Statistica Sinica **11**(2001), 409-428

ADAPTIVE ORTHOGONAL SERIES ESTIMATION IN ADDITIVE STOCHASTIC REGRESSION MODELS

Jiti Gao, Howell Tong and Rodney Wolff

The University of Western Australia, The University of Hong Kong and The London School of Economics and Queensland University of Technology

Abstract: In this paper, we consider additive stochastic nonparametric regression models. By approximating the nonparametric components by a class of orthogonal series and using a generalized cross-validation criterion, an adaptive and simultaneous estimation procedure for the nonparametric components is constructed. We illustrate the adaptive and simultaneous estimation procedure by a number of simulated and real examples.

Key words and phrases: Adaptive estimation, additive model, dependent process, mixing condition, nonlinear time series, nonparametric regression, orthogonal series, strict stationarity, truncation parameter.

1. Introduction

Recent developments in additive nonparametric regression and autoregressive models have provided a practical and efficient way to model multivariate data sets. Hastie and Tibshirani (1990) proposed a number of iterative estimation procedures to estimate additive and generalized additive regression functions, which can avoid the difficulty of the "curse of dimensionality". See Hastie and Tibshirani (1990) and others for more details. Chen and Tsay (1993) considered a class of additive autoregressive models and proposed two kernel-based iterative estimation procedures for the nonparametric components.

Let us now consider a general stochastic nonparametric regression model $Y = m(X) + e$, where $X = (X^1, \ldots, X^p)^\tau$ is a vector of p stochastic regressors $\{X^i : 1 \le i \le p\}$. Given the data $\{(Y_t, X_t) : t \ge 1\}$, a crucial problem is how to analyse the structure of (Y_t, X_t) and how to determine the relationship between the present observation Y_t and the regressor vector $X_t = (X_{t1}, \ldots, X_{tp})^\tau$. Yao and Tong (1994) proposed using a consistent cross-validation (CV) criterion to select an optimum subset of regressors. See also Vieu (1994). More recently, Gao, Wolff and Anh (1999) considered using a CV criterion to select an optimum subset of linear regressors from X_t and illustrated the CV criterion on several real data sets.

In this paper, we consider the following p-order additive stochastic nonparametric regression model:

$$Y_t = \sum_{i=1}^{p} g_i(X_{ti}) + e_t, \qquad t = 1, \ldots, T, \tag{1.1}$$

where T is the number of observations, $p \geq 1$ is an integer, $\{g_i(\cdot) : 1 \leq i \leq p\}$ are unknown functions over R^1, $\{e_t\}$ is a sequence of martingale differences. For $Y_t = y_{t+p}$ and $X_{ti} = y_{t+p-i}$, (1.1) is a p-order additive nonparametric autoregressive model studied extensively by Chen and Tsay (1993). See also Wong and Kohn (1996), who considered using a Bayesian approach to estimating and forcasting a special case of (1.1) with $p = 2$, $Y_t = y_{t+2}$ and $X_{ti} = y_{t+2-i}$. Recently, Gao and Liang (1995) studied another special case of (1.1) with $p = 2$, $Y_t = y_{t+2}$, $X_{ti} = y_{t+2-i}$, $g_1(y_{t+1}) = \beta y_{t+1}$ and $g_2(\cdot)$ approximated by a piecewise polynomial function. See also Gao and Yee (2000), who constructed a data-based adaptive estimator for β based on g_2 being estimated by a kernel function. See Härdle, Liang and Gao (2000) for recent developments in partially linear models.

Here we consider (1.1). For the sake of identifiability, we only need to consider the transformed model $Y_t = \alpha + \tilde{g}_1(X_{t1}) + \cdots + \tilde{g}_p(X_{tp}) + e_t, t = 1, \ldots, T$, where $\alpha = \sum_{i=1}^{p} E[g_i(X_{ti})]$ is an unknown parameter and $\tilde{g}_i(X_{ti}) = g_i(X_{ti}) - E[g_i(X_{ti})]$ satisfies $E[\tilde{g}_i(X_{ti})] = 0$. It is obvious from the proofs in the Appendix that the conclusion of Theorems 2.1-2.3 remains unchanged when Y_t is replaced by $\tilde{Y}_t = Y_t - \hat{\alpha}$, where $\hat{\alpha} = \frac{1}{T}\sum_{t=1}^{T} Y_t$ is defined as the estimator of α.

In this paper, we propose using the orthogonal series approach to constructing adaptive estimators for (1.1). As suggested in the recent econometric literature (see Eastwood and Gallant (1991)), the orthogonal series method can be an alternative to the kernel estimation method. Recently, Linton and Nielson (1995), Linton (1997), and Fan, Härdle and Mammen (1998) proposed the so-called "marginal integration method" coupled with the Nadaraya-Watson approach for the independent and identically distributed (i.i.d.) case, and Tjøstheim and Auestad (1994a, 1994b) proposed a "projection method" for the time series situation. See also Masry and Tjøstheim (1995, 1997). However, these methods (even the latest paper by Fan, Härdle and Mammen (1998)) have not extensively considered selecting the bandwidth parameters involved in their methods. By contrast, the proposed orthogonal series method provides not only a simultaneous estimation procedure for all the nonparametric components, but also a practical procedure for selecting the truncation parameters involved in the orthogonal series.

By approximating each $g_i(\cdot)$ by an orthogonal series $\sum_{j=1}^{h_i} f_{ij}(\cdot)\theta_{ij}$, we have the following approximate model

$$Y_t = \sum_{i=1}^{p}\sum_{j=1}^{h_i} f_{ij}(X_{ti})\theta_{ij} + e_t, \tag{1.2}$$

which is a natural extension of an additive linear model. Therefore, some existing estimation procedures can be used to obtain explicit estimators for $\{g_i(\cdot) : 1 \leq i \leq p\}$. In the meantime, we propose a data-based criterion to determine the truncation parameters $\{h_i : 1 \leq i \leq p\}$. We illustrate the estimation procedure by simulated and real examples later.

The organisation of this paper is as follows: Section 2 proposes an adaptive and simultaneous estimation procedure for $\{g_i(\cdot) : 1 \leq i \leq p\}$. Illustrations of the proposed estimation procedure are given in Section 3. Mathematical details are given in the Appendix.

2. Adaptive and Simultaneous Estimation Procedure

The approach taken in this section is to approximate each $g_i(\cdot)$ by the orthogonal series $\sum_{j=1}^{h_i} f_{ij}(\cdot)\theta_{ij}$, where $\{f_{ij}(\cdot) : 1 \leq j \leq h_i\}$ are prespecified families of continuous functions from R^1 to R^1, $\theta_i = (\theta_{i1}, \ldots, \theta_{ih_i})^\tau$ is a vector of unknown parameters and $h_i = h_i(T)$ is the truncation parameter.

Based on (1.2), we define the least squares (LS) estimator $\hat{\theta}(h) = (\hat{\theta}_1(h)^\tau, \ldots, \hat{\theta}_p(h)^\tau)^\tau$ of $\theta = (\theta_1^\tau, \ldots, \theta_p^\tau)^\tau$ as the solution of

$$\sum_{t=1}^{T} \Big(Y_t - \sum_{i=1}^{p} F_i(X_{ti})^\tau \theta_i\Big)^2 = \min!, \tag{2.1}$$

where $h = (h_1, \ldots, h_p)^\tau$ and $F_i(X_{ti}) = F_{ih_i}(X_{ti}) = (f_{i1}(X_{ti}), \ldots, f_{ih_i}(X_{ti}))^\tau$.

