MODEL SELECTION AND INFERENCE IN SEMIPARAMETRIC REGRESSION [1]

By Jiti Gao[2] and Howell Tong

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

It is known that semiparametric 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 spareness of the data. This is partly because the curse of dimensionality problem may still arise from using a semiparametric 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 components. We develop a novel cross-validation based model selection procedure for the choice of both the parametric and nonparametric regressors in semiparametric regression, and then suggest using a combination of the cross-validation based model selection and residual diagnostics in the selection of completely linear, partially linear and fully nonparametric regression models. We illustrate the cross-validation selection procedure through simulated and real data. Our small sample studies show that the proposed cross-validation selection procedure works well numerically.

1. Introduction. In multivariate regression problems one of the tasks is to study the structural relationship between the response variable Y and the vector of covariates $X = (X_1, \ldots, X_q)^\tau$ via

$$m(x) = E(Y|X = x),$$

where $x = (x_1, \ldots, x_q)^\tau$ and $m(x) = m(x_1, \ldots, x_q)$. The problem then is to fit a q–dimensional surface to the observed data $\{(X_{t1}, \ldots, X_{tq}; Y_t) : t \geq 1\}$. In practice, there is a growing number of enormous (in millions) data sets in finance, business, economics, etc., that allow us to implement multivariate surface smoothing techniques for exploring fine structural architecture, even when there are several covariates. Although surface smoothing techniques appear to be feasible, there is a serious problem: the so-called curse of dimensionality. This problem has been discussed and illustrated in many monographs, among which are Eubank (1988), Härdle (1990), Hastie and Tibshirani (1990), Wahba (1990), Green and Silverman (1994), Fan and Gijbels (1996), and Härdle, Liang and Gao (2000). Recently, several nonparametric and semiparametric approaches have been proposed to deal with the curse of dimensionality problem as well as some related problems. These include additive partially linear regression modelling and nonparametric regression variable selection. For additive partially linear regression modelling, interests focus on the construction of efficient estimation procedures for both the parametric and nonparametric components. For example, Stone (1985, 1986), Heckman (1986), Rice (1986), Robinson (1988), Speckman (1988), Bhattacharya and Zhao (1997), Fan and Gijbels (1996), Carroll, Fan, Gijbels and Wand (1997), Gao and Liang (1997), Gao and Shi (1997), Mammen and van de Geer (1997), Fan, Härdle and Mammem (1998), Shi and Tsai (1999), Liang, Härdle and Carroll (1999), Härdle, Liang and Gao (2000), and others have

[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

constructed estimation procedures based on either the kernel method, the local linear method, the orthogonal series approach or the spline approximation technique for additive partially linear models with independent and identically distributed (i.i.d.) observations; Robinson (1988), Teräsvirta, Tjøstheim and Granger (1994), Gao and Liang (1995), Gao and Tong (1999), Gao and Anh (1999), Gao and Yee (2000), Härdle, Liang and Gao (2000), Gao, Tong and Wolff (2001a, 2001b), and others have applied either the kernel approach or the orthogonal series approximation method to estimate additive semiparametric time series regression. For the choice of regressors in nonparametric regression, several papers have investigated the use of cross-validation (CV), generalized cross-validation (GCV) and some other methods to construct consistent order selections for nonparametric regression models under i.i.d. observations or dependent processes. For example, Zhang (1991), Cheng and Tong (1992, 1993), Tjøstheim and Auestad (1994a, 1994b), Vieu (1994, 1995), Yao and Tong (1994), Tjøstheim (1999), and Gao, Wolff and Anh (2001) have considered the selection of optimum regression variables for the i.i.d. case and the selection of significant lags for the time series case by employing kernel-based determination criteria.

Recently, Fan, Härdle and Mammen (1998) have provided an efficient and direct way to deal with the dimensionality reduction problem. In practice, however, before applying the additive nonparametric regression technique to model real data, a crucial problem is whether an additive nonparametric regression model is appropriate for a given set of data. In other words, we should test for nonparametric additivity before using additive nonparametric regression to model a given set of data. When an additive nonparametric regression model is not appropriate for a given set of data, we need to find alternative methods to solve the dimensionality reduction problem. As an alternative, this paper suggests combining the additive partially linear regression modelling and nonparametric regression variable selection together to deal with the dimensionality reduction problem. In theory, we can assume that the process (Y_t, U_t, X_t) satisfies the following model

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

where $U_t = (U_{t1}, \ldots, U_{tp})^\tau$ and $X_t = (X_{t1}, \ldots, X_{tq})^\tau$ are both time series, $\beta = (\beta_1, \ldots, \beta_p)^\tau$ is a vector of unknown parameters, ϕ is an unknown function defined over R^q, and the error process e_t satisfies $E[e_t] = 0$ and $0 < E[e_t^2] < \infty$. In practice, a crucial problem is how to identify U_t and X_t before applying model (1.1). For some cases, the identification problem can be solved easily by using empirical studies. For example, when modelling electricity sales, it is natural to assume the impact of temperature on electricity consumption to be nonlinear, as both high and low temperatures lead to increased consumption, whereas a linear relationship may be assumed for other regressors. See Engle, Granger, Rice and Weiss (1986). Similarly, when modelling the dependence of earnings on qualification and labour market experience, our research [see Härdle, Liang and Gao (2000)] shows that the impact of qualification on earnings to be linear, while the dependence of earnings on labour market experience appears to be nonlinear. For many other cases, the identification problem should be solved theoretically before using model (1.1). More recently, Härdle, Liang and Gao (2000, §6.2) have extended the discussion of Chen and Chen (1991) for the i.i.d. case to the time series case.

