Selecting Regressors in Partially Linear Models: A technical report

Panagiotis Avramidis

Department of Statistics, London School of Economics,

Houghton Street, London, WC2A 2AE

P.Avramidis@lse.ac.uk

January, 2003

Abstract

In an earlier work, it was shown that the combination of the leave one out Cross Validation criterion introduced for nonparametric models and the leave-n_v-out Cross Validation criterion used for linear regressors selection has successfully identified the nonlinear and the linear variables of a partially linear model. This technical report deals with the mathematical issues of the proposed two step procedure. Under certain regularity conditions we prove analytically the consistency of the regressors estimators.

KEY WORDS: Bandwidth, cross validation, Monte Carlo CV, ergodic theorem, absolute regular process, consistency.

1 Introduction

We begin with the introduction of the partially linear model. 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. We begin with 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 $g : \mathbb{R}^Q \to \mathbb{R}$ an unknown function, $\theta = (\theta_1, \ldots, \theta_P)^T$ the vector of the parameters and $\epsilon_t = Y_t - E(Y_t|\mathbf{X}_t, \mathbf{Z}_t)$ an error term. Then, it follows that $E(\epsilon_t|\mathbf{X}_t, \mathbf{Z}_t) = 0$. Call $U_t = Y_t - \mathbf{X}_t^T\theta$ then equation (1) yields $E(U_t|\mathbf{Z}_t) = g(\mathbf{Z}_t)$. The proposed selection procedure is using the leave-one-out cross validation criterion to select the nonlinear regressors, while the linear regressors are identified using a Monte Carlo estimate of the leave-n_v-out cross validation criterion. Under mild regularity conditions, it was argued that the two regressors subset estimators will converge in probability to the true regressors subsets i.e. the estimators are consistent. One of the most important condition is the existence and identifiability of the true underlying model implied from assumption A1,

A1 We assume that the true model is the model with the optimal nonparametric $\mathbf{Z}_t^q = \{Z_{t,1}, \ldots, Z_{t,q}\}$ and parametric $\mathbf{X}_t^p = \{X_{t,1}, \ldots, X_{t,p}\}$ components

where definition of the optimal nonparametric and parametric components can be found in Yao and Tong (1994). In this report, we are presenting the technical arguments used for the derivation of the asymptotic properties of the regressors subset estimators as shown in an earlier work (Avramidis 2003). In particular, the calculations reveal that after removing the linear effect, application of the leave-one-out cross validation criterion yields similar results to a fully nonparametric model as shown by Yao and Tong (1994). In other words, the linear term does not affect the asymptotic behavior of

the nonparametric subset estimator. An equivalent observation holds for the parametric selection criterion MCCV for a specific range of nonparametric bandwidths. This observation is more surprising as we expect that the estimator of the nonparametric component will converge slower that the parametric estimator. However, similar observations have been made by Speckman (1988) and Härdle, Liang, and Gao (2000).

The following sections involve the mathematical results for the lemmas and theorems that appear in the earlier discussion paper. Details for the overall evaluation of the selection procedure, the intuitive arguments and for numerical results see Avramidis (2003).

2 Consistency of the nonparametric regressors estimator

We now focus on the selection of the nonparametric component. Call $\hat{\theta}$ the parameter estimator calculated by regressing Y_t against $X_{t,1},\ldots,X_{t,P}$ and $\hat{U}_t = Y_t - \mathbf{X}_t^T\hat{\theta}$ the residuals. For any $\{i_1,\ldots,i_k\} \subseteq \{1,\ldots,Q\}$ let $\mathbf{Z}_t^k = (Z_{t,i_1},\ldots,Z_{t,i_k})^T$. The standard Nadaraya-Watson estimation for $g(.)$ is $g_n(\mathbf{z}) = \sum_{t=1}^n w_{t,k}(\mathbf{z})(Y_t - \mathbf{X_t}^T\theta)$ and if we replace θ by its estimator $\hat{\theta}$ we have $\hat{g}_n(\mathbf{z}) = \sum_{t=1}^n w_{t,k}(\mathbf{z})(Y_t - \mathbf{X}_t^T\hat{\theta})$ with $w_{t,k} : \mathbb{R}^k \to \mathbb{R}$, a weight function $w_{t,k}(\mathbf{z}) = K_h(\mathbf{Z}_t^k - \mathbf{z})/\sum_{r=1}^n K_h(\mathbf{Z}_r^k - \mathbf{z})$ and $K_h(.) = 1/h^k K(./h)$, $K : \mathbb{R}^k \to \mathbb{R}$ a k-dimensional kernel function. The leave-one-out estimators are $g_n^{(-s)}(\mathbf{z}) = \sum_{t=1,t\neq s}^n w_{t,k}^{(-s)}(\mathbf{z})(Y_t - \mathbf{X}_t^T\theta)$ and $\hat{g}_n^{(-s)}(\mathbf{z}) = \sum_{t=1,t\neq s}^n w_{t,k}^{(-s)}(\mathbf{z})(Y_t - \mathbf{X}_t^T\hat{\theta})$ respectively, where $w_{t,k}^{(-s)}(\mathbf{z}) = K_h(\mathbf{Z}_t^k - \mathbf{z})/\sum_{r=1,r\neq s}^n K_h(\mathbf{Z}_r^k - \mathbf{z})$. The cross validation function is defined:

$$CV(i_1,\ldots,i_k) = \frac{1}{n}\sum_{s=1}^n \{\hat{U}_s - \hat{g}_n^{(-s)}(Z_{s,i_1},\ldots,Z_{s,i_k})\}^2 \text{ for all } \{i_1,\ldots,i_k\}. \quad (2)$$

The estimator for the optimal regression subset of the nonparametric component is defined as

$$\hat{Q}_0 = \arg \min_{\{i_1,\ldots,i_k\}\subseteq\{1,\ldots,Q\},1\leq k\leq Q} CV(i_1,\ldots,i_k). \quad (3)$$

The following assumptions are necessary for the derivation of the asymptotic results. Let $C > 0$ be a constant that can take different values in different places.

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

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

A4 For the strictly stationary process $\{(Y_t, \mathbf{X}_t, \mathbf{Z}_t) : t = 1, 2, \ldots\}$ let $\beta(n) = \sup_{k \geq 1} E\{\sup_{A \in \Im_{k+n}^{\infty}} |P(A|\Im_1^k) - P(A)|\}$ 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 as $\limsup_{n \to \infty} (1 + 6\sqrt{e}\beta(m_n)^{1/(1+l_n)})^{l_n} < \infty$.

