



Munich Personal RePEc Archive

Nonparametric and semiparametric regression model selection

Gao, Jiti and Tong, Howell

The University of Adelaide, London School of Economics

May 2002

Online at <https://mpra.ub.uni-muenchen.de/11987/>
MPRA Paper No. 11987, posted 09 Dec 2008 00:06 UTC

NONPARAMETRIC AND SEMIPARAMETRIC REGRESSION MODEL SELECTION ¹

BY JITI GAO² AND HOWELL TONG

*The University of Western Australia; The London School of Economics and The
University of Hong Kong*

ABSTRACT: It is known that semiparametric time series regression is often used without checking its suitability and compactness. In theory, this may result in dealing with an unnecessarily complicated model. In practice, one may encounter the computational difficulty caused by the sparseness of the data. This is partly because the curse of dimensionality problem may still arise from using a semiparametric time series regression model. This paper suggests that in order to provide more precise predictions we need to choose the most significant regressors for both the parametric and nonparametric time series components. We develop a novel cross-validation based model selection procedure for the choice of both the parametric and nonparametric time series components in semiparametric time series regression, and then establish some asymptotic properties of the proposed model selection procedure. In addition, we demonstrate how to implement the model selection procedure in practice through using both simulated and real examples. Our empirical studies show that the proposed cross-validation selection procedure works well numerically.

1. Introduction

In modelling nonlinear time series data one of the tasks is to study the structural relationship between the present observation and the history of the data set. The problem then is to fit a high dimensional surface to a nonlinear time series data set. Since the publication of Tong (1990), nonparametric techniques have been used extensively to model nonlinear time series data (see the two review papers: Tjøstheim 1994; Härdle,

¹*Key words and phrases:* Linear model, model selection, mixing process, nonlinear time series, nonparametric regression, semiparametric regression, strictly stationary process, variable selection.

AMS 1991 subject classifications. Primary 62G07; secondary 62G05.

²*Address for correspondence:* Dr Jiti Gao, Department of Mathematics and Statistics, The University of Western Australia, Crawley WA 6009, Australia. Email: jiti@maths.uwa.edu.au

Lütkepohl and Chen 1997; Chapter 6 of Fan and Gijbels 1996; and the references included in both the papers and the book). Although nonparametric techniques appear to be feasible, there is a serious problem: the so-called curse of dimensionality. For the independent and identically distributed (i.i.d.) case, this problem has been discussed and illustrated in several monographs and many papers (see Fan and Gijbels 1996 for example). In order to deal with the curse of dimensionality problem for the time series case, several nonparametric and semiparametric approaches have been proposed. These include: (i) nonparametric time series single-index and projection pursuit modelling (see Xia, Tong, Li and Zhu 2002 for the time series case); (ii) additive nonparametric time series modelling (see Chen and Tsay 1993; Masry and Tjøstheim 1995, 1997; Gao, Tong and Wolff 2002a, 2002b for the time series case); (iii) semiparametric time series modelling (see Gao 1998; Chapter 6 of Härdle, Liang and Hua 2000 for example); and (iv) nonparametric time series variable selection (see Cheng and Tong 1992, 1993; Tjøstheim and Auestad 1994a, 1994b; Yao and Tong 1994; Tjøstheim 1999; Gao, Wolff and Anh 2001, and others).

In theory, one may suggest using one of the methods to deal with the dimensionality reduction problem. In practice, however, one needs to check whether the method used is appropriate for a given set of data before using the method. For example, before applying additive nonparametric time series regression modelling, a crucial problem is whether an additive nonparametric time series model is appropriate for a given set of time series data. In other words, we should test for nonparametric additivity before using an additive nonparametric time series model to fit a given set of time series data (see Chen, Liu and Tsay 1995; Gao, Tong and Wolff 2002b for the time series case). If an additive nonparametric time series model is appropriate (once the additivity is tested and not rejected) for a high-dimensional time series data, then it can be entertained. Otherwise, we could check whether a partially linear time series model is appropriate. Although partially linear time series modelling may not be capable of reducing the nonparametric time series regression into a sum of one-dimensional nonparametric functions of individual lags, they can reduce the dimensionality significantly for some cases. Moreover, a feature of partially linear time series modelling is that it takes the true structure of the time series data into account and avoids neglecting some existing information on the

linearity of the data. This paper then suggests combining semiparametric time series modelling and nonparametric time series variable selection together to deal with the dimensionality reduction problem. We assume that a time series data set (Y_t, U_t, X_t) satisfies a partially linear time series model of the form

$$Y_t = U_t^\tau \beta + \phi(X_t) + e_t, \quad (1.1)$$

where $U_t = (U_{t1}, \dots, U_{tp})^\tau$ and $X_t = (X_{t1}, \dots, X_{tq})^\tau$ are both time series, U_t and X_t may be two different time series, $\beta = (\beta_1, \dots, \beta_p)^\tau$ 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. In model (1.1), the linear time series component is $U_t^\tau \beta$ and $\phi(X_t)$ is called the nonparametric time series component.

Model (1.1) covers some existing nonlinear time series cases. See for example, Robinson (1988), Teräsvirta, Tjøstheim and Granger (1994), Gao and Liang (1995), Gao (1998), Gao and Yee (2000), and others. In theory, model (1.1) can be used to overcome the dimensionality problem. In practice, however, model (1.1) itself may still suffer from the "curse of dimensionality". Thus, before using model (1.1) one needs to consider a model selection problem. In other words, we need to determine whether both the linear component and the nonparametric component are of the smallest possible dimensions. For the partially linear model case, the conventional nonparametric cross-validation model selection function simply cannot take the given linear component into account but treats each linear regressor as a nonparametric regressor. As a result, the conventional nonparametric cross-validation model selection function could neglect existing information about the linear component and therefore cause model misspecification problem. Hence, we need to consider an extension of existing parametric and nonparametric cross-validation model selection criteria to the semiparametric time series setting. Recently, Gao and Tong (2004) develop a simultaneous semiparametric leave-more-out cross-validation selection method for the optimum choice of both U_t and X_t . As observed in the simulations in Section 3.1 of Gao and Tong (2004), the number of observations used to fit the model is, however, quite small (with $T_c = 69$ in Section 3.1 for the semiparametric case), while the number of observations used to validate the proposed method

is relatively large ($T_v = 219$, respectively). This may impede the implementation of the semiparametric-based selection method in practice because the theory requires $T_c \rightarrow \infty$.

In order to avoid using more data for model validation, this paper develops a novel model selection procedure combining the leave-one-out cross-validation (abbreviated as CV1) function for the choice of the nonparametric regressors and the leave- T_v -out cross-validation (abbreviated as CV T_v) function for the choice of parametric regressors, where $T_v > 1$ is a positive interger satisfying $T_v \rightarrow \infty$ as the number of observations, T , converges to ∞ . Our proposed semiparametric cross-validation (CV) based time series model selection procedure has the following features:

(i) It provides a general model selection procedure in determining asymptotically whether both the linear time series component and the nonparametric time series component are of the smallest possible dimensions. The procedure can select the true form of the linear time series component. Moreover, it could overcome the difficulty known as the "curse of dimensionality" arising from using nonparametric techniques to estimate the nonparametric time series component in (1.1).

(ii) It extends the leave- T_v -out cross-validation (CV) selection criterion for classical linear regression (see Shao 1993; Zhang 1993) and the leave-one-out cross-validation selection criterion (see Vieu 1994; and Yao and Tong 1994) for purely nonparametric regression to the semiparametric time series setting.

(iii) It is applicable to a wide variety of models, which include additive partially linear models for both the i.i.d. case and the time series case. As a result, the proposed model selection procedure is capable of selecting the most significant lags for both the parametric and nonparametric components. Both the methodology and theoretical techniques developed in this paper can be used to improve statistical model building and forecasting.

In this paper, we propose the combined cross-validation (CV) based nonparametric and parametric regression model selection procedure and develop the related theory. Moreover, we illustrate the CV criterion with simulated and real data sets. The organization of this paper is as follows. Section 2 proposes two CV based selection criteria. Applications and illustrations of the criteria are given in Section 3. The paper concludes with a discussion in Section 4. Assumptions and mathematical proofs are given

in Appendices A–C.

2. CV criteria for semiparametric time series regression

Although concepts like the Akaike’s information criterion (AIC) and maximum likelihood do not carry over to the nonparametric situation in a straightforward fashion, it makes sense to talk about prediction error and cross-validation in the general framework. The equivalence of AIC and CV criterion for the parametric autoregressive model selection was alluded by Tong (1976) and established by Stone (1977). Since then, Zhang (1991), Bickel and Zhang (1992), Cheng and Tong (1992, 1993), Vieu (1994), Yao and Tong (1994), and others have studied the behavior of the CV criterion in nonparametric regression for both the i.i.d. and time series cases.

Before establishing our general framework for the semiparametric time series case, we need to introduce some notation.

Let $A_0 = \{1, 2, \dots, p\}$, $D_q = \{1, 2, \dots, q\}$, \mathcal{A} denote all nonempty subsets of A_0 and \mathcal{D} denote all nonempty subsets of D_q . For any subset $A \in \mathcal{A}$, U_{tA} is defined as a column vector consisting of $\{U_{ti}, i \in A\}$. For any subset $D \in \mathcal{D}$, X_{tD} is defined as a column vector consisting of $\{X_{ti}, i \in D\}$. Throughout this paper, $B \subseteq C$ means that B can be the maximum subset C , and $B \subset C$ means that B cannot attain the maximum subset C . We use $d_E = |E|$ to denote the cardinality of a set E .

In this paper, we assume that there is a unique pair (A_*, D_*) with $A_* \in \mathcal{A}$ and $D_* \in \mathcal{D}$ such that there is a true and compact version of model (1.1) defined by

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

where β_* is a vector of unknown parameters, $\phi_*(\cdot)$ is an unknown function over $R^{|D_*|}$, and $e_t^* = Y_t - E[Y_t|U_t, X_t]$.

Detailed conditions for the existence and the uniqueness of A_* and D_* are discussed in Sections 2.1 and 2.2 below. This paper then considers the following cases:

- **Case I:** If the linear component of model (1.1) is already compact but the nonparametric component is not compact, we then take $A_* = A_0 = \{1, 2, \dots, p\}$ and estimate D_* in Section 2.1 below. We will use the notation of $D_* = D_0 = D_0(A_0)$ in Section 2.1.

- **Case II:** If both the linear and nonparametric components are not compact, we then estimate both A_* and D_* in Section 2.2 below. Note that the notation of $D_* = D_0(A_*)$ will be used in Section 2.2.
- **Case III:** If model (1.1) is already compact, then $A_* = A_0$ and $D_* = D_q$. For this case, no model selection is needed.
- **Case IV:** If the nonparametric component of model (1.1) is already compact but the parametric component is not compact, we take $D_* = D_q$ and then estimate A_* . As this is a special case of Case II with $D_* = D_q$ and the detailed discussion for this case is very similar but less difficult than that for Case II, we shall not discuss it in detail.

For Case I, Theorem 2.1 in Section 2.1 below extends some existing results for the nonparametric time series case under the β -mixing condition to the semiparametric time series case under the α -mixing condition. For Case II, Theorem 2.3 in Section 2.2 below shows that if a given data set (Y_t, U_t, X_t) satisfies a partially linear model of the form (1.1), the proposed nonparametric CV1 and parametric CVT_v selection procedure suggests that we need only to consider the selection of $(2^q - 1) \times (2^p - 1)$ possible models of the form (1.1). If we choose to use either the purely nonparametric CV1 selection procedure or the completely parametric CVT_v selection procedure for the selection of an optimum set of (U_t, X_t) , we need to consider the selection of $2^{p+q} - 1$ possible models. Consequently, in theory a completely linear model or a purely nonparametric regression model may be either too simple or too general for a given time series data. In practice, the computation of selecting $2^{p+q} - 1$ possible models is more expensive than that of selecting $(2^q - 1) \times (2^p - 1)$ possible models when p and q are large.

2.1. CV criterion for nonparametric regressors

Assume that the data set $\{(Y_t, U_t, X_t) : t \geq 1\}$ satisfies model (1.1). In this section, we assume that the linear component is already compact in the selection of nonparametric regressors.

Assume that the data set $\{(Y_t, U_t, X_{tD}) : t \geq 1\}$ satisfies

$$Y_t = U_t^T \beta(D) + \phi_D(X_{tD}) + e_{tD}, \quad (2.2)$$

where e_{tD} is an error process, $\beta(D) = (\beta_1(D), \dots, \beta_p(D))^\tau$ is a vector of unknown parameters, and $\phi_D(\cdot)$ is an unknown function over R^{d_D} . Note that $\beta(D)$ is still a vector of p unknown parameters, but may depend on D .

In order to ensure that model (2.2) is identifiable for each given $D \in \mathcal{D}$, one needs to define (see §1.2 of Härdle, Liang and Gao 2000)

$$\beta(D) = \{E(U_t - E[U_t|X_{tD}]) (U_t - E[U_t|X_{tD}])^\tau\}^{-1} E(U_t - E[U_t|X_{tD}]) (Y_t - E[Y_t|X_{tD}])$$

and

$$\phi_D(X_{tD}) = \phi_D(X_{tD}, \beta(D)) = E\{(Y_t - U_t^\tau \beta(D)) | X_{tD}\} = \phi_1(X_{tD}) - \phi_2(X_{tD})^\tau \beta(D), \quad (2.3)$$

under Assumption 2.1(i) below, where $\phi_1(X_{tD}) = E[Y_t | X_{tD}]$ and $\phi_2(X_{tD}) = E[U_t | X_{tD}]$.