It is obvious that

$$\hat{\theta}(h) = (F^\tau F)^+ F^\tau Y, \tag{2.2}$$

provided the right-hand side is well defined, where $Y = (Y_1, \ldots, Y_T)^\tau$, $F = (F_1, \ldots, F_p)$, $F_i = F_{ih_i} = (F_i(X_{1i}), \ldots, F_i(X_{Ti}))^\tau$, and $(\cdot)^+$ denotes the Moore-Penrose inverse.

In the case of $p = 2$, by Theorem 3.7 of Seber (1977), we obtain the LS estimators $\hat{\theta}_1(h) = (\hat{F}_1^\tau \hat{F}_1^\tau)^+ \hat{F}_1^\tau Y$ and $\hat{\theta}_2(h) = (F_2^\tau F_2)^+ F_2^\tau (I - F_1(\hat{F}_1^\tau \hat{F}_1^\tau)^+ \hat{F}_1^\tau)Y$, where $\hat{F}_1 = (I - P_2)F_1$ and $P_2 = P_2(h) = F_2(F_2^\tau F_2)^+ F_2^\tau$. If F_2 is of full column rank h_2, then $(F_2^\tau F_2)^{-1}$ exists.

For the given truncation parameters $\{h_i : 1 \leq i \leq p\}$, we propose the prediction equation

$$\hat{g}(X_t, h) = \sum_{i=1}^{p} F_i(X_{ti})^\tau \hat{\theta}_i(h), \tag{2.3}$$

where $X_t = (X_{t1}, \ldots, X_{tp})^\tau$.

It follows from (2.3) that the prediction equation depends on not only the series functions $\{f_{ij} : 1 \leq j \leq h_i, 1 \leq i \leq p\}$ but also h, the vector of truncation parameters. Our experience suggests that the choice of the series functions is

much less critical than that of the vector of truncation parameters. The series functions used in this paper need to satisfy Assumptions 2.2 and 2.3 below, which hold when each f_{ij} belongs to either the class of trigonometric series used by Eastwood and Gallant (1991), or the general class of orthogonal series presented by Kashin and Saakyan (1989). See Examples 3.1-3.2 and Remark 2.1 below for more details. Therefore, a crucial problem is how to select h practically. Li (1987) has discussed the asymptotic optimality of a generalized cross-validation (GCV) criterion as well as other model selection criteria. See also Li (1985, 1986) and Shao (1997). Wahba (1990) provided a survey of nonparametric smoothing spline literature up to 1990. Chen and Chen (1991) considered using a generalized cross-validation criterion for selecting an optimum subset for the i.i.d. case. Gao (1998) applied a generalized cross-validation criterion for smoothing truncation parameters for the time series case. More recently, Shi and Tsai (1999) considered semiparametric regression model selections and proposed a so-called "AICC small-sample criterion" coupled with the B-spline approach for the i.i.d. case. They showed that their criterion has advantages over some existing criteria. In this paper, we apply a generalized cross-validation method to estimate h and then determine model (1.1).

Let

$$\hat{D}(h) = \frac{1}{T}\sum_{t=1}^{T}\Big\{\sum_{i=1}^{p} F_i(X_{ti})^{\tau}\hat{\theta}_i(h) - \sum_{i=1}^{p} g_i(X_{ti})\Big\}^2. \tag{2.4}$$

Before establishing the main results of this paper, we first need to introduce the following assumptions and definitions.

Assumption 2.1. (i) Assume the process (X_t, Y_t) is strictly stationary and α-mixing with mixing coefficient $\alpha(T) = C_\alpha \eta^T$, where $0 < C_\alpha < \infty$ and $0 < \eta < 1$ are constants.

(ii) Assume $e_t = Y_t - E[Y_t|X_t]$ satisfies $E[e_t|\Omega_{t-1}] = 0$, $E[e_t^2|\Omega_{t-1}] = E[e_t^2]$ $a.s.$, and $E[e_t^4|\Omega_{t-1}] < \infty$ for all $t \geq 1$, where $\Omega_t = \sigma\{(X_{s+1}, Y_s) : 1 \leq s \leq t\}$ is a sequence of σ-fields generated by $\{(X_{s+1}, Y_s) : 1 \leq s \leq t\}$.

Let $g_i^{(m_i)}$ be the m_i-order derivative of the function g_i and M_{0i} be a constant. Let

$$G_{m_i}(S_i) = \Big\{g : |g_i^{(m_i)}(s) - g_i^{(m_i)}(s')| \leq M_{0i}|s - s'|, s, s' \in S_i \subset R^1\Big\}$$

where $m_i \geq 1$ is an integer, $0 < M_{0i} < \infty$, and each S_i is a compact subset of $R^1 = (-\infty, \infty)$.

Assumption 2.2. (i) For $g_i \in G_{m_i}(S_i)$ and $\{f_{ik}(\cdot) : k = 1, \ldots\}$ given above, there exists a vector of unknown parameters $\theta_i = (\theta_{i1}, \ldots, \theta_{ih_i})^{\tau}$ such that for a

sequence of constants $\{B_i : 1 \le i \le p\}$ ($0 < B_i < \infty$ independent of T) and large enough T,

$$\sup_{x_i \in S_i} |F_i(x_i)^\tau \theta_i - g_i(x_i)| \le B_i h_i^{-(m_i+1)} \tag{2.5}$$

uniformly over $h_i \in H_{iT}$ and $1 \le i \le p$. Here $H_{iT} = \{p_{iT}, p_{iT}+1, \ldots, q_{iT}\}$, in which $p_{iT} = [a_i T^{d_i}]$, $q_{iT} = [b_i T^{c_i}]$, $0 < a_i < b_i < \infty$, $0 < d_i < c_i < \frac{1}{2(m_i+1)}$ are constants, and $[x] \le x$ denotes the largest integer part of x.

(ii) There exists a sequence of constants $\{C_i : 1 \le i \le p\}$ ($0 < C_i < \infty$ independent of T) such that, for large enough T,

$$h_i^{2(m_i+1)} E\left\{F_i(X_{ti})^\tau \theta_i - g_i(X_{ti})\right\}^2 \approx C_i \tag{2.6}$$

uniformly over $h_i \in H_{iT}$ and $1 \le i \le p$. ("$\approx$" indicates that the ratio of the left-hand side and the right-hand side tends to one as $T \to \infty$).

Assumption 2.3. (i) F_i is of full column rank $h_i \in H_{iT}$ for T large enough, $\{f_{ij} : 1 \le j \le h_i, 1 \le i \le p\}$ are continuous functions with $\sup_x \sup_{i,k \ge 1} |f_{ik}(x)| \le c_0 < \infty$.

(ii) For all $1 \le i, j \le p$ and $s \ne t$, $E[f_{ij}(X_{si}) f_{ij}(X_{ti})] = 0$, and for all $t \ge 1$

$$E[f_{ik}(X_{ti}) f_{jl}(X_{tj})] = \begin{cases} c_{ik}^2, & \text{if } i = j \text{ and } k = l \\ 0, & \text{if } (i,j,k,l) \in IJKL, \end{cases}$$

where $IJKL = \{(i,j,k,l) : 1 \le i, j \le p, 1 \le k \le h_i, 1 \le l \le h_j\} - \{(i,j,k,l) : 1 \le i = j \le p, 1 \le k = l \le h_i\}$.

Assumption 2.4. There are positive constants $\{C_K : K \ge 1\}$ such that for $K = 1, 2, \ldots$, $\sup_x E(|Y_t|^K | X_t = x) \le C_K < \infty$.