A5 For $1 \leq k \leq Q$ the kernel function $K_h = K(./h)$ where $K : \mathbb{R}^k \to \mathbb{R}$ a symmetric density function with bounded support satisfying the 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 $E|Y_t|^6 < \infty$, $E \parallel \mathbf{X}_t \parallel^6 < \infty$ and $E(Y_t|\mathbf{X}_t, \ldots, \mathbf{X}_1, \mathbf{Z}_t, \ldots, \mathbf{Z}_1) = E(Y_t|\mathbf{X}_t, \mathbf{Z}_t)$ 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 For the $g(\mathbf{z}) = E(U_t|\mathbf{Z}_t^k = \mathbf{z})$,$1 \leq k \leq Q$ it holds that $|g(\mathbf{z}_1) - g(\mathbf{z}_2)| \leq C \parallel \mathbf{z}_1 - \mathbf{z}_2 \parallel^{\gamma}$ where $\gamma > 0$ constant.

A9 For γ in A8, $(k+\gamma)\lambda(k) > 1/2$ for all $1 \leq k \leq Q$ and $k\lambda(k)$ is a strictly increasing function of k.

Lemma 1 *Under assumptions A2-A9 it holds that*

(a) For any $\{i_1, \ldots, i_k\}$, $1 \leq k \leq q$: $CV(i_1, \ldots, i_k) \xrightarrow{P} \sigma^2(i_1, \ldots, i_k)$ *where* $\sigma^2(i_1, \ldots, i_k) = E[U_t - E(U_t|Z_{t,i_1}, \ldots, Z_{t,i_k})]^2$.

(b) *If for some* $\{i_1,\ldots,i_k\}$ *holds that* $E(U_t|Z_{i_1},\ldots,Z_{i_k}) = E(U_t|Z_1,\ldots,Z_q)$ *a.s. then*

$$CV(i_1,\ldots,i_k) = \frac{1}{n}\sum_{s=1}^{n}\epsilon_s^2 + \frac{1}{nh^k}E(\epsilon_t^2/f(\mathbf{Z}_t^k))\int K^2(u)du + o_p(n^{-1}h^{-k}).$$

Proof of lemma 1 Note that: $CV(i_1,\ldots,i_k) = \frac{1}{n}\sum_{s=1}^{n}\{\hat{U}_s - U_s\}^2 + \frac{1}{n}\sum_{s=1}^{n}\{U_s - \hat{g}_n^{(-s)}(\mathbf{Z}_s^k)\}^2 + \frac{2}{n}\sum_{s=1}^{n}\{\hat{U}_s - U_s\}\{U_s - \hat{g}_n^{(-s)}(\mathbf{Z_s}^k)\} = I_1 + I_2 + I_3$. Assumptions A2-A4 along with Slutsky's theorem yield

$$I_1 = \frac{1}{n}\sum_{s=1}^{n}\{\mathbf{X}_t^T(\hat{\theta}-\theta)\}^2 = \frac{1}{n}(\hat{\theta}-\theta)^T\sum_{s=1}^{n}\mathbf{X}_t^T\mathbf{X}_t(\hat{\theta}-\theta) = o_p(n^{-1})$$

Call $\epsilon_s^{(i_1,\ldots,i_k)} = U_s - g(\mathbf{Z}_s^k)$, then $I_2 = \sum_{j=1}^{3} I_{2,j}$ with $I_{2,1} = 1/n\sum_{s=1}^{n}(\epsilon_s^{(i_1,\ldots,i_k)})^2$, $I_{2,2} = 1/n\sum_{s=1}^{n}\{g(\mathbf{Z}_s^k) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^k)\}^2$ $I_{2,3} = 2/n\sum_{s=1}^{n}\epsilon_s^{(i_1,\ldots,i_k)}\{g(\mathbf{Z}_s^k) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^k)\}$. Using the ergodic theorem, (see Fan and Yao (2003), chap.2 prop 2.8) we have that

$$I_{2,1} \xrightarrow{P} E[U_t - E(U_t|Z_{t,i_1},\ldots,Z_{t,i_k})]^2 = \sigma^2(i_1,\ldots,i_k) \tag{4}$$

For the remaining two terms, lemma 2 below shows that they converge in probability to 0. Hence, we proved (a). Further, note that if

$$E(U_t|Z_{t,i_1},\ldots,Z_{t,i_k}) = E(U_t|Z_{t,1},\ldots,Z_{t,q}) \quad a.s. \text{ then } \epsilon_s^{(i_1,\ldots,i_k)} = \epsilon_s$$

so $I_{2,1} = \frac{1}{n}\sum_{s=1}^{n}\epsilon_s^2$. This along with the results in lemma 3 yields (b) and the proof of lemma 1 is complete.

Lemma 2 *Suppose A2-A8 hold. Then for any* $\{i_1,\ldots,i_k\}$, $1 \le k \le q$

(a) $\frac{1}{n}\sum_{s=1}^{n}\epsilon_s^{(i_1,\ldots,i_k)}\{g(\mathbf{Z}_s^k) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^k)\} \xrightarrow{P} 0$

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

Proof of lemma 2 (a): Note that $n^{-1}\sum_{s}^{n}\epsilon_s^{(i_1,\ldots,i_k)}\{g(\mathbf{Z}_s^k) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^k)\} = n^{-1}(\sum_{s}^{n}\epsilon_s^{(i_1,\ldots,i_k)}\{g(\mathbf{Z}_s^k) - g_n^{(-s)}(\mathbf{Z}_s^k)\} + \sum_{s}^{n}\epsilon_s^{(i_1,\ldots,i_k)}\{g_n^{(-s)}(\mathbf{Z}_s^k) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^k)\}) = J_1 + J_2$. Concentrate on J_1: $n^{-1}|\sum_{r=1,r\neq s}^{n}K_h(\mathbf{Z}_s^k - \mathbf{Z}_r^k) - \sum_{r=1}^{n}K_h(\mathbf{Z}_s^k - \mathbf{Z}_r^k)| = o_p(n^{-1+k\lambda(k)}) = o_p(1)$ and $\sup_{\mathbf{z}:f_k(\mathbf{z})>0}|\frac{1}{n}\sum_{r=1}^{n}K_h(\mathbf{z} - \mathbf{Z}_r^k) - f_k(\mathbf{z})| \to 0$. The

latter is implied from assumption A3-A6 and theorem 3.1 Roussas (1988). See also Yao and Tong (1994) or Cheng and Tong (1992). Hence it follows that $g(\mathbf{Z}_s^k) - g_n^{(-s)}(\mathbf{Z}_s^k) =$

$$\Big(g(\mathbf{Z}_s^k) - g_n^{(-s)}(\mathbf{Z}_s^k)\Big)\Big(\frac{1}{n}\sum_{r=1,r\neq s}^{n} K_h(\mathbf{Z}_s^k - \mathbf{Z}_r^k)\Big)(f_k(\mathbf{Z}_s^k))^{-1}(1+o_p(1))$$

