Statistica Sinica **17**(2007), 99-114

SEMIPARAMETRIC PENALTY FUNCTION METHOD IN PARTIALLY LINEAR MODEL SELECTION

Chaohua Dong, Jiti Gao and Howell Tong

Shanxi University of Economics and Finance,
University of Western Australia and London School of Economics

Abstract: Model selection in nonparametric and semiparametric regression is of both theoretical and practical interest. Gao and Tong (2004) proposed a semiparametric leave–more–out cross–validation selection procedure for the choice of both the parametric and nonparametric regressors in a nonlinear time series regression model. As recognized by the authors, the implementation of the proposed procedure requires the availability of relatively large sample sizes. In order to address the model selection problem with small or medium sample sizes, we propose a model selection procedure for practical use. By extending the so–called penalty function method proposed in Zheng and Loh (1995, 1997) through the incorporation of features of the leave-one-out cross-validation approach, we develop a semiparametric, consistent selection procedure suitable for the choice of optimum subsets in a partially linear model. The newly proposed method is implemented using the full set of data, and simulations show that it works well for both small and medium sample sizes.

Key words and phrases: Linear model, model selection, nonparametric method, partially linear model, semiparametric method.

1. Introduction

The problem of model selection in linear, nonlinear and partially linear regression models has attracted much attention in recent years. Many selection methods have been developed since the pioneering notions of Akaike's information criterion (AIC) (Akaike (1973) and Shibata (1976)) and Mallows' C_p (Mallows (1973)). Recent developments that address the nonlinear and partially linear cases include leave–one–out cross-validation (CV1) (e.g., Cheng and Tong (1992), Vieu (1994) and Yao and Tong (1994)), and leave–more–out cross–validation (e.g., Shao (1993), Zhang (1993) and Gao and Tong (2004)). Usually a model selection method is to select the true subset of covariates for either a given dependent variable or a collection of covariates. The most commonly adopted strategy consists of two components: the residual sum of squares (RSS) (as a measure of the goodness of fit) and some function of the number of covariates in the candidate model (as a measure of the penalty for model complexity). By minimizing the sum of these two components, sometimes called the final prediction error, we aim to identify the true model.

Asymptotic consistency is obtained for the fully nonparametric CV1 model selection (e.g., Cheng and Tong (1992)), but in a parametric setting the leave–one–out cross–validation tends to lead to an inconsistent selection and is considered to be too conservative in the sense that it tends to select an unnecessarily large model. In order to restore consistency, Shao (1993) proposed the so–called leave–more–out cross–validation, denoted by $CV(T_v)$ with $T_v/T \to 1$, where T is the number of observations and T_v the number of observations removed for cross-validation. Recently, Gao and Tong (2004) extended this method to semiparametric time series modelling. As observed in the simulations in both Shao (1993) and Gao and Tong (2004), the number of observations, T_c, used to fit the model is quite small (with $T = 40$ and $T_c = 15$ in Shao (1993), and $T = 288$ and $T_c = 69$ in Gao and Tong (2004)), while the number of observations, T_v, used to validate the proposed method is relatively large (with $T_v = 25$ and $T_v = 219$, respectively). This may impede the implementation of the above methodology in practice because the theory requires $T_c \to \infty$. To ensure that more data are used to construct the model, Zheng and Loh (1995, 1997) proposed using a parametric penalty function method. Such observations have motivated us to consider combining the penalty function method proposed in Zheng and Loh (1995, 1997) with the CV1 method for semiparametric modelling. The use of semiparametric models has become increasingly popular in statistical modelling, mainly because these models are quite effective in dealing with the curse of dimensionality that often arises from using fully nonparametric models.

Specifically, we consider a partial linear model of the form

$$Y_t = U_t^\tau \beta + \phi(X_t) + e_t, \quad t = 1, \ldots, T, \tag{1.1}$$

where $U_t = (U_{t1}, \ldots, U_{tp})^\tau$ and $X_t = (X_{t1}, \ldots, X_{tq})^\tau$ are vectors of either independent or dependent observations (U_t and X_t may be two different time series), $\beta = (\beta_1, \ldots, \beta_p)^\tau$ is a vector of unknown parameters, $\phi(\cdot)$ is an unknown and possibly nonlinear function defined over R^q, and the error process $\{e_t\}$ satisfies $E[e_t] = 0$, $0 < E[e_t^2] < \infty$ and some other mild conditions to be specified later. We propose a semiparametric penalty-function-based model selection criterion, by incorporating essential features of the CV1 selection method for the choice of both the parametric and the nonparametric regressors in model (1.1). Our contributions include proposing a new selection criterion, developing the associated theory, and demonstrating the key feature of easy implementation of the proposed semiparametric penalty function method through using simulated examples. The organization of this paper is as follows. Section 2 proposes the penalty–function–based selection criterion and develops its asymptotic theory.

Two simulated examples are given in Section 3. Assumptions and proofs are given in the appendix.

2. Semiparametric Penalty Function Method

First we introduce some notation. Let $A_p = \{1, \ldots, p\}$, $D_q = \{1, \ldots, q\}$, $\mathcal{A}$ denote all nonempty subsets of A_p, and $\mathcal{D}$ denote all nonempty subsets of D_q. For any subset $A \in \mathcal{A}, U_{tA}$ is the column vector consisting of $\{U_{ti}, i \in A\}$, and β_A is the column vector consisting of $\{\beta_i, i \in A\}$. For any subset $D \in \mathcal{D}, X_{tD}$ is the column vector consisting of $\{X_{ti}, i \in D\}$. We use $d_E = |E|$ to denote the cardinality of a set E. Let

$$\begin{aligned}
\mathcal{A}_1 &= \{A : A \in \mathcal{A} \ \text{ such that at least one nonzero component of } \beta \text{ is not in } \beta_A\}\\
\mathcal{A}_2 &= \{A : A \in \mathcal{A} \ \text{ such that } \beta_A \text{ contains all nonzero components of } \beta\}\\
\mathcal{D}_1 &= \{D : D \in \mathcal{D} \ \text{ such that } E[Y_t|X_{tD}] = E[Y_t|X_t]\}\\
\mathcal{D}_2 &= \{D : D \in \mathcal{D} \ \text{ such that } E[U_t^\tau \beta|X_{tD}] = E[U_t^\tau \beta|X_t]\}\\
\mathcal{B}_1 &= \{(A, D) : A \in \mathcal{A}_2 \ \text{ and } \ D \in \mathcal{D}_1 \cap \mathcal{D}_2\}.
\end{aligned}$$

Obviously, the subsets $A \in \mathcal{A}_1$ and $D \in \mathcal{D}_1^c = \mathcal{D} - \mathcal{D}_1$ correspond to incorrect models. The correct models correspond to $(A_0, D_0) \in \mathcal{B}_1$ such that both A_0 and D_0 are of the smallest dimension. To ensure the existence and uniqueness of such a pair (A_0, D_0), we need the same conditions as required in Gao and Tong (2004). For the sake of being self–contained, we recite them here.