Remark 2.1. (i) Assumption 2.1 is quite common in such problems. See Doukhan (1995) for the advantages of the geometric mixing. However it would be possible, but with more tedious proofs, to obtain Theorems 2.1-2.3 below under less restrictive assumptions that include some algebraically decaying rates. If e_t is i.i.d. and e_t is independent of X_t, then Assumption 2.1(i) only requires that the process X_t is strictly stationary and α-mixing, and Assumption 2.1(ii) yields $E[e_t] = 0$ and $E[e_t^4] < \infty$. This is a natural condition in nonparametric autoregression. See for example, Assumption 2.1 of Gao (1998). For the heteroscedastic case, one needs to modify both Assumptions 2.1(i) and 2.1(ii). See for example, Conditions (A2) and (A4) of Hjellvik, Yao and Tjøstheim (1998).

(ii) Assumption 2.2(i) is imposed to exclude the case that each g_i is already a linear combination of $\{f_{ij} : 1 \le j \le h_i\}$. If (1.1) is an additive polynomial regression, the choice of h_i is a model selection problem already discussed by Li (1987).

(iii) The purpose of introducing Assumptions 2.2(i) and 2.2(ii) is to replace the unknown functions by finite series sums together with vectors of unknown parameters. Equation (2.5) is a standard smoothness condition in approximation theory. See Corollary 6.21 of Schumaker (1981) for the B-spline approximation, Chapter IV of Kashin and Saakyan (1989) for the trigonometric approximation, Theorem 0 of Gallant and Souza (1991) for the flexible Fourier form, and Chapter 7 of DeVore and Lorentz (1993) for the general orthogonal series approximation. If Assumption 2.2(ii) holds, then (2.6) is equivalent to

$$h_i^{2(m_i+1)} \int [F_i(u_i)^\tau \theta_i - g_i(u_i)]^2 p_i(u_i) du_i \approx C_i, \tag{2.7}$$

where $p_i(u_i)$ denotes the density function of X_{ti}. Equation (2.7) is a standard smoothness condition in approximation theory. See Theorems 3.1 and 4.1 of Agarwal and Studden (1980) for the B-spline approximation, and §3.2 of Eastwood and Gallant (1991) for the trigonometric approximation.

(iv) A technical advantage of Assumption 2.2(i) over previous assumptions of this type (see Assumption 2.2 of Gao (1998)) is that the range of h_i under consideration has been extended from $\{[a_i T^{\frac{1}{2m_i+3}-\epsilon_i}], \ldots, [b_i T^{\frac{1}{2m_i+3}+\epsilon_i}]\}$ with $0 < \epsilon_i < (2m_i - 1)/[4(2m_i + 3)]$ to $\{p_{iT}, \ldots, q_{iT}\}$. This provides more security and theoretical underpinning for consideration of h_i both large and small. The choice of d_i and c_i is due to the fact that each theoretical optimum value of h_i is proportional to $[T^{\frac{1}{2(m_i+1)+1}}]$. See the proof of Theorem 2.1 below. This choice is more reasonable in practice.

(v) A technical restriction of Assumption 2.2 is that each g_i is defined on the compact subset S_i. As discussed in the references cited in this paper, compactness is a very natural condition in approximation theory. But it can be weakened by introducing a weight function into (2.4). Details are similiar to those used in nonparametric kernel regression. See Härdle and Vieu (1992).

(vi) Assumption 2.3 is a kind of orthogonality condition, which holds when the process X_t is strictly stationary and $\{f_{ij} : 1 \le j \le h_i, 1 \le i \le p\}$ is either in the family of trigonometric series or of Gallant (1981)'s flexible Fourier form. For example, the orthogonality condition holds when X_{t1} is strictly stationary and distributed uniformly over $[-1, 1]$ and $f_{1k}(X_{t1}) = \sin(k\pi X_{t1})$ or $\cos(k\pi X_{t1})$. Moreover, orthogonality is a natural condition in nonparametric series regression.

(vii) Assumption 2.4 is required to deal with this kind of problem. Many authors have used similar conditions. See for example, (C.7) of Härdle and Vieu (1992).

Definition 2.1. A data-driven estimator $\hat{h}$ is *asymptotically optimal* if $\hat{D}(\hat{h})/\inf_{h \in H_T} \hat{D}(h) \to_p 1$, where $\hat{h} = (\hat{h}_1, \ldots, \hat{h}_p)^\tau$, $h \in H_T = \{h = (h_1, \ldots, h_p)^\tau : h_i \in H_{iT}\}$ and H_{iT} is defined in Assumption 2.2.

Definition 2.2. Select h, denoted by $\hat{h}_G = (\hat{h}_{1G}, \ldots, \hat{h}_{pG})^\tau$, so that

$$GCV(\hat{h}_G) = \inf_{h \in H_T} GCV(h) = \inf_{h \in H_T} \frac{\hat{\sigma}^2(h)}{[1 - \frac{1}{T}\sum_{i=1}^p h_i]^2}, \tag{2.8}$$

where $\hat{\sigma}^2(h) = \frac{1}{T}\sum_{t=1}^T \left\{Y_t - \sum_{i=1}^p F_i(X_{ti})^\tau \hat{\theta}_i(h)\right\}^2$.

Theorem 2.1. (i) *Let Assumptions* 2.1-2.2(i), 2.3 *and* 2.4 *hold. Then*

$$\hat{D}(h) = \frac{\sigma^2}{T}\sum_{i=1}^p h_i + \frac{1}{T}E[\Delta^\tau \Delta] + o_p(\hat{D}(h)), \tag{2.9}$$

where $\Delta = \sum_{i=1}^p [F_i\theta_i - G_i]$, $F_i = (F_i(X_{1i}), \ldots, F_i(X_{Ti}))^\tau$ *and* $G_i = (g_i(X_{1i}), \ldots, g_i(X_{Ti}))^\tau$.

(ii) *In addition if Assumption* 2.2(*ii*) *holds, then*

$$\hat{D}(h) = \frac{\sigma^2}{T}\sum_{i=1}^p h_i + \sum_{i=1}^p C_i h_i^{-2(m_i+1)} + o_p(\hat{D}(h)) \tag{2.10}$$

uniformly over $h \in H_T$, *where* $\sigma^2 = E[e_t^2] < \infty$ *and* m_i *is the smoothness order of* g_i.

Theorem 2.2. (i) *Under the conditions of Theorem* 2.1(*i*), $\hat{h}_G$ *is asymptotically optimal.*

(ii) *Under the conditions of Theorem* 2.1(ii)

$$\frac{\hat{D}(\hat{h}_G)}{\hat{D}(\hat{h}_D)} - 1 = o_p(T^{-\tau}), \tag{2.11}$$

$$\sum_{i=1}^p \left|\frac{\hat{h}_{iG}}{\hat{h}_{iD}} - 1\right| = o_p(T^{-\tau}), \tag{2.12}$$

where $\hat{h}_{iD}$ *is the i-th component of* $\hat{h}_D = (\hat{h}_{1D}, \ldots, \hat{h}_{pD})^\tau$ *that minimises* $\hat{D}(h)$ *over* H_T, $0 < \tau = \min(\tau_1 - \epsilon_1, \tau_2 - \epsilon_2)$, *in which* $\tau_1 = \frac{1}{2}d$, $\tau_2 = \frac{1}{2} - 2c$, *both* ϵ_1 *and* ϵ_2 *satisfying* $0 < \epsilon_1 < \tau_1$ *and* $0 < \epsilon_2 < \tau_2$ *are arbitrarily small,* $d = \min_{1 \le i \le p} d_i$ *and* $c = \max_{1 \le i \le p} c_i$.

Proofs of Theorems 2.1 and 2.2 are relegated to the Appendix.

