Statistica Sinica 4(1994), 51-70

ON SUBSET SELECTION IN NON-PARAMETRIC STOCHASTIC REGRESSION

Qiwei Yao

University of Kent
and
Southeast University

Howell Tong

University of Kent

Abstract: This paper is concerned with the use of a cross-validation method based on the kernel estimate of the conditional mean for the subset selection of stochastic regressors within the framework of non-linear stochastic regression. Under the assumption that the observations are strictly stationary and absolutely regular, we show that the cross-validatory selection is consistent. Furthermore, two kinds of asymptotic efficiency of the selected model are proved. Both simulated and real data are used as illustrations.

Key words and phrases: Absolutely regular, cross-validation, efficiency, kernel estimation, heteroscedasticity, non-linear stochastic regression, subset selection.

0. Introduction

Until quite recently non-linear time series modeling has been dominated by the parametric approach (see, e.g., Tong (1990), for an introduction). The non-parametric approach has a much shorter history, although there has been a rapidly increasing literature on the kernel approach to non-linear time series modeling; see e.g. Auestad and Tjøstheim (1990), Cheng and Tong (1992), Masry (1989) and Truong and Stone (1990), to name just a few. The class of non-linear autoregressive models appears to be the most readily amenable to the kernel treatment by virtue of its affinity with non-linear regression models. As a result, much of the non-parametric literature in this area concentrates on the (univariate) non-linear autoregressive models. However, within this context, only recently has the important problem of order determination been addressed by Auestad and Tjøstheim (1990), Tong (1990), Tjøstheim