Assumption 2.1. (i) $\Delta_{A,D} = E\{U_{tA} - E[U_{tA}|X_{tD}]\}\{U_{tA} - E[U_{tA}|X_{tD}]\}^\tau$ is a positive definite matrix of order $d_A \times d_A$ for each pair $A \in \mathcal{A}$ and $D \in \mathcal{D}$.

(ii) Let $\mathcal{B}_0 = \{(A_0, D_0) \in \mathcal{B}_1, \text{such that } |A_0| + |D_0| = \min_{(A,D)\in\mathcal{B}_1}[|A| + |D|]\}$. The pair (A_0, D_0) is the unique element of $\mathcal{B}_0$, denoted by (A_*, D_*).

Assumption 2.2. There is a unique pair (β_*, ϕ_*) such that the true and compact version of model (1.1) is

$$Y_t = U_{tA_*}^\tau \beta_* + \phi_*(X_{tD_*}) + e_t, \tag{2.1}$$

where $e_t = Y_t - E[Y_t|U_t, X_t]$.

Assumption 2.3. Define $\theta_j(X_{tj}) = E[\phi_*(X_{tD_*})|X_{tj}]$ for $j \in D_q - D_*$. There exists an absolute constant M_0 such that

$$\min_{j\in D_q-D_*} \min_{\alpha,\beta} E\left[\theta_j(X_{tj}) - \alpha - \beta X_{tj}\right]^2 \geq M_0. \tag{2.2}$$

Assumption 2.1 is a standard condition in this kind of problem. Assumption 2.2 is to ensure the existence and uniqueness of the true model. Assumption 2.3

is imposed to ensure that model (2.1) is identifiable and to exclude the case in which ϕ_* itself is a linear function in X_{tj} for $j \in D_q - D_*$. More details are available in Gao and Tong (2004).

For any given pair $A \in \mathcal{A}$ and $D \in \mathcal{D}$, we consider a partial linear model of the form

$$Y_t = U_{tA}^\tau \beta_A + \phi_D(X_{tD}) + e_t(A, D), \tag{2.3}$$

where $e_t(A, D) = Y_t - E[Y_t | U_{tA}, X_{tD}], \beta_A$ is as defined before, and $\phi_D(\cdot)$ is an unknown function on $\boldsymbol{R}^{|D|}$.

To use CV1, we introduce some notation to smooth and linearize the semiparametric model (2.3):

$$\begin{aligned}
&\hat{\phi}_{1t}(D) = \sum_{s=1, s\neq t}^{T} W_D^{(-t)}(t,s) Y_s, && \hat{\phi}_{2t}(A,D) = \sum_{s=1, s\neq t}^{T} W_D^{(-t)}(t,s) U_{sA},\\
&Z_t(D) = Y_t - \hat{\phi}_{1t}(D), && Z(D) = (Z_1(D), \ldots, Z_T(D))^\tau,\\
&V_t(A,D) = U_{tA} - \hat{\phi}_{2t}(A,D), && V(A,D) = (V_1(A,D), \ldots, V_T(A,D))^\tau,\\
&\phi_1(X_t) = E[Y_t|X_t], && \phi_2(X_t) = E[U_t|X_t],\\
&V_t = U_t - \phi_2(X_t), && V = (V_1, \ldots, V_T)^\tau,
\end{aligned} \tag{2.4}$$

where

$$W_D^{(-t)}(t,s) = \frac{K_D(\frac{X_{tD} - X_{sD}}{h})}{\sum_{l=1, l\neq t}^{T} K_D(\frac{X_{tD} - X_{lD}}{h})},$$

in which T is the number of observations, K_D is a multivariate kernel function defined on $\boldsymbol{R}^{|D|}$, and $h = h_D$ is a bandwidth parameter satisfying $h \in H_{TD} = [h_{\min}(T,D), h_{\max}(T,D)]$, where both $0 < h_{\min}(T,D), h_{\max}(T,D) \leq 1$ are suitable functions of (D,T) such that the optimal order $T^{-1/[4+|D|]}$ is included in H_{TD}. In the conventional cross–validation selection of an optimal bandwidth for the autoregressive case, the optimal bandwidth is proportional to $\sigma\ T^{-1/5}$ for example, where σ is the standard deviation of the data.

A technical advantage of using H_{TD} over previous assumptions of this type, such as an interval of the form $\left[c_1\ T^{-7/[6(4+|D|)]}, c_2\ T^{-5/[6(4+|D|)]}\right]$ with $0 < c_1 < c_2 < \infty$, as proposed in Gao and Yee (2000), is that the range of h under consideration has been extended considerably. This provides more security and theoretical underpinning for consideration of h both large and small, as initially discussed in Härdle, Hall and Marron (2002). In addition, Lemma A.1(i) in the appendix shows that an optimal choice of h depends mainly on the choice of D and T. This is why we do not involve h in the penalty function $\Lambda_T(A,D)$ to be introduced in Assumption 2.4 below.

We set up the linearized model

$$Z_t(D) = V_t(A, D)\beta_A + e_t(A, D), \tag{2.5}$$

and give the least squares estimator of β_A as

$$\hat{\beta}(A, D) = \Big(V(A, D)^\tau V(A, D)\Big)^+ V(A, D)^\tau Z(D), \tag{2.6}$$

where $(\cdot)^+$ is the Moore-Penrose inverse.