For any $D \in \mathcal{D}$, define $\psi_D(U_t, X_{tD}) = U_t^\tau \beta(D) + \phi_D(X_{tD})$ and $\Psi(U_t, X_t) = E[Y_t | U_t, X_t]$. The following assumption imposes some existence and uniqueness conditions on model (2.2).

ASSUMPTION 2.1. (i) Assume that $\Delta_D = E\{U_t - E[U_t|X_{tD}]\} \{U_t - E[U_t|X_{tD}]\}^\tau$ is a positive definite matrix with order $d_D \times d_D$ for each given $D \in \mathcal{D}$.

(ii) Let $\mathcal{D}_1 = \{D \in \mathcal{D}, \text{ such that } \psi_D = \Psi\}$ and $\mathcal{D}_0 = \{D_0 \in \mathcal{D}_1, \text{ such that } |D_0| = \min_{D \in \mathcal{D}_1} |D|\}$. Assume that D_0 is the unique element of \mathcal{D}_0 and that $\phi_{D_0}(X_{tD_0})$ is an unknown nonparametric function.

REMARK 2.1. (i) Assumption 2.1(i) then requires the positivity of the matrix even when both X_t and U_t are dependent time series. When X_t and U_t are two independent time series, $\Delta_D = E\{U_t - E[U_t]\} \{U_t - E[U_t]\}^\tau$, which corresponds to the linear time series case. We should point out that when U_t and X_t have common components, Assumption 2.1 needs to be slightly modified. An obvious remedy for this case is to put $\beta_j(D) = 0$ when U_{tj} is equal to a component of X_{tD} .

(ii) Assumption 2.1(ii) assumes both the existence and uniqueness of D_0 . It might be possible that there exists another subset $D_1 \neq D_0$ such that $|D_1| = |D_0|$. This makes our discussion more complicated. Since it is not a likely case in practice, we agree to discard this case. In order to avoid this case, Assumption 2.1(ii) requires the uniqueness of the true nonparametric component and ensures that the true nonparametric component is

of the smallest possible dimension or compact. Assumption 2.1(ii) is also imposed to exclude the case where $\phi_{D_0}(\cdot)$ is a known parametric function. This is just for rigorousness consideration. Conventionally, the nonparametric component of a partially linear model is viewed as a nonparametric and unknown function.

(iii) Assumption 2.1 also implies that if there is another pair $(\beta'(D_0), \phi'_{D_0})$ such that

$$U_t^\tau \beta(D_0) + \phi_{D_0}(X_{tD_0}) = U_t^\tau \beta'(D_0) + \phi'_{D_0}(X_{tD_0}) \quad \text{almost surely,}$$

then $\beta(D_0) = \beta'(D_0)$ and $\phi_{D_0} = \phi'_{D_0}$. Thus Assumption 2.1 guarantees that the true regression function $U_t^\tau \beta(D_0) + \phi_{D_0}(X_{tD_0})$ is identifiable, i.e., $\beta(D_0)$ and ϕ_{D_0} are uniquely determined up to a set of measure zero.

It follows from (2.1)–(2.3) and Assumption 2.1 that for Case I we may define the true model as

$$Y_t = U_t^\tau \beta(D_0) + \phi_{D_0}(X_{tD_0}) + e_{tD_0}, \quad (2.4)$$

where $e_{tD_0} = Y_t - E[Y_t|U_t, X_t]$. Note that model (2.4) is a special case of (2.1) where $A_* = A_0$, $U_{tA_0} = U_t$, $\beta_* = \beta(D_0)$, $\phi_{D_0} = \phi_*$, $D_* = D_0$, $X_{tD_0} = X_{tD_*}$, and $e_{tD_0} = e_t^*$.

For the given D_0 , we define the least squares estimator, $\tilde{\beta}(D_0, h)$, of $\beta(D_0)$ as the solution of (see §1.2 of Härdle, Liang and Gao 2000 for example)

$$\sum_{t=1}^T \left\{ Y_t - U_t^\tau \tilde{\beta}(D_0, h) - \hat{\phi}(X_{tD_0}, \tilde{\beta}(D_0, h)) \right\}^2 = \min!, \quad (2.5)$$

where

$$\hat{\phi}(X_{tD}, \beta) = \sum_{s=1}^T W_D(t, s) (Y_s - U_s^\tau \beta), \quad \text{in which } W_D(t, s) = \frac{K_D((X_{tD} - X_{sD})/h)}{\sum_{l=1}^T K_D((X_{tD} - X_{lD})/h)},$$

T is the number of observations, K_D is a multivariate kernel function, and h is a bandwidth parameter satisfying $h = h_T \rightarrow 0$ as $T \rightarrow \infty$.

It follows from (2.4) that

$$\tilde{\beta}(D, h) = (\tilde{\Sigma}(D, h))^+ \sum_{t=1}^T \tilde{U}_t(D, h) (Y_t - \hat{\phi}_1(X_{tD}, h)), \quad (2.6)$$

where $(\cdot)^+$ denotes the Moore–Penrose inverse,

$$\tilde{\Sigma}(D, h) = \sum_{t=1}^T \tilde{U}_t(D, h) \tilde{U}_t(D, h)^\tau, \quad \tilde{U}_t(D, h) = U_t - \hat{\phi}_2(X_{tD}, h),$$

$$\hat{\phi}_1(X_{tD}, h) = \sum_{s=1}^T W_D(t, s) Y_s \text{ and } \hat{\phi}_2(X_{tD}, h) = \sum_{s=1}^T W_D(t, s) U_s.$$

In order to select both h and D_0 , we introduce several leave-one-out estimates. For any $D \in \mathcal{D}$, equations (2.3)–(2.4) suggest the leave-one-out estimator

$$\hat{\phi}_t(X_{tD}, \beta) = \hat{\phi}_{1t}(X_{tD}, h) - \hat{\phi}_{2t}(X_{tD}, h)^\tau \beta,$$

where

$$\hat{\phi}_{1t}(X_{tD}, h) = \sum_{s=1, s \neq t}^T W_D^{(-t)}(t, s) Y_s \text{ and } \hat{\phi}_{2t}(X_{tD}, h) = \sum_{s=1, s \neq t}^T W_D^{(-t)}(t, s) U_s,$$

in which

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

Then, we define the leave-one-out least squares (LS) estimator $\hat{\beta}(D, h)$ of $\beta(D)$ as the solution of

$$\sum_{t=1}^T \left\{ Y_t - U_t^\tau \hat{\beta}(D, h) - \hat{\phi}_t(X_{tD}, \hat{\beta}(D, h)) \right\}^2.$$

For any given $D \in \mathcal{D}$, the leave-one-out LS estimator is

$$\hat{\beta}(D, h) = (\tilde{\Sigma}(D, h))^+ \sum_{t=1}^T \tilde{U}_t(D, h) (Y_t - \hat{\phi}_{1t}(X_{tD}, h)), \quad (2.7)$$

where $\tilde{U}_t(D, h) = U_t - \hat{\phi}_{2t}(X_{tD}, h)$, $\tilde{\Sigma}(D, h) = \sum_{t=1}^T \tilde{U}_t(D, h) \tilde{U}_t(D, h)^\tau$. It is noted that the LS estimator $\tilde{\beta}(D_0, h)$ of (2.6) is asymptotically equivalent to the leave-one-out least squares (LS) estimator $\hat{\beta}(D_0, h)$ of (2.7). In defining the following leave-one-out cross-validation, we use the latter.

We now introduce a version of the leave-one-out cross-validation, abbreviated as CV1. For any $D \in \mathcal{D}$, we define

$$\text{CV1}(D, h) = \frac{1}{T} \sum_{t=1}^T \left\{ Y_t - U_t^\tau \hat{\beta}(D, h) - \hat{\phi}_t(X_{tD}, \hat{\beta}(D, h)) \right\}^2 w(X_t), \quad (2.8)$$

where $w(\cdot)$ is a weight function defined on R^q .

Let \hat{D}_0 and \hat{h} denote the estimators of D_0 and h , respectively, which are obtained by minimising the $\text{CV1}(D, h)$ function over $D \in \mathcal{D}$ and $h \in H_{TD}$, and written as

$$(\hat{D}_0, \hat{h}) = \operatorname{argmin}_{\{D \in \mathcal{D}, h \in H_{TD}\}} \text{CV1}(D, h), \quad (2.9)$$

where

$$H_{TD} = \left[a_D T^{-\frac{1}{4+|D|}-c_D}, b_D T^{-\frac{1}{4+|D|}+c_D} \right],$$

in which the constants a_D , b_D and c_D satisfy $0 < a_D < b_D < \infty$ and $0 < c_D < \frac{1}{2(4+|D|)}$.

REMARK 2.2. The cross-validation function CV1 of (2.8) generalises the conventional CV1 cross-validation function for purely nonparametric regression to the semiparametric setting. When $\beta(D) = 0$, the CV1 function reduces to the conventional leave-one-out cross-validation for purely nonparametric regression model selection. Similar to Vieu (1994), we integrate the weight function w not depending on D into CV1. Under an additional but complicated condition similar to condition (G) of Zhang (1991), however, we can integrate a weight function w_D depending on D into the CV1 function. Cheng and Tong (1993) also considered a special weight function. Yao and Tong (1994) avoided using such a weight function by assuming that the marginal density of X_t has a compact support.

We now state the first result of this paper and its proof is relegated to Appendix B.

THEOREM 2.1. *Assume that Assumptions 2.1 and A.1–A.4 listed in Appendix A hold with $A = A_0$. Then*

$$\lim_{T \rightarrow \infty} P(\hat{D}_0 = D_0) = 1 \quad \text{and} \quad \frac{\hat{h}}{h_0} \rightarrow_p 1$$

as $T \rightarrow \infty$, where h_0 is the minimizer of the mean average squared error (MASE) given by

$$\text{MASE}(D_0, h) = \frac{1}{T} \sum_{t=1}^T E \left\{ U_t^\tau \tilde{\beta}(D_0, h) + \hat{\phi}(X_{tD_0}, \tilde{\beta}(D_0, h)) - U_t^\tau \beta(D_0) - \phi_{D_0}(X_{tD_0}) \right\}^2.$$

REMARK 2.3. It can be shown that $h_0 = C_{D_0} T^{-\frac{1}{4+|D_0|}}$ and $C_{D_0} > 0$ is a constant independent of T . Due to this property, instead of defining h_0 as the minimizer of certain MASE we shall use this explicit form for h_0 throughout the rest of the paper. Theorem 2.1 shows that the true and unique subset D_0 can be identified asymptotically. Moreover, the criterion can also determine the bandwidth asymptotically.

When $\beta = 0$ in (1.1) and therefore $\beta(D) = 0$ in (2.2), we have the following result for purely nonparametric regression model selection.

COROLLARY 2.1. *For the purely nonparametric regression case, the conclusion of Theorem 2.1 holds.*

This result extends some existing results for nonparametric regression model selection for both the i.i.d. case and the β -mixing time series case to the α -mixing time series case.

Based on \hat{D}_0 and \hat{h} of (2.9), we define the following prediction equation

$$\hat{m}_{\hat{D}_0}(U_t, X_{t\hat{D}_0}) = U_t^\tau \tilde{\beta}(\hat{D}_0, \hat{h}) + \hat{\phi}(X_{t\hat{D}_0}, \tilde{\beta}(\hat{D}_0, \hat{h})). \quad (2.10)$$

COROLLARY 2.2. *Under the conditions of Theorem 2.1, we have as $T \rightarrow \infty$*

$$\frac{1}{T} \sum_{t=1}^T \{Y_t - \hat{m}_{\hat{D}_0}(U_t, X_{t\hat{D}_0})\}^2 \rightarrow_p \sigma_{1D_0}^2 = E\{Y_t - U_t^\tau \beta(D_0) - \phi_{D_0}(X_{tD_0})\}^2.$$

Corollary 2.2 shows that the semiparametric estimator of (2.10) is asymptotically close to the true regression function. The proofs of Corollaries 2.1–2.2 are relegated to Appendix B.

In Section 2.1, we have considered Case I where the linear component is already compact and then propose the leave-one-out cross-validation for the selection of nonparametric regressors. In Section 2.2 below, we consider the selection of both parametric and nonparametric regressors. Since for the selection of parametric regressors the leave-one-out cross-validation is asymptotically inconsistent (see Zhang 1993; Shao 1993), we need to consider using the leave- T_v -out cross-validation for the selection of parametric regressors. Moreover, because the theory of the leave- T_v -out cross-validation is different to that of the leave-one-out cross-validation and much more complicated, we consider Case II separately.

2.2. CV criterion for the selection of parametric regressors

As can be seen in Section 2.1, the selected \hat{D}_0 and \hat{h} depend on A_0 . Thus we can rewrite $\hat{D}_0 = \hat{D}_0(A_0)$ and $\hat{h} = \hat{h}(A_0)$. Let \mathcal{A} denote all nonempty subsets of A_0 . For $A \in \mathcal{A}$, let β_A be a column vector consisting of $\{\beta_i : i \in A\}$. Denote U_{tA} with $A = A_0$ by U_t and β_A with $A = A_0$ by $\beta = (\beta_1, \dots, \beta_p)^\tau$.