Model (1.1) covers some existing cases. See for example, Robinson (1988), Speckman (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 de-

termine 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 setting. This paper proposes a novel model selection procedure combining the leave–one–out cross-validation function for the choice of the nonparametric regressors and the leave–T_v–out cross-validation function for the choice of parametric regressors, where $T_v > 1$ is a positive interger satisfying $T_v \to \infty$ as the number of observations, T, converges to ∞. Our proposed semiparametric cross-validation (CV) based model selection procedure has the following features:

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

(ii) It extends the leave–T_v–out cross-validation (CV) selection criterion for classical linear regression (see Shao 1993, 1997; Zhang 1993) and the leave-one-out cross-validation selection criterion (see Zhang 1991) for purely nonparametric regression to the semiparametric 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. 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 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). Zhang (1991), Bickel and Zhang (1992), Cheng and Tong (1992, 1993), Vieu (1994), and Yao and Tong (1994) have studied the behavior of the CV criterion in nonparametric regression. In the following, we construct a leave–one–out cross-validation criterion for the selection of nonparametric regressors in Section 2.1 and then use a leave–T_v–out cross-validation criterion for the selection of parametric regressors in Section 2.2.

2.1. *CV criterion for nonparametric regressors.* Although the linear regression normally fails to fit nonlinear phenomena, the linear regression has still a useful role as a starting model. We therefore

suggest using a partially linear model of the form (1.1) to model nonlinear phenomena. On the one hand, we need to take all possible factors into consideration when selecting the parametric and nonparametric regressors. On the other hand, in order to avoid the computational difficulty caused by the spareness of the data and to provide more precise prediction, we need only to choose the most significant factors. For example, in searching for important factors which cause air pollution, in principle we can assume that air pollution could be caused by high wind speed, extremely high temperature and many other factors. In practice, however, only a small number of factors are often selected as the true regressors.

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

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

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

$$Y_t = U_t^\tau \beta(D) + \phi_D(X_{tD}) + e_{tD}, \tag{2.1}$$

where e_{tD} is an error process, $\beta(D) = (\beta_1(D), \ldots, \beta_p(D))^\tau$ is a vector of unknown parameters defined by

$$\beta(D) = \{E\left(U_t - E[U_t|X_{tD}]\right)\left(U_t - E[U_t|X_{tD}]\right)^\tau\}^{-1} E\left(U_t - E[U_t|X_{tD}]\right)\left(Y_t - E[Y_t|X_{tD}]\right) \tag{2.2}$$

under Assumption 2.1(i) below, and ϕ_D is an unknown function over R^{d_D} defined by

$$\phi_D(X_{tD}) = \phi_D(X_{tD}, \beta(D)) = E\left\{(Y_t - U_t^\tau \beta(D))|X_{tD}\right\}. \tag{2.3}$$

For any $D \in \mathcal{D}$, define

$$\psi_D(U_t, X_{tD}) = U_t^\tau \beta(D) + \phi_D(X_{tD}),\ \Psi(U_t, X_t) = E[Y_t|U_t, X_t],$$

$$e_{tD} = Y_t - U_t^\tau \beta(D) - \phi_D(X_{tD}), \sigma_{1D}^2 = E\{e_{tD}^2\} \text{ and } \sigma_{10}^2 = E\left\{Y_t - E[Y_t|U_t, X_t]\right\}^2.$$

The following assumption imposes some existence and uniqueness conditions on the true version of D.

ASSUMPTION 2.1. (i) For every $D \in \mathcal{D}$, $\Delta_{1D} = E\left\{U_t - E[U_t|X_{tD}]\right\}\left\{U_t - E[U_t|X_{tD}]\right\}^\tau$ is a positive definite matrix with order $d_D \times d_D$.

(ii) Let $\mathcal{D}_1 = \{D \in \mathcal{D}$, such that $\psi_D = \Psi\}$ and $\mathcal{D}_0 = \{D_0 \in \mathcal{D}_1$, such that $|D_0| = \min_{D \in \mathcal{D}_1} |D|\}$. Assume that D_0 is the unique element of $\mathcal{D}_0$.

REMARK 2.1. (i) We should point out that when U_t and X_t have common components, Assumption 2.1 is not satisfied. The obvious remedy for this case is to put $\beta_j(D) = 0$ when U_{tj} is equal to a component of X_{tD}.

(ii) From now on, we always assume both the existence and uniqueness of a true model. 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 that if there is another $D_2 \neq D_0$,

$D_2 \in \mathcal{D}$ such that $\sigma^2_{1D_2} = \sigma^2_{1D_0}$, then $|D_2| > |D_0|$. This defines the uniqueness of the true regression function and furthermore requires that the true nonparametric regression function is of the smallest possible dimension. For example, for the case where $p = 4$ and $\mathcal{Q} = \{1, 2, 3, 4\}$ if both

$$\sum_{i=1}^{4} U_{ti}\beta_{1i} + g_1(X_{t1}, X_{t2}) \text{ and } \sum_{i=1}^{4} U_{ti}\beta_{2i} + g_2(X_{t1}, X_{t2}, X_{t4})$$

satisfy

$$E\left\{Y_t - \sum_{i=1}^{4} U_{ti}\beta_{1i} - g_1(X_{t1}, X_{t2})\right\}^2 = E\left\{Y_t - \sum_{i=1}^{4} U_{ti}\beta_{2i} - g_2(X_{t1}, X_{t2}, X_{t4})\right\}^2 = \sigma^2_{10},$$

then only $\sum_{i=1}^{4} U_{ti}\beta_{1i} + g_1(X_{t1}, X_{t2})$ is defined as the true regression function.