After fitting (2.5) to the data $\{(Y_t, U_t, X_t) : t = 1, \ldots, T\}$, the residual sum of squares is

$$\begin{aligned} RSS(A, D; h) &= \sum_{t=1}^{T} \Big[Z_t(D) - V_t(A, D)^\tau \hat{\beta}(A, D)\Big]^2 \\ &= \Big(Z(D) - V(A, D)\hat{\beta}(A, D)\Big)^\tau \Big(Z(D) - V(A, D)\hat{\beta}(A, D)\Big) \\ &= \varepsilon^\tau R(A, D)\varepsilon + \tilde{\Phi}(D)^\tau R(A, D)\tilde{\Phi}(D) + (V\beta)^\tau R(A, D)(V\beta) + \Delta_T(A, D; h), \end{aligned} \tag{2.7}$$

where $\varepsilon = (e_1, \ldots, e_T)^\tau$, $P(A, D) = V(A, D)(V(A, D)^\tau V(A, D))^+ V(A, D)^\tau$, $R(A, D) = I_T - P(A, D)$, $\tilde{\Phi}(D) = (\tilde{\phi}_1(D), \ldots, \tilde{\phi}_T(D))^\tau$, $\tilde{\phi}_t(D) = \phi_1(X_t) - \hat{\phi}_{1t}(D)$, I_T is the identity matrix of order $T \times T$, and the reminder term is

$$\Delta_T(A, D; h) = 2\varepsilon^\tau R(A, D)\tilde{\Phi}(D) + 2\tilde{\Phi}^\tau(D)R(A, D)(V\beta) + 2\varepsilon^\tau R(A, D)(V\beta). \tag{2.8}$$

It may be shown from (2.7) that the following equations hold uniformly in $h \in H_{TD}$:

$$\begin{aligned} RSS(A, D; h) &= \varepsilon^\tau R(A, D)\varepsilon + \tilde{\Phi}(D)^\tau R(A, D)\tilde{\Phi}(D) + (V\beta)^\tau R(A, D)(V\beta) \\ &\quad + o_P\left(RSS(A, D; h)\right), \\ M_T(A, D; h) &= E\left[RSS(A, D; h)\right] = (T - |A|)\sigma^2 + P_T(A, D) + N_T(A, D) \\ &\quad + o\left(M_T(A, D)\right), \end{aligned} \tag{2.9}$$

where $P_T(A, D) = E\left[(V\beta)^\tau R(A, D)(V\beta)\right]$ and $N_T(A, D) = E[\tilde{\Phi}(D)^\tau R(A, D) \tilde{\Phi}(D)]$. The equations in (2.9) reflect the errors incurred in model selection and estimation. The proof of (2.9) is relegated to Lemma A.3 in the appendix.

The penalty function $\Lambda_T(A, D)$ is now defined as

$$\Lambda_T(A, D) : \mathcal{A} \times \mathcal{D} \to \boldsymbol{R},$$

and satisfies the following.

Assumption 2.4. (i) Let $\Lambda_T(\emptyset, D) = 0$ for $D = \emptyset$ or any given $D \in \mathcal{D}$.

(ii) For any subsets $A_1, A_2 \in \mathcal{A}, D_1, D_2 \in \mathcal{D}$ satisfying $A_2 \supset A_1, D_2 \supset D_1$,

$$\lim_{T\to\infty} \inf \frac{\Lambda_T(A_1, D_1)}{\Lambda_T(A_2, D_2)} < 1.$$

(iii) For any nonempty sets A and D,

$$\lim_{T\to\infty} \Lambda_T(A, D) = \infty \quad \text{and} \quad \lim_{T\to\infty} \frac{\Lambda_T(A, D)}{T} = 0.$$

Remark 2.1. (i) It should be noted that $\Lambda_T(A, D)$ can be chosen quite generally. It is a function of sets in theory, but in practice we can define it as a function of $|A|$ and $|D|$ satisfying Assumption 2.4. Obviously, the definition of $\Lambda_T(A, D)$ generalizes the function $h_n(k)$ in Zheng and Loh (1997).

(ii) Assumption 2.4 regularizes the penalty function so as to avoid any problem of over-fitting or under-fitting.

We now extend the penalty function method from linear model selection to the partial linear model selection. Define

$$(\hat{A}, \hat{D}, \hat{h}) = \arg \min_{A\in\mathcal{A}, D\in\mathcal{D}, h\in H_{TD}} \{RSS(A, D; h) + \Lambda_T(A, D)\widehat{\sigma}^2\}, \tag{2.10}$$

where $\widehat{\sigma}^2 = RSS(A_p, D_q; h)/(T - p)$ is the usual consistent estimate of $\mathrm{Var}[e_t] = \sigma^2$. It may be shown from equation (14) of Zheng and Loh (1997) that

$$\widehat{\sigma}^2 = \sigma^2 + o_P(1) \tag{2.11}$$

uniformly in $h \in H_{TD}$. The proof of (2.11) is similar to, but simpler, than that of (2.9) and is therefore omitted.

Remark 2.2. The method proposed in this paper generalizes those in Zheng and Loh (1995, 1997), Yao and Tong (1994) and Vieu (1994). For example, if D_* is already identified, then the problem becomes a model selection problem for linear models as discussed in Zheng and Loh (1995, 1997). If A is already identified as A_*, and we need only to select D for (A_*, D), then the model selection reduces to a purely nonparametric leave–one–out cross–validation selection problem. This is because, as shown in Section 2.1 of Gao and Tong (2002), the leading term $RSS(A_*, D; h)/T$ is asymptotically equivalent to a $CV1(D, h)$ function of (D, h) defined as

$$\mathrm{CV1}(D, h) = \frac{1}{T}\sum_{t=1}^{T} \left\{Y_t - U_t^\tau \hat{\beta}(A_p, D) - \widehat{\phi}_t(X_{tD}, \hat{\beta}(A_p, D))\right\}^2, \tag{2.12}$$

where $\widehat{\phi}_t(X_{tD}, \beta) = \hat{\phi}_{1t}(D) - \hat{\phi}_{2t}(A, D)^\tau \beta$ and $\hat{\beta}(A, D)$ are given at (2.6). Thus, in the case where A is already identified as A_*, we may choose (D, h) as

$$(\hat{D}, \hat{h}) = \arg \min_{D\in\mathcal{D}, h\in H_{TD}} \mathrm{CV1}(D, h).$$

We now state the main result of this paper; its proof is relegated to the Appendix.

