection procedure that takes into account the linear, in respect to the squared Y_t, component of the underlying model.

Appendix

Lemma 1. *Under* A2−A9,

(a) *for any* $A = \{i_1, \ldots, i_k\}$, $1 \le k \le q = |A_0|$, $CV(A) \xrightarrow{P} \sigma^2(A)$, $\sigma^2(A)$ *as in* (2);

(b) *if for some* $A = \{i_1, \ldots, i_k\}$, $E(U_t|\mathbf{Z}_t^A) = E(U_t|\mathbf{Z}_t)$ *a.s., then*

$$CV(A) = \frac{1}{n}\sum_{s=1}^{n} \epsilon_s^2 + \frac{1}{nh^k} E(\epsilon_t^2 / f(\mathbf{Z}_t^A)) \int K^2(u)du + o_p(n^{-1}h^{-k}).$$

Lemma 2. *If* A2−A9 *hold, then for any* $A = \{i_1, \ldots, i_k\}$, $1 \le k \le q$, *it follows that*

(a) $n^{-1}\sum_{s=1}^{n} \epsilon_s^A \{g(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\} \xrightarrow{P} 0$,

(b) $n^{-1}\sum_{s=1}^{n} \{g(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}^2 \xrightarrow{P} 0$.

Lemma 3. *Suppose* A2−A9 *hold and that, for some* $A = \{i_1, \ldots, i_k\}$,

$$E(Y_t|\mathbf{X}_t, \mathbf{Z}_t^A) = E(Y_t|\mathbf{X}_t, \mathbf{Z}_t) \quad a.s. \tag{16}$$

(a) $n^{-1}\sum_{s=1}^{n} \epsilon_s^A \{g(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\} = o_p(n^{-1}h^{-k})$,

(b) $n^{-1}\sum_{s=1}^{n} \{g(\mathbf{Z}_s^A) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^A)\}^2 = n^{-1}h^{-k}\mu + o_p(n^{-1}h^{-k})$ *with* $\mu = E(\epsilon_t^2 / f(\mathbf{Z}_t^A)) \int K^2(u)du$.

Lemma 1 plays a key role in the proof of Theorem 1 while Lemmas 2 and 3 imply Lemma 1. All proofs here can be found in a technical report (Avramidis (2003)).

Proof of Theorem 1. For any $A = \{i_1, \ldots, i_k\} \subset \{1, \ldots, Q\}$ with $1 \le k \le Q$, if $\sigma^2(A) > \sigma^2(1, \ldots, Q) = \sigma^2(A_0)$, then from Lemma 1(a) it follows that $P(CV(A_0) < CV(A)) \to 1$. Alternatively, if $\sigma^2(A) = \sigma^2(1, \ldots, Q) = \sigma^2(A_0)$, then (16) in Lemma 3 holds. Note also that by definition $|A| = k > q = |A_0|$. Hence, from A10,

$$h^q / h^k = n^{k\lambda(k) - q\lambda(q)} \to \infty \text{ as } n \to \infty. \tag{17}$$

Thus Lemma 1(b), along with (17), yields $P(CV(A)\text{–}CV(A_0)>0)=P(\int K^2(u)du\{(h^q/h^k)E(\epsilon_t^2/f(\mathbf{Z}_t^A))-E(\epsilon_t^2/f(\mathbf{Z}_t^{A_0}))\}+o_p(h^{q-k})>0)\to 1\Rightarrow P(\hat{A}=A_0)\to 1$ as $n\to\infty$.

Proof of Proposition 1. Define $X_i^{'}=X_iI(|X_i|\le n^{1/r})$ and $X_i^{''}=X_i-X_i^{'}$. Note that $\sup_{1\le i\le n}|\alpha_{i,j}X_i^{'}|<Cn^{-p}n^{1/r}\equiv M$. The exponential-type inequality in Theorem 1.3 in Bosq (1998), with $\varepsilon=n^{-p-2/3+1/r}\log n$ and $q=n^{2/3}$, yields $P(\max_{1\le j\le n}|\sum_{i=1}^n\alpha_{i,j}(X_i^{'}-EX_i^{'})|>n\varepsilon)\le\sum_{j=1}^nP(|\sum_{i=1}^n\alpha_{i,j}(X_i^{'}-E(X_i^{'}))|>n\varepsilon)\le 4n\exp(-(\varepsilon^2q)/(8v^2(q)))+22n^{1+2/3}(1+(4M/\varepsilon)^{1/2}\alpha([n^{1/3}/2])$, where $\alpha(k)$ is the mixing coefficient and $v^2(q)\le 8n^{-2/3}\{\max_{0\le t\le n}E(\alpha_{i,j}(X_i^{'}-EX_i^{'}))^2+8M^2\sum_{k=1}^{[n^{2/3}]+1}\alpha(k)\}+M\varepsilon/2$. Note that $v^2(q)\le CM^2n^{-2/3}+(1/2)M\varepsilon\le Cn^{-2p+2/r-2/3}$. Thus, we have that