almost surely. Note that $(g(\mathbf{Z}_s^k) - g_n^{(-s)}(\mathbf{Z}_s^k))\frac{1}{n}\sum_{r=1,r\neq s}^{n} K_h(\mathbf{Z}_s^k - \mathbf{Z}_r^k) =$

$$\frac{1}{n}\sum_{t=1,t\neq s}^{n} K_h(\mathbf{Z}_s^k - \mathbf{Z}_t^k)\{g(\mathbf{Z}_s^k) - U_t\} = \frac{1}{n}\sum_{t=1}^{n} K_h(\mathbf{Z}_s^k - \mathbf{Z}_t^k)\{g(\mathbf{Z}_s^k) - g(\mathbf{Z}_t^k)\}$$

$$+\frac{1}{n}K_h(0)g(\mathbf{Z}_s^k) - \frac{1}{n}\sum_{t=1,t\neq s}^{n} K_h(\mathbf{Z}_s^k - \mathbf{Z}_t^k)\{U_t - g(\mathbf{Z}_t^k)\} =$$

$$= \frac{1}{nh^k}\Big(\sum_{t=1}^{n} C_{s,t} - \sum_{t=1,t\neq s}^{n} \epsilon_t^{(i_1,\dots,i_k)} d_{s,t} + K(0)g(\mathbf{Z}_s^k)\Big)$$

where $d_{s,t} = K((\mathbf{Z}_s^k - \mathbf{Z}_t^k)/h)$ and $C_{s,t} = d_{s,t}\{g(\mathbf{Z}_s^k) - g(\mathbf{Z}_t^k)\}$. Hence, we write $g(\mathbf{Z}_s^k) - g_n^{(-s)}(\mathbf{Z}_s^k) =$

$$\frac{(f_k(\mathbf{Z}_s^k))^{-1}}{nh^k}\Big(\sum_{t=1}^{n} C_{s,t} - \sum_{t=1,t\neq s}^{n} \epsilon_t^{(i_1,\dots,i_k)} d_{s,t} + K(0)g(\mathbf{Z}_s^k)\Big)(1+o_p(1))$$

and therefore $J_1 = n^{-2}h^{-k}\Big(\sum_{s=1}^{n}\sum_{t=1,t\neq s}^{n} \epsilon_s^{(i_1,\dots,i_k)}\{C_{s,t} - \epsilon_t^{(i_1,\dots,i_k)} d_{s,t}\} f_k^{-1}(\mathbf{Z}_s^k)$ $+K(0)\sum_{s=1}^{n} \epsilon_s^{(i_1,\dots,i_k)} f_k^{-1}(\mathbf{Z}_s^k) g(\mathbf{Z}_s^k)\Big)(1+o_p(1)) = J_{1,1} + J_{1,2} + o_p(J_{1,1}+J_{1,2})$. From the ergodic theorem $nh^k J_{1,2} \xrightarrow{P} E(\epsilon_s^{(i_1,\dots,i_k)} f_k^{-1}(\mathbf{Z}_s^k) g(\mathbf{Z}_s^k))$ and using conditional argument and the fact that $nh^k \to \infty$ from A5, we prove $J_{1,2} \xrightarrow{P} 0$. For $J_{1,1}$ we follow Yao and Tong (1994) and apply the decomposition of U-statistics as proposed by Yoshihara (1976). Particularly, under certain conditions, Yoshihara showed that a U-statistics $U_n = 2!(n-2)!/n!\sum_i^n g(\xi_{i_1}, \xi_{i_2})$ which is an estimator of a functional form $\vartheta(F) = \int_{R^{2p}} g(x_1,x_2)dF(x_1)dF(x_2)$, can be decomposed as

$$\vartheta(F) + \sum_{1\le i_1<i_2\le n}\{g(\xi_{i_1},\xi_{i_2}) - \int g(\xi_{i_1},\xi_{i_2})dF(\xi_{i_2}) - \int g(\xi_{i_1},\xi_{i_2})dF(\xi_{i_1}) + \vartheta(F)\}.$$

Call $\eta_t = (\mathbf{Z}_t^k, \epsilon_s^{(i_1,\ldots,i_k)})^T$ and define $H(\eta_t, \eta_s) =$

$$\epsilon_s^{(i_1,\ldots,i_k)}\Big(C_{s,t} - \epsilon_t^{(i_1,\ldots,i_k)} d_{s,t}\Big) f_k^{-1}(\mathbf{Z}_s^k) + \epsilon_t^{(i_1,\ldots,i_k)}\Big(C_{t,s} - \epsilon_s^{(i_1,\ldots,i_k)} d_{t,s}\Big) f_k^{-1}(\mathbf{Z}_t^k)$$

and $H(\eta_t) = \int H(\eta_t, \eta_s) dP(\eta_s) = \epsilon_t^{(i_1,\ldots,i_k)} f_k^{-1}(\mathbf{Z}_t^k) \int C_{t,s} dP(\mathbf{Z}_s^k)$ then we can write

$$J_{1,1} = \frac{1}{2n^2h^k} \sum_{t \neq s}^{n} \{H(\eta_t, \eta_s) - H(\eta_s) - H(\eta_t)\} + \frac{n-1}{n^2h^k} \sum_{t=1}^{n} H(\eta_t).$$

Note also that $H(\eta_s, \eta_t)$ is symmetric and η_t is a strictly stationary absolutely regular process. Hence, applying the above results for $g(\xi_{i_1}, \xi_{i_2}) = H(\eta_s, \eta_t)$ and noting that $\int H(\eta_t) dP(\eta_t) = 0 \Rightarrow \vartheta(F) = 0$ we have that the first term of $J_{1,1}$ is equal to the remainder in Hoeffding's projection decomposition of the U-statistics. Further, assumption A3 yields

$$|H(\eta_t, \eta_s)| \leq C\Big(|\epsilon_s^{(i_1,\ldots,i_k)}| + |\epsilon_t^{(i_1,\ldots,i_k)}|\Big) + C|\epsilon_s^{(i_1,\ldots,i_k)}||\epsilon_t^{(i_1,\ldots,i_k)}| \ \ a.s.$$

and with assumption A7 $\int |H(\eta_t, \eta_s)|^3 dP(\eta_t) dP(\eta_s) < \infty$ we have that lemma 1 of Yoshihara (1976) yields $\sup_{s<t} E|H(\eta_t, \eta_s)|^3 < \infty$. The latter along with lemma 2 of Yoshihara (1976) implies $E(n^{-2}\sum_{t\neq s}^{n}\{H(\eta_t, \eta_s) - H(\eta_s) - H(\eta_t)\})^2 = o(n^{-2})$ thus $n^{-2}h^{-k}\sum_{t\neq s}^{n}\{H(\eta_t, \eta_s) - H(\eta_s) - H(\eta_t)\} = o_p(n^{-1}h^{-k})$. For the second term note that A3, A8 yield $|H(\eta_t)| \leq Ch^{k+\gamma}|\epsilon_t^{(i_1,\ldots,i_k)}|$ a.s. so $(n-1)/(n^2h^k)|\sum_{t=1}^{n} H(\eta_t)| \leq Ch^{\gamma} n^{-1} \sum_{t=1}^{n} |\epsilon_t^{(i_1,\ldots,i_k)}|$ and the latter converges a.s. to zero by the ergodic theorem. Thus, we have shown that $J_{1,1} \xrightarrow{P} 0$ which completes the proof of $J_1 \xrightarrow{P} 0$.