Theorem 2.1. *If the Assumptions* 2.1−2.4 *and* A.1−A.5 *of the Appendix hold, then*

$$\lim_{T\to\infty} P\left(\hat{A}=A_*,\hat{D}=D_*\right)=1 \quad \text{and} \quad \frac{\widehat{h}}{h_*}\to_P 1$$

as $T\to\infty$, *where* $h_*=c_*T^{-(4+|D_*|)^{-1}}$ *and* c_* *is a positive constant.*

3. Some Simulation Results

In this section we illustrate Theorem 2.1 using two simulated examples. Our simulation results support the asymptotic theory and the use of the semiparametric penalty function method for practical partial linear model selection.

Example 3.1. Consider a nonlinear time series model of the form

$$Y_t=0.47U_{t-1}-0.45U_{t-2}+\frac{0.5X_{t-1}}{1+X_{t-1}^2}+e_t,$$

where $U_t=0.55U_{t-1}-0.12U_{t-2}+\delta_t$ and $X_t=0.3\sin(2\pi X_{t-1})+\epsilon_t$, in which δ_t, ϵ_t and e_t are mutually independent and are identically distributed as uniform $(-1,1)$, uniform $(-0.5,0.5)$, and $N(0,1)$, respectively, U_1,U_2,X_1,X_2 are independent and identically distributed as uniform $(-1,1)$, U_s and X_t are mutually independent for all $s,t\geq 3$, and the processes $\{(\delta_t,\epsilon_t,e_t):t\geq 3\}$ are independent of both (U_1,U_2) and (X_1,X_2).

For Example 3.1, the strict stationarity and mixing condition can be justified by using existing results (e.g., Masry and Tjøstheim (1995, 1997)). Thus, Assumption A.1 holds. For an application of Theorem 2.1, let

$$\beta=(\beta_1,\beta_2)^\tau=(0.47,-0.45)^\tau \quad \text{and} \quad \phi(X_{t-1})=\frac{0.5X_{t-1}}{1+X_{t-1}^2}.$$

In this example, we consider the case where X_t and X_{t-1} are selected as candidate nonparametric regressors and U_{t-1} and U_{t-2} as candidate parametric regressors. Then we use (2.10) to check if $(U_{t-1},U_{t-2},X_{t-1})$ is the true set of semiparametric regressors. For this case, there are $2^2-1=3$ possible nonparametric regressors and $2^2-1=3$ possible parametric regressors, respectively. Thus, there are 9 posisble candidates for the true model, since U_t and X_t are independent. Let $D_0=\{1\},D_1=\{0\}$, $D_2=\{0,1\}$, $\mathcal{D}=\{D_i:0\leq i\leq 2\}$, $X_{tD_0}=X_{t-1}$, $X_{tD_1}=X_t$, $X_{tD_2}=(X_t,X_{t-1})^\tau,A_0=\{1,2\},A_1=\{1\},A_2=\{2\}$, $\mathcal{A}=\{A_i:i=0,1,2\}$, $U_{tA_0}=(U_{t-1},\ U_{t-2})^\tau$, $U_{tA_1}=U_{t-1}$, and $U_{tA_2}=U_{t-2}$. It follows that both $D_*=D_0$ and $A_*=A_0$ are unique. Assumptions 2.1−2.3 therefore hold.

We use $\Lambda_T(A,D) = (|A| + |D|) \cdot T^{0.5}$ as the penalty function. Thus, Assumption 2.4 holds automatically. In addition to this choice of $\Lambda_T(A,D)$, we also considered several other forms for $\Lambda_T(A,D)$; as the resulting simulated frequencies are very similar, we do not report them here. Throughout Example 3.1, we use

$$h \in H_{TD_0} = \left[c_1\ T^{-\frac{7}{6(4+|D_0|)}}, c_2\ T^{-\frac{5}{6(4+|D_0|)}}\right] = \left[T^{-\frac{7}{30}},\ 2 \cdot T^{-\frac{1}{6}}\right],$$

based on the method proposed in Gao and Yee (2000) with $c_1 = 1$ and $c_2 = 2$.

For the multivariate kernel function $K(\cdot)$ involved in $W_D(t,s)$, define $k(u) = (\sqrt{2\pi})^{-1}e^{-u^2/2}$ and $K(u_1,\ldots,u_j) = \Pi_{i=1}^{j} k(u_j)$ for $j = 1,2,3$.

It follows that Assumptions A.1−A.4 are all satisfied. Now we prove that Assumption A.5 is also satisfied. Assumption A.5 (i) is checked in Gao and Tong (2004). The independence between U_t and X_t implies that $E[U_{tA_i}|X_{tD_j}] = E[U_{tA_i}]$ for all $i,j = 0,1,2$. Thus, we need only to introduce the following notations. For $i = 0,1,2$, let $\eta_t(A_i) = U_{tA_i} - E[U_{tA_i}]$, $\eta(A_i) = (\eta_1(A_i),\ldots,\eta_T(A_i))^\tau$, $\eta_t = \eta_t(A_0)$, and $\eta = (\eta_1,\ldots,\eta_T)^\tau$.

Let $\eta\beta = (\alpha_1,\ldots,\alpha_T)^\tau$, where $\alpha_t = \eta_t(A_1)\beta_1 + \eta_t(A_2)\beta_2$. A calculation yields that

$$\begin{aligned}(\eta\beta)^\tau[I - Q(A,D)](\eta\beta) &= (\eta\beta)^\tau \left(I - \eta(A_i)(\eta(A_i)^\tau\eta(A_i))^{-1}\eta(A_i)^\tau\right)(\eta\beta) \\ &= \frac{\sum_{t=3}^T \eta_t^2(A_i)\sum_{t=3}^T \alpha_t^2 - \left[\sum_{t=3}^T \eta_t(A_i)\alpha_t\right]^2}{\sum_{t=3}^T \eta_t^2(A_i)} > 0\end{aligned}$$

with probability one for all $i = 1,2$, because $P(\eta_t(A_i) = \alpha_t) = 0$. Thus Assumption A.5 holds.

In order to assess both the small and medium sample properties of our theory, we took $T = 52$ and $T = 552$. For the $T = 52, 127, 272$ and 552, Table 3.1 gives the relative frequencies of the selected parametric and nonparametric regressors in $1,000$ replications.

Table 3.1. Frequencies of semiparametric model selection.