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

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

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 we may define the true model as

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

where $e_{tD_0} = Y_t - E[Y_t|U_t, X_t]$.

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)]

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

where

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

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

It follows from (2.5) that

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

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

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

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

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

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

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

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

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 $CV1(D, h)$ function over $D \in \mathcal{D}$ and $h \in H_{TD}$, and written as

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

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 (3.1) of Vieu (1994), we integrate the weight function w not depending on D into CV1. Under an additional 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.

Before proposing the first theorem of this section, we state the main lemmas of this section. Their proofs are relegated to Appendix B below.

LEMMA 2.1. *Assume that Assumptions 2.1 and A.1–A.4 listed in Appendix A hold. Then*

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

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 $h \in H_{TD}$

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

LEMMA 2.2. *Assume that the conditions of Lemma 2.1 hold. Then for every $D \in \mathcal{D}_1$, $h \in H_{TD}$ and $T \to \infty$*

$$\sqrt{T}\Delta_{1D}(\hat{\beta}(D,h) - E[\hat{\beta}(D,h)]) \to N(0, E[e_{tD}^2 \xi_{tD}\xi_{tD}^\tau])$$

and

$$E[\hat{\beta}(D,h)] - \beta(D) = O(h^4) + O(h^2(Th^{|D|})^{-1/2}),$$

where $\xi_{tD} = U_t - E[U_t|X_{tD}]$ and $\Delta_{1D} = E[\xi_{tD}\xi_{tD}^\tau]$ as defined in Assumption 2.1.

REMARK 2.3. Lemma 2.2 shows that the rate of the parametric convergence is much faster than that of the nonparametric convergence. See Theorems 2 and 4 of Speckman (1988) for similar results in the i.i.d. case. Because the rate of the parametric convergence is asymptotically negligible compared with the rate of the nonparametric convergence, the proposed selection function CV1 cannot be directly applied to select parametric regressors. The discussion of selecting parametric regressors is given in Section 2.2 below.

THEOREM 2.1. *Assume that the conditions of Lemma 2.1 hold. Then*

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

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

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

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 the assumption about the existence and uniqueness is not satisfied, we have the following corollary. The practical importance of the corollary is that the selection criterion can be extended to the case where a 'true' model does not necessarily exist.

COROLLARY 2.1. *Assume that Assumptions 2.1(i) and A.1–A.4 hold. Then*

$$\lim_{T\to\infty} P(\hat{D}_0 \in \mathcal{D}_0) = 1,$$

where $\mathcal{D}_0$ is as defined in Assumption 2.1. Moreover, as $T \to \infty$

$$\frac{\hat{h}}{h_0} \to_p 1.$$

When $\beta(D) = 0$ in (2.1), we have the following result for purely nonparametric regression model selection. 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.

COROLLARY 2.2. *For the nonparametric regression case, the conclusion of Theorem 2.1 holds.*

The proofs of Theorem 2.1 and Corollaries 2.1–2.2 are relegated to Appendix B.

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

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

We now have the following corollary and its proof is given in Appendix B.

COROLLARY 2.3. *Under the conditions of Theorem 2.1, we have as* $T \to \infty$

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

Corollary 2.3 shows that the corresponding semiparametric estimator of (2.10) is asymptotically close to the true regression function.

In Section 2.1, we assume that the linear component is already compact and then propose the leave–one–out cross-validation for the selection of nonparametric regressors. In both theory and practice, we need to consider selecting parametric regressors as well when the parametric component is not compact. 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 (e.g. 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 selecting parametric regressors separately.

2.2. *CV criterion for the selection of parametric regressors.* This section considers using a cross-validation function to choose an optimum linear component for model (2.1).

As can be seen in Section 2.1, the selected $\hat{D}_0$ and $\hat{h}$ depend on $A_0 = \{1, 2, \ldots, p\}$. 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 $A_0 = \{1, 2, \ldots, p\}$. For $A \in \mathcal{A}$, let U_{tA} be a column vector consisting of $\{U_{ti} : i \in A\}$ and 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 $\beta = (\beta_1, \ldots, \beta_p)^\tau$.