It remains to prove $J_2 \xrightarrow{P} 0$ in order to complete the proof of (a). From assumption A3 it follows that X_t is bounded process i.e. $\parallel X_t \parallel \leq M < \infty$ a.s. Consequently, from

$$\frac{1}{n}\Big|\sum_{s=1}^{n} \epsilon_s^{(i_1,\ldots,i_k)}\Big(\sum_{t=1, t\neq s}^{n} w_{t,k}^{(-s)}(\mathbf{Z}_s^k) X_t^T(\hat{\theta} - \theta)\Big)\Big| \leq M \parallel \hat{\theta} - \theta \parallel \frac{1}{n}\sum_{s=1}^{n} |\epsilon_s^{(i_1,\ldots,i_k)}|$$

ergodic theorem and A2 yield $J_2 \xrightarrow{P} 0$. The proof of (b) contains similar technical details with the proof of (a) so omitted and lemma 2 is complete.

Lemma 3 *Suppose A2-A9 hold and that for some $\{i_1, \ldots, i_k\}$ and $1 \leq k \leq q$*

$$E(Y_t|\mathbf{X}_t, Z_{i_1}, \ldots, Z_{i_k}) = E(Y_t|\mathbf{X}_t, Z_1, \ldots, Z_q) \quad a.s. \tag{5}$$

(a) $n^{-1} \sum_{s=1}^{n} \epsilon_s^{(i_1,\ldots,i_k)} \{g(\mathbf{Z}_s^k) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^k)\} = o_p(n^{-1}h^{-k})$

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

Proof of lemma 3 (a): From lemma 2, $J_{1,2} = o_p(n^{-1}h^{-k})$ while for $J_{1,1}$ we have that $n^{-2}h^{-k} \sum_{t \neq s}^{n} \{H(\eta_t, \eta_s) - H(\eta_s) - H(\eta_t)\} = o_p(n^{-1}h^{-k})$. Note that under (5) we have that $E(H(\eta_t)H(\eta_s)) = 0$ for $s < t$ the latter from the fact that $E(H(\eta_t)|\mathbf{X}_t, \ldots, \mathbf{X}_1, \mathbf{Z}_t, \ldots, \mathbf{Z}_1) = C(\mathbf{Z}_t)E(\epsilon_t|\mathbf{X}_t, \mathbf{Z}_t) = 0$. It follows that $E(n^{-1} \sum_{t=1}^{n} H(\eta_t))^2 = n^{-1}EH^2(\eta_t) \leq Ch^{2k+2\gamma}n^{-1}$ which along with A9 yields

$$\frac{n-1}{n^2h^k} \sum_{t=1}^{n} H(\eta_t) = O_p(\frac{n-1}{nh^k}(\frac{h^{2k+2\gamma}}{n})^{1/2}) = O_p(h^{\gamma}n^{-1/2}) = o_p(n^{-1}h^{-k}).$$

We conclude that $J_1 = o_p(n^{-1}h^{-k})$. Similarly, note that assumption A7 implies $E(\epsilon_t\epsilon_s) = 0$ so we can write $E|J_2|^2 \leq n^{-1}M^2E \parallel \hat{\theta} - \theta) \parallel^2 E(\epsilon_t^2) \Rightarrow E|J_2|^2 = o(n^{-2})$ the latter from A2. Therefore, $J_2 = O_p(E|J_2|^2)^{1/2} = O_p(n^{-1}) = o_p(h^{-k}n^{-1})$ and conclude. For (b): $n^{-1} \sum_{s=1}^{n} \{g(\mathbf{Z}_s^k) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^k)\}^2 =$

$$\frac{1}{n} \sum_{s=1}^{n} \{g(\mathbf{Z}_s^k) - g_n^{(-s)}(\mathbf{Z}_s^k)\}^2 + \frac{1}{n} \sum_{s=1}^{n} \{g_n^{(-s)}(\mathbf{Z}_s^k) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^k)\}^2$$

$$+\frac{2}{n} \sum_{s=1}^{n} \{g(\mathbf{Z}_s^k) - g_n^{(-s)}(\mathbf{Z}_s^k)\}\{g_n^{(-s)}(\mathbf{Z}_s^k) - \hat{g}_n^{(-s)}(\mathbf{Z}_s^k)\} = \sum_{j=1}^{3} R_j$$

First, concentrate on R_1. Recall lemma 2, then we can easily see that

$$R_1 = \frac{1}{n^3h^{2k}} \sum_{s=1}^{n} \sum_{t=1, t \neq s}^{n} (C_{s,t} - \epsilon_t d_{s,t})^2 f^{-2}(\mathbf{Z}_s^k) +$$

$$+\frac{1}{n^3h^{2k}} \sum_{s=1}^{n} \sum_{t_1,t_2=1, t_1 \neq t_2 \neq s}^{n} (C_{s,t_1} - \epsilon_{t_1}d_{s,t_1})(C_{s,t_2} - \epsilon_{t_2}d_{s,t_2}) f^{-2}(\mathbf{Z}_s^k)$$

$$+\frac{2}{n^3h^{2k}}K^2(0)\sum_{s=1}^{n}\sum_{t=1,t\neq s}^{n}g(\mathbf{Z}_s^k)(C_{s,t}-\epsilon_t d_{s,t})f^{-2}(\mathbf{Z}_s^k)$$

$$+\frac{1}{n^3h^{2k}}K^2(0)\sum_{s=1}^{n}g^2(\mathbf{Z}_s^k)f^{-2}(\mathbf{Z}_s^k)+o_p(R_1)=\sum_{j=1}^{4}R_{1,j}(1+o_p(1))$$