To extend Assumption 2.1 to the case where both the linear and nonparametric components are not compact, one needs to restate some notation.

For each $A \in \mathcal{A}$ and $D \in \mathcal{D}$, define $\psi_{A,D}(U_{tA}, X_{tD}) = U_{tA}^\tau \beta_A + \phi_D(X_{tD})$ and $\Psi(U_t, X_t) = E[Y_t | U_t, X_t]$. The following assumption imposes some existence and uniqueness conditions on the true versions of A and D . Detailed explanation for Assumption 2.2 below will be similar to Remark 2.1.

ASSUMPTION 2.2. (i) Assume that $\Delta_{A,D} = E \{U_{tA} - E[U_{tA} | X_{tD}]\} \{U_{tA} - E[U_{tA} | X_{tD}]\}^\tau$ is a positive definite matrix with order $d_D \times d_D$ for each given $A \in \mathcal{A}$ and $D \in \mathcal{D}$.

(ii) For each given $A \in \mathcal{A}$, let $\mathcal{D}_{1A} = \{D \in \mathcal{D}, \text{ such that } \psi_{A,D} = \Psi\}$ and $\mathcal{D}_{0A} = \{D_0(A) \in \mathcal{D}_{1A}, \text{ such that } |D_0(A)| = \min_{D \in \mathcal{D}_{1A}} |D|\}$. Assume that $D_0(A)$ is the unique element of \mathcal{D}_{0A} and that $\phi_{D_0(A)}(X_{tD_0(A)})$ is an unknown nonparametric function for each given $A \in \mathcal{A}$.

Following Assumption 2.2, for each $A \in \mathcal{A}$ we can define the corresponding $D_0(A)$. Theorem 2.1 then shows that

$$\lim_{T \rightarrow \infty} P(\hat{D}_0(A) = D_0(A)) = 1 \text{ and } \frac{\hat{h}(A)}{h_0(A)} \rightarrow_p 1$$

as $T \rightarrow \infty$, where $h_0(A) = C_{D_0(A)} T^{-\frac{1}{4+|D_0(A)|}}$.

For simplicity and convenience, we introduce the following notation:

$$\begin{aligned} \hat{\psi}_1(t, A) &= \hat{\phi}_1(X_{t\hat{D}_0(A)}, \hat{h}(A)) = \sum_{s=1}^T W_{\hat{D}_0(A)}(t, s) Y_s, \\ \hat{\psi}_2(t, A) &= \hat{\phi}_2(X_{t\hat{D}_0(A)}, \hat{h}(A)) = \sum_{s=1}^T W_{\hat{D}_0(A)}(t, s) U_{sA}, \\ \eta_{tA} &= U_{tA} - E[U_{tA} | X_{tD_0(A)}], \quad \delta_{tA} = E[U_{tA} | X_{tD_0(A)}] - \hat{\psi}_2(t, A), \\ V_{tA} &= \eta_{tA} + \delta_{tA} = U_{tA} - \hat{\psi}_2(t, A), \quad V_A = (V_{1A}, \dots, V_{TA})^\tau, \\ \hat{\psi}_1(t) &= \hat{\psi}_1(t, A_0), \quad \hat{\psi}_2(t) = \hat{\psi}_2(t, A_0), \quad \eta_t = U_t - E[U_t | X_{tD_0}], \quad \delta_t = E[U_t | X_{tD_0}] - \hat{\psi}_2(t), \\ V_t &= \eta_t + \delta_t = U_t - \hat{\psi}_2(t), \quad V = (V_1, \dots, V_T)^\tau, \quad Z_t = Y_t - \hat{\psi}_1(t), \text{ and } Z = (Z_1, \dots, Z_T)^\tau, \end{aligned} \tag{2.11}$$

where $D_0 = D_0(A_0)$ is as defined in Assumption 2.1.

Because some of the components of β may be zero, the following model

$$Y_t = U_{tA}^\tau \beta_A + \phi_{D_0(A)}(X_{tD_0(A)}) + \epsilon_{tA}, \text{ where } \epsilon_{tA} \text{ is an error process,} \quad (2.12)$$

might be more compact than model (2.4) given by $Y_t = U_t^\tau \beta(D_0) + \phi_{D_0}(X_{tD_0}) + e_{tD_0}$. Note that $\beta(D)$ signifies that $\beta(D)$ may depend on D while the notation of β_A means that β_A is a subset of β .

As mentioned earlier, for each $A \in \mathcal{A}$ it is natural to estimate each $D_0(A)$ by $\hat{D}_0(A)$. The definition of $\hat{\phi}(X_{tD}, \beta)$ of (2.5) then suggests estimating (see §1.2 of Härdle, Liang and Gao 2000 for example)

$$\phi_{D_0(A)}(X_{tD_0(A)}) = \phi_{D_0(A)}(X_{tD_0(A)}, \beta_A) \text{ by } \hat{\phi}(X_{t\hat{D}_0(A)}, \beta_A) = \hat{\psi}_1(t, A) - \hat{\psi}_2(t, A)^\tau \beta_A.$$

Thus, using (2.11), model (2.12) can be rewritten as

$$\begin{aligned} Y_t - \hat{\psi}_1(t, A) &= \beta_A^\tau (U_{tA} - \hat{\psi}_2(t, A)) + \phi_{D_0(A)}(X_{tD_0(A)}) - \hat{\phi}(X_{t\hat{D}_0(A)}, \beta_A) + \epsilon_{tA} \\ &= V_{tA}^\tau \beta_A + \epsilon_{tA} + o_p(1) \end{aligned}$$

using the fact that the rate of uniform convergence of $\hat{\phi}(X_{t\hat{D}_0(A)}, \beta_A)$ to $\phi_{D_0(A)}(X_{tD_0(A)})$ is of order $o_p(1)$ (see Theorem 3.2.2 of Härdle, Liang and Gao 2000 for example).

This suggests using a linear model of the form

$$Y_t - \hat{\psi}_1(t, A) = V_{tA}^\tau \beta_A + \epsilon_{tA} \quad (2.13)$$

to approximate model (2.12) in the selection of A without changing the true version of A . Obviously, there are $2^p - 1$ possible models of the form (2.13), each of which corresponds to a subset A and is defined by \mathcal{M}_A . The dimension of \mathcal{M}_A is defined to be d_A , the number of predictors in \mathcal{M}_A . If we know whether each component of β is zero or not, then the models \mathcal{M}_A can be classified into two categories:

- Category I: At least one nonzero component of β is not in β_A .
- Category II: β_A contains all nonzero components of β .

Clearly, the models in Category I are incorrect models, and the models in Category II may be inefficient because of their unnecessarily large sizes. The optimum model, denoted by \mathcal{M}_* , is the model in Category II with the smallest dimension.

Let A_* correspond to \mathcal{M}_* . For Case II, we may define the true model as

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

where $D_* = D_0(A_*)$, $\phi_* = \phi_{D_*} = \phi_{D_0(A_*)}$, and e_t^* is as defined in (2.1). Note that this is the true model we have assumed in (2.1) for Case II.

Thus, in order to determine the true model (2.1) for Case II, one needs to estimate A_* . The selection of A is carried out by using the data $\{(Z_t, V_t) : t = 1, 2, \dots, T\}$ satisfying

$$Z_t = V_t^\tau \beta + \epsilon_t,$$

where ϵ_t is an error process. Under model \mathcal{M}_A , the least squares estimator of β_A is

$$\hat{\beta}_A = (V_A^\tau V_A)^+ V_A^\tau Z,$$

where Z and V_A are as defined in (2.11).

Using model \mathcal{M}_A fitted based on the data $\{(Z_t, V_t) : t = 1, 2, \dots, T\}$, the average squared prediction error is

$$\begin{aligned} L_T(A) &= \frac{1}{T} \sum_{t=1}^T [Z_t - V_{tA}^\tau \hat{\beta}_A]^2 = \frac{1}{T} (Z - V_A \hat{\beta}_A)^\tau (Z - V_A \hat{\beta}_A) \\ &= \frac{1}{T} \epsilon^\tau \epsilon + \frac{1}{T} \epsilon^\tau P_A \epsilon + \frac{1}{T} (V\beta)^\tau R_A (V\beta) + \frac{2}{T} \epsilon^\tau R_A (V\beta), \end{aligned} \quad (2.14)$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_T)^\tau$, $P_A = V_A (V_A^\tau V_A)^+ V_A^\tau$, $R_A = I_T - P_A$, and I_T is the identity matrix of order T .

It follows from (2.14) that because of Assumption A.1, the conditionally expected average squared error is

$$\begin{aligned} R_T(A, V) &= E[L_T(A)|V] = \frac{1}{T} E[\epsilon^\tau \epsilon | V] + \frac{1}{T} E[\epsilon^\tau P_A \epsilon | V] + \frac{1}{T} (V\beta)^\tau R_A (V\beta) + \frac{2}{T} E[\epsilon^\tau R_A (V\beta) | V] \\ &= \sigma_\epsilon^2 + \frac{1}{T} d_A \sigma_\epsilon^2 + \Delta_{T,A}, \text{ with probability one,} \end{aligned} \quad (2.15)$$

where $\sigma_\epsilon^2 = E[\epsilon^\tau \epsilon]$ and $\Delta_{T,A} = \frac{1}{T} (V\beta)^\tau R_A (V\beta)$.

When \mathcal{M}_A is in Category I, we assume that

$$\liminf_{T \rightarrow \infty} \Delta_{T,A} > 0 \text{ in probability.} \quad (2.16)$$

When \mathcal{M}_A is in Category II, it follows from (2.14) and (2.15) that because $V\beta = V_A\beta_A$,

$$L_T(A) = \frac{1}{T}\epsilon^\tau\epsilon + \frac{1}{T}\epsilon^\tau P_A\epsilon + \frac{2}{T}\epsilon^\tau R_A(V\beta) \quad \text{and} \quad R_T(A, V) = \frac{1}{T}(T - d_A)\sigma_\epsilon^2.$$

We now have the following remark.

REMARK 2.4. As argued in Shao (1993), condition (2.16) is a type of asymptotic model identifiability condition and is very minimal for asymptotic analysis. It can be shown that for (2.16) to hold, it suffices to assume that for \mathcal{M}_A in Category I

$$\liminf_{T \rightarrow \infty} \frac{1}{T}(\eta\beta)^\tau(I_T - \eta_A(\eta_A^\tau\eta_A)^+\eta_A^\tau)\eta\beta > 0 \text{ in probability,} \quad (2.17)$$

where $\eta = (\eta_1, \dots, \eta_T)^\tau$, $\eta_A = (\eta_{1A}, \dots, \eta_{TA})^\tau$, $\eta_t = U_t - E[U_t|X_{tD_0}]$ and $\eta_{tA} = U_{tA} - E[U_{tA}|X_{tD_0(A)}]$ are as defined in (2.11). It follows that when U_t and X_t are independent, we have

$$\eta_t = U_t - E[U_t] \quad \text{and} \quad \eta_{tA} = U_{tA} - E[U_{tA}].$$

Thus, condition (2.17) imposes only an asymptotic model identifiability condition on the linear component and is a natural extension of condition (2.5) of Shao (1993) to the semiparametric time series setting.

We now propose our cross-validation criterion for the selection of $A \in \mathcal{A}$. Suppose that we split the data set into two parts: $\{(Z_t, V_t) : t \in S\}$ and $\{(Z_t, V_t) : t \in S^c\}$, where S is a subset of $\{1, 2, \dots, T\}$ containing T_v integers and S^c is its complement containing T_c integers, $T_v + T_c = T$. The model \mathcal{M}_A is fitted using the construction data $\{(Z_t, V_t) : t \in S^c\}$ and the prediction error is assessed using the validation data $\{(Z_t, V_t) : t \in S\}$, treated as if they were future values. The average squared prediction error is

$$\text{CV}(T_v) = \text{CV}_{A,S}(T_v) = \frac{1}{T_v} \left\| Z_S - \hat{Z}_{A,S^c} \right\|^2 = \frac{1}{T_v} \left\| (I_{T_v} - Q_{A,S})^+(Z_S - V_{A,S}\hat{\beta}_A) \right\|^2,$$

where $\|x\| = \sqrt{x^\tau x}$ for a vector x , Z_S is the column vector containing the components of Z indexed by $t \in S$, $V_{A,S}$ is the $T_v \times d_A$ matrix containing the rows of V_A indexed by $t \in S$, \hat{Z}_{A,S^c} is the prediction of Z_S using the construction data and the least squares method under model \mathcal{M}_A , $Q_{A,S} = V_{A,S}(V_A^\tau V_A)^+ V_{A,S}^\tau$, and $\hat{\beta}_A$ is as defined before.

The $CV_{A,S}(T_v)$ function is called the leave- T_v -out cross-validation, abbreviated as $CV(T_v) = CVT_v$. From the computational point of view, the simplest CVT_v is the one with $T_v \equiv 1$ and $S = \{t\}$; that is, the CV1. As the CV1 is asymptotically inconsistent, we adopt the following Monte Carlo CVT_v in the selection of A .