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

$$\lim_{T\to\infty} P\left(\hat{D}_0(A) = D_0(A)\right) = 1 \text{ and } \frac{\hat{h}(A)}{h_0(A)} \to_p 1$$

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

For simplicity and convenience, we introduce the following notation.

$$\hat{\psi}_1(t, A) = \hat{\phi}_1\left(X_{t\hat{D}_0(A)}, \hat{h}(A)\right) = \sum_{s=1}^{T} W_{\hat{D}_0(A)}(t, s) Y_s,$$

$$\hat{\psi}_2(t, A) = \hat{\phi}_2\left(X_{t\hat{D}_0(A)}, \hat{h}(A)\right) = \sum_{s=1}^{T} W_{\hat{D}_0(A)}(t, s) U_{sA},$$

$$\eta_{tA} = U_{tA} - E[U_{tA}|X_{tD_0(A)}],\ \delta_{tA} = E[U_{tA}|X_{tD_0(A)}] - \hat{\psi}_2(t, A),$$

$$V_{tA} = \eta_{tA} + \delta_{tA} = U_{tA} - \hat{\psi}_2(t, A),\ V_A = (V_{1A}, \ldots, V_{TA})^\tau,$$

$$\hat{\psi}_1(t) = \hat{\psi}_1(t, A_0),\ \hat{\psi}_2(t) = \hat{\psi}_2(t, A_0),\ \eta_t = U_t - E[U_t|X_{tD_0}],\ \delta_t = E[U_t|X_{tD_0}] - \hat{\psi}_2(t),$$

$$V_t = \eta_t + \delta_t = U_t - \hat{\psi}_2(t), V = (V_1, \ldots, V_T)^\tau,\ Z_t = Y_t - \hat{\psi}_1(t), \text{ and } Z = (Z_1, \ldots, Z_T)^\tau, \quad (2.11)$$

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

$$Y_t = U_{tA}^\tau \beta_A + \phi_{D_0(A)}(X_{tD_0(A)}) + \epsilon_{tA}, \quad (2.12)$$

where ϵ_{tA} is an error process, might be more compact than model (2.4) given by

$$Y_t = U_t^\tau \beta(D_0) + \phi_{D_0}(X_{tD_0}) + e_{tD_0},\ D_0 \in \mathcal{D}_0.$$

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

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

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. 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.

Now the selection of A is carried out by using the data $\{(Z_t, V_t) : t = 1, 2, \ldots, T\}$ 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 $\{(Z_t, V_t) : t = 1, 2, \ldots, T\}$, the average squared prediction error is

$$L_T(A) = \frac{1}{T}\sum_{t=1}^{T}\left[Z_t - V_{tA}^\tau \hat{\beta}_A\right]^2 = \frac{1}{T}\left(Z - V_A\hat{\beta}_A\right)^\tau \left(Z - V_A\hat{\beta}_A\right)$$

$$= \frac{1}{T}\epsilon^\tau\epsilon - \frac{1}{T}\epsilon^\tau P_A\epsilon + \frac{1}{T}(V\beta)^\tau R_A(V\beta) + \frac{2}{T}\epsilon^\tau R_A(V\beta), \tag{2.14}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_T)^\tau$, $P_A = V_A\,(V_A^\tau V_A)^+ V_A^\tau$, $R_A = I_T - P_A$, 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

$$R_T(A, V) = E[L_T(A)|V] = \frac{1}{T}E[\epsilon^\tau\epsilon|V] - \frac{1}{T}E[\epsilon^\tau P_A\epsilon|V] + \frac{1}{T}E[(V\beta)^\tau R_A(V\beta)|V] + \frac{2}{T}E\left[\epsilon^\tau R_A(V\beta)|V\right]$$

$$= \sigma_\epsilon^2 - \frac{1}{T}d_A\sigma_\epsilon^2 + \Delta_{T,A}, \text{ with probability one,} \tag{2.15}$$

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

$$\liminf_{T\to\infty} \Delta_{T,A} > 0 \text{ in probability.} \tag{2.16}$$

When $\mathcal{M}_A$ is in Category II, it follows from (2.14) and (2.15) that

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

because $V\beta = V_A\beta_A$.

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

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

where $\eta = (\eta_1, \ldots, \eta_T)^\tau$, $\eta_A = (\eta_{1A}, \ldots, \eta_{TA})^\tau$, $\eta_t = U_t - E[U_t|X_{tD_0}]$ 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 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 $\{1, 2, \ldots, T\}$ 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

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

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 $\mathrm{CV}_{A,S}(T_v)$ function is called the leave–T_v–out cross-validation, abbreviated as $\mathrm{CV}(T_v) = \mathrm{CV}T_v$. From the computational point of view, the simplest $\mathrm{CV}T_v$ 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 $\mathrm{CV}T_v$ in the selection of A.

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

$$\mathrm{MCCV}(A, T_v) = \frac{1}{b} \sum_{S \in \mathcal{R}} \mathrm{CV}_{A,S}(T_v) = \frac{1}{bT_v} \sum_{S \in \mathcal{R}} \left|\left| Z_S - \hat{Z}_{A,S^c} \right|\right|^2. \tag{2.18}$$

This method is called the Monte Carlo $\mathrm{CV}T_v$, abbreviated as $\mathrm{MCCV}T_v$, 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 the conditions of Theorem 2.1 hold. In addition, Assumption A.5 holds. Then we have the following conclusions*:

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

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

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*

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

(iii) *Consequently*,

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

Let $\hat{A}$ and A_* correspond to the selected model and $\mathcal{M}_*$ respectively. Then following Theorems 2.1 and 2.2, we have the following main result.

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

$$\lim_{T\to\infty} P(\hat{A} = A_*,\ \hat{D}_0(\hat{A}) = D_0(A_*)) = 1 \text{ and } \frac{\hat{h}(\hat{A})}{h_0(A_*)} \to_p 1$$

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

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

REMARK 2.5. Sections 2.1 and 2.2 show 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 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.