The process η_t defined in lemma 2 can be written as $\eta_t=(\mathbf{Z}_t^k,\epsilon_t)^T$ under (5) and hence define

$$H_1(\eta_t,\eta_s)=(C_{s,t}^2-2C_{s,t}\epsilon_t d_{s,t}+\epsilon_t^2 d_{s,t}^2)f^{-2}(\mathbf{Z}_s^k)+(C_{t,s}^2-2C_{t,s}\epsilon_s d_{t,s}+\epsilon_s^s d_{t,s}^2)f^{-2}(\mathbf{Z}_t^k)$$

and $H_1(\eta_t)=\int C_{s,t}^2 f^{-2}(\mathbf{Z}_s^k)dP(\mathbf{Z}_s^k)-2\epsilon_t\int C_{s,t}d_{s,t}f^{-2}(\mathbf{Z}_s^k)dP(\mathbf{Z}_s^k)$

$$+\epsilon_t^2\int d_{s,t}^2 f^{-2}(\mathbf{Z}_s^k)dP(\mathbf{Z}_s^k)+f^{-2}(\mathbf{Z}_t^k)\int C_{t,s}^2 dP(\mathbf{Z}_s^k)+f^{-2}(\mathbf{Z}_t^k)\int \epsilon_s^2 d_{t,s}^2 dP(\mathbf{Z}_s^k)$$

$$\text{thus}\quad R_{1,1}=\frac{1}{2n^3h^{2k}}\sum_{t\neq s}^{n}\{H_1(\eta_t,\eta_s)-H_1(\eta_t)-H_1(\eta_s)-\int H_1(\eta_t)dP(\eta_t)\}$$

$$+\frac{1}{2nh^{2k}}\int H_1(\eta_t)dP(\eta_t)+\frac{n-1}{n^3h^{2k}}\sum_{t=1}^{n}H_1(\eta_t)$$

Note that like in lemma 2 the first term can be interpreted as the remainder in Hoeffding's projection decomposition of the U-statistics generated by $H_1(\eta_t,\eta_s)$ and using similar arguments we show that

$$\frac{1}{2n^3h^{2k}}\sum_{t\neq s}^{n}\{H_1(\eta_t,\eta_s)-H_1(\eta_t)-H_1(\eta_s)-\int H_1(\eta_t)dP(\eta_t)\}=o_p(n^{-1}h^{-k}). \tag{6}$$

For the second term note $n^{-1}h^{-2k}\int H_1(\eta_t)dP(\eta_t)=$

$$\frac{1}{nh^{2k}}\int C_{s,t}^2 f^{-2}(\mathbf{Z}_t^k)dP(\mathbf{Z}_s^k)dP(\mathbf{Z}_t^k)+\frac{1}{nh^{2k}}\int \epsilon_t^2 d_{s,t}^2 f^{-2}(Z_s^k)dP(\mathbf{Z}_s^k)dP(\mathbf{Z}_t^k)$$

but from A3 and A8-A9 the first part is $O(h^{2\gamma+2k}n^{-1}h^{-2k})=o(n^{-1}h^{-k})$ while the second part is

$$\frac{1}{nh^{2k}}\int \epsilon_t^2 d_{s,t}^2 f^{-2}(\mathbf{Z}_s^k)dP(\mathbf{Z}_s^k)dP(\mathbf{Z}_t^k)\sim n^{-1}h^{-k}E(\epsilon_t^2/f(\mathbf{Z}_t^k))\int K^2(u)du$$

where with $\sim$ we imply that they are asymptotically equivalent. Thus

$$\frac{1}{2nh^{2k}}\int H_1(\eta_t)dP(\eta_t) = n^{-1}h^{-k}E(\epsilon_t^2/f(\mathbf{Z}_t^k))\int K^2(u)du + o_p(n^{-1}h^{-k}). \quad (7)$$

For the third term of $R_{1,1}$ note that

$$E\Big(\frac{1}{n}\sum_{t=1}^{n} H_1(\eta_t)\Big)^2 = E\Big(\frac{1}{n^2}\sum_{t=1}^{n} H_1^2(\eta_t)\Big) + E\Big(\frac{2}{n}\sum_{t=1}^{n-1}(1-\frac{t}{n})H_1(\eta_1)H_1(\eta_{t+1})\Big)$$

$$= \frac{1}{n}E(H_1^2(\eta_t)) + E(\frac{2}{n}\sum_{t=1}^{n-1}(1-\frac{t}{n})H_1(\eta_1)H_1(\eta_{t+1})) = O(h^{4k})$$

the latter from A8 and A9. Hence

$$\frac{n-1}{n^3h^{2k}}\sum_{t=1}^{n} H_1(\eta_t) = O_p(n^{-1}) = o_p(n^{-1}h^{-k}). \quad (8)$$

(6),(7) and (8) yield $R_{1,1} = n^{-1}h^{-k}E(\epsilon_t^2/f(\mathbf{Z}_t^k))\int K^2(u)du + o_p(n^{-1}h^{-k})$, Using similar arguments based on the decomposition of U-statistics we can show that $R_{1,j} = o_p(n^{-1}h^{-k})$ for $j = 2, 3$ while the standard ergodic theorem yields $R_{1,4} = o_p(n^{-1}h^{-k})$. Therefore we proved

$$R_1 = n^{-1}h^{-k}E(\epsilon_t^2/f(\mathbf{Z}_t^k))\int K^2(u)du + o_p(n^{-1}h^{-k}). \quad (9)$$

For R_2 note that $R_2 = n^{-1}\sum_{s=1}^{n}(\sum_{t=1,t\neq s}^{n} w_{t,k}^{(-s)}(\mathbf{Z}_s^k)\mathbf{X}_t^T(\hat{\theta}-\theta))^2$ thus $|R_2| \leq C \parallel \hat{\theta}-\theta \parallel^2 n^{-1}\sum_{s=1}^{n}|\sum_{t=1,t\neq s}^{n} w_{t,k}^{(-s)}(\mathbf{Z}_s^k)|^2$ and since $\sum_{t=1,t\neq s}^{n} w_{t,k}^{(-s)}(\mathbf{Z}_s^k) = O_p(1)$ assumption A2 yields $E|R_2| = O(n^{-1}) \Rightarrow$

$$R_2 = O_p(n^{-1}) = o_p(n^{-1}h^{-k}). \quad (10)$$

Note that Cauchy-Schwartz inequality yields $E|R_3| \leq (E(R_1))^{1/2}(E(R_2))^{1/2}$ $\Rightarrow E|R_3| = O(h^{-k/2}n^{-1/2})o(h^{-k/2}n^{-1/2}) = o(n^{-1}h^{-k})$ so

$$R_3 = O_p(E|R_3|) = o_p(n^{-1}h^{-k}) \quad (11)$$

and the proof of (b) concludes from (9),(10) and (11).