Randomly draw a collection \mathcal{R} of b subsets of $\{1, 2, \dots, T\}$ that have size T_v and select a model by minimizing

$$\text{MCCV}(A, T_v) = \frac{1}{b} \sum_{S \in \mathcal{R}} CV_{A,S}(T_v) = \frac{1}{bT_v} \sum_{S \in \mathcal{R}} \|Z_S - \hat{Z}_{A,S^c}\|^2. \quad (2.18)$$

This method is called the Monte Carlo CVT_v , abbreviated as $\text{MCCV}T_v$, as (2.18) is obtained by randomly splitting the data b times and averaging the squared prediction errors over the splits.

We now have the following result.

THEOREM 2.2. *Assume that Assumption 2.2 and Assumptions A.1–A.5 hold. Then we have the following conclusions:*

(i) *If \mathcal{M}_A is in Category I, then there exists $R_T \geq 0$ such that*

$$\text{MCCV}(A, T_v) = \frac{1}{T_v b} \sum_{S \in \mathcal{R}} \epsilon_S^\tau \epsilon_S + \Lambda_{T,A} + o_p(1) + R_T,$$

where $\epsilon_S = V_S - Z_S \beta$ and $\Lambda_{T,A} = \frac{1}{T} (\eta \beta)^\tau (I_T - \eta_A (\eta_A^\tau \eta_A)^+ \eta_A^\tau) \eta \beta$.

(ii) *If \mathcal{M}_A is in Category II, then*

$$\text{MCCV}(A, T_v) = \frac{1}{T_v b} \sum_{S \in \mathcal{R}} \epsilon_S^\tau \epsilon_S + \frac{d_A}{T_c} \sigma_\epsilon^2 + o_p\left(\frac{1}{T_c}\right).$$

(iii) *Consequently,*

$$\lim_{T \rightarrow \infty} P(\text{the selected model is } \mathcal{M}_*) = 1.$$

Let \hat{A} correspond to the selected model. Then, Theorems 2.1 and 2.2 imply the following main result of this paper.

THEOREM 2.3. *Assume that the conditions of Theorem 2.2 hold. Then*

$$\lim_{T \rightarrow \infty} P(\hat{A} = A_*, \hat{D}_0(\hat{A}) = D_*) = 1 \text{ and } \frac{\hat{h}(\hat{A})}{h_0(A_*)} \rightarrow_p 1$$

as $T \rightarrow \infty$, where $h_0(A_*) = C_{D_*} T^{-\frac{1}{4+|D_*|}}$.

The proofs of Theorems 2.2 and 2.3 are relegated to Appendices B and C.

REMARK 2.5. As can be seen, Theorems 2.2 and 2.3 not only extend the model selection results of Shao (1993) for the fixed design linear model to the selection of both parametric and nonparametric regressors in semiparametric time series regression, but also cover the model selection results for the nonparametric time series regression for the α -mixing case. We should also point out that due to the complexity of a partially linear model of the form (1.1), the extension of the results of Shao (1993) is not trivial and direct, but requires the establishment of a novel framework for the semiparametric time series case. For example, the construction of equations (2.13)–(2.18) requires some delicate and deep understanding of both parametric model and semiparametric frameworks, although they look similar to the parametric framework of Shao (1993). In particular, the key condition (2.17) involves only the main components, $U_t - E[U_t|X_{tD_0}]$ and $U_{tA} - E[U_{tA}|X_{tD_0(A)}]$, of U_t and U_{tA} rather than U_t and U_{tA} themselves. This is part of the reason we think that (2.17) is the weakest possible condition and a natural extension of (2.5) of Shao (1993) to the semiparametric time series setting. The proof of Theorem 2.2 also requires some delicate reasoning.

REMARK 2.6. It should be noted that the multifold cross-validation (MCV) criterion proposed by Zhang (1993) can also be employed to select the parametric regressors. The detailed employment of the MCV is very similar to that of the CVT_v . Another related criterion is the modified final prediction error (MFPE) criterion proposed by Zheng and Loh (1997). We should point out that when p , the number of the parametric regressors in model (1.1), depends on T and increases as T increases, the parametric leave-one-out cross-validation is consistent and asymptotically optimal in some sense. The discussion for this case is quite different and requires some detailed extension of Shao (1997) and Gao, Tong and Wolff (2001a).

In summary, Theorems 2.1–2.3 not only provide the asymptotic consistency of the combined nonparametric CV1 and parametric CVT_v selection procedure, but also show that if a partially linear model of the form (1.1) within the context tried is the truth, then the combined selection procedure will find it asymptotically. In Section 3 below, we will show how to implement the proposed selection procedure in practice.

3. Examples and applications

In this section, we apply Theorems 2.1–2.3 to determine simulated models and to fit a set of real data.

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

$$Y_t = 0.35Y_{t-1} - 0.15Y_{t-2} + 0.5\frac{X_t}{1 + X_t^2} + e_t,$$

where

$$X_t = 0.3X_{t-1} + 0.2X_{t-2} + \epsilon_t, \quad t = 3, 4, \dots, T,$$

in which e_t and ϵ_t are mutually independent and identically distributed over uniform distributions $(-0.25, 0.25)$ and $(-0.5, 0.5)$ respectively, X_1, X_2, Y_1, Y_2 are i.i.d. over uniform distribution $(-1, 1)$, and the processes $\{(\epsilon_t, e_t) : t \geq 3\}$ are independent of both (X_1, X_2) and (Y_1, Y_2) .

It follows from the definition of Y_t that Assumption 2.1(i) holds. For Example 3.1, the strict stationarity and mixing condition can be justified by using Assumption 3.3 and Lemma 3.1 of Masry and Tjøstheim (1997). Thus, Assumption A.1 holds. For an application of Theorem 2.1, denote

$$U_t = (Y_{t-1}, Y_{t-2})^\tau, \quad \beta = (\beta_1, \beta_2)^\tau = (0.35, -0.15)^\tau, \quad \phi(X_t) = 0.5\frac{X_t}{1 + X_t^2}.$$

Throughout Example 3.1, we consider using $h \in H_T = [0.3 \cdot T^{-\frac{7}{30}}, 2 \cdot T^{-\frac{1}{6}}]$ and the following weight function

$$w(u) = \begin{cases} 1 & \text{if } |u| \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

For the multivariate kernel function $K(\cdot)$ involved in $W_D^{(-t)}(t, s)$ and $W_D(t, s)$, define $K(u_1, u_2) = \prod_{i=1}^2 k(u_i)$, where

$$k(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right).$$

It follows that Assumptions A.1–A.4 are all satisfied.

In this example, we consider the case where X_t and X_{t-1} are selected as the candidates of nonparametric regressors and use the CV1 function of (2.8) to determine

whether X_t is the optimum nonparametric regressor. We then further use the $\text{MCCV}(T_v)$ function of (2.18) to check if Y_{t-1} and Y_{t-2} are the true parametric regressors. Let $\mathcal{D} = \{\{0, 1\}, \{0\}, \{1\}\}$, $X_{tD_0} = X_t$, $X_{tD_1} = X_{t-1}$, and $X_{tD_2} = (X_t, X_{t-1})^\tau$, $\mathcal{A} = \{\{1, 2\}, \{1\}, \{2\}\}$, $U_{tA_0} = (Y_{t-1}, Y_{t-2})^\tau$, $U_{tA_1} = Y_{t-1}$ and $U_{tA_2} = Y_{t-2}$. Then $|D_2| = |A_0| = 2$ and $|D_0| = |D_1| = |A_1| = |A_2| = 1$. In the detailed calculation of $\text{MCCV}(T_v)$, we choose $b = T$, $T_v = T - T_c$ and $T_c = \lceil T^{3/4} \rceil$, the largest integer part of $T^{3/4}$.

Now \mathcal{D}_0 in Assumption 2.1(ii) has the unique element $D_0 = \{0\}$. Assumption A.5(ii) follows immediately from the choice of $b = T$ and T_v . Before justifying Assumption A.5(i), we introduce the following notation.

$$\eta_{t1} = Y_{t-1} - E[Y_{t-1}|X_t], \quad \eta_{t2} = Y_{t-2} - E[Y_{t-2}|X_t], \quad \eta_t = (\eta_{t1}, \eta_{t2})^\tau, \quad \eta = (\eta_1, \dots, \eta_T)^\tau$$

$$\eta_{tA_0} = \eta_t, \quad \eta_{tA_1} = \eta_{t1}, \quad \eta_{tA_2} = \eta_{t2}, \quad \eta_{A_i} = (\eta_{1A_i}, \dots, \eta_{TA_i})^\tau, \quad i = 1, 2.$$

A detailed calculation yields that for $i = 1, 2$

$$(\eta\beta)^\tau \left(I_T - \eta_{A_i} (\eta_{A_i}^\tau \eta_{A_i})^+ \eta_{A_i}^\tau \right) (\eta\beta) = \frac{\sum_{t=3}^T \eta_{t1}^2 \sum_{t=3}^T \left(\sum_{j=1}^2 \beta_j \eta_{tj} \right)^2 - \left[\sum_{t=3}^T \eta_{t1} \left(\sum_{j=1}^2 \beta_j \eta_{tj} \right) \right]^2}{\sum_{t=3}^T \eta_{ti}^2} > 0$$

with probability one, because $P(\eta_{t1} = \eta_{t2}) = 0$. This shows that Assumption A.5(i) holds. Therefore, Assumptions 2.1 and A.1–A.5 all hold.

In order to compare the semiparametric model selection function CV1 with its special case, namely the nonparametric model selection function, we calculate the following sample average squared error (ASE) over 150 replications,

$$\text{ASE} = \frac{1}{150} \sum_{150 \text{ replications}} \left\{ \frac{1}{T-2} \sum_{t=3}^T [\hat{m}(Z_t) - m(Z_t)]^2 \right\},$$

where $m(Z_t) = 0.35Y_{t-1} - 0.15Y_{t-2} + 0.5 \frac{X_t}{1+X_t^2}$, $\hat{m}(Z_t)$ is a semiparametric regression estimator or a nonparametric regression estimator of $m(Z_t)$, and $Z_t = (Y_{t-1}, Y_{t-2}, X_t)$.

For the three sample sizes $T = 22$, $T = 72$ and $T = 152$, we calculated the probabilities of the selected parametric and nonparametric regressors in 150 replications. In addition, for each case we calculated the sample average square error (ASE). Table 3.1 below reports the results of the simulation for the semiparametric leave-one-out cross-validation function CV1. Table 3.2 below reports the results of the simulation for the parametric leave- T_v -out cross-validation function $\text{MCCV}(T_v)$ and the corresponding parametric leave-one-out cross-validation function CV1 for empirical comparison.

Table 3.1. The semiparametric CV1 based probabilities and ASEs for Example 3.1

Parametric	Nonparametric	Probability			ASE value		
		$T = 22$	$T = 72$	$T = 152$	$T = 22$	$T = 72$	$T = 152$
subset	subset						
$\{Y_{t-1}, Y_{t-2}\}$	$\{X_t, X_{t-1}\}$	0.162	0.114	0.003	0.0164	0.0162	0.0163
	$\{X_{t-1}\}$	0.306	0.238	0.052	0.0157	0.0154	0.0153
	$\{X_t\}$	0.532	0.648	0.945	0.0051	0.0017	0.0010
$\{Y_{t-1}\}$	$\{X_t, X_{t-1}\}$	0.177	0.158	0.012	0.0171	0.0167	0.0168
	$\{X_{t-1}\}$	0.324	0.215	0.104	0.0165	0.0159	0.0157
	$\{X_t\}$	0.499	0.627	0.884	0.0054	0.0024	0.0018
$\{Y_{t-2}\}$	$\{X_t, X_{t-1}\}$	0.214	0.131	0.076	0.0248	0.0251	0.0254
	$\{X_{t-1}\}$	0.376	0.287	0.219	0.0197	0.0188	0.0183
	$\{X_t\}$	0.410	0.582	0.705	0.0095	0.0061	0.0054

Table 3.2. The parametric $MCCV(T_v)$ and CV1 based probabilities for Example 3.1

Parametric and nonparametric	$MCCV(T_v)$			CV1		
	$T = 22$	$T = 72$	$T = 152$	$T = 22$	$T = 72$	$T = 152$
subset						
$\{Y_{t-1}, Y_{t-2}, X_t\}$	0.617	0.681	0.938	0.531	0.611	0.769
$\{Y_{t-1}, X_t\}$	0.242	0.212	0.053	0.312	0.274	0.187
$\{Y_{t-2}, X_t\}$	0.141	0.107	0.009	0.157	0.115	0.044

REMARK 3.1. (i) First, Tables 3.1 and 3.2 show that both the CV1 function and the $MCCV(T_v)$ function can be implemented in practice. Second, Table 3.1 supports the validity of our definition of optimum subset (see Assumption 2.1). Third, the detailed simulation results show that \hat{D}_0 is a reasonably good estimator of D_0 even when the sample size T is as small as 22 as shown in Table 3.1. Fourth, Table 3.2 shows that both the $MCCV(T_v)$ function and the CV1 function can identify the optimum parametric regressor $\{Y_{t-1}, Y_{t-2}\}$. Finally, the performance of the $MCCV(T_v)$ is better than the CV1: this is a reflection of the fact that $MCCV(T_v)$ leads to a consistent subset selection while CV1 does not.