Parametric and Nonparametric	Frequencies			
subset	$T = 52$	$T = 127$	$T = 272$	$T = 552$
$\{U_{t-2}, U_{t-1}, X_{t-1}\}$	0.609	0.746	0.873	0.971
$\{U_{t-2}, U_{t-1}, X_t\}$	0.350	0.235	0.123	0.028
$\{U_{t-2}, X_{t-1}, X_t\}$	0.024	0.014	0.002	0.001
$\{U_{t-1}, X_{t-1}, X_t\}$	0.017	0.005	0.002	0.000

Remark 3.1. As can be seen from Table 3.1, the true set of regressors $\{U_{t-2}, U_{t-1}, X_{t-1}\}$ is selected with increasing frequencies from 0.609 to 0.971 as the

sample size increases from $T = 52$ to $T = 552$. The model $\{U_{t-2}, U_{t-1}, X_t\}$, one of the closest to the true model, is selected with frequencies decreasing from 0.350 to 0.028. This lends support to the efficacy of combining the penalty function method with the leave–one–out cross validation (CV1).

Table 3.1 shows that the proposed semiparametric model selection method works well numerically when the true model is a partial linear model. Tables 3.2 and 3.3 show that the proposed model selection method is much more effective than existing model selection methods by comparing it with the penalty function method for linear models proposed by Zheng and Loh, as well as with the conventional nonparametric leave–one–out cross–validation function CV1. The candidate variables are still $\{U_{t-2}, U_{t-1}, X_{t-1}, X_t\}$. Both the penalty function method for linear model selection and the CV1 selection procedure consider all possible 15 models. As many insignificant regressors have tiny probabilities of being selected, Tables 3.2 and 3.3 below provide only the relevant frequencies for the significant regressors.

Table 3.2. Frequencies of parametric model selection.

Parametric	Frequencies			
subset	$T = 52$	$T = 127$	$T = 272$	$T = 552$
$\{U_{t-2}, U_{t-1}, X_{t-1}\}$	0.055	0.142	0.383	0.684
$\{U_{t-2}, U_{t-1}\}$	0.238	0.464	0.525	0.310
$\{U_{t-1}, U_{t-2}, X_t\}$	0.024	0.013	0.009	0.003
$\{U_{t-2}\}$	0.180	0.096	0.015	0.000
$\{U_{t-1}\}$	0.193	0.112	0.020	0.000
$\{X_{t-1}\}$	0.178	0.101	0.025	0.000
$\{X_t\}$	0.067	0.024	0.001	0.000
$\{U_{t-2}, X_{t-1}\}$	0.019	0.014	0.007	0.000
$\{U_{t-1}, X_{t-1}\}$	0.022	0.024	0.008	0.000
$\{U_{t-2}, U_{t-1}, X_{t-1}, X_t\}$	0.003	0.004	0.005	0.003

Table 3.3. Frequencies of nonparametric model selection.

Nonparametric	Frequencies			
subset	$T = 52$	$T = 127$	$T = 272$	$T = 552$
$\{U_{t-2}, U_{t-1}, X_{t-1}\}$	0.103	0.288	0.464	0.652
$\{U_{t-2}, U_{t-1}, X_{t-1}, X_t\}$	0.050	0.103	0.184	0.196
$\{U_{t-2}, U_{t-1}\}$	0.135	0.205	0.194	0.117
$\{U_{t-2}\}$	0.064	0.022	0.002	0.000
$\{U_{t-1}\}$	0.102	0.035	0.004	0.000
$\{X_{t-1}\}$	0.102	0.048	0.008	0.000
$\{X_t\}$	0.059	0.011	0.000	0.000
$\{U_{t-2}, X_{t-1}\}$	0.053	0.028	0.009	0.000
$\{U_{t-1}, X_{t-1}\}$	0.058	0.060	0.014	0.002
$\{U_{t-1}, U_{t-2}, X_t\}$	0.038	0.019	0.092	0.001

Remark 3.2. Tables 3.2 and 3.3 show that the penalty function method for linear models and the conventional nonparametric CV1 method have highest frequencies that the true model is selected of 0.383 and 0.464, respectively, when the sample size is $T \leq 272$. Although their performance improves when the sample size increases to 552, there is still a huge difference between their performance and the performance here. Our findings justify the new efficient selection method for problems that cannot be solved using existing selection methods for either completely linear models or fully nonparametric models.

The above simulations are based on the assumption that the true model is a partial linear model, for which our method is designed. If the true model is either a parametric model or a fully nonparametric model, our method performs reasonably well. Example 3.2 below considers the case where the true model is a parametric linear model and then applies both the parametric selection proposed by Zheng and Loh (1995) and our own semiparametric selection procedure. When using the proposed semiparametric selection method, our preliminary computation suggests involving the same kernel function and bandwidth interval as used in Example 3.1 for the simulation in Example 3.2 below.

Example 3.2. Consider a linear time series model of the form $Y_t = 0.47U_{t-1} - 0.45U_{t-2} + 0.5X_{t-1} + e_t$, where $U_t = 0.55U_{t-1} - 0.12U_{t-2} + \delta_t$ and $X_t = 0.3 \sin(2\pi X_{t-1}) + \epsilon_t$, in which δ_t, ϵ_t and e_t are mutually independent and identically distributed as uniform $(-1, 1)$, uniform $(-0.5, 0.5)$, and $N(0, 1)$, respectively, U_1, U_2, X_1, X_2 are independent and identically distributed as uniform $(-1, 1)$, U_s and X_t are mutually independent for all $s, t \geq 3$, and the processes $\{(\delta_t, \epsilon_t, e_t) : t \geq 3\}$ are independent of both (U_1, U_2) and (X_1, X_2).

For $T = 52, 127, 272$ and 552, we chose the penalty function $\Lambda_T(A, D) = (|A| + |D|)\, T^{0.5}$ and then calculated the relative frequencies of the selected parametric and semiparametric regressors in $1,000$ replications. The results are in Tables 3.4 and 3.5.

Table 3.4. Frequencies of semiparametric model selection.

Parametric and Nonparametric	Frequencies			
subset	$T = 52$	$T = 127$	$T = 272$	$T = 552$
$\{U_{t-2}, U_{t-1}, X_{t-1}\}$	0.070	0.172	0.432	0.836
$\{U_{t-2}, U_{t-1}, X_t\}$	0.038	0.040	0.028	0.009
$\{U_{t-2}, X_{t-1}, X_t\}$	0.002	0.004	0.000	0.000
$\{U_{t-1}, X_{t-1}, X_t\}$	0.001	0.000	0.000	0.000
$\{U_{t-1}, U_{t-2}, X_{t-1}, X_t\}$	0.001	0.001	0.001	0.000
$\{U_{t-2}, X_{t-1}\}$	0.280	0.293	0.208	0.051
$\{U_{t-2}, X_t\}$	0.147	0.059	0.007	0.000
$\{U_{t-1}, X_{t-1}\}$	0.310	0.369	0.308	0.104
$\{U_{t-1}, X_t\}$	0.151	0.062	0.016	0.000

Table 3.5. Frequencies of parametric model selection.