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

Theorem 1 *Under assumptions A1-A9, it holds that*

$$\lim_{n\to\infty} P(\hat{Q}_0 = \{1,\ldots,q\}) = 1$$

Proof of theorem 1 For any $\{i_1,\ldots,i_k\}$ subset of $\{1,\ldots,Q\}$ and $1 \le k \le Q$, if $\sigma^2(i_1,\ldots,i_k) > \sigma^2(1,\ldots,Q) = \sigma^2(1,\ldots,q)$ then from lemma 1(a) it follows that $P\Big(CV(1,\ldots,q) < CV(i_1,\ldots,i_k)\Big) \to 1$. Alternatively, if $\sigma^2(i_1,\ldots,i_k) = \sigma^2(1,\ldots,Q) = \sigma^2(1,\ldots,q)$ then condition (5) in lemma 3 holds. Note also that by definition $k > q$. Hence, from assumption A9,

$$h^q/h^k = n^{k\lambda(k)-q\lambda(q)} \to \infty \text{ as } n \to \infty \tag{12}$$

thus lemma 1 (b) along with (12): $P\Big(CV(i_1,\ldots,i_k) - CV(1,\ldots,q) > 0\Big) =$

$$P\Big(nh^q(CV(i_1,\ldots,i_k) - CV(1,\ldots,q)) > 0\Big)$$

$$= P\Big(\frac{h^q}{h^k}\int K^2(u)du\{E(\epsilon_t^2/f(\mathbf{Z}_t^k)) - E(\epsilon_t^2/f(\mathbf{Z}_t^q))\} + o_p(h^{q-k}) > 0\Big) \to 1$$

$\Rightarrow P(\hat{Q}_0 = \{1,\ldots,q\}) \to 1$ as $n \to \infty$ and theorem's proof is complete.

Note here that the main terms of the decomposition of the CV-function are the same with those derived in Yao and Tong (1994) for a fully nonparametric model.

3 Consistency of the parametric regressors estimator

For any $M \subset \{1,\ldots,P\}$ we write $Y_t = (\mathbf{X}_t^M)^T\theta_M + g(\mathbf{Z}_t^q) + \epsilon_{t,M}$ where $\mathbf{X}_t^M = (X_{t,i} : i \in M)^T$ and $\epsilon_{t,M} = Y_t - E(Y_t|\mathbf{X}_t^M, \mathbf{Z}_t^q)$. Substituting $g(.)$ with the nonparametric estimator $g_n(\mathbf{Z}_t^q) = \sum_{s=1}^n w_{s,q}(\mathbf{Z}_t^q)(Y_s - \mathbf{X}_s^T\theta)$ yields that the least squares estimator of θ_M is

$$\hat{\theta}_M = (\tilde{\mathbf{X}}_M^T\tilde{\mathbf{X}}_M)^{-1}\tilde{\mathbf{X}}_M^T\tilde{\mathbf{Y}}. \tag{13}$$

where $\tilde{Y}_t = Y_t - \sum_{s=1}^n w_{s,q}(\mathbf{Z}_t^q)Y_s$ and $\tilde{\mathbf{X}}_t^M = \mathbf{X}_t^M - \sum_{s=1}^n w_{s,q}(\mathbf{Z}_t^q)\mathbf{X}_s^M$. Let M_0 denote the true underlying regressors subset. The conditional expected

mean square error is found to be $EMSE_n(M) = \sigma^2_{\tilde{\epsilon}} - \frac{m}{n}\sigma^2_{\tilde{\epsilon}} + \Omega_{n,M}$ where $\sigma^2_{\tilde{\epsilon}} = \frac{1}{n}E(\tilde{\epsilon}^T\tilde{\epsilon})$ and $\Omega_{n,M} = \frac{1}{n}\theta^T\tilde{\mathbf{X}}^T\mathbf{H}_M\tilde{\mathbf{X}}\theta$. Details can be found in Avramidis (2003). The important observation is that the parametric estimator is not affected by the slower rate of convergence of the nonparametric estimator, a result already stated in earlier works by Speckman (1988) and Härdle, Liang, and Gao (2000). In addition to A1-A9, we impose the following assuption

A10 For every $M \subset \{1,\dots,P\}$ $E(\tilde{\mathbf{X}}^T_M\tilde{\mathbf{X}}_M)$ is a positive definite matrix with order $m \times m$. Further, for all M such that $M_0 \not\subseteq M$ it holds that $\liminf_{n\to\infty}\Omega_{n,M} > 0$ in probability.

Note that for every $M \subset \{1,\dots,P\}$ with $M_0 \subset M$ $EMSE_n(M) = \frac{1}{n}(n - m)\sigma^2_{\tilde{\epsilon}}$ from $\tilde{\mathbf{X}}\theta = \tilde{\mathbf{X}}_M\theta_M$. Further let $d_j(\mathbf{z}) = E(X_{t,j}|\mathbf{Z}^q_t = \mathbf{z})$ and define $u_{t,j} = X_{t,j} - d_j(\mathbf{Z}^q_t)$ for $j = 1,\dots,P$ and $\mathbf{u}_t = (u_{t,1},\dots,u_{t,P})^T$. The leave-$n_v$-out cross validation function is defined by $CV(M, n_v) = n_v^{-1} \parallel \tilde{Y}_N - \tilde{\mathbf{X}}_{N,M}\hat{\theta}_{N^c,M} \parallel^2$ where the data is split to $\{(\tilde{Y}_t, \tilde{\mathbf{X}}_t) : t \in N\}$ the validation data used for the calculation of the cross validation function and $\{(\tilde{Y}_t, \tilde{\mathbf{X}}_t) : t \in N^c\}$, the construction data used for fitting the model. Note here that N^c is the compliment of $N \subset \{t : t = 1,\dots,n\}$ and n_v is the dimensionality of N. To calculate the leave-n_v-out CV function we employ the Monte Carlo technique. We randomly draw a collection $\mathcal{B}$ of b subsets of $\{1,\dots,n\}$ each one with size n_v then the estimator for the optimal regression subset of the linear component is defined as

$$\hat{M} = \arg\min_{M\subset\{1,\dots,P\}} MCCV(M, n_v). \tag{14}$$

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

Theorem 2 *Assume that A1-A10 hold and that $n_v/n \to 1$, $n_c = n - n_v \to \infty$ and $n^2/n_c^2 b \to 0$ as $n \to \infty$, then we have*

(1) If $M_0 \not\subseteq M$, then there exists $R_n \geq 0$ such that

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

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

(2) If $M_0 \subset M$, 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) Combining (1) and (2) we conclude $\lim_{n\to\infty} P(\hat{M} = M_0) = 1$.