(ii) In addition, the ASE values in Table 3.1 also highlight the small sample performance of the the semiparametric CV1 function. For example, for the case where $T = 22$,

the ASE value for the true model (see the fifth row and sixth column) is 0.0051 and smaller than 0.0054, the ASE for the second best model (see the eighth row and sixth column). For the same model, the ASE decreases when T increases. For example, when $T = 152$, the ASE for the true model (see the fifth row and eighth column) is already as small as 0.0010.

(iii) Before using the standard normal kernel function $k(\cdot)$, we also calculated the corresponding probabilities and ASEs for a uniform kernel function. Our computation shows that the small sample results for the standard normal kernel function are much better and more stable than those for the uniform kernel. In the meantime, besides the bandwidth interval H_T , we also calculated the CV1 function over all possible intervals. Our computation indicates that H_T is the smallest possible interval, on which the CV1 function for each possible model can attain the smallest value.

(iv) Throughout Example 3.1, we point out that Assumptions 2.1 and A.1–A.5 are satisfied. In theory, Assumption A.5(i) is a very minimal model identifiability condition. In practice, it is not easy to justify the model identifiability condition. For Example 3.1, however, we have been able to calculate the quadratic form explicitly and to show that the quadratic form is positive with probability one.

We now compare our semiparametric model selection function CV1 with the fully nonparametric model selection function. For the same Example 3.1, consider the case where Y_{t-1} , Y_{t-2} , X_t and X_{t-1} are selected as the candidates of nonparametric regressors. For an application of the CV1 function, we choose the same w , k and h defined as above, and define for $j = 2, 3, 4$

$$K_j(u_1, u_2, \dots, u_j) = \prod_{i=1}^j k(u_i)$$

for the multivariate kernel function involved in $W_D^{(-t)}(t, s)$ and $W_D(t, s)$.

Let $X_{tD_1} = (Y_{t-1}, Y_{t-2}, X_t, X_{t-1})^\tau$, $X_{tD_2} = (Y_{t-1}, Y_{t-2}, X_t)^\tau$, $X_{tD_3} = (Y_{t-1}, Y_{t-2}, X_{t-1})^\tau$, $X_{tD_4} = (Y_{t-1}, X_t, X_{t-1})^\tau$, $X_{tD_5} = (Y_{t-2}, X_t, X_{t-1})^\tau$, $X_{tD_6} = (Y_{t-1}, Y_{t-2})^\tau$, $X_{tD_7} = (Y_{t-1}, X_t)^\tau$, $X_{tD_8} = (Y_{t-1}, X_{t-1})^\tau$, $X_{tD_9} = (Y_{t-2}, X_t)^\tau$, $X_{tD_{10}} = (Y_{t-2}, X_{t-1})^\tau$, $X_{tD_{11}} = (X_t, X_{t-1})^\tau$, $X_{tD_{12}} = Y_{t-1}$, $X_{tD_{13}} = Y_{t-2}$, $X_{tD_{14}} = X_{t-1}$, and $X_{tD_{15}} = X_t$. Then $|D_1| = 4$, $|D_2| = |D_3| = |D_4| = |D_5| = 3$, $|D_6| = |D_7| = |D_8| = |D_9| = |D_{10}| = |D_{11}| = 2$, and $|D_{12}| = |D_{13}| = |D_{14}| = |D_{15}| = 1$. For all the subsets, we calculated all the correspond-

ing CV1 values. For each of the three sample sizes $T = 22, 72$ and 152 , we calculated CV1 value and the sample average squared error (ASE). Table 3.3 below provides the CV1 values and the corresponding ASE values for the nonparametric model selection.

Table 3.3. The nonparametric CV1 based minimum CV values and ASEs for Example 3.1

Nonparametric subset	CV value			ASE value		
	$T = 22$	$T = 72$	$T = 152$	$T = 22$	$T = 72$	$T = 152$
$\{Y_{t-1}, Y_{t-2}, X_t, X_{t-1}\}$	0.04702	0.04608	0.04535	0.02576	0.02471	0.02315
$\{Y_{t-1}, Y_{t-2}, X_t\}$	0.02445	0.02365	0.02219	0.00846	0.00578	0.00509
$\{Y_{t-1}, Y_{t-2}, X_{t-1}\}$	0.03641	0.03406	0.03321	0.01908	0.01823	0.01793
$\{Y_{t-1}, X_t, X_{t-1}\}$	0.02648	0.02569	0.02423	0.01051	0.00984	0.00715
$\{Y_{t-2}, X_t, X_{t-1}\}$	0.03644	0.03506	0.03377	0.01921	0.01839	0.01794
$\{Y_{t-1}, Y_{t-2}\}$	0.04605	0.04511	0.04435	0.02626	0.02571	0.02485
$\{Y_{t-1}, X_t\}$	0.04603	0.04505	0.04434	0.02624	0.02570	0.02486
$\{Y_{t-1}, X_{t-1}\}$	0.04606	0.04507	0.04436	0.02626	0.02572	0.02487
$\{Y_{t-2}, X_t\}$	0.04604	0.04506	0.04435	0.02624	0.02571	0.02484
$\{Y_{t-2}, X_{t-1}\}$	0.04605	0.04508	0.04437	0.02629	0.02573	0.02488
$\{X_t, X_{t-1}\}$	0.04606	0.04509	0.04439	0.02628	0.02575	0.02489
$\{Y_{t-1}\}$	0.04884	0.04664	0.04571	0.02552	0.02468	0.02343
$\{Y_{t-2}\}$	0.04414	0.03971	0.03874	0.01830	0.01719	0.01637
$\{X_{t-1}\}$	0.04454	0.04094	0.03967	0.01963	0.01874	0.01721
$\{X_t\}$	0.03128	0.02912	0.02716	0.01191	0.01011	0.00737

REMARK 3.2. First, Table 3.3 shows that the true subset $\{Y_{t-1}, Y_{t-2}, X_t\}$ is readily selected using our method. Second, for each case the ASE of the true nonparametric model is always larger than that of the corresponding semiparametric model (see the sixth–eighth columns of Table 3.1 and the fifth–seventh columns of Table 3.3). For example, for the case of $T = 22$, the ASE of the true nonparametric model in the fifth column of Table 3.3 is 0.00846, which is larger than 0.0051 in the sixth column of Table 3.1, the ASE of the true semiparametric model. Moreover, by comparing the CPU hours

for Tables 3.1 and 3.3, we know that the computation of the semiparametric model selection function CV1 is much less expensive than that of the nonparametric model selection function. Therefore we conclude that when selecting an optimum subset of nonparametric regressors for a partially linear model, the semiparametric model selection function CV1 is much more efficient than the usual nonparametric model selection function.

EXAMPLE 3.2. Fisheries Western Australia (WA) manages commercial fishing in Western Australia. Simple Catch and Effort statistics are often used in regulating the amount of fish that can be caught and the number of boats that are licensed to catch them. The establishment of the relationship between the Catch (in kilograms) and Effort (the number of days the fishing vessels spent at sea) is very important both commercially and ecologically. This example considers using the proposed model selection procedure to find a best possible model for the relationship between catch and effort.

The historical monthly fishing data from January 1976 through to December 1999 available to us comes from the Fisheries WA Catch and Effort Statistics (CAES) database. Existing studies from the Fisheries suggest that the relationship between the catch and the effort does not look like linear while the dependence of the current catch on the past catch appears to be linear. This suggests using a partially linear model of the form

$$C_t = \beta_1 C_{t-1} + \dots + \beta_p C_{t-p} + \phi(E_t, E_{t-1}, \dots, E_{t-q+1}) + \epsilon_t,$$

where ϵ_t is a random error, C_t and E_t represent the catch and the effort at time t , respectively. In the detailed computation, we use the transformed data $Y_t = \log_{10}(C_t)$ and $X_t = \log_{10}(E_t)$ satisfying the following model

$$Y_{t+r} = \beta_1 Y_{t+r-1} + \dots + \beta_p Y_{t+r-p} + \phi(X_{t+r}, \dots, X_{t+r-q+1}) + e_t, \quad (3.1)$$

where $r = \max(p, q)$ and e_t is a random error with zero mean and finite variance.

Before using model (3.1), we need to choose an optimum and compact form of model (3.1). We consider the case of $p = 4$ and $q = 5$ and then find an optimum model. For this case, there are $2^4 - 1 = 15$ different parametric regressors and $2^5 - 1 = 31$ different nonparametric regressors for model (3.1).

Similar to Example 3.1, we define the parametric candidates U_{tA_i} for $1 \leq i \leq 15$ and the nonparametric candidates X_{tD_j} for $1 \leq j \leq 31$. It follows that

$$Y_{t+5} = U_{tA_i}^r \beta_{A_i} + \phi_{D_j}(X_{tD_j}) + e_{tij}, \quad (3.2)$$

where β_{A_i} and ϕ_{D_j} are similar to those of β_A and ϕ_D , and each e_{tij} is assumed to be an i.i.d. random error with zero mean and finite variance.

For this case, we consider using $K_4(u_1, \dots, u_j) = \prod_{i=1}^j k(u_i)$ for $j = 1, 2, \dots, 5$ for the multivariate kernel function involved in $W_D^{(-t)}(t, s)$ and $W_D(t, s)$. We use the same $w(\cdot)$, $k(\cdot)$ and H_T as in Example 3.1.

First, we use the first 144 observations of the data from January 1976 to December 1987 for the selection of a best possible partially linear model. In the detailed calculation of the $MCCV(T_v)$ function, we choose $b = T = 144$, $T_c = [T^{3/4}] = 41$ and $T_v = T - T_c = 103$. The semiparametric CV1 and parametric $MCCV(T_v)$ values for model (3.2) are calculated. The combined semiparametric CV1 and parametric $MCCV(T_v)$ selection procedure then suggests using the following partially linear prediction model

$$Y_{t+5} = \hat{\beta}Y_{t+4} + \hat{\phi}(X_{t+5}, X_{t+3}), \quad 1 \leq t \leq 144, \quad (3.3)$$

where $\hat{\beta} = 0.2098$ and $\hat{\phi}(\cdot, \cdot)$ is as defined before. The optimum value for the bandwidth involved in (3.3) is $\hat{h}_1 = 0.080088$.

We also consider using the nonparametric CV selection function for the same part of the data for the case where Y_{t+i} for $1 \leq i \leq 4$ and X_{t+j} for $1 \leq j \leq 5$ are candidates of nonparametric regressors. The nonparametric CV selection function suggests the following nonparametric prediction model

$$Y_{t+5} = \hat{m}(Y_{t+4}, X_{t+5}, X_{t+3}), \quad 1 \leq t \leq 144, \quad (3.4)$$

where $\hat{m}(\cdot, \cdot, \cdot)$ is the usual nonparametric regression estimator as defined before. The optimum value for the bandwidth involved in (3.4) is $\hat{h}_2 = 0.08011$.

When we assume that the dependence of Y_{t+5} on Y_{t+i} for $1 \leq i \leq 4$ and X_{t+j} for $1 \leq j \leq 5$ is linear, the conventional AIC criterion suggests the following linear prediction model for the first part of the data

$$Y_{t+5} = \hat{\beta}_1 Y_{t+4} + \hat{\beta}_2 X_{t+5} + \hat{\beta}_3 X_{t+3}, \quad 1 \leq t \leq 144, \quad (3.5)$$

where $\hat{\beta}_1 = 0.4944$, $\hat{\beta}_2 = 0.8740$ and $\hat{\beta}_3 = -0.1923$.

We then use the second part of the data from January 1988 to December 1999 for the validation of the selected models (3.3)–(3.5). The validation supports the use of the

selected models. For the whole data set, the estimated error variances for the partially linear model (3.3), the fully nonparametric model (3.4) and the completely linear model (3.5) were 0.00935, 0.01508 and 0.02661, respectively.

REMARK 3.4. Example 3.2 shows that if a partially linear model among the possible partially linear models is an appropriate model for the data, then the combined semiparametric CV1 and $MCCV(T_v)$ selection procedure is capable of finding it. Furthermore, when using both the nonparametric CV1 selection criterion and a parametric AIC selection criterion to check whether the partially linear model (3.3) is the best possible model, both the nonparametric and parametric selection criteria support the selection of the regressors. In addition, the estimated error variance for the partially linear model is the smallest one among the partially linear model (3.3), the nonparametric regression model (3.4) and the parametric linear model (3.5). Our findings in Example 3.2 are consistent with existing studies from the Fisheries in that the relationship between the catch and the effort appears to be nonlinear while the current catch depends linearly on the past catch.

REMARK 3.5. As expected, there is no evidence of conditional heteroscedasticity in the catch–effort data. In both theory and practice, however, we need to consider the heteroscedastic case. As the homoskedasticity assumption given in Assumption A.1 is a convenient but not vital condition, we can relax it and obtain similar model selection functions and the corresponding consistency results of Theorems 2.1–2.2, but the proofs of Theorems 2.1–2.2 would be extremely technical.

REMARK 3.6. This paper only considers using the Nadaraya-Watson (NW) kernel based weight function, as the corresponding weight function based on the local polynomial kernel proposed by Fan (1992) involves multivariate polynomials, and therefore the computation of the corresponding CV functions is much more complicated than that of those based on the NW kernel in general. For Example 3.1, however, we have been able to make some comparisons among the NW, the Gasser–Müller (GM) and the local polynomial kernel (LPK) based criteria. Our empirical studies show that both the GM and the LPK based criteria support the true model selected by using the NW based criterion. Moreover, for each case the estimator of the error variance of the LPK estimator is smaller than that of the GM estimator. This is one of the properties which suggest

that the LPK estimation method is superior to the GM estimation method.