Parametric	Frequencies			
subset	$T = 52$	$T = 127$	$T = 272$	$T = 552$
$\{U_{t-2}, U_{t-1}, X_{t-1}\}$	0.086	0.325	0.742	0.942
$\{U_{t-2}, U_{t-1}\}$	0.213	0.327	0.198	0.035
$\{U_{t-1}, U_{t-2}, X_t\}$	0.032	0.014	0.006	0.003
$\{U_{t-2}\}$	0.129	0.041	0.003	0.000
$\{U_{t-1}\}$	0.159	0.072	0.004	0.000
$\{X_{t-1}\}$	0.192	0.121	0.012	0.000
$\{X_t\}$	0.073	0.006	0.001	0.000
$\{U_{t-2}, X_{t-1}\}$	0.040	0.019	0.001	0.000
$\{U_{t-1}, X_{t-1}\}$	0.028	0.038	0.011	0.000
$\{X_{t-1}, X_t\}$	0.013	0.003	0.000	0.000
$\{U_{t-2}, U_{t-1}, X_{t-1}, X_t\}$	0.011	0.018	0.022	0.020

Remark 3.3. Tables 3.2 and 3.4 show that semiparametric penalty function method has a similar performance to that of the parametric penalty function method for the cases of $T = 52, T = 127$ and $T = 272$. When $T = 552$, the semiparametric penalty function method does better.

Acknowledgements

The authors would like to thank the Co–Editors, the Guest Editors, and the referees for their constructive comments and suggestions on earlier versions of this paper. Thanks also go to the Australian Research Council Discovery Grants Program for its financial support.

Appendix

There are several basic assumptions stated here. Throughout this appendix, let $C(0 < C < \infty)$ denote a constant which may have different values at each appearance.

Assumption A.1. Assume that the stochastic process (Y_t, U_t, X_t) is strictly stationary and α–mixing with mixing coefficient $\alpha(T) \leq C\eta^T$, where $0 < \eta < 1$ is a constant. In addition, assume $\{e_t\}$ is a stationary martingale difference with respect to $\Omega_t = \sigma\{(Y_s, U_{s+1}, X_{s+1}) : 1 \leq s \leq t-1\}$, a sequence of σ-fields generated by $\{(Y_s, U_{s+1}, X_{s+1}) : 1 \leq s \leq t-1\}$. Suppose $P(E[e_t^2|\Omega_t] = \sigma^2) = 1$, where $0 < \sigma^2 = E[e_t^2] < \infty$.

Assumption A.2. For every $D \in \mathcal{D}$, K_D is a $|D|$-dimensional symmetric, Lipschitz continuous probability kernel function with $\int ||u||^2 K_D(u)du < \infty$, and K_D has an absolutely integrable Fourier transform, where $|| \cdot ||$ denote the Euclidean norm.

Assumption A.3. Let S_w be a compact subset of $\boldsymbol{R}^q$, $w(x)$ be a weight function supported on S_w with $0 < w(x) \leq C$ for some constant C. For every $D \in \mathcal{D}$, let $\boldsymbol{R}_{X,D} \subset \boldsymbol{R}^{|D|} = (-\infty, \infty)^{|D|}$ be the subset such that $X_{tD} \in \boldsymbol{R}_{X,D}$ and let S_D be the projection of S_w in $\boldsymbol{R}_{X,D}$ (that is, $S_D = \boldsymbol{R}_{X,D} \cap S_w$). Assume that the marginal density function, $f_D(\cdot)$, of X_{tD}, and the first two derivatives of $f_D(x), \varphi_1(x; D)$ and $\varphi_2(x; A, D)$ are all continuous on $x \in \boldsymbol{R}_{X,D}$, and on S_D the density function $f_D(x)$ is bounded below by C_D and above by C_D^{-1} for some $C_D > 0$, where $\varphi_1(x; D) = E[Y_t | X_{tD} = x]$ and $\varphi_2(x; A, D) = E[U_{tA} | X_{tD} = x]$ for every $A \in \mathcal{A}$ and $D \in \mathcal{D}$.

Assumption A.4. There exists constants $0 < C_1, C_2 < \infty$ such that for any integer $l \geq 1$,

$$\sup_x \sup_{A \in \mathcal{A}, D \in \mathcal{D}} E\left(|Y_t - E[Y_t|(U_{tA}, X_{tD})]|^l | X_{tD} = x \right) \leq C_1,$$

$$\sup_x \sup_{A \in \mathcal{A}, D \in \mathcal{D}} E\left(\|U_{tA}\|^l | X_{tD} = x \right) \leq C_2.$$

Assumption A.5. For

$$\begin{aligned} &\eta_t(A, D) = U_{tA} - E[U_{tA}|X_{tD}], \quad \eta(A, D) = (\eta_1(A, D), \ldots, \eta_T(A, D))^\tau, \\ &\eta_t = U_t - E[U_t|X_t], \quad \eta = (\eta_1, \ldots, \eta_T)^\tau, \\ &Q(A, D) = \eta(A, D) \left(\eta(A, D)^\tau \eta(A, D) \right)^+ \eta(A, D)^\tau, \end{aligned}$$

and $P_{1T}(A, D) = T^{-1}(\eta\beta)^\tau [I_T - Q(A, D)](\eta\beta)$, assume that for any given $A \in \mathcal{A}_1$ and $D \in \mathcal{D}$, $\liminf_{T\to\infty} P_{1T}(A, D) > 0$ in probability.

Assumptions A.1−A.5 are standard conditions in this kind of problem. Remark A.1 of Gao and Tong (2002) gives detailed justification for Assumptions A.1−A.5.

To prove Theorem 2.1, the following lemmas are required. Similar to Lemma B.1 of Gao and Tong (2004) and Lemma B.3 of Gao and Tong (2002), we have the Lemmas A.1 and A.2. In addition, we include Lemma A.3 to ensure that (2.9) holds.