Proof of theorem 2 The proof is based on the proof of theorem 2 in Shao (1993). Also, similar result for the partial linear model can be found in theorem 2.2 Gao and Tong (2002). Hence we only present an outline of the proof and particularly we show that conditions in theorem 2 Shao (1993) hold. Indeed condition 2.5 3.12 and 3.22 have been introduced in directly in the theorem. Hence, it remains to show

$$\max_{N\in\mathcal{B}} \parallel \frac{1}{n_v}\sum_{t\in N} \tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - \frac{1}{n_c}\sum_{t\in N^c} \tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T \parallel = o_p(1) \tag{15}$$

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = O_p(n),\ (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1} = O_p(n^{-1}) \text{ and} \tag{16}$$

$$\lim_{n\to\infty} \max p_{t,M} = 0 \text{ for all } M \tag{17}$$

where $p_{t,M}$ is the t-th diagonal element of $\mathbf{P}_M$. Lemma 5 establishes the above conditions so the outline of the proof of theorem 2 is complete. The proof of theorem 2 is based on the following lemma which is an extension of lemma A.3 in Härdle, Liang and Gao (2000) for α-mixing processes. Proofs are postponed to Appendix A.

Lemma 4 *Let $X_1, \ldots, X_n$ be zero mean, strictly stationary, α-mixing, real valued r.v. Let the mixing coefficients follow $t^6\alpha(t) \to 0$. Suppose that for some $r > 2$, $\sup_{1\le i\le n} E|X_i|^r < C < \infty$ and denote with $\alpha_{i,j}$ with $i, j = 1, \ldots, n$ 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 it holds that $\max_{1\le j\le n} |\sum_{i=1}^n \alpha_{i,j}X_i| = o(n^{-p+1/3+1/r}\log n)$.*

Proof of lemma 4 Define $X_i' = X_i I(|X_i| \le n^{1/r})$ and $X_i'' = X_i - X_i'$. Note that $\sup_{1\le i\le n} |\alpha_{i,j}X_i'| < Cn^{-p}n^{1/r} \equiv M$. Then 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:

$$P\Big(\max_{1\le j\le n} |\sum_{i=1}^n \alpha_{i,j}(X_i' - EX_i')| > n\varepsilon\Big) \le \sum_{j=1}^n P\Big(|\sum_{i=1}^n \alpha_{i,j}(X_i' - E(X_i'))| > n\varepsilon\Big)$$

$$\leq 4n\exp\left(-\frac{\varepsilon^2 q}{8v^2(q)}\right)+22n^{1+2/3}(1+\frac{4M}{\varepsilon})^{1/2}\alpha([n^{1/3}/2])$$

where $\alpha(k)$ we denote the mixing coefficient and

$$v^2(q)\leq\frac{8}{n^{2/3}}\{\max_{0\leq t\leq n}E(\alpha_{i,j}(X_i^{'}-EX_i^{'}))^2+8M^2\sum_{k=1}^{[n^{2/3}]+1}\alpha(k)\}+\frac{M\varepsilon}{2}.$$

Note that $v^2(q)\leq CM^2n^{-2/3}+\frac{1}{2}M\varepsilon\leq Cn^{-2p+2/r-2/3}$ thus, we have that

$$P\Big(\max_{1\leq j\leq n}|\sum_{i=1}^n\alpha_{i,j}(X_i^{'}-EX_i^{'})|>n\varepsilon\Big)\leq 4n\exp\Big(-\frac{Cn^{-2p-2+4/3+2/r}\log^2 n}{n^{-2p-2/3+2/r}}\Big)$$

$$+22n^{5/3}(1+\frac{4Cn^{-p+1/r}}{n^{-p-2/3+1/r}\log n})^{1/2}\alpha([n^{1/3}/2])\leq 4n\exp\big(-\log^2 n\big)$$

$$+22n^2(n^{-2/3}+4C\log^{-1}n)^{1/2}\alpha([n^{1/3}/2])\leq n^{1-\log n}+C_2n^2\alpha(n^{1/3})\to 0$$

for the mixing coefficients holds that $t^6\alpha(t)\to 0$ when $t\to\infty$. Thus, Borel-Cantelli lemma yields

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

For the second process $X_i^{''}$, note that Hölder's inequality with m,l such as $1/m\leq 1/3$ and $1/m+1/l=1$ yields

$$\max_{1\leq j\leq n}|\sum_{i=1}^n\alpha_{i,j}(X_i^{''}-EX_i^{''})|\leq\max_{1\leq j\leq n}\Big(\sum_{i=1}^n|\alpha_{i,j}|^m\Big)^{\frac{1}{m}}\Big(\sum_{i=1}^n|X_i^{''}-EX_i^{''}|^l\Big)^{\frac{1}{l}}$$

$\leq Cn^{-p+1/m}\Big(\sum_{i=1}^n|X_i^{''}-EX_i^{''}|^l\Big)^{\frac{1}{l}}$. 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 $X_i^{''}=X_i-X_i^{'}=X_i-X_iI(|X_i|\leq n^{1/r})=X_iI(|X_i|\geq n^{1/r})$ and $E|X_i^{''}|^l=E(|X_i|^lI(|X_i|\geq n^{1/r}))\leq\Big(E|X_i|^r\Big)^{l/r}\Big(E(I(|X_i|\geq n^{1/r}))\Big)^{1-l/r}$

$$=\Big(E|X_i|^r\Big)^{l/r}\Big(P(|X_i|\geq n^{1/r})\Big)^{1-l/r}\leq\Big(E|X_i|^r\Big)^{l/r}\Big(\frac{E|X_i|^r}{n}\Big)^{1-l/r}$$

the latter from Markov inequality, thus we have $E|X_i^{''}|^l \leq E|X_i|^r n^{l/r-1}$. Hence, from $E|X_i^{''} - EX_i^{''}|^l \leq CE|X_i^{''}|^l \leq CE|X_i|^r n^{l/r-1} \leq Cn^{l/r-1}$ along with (19) we prove that $\sum_{i=1}^n |X_i^{''} - EX_i^{''}|^l \leq Cn^{l/r}$ a.s. hence

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

and lemma concludes from (18) and (20).