We now define the adaptive and simultaneous estimation procedure as follows:

(i) solve the LS estimator $\hat{\theta}(h)$ by (2.2);
(ii) define the prediction equation by (2.4);
(iii) solve the GCV-based $\hat{h}_G$ from (2.8);

(iv) define the following adaptive and simultaneous prediction equation

$$\hat{g}(X_t, \hat{h}_G) = \sum_{i=1}^{p} F_{i\hat{h}_{iG}}(X_{ti})^\tau \hat{\theta}_i(\hat{h}_G).$$

If σ^2 is unknown, it estimated by $\hat{\sigma}^2(\hat{h}_G) = (1/T)\sum_{t=1}^{T}\{Y_t - \hat{g}(X_t, \hat{h}_G)\}^2$.

Theorem 2.3. *Under the conditions of Theorem* 2.1(*i*), *as* $T \to \infty$, $\sqrt{T}(\hat{\sigma}^2(\hat{h}_G) - \sigma^2) \to N(0, \mathrm{var}(e_1^2))$.

The proof of Theorem 2.3 is postponed to the Appendix.

Remark 2.2. Equations (2.9) and (2.10) provide asymptotic representations for the average squared error $\hat{D}(h)$. See Härdle, Hall and Marron (1988) for an equivalent result in nonparametric kernel regression, and Hall and Patil (1995) for a corresponding form in non-linear wavelet estimation. In addition, Theorem 2.2(i) shows that the GCV based $\hat{h}_G$ is asymptotically optimal. This conclusion is equivalent to Corollary 3.1 of Li (1987) in the model selection problem. However, the fundamental difference between our paper and Li (1987) is that this paper uses the GCV method to determine how many terms are required to ensure that each nonparametric function can be approximated optimally, while Li (1987) suggested using the GCV selection criterion to determine how many variables should be employed in a linear model. Due to the different objectives, our conditions and conclusions are different from those of Li (1987), although there are some similarities.

Remark 2.3. Theorem 2.2(ii) not only establishes the asymptotic optimality but also provides the rate of convergence. This rate of convergence is equivalent to that of bandwidth estimates in nonparametric kernel regression. See Härdle, Hall and Marron (1992). More recently, Hurvich and Tsai (1995) have established a similar result for a linear model selection. Moreover, it follows from Theorem 2.2(ii) that the rate of convergence depends heavily on d_i and c_i. Let $d_i = \frac{1}{2m_i+3}$ and $c_i = \frac{1}{2m_i+3} + \eta_i$ for arbitrarily small $\eta_i > 0$. Then the rate of convergence will be of order

$$\min\Big(\min_{1\le i\le p}\Big(\frac{1}{2(2m_i+3)}\Big), \max_{1\le i\le p}\Big(\frac{2m_i-1}{2(2m_i+3)}\Big)\Big) - \epsilon$$

for some arbitrarily small $\epsilon > 0$. Obviously, if each g_i is continuously differentiable, the rate of convergence will be close to $\frac{1}{10} - \epsilon$.

Remark 2.4. In this paper, we assume the data $\{(Y_t, X_t) : t \ge 1\}$ satisfy (1.1) and then propose the orthogonal series method to model the data. In practice, before applying the estimation procedure to model the data, a crucial problem

is how to test additivity. Related results for additive nonparametric regression have been given by some authors. See for example, Gao, Tong and Wolff (2000). In this paper, we do not discuss the problem further.

Remark 2.5. This paper only considers the case where $\{e_t\}$ is a sequence of martingale differences. In practice, there are data sets where these assumptions are far from being satisfied, and there is an increasing awareness of deviations from these assumptions in general. In particular, econometricians have assembled increasing evidence for non-constant conditional variance describing a fluctuating risk structure for financial time series (see Bera and Higgins (1993)). For example, Tjøstheim and Auestad (1994a, 1994b) mention $\sigma^2(x_t) = E[(Y_t - E(Y_t|X_t))^2|X_t = x_t] = \sum_{j=1}^{p} \sigma_j^2(x_{tj})$, where $x_t = (x_{t1}, \ldots, x_{tp})^\tau$ and $\{\sigma_i^2(\cdot) : 1 \le i \le p\}$ are unknown functions. In this case, we need to approximate each $\sigma_i^2(\cdot)$ by an orthogonal series and then construct the weighted LS estimators $\hat{\theta}$ and $\hat{\gamma}$ as the solution of

$$\sum_{t=1}^{T} \Big(\frac{Y_t - F(X_t)^\tau \theta}{\Sigma(X_t)^\tau \gamma}\Big)^2 = \min!,$$

where $\Sigma(\cdot)^\tau\gamma$ is used to approximate $\sigma(\cdot)$, $\Sigma(\cdot)$ is a vector of known functions and γ is a vector of unknown parameters.

Analogous to Theorems 2.1-2.3, we can establish corresponding results.

Remark 2.6. As mentioned in Tjøstheim (1994), the following additive model is very useful in economic time series analysis: $Y_t = \sum_{i=1}^{p} g_i(\alpha_i^\tau X_t) + e_t$, where $X_t = (X_{t1}, \ldots, X_{td})^\tau$, $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{id})^\tau$ is a vector of unknown parameters, $\{g_i(\cdot) : 1 \le i \le q\}$ are unknown functions over R^1, and p and d are positive integers. Recently, Gao and Liang (1997) considered a special case of this model and constructed series estimators through approximating each $g_i(\cdot)$ by a finite series. Similar to the discussion of (1.1), we can construct some explicit estimators for $\{g_i(\cdot) : 1 \le i \le p\}$, but can only provide some iterative estimators for all $\{\alpha_i : 1 \le i \le d\}$. Existing iterative estimation procedures for $\{\alpha_i : 1 \le i \le d\}$ can be found in Seber and Wild (1989).

Remark 2.7. We choose the traditional LS method. However, it is well known that estimators based on the LS method are sensitive to outliers and that the error distribution may be heavy-tailed. Thus a more robust estimation procedure for all $\{g_i(\cdot) : 1 \le i \le p\}$ might be worth study to achieve desirable robustness properties. A recent paper by Gao and Shi (1997) on M-type smoothing splines for nonparametric and semiparametric regression may be useful to construct and study the following M-type estimator $\hat{\theta}_M(h) = (\hat{\theta}_{M1}(h), \ldots, \hat{\theta}_{Mp}(h))^\tau$:

$$\sum_{t=1}^{T} \rho\Big(Y_t - \sum_{i=1}^{p} F_i(X_{ti})^\tau \hat{\theta}_{Mi}(h)\Big) = \min!,$$

where $\rho(\cdot)$ is a convex function.

3. Applications and Examples

In this section, we illustrate the above estimation procedure by two examples.

Example 3.1. Consider the model

$$Y_t = 0.25Y_{t-1} + 0.25\frac{Y_{t-2}}{1+Y_{t-2}^2} + \frac{1}{8\pi}X_t^2 + e_t,\ t = 3, 4, \ldots, T, \tag{3.1}$$

where e_t is uniformly distributed over $(-0.5\pi, 0.5\pi)$, Y_1 and Y_2 are mutually independent and uniformly distributed over $[1/128, 2\pi - 1/128]$, (Y_1, Y_2) is independent of e_t for $t \geq 3$;

$$X_t = 0.25X_{t-1} - 0.25X_{t-2} + \epsilon_t, \tag{3.2}$$

in which ϵ_t is uniformly distributed over $(-0.5\pi, 0.5\pi)$, X_1 and X_2 are mutually independent and uniformly distributed over $[1/128, 2\pi - 1/128]$, and ϵ_t is independent of (X_1, X_2) and e_t for all $t \geq 3$.