Lemma A.1 (i) *Assume that the conditions of Theorem* 2.1 *hold. If* $A \in \mathcal{A}_1$ *and* $D \in \mathcal{D}$*, then there exists* $R_{1T} \geq 0$ *such that*

$$RSS(A, D; h) = \sum_{t=1}^T e_t^2 + T \cdot P_{1T}(A, D) + T \cdot N_{1T}(D, h) + R_{1T} + o_P(T) \qquad \text{(A.1)}$$

uniformly in $h \in H_{TD}$*, where* R_{1T} *is independent of* (A, D)*,* P_{1T} *is as defined in Assumption* A.5*, and*

$$N_{1T}(D,h) = \begin{cases} c_1(D)\frac{1}{Th^{|D|}} + c_2(D)h^4 + o_p\left(\frac{1}{Th^{|D|}}\right) + o_p(h^4) & \text{if } D \in \mathcal{D}_1 \text{ and } h \in H_{TD}, \\ E\{E[Y_t|X_{tD}] - E[Y_t|X_t]\}^2 + o(1) & \text{if } D \notin \mathcal{D}_1 \text{ and } h \in H_{TD}. \end{cases}$$

Here both $c_1(D)$ and $c_2(D)$ are positive constants depending on $D \in \mathcal{D}_1$.

(ii) *Assume that the conditions of Theorem* 2.1 *hold. If $A \in \mathcal{A}_2$ and $D \in \mathcal{D}$, then*

$$RSS(A, D; h) = \sum_{t=1}^{T} e_t^2 + d_A \sigma^2 + T \cdot N_{1T}(D, h) + o_p(1) \tag{A.2}$$

uniformly in $h \in H_{TD}$, where $N_{1T}(D, h)$ is defined as above.

Similar to the proof of Lemma 1(a) of Gao and Tong (2004), it may be shown that for each given D, there exist some positive constants $d_1(D)$ and $d_2(D)$ such that $\bar{h}_D = d_1(D) T^{-(4+|D|)^{-1}}$ and

$$N_{1T}(D, \bar{h}_D) = \min_{h \in H_{TD}} N_{1T}(D, h) = \begin{cases} d_2(D) T^{-\frac{4}{4+|D|}} + o_p\left(T^{-\frac{4}{4+|D|}}\right), & \text{if } D \in \mathcal{D}_1, \\ E\{E[Y_t|X_{tD}] - E[Y_t|X_t]\}^2 + o_p(1), & \text{if } D \notin \mathcal{D}_1. \end{cases} \tag{A.3}$$

Lemma A.2. *Assume that the conditions of Theorem* 2.1 *hold. If $A \in \mathcal{A}_1$ and $D \in \mathcal{D}$ then*

$$\liminf_{T \to \infty} Q_{1T}(A, D) = \liminf_{T \to \infty} P_{1T}(A, D) > 0 \ \text{ in probability,} \tag{A.4}$$

where $Q_{1T}(A, D) = T^{-1}(V\beta)^\tau R(A, D)(V\beta)$, $P_{1T}(A, D)$ and $R(A, D)$ are defined as before.

Lemma A.3. *Assume that the conditions of Theorem* 2.1 *hold. Then, uniformly in $h \in H_{TD}$,*

$$\begin{aligned} RSS(A, D; h) &= \varepsilon^\tau R(A, D)\varepsilon + \tilde{\Phi}(D)^\tau R(A, D)\tilde{\Phi}(D) + (V\beta)^\tau R(A, D)(V\beta) \\ &\quad + o_P\left(RSS(A, D; h)\right), & \text{(A.5)} \\ M_T(A, D; h) &= E\left[RSS(A, D; h)\right] = (T - |A|)\sigma^2 + P_T(A, D) + N_T(A, D) \\ &\quad + o\left(M_T(A, D; h)\right). & \text{(A.6)} \end{aligned}$$

Proof. Write

$$\Delta_T(A, D; h) = 2\varepsilon^\tau R(A, D)\tilde{\Phi}(D) + 2\tilde{\Phi}^\tau(D) R(A, D)(V\beta) + 2\varepsilon^\tau R(A, D)(V\beta). \tag{A.7}$$

In order to prove (A.5), it suffices to show that for sufficiently large T

$$\sup_{h \in H_{TD}} \frac{|\Delta_T(A, D; h)|}{RSS(A, D; h)} = o_P(1), \tag{A.8}$$

which follows from

$$\sup_{h\in H_{TD}} \frac{\left|\varepsilon^\tau R(A,D)\tilde{\Phi}(D)\right|}{RSS(A,D;h)} = o_P(1), \quad \sup_{h\in H_{TD}} \frac{\left|\tilde{\Phi}^\tau(D)R(A,D)(V\beta)\right|}{RSS(A,D;h)} = o_P(1), \tag{A.9}$$

$$\sup_{h\in H_{TD}} \frac{|\varepsilon^\tau R(A,D)(V\beta)|}{RSS(A,D;h)} = o_P(1). \tag{A.10}$$

The proofs of (A.9) and (A.10) are quite standard in this kind of problem. The details are as in the proof of (A.18) of Gao and Yee (2000) for example, and are omitted here. The proof of (A.6) follows from (A.5) using the Dominated Convergence Theorem.

Proof of Theorem 2.1. Let $RSS(A,D) = \min_{h\in H_{TD}} RSS(A,D;h)$. In view of (A.1)−(A.3), we can write $RSS(A,D)$ as

$$RSS(A,D) = \begin{cases} \sum_{t=1}^T e_t^2 + T\cdot P_{1T}(A,D) + T\cdot N_{1T}(D,\bar{h}_D) + R_{1T} + o_p(T) & A\in\mathcal{A}_1, D\in\mathcal{D}, \\ \sum_{t=1}^T e_t^2 + d_A\sigma^2 + T\cdot N_{1T}(D,\bar{h}_D) + o_p(1) & A\in\mathcal{A}_2, D\in\mathcal{D}. \end{cases}$$

It follows immediately from $P_{1T}(A_*,D_*)=0$ that

$$\begin{aligned} &RSS(A,D) - RSS(A_*,D_*) \\ &= \begin{cases} T\cdot P_{1T}(A,D) - d_{A_*}\sigma^2 + T(N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*})) + o_p(T) & A\in\mathcal{A}_1, D\in\mathcal{D}, \\ (d_A - d_{A_*})\sigma^2 + T(N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*})) + o_p(1) & A\in\mathcal{A}_2, D\in\mathcal{D}. \end{cases} \end{aligned}$$