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. See for example, Li (1987), Shao (1997), and Gao, Tong and Wolff (2001a). The discussion for this case is quite different and the detailed discussion is similar to 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

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

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

where 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

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

Throughout Example 3.1, we consider using $h \in H_{T1} = [0.3T^{-7/30}, 2 \cdot T^{-1/6}]$ and the following weight function

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

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 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 MCCV(T_v), we choose $b = T$, $T_v = T - T_c$ and $T_c = [T^{3/4}]$, 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.

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

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

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

$$(\eta\beta)^\tau \left(I_T - \eta_{A_i}(\eta_{A_i}^\tau \eta_{A_i})^+ \eta_{A_i}^\tau\right)(\eta\beta) = \beta_{3-i}^2 \frac{\sum_{t=3}^T \eta_{t1}^2 \sum_{t=3}^T \eta_{t2}^2 - [\sum_{t=3}^T \eta_{t1}\eta_{t2}]^2}{\sum_{t=3}^T \eta_{ti}^2} > 0$$

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 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		
subset	subset	$T=22$	$T=72$	$T=152$	$T=22$	$T=72$	$T=152$
$\{Y_{t-1}, Y_{t-2}\}$	$\{X_t, X_{t-1}\}$	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		
subset	$T=22$	$T=72$	$T=152$	$T=22$	$T=72$	$T=152$
$\{Y_{t-1}, Y_{t-2}, X_t\}$	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 decreses 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 $j = 2, 3, 4$

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

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 corresponding 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	CV value			ASE value		
subset	$T = 22$	$T = 72$	$T = 152$	$T = 22$	$T = 72$	$T = 152$
$\{Y_{t-1}, Y_{t-2}, X_t, X_{t-1}\}$	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 columin 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 commerically 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

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

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

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

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 the parametric and nonparametric regressors. 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 $1 \le i \le 15$ and the nonparametric candidates X_{tD_j} for $1 \le j \le 31$. It follows that

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

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 $K_4(u_1, \ldots, u_j) = \prod_{i=1}^{j} k(u_i)$ for $j = 1, 2, \cdots, 5$ for the multivariate kernel function involved in $W_D^{(-t)}(t, s)$ and $W_D(t, s)$. We use the same w, k and H_{T1} 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

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

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 $1 \le i \le 4$ and X_{t+j} for $1 \le j \le 5$ are candidates of nonparametric regressors. The nonparametric CV selection function suggests the following nonparametric prediction model

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

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 $1 \le i \le 4$ and X_{t+j} for $1 \le j \le 5$ is linear, the conventional AIC criterion suggests the following linear prediction model for the first part of the data

$$Y_{t+5} = \hat{\beta}_1 Y_{t+4} + \hat{\beta}_2 X_{t+5} + \hat{\beta}_3 X_{t+3},\ 1 \le t \le 144, \tag{3.5}$$

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. Moreover, we produce the corresponding plots based on the whole data set from January 1976 to December 1999 in Figure 1. Parts (a)–(c) present time plots of the common-log-transformed catch data, the common-log-transformed effort data, and the transformed catch data against the transformed effort data, respectively; parts (d), (e) and (f) give plots of the fitted values (lines) and the catch data (dots) for the partially linear model (3.3), the purely nonparametric model (3.4) and the completely linear model (3.5), respectively.

Figure 1 near here

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. Our experience suggests 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 more complicated than that of those based on the NW kernel. For Example 3.1, however, we made some comparisons among the NW, the Gasser–Müller (GM) and the local polynomial kernel (LPK) based criteria. Our 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 regression model for a given set of data, our research suggests using the computer-intensive semiparametric model selection criterion to determine whether a partially linear model is more appropriate than the fully nonparametric regression 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 haven't 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 $P(E[e_t^2|\Omega_t] = \sigma_{10}^2) = 1$, where $0 < \sigma_{10}^2 = E[e_t^2] < \infty$.

ASSUMPTION A.2. For every $E \in \mathcal{D}$, K_E is a $|E|$–dimensional symmetric, Lipschitz continuous probability kernel function with $\int ||u||^2 K_E(u)du < \infty$, and K_E 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 $E \in \mathcal{D}$, let $R_X^E \subseteq R^{|E|} = (-\infty, \infty)^{|E|}$ be the subset such that $X_{tE} \in R_X^E$ and S_E be the projection of S_ω in R_X^E. Assume that the marginal density function, f_E, of X_{tE}, and all the first two derivatives of f_E and g_{iE}, $i = 1, 2$, are continuous on R_X^E, and on S_E the density function f_E is bounded below by C_E and above by C_E^{-1} for some $C_E > 0$, where $g_{1E}(x) = E[Y_t|X_{tE} = x]$ and $g_{2E}(x) = E[U_t|X_{tE} = x]$.

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

$$\sup_x \sup_{E \subset \mathcal{R}} E\left\{|Y_t - E[Y_t|(U_t, X_t)]|^l \, |X_{tE} = x\right\} \leq C_1 \text{ and } \sup_x \sup_{E \subset \mathcal{R}} E\left\{||U_t||^l | X_{tE} = x\right\} \leq C_2.$$

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

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

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

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

REMARK A.1. Assumption A.1 is standard in this kind of problem. See (A.1) of Cheng and Tong (1993). Note that we have not assumed 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. 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 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 equivalent to Assumption C of Zhang (1993).