First, it follows from Lemma 3.1 of Masry and Tjøstheim (1997) that both the stationarity and the mixing condition are met. See also Chapter 4 of Tong (1990), §2.4 of Tjøstheim (1994) and §2.4 of Doukhan (1995). Thus, Assumption 2.1(i) holds. Second, it follows from (3.1) and (3.2) that Assumption 2.1(ii) holds immediately. Third, let

$$g_1(x) = 0.25x,\ g_2(x) = 0.25\frac{x}{1+x^2} \text{ and } g_3(x) = \frac{1}{8\pi}x^2. \tag{3.3}$$

Since $\{g_i : 1 \leq i \leq 3\}$ are continuously differentiable on $(-\infty, \infty)$, there exist three corresponding periodic functions defined on $[0, 2\pi]$ that are continuously differentiable on $[0, 2\pi]$ and coincide with $\{g_i : 1 \leq i \leq 3\}$ (see Hong and White (1995, p.1141)). Similar to §3.2 of Eastwood and Gallant (1991), we can show that there exist the following three corresponding trigonometric polynomials

$$g_1^*(x) = \sum_{j=1}^{h_1} \sin(jx)\theta_{1j},\ g_2^*(x) = \sum_{j=1}^{h_2} \sin(jx)\theta_{2j} \text{ and } g_3^*(x) = \sum_{j=1}^{h_3} \cos(jx)\theta_{3j} \tag{3.4}$$

such that Assumptions 2.2(i) and 2.2(ii) are satisfied, and the same convergence rate can be obtained as in the periodic case. Obviously, it follows from (3.3) and (3.4) that Assumption 2.2(i) holds. Fourth, Assumption 2.3 is satisfied due to (3.4) and the orthogonality of trigonometric series. Finally, Assumption 2.4 holds due to the fact that $\sup_{t\geq 1} |Y_t| \leq 2\pi$.

We now define g_1^*, g_2^* and g_3^* as the corresponding approximations of g_1, g_2 and g_3 with

$$x \in S = [1/128, 2\pi - 1/128] \text{ and } h_i \in H_{iT} = \{[a_i T^{d_i}], \ldots, [b_i T^{c_i}]\}, \tag{3.5}$$

in which $i = 1, 2, 3$, $d_i = \frac{1}{2m_i+3}$ and $c_i = \frac{1}{2m_i+3} + \frac{2m_i-1}{6(2m_i+3)}$.

In the following simulation, we consider the case where $a_i = 1$, $b_i = 2$ and $m_i = 1$ for $i = 1, 2, 3$. Let $F_1(x) = (\sin(x), \sin(2x), \ldots, \sin(h_1 x))^\tau$, $F_2(x) = (\sin(x), \sin(2x), \ldots, \sin(h_2 x))^\tau$ and $F_3(x) = (\cos(x), \cos(2x), \ldots, \cos(h_3 x))^\tau$.

The LS estimator $\hat{\theta}(h) = (\hat{\theta}_1(h)^\tau, \hat{\theta}_2(h)^\tau, \hat{\theta}_3(h)^\tau)^\tau$ of $\theta = (\theta_1^\tau, \theta_2^\tau, \theta_3^\tau)^\tau$ can be computed from (2.2) and (3.1)-(3.4). In the meantime, the optimum value $\hat{h}_D$ can be solved from minimising $\hat{D}(h)$ over H_T.

(i) Compute $\hat{\sigma}^2(h) = \frac{1}{N}\sum_{n=1}^{N}\Big\{Y_{n+2} - \Big[F_1(Y_{n+1})^\tau\hat{\theta}_1(h) + F_2(Y_n)^\tau\hat{\theta}_2(h) + F_3(X_{n+2})^\tau\hat{\theta}_3(h)\Big]\Big\}^2$.

(ii) Compute $GCV(h) = \frac{\hat{\sigma}^2(h)}{\left(1 - \frac{h_1+h_2+h_3}{N}\right)^2}$.

Now, the minimizer $\hat{h}_G$ is $\hat{h}_G = \arg\min_{h\in H_T}\{GCV(h)\}$.

(iii) For the cases of $T = 102, 252, 402, 502,$ and 752, compute for $i = 1, 2, 3$,

$$d_i(\hat{h}_{iG}, \hat{h}_{iD}) = \frac{\hat{h}_{iG}}{\hat{h}_{iD}} - 1, \; d_4(\hat{h}_G, \hat{h}_D) = \frac{\hat{D}(\hat{h}_G)}{\hat{D}(\hat{h}_D)} - 1,$$

$$ASE_i(\hat{h}_G) = \frac{1}{N}\sum_{n=1}^{N}\Big\{F_i(Z_{ni})^\tau\hat{\theta}_i(\hat{h}_G) - g_i(Z_{ni})\Big\}^2,$$

$$ASE_4(\hat{h}_G) = \frac{1}{N}\sum_{n=1}^{N}\Big\{\sum_{i=1}^{3}\Big(F_i(Z_{ni})^\tau\hat{\theta}_i(\hat{h}_G) - g_i(Z_{ni})\Big)\Big\}^2,$$

$$VAR(\hat{h}_G) = \Big|\hat{\sigma}^2(\hat{h}_G) - \sigma^2\Big|,$$

where $\sigma^2 = \frac{\pi^2}{12} = 0.822467$, $\hat{h}_G = (\hat{h}_{1G}, \hat{h}_{1G}, \hat{h}_{3G})^\tau$, $Z_{n1} = Y_{n+1}$, $Z_{n2} = Y_n$ and $Z_{n3} = X_{n+2}$.

The simulation results below were performed 1000 times using the Splus commands (see Becker, Chamber and Wilks (1988)) and the means are tabulated in Table 1 below.

Remark 3.1. Both Theorem 2.2(ii) and Table 1 demonstrate that the rate of convergence of the GCV-based d_i for $i = 1, 2, 3, 4$ is of order $T^{-1/10}$. This suggests the question of whether any better selection rule exists for the truncation parameters. This is beyond the scope of this paper. In addition, both (2.10) and

the simulation results for $ASE_i(\hat{h}_G)$ given in Table 1 show that when h_i is of order $T^{-1/5}$, the rate of convergence of each ASE_i is of order $T^{-4/5}$.

Table 1. Simulation Results for Example 3.1

N	100	250	400	500	750
H_{iT}	{1,...,5}	{1,...,6}	{1,...,6}	{1,...,6}	{1,...,7}
$d_1(\hat{h}_{1G}, \hat{h}_{1D})$	0.10485	0.08755	0.09098	0.08143	0.07943
$d_2(\hat{h}_{2G}, \hat{h}_{2D})$	0.11391	0.07716	0.08478	0.08964	0.07983
$d_3(\hat{h}_{3G}, \hat{h}_{3D})$	0.09978	0.08155	0.08173	0.08021	0.08371
$d_4(\hat{h}_G, \hat{h}_D)$	0.32441	0.22844	0.24108	0.22416	0.22084
$ASE_1(\hat{h}_G)$	0.03537	0.01755	0.01123	0.00782	0.00612
$ASE_2(\hat{h}_G)$	0.02543	0.01431	0.00861	0.00609	0.00465
$ASE_3(\hat{h}_G)$	0.02507	0.01348	0.00795	0.00577	0.00449
$ASE_4(\hat{h}_G)$	0.06067	0.03472	0.02131	0.01559	0.01214
$VAR(\hat{h}_G)$	0.05201	0.03361	0.01979	0.01322	0.01086

Example 3.2. In this example, we consider the Canadian lynx data. This data set is the annual record of the number of Canadian lynx trapped in the MacKenzie River district of North-West Canada for the years 1821 to 1934. Tong (1977) fitted an eleventh-order linear Gaussian autoregressive model to $Y_t = \log_{10}\{\text{number of lynx trapped in the year } (1820 + t)\}$ for $t = 1, 2, \ldots, 114$ ($T = 114$). It follows from the definition of $(Y_t, 1 \le t \le 114)$ that all the transformed values $(Y_t : t \ge 1)$ are bounded.