Lemma 5 *Under assumptions A4-A5 and A7 it holds that*

(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 $M \subset \{1,\dots,P\}$ *and*

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

Proof of lemma 5 Start with (a). Note that $u_{t,j} = X_{t,j} - d_j(\mathbf{Z}_t^q)$ for $j = 1,\dots,P$, $d_j(z) = E(X_{t,j}|\mathbf{Z}_t^q = z)$ is a strictly stationary α-mixing process with mixing coeficient $\alpha(t) = \alpha^t$ with $0 < \alpha < 1$. Call $A_n = n^{-1}\sum_{t=1}^n \tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T$ then we prove that for $i,j = 1,\dots,P$

$$A_{n,i,j} = \frac{1}{n}\sum_{t=1}^n \tilde{X}_{t,i}\tilde{X}_{t,j} \to A_{i,j}$$

such as $A = [A_{i,j}]$ is a positive definite matrix (see A10). Indeed, note that

$$\tilde{X}_{t,j} = u_{t,j} - \sum_{s=1}^n w_s(\mathbf{Z}_t^q)u_{s,j} + D_{t,j} \text{ with } D_{t,j} = d_j(\mathbf{Z}_t^q) - \sum_{s=1}^n w_s(\mathbf{Z}_t^q)d_j(\mathbf{Z}_t^q).$$

$$A_{n,i,j} = \frac{1}{n}\sum_{t=1}^n u_{t,i}u_{t,j} + \frac{1}{n}\sum_{t=1}^n D_{t,j}D_{t,i} + \frac{1}{n}\sum_{t=1}^n \left(\sum_{s=1}^n w_s(\mathbf{Z}_t^q)u_{s,j}\sum_{k=1}^n w_k(\mathbf{Z}_t^q)u_{k,i}\right)$$
$$-\frac{1}{n}\sum_{t=1}^n D_{t,j}\sum_{k=1}^n w_k(\mathbf{Z}_t^q)u_{k,i} - \frac{1}{n}\sum_{t=1}^n D_{t,i}\sum_{s=1}^n w_k(\mathbf{Z}_t^q)u_{s,j} - \frac{1}{n}\sum_{t=1}^n u_{t,j}\sum_{k=1}^n w_k(\mathbf{Z}_t^q)u_{k,i}$$
$$-\frac{1}{n}\sum_{t=1}^n u_{t,i}\sum_{s=1}^n w_k(\mathbf{Z}_t^q)u_{s,j} + \frac{1}{n}\sum_{t=1}^n D_{t,j}u_{t,i} + \frac{1}{n}\sum_{t=1}^n D_{t,i}u_{t,j} = \sum_{m=0}^8 J_m.$$

J_0 converges from A4, A7 and ergodic theorem $n^{-1}\sum_{t=1}^n u_{t,j}u_{t,i} \xrightarrow{P} E(u_{1,j}u_{1,i})$. Lipschitz kernel and $\max_{1\leq t\leq n}\sum_{s=1}^n w_s(\mathbf{Z}_t^q)I(\parallel \mathbf{Z}_t^q - \mathbf{Z}_s^q \parallel > n^{-1/2}) = O_p(n^{-1/2})$ yield

$$\max_{1\leq t\leq n}|d_j(\mathbf{Z}_t^q) - \sum_{s=1}^n w_s(\mathbf{Z}_t^q)d_j(\mathbf{Z}_t^q)| = O_p(n^{-1/2}) \tag{21}$$

Then, using Abel's inequality and the latter equation (21), note that $J_1 = O_p(n^{-1}) = o(1)$. Further, lemma 4 for $p = 1 - k\lambda(k)$, $r = 6$, yields

$$|\max_{1\leq t\leq n}\sum_{s=1}^n w_s(\mathbf{Z}_t^q)u_{s,j}| = O_p(n^{k\lambda(k)-1/2}\log n) \tag{22}$$

which along with Abel's inequality yields $J_2 = O_p(n^{2k\lambda(k)-1}\log n) = o_p(1)$ since in A5 $k\lambda(k) < 1/2$. Similarly for $m = 3, 4$ from equation (21) and (22)

$$J_m \leq C\max_{1\leq t\leq n}|D_{t,j}|\max_{1\leq t\leq n}|\sum_{s=1}^n w_s(\mathbf{Z}_t^q)u_{s,i}| = O_p(n^{k\lambda(k)-1}\log n) = o_p(1)$$

while Cauchy-Swartz inequality for $m = 5, 6$ along with the ergodic theorem and equation (22)

$$J_m \leq Cn^{-1/2}\Big(\sum_{t=1}^n u_{t,i}^2\Big)^{1/2}\max_{1\leq t\leq n}|\sum_{s=1}^n w_k(\mathbf{Z}_t^q)u_{s,j}| = O_p(n^{k\lambda(k)-1/2}\log n) = o_p(1)$$

from A5. Finally Cauchy-Swartz inequality, ergodic theorem and (21) for $m = 7, 8$ $J_m \leq Cn^{-1/2}(\sum_{t=1}^n u_{t,i}^2)^{1/2}\max_{1\leq t\leq n}|D_{t,j}| = O_p(n^{-1/2}) = o_p(1)$. Hence conclude (a). It can be easily seen that the t-th diagonal element is $p_{t,M} = \sum_{r,j=1}^m \tilde{\mathbf{X}}_{t,i_r}^2 C_{t,j}$ where $C_{t,j} = O_p(n^{-1})$ from (a), hence under assumption A4, A5 and A7 conclude (b). Concentrate on (c):
$\max_{N\in\mathcal{B}} \parallel \frac{1}{n_u}\sum_{t\in N}\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - \frac{1}{n_c}\sum_{t\in N^c}\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T \parallel \leq$

$$\max_{N\in\mathcal{B}} \parallel \frac{1}{n_u}\sum_{t\in N}\{\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - E(\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T)\} \parallel + \max_{N\in\mathcal{B}} \parallel \frac{1}{n_c}\sum_{t\in N^c}\{\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - E(\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T)\} \parallel$$

Call $\tau_t = \tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T - E(\tilde{\mathbf{X}}_t\tilde{\mathbf{X}}_t^T)$ then we prove $\max_{N\in\mathcal{B}} \parallel \sum_{t\in N}\tau_t \parallel = o_p(n_u)$, $\max_{N\in\mathcal{B}} \parallel \sum_{t\in N^c}\tau_t \parallel = o_p(n_c)$ using similar arguments like in (a). Details are omitted. The proof is now complete.

Acknowledgement The author wishes to thank Professor Q. Yao for his helpful comments. This report was produced parallel to another work and it provides the technical details of theoretical findings.

References

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

Bosq, D. (1998). *Nonparametric statistics for stochastic processes: estimation and prediction*. New York: Springer.

Cai, Z., J. Fan, and Q. Yao (2003). Adaptive varying-coefficient linear models. To appear in J. Roy. Statist. Soc. B.

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

Fan, J. and Q. Yao (2003). *Nonlinear time series: nonparametric and parametric methods*. New York: Springer Series in Contributions to Statistics Physica-Verlag.

Härdle, W., H. Liang, and J. Gao (2000). *Partially Linear Models*. New York: Springer Series in Contributions to Statistics Physica-Verlag.

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

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

Yao, Q. and H. Tong (1994). On subset selection in non-parametric stochastic regression. *Statistica 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.