APPENDIX B

In this appendix, we give only a sketch proof of Theorem 2.1, because the detailed proof is very tedious. The following lemmas are required to prove Theorem 2.1.

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

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

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 $D \notin \mathcal{D}_1$,

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

The following lemmas are needed to complete the proof of Lemma B.1. The proof of Lemma 2.2 follows from that of Lemma B.2 below.

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

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

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

where $\phi_1(X_{tD}) = E[Y_t|X_{tD}]$ and $\phi_2(X_{tD}) = E[U_t|X_{tD}]$.

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

$$\sqrt{T}\Delta_{1D}\left(\hat{\beta}(D,h) - E[\hat{\beta}(D,h)]\right) \to N(0, E[e_{tD}^2 \xi_{tD} \xi_{tD}^{\tau}])$$

and

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

where $\xi_{tD} = U_t - E[U_t|X_{tD}]$ as defined before.

Obviously, Lemma 2.2 follows from (B.2). In order to prove (B.2), it suffices to show that

$$\frac{1}{T}\tilde{\Sigma}(D,h) = \frac{1}{T}\sum_{t=1}^{T} \tilde{U}_t(D,h)\tilde{U}_t(D,h)^{\tau} \to_p \Delta_{1D}, \tag{B.3}$$

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

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

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

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

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

where $r_{tD} = e_{tD}$ or ξ_{tD}^{τ} and Δ_{1D} is as defined in Assumption 2.1(i).

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)]. The proof of (B.3) follows from Assumption A.4, (B.6), (B.7), the Cauchy-Schwarz inequality and

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

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

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

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

$$V_2(D,h)=\frac{1}{T}\sum_{t=1}^{T}\left(\phi_2(X_{tD})-\hat{\phi}_{2t}(X_{tD},h)\right)\left(\phi_2(X_{tD})-\hat{\phi}_{2t}(X_{tD},h)\right)^{\tau}w(X_t)$$

$$=d_3(D,h)\frac{1}{Th^{|D|}}+d_4(D,h)h^4+o_p\{V_2(D,h)\},\qquad (B.10)$$

where $\{d_i(D,h):1\le i\le 2\}$ *are positive constants and* $\{d_j(D,h):3\le j\le 4\}$ *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* $h\in H_{TD}$

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

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

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 $h\in H_{TD}$

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

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

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

and

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

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*

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

$$=d_5(D,h)\frac{1}{Th^{|D|}}+d_6(D,h)h^4+o_p(V(D,h))\qquad (B.16)$$

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

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

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,

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

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

$$\equiv V(D,h)_1+V(D,h)_2+V(D,h)_3,\qquad (B.18)$$

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 $h \in H_{TD}$

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

and

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

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 $h \in H_{TD}$

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

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

Proof of Lemmas 2.1 and B.1. It follows from the definition of $\mathrm{CV1}(D,h)$ that

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

$$\equiv \frac{1}{T}\sum_{t=1}^{T} e_{tD_0}^2 w(X_t) + V(D,h) + R(D,h), \tag{B.22}$$

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 $h \in H_{TD}$

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

Therefore, Lemma B.4 and equation (B.24) imply Lemmas 2.1 and 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

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

and

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

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

Using the fact that Assumption 2.1 implies $|D| > |D_0|$, by (B.25) we have as $T \to \infty$

$$P\left(\mathrm{CV1}(D) > \mathrm{CV1}(D_0)\right) = P\left(T^{\frac{4}{4+|D_0|}}\left(\mathrm{CV1}(D) - \mathrm{CV1}(D_0)\right) > 0\right)$$

$$= P\left(C_{2D}T^{\frac{4(|D|-|D_0|)}{(4+|D|)(4+|D_0|)}} - C_{2D_0} + o_p\left(T^{\frac{4(|D|-|D_0|)}{(4+|D|)(4+|D_0|)}}\right) > 0\right) \to 1. \tag{B.26}$$

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

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

in probability as $T \to \infty$.

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

$$\lim_{T\to\infty} P(\hat{D}_0 = D_0) = 1. \tag{B.28}$$

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 $T \to \infty$

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

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. The proof is similar to the proof of Theorem 3.2 of Vieu (1994) and Theorem 2.1 above.

Proof of Corollary 2.2. It is a special case of Theorem 2.1.

Proof of Corollary 2.3. The proof is similar to the proofs of Lemmas B.1 and B.4.

Proof of Theorem 2.2. As the proof of Theorem 2.2 is based on those of Theorems 1 and 2 of Shao (1993), we give only an outline. 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 and 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) can be replaced by condition (2.16). In other words, we need to prove (2.16) and the following conditions:

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

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

$$\lim_{T\to\infty} \max_{t\le T} p_{tA} = 0 \text{ for any } A \in \mathcal{A}, \tag{B.31}$$

where p_{tA} is the tth 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,

$$E\left[\epsilon^\tau P_A \epsilon\right] = E\left\{E\left[\epsilon^\tau P_A \epsilon | V\right]\right\} = d_A \sigma_\epsilon^2$$

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 $T \to \infty$

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

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. Now the outline proof is completed.

Proof of Theorem 2.3. The proof follows from those of Theorems 2.1 and 2.2.