We apply the above estimation procedure to fit the real data set listed in Example 3.2 by a third-order additive autoregressive model of the form

$$Y_{n+3} = g_1(Y_{n+2}) + g_2(Y_{n+1}) + g_3(Y_n) + e_{n+3},\ n = 1, \ldots, N, \tag{3.6}$$

where $N = T - 3$, $\{g_i : i = 1, 2, 3\}$ are unknown functions, and e_{n+3} is assumed to be independent random error with zero mean and finite variance.

Similarly, we approximate g_1, g_2 and g_3 by

$$g_1^*(u) = \sum_{j=1}^{h_1} f_{1j}(u)\theta_{1j},\ g_2^*(v) = \sum_{j=1}^{h_2} f_{2j}(v)\theta_{2j} \text{ and } g_3^*(w) = \sum_{j=1}^{h_3} f_{3j}(w)\theta_{3j} \tag{3.7}$$

respectively, where $f_{1j}(u) = \sin(ju)$ for $1 \le j \le h_1$, $f_{2j}(v) = \sin(jv)$ for $1 \le j \le h_2$, $f_{3j}(w) = \cos(jw)$ for $1 \le j \le h_3$, and $h_j \in H_{jT} = \{[T^{0.2}], \ldots, [2T^{7/30}]\}$.

Through computing the LS estimator defined by (2.2), the GCV function $GCV(h)$ and the estimator of the error variance (VAR) $\hat{\sigma}^2(h)$ defined before, we

obtain the following polynomial prediction

$$\hat{Y}_{n+3} = \sum_{j=1}^{\hat{h}_{1G}} \sin(jY_{n+2})\theta_{1j} + \sum_{j=1}^{\hat{h}_{2G}} \sin(jY_{n+1})\theta_{2j} + \sum_{j=1}^{\hat{h}_{3G}} \cos(jY_n)\theta_{3j}, \tag{3.8}$$

where $\hat{h}_{1G} = 5$, $\hat{h}_{2G} = \hat{h}_{3G} = 6$, and the coefficients are given in the following Table 2.

Table 2. Coefficients for Equation (3.8)

$\theta_1 = (\theta_{11}, \ldots, \theta_{15})^\tau$	$\theta_2 = (\theta_{21}, \ldots, \theta_{26})^\tau$	$\theta_3 = (\theta_{31}, \ldots, \theta_{36})^\tau$
11.877	−2.9211	−6.8698
18.015	−5.4998	−7.8529
10.807	−4.9084	−7.1952
4.1541	−3.1189	−4.8019
0.7997	−1.2744	−2.0529
	−0.2838	−0.4392

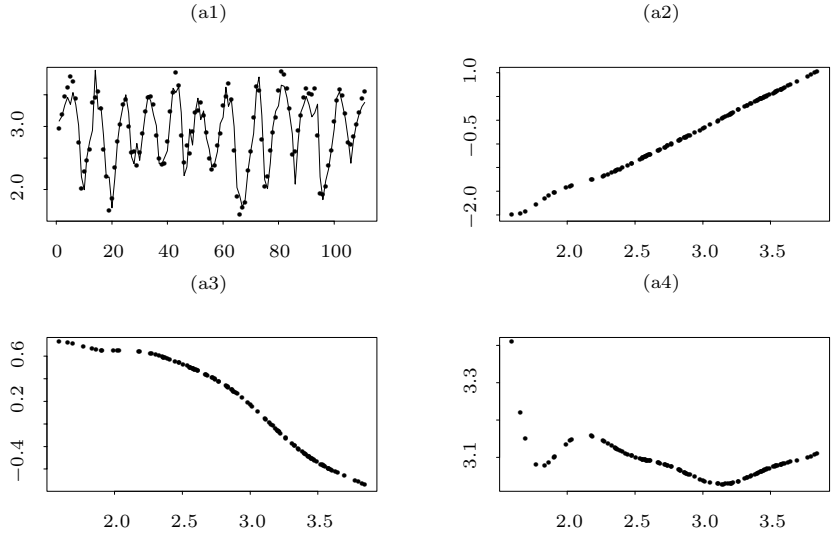

Figure 1. (a1) Fitted values (solid lines) for model (3.6) and the data (dots). (a2) Partial plot of the nonparametric estimator $g_1^*(Y_{n+2})$ versus Y_{n+2}. (a3) Partial plot of the nonparametric estimate $g_2^*(Y_{n+1})$ versus Y_{n+1}. (a4) Partial plot of the nonparametric estimate $g_3^*(Y_n)$ versus Y_n.

The estimator of the error variance was 0.0418. Some plots for Example 3.2 are given in Figure 1 above. Part (a1) provides fitted values (solid lines) for model (3.6) and the data (dots). Partial plot of the nonparametric estimator $g_1^*(Y_{n+2})$ versus Y_{n+2} is given in (a2). Part (a3) presents a partial plot of the nonparametric

estimate $g_2^*(Y_{n+1})$ versus Y_{n+1}. A partial plot of the nonparametric estimate $g_3^*(Y_n)$ versus Y_n is given in (a4).

Remark 3.2. For the Canadian lynx data, Tong (1977) fitted an eleventh–order linear Gaussian autoregressive model to the data, and the estimate of the error variance was 0.0437. Equation (3.8) and Figure 1 show that when using equation (3.6) to fit the data, the estimator of g_1 is almost linear while the estimators of both g_2 and g_3 appear to be nonlinear. This finding is the same as the conclusion reached by Wong and Kohn (WK) (1996), who used a Bayesian based iterative procedure to do the fit. Their estimator of the error variance was 0.0421, comparable to our variance estimator of 0.0418. Moreover, our estimation procedure provides the explicit equation (3.8) and the CPU time for Example 3.2 took just about 2 minutes. By contrast, WK can only provide an iterative estimation procedure for each g_i since their approach depends heavily on the Gibbs sampler.

Remark 3.3. Both Examples 3.1 and 3.2 demonstrate that the explicit estimation procedure can provide some additional information for further diagnostics and statistical inference, as well as produce models with better predictive power than is available from linear models. For example, (3.8) is more appropriate than a completely linear model for the lynx data as mentioned in Remark 3.2. Moreover, (3.8) not only can be calculated at a new design point with the same convenience as in linear models, but also provides the individual coefficients. That can be used to measure the individual influence of each Y_{n+i} for $i = 0, 1, 2$.

Remark 3.4. As mentioned before, some special cases of (1.1) have been discussed through using either the kernel estimation method or the orthogonal series method. More recently, Gao and Yee (2000) have constructed a kernel-based estimation procedure for a partially linear model, conducted a small sample study for (3.1) with $X_t \equiv 0$, and obtained similar large and small sample results. Through comparing the small sample simulation results for the same model, we have found that both the kernel method and the orthogonal series method work well numerically, and there is little difference between the two different methods for the same model. However, the kernel estimation method has not been applied extensively to estimate (1.1). By contrast, the orthogonal series method has been successfully applied to determine (1.1).

Acknowledgements

The authors thank the Editors, the Associate Editors and two anonymous referees for their constructive comments. Thanks from the first author also go to Dr Qiwei Yao and Professor Enno Mammen for sending discussion papers. The second author acknowledges financial support from the EU under the Human

Capital Programme (CHRX-CT 94-0693), the Engineering and Physical Science Research Council of UK and the Hong Kong University CRCG award. The third author thanks the Australian Research Council for a large grant support.