4. Discussion

In recent years, there have been growing interests in applying iterative algorithms in nonparametric and semiparametric smoothing. However, such techniques cannot provide a 'model' whose value can be calculated at a new design point with the same convenience as in linear models. Before selecting a fully nonparametric time series model for a given set of data, our research suggests using the computer-intensive semiparametric time series model selection to determine whether a partially linear time series model is more appropriate than the fully nonparametric time series model for the given set of data, as semiparametric methods can provide a 'model' with better predictive power than is available from nonparametric methods (see Example 3.2).

We acknowledge the computing expenses of the CV based selection procedure. In our detailed simulation and computing for Examples 3.1 and 3.2, we have used some optimal algorithms, such as some vectorised algorithms in the calculation of the CV1 function and the MCCV function of many possible candidates. The final computing time for each example is reasonable. We have not tried the backward or forward selection suggested by Shao (1993), although we think it might be less expensive in terms of computing time. We think that further discussion of computing algorithms is beyond the scope of this paper.

APPENDIX A

Throughout Appendices A–C, let C ($C < \infty$) denote a positive 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 the mixing coefficient $\alpha(T) = C\eta^T$, where $0 < C < \infty$ and $0 < \eta < 1$ are constants. In addition, $e_t = Y_t - E[Y_t|U_t, X_t]$ is a stationary martingale difference with respect to $\Omega_t = \sigma\{(Y_s, U_{s+1}, X_{s+1}) : 1 \leq s \leq t-1\}$, which is a sequence of σ -fields generated by $\{(Y_s, U_{s+1}, X_{s+1}) : 1 \leq s \leq t-1\}$. Suppose that $P(E[e_t^2|\Omega_t] = \sigma_0^2) = 1$, where $0 < \sigma_0^2 = E[e_t^2] < \infty$.

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 $\|\cdot\|$ denotes the Euclidean norm.

ASSUMPTION A.3. Let S_w be a compact subset of R^q and w be a weight function supported on S_w and $w \leq C$ for some constant C . For every $D \in \mathcal{D}$, let $R_{X,D} \subseteq R^{|D|} = (-\infty, \infty)^{|D|}$ be the subset such that $X_{tD} \in R_{X,D}$ and S_D be the projection of S_w in $R_{X,D}$ (that is, $S_D = R_{X,D} \cap S_w$). Assume that the marginal density function, $f_D(\cdot)$, of X_{tD} , and all the first two derivatives of $f_D(\cdot)$, $\phi_1(\cdot)$ and $\phi_{A,D}(\cdot)$, are continuous on $R_{X,D}$, and on S_D the density function $f_D(\cdot)$ is bounded below by C_D and above by C_D^{-1} for some $C_D > 0$, where $\phi_1(x) = E[Y_t | X_{tD} = x]$ and $\phi_{A,D}(x) = E[U_{tA} | X_{tD} = x]$ for every $A \in \mathcal{A}$ and $D \in \mathcal{D}$.

ASSUMPTION A.4. There exist absolute constants $0 < C_1 < \infty$ and $0 < C_2 < \infty$ such that for any integer $l \geq 1$

$$\sup_x \sup_{A \in \mathcal{A}, D \in \mathcal{D}} E \left\{ |Y_t - E[Y_t | (U_{tA}, X_{tD})]|^l | X_{tD} = x \right\} \leq C_1 \text{ and } \sup_x \sup_{A \in \mathcal{A}, D \in \mathcal{D}} E \left\{ \|U_{tA}\|^l | X_{tD} = x \right\} \leq C_2.$$

ASSUMPTION A.5. (i) For η_{tA} and η_t defined in (2.11), let $\eta_A = (\eta_{1A}, \dots, \eta_{TA})^\tau$ and $\eta = (\eta_1, \dots, \eta_T)^\tau$. Assume that when \mathcal{M}_A is in Category I,

$$\liminf_{T \rightarrow \infty} \frac{1}{T} (\eta\beta)^\tau (I_T - \eta_A (\eta_A^\tau \eta_A)^+ \eta_A^\tau) \eta\beta > 0 \text{ in probability.}$$

(ii) As $T \rightarrow \infty$,

$$\frac{T_v}{T} \rightarrow 1, T_c = T - T_v \rightarrow \infty \text{ and } \frac{T^2}{T_c^2 b} \rightarrow 0.$$

REMARK A.1. Assumptions A.1–A.4 are standard in this kind of problem. See (A.1) of Cheng and Tong (1993). Due to Assumption A.3, we do not need to assume that the marginal density of X_t has a compact support. Assumptions A.2–A.4 are a set of extensions of some existing conditions to the α -mixing time series case. See for example, (A)–(E) of Zhang (1991), (A2)–(A5) of Cheng and Tong (1993), and (C.2)–(C.5) of Vieu (1994). As pointed out in Remark 2.4(ii), when X_t and U_t are independent, Assumption A.5(i) imposes only an asymptotic and minimal model identifiability condition on the linear component. This means that Assumption A.5(i) is a natural extension of condition (2.5) of Shao (1993) to the semiparametric time series setting. Assumption A.5(ii) corresponds to conditions (3.12) and (3.22) of Shao (1993) for the linear model case. In addition, Assumption A.5(i) is also equivalent to Assumption C of Zhang (1993) for the linear model case.

APPENDIX B

In this appendix, we provide a detailed proof for Theorem 2.1. The following lemmas are required to prove Theorem 2.1.

LEMMA B.1. *Under the conditions of Theorem 2.1, we have for every $D \in \mathcal{D}$*

$$\text{CV1}(D, h) = \frac{1}{T} \sum_{t=1}^T e_{tD_0}^2 w(X_t) + V(D, h) + o_p(V(D, h)),$$

where for every $D \in \mathcal{D}_1$ and $h \in H_{TD}$

$$V(D, h) = a_1(D, h) \frac{1}{Th^{|D|}} + a_2(D, h)h^4 + o_p(V(D, h)),$$

in which $a_1(D, h)$ and $a_2(D, h)$ are positive constants depending only on (D, h) , and for every $D \in \mathcal{D}$, $D \notin \mathcal{D}_1$, and $h \in H_{TD}$

$$V(D, h) = E \left\{ [U_t^T (\beta(D) - \beta(D_0)) + \phi_D(X_{tD}) - \phi_{D_0}(X_{tD_0})]^2 w(X_t) \right\} + o_p(1).$$

Then, we have

$$\text{CV1}(D) = \inf_{h \in H_{TD}} \text{CV1}(D, h) = \frac{1}{T} \sum_{t=1}^T e_{tD_0}^2 w(X_t) + R(D) + o_p(1),$$

where e_{tD_0} is as defined in (2.4), for $D \in \mathcal{D}_1$

$$R(D) = C_D T^{-\frac{4}{4+|D|}} + o_p(T^{-\frac{4}{4+|D|}}),$$

where C_D is a positive constant depending only on D , and for $D \in \mathcal{D}$ but $D \notin \mathcal{D}_1$,

$$R(D) = E \left\{ [U_t^T (\beta(D) - \beta(D_0)) + \phi_D(X_{tD}) - \phi_{D_0}(X_{tD_0})]^2 w(X_t) \right\} + o_p(1).$$

The following lemmas are needed to complete the proof of Lemma B.1.

LEMMA B.2. *Under the conditions of Theorem 2.1, we have*

$$\delta(D, h)(\hat{\beta}(D, h) - \beta(D)) = o_p(1) \tag{B.1}$$

uniformly over $D \in \mathcal{D}$ and $h \in H_{TD}$, where $\delta(D, h) = \max\{(Th^{|D|})^{1/2}, h^{-2}\}$.

Proof. It follows from (2.7) that

$$\hat{\beta}(D, h) - \beta(D) = (\tilde{\Sigma}(D, h))^+ \sum_{t=1}^T \tilde{U}_t(D, h) e_{tD}$$

$$+(\tilde{\Sigma}(D, h))^+ \sum_{t=1}^T \tilde{U}_t(D, h)(\hat{\phi}_{2t}(X_{tD}, h) - \phi_2(X_{tD}))^\tau \beta(D) + (\tilde{\Sigma}(D, h))^+ \sum_{t=1}^T \tilde{U}_t(D, h)(\phi_1(X_{tD}) - \hat{\phi}_{1t}(X_{tD}, h)).$$

In order to prove (B.1), it suffices to show that as $T \rightarrow \infty$

$$\sqrt{T} \Delta_D \left(\hat{\beta}(D, h) - E[\hat{\beta}(D, h)] \right) \rightarrow N(0, E[e_{tD}^2 \xi_{tD} \xi_{tD}^\tau])$$

and

$$E[\hat{\beta}(D, h)] - \beta(D) = O(h^4) + O(h^2(T h^{|D|})^{-1/2}) = o(\delta(D, h)^{-1}), \quad (\text{B.2})$$

where $\xi_{tD} = U_t - E[U_t | X_{tD}]$ and $\Delta_D = E[\xi_{tD} \xi_{tD}^\tau]$ is as defined in Assumption 2.1(i).

In order to prove (B.2), it suffices to show that

$$\frac{1}{T} \tilde{\Sigma}(D, h) = \frac{1}{T} \sum_{t=1}^T \tilde{U}_t(D, h) \tilde{U}_t(D, h)^\tau \rightarrow_p \Delta_D, \quad (\text{B.3})$$

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \xi_{tD} e_{tD} \rightarrow N(0, E[e_{tD}^2 \xi_{tD} \xi_{tD}^\tau]), \quad (\text{B.4})$$

$$\frac{1}{T} \sum_{t=1}^T (\phi_1(X_{tD}) - \hat{\phi}_{1t}(X_{tD}, h))^2 = o_p(\delta(D, h)^{-1}), \quad (\text{B.5})$$

$$\frac{1}{T} \sum_{t=1}^T (\phi_2(X_{tD}) - \hat{\phi}_{2t}(X_{tD}, h))(\phi_2(X_{tD}) - \hat{\phi}_{2t}(X_{tD}, h))^\tau = o_p(\delta(D, h)^{-1}), \quad (\text{B.6})$$

$$\frac{1}{T} \sum_{t=1}^T (\phi_2(X_{tD}) - \hat{\phi}_{2t}(X_{tD}, h)) r_{tD} = o_p(\delta(D, h)^{-1}), \quad (\text{B.7})$$

$$\frac{1}{T} \sum_{t=1}^T (\phi_1(X_{tD}) - \hat{\phi}_{1t}(X_{tD}, h)) \xi_{tD} = o_p(\delta(D, h)^{-1}), \quad (\text{B.8})$$

where $r_{tD} = e_{tD}$ or ξ_{tD}^τ and Δ_D are as defined above.

The proofs of (B.5)–(B.8) are standard. The details are similar to Lemma A.2(ii) of Gao and Yee (2000). In the proof of (B.5)–(B.8), Proposition 14.1 of Cheng and Tong (1993) is used repeatedly. Using Assumptions A.1 and A.4 and applying the fact that $E[\xi_{tD} e_{tD}] = E\{\xi_{tD} E[e_{tD} | (U_t, X_{tD})]\} = 0$, we can prove (B.4) by applying the classical martingale limit theorem (see Lemma 3.3 of Gao and Liang 1995 for example). The proof of (B.3) follows from Assumption A.4, (B.6), (B.7), the Cauchy-Schwarz inequality and

$$\frac{1}{T} \sum_{t=1}^T \tilde{U}_t(D, h) \tilde{U}_t(D, h)^\tau = \frac{1}{T} \sum_{t=1}^T \xi_{tD} \xi_{tD}^\tau + \frac{1}{T} \sum_{t=1}^T (\phi_2(X_{tD}) - \hat{\phi}_{2t}(X_{tD}, h))(\phi_2(X_{tD}) - \hat{\phi}_{2t}(X_{tD}, h))^\tau$$

$$+\frac{1}{T} \sum_{t=1}^T (\phi_2(X_{tD}) - \hat{\phi}_{2t}(X_{tD}, h)) \xi_{tD}^\tau + \frac{1}{T} \sum_{t=1}^T \xi_{tD} (\phi_2(X_{tD}) - \hat{\phi}_{2t}(X_{tD}, h))^\tau.$$

The first term then converges to Δ_D in probability and the other terms converge to zero in probability due to the use of some convergence results of nonparametric estimates (see Robinson 1983 for example).

LEMMA B.3. (i) *Under the conditions of Theorem 2.1, we have for every given $D \in \mathcal{D}_1$ and $h \in H_{TD}$*

$$V_1(D, h) = \frac{1}{T} \sum_{t=1}^T \left\{ \phi_1(X_{tD}) - \hat{\phi}_{1t}(X_{tD}, h) \right\}^2 w(X_t) = d_1(D, h) \frac{1}{Th^{|D|}} + d_2(D, h) h^4 + o_p\{V_1(D, h)\}, \quad (B.9)$$

$$\begin{aligned} V_2(D, h) &= \frac{1}{T} \sum_{t=1}^T \left(\phi_2(X_{tD}) - \hat{\phi}_{2t}(X_{tD}, h) \right) \left(\phi_2(X_{tD}) - \hat{\phi}_{2t}(X_{tD}, h) \right)^\tau w(X_t) \\ &= d_3(D, h) \frac{1}{Th^{|D|}} + d_4(D, h) h^4 + o_p\{V_2(D, h)\}, \end{aligned} \quad (B.10)$$

where $\{d_i(D, h) : 1 \leq i \leq 2\}$ are positive constants and $\{d_j(D, h) : 3 \leq j \leq 4\}$ are positive definite matrices.