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

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

$$\frac{1}{T}\eta_A^\tau \eta_A \to_p E\left[\eta_A^\tau \eta_A\right], \ \frac{1}{T}\eta^\tau \eta \to_p E\left[\eta^\tau \eta\right], \tag{C.1}$$

$$T\left(\eta_A^\tau \eta_A\right)^+ \to_p \left(E\left[\eta_A^\tau \eta_A\right]\right)^+, \ T\left(\eta^\tau \eta\right)^+ \to_p \left(E\left[\eta^\tau \eta\right]\right)^+, \tag{C.2}$$

$$\frac{1}{T}V_A^\tau V_A \to_p E\left[\eta_A^\tau \eta_A\right], \ \frac{1}{T}V^\tau V \to_p E\left[\eta^\tau \eta\right], \tag{C.3}$$

$$T\left(V_A^\tau V_A\right)^+ \to_p \left(E\left[\eta_A^\tau \eta_A\right]\right)^+, \ T\left(V^\tau V\right)^+ \to_p \left(E\left[\eta^\tau \eta\right]\right)^+, \tag{C.4}$$

$$\frac{1}{T}(\delta\beta)^\tau (\delta\beta) \to_p 0. \tag{C.5}$$

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 $T \to \infty$

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

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

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

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

$$P\left(\left\| \max_{S\in\mathcal{R}} \sum_{t\in S} (\zeta_t'' - E[\zeta_t'']) \right\| > \xi T_v \right) \leq \max_{S\in\mathcal{R}} P\left(\left\| \sum_{t\in S} (\zeta_t'' - E[\zeta_t'']) \right\| > \xi T_v \right)$$

$$\leq C T_v^{-1} \max_{S\in\mathcal{R}} \sum_{t\in S} E\left\|\zeta_t''\right\| \leq C T_v^{-1} \max_{S\in\mathcal{R}} \sum_{t\in S} E\left(\frac{\|\zeta_t\|^6}{\|\zeta_t\|^5} I[\|\zeta_t\| > T_v^{1/2}] \right)$$

$$\leq C T_v^{-1} \max_{S\in\mathcal{R}} \sum_{t\in S} E[\|\zeta_t\|^6] T_v^{-5/2} \leq C b T_v^{-5/2}, \tag{C.7}$$

using Assumption A.4.

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

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

Equation (C.8) implies that as $T_v \to \infty$

$$\frac{1}{T_v}\max_{S\in\mathcal{R}}\left\|\sum_{t\in S}\zeta_t\right\| \to 0$$

holds with probability one. Thus, equation (B.29) holds in probability. Note that we have actually shown that (B.29) holds with probability one. Thus we may conclude that if $x_1, \ldots, x_n$ are random, (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

$$\begin{aligned} E\left[\sum_{t\in\mathcal{S}} p_{tA}\epsilon_t^2\right]^2 &= \sum_{t\in\mathcal{S}} E\left[p_{tA}^2\epsilon_t^4\right] + \sum_{s\neq t\in\mathcal{S}} E\left[p_{sA}p_{tA}\epsilon_s^2\epsilon_t^2\right] \\ &= \sum_{t\in\mathcal{S}} E\left\{E\left[p_{tA}^2\epsilon_t^4|V\right]\right\} + \sum_{s\neq t\in\mathcal{S}} E\left\{E\left[p_{sA}p_{tA}\epsilon_s^2\epsilon_t^2|V\right]\right\} \\ &\leq C_1(\epsilon)E\left[\sum_{t\in\mathcal{S}} p_{tA}^2\right] + C_2E\left[\sum_{s\neq t\in\mathcal{S}} p_{sA}p_{tA}\right] \leq C, \end{aligned} \qquad (C.9)$$

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

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

imply the first part of (B.32).

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

$$\frac{\sum_{s\neq t\in\mathcal{S}} p_{sA}p_{tA}\epsilon_s\epsilon_t}{\sqrt{\mathrm{Var}\left(\sum_{s\neq t\in\mathcal{S}} p_{sA}p_{tA}\epsilon_s\epsilon_t\right)}} \to N(0,1). \qquad (C.10)$$

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. We would like to thank the Editors, the Associate Editors and the referees for their comments and suggestions. 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

Bhattacharya, P. K. and Zhao, P. L. (1997). Semiparametric inference in a partial linear model. *Ann. Statist.* **25** 244–262.

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.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489.

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

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

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.

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

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

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.

Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low dimensional components in additive models. *Ann. Statist.* **26** 943–971.

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

Gao, J. and Anh, V. (1999). Semiparametric regression with long-range dependent error processes. *J. Statist. Plann. Inference* **80** 37–57.

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 Liang, H. (1997). Statistical inference in single-index and partially linear regression models. *Ann. Inst. Statist. Math.* **49** 493–517.

Gao, J. and Shi, P. (1997). M-type smoothing splines in nonparametric and semiparametric regression models. *Statistica Sinica* **7** 1155–1169.

Gao, J., Tong, H. and Wolff, R. (2001a). Adaptive series estimation in additive stochastic regression models. *Statistica Sinica* **12** 317–336.

Gao, J., Tong, H. and Wolff, R. (2001b). Model specification tests in nonparametric stochastic regression models. *J. Multivariate. Anal.* **82** 251–286.