Appendix

A.1. Technical lemmas

For simplicity, let C ($C < \infty$) denote a positive constant which may have different values at each appearance throughout this section.

As the proof of Theorems 2.1–2.3 is extremely technical, only an outline is given. Detailed proofs can be obtained from Gao, Tong and Wolff (2000), available upon request.

Let $c^2_{\max} = \max_{1\le i\le p}\max_{1\le j\le h_i} c^2_{ij}$ and $c^2_{\min} = \min_{1\le i\le p}\min_{1\le j\le h_i} c^2_{ij}$.

Lemma A.1. *Let $\delta(h) > 0$ be a sequence satisfying $\inf_{h\in H_T}(\delta(h)/M(h))\sqrt{T} > 0$ and $T^\tau \sum_{h\in H_T}\delta(h) \to 0$ as $T \to \infty$. Assume the conditions of Theorem 2.1 hold. Then*

$$c^2_{\min} + o_p(\delta(h)) \le \lambda_{\min}\Big(\frac{1}{T}F^\tau F\Big) \le \lambda_{\max}\Big(\frac{1}{T}F^\tau F\Big) \le c^2_{\max} + o_p(\delta(h))$$

and for all $k = 1, \dots, M(h)$,

$$\lambda_k\Big(\frac{1}{T}F^\tau F - I(h)\Big) = o_p(\delta(h)).$$

Here $M(h) = \sum_{i=1}^p h_i$, $I(h) = diag(c^2_{11}, \dots, c^2_{1h_1}; c^2_{21}, \dots, c^2_{2h_2}; \dots, c^2_{p1}, \dots, c^2_{ph_p})$ is a $M(h)\times M(h)$ order diagonal matrix, $\lambda_{\min}(B)$ and $\lambda_{\max}(B)$ denote the smallest and largest eigenvalues of matrix B respectively, and $\{\lambda_k(D)\}$ denotes the kth eigenvalue of matrix D.

Proof. See Lemma A.2 of Gao, Tong and Wolff (2000).

Lemma A.2. *Assumptions 2.1–2.4 hold. Let $\{j(1), \cdots, j(r)\}$ be r distinct positive integers, and define*

$$\phi(X_{j(1)}, \dots, X_{j(r)}) = \prod_{t=1}^{q}\prod_{s=1,\neq t}^{q} A(X_{j(t)}, X_{j(s)})^{l_{t,s}} \prod_{i=1}^{r}\phi_{j(i)}(X_{j(i)}),$$

where $X_t = (X_{t1}, \dots, X_{tp})^\tau$, $A(X_s, X_t) = \sum_{i=1}^p \sum_{j=1}^{h_i} c_{ij}^{-2} f_{ij}(X_{si}) f_{ij}(X_{ti})$, $\{\phi_{j(k)} : k \ge 1\}$ are real-valued functions such that $|\phi_{j(k)}(\cdot)| \le M_k < \infty$, $\{l_{t,s} : t, s \ge 1\}$ are nonnegative integers, and $q \le r$. Let $A_1, \dots, A_v$ be a partition of $\{j(1), \dots, j(r)\}$. Then there exists a finite positive constant c such that

$$\Big|\int \phi dP_{(X_{j(1)},\dots,X_{j(r)})} - \int \phi dP_{(X_t, t\in A_1)} \cdots dP_{(X_t, t\in A_v)}\Big| \le C \cdot M(h)^l \tilde{\alpha}(d),$$

where $d = \inf\{d(A_i, A_j) : i, j = 1, 2, \cdots, v; i \le j\}$, $d(A_i, A_j) = \inf\{|s - t|, s \in A_i, t \in A_j\}$, $\tilde{\alpha}(d) = \sup_{j \ge d} \alpha(j)$, *and* $l = \sum_{t=1}^{q} \sum_{s=1, \neq t}^{q} l_{t,s}$.

Proof. The proof is similar to that of Theorem 6.2.1 of Györfi, Härdle, Sarda and Vieu (1989). See Lemma A.3 of Gao, Tong and Wolff (2000) for more details.

Remark A.1. Lemma A.2 is useful in itself for dealing with the estimation of strictly stationary and mixing processes. It is as important as Proposition 1 of Hart and Vieu (1990) in the kernel estimation case.

A.2. Proofs of Theorems 2.1–2.3

A.2.1. Proof of Theorem 2.1

By (2.2) and (2.4), we have uniformly over $h \in H_T$,

$$\hat{D}(h) = \frac{1}{T}\{e^\tau P(h) e + \Delta^\tau \Delta - \Delta^\tau P(h) \Delta\}, \tag{A.1}$$

$$D(h) = E[\hat{D}(h)] = \frac{1}{T}\left\{E[e^\tau P(h) e] + E[\Delta^\tau \Delta] - E[\Delta^\tau P(h) \Delta]\right\}, \tag{A.2}$$

where $e = (e_1, \ldots, e_T)^\tau$, $P(h) = F(F^\tau F)^+ F^\tau$, $\Delta = F\theta - G = \sum_{i=1}^{p} [F_i \theta_i - G_i]$, $F_i = (F_i(X_{1i}), \ldots, F_i(X_{Ti}))^\tau$, and $G_i = (g_i(X_{1i}), \ldots, g_i(X_{Ti}))^\tau$.

Applying Lemma A.1, one can prove, uniformly over $h \in H_T$, $\Delta^\tau P(h) \Delta = \Delta^\tau F(F^\tau F)^+ F^\tau \Delta \le \lambda_{\max}((F^\tau F)^+) \Delta^\tau F F^\tau \Delta = o_p(T^{-\tau} \Delta^\tau \Delta)$ using the fact that $(1/T)\lambda_{\max}(FF^\tau) = o_p(T^{-\tau})$.

Similarly, one can show that, as $T \to \infty$,

$$\frac{E[\Delta^\tau P \Delta]}{E[\Delta^\tau \Delta]} = o(T^{-\tau}). \tag{A.3}$$

Let $\hat{D}_1(h) = \frac{1}{T}\{e^\tau P(h) e + \Delta^\tau \Delta\}$ and $D_1(h) = E[\hat{D}_1(h)] = \frac{1}{T}\{E[e^\tau P(h) e] + E[\Delta^\tau \Delta]\} = \frac{\sigma^2}{T} \sum_{i=1}^{p} h_i + \frac{1}{T} E[\Delta^\tau \Delta]$.

If (2.6) holds, then we have

$$D_1(h) = E[\hat{D}_1(h)] \approx \frac{\sigma^2}{T} \sum_{i=1}^{p} h_i + \sum_{i=1}^{p} C_i h_i^{-2(m_i+1)}. \tag{A.4}$$

Obviously, (A.1)–(A.4) imply

$$\sup_{h \in H_T} \frac{|\hat{D}(h) - \hat{D}_1(h)|}{\hat{D}_1(h)} = o_p(T^{-\tau}), \tag{A.5}$$

$$\sup_{h \in H_T} \frac{|D(h) - D_1(h)|}{D_1(h)} = o(T^{-\tau}). \tag{A.6}$$

It follows from (A.4) and (A.6) that the minimizer h_D of $D(h)$ over H_T is a vector of the minimizers $\{h_{iD} : 1 \le i \le p\}$, in which h_{iD} is proportional to $\left[T^{\frac{1}{2(m_i+1)+1}}\right] \in H_{iT}$. This suggests defining H_{iT} in Assumption 2.2.