(ii) *Under the conditions of Theorem 2.1, we have for every given $D \in \mathcal{D}$, $D \notin \mathcal{D}_1$ and $h \in H_{TD}$*

$$V_1(D, h) = E \left\{ [\phi_1(X_{tD}) - \phi_1(X_{tD_0})]^2 w(X_t) \right\} + o_p(1), \quad (B.11)$$

$$V_2(D, h) = E \left\{ (\phi_2(X_{tD}) - \phi_2(X_{tD_0})) (\phi_2(X_{tD}) - \phi_2(X_{tD_0}))^\tau w(X_t) \right\} + o_p(1). \quad (B.12)$$

Proof. We prove only (B.9) and (B.11) and the others follow similarly. In order to prove (B.9) and (B.11), it suffices to show that for $D \in \mathcal{D}_1$ and $h \in H_{TD}$

$$\bar{V}_1(D, h) = \frac{1}{T} \sum_{t=1}^T \left\{ \hat{\phi}_1(X_{tD}, h) - \phi_1(X_{tD}) \right\}^2 w(X_t) = d_1(D, h) \frac{1}{Th^{|D|}} + d_2(D, h) h^4 + o_p(\bar{V}_1(D, h)), \quad (B.13)$$

and for $D \notin \mathcal{D}_1$ and $h \in H_{TD}$

$$\bar{V}_1(D, h) = E \left\{ [\phi_1(X_{tD}) - \phi_1(X_{tD_0})]^2 w(X_t) \right\} + o_p(1) \quad (B.14)$$

and

$$\sup_{D \in \mathcal{D}} \sup_{h \in H_{TD}} \frac{|\bar{V}_1(D, h) - V_1(D, h)|}{\bar{V}_1(D, h)} = o_p(1), \quad (B.15)$$

where $\hat{\phi}_1(X_{tD}, h) = \sum_{s=1}^T W_{sD}(X_{tD}, h) Y_s$.

Similar to the proofs of Lemmas 14.7 and 14.4 of Cheng and Tong (1993), equations (B.13)–(B.15) can be proved. Similar lemmas for the i.i.d. case and the time series case can be found in equations (5.3) and (5.4) of Vieu (1994), and Lemmas 2 and 8 of Härdle and Vieu (1992), respectively.

LEMMA B.4. *Under the conditions of Theorem 2.1, we have*

$$\begin{aligned} V(D, h) &= \frac{1}{T} \sum_{t=1}^T \left\{ \left[U_t^\tau \hat{\beta}(D, h) + \hat{\phi}_t(X_{tD}, \hat{\beta}(D, h)) \right] - \left[U_t^\tau \beta(D_0) + \phi_{D_0}(X_{tD_0}) \right] \right\}^2 w(X_t) \\ &= d_5(D, h) \frac{1}{T|h^{|D|}} + d_6(D, h)h^4 + o_p(V(D, h)) \end{aligned} \quad (B.16)$$

for every $D \in \mathcal{D}_1$ and $h \in H_{TD}$, and

$$V(D, h) = E \left\{ \left[U_t^\tau (\beta(D) - \beta(D_0)) + \phi_D(X_{tD}) - \phi_{D_0}(X_{tD_0}) \right]^2 w(X_t) \right\} + o_p(1) \quad (B.17)$$

for every $D \in \mathcal{D}$, $D \notin \mathcal{D}_1$ and $h \in H_{TD}$, where $d_5(D, h)$ and $d_6(D, h)$ are positive constants only depending on (D, h) .

Proof. Obviously,

$$\begin{aligned} V(D, h) &= \frac{1}{T} \sum_{t=1}^T \left\{ U_t^\tau \left(\hat{\beta}(D, h) - \beta(D_0) \right) \right\}^2 w(X_t) + \frac{1}{T} \sum_{t=1}^T \left\{ \hat{\phi}_t(X_{tD}, \hat{\beta}(D, h)) - \phi_{D_0}(X_{tD_0}) \right\}^2 w(X_t) \\ &\quad + \frac{2}{T} \sum_{t=1}^T \left\{ U_t^\tau \left(\hat{\beta}(D, h) - \beta(D_0) \right) \right\} \left\{ \hat{\phi}_t(X_{tD}, \hat{\beta}(D, h)) - \phi_{D_0}(X_{tD_0}) \right\} w(X_t) \\ &\equiv V(D, h)_1 + V(D, h)_2 + V(D, h)_3, \end{aligned} \quad (B.18)$$

where the symbol " \equiv " indicates that the terms of the left-hand side are represented by those of the right-hand side correspondingly.

Similar to the proof of Lemmas B.2 and B.3, we have for every $D \in \mathcal{D}_1$ and $h \in H_{TD}$

$$V(D, h)_2 = V_1(D, h) + \beta(D)^\tau V_2(D, h) \beta(D) + o_p(V(D, h)_2), \quad (B.19)$$

and

$$\sup_{D \in \mathcal{D}} \sup_{h \in H_{TD}} \frac{V(D, h)_1}{V(D, h)_2} = o_p(1), \quad \sup_{D \in \mathcal{D}} \sup_{h \in H_{TD}} \frac{V(D, h)_3}{V(D, h)_2} = o_p(1). \quad (B.20)$$

On the other hand, using Lemmas B.2 and B.3 again, we have for every $D \in \mathcal{D}$, $D \notin \mathcal{D}_1$ and $h \in H_{TD}$

$$V(D, h) = E \left\{ \left[U_t^\tau (\beta(D) - \beta(D_0)) + \phi_D(X_{tD}) - \phi_{D_0}(X_{tD_0}) \right]^2 w(X_t) \right\} + o_p(1). \quad (B.21)$$

Therefore, equations (B.18)–(B.21) complete the proof of (B.16) and (B.17).

Proof of Lemma B.1. It follows from the definition of $\text{CV1}(D, h)$ that

$$\begin{aligned}\text{CV1}(D, h) &= \frac{1}{T} \sum_{t=1}^T \left\{ Y_t - U_t^\tau \hat{\beta}(D, h) - \hat{\phi}_t(X_{tD}, \hat{\beta}(D, h)) \right\}^2 w(X_t) \\ &\equiv \frac{1}{T} \sum_{t=1}^T e_{tD_0}^2 w(X_t) + V(D, h) + R(D, h),\end{aligned}\tag{B.22}$$

where $R(D, h) = \frac{2}{T} \sum_{t=1}^T \left\{ U_t^\tau \left(\beta(D_0) - \hat{\beta}(D, h) \right) + \phi_{D_0}(X_{tD_0}) - \hat{\phi}_t(X_{tD}, \hat{\alpha}(D, h)) \right\} e_{tD_0} w(X_t)$.

Analogous to the proof of (14.25) of Cheng and Tong (1993) (see also (A.25) of Gao and Yee 2000), we have

$$\sup_{D \in \mathcal{D}} \sup_{h \in H_{TD}} \frac{|R(D, h)|}{V(D, h)} = o_p(1).\tag{B.23}$$

Thus, equations (B.22) and (B.23) imply for every $D \in \mathcal{D}$ and $h \in H_{TD}$

$$\text{CV1}(D, h) = \frac{1}{T} \sum_{t=1}^T e_{tD_0}^2 w(X_t) + V(D, h) + o_p(V(D, h)).\tag{B.24}$$

Therefore, Lemma B.4 and equation (B.24) imply Lemma B.1.

Proof of Theorem 2.1. Since equation (B.24) holds for every $D \in \mathcal{D}$, we have that there exists $\bar{h}_D \in H_{TD}$ such that

$$\text{CV1}(D, \bar{h}_D) = \inf_{h \in H_{TD}} \text{CV1}(D, h)$$

and

$$\text{CV1}(D) = \text{CV1}(D, \bar{h}_D) = \inf_{h \in H_{TD}} \text{CV1}(D, h) = \frac{1}{T} \sum_{t=1}^T e_{tD_0}^2 w(X_t) + C_D T^{-\frac{4}{4+|D|}} + o_p\left(T^{-\frac{4}{4+|D|}}\right)\tag{B.25}$$

for every $D \in \mathcal{D}_1$, where C_D is a positive constant possibly depending on D .

Using the fact that Assumption 2.1 implies $|D| > |D_0|$ for every $D \in \mathcal{D}_1$, by (B.25) we have as $T \rightarrow \infty$

$$\begin{aligned}P(\text{CV1}(D) > \text{CV1}(D_0)) &= P\left(T^{\frac{4}{4+|D_0|}}(\text{CV1}(D) - \text{CV1}(D_0)) > 0\right) \\ &= P\left(C_D T^{\frac{4(|D|-|D_0|)}{(4+|D|)(4+|D_0|)}} - C_{D_0} + o_p\left(T^{\frac{4(|D|-|D_0|)}{(4+|D|)(4+|D_0|)}}\right) > 0\right) \rightarrow 1.\end{aligned}\tag{B.26}$$

On the other hand, for every $D \in \mathcal{D}$ but $D \notin \mathcal{D}_1$, we obtain by (B.24) and (B.17) that there exists a positive constant $\pi(D, D_0)$ depending only on (D, D_0) such that

$$\text{CV1}(D) - \text{CV1}(D_0) \rightarrow \pi(D, D_0) > 0\tag{B.27}$$

in probability as $T \rightarrow \infty$.

Each of (B.26) and (B.27) implies

$$\lim_{T \rightarrow \infty} P(\hat{D}_0 = D_0) = 1. \quad (\text{B.28})$$

Furthermore, similar to the proof of (2.3) of Härdle, Hall and Marron (1988), using equations (B.25) and (B.28) we can show that as $T \rightarrow \infty$

$$\frac{\hat{h}}{h_0} \rightarrow_p 1,$$

where $\hat{h} = \bar{h}_{\hat{D}_0}$ and $h_0 = C_{D_0} T^{-\frac{1}{4+|D_0|}}$.

The proof of Theorem 2.1 is finally completed.

Proof of Corollary 2.1. It is a special case of Theorem 2.1.

Proof of Corollary 2.2. The proof is similar to the proofs of Lemmas B.1 and B.4.

Proof of Theorem 2.2. In view of the proofs of Theorems 1 and 2 of Shao (1993), in order to prove Theorem 2.2, it suffices to show that conditions (3.3) and (3.4) of Shao (1993) hold in probability with respect to the probability measure of (Y_t, U_t, X_t) and that condition (3.21) of Shao (1993) holds in probability with respect to both the probability measure of (Y_t, U_t, X_t) and the random selection of \mathcal{R} . Condition (2.5) of Shao (1993) now corresponds to condition (2.16) for our case. In other words, we need to prove (2.16) and the following conditions:

$$\max_{S \in \mathcal{R}} \left\| \frac{1}{T_v} \sum_{t \in S} V_t V_t^\tau - \frac{1}{T_c} \sum_{t \in S^c} V_t V_t^\tau \right\| = o_p(1), \quad (\text{B.29})$$

$$V^\tau V = O_p(T) \quad \text{and} \quad (V^\tau V)^{-1} = O_p(T^{-1}), \quad (\text{B.30})$$

$$\lim_{T \rightarrow \infty} \max_{t \leq T} p_{tA} = 0 \quad \text{for any } A \in \mathcal{A}, \quad (\text{B.31})$$

where p_{tA} is the t th diagonal element of the projection matrix P_A defined in (2.14).

The proofs of (2.16) and (B.29)–(B.31) are relegated to Appendix C below. In view of the conditions of Theorem 2.2, we modify some parts of the proofs of Theorems 1 and 2 of Shao (1993). For example, n_c , n_v , n and the term $o\left(\frac{n_c}{n}\right)$ involved in the proofs of Theorems 1 and 2 of Shao (1993) need to be replaced by T_c , T_v , T and $o_p\left(\frac{T_c}{T}\right)$ respectively. Some notational changes are incurred. Note also that under Assumption A.1, $E[\epsilon^\tau P_A \epsilon] = E\{E[\epsilon^\tau P_A \epsilon | V]\} = d_A \sigma_\epsilon^2$ is used in the proof of Theorem 2.2.

In view of (A.13)–(A.15) of Shao (1993), we need to show that as $T \rightarrow \infty$

$$\sum_{t=1}^T p_{tA} \epsilon_t^2 = O_p(1) \text{ and } \sum_{s,t \in S, s \neq t} p_{stA} \epsilon_s \epsilon_t = O_p(1), \quad (\text{B.32})$$

where p_{stA} is the (s, t) th element of P_A of (2.14). The proof of (B.32) is relegated to Appendix C. The outline of the proof is now completed.

Proof of Theorem 2.3. For the detailed proof, one needs to modify the proof of Theorems 2.1 and then can show that for each $A \in \mathcal{A}$

$$\lim_{T \rightarrow \infty} P \left(\hat{D}_0(A) = D_0(A) \right) = 1 \text{ and } \frac{\hat{h}(A)}{h_0(A)} \rightarrow_p 1 \quad (\text{B.33})$$

as $T \rightarrow \infty$, where $h_0(A) = C_{D_0(A)} T^{-\frac{1}{4+|D_0(A)|}}$. The proof of Theorem 2.3 then follows from Theorem 2.2(iii) and (B.33).