$$
\begin{aligned}
&P\Big(\max_{1\le j\le n}|\sum_{i=1}^n\alpha_{i,j}(X_i^{'}-EX_i^{'})|>n\varepsilon\Big)\\
&\le 4n\exp\Big(-\frac{Cn^{-2p-2+4/3+2/r}\log^2 n}{n^{-2p-2/3+2/r}}\Big)+22n^{5/3}(1+\frac{4Cn^{-p+1/r}}{n^{-p-2/3+1/r}\log n})^{1/2}\alpha([\frac{n^{1/3}}{2}])\\
&\le n^{1-\log n}+C_2n^2\alpha(n^{1/3})\to 0
\end{aligned}
$$

since $t^6\alpha(t)\to 0$ when $t\to\infty$. Then the Borel-Cantelli Lemma yields

$$\max_{1\le j\le n}|\sum_{i=1}^n\alpha_{i,j}(X_i^{'}-EX_i^{'})|=o_p(n^{-p+1/r+1/3}\log n). \tag{18}$$

Further note that Hölder's inequality, with m,l such that $1/m\le 1/3$ and $1/m+1/l=1$, yields $\max_{1\le j\le n}|\sum_{i=1}^n\alpha_{i,j}(X_i^{''}-EX_i^{''})|\le\max_{1\le j\le n}(\sum_{i=1}^n|\alpha_{i,j}|^m)^{1/m}(\sum_{i=1}^n|X_i^{''}-EX_i^{''}|^l)^{1/l}\le Cn^{-p+1/m}(\sum_{i=1}^n|X_i^{''}-EX_i^{''}|^l)^{1/l}$. The Ergodic Theorem yields

$$\frac{1}{n}\sum_{i=1}^n\Big(|X_i^{''}-EX_i^{''}|^l-E|X_i^{''}-EX_i^{''}|^l\Big)\overset{a.s.}{\to}0. \tag{19}$$

Note that $X_i^{''}=X_i-X_i^{'}=X_i-X_iI(|X_i|\le n^{1/r})=X_iI(|X_i|\ge n^{1/r})$ and $E|X_i^{''}|^l=E(|X_i|^lI(|X_i|\ge n^{1/r}))\le(E|X_i|^r)^{l/r}(E(I(|X_i|\ge n^{1/r})))^{1-l/r}=(E|X_i|^r)^{l/r}(P(|X_i|\ge n^{1/r}))^{1-l/r}\le(E|X_i|^r)^{l/r}(E|X_i|^r/n)^{1-l/r}$, the latter from the Markov inequality. Thus we have $E|X_i^{''}|^l\le E|X_i|^rn^{l/r-1}$. Hence, from $E|X_i^{''}-EX_i^{''}|^l\le CE|X_i^{''}|^l\le CE|X_i|^rn^{l/r-1}\le Cn^{l/r-1}$, along with (19), we prove that $\sum_{i=1}^n|X_i^{''}-EX_i^{''}|^l\le Cn^{l/r}$ a.s.. Hence

$$\max_{1\le j\le n}|\sum_{i=1}^n\alpha_{i,j}(X_i^{''}-EX_i^{''})|\le Cn^{-p+1/m+1/r}=o_p(n^{-p+1/3+1/r}\log n), \tag{20}$$

and the Lemma follows from (18) and (20).