In view of (A.5) and (A.6), in order to prove

$$\sup_{h\in H_T} \frac{|\hat{D}(h) - D(h)|}{D(h)} = o_p(T^{-\tau}), \tag{A.7}$$

it suffices to show that

$$\sup_{h\in H_T} \frac{|\hat{D}_1(h) - D_1(h)|}{D_1(h)} = o_p(T^{-\tau}). \tag{A.8}$$

First, we show that

$$\sup_{h\in H_T} \frac{|e^\tau P(h)e - \sum_{1\le s,t\le T} a_{st}e_s e_t|}{M(h)} = o_p(T^{-\tau}), \tag{A.9}$$

where $a_{st} = \frac{1}{T}\sum_{i=1}^{p}\sum_{j=1}^{h_i} c_{ij}^{-2} f_{ij}(X_{si}) f_{ij}(X_{ti})$ and $M(h) = \sum_{i=1}^{p} h_i$. In order to prove (A.9), it suffices to show that

$$\sup_{h\in H_T} M(h)^{-1}|e^\tau(P(h) - Q(h))e| = o_p(T^{-\tau}), \tag{A.10}$$

where $Q(h) = \{a_{st}\}_{1\le s,t\le T}$ is a matrix of order $T\times T$. The proof of (A.10) follows from the Markov inequality and Assumption 2.3.

Second, one needs to show

$$\sup_{h\in H_T} \frac{|\sum_{1\le s,t\le T} a_{st}e_s e_t - M(h)\sigma^2|}{d(h)} = o_p(T^{-\tau}), \tag{A.11}$$

where $d(h) = M(h)\sigma^2 + E[\Delta^\tau\Delta]$. Applying Lemmas A.1 and A.2, one can prove (A.11).

With the proof of (A.11), (A.7) holds. Thus the proof of Theorem 2.1 is completed.

A.2.2. Proof of Theorem 2.2

As Theorem 2.2(ii) implies Theorem 2.2(i), we prove only Theorem 2.2(ii). In order to prove (2.11), in view of (2.4) and (2.8), it suffices to show that

$$\sup_{h,h'\in H_T} \frac{|\hat{D}(h) - \hat{D}(h') - [GCV(h) - GCV(h')]|}{\hat{D}(h)} = o_p(T^{-\tau}). \tag{A.12}$$

The proof of (A.12) is similar to that of (4.29) of Gao (1998), details are in Gao, Tong and Wolff (2000). Thus, the proof of (2.11) is completed.

We now finish the proof of (2.12). Observe that

$$\frac{D_1(\hat{h}_G) - D_1(\hat{h}_D)}{D_1(\hat{h}_D)}$$
$$= \frac{D_1(\hat{h}_G)-\hat{D}_1(\hat{h}_G)}{D_1(\hat{h}_G)}\frac{D_1(\hat{h}_G)}{D_1(\hat{h}_D)} + \frac{\hat{D}_1(\hat{h}_G)-\hat{D}_1(\hat{h}_D)}{\hat{D}_1(\hat{h}_D)}\frac{\hat{D}_1(\hat{h}_D)}{D_1(\hat{h}_D)} + \frac{\hat{D}_1(\hat{h}_D)-D_1(\hat{h}_D)}{D_1(\hat{h}_D)}.$$

To prove (2.12), in view of (A.8), it suffices to show that

$$\frac{[\hat{D}_1(\hat{h}_G) - \hat{D}_1(\hat{h}_D)]}{\hat{D}_1(\hat{h}_D)} = o_p(T^{-\tau}),$$

which follows from (2.11) and (A.5).

A.2.3. Proof of Theorem 2.3

The proof of Theorem 2.3 follows from the definition of $\hat{\sigma}^2(h)$ and the proof of Theorem 2.2.

References

Agarwal, G. G. and Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.* **8**, 1307-1325.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The New S Language*. Wadsworth and Brooks Advanced Books and Software. Pacific Grove, California.

Bera, A. K. and Higgins, M. L. (1993). A survey of ARCH models: properties, estimation and testing. *J. Econom. Surveys* **7**, 305-366.

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.

Chen, R. and Tsay, R. (1993). Nonlinear additive ARX models. *J. Amer. Statist. Assoc.* **88**, 955-967.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation*. Springer, New York.

Doukhan, P. (1995). Mixing: Properties and Examples. *Lecture Notes in Statistics* **85**. Springer, New York.

Eastwood, B. and Gallant, R. (1991). Adaptive truncation rules for seminonparametric estimates that achieve asymptotic normality. *Econom. Theory* **7**, 307-340.

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.

Gallant, A. R. and Souza, G. (1991). On the asymptotic normality of Fourier flexible form estimates. *J. Econom.* **50**, 329-353.

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

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 semiparametric single-index and partially nonlinear 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. *Statist. Sinica* **7**, 1155-1169.

Gao, J., Tong, H. and Wolff, R. (1999). Model specification tests in nonparametric stochastic regression models. *J. Multivar. Anal.* **81** (in press).

Gao, J., Tong, H. and Wolff, R. (2000). Adaptive orthogonal series estimation in additive stochastic regression models. Technical report, available at www.maths.uwa.edu.au/$\tilde{}$jiti.

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.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989). Nonparametric Curve Estimation for Time Series. *Lecture Notes in Statistics* **60**. Springer, New York.

Hall, P. and Patil, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.* **23**, 905-928.

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

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., Hall, P. and Marron, J. (1992). Regression smoothing parameters that are not far from their optimum. *J. Amer. Statist. Assoc.* **87**, 227-233.

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

Härdle, W. and Vieu, P. (1992). Kernel regression smoothing of time series. *J. Time Ser. Anal.* **13**, 209-232.

Hart, J. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.* **18**, 873-890.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hjellvik, V., Yao, Q. W. and Tjøstheim, D. (1998). Linearity testing using local polynomial approximation. *J. Statist. Plann. Inference* **68**, 295-321.

Hurvich, C. and Tsai, C. L. (1995). Relative rates of convergence for efficient model selection criteria in linear regression. *Biometrika* **82**, 418-425.

Kashin, B. S. and Saakyan, A. A. (1989). *Orthogonal Series*. Translations of Mathematical Monographs, Vol. **75**, American Mathematical Society.

Li, K. C. (1985). From Stein's unbiased risk estimates to the method of generalized cross-validation. *Ann. Statist.* **13**, 1352-1377.

Li, K. C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14**, 1101-1112.

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.

Linton, O. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* **84**, 469-473.

Linton, O. and Nielson, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93-100.

Masry, E. and Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series. *Econom. Theory* **11**, 228-289.

Masry, E. and Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econom. Theory* **13**, 214-251.

Schumaker, L. (1981). *Spline Functions*. John Wiley, New York.

Seber, G. A. F. (1977). *Linear Regression Analysis*. John Wiley, New York.
Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*. John Wiley, New York.
Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statist. Sinica* **7**, 221-264.
Shi, P. and Tsai, C. L. (1999). Semiparametric regression model selections. *J. Statist. Plann. Inference* **77**, 119-139.
Tjøstheim, D. (1994). Nonlinear time series: a selective review. *Scand. J. Statist.* **21**, 97-130.
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. (1977). Some comments on the Canadian lynx data (with discussion). *J. R. Statist. Soc. Ser. A* **140**, 432-435, 448-468.
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.
Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.
Wong, C. M. and Kohn, R. (1996). A Bayesian approach to estimating and forecasting additive nonparametric autoregressive models. *J. Time Ser. Anal.* **17**, 203-220.
Yao, Q. W. and Tong, H. (1994). On subset selection in nonparametric stochastic regression. *Statist. Sinica* **4**, 51-70.

Department of Mathematics and Statistics, The University of Western Australia, Crawley WA 6009, Australia.

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

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong.

E-mail: htong@hku.hk

School of Mathematical Science, Queensland University of Technology, GPO Box 2434, Brisbane Qld 4001, Australia.

E-mail: r.wolff@qut.edu.au

(Received February 1998; accepted May 2001)