APPENDIX C

This appendix supplements the proofs of (2.16) and (B.29)–(B.32).

Proof of (2.16). It follows from (2.11) that

$$V_A = \eta_A + \delta_A \text{ and } V = \eta + \delta,$$

where $\delta_A = (\delta_{1A}, \dots, \delta_{TA})^\tau$ and $\delta = (\delta_1, \dots, \delta_T)^\tau$.

To prove (2.16), it now suffices to show that as $T \rightarrow \infty$

$$\frac{1}{T} \eta_A^\tau \eta_A \rightarrow_p E[\eta_A^\tau \eta_A], \quad \frac{1}{T} \eta^\tau \eta \rightarrow_p E[\eta^\tau \eta], \quad (\text{C.1})$$

$$T(\eta_A^\tau \eta_A)^+ \rightarrow_p (E[\eta_A^\tau \eta_A])^+, \quad T(\eta^\tau \eta)^+ \rightarrow_p (E[\eta^\tau \eta])^+, \quad (\text{C.2})$$

$$\frac{1}{T} V_A^\tau V_A \rightarrow_p E[\eta_A^\tau \eta_A], \quad \frac{1}{T} V^\tau V \rightarrow_p E[\eta^\tau \eta], \quad (\text{C.3})$$

$$T(V_A^\tau V_A)^+ \rightarrow_p (E[\eta_A^\tau \eta_A])^+, \quad T(V^\tau V)^+ \rightarrow_p (E[\eta^\tau \eta])^+, \quad (\text{C.4})$$

$$\frac{1}{T} (\delta\beta)^\tau (\delta\beta) \rightarrow_p 0. \quad (\text{C.5})$$

The detailed proofs of (C.1)–(C.5) are similar to that of Lemma B.2.

Proof of (B.29). Observe that

$$\left\| \frac{1}{T_v} \sum_{t \in S} V_t V_t^\tau - \frac{1}{T_c} \sum_{t \in S^c} V_t V_t^\tau \right\|$$

$$\leq \left\| \frac{1}{T_v} \sum_{t \in S} (V_t V_t^\tau - E[V_1 V_1^\tau]) \right\| + \left\| \frac{1}{T_c} \sum_{t \in S^c} (V_t V_t^\tau - E[V_1 V_1^\tau]) \right\|.$$

Let $\zeta_t = V_t V_t^\tau - E[V_1 V_1^\tau]$. To prove (B.29), it suffices to show that as $T \rightarrow \infty$

$$\max_{S \in \mathcal{R}} \left\| \sum_{t \in S} \zeta_t \right\| = o_p(T_v) \text{ and } \max_{S \in \mathcal{R}} \left\| \sum_{t \in S^c} \zeta_t \right\| = o_p(T_c).$$

We prove only the first one, as the proof of the other follows similarly.

Let $\zeta'_t = \zeta_t I[|\zeta_t| \leq T_v^{1/2}]$ and $\zeta''_t = \zeta_t I[|\zeta_t| > T_v^{1/2}]$. For any given constant $\xi > 0$, applying Lemma 3.1 of Boente and Fraiman (1988) one can have

$$P \left(\left\| \max_{S \in \mathcal{R}} \sum_{t \in S} (\zeta'_t - E[\zeta'_t]) \right\| > \xi T_v \right) \leq \max_{S \in \mathcal{R}} P \left(\left\| \sum_{t \in S} (\zeta'_t - E[\zeta'_t]) \right\| > \xi T_v \right) \leq C_1 b \exp(-C_2 \xi^{1/2} T_v^{1/4}). \quad (C.6)$$

For any given constant $\xi > 0$, we have

$$\begin{aligned} P \left(\left\| \max_{S \in \mathcal{R}} \sum_{t \in S} (\zeta''_t - E[\zeta''_t]) \right\| > \xi T_v \right) &\leq \max_{S \in \mathcal{R}} P \left(\left\| \sum_{t \in S} (\zeta''_t - E[\zeta''_t]) \right\| > \xi T_v \right) \\ &\leq C T_v^{-1} \max_{S \in \mathcal{R}} \sum_{t \in S} E \|\zeta''_t\| \leq C T_v^{-1} \max_{S \in \mathcal{R}} \sum_{t \in S} E \left(\frac{\|\zeta_t\|^6}{\|\zeta_t\|^5} I[|\zeta_t| > T_v^{1/2}] \right) \\ &\leq C T_v^{-1} \max_{S \in \mathcal{R}} \sum_{t \in S} E[|\zeta_t|^6] T_v^{-5/2} \leq C b T_v^{-5/2}, \end{aligned} \quad (C.7)$$

using Assumption A.4.

Equations (C.6) and (C.7) imply

$$\sum_{T_v=1}^{\infty} P \left(\left\| \max_{S \in \mathcal{R}} \sum_{t \in S} \zeta_t \right\| > \xi T_v \right) < \infty. \quad (C.8)$$

Equation (C.8) implies that $\lim_{T_v \rightarrow \infty} \frac{1}{T_v} \max_{S \in \mathcal{R}} \|\sum_{t \in S} \zeta_t\| = 0$ holds with probability one. Thus, equation (B.29) holds in probability.

REMARK B.1. Note that we have actually shown that (B.29) holds with probability one. As a result, we may conclude that if x_1, \dots, x_n are either independent random variables or stationary time series, equation (3.11) of Shao (1993) holds with probability one.

Proof of (B.30) and (B.31). The proof follows from (C.1)–(C.4) above.

Proof of (B.32). Note that

$$E \left[\sum_{t \in S} p_{tA} \epsilon_t^2 \right]^2 = \sum_{t \in S} E \left[p_{tA}^2 \epsilon_t^4 \right] + \sum_{s \neq t \in S} E \left[p_{sA} p_{tA} \epsilon_s^2 \epsilon_t^2 \right]$$

$$\begin{aligned}
&= \sum_{t \in \mathcal{S}} E \left\{ E \left[p_{tA}^2 \epsilon_t^4 | V \right] \right\} + \sum_{s \neq t \in \mathcal{S}} E \left\{ E \left[p_{sA} p_{tA} \epsilon_s^2 \epsilon_t^2 | V \right] \right\} \\
&\leq C_1(\epsilon) E \left[\sum_{t \in \mathcal{S}} p_{tA}^2 \right] + C_2 E \left[\sum_{s \neq t \in \mathcal{S}} p_{sA} p_{tA} \right] \leq C,
\end{aligned} \tag{C.9}$$

using Assumptions A.1 and A.4, and $\sum_{t=1}^T p_{tA} = d_A$.

Equation (C.9) and the following central limit theorem

$$\frac{\sum_{t \in \mathcal{S}} p_{tA} \epsilon_t^2 - E[p_{tA} \epsilon_t^2]}{\sqrt{\text{Var}(\sum_{t \in \mathcal{S}} p_{tA} \epsilon_t^2)}} \rightarrow N(0, 1) \text{ as } T \rightarrow \infty$$

imply the first part of (B.32).

Similar to the proof of Theorem 2.1 of Gao and Anh (2000), we have that as $T \rightarrow \infty$

$$\frac{\sum_{s \neq t \in \mathcal{S}} p_{sA} p_{tA} \epsilon_s \epsilon_t}{\sqrt{\text{Var}(\sum_{s \neq t \in \mathcal{S}} p_{sA} p_{tA} \epsilon_s \epsilon_t)}} \rightarrow N(0, 1). \tag{C.10}$$

In the detailed proof of (C.10), the mixing condition assumed in Assumption A.1 is used. Details are similar to (A.17) and (A.18) of Gao and Anh (2000). Thus the second part of (B.32) is proved.

Acknowledgements. The first author would like to thank the Australian Research Council for its financial support. The second author acknowledges financial support from the BBSRC/EPSRC of UK and the Hong Kong Research Grants Council.

REFERENCES

- BICKEL, P. AND ZHANG, P. (1992) Variable selection in nonparametric regression with categorical covariates. *J. Amer. Statist. Assoc.*, **87**, 90–97.
- BOENTE, G. AND FRAIMAN, R. (1988). Consistency of a nonparametric estimate of a density function for dependent variables. *J. Multi. Anal.*, **25**, 90–99.
- CHEN, R., LIU, J. AND TSAY, R. (1995) Additivity tests for nonlinear autoregression. *Biometrika*, **82**, 369–383.
- CHEN, R. AND TSAY, R. (1993) Nonlinear additive ARX models. *J. Amer. Statist. Assoc.*, **88**, 955–967.
- CHENG, B. AND TONG, H. (1992) On consistent nonparametric order determination and chaos. *J.*

Roy. Statist. Soc. Ser. B **54** 427–449.

CHENG, B. AND TONG, H. (1993) Nonparametric function estimation in noisy chaos. *Developments in Time Series Analysis* (ed. T. Subba Rao), 183–206. Chapman and Hall, London.

FAN, J. (1992) Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.*, **87**, 998–1004.

FAN, J. AND GIJBELS, I. (1996) *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

GAO, J. (1998) Semiparametric regression modelling of nonlinear time series. *Scand. J. Statist.*, **25**, 521–539.

GAO, J. AND ANH, V. (2000) A central limit theorem for a random quadratic form of strictly stationary processes. *Statist. & Probab. Lett.*, **49**, 69–79.

GAO, J. AND LIANG, H. (1995) Asymptotic normality of pseudo-LS estimator for partially linear autoregressive models. *Statist. Probab. Lett.*, **23**, 27–34.

GAO, J. AND TONG, H. (2004) Semiparametric nonlinear time series model selection. *J. Roy. Statist. Soc. Ser. B*, **66**, 321–336.

GAO, J., TONG, H. AND WOLFF, R. (2002a) Adaptive series estimation in additive stochastic regression models. *Statistica Sinica*, **12**, 409–428.

GAO, J., TONG, H. AND WOLFF, R. (2002b) Model specification tests in nonparametric stochastic regression models. *J. Multivariate. Anal.* **83**, 324–359.

GAO, J., WOLFF, R., ANH, V. (2001) Semiparametric methods in multivariate approximation problems. *J. Complexity*, **17**, 754–772.

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. (1988) How far are automatically chosen regression smoothing parameters from their optimum (with discussion) ? *J. Amer. Statist. Assoc.*, **83**, 86–99.

HÄRDLE, W., LIANG, H. AND GAO, J. (2000) *Partially Linear Models*. Springer Series in Contributions to Statistics. Physica-Verlag, New York.

HÄRDLE, W., LÜTKEPOHL, H. AND CHEN, R. (1997) A review of nonparametric time series analysis. *International Statist. Rev.*, **65**, 49–72.

HÄRDLE, W. AND VIEU, P. (1992) Kernel regression smoothing of time series. *J. Time Ser. Anal.*, **13**, 209–232.

MASRY, E. AND TJØSTHEIM, D. (1995) Nonparametric estimation and identification of nonlinear ARCH time series. *Econometric Theory* **11**, 258–289.

- MASRY, E. AND TJØSTHEIM, D. (1997) Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, **13**, 214–252.
- ROBINSON, P. (1983) Nonparametric estimation for time series models. *J. Time Ser. Anal.*, **4**, 185–208.
- ROBINSON, P. (1988) Root-N-consistent semiparametric regression. *Econometrica*, **56**, 931–964.
- SHAO, J. (1993) Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, **422**, 486–494.
- SHAO, J. (1997) An asymptotic theory for linear model selection (with comments). *Statistica Sinica*, **7**, 221–264.
- STONE, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike’s criterion. *J. Roy. Statist. Soc. Ser. B*, **39**, 44–47.
- TERÄSVIRTA, T., TJØSTHEIM, D. AND GRANGER, C. W. J. (1994) Aspects of modelling nonlinear time series, in R. F. Engle and D. L. McFadden (eds), *Handbook of Econometrics*, **4**, 2919–2957.
- TJØSTHEIM, D. (1994) Nonlinear time series: a selective review. *Scand. J. Statist.*, **21**, 97–130.
- TJØSTHEIM, D. (1999) Nonparametric specification procedures for time series. *Asymptotics, nonparametrics, and time series*, **158**, 149–199. *Statistics: Textbooks and Monographs*. Dekker, New York.
- TJØSTHEIM, D. AND AUESTAD, B. (1994a) Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.*, **89**, 1398–1409.
- TJØSTHEIM, D. AND AUESTAD, B. (1994b) Nonparametric identification of nonlinear time series: selecting significant lags. *J. Amer. Statist. Assoc.*, **89**, 1410–1419.
- TONG, H. (1976) Fitting a smooth moving average to noisy data. *IEEE Trans. Inf. Theory*, **IT-26**, 493–496.
- TONG, H. (1990) *Nonlinear Time Series*. Oxford University Press, Oxford.
- VIEU, P. (1994) Choice of regressors in nonparametric estimation. *Computat. Statist. & Data Anal.*, **17**, 575–594.
- XIA, Y., TONG, H., LI, W. K. AND ZHU, L. (2002) An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B*, **64** (in press).
- YAO, Q. AND TONG, H. (1994) On subset selection in nonparametric stochastic regression. *Statistica Sinica*, **4**, 51–70.
- 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.
- ZHENG, X. AND LOH, W. Y. (1997) A consistent variable selection criterion for linear models with high-dimensional covariates. *Statistica Sinica*, **7**, 311–326.