Proof of Theorem 2. The proof is based on Theorem 2 in Shao (1993). Also, similar results for the partial linear model can be found in Theorem 2.2 of Gao and Tong (2004). Hence we only present an outline of the proof and, in particular, we show that conditions in Theorem 2 Shao (1993) hold. Indeed condition 2.5, 3.12 and 3.22 have been introduced in B1 and B2. Hence, it remains to show that

$$\max_{N\in\mathcal{B}} \parallel \frac{1}{n_v}\sum_{t\in N}\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - \frac{1}{n_c}\sum_{t\in N^c}\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T \parallel = o_p(1), \tag{21}$$

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = O_p(n),\ (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}^{-1} = O_p(n^{-1}),\ \text{ and } \lim_{n\to\infty}\max p_{t,M} = 0 \tag{22}$$

for all M, where $p_{t,M}$ is the tth diagonal element of $\mathbf{P}_M$. Lemma 4 establishes (21) and (22).

Lemma 4. *Under Assumptions* A4−A5 *and* A7,
(a) $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = O_p(n)$, $(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1} = O_p(n^{-1})$,
(b) $\lim_{n\to\infty}\max p_{t,M} = 0$ *for all* $M \subset \{1,\ldots,P\}$, *and*
(c) $\max_{N\in\mathcal{B}} \parallel \frac{1}{n_v}\sum_{t\in N}\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - \frac{1}{n_c}\sum_{t\in N^c}\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T \parallel = o_p(1)$.

Acknowledgement

The author wishes to thank Professor Qiwei Yao for his help. The comments and suggestions from an associate editor and the referee, that helped to substantially improve this paper, are gratefully acknowledged. This research was supported by an ESRC grant.

References

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

Avramidis, P. (2003). Selecting regressors in partially linear models: A technical report. London School of Economics. www.lse.ac.uk/collections/statistics/documents/researchreport92.pdf.

Bickel, P. and Zhang, P. (1992). Variable selection in nonparametric regression with categorical covariates. *J. Amer. Statist. Assoc.* **87**, 90-97.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Springer, New York.

Chen, H. and Chen, K. (1991). Selection of the splined variables and convergence rates in a partial spline model. *Canad. J. Statist.* **19**, 323-339.

Cheng, B. and Tong, H. (1992). On consistent nonparametric order determination and chaos. *J. Roy. Statist. Soc. Ser. B* **54**, 427-474.

Cheng, B. and Tong, H. (1993). On residual sums of squares in non-parametric autoregression. *Stochastic Process. Appl.* **48**, 154-174.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377-403.

Fan, J. and Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *J. Roy. Statist. Soc. Ser. B* **65**, 57-80.

Gao, J. and Tong, H. (2004). Semiparametric nonlinear time series model selection. *J. Roy. Statist. Soc. Ser. B* **66**, 321-336.

Härdle, W. and Liang, H. and Gao, J. (2000). *Partially Linear Models*. Springer, New York.

Roussas, G. G. (1988). Non-parametric estimation in mixing sequencies of random variables. *J. Statist. Plann. Inference* **18**, 135-149.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

Shao, J. (1997). An asymptotic theory for linear model selection (with comments). *Statist. Sinica* **7**, 221-264.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50**, 413-436.

Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction. *J. Roy. Statist. Soc. Ser. B* **36**, 111-147.

Tjøstheim, D. and Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: selecting significant lags. *J. Amer. Statist. Assoc.* **89**, 1410-1419.

Tschernig, R. and Yang, L. (2000). Nonparametric lag selection for time series. *J. Time Ser. Anal.* **21**, 457-487.

Vieu, P. (1994). Choice of regressors in nonparametric estimation. *Comput. Statist. Data Anal.* **17**, 575-594.

Vieu, P. (1995). Order choice in nonlinear autoregressive models. *Statistics* **27**, 307-328.

Wei, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20**, 1-42.

Yang, Y. (1999). Model selection for nonparametric regression. *Statist. Sinica* **9**, 475-499.

Yao, Q. and Tong, H. (1994). On subset selection in non-parametric stochastic regression *Statist. Sinica* **4**, 51-70.

Yoshihara, K. I. (1976). Limiting behavior of U-statistics for stationary, absolutely regular process. *Z. Wahrsch. Verw. Gebiete* **35**, 237-252.

Zhang, P. (1991). Variable selection in nonparametric regression with continuous covariates. *Ann. Statist.* **19**, 1869-1882.

Zhang, P. (1993). Model selection via multifold cross-validation. *Ann. Statist.* **21**, 299-313.

Department of Statistics, London School of Economics, Houghton Street, London, WC2A 2AE, U.K.

E-mail: p.avramidis@lse.ac.uk

(Received March 2003; accepted November 2004)