If $A\in\mathcal{A}_2$ and $D\in\mathcal{D}$, then as $T\to\infty$,

$$\begin{aligned} &1 - P\left\{RSS(A,D) - RSS(A_*,D_*) + (\Lambda_T(A,D) - \Lambda_T(A_*,D_*))\widehat{\sigma}^2 > 0\right\} \\ &= P\Big\{(d_A - d_{A_*})\sigma^2 + T(N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*})) + o_p(1) \\ &\quad +(\Lambda_T(A,D) - \Lambda_T(A_*,D_*))\widehat{\sigma}^2 \le 0\Big\} \\ &= P\left\{N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*}) \le -\frac{d_A - d_{A_*}}{T}\sigma^2 - \frac{\Lambda_T(A,D) - \Lambda_T(A_*,D_*)}{T}\widehat{\sigma}^2\right\} \\ &\le P\Big\{N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*}) \\ &\qquad \le -\frac{d_A - d_{A_*}}{T}\sigma^2 - \frac{\Lambda_T(A,D)}{T}\left(1 - \frac{\Lambda_T(A_*,D_*)}{\Lambda_T(A,D)}\right)\cdot\frac{1}{2}\sigma^2\Big\} + o(1) \\ &\le P\left\{N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*}) \le -\frac{\Lambda_T(A,D)}{T}\left(1 - \frac{\Lambda_T(A_*,D_*)}{\Lambda_T(A,D)}\right)\cdot\frac{1}{2}\sigma^2\right\} \\ &\quad +o(1) \to 0 \end{aligned} \tag{A.11}$$

because $N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*}) > 0$ for all $A \in \mathcal{A}_1$ and either $D \in \mathcal{D}_1$ or $D \in \mathcal{D} - \mathcal{D}_1$.

If $A \in \mathcal{A}_1$ and $D \in \mathcal{D}$, then we obtain as $T \to \infty$,

$$\begin{aligned}
1 - P&\left\{RSS(A,D) - RSS(A_*,D_*) + (\Lambda_T(A,D) - \Lambda_T(A_*,D_*))\widehat{\sigma}^2 > 0\right\} \\
&= P\Big\{P_{1T}(A,D) + \big(N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*})\big) \\
&\qquad\qquad + \frac{R_{1T} - d_{A_*}\sigma^2 + (\Lambda_T(A,D) - \Lambda_T(A_*,D_*))\,\widehat{\sigma}^2}{T} + o_P(1) \le 0\Big\} \\
&= P\Big\{P_{1T}(A,D) + N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*}) \\
&\qquad\qquad \le -\frac{o_p(T)}{T} + \frac{d_{A_*}\sigma^2}{T} - \frac{R_{1T}}{T} - \frac{(\Lambda_T(A,D) - \Lambda_T(A_*,D_*))\,\widehat{\sigma}^2}{T}\Big\} + o(1) \\
&= P\Big\{P_{1T}(A,D) + N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*}) \\
&\qquad\qquad \le -\frac{\Lambda_T(A,D)}{T}\left(1 - \frac{\Lambda_T(A_*,D_*)}{\Lambda_T(A,D)}\right) \cdot \frac{1}{2}\sigma^2\Big\} + o(1) \to 0 \qquad \text{(A.12)}
\end{aligned}$$

because $\liminf_{T\to\infty} P_{1T}(A,D) > 0$ and $N_{1T}(D,\bar{h}_D) - N_{1T}(D_*,\bar{h}_{D_*}) > 0$ for all $D \in \mathcal{D}_1$ or $D \in \mathcal{D} - \mathcal{D}_1$.

Consequently, as $T \to \infty$,

$$\begin{aligned}
1 &\ge P(\hat{A} = A_*, \hat{D} = D_*) \\
&\ge P\left\{RSS(A,D) - RSS(A_*,D_*) + (\Lambda_T(A,D) - \Lambda_T(A_*,D_*))\widehat{\sigma}^2 > 0\right\} \to 1.
\end{aligned}$$

This completes the proof of the first part of Theorem 2.1. Now if $\hat{h} = \bar{h}_{\hat{D}}$, $c_* = c_{D_*}$ and $h_* = \bar{h}_{D_*} = c_* T^{-(4+|D_*|)^{-1}}$, it follows from the above proof and (A.3) that $\hat{h}/h_* \to_p 1$ as $T \to \infty$. This completes the proofs.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2*nd International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csáki), 267-281. Akadémiai Kiado, Budapest.

Cheng, B. and Tong, H. (1992). On consistent nonparametric order determination and chaos. *J. Roy. Statist. Soc. Ser. B* **54**, 427-449.

Gao, J. and Tong, H. (2002). Nonparametric and semiparametric regression model selection. Working paper available from www.maths.uwa.edu.au/˜jiti/kao43.pdf.

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

Gao, J. and Yee, T. (2000). Adaptive estimation in partially linear (semiparametric) autoregressive models. *Canad. J. Statist.* **28**, 571-586.

Härdle, W., Hall, P. and Marron, J. (2002). Regression smoothing parameters that are not far from their optimum. *J. Amer. Statist. Assoc.* **87**, 227-233.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

Masry, E. and Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series. *Econom. Theory* **11**, 258-289.

Masry, E. and Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econom. Theory* **13**, 214-252.

Shao, J. (1993). Linear model selection by cross–validation. *J. Amer. Statist. Assoc.* **422**, 486-494.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117-126.

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

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

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

Zheng, X. and Loh, W. Y. (1995). Consistent variable selection in linear models. *J. Amer. Statist. Assoc.* **90**, 151-156.

Zheng, X. and Loh, W. Y. (1997). A consistent variable selection criterion for linear models with high–dimensional covariates. *Statist. Sinica* **7**, 311-325.

Department of Applied Mathematics, Shanxi University of Economics and Finance, P. R. China.

E-mail: dchaohua@hotmail.com

School of Mathematics and Statistics, The University of Western Australia, Australia.

E-mail: jiti.gao@uwa.edu.au

Department of Statistics, London School of Economics, London, England.

E-mail: h.tong@lse.ac.uk

(Received February 2006; accepted August 2006)