Gao, J., Wolff, R. and Anh, V. (2001). Semiparametric approximation methods in multivariate model selection. *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.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall, London.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Boston.

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

HASTIE, T. AND TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

HECKMAN, N. (1986). Spline smoothing in a partly linear model. *J. Roy. Statist. Soc. Ser. B* **48** 244–248.

LI, K. C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15** 958–975.

LIANG, H., HÄRDLE, W. AND CARROLL, R. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *Ann. Statist.* **27** 1519–1535.

MAMMEN, E. AND VAN DE GEER, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* **25** 1014–1035.

MASRY, E. AND TJØSTHEIM, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory* **13** 214–252.

RICE, J. (1986). Convergence rates for partially spline models. *Statist. Probab. Lett.* **4** 203–208.

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.

SHI, P. AND TSAI, C. L. (1999). Semiparametric regression model selections. *J. Statist. Plann. & Inference* **77** 119–139.

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

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 685–705.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 592–606.

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. (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.

VIEU, P. (1995). Order choice in nonlinear autoregressive models. *Statistics* **26** 307–328.

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

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.

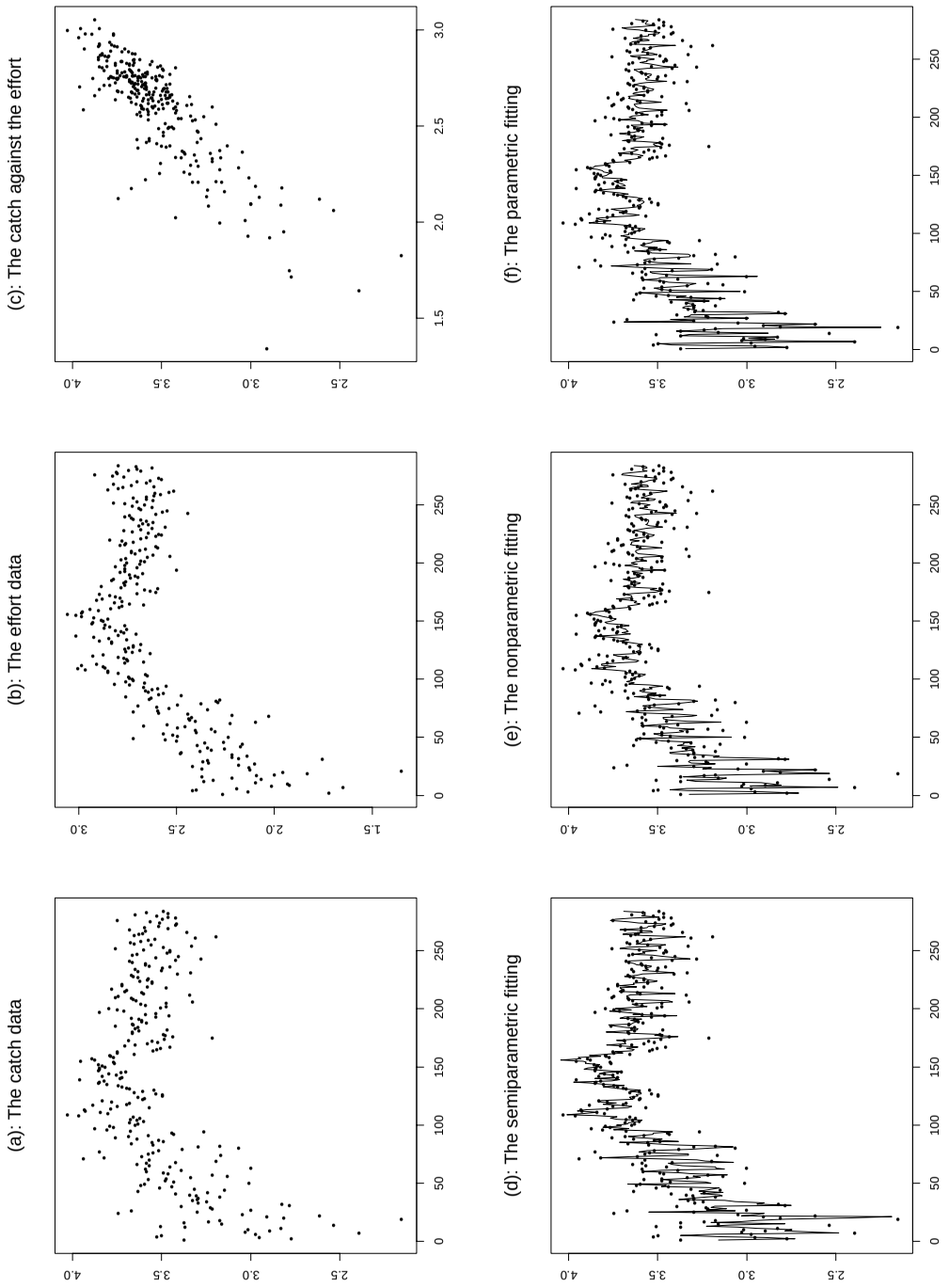

Figure 1: For the whole data set, parts (a)–(c) present time plots of the common-log-transformed catch data, the common-log-transformed effort data, and the transformed catch data against the transformed effort data, respectively; parts (d), (e) and (f) give plots of the fitted values (lines) and the catch data (dots) for the partially linear model (3.3), the purely nonparametric model (3.4) and the completely linear model (3.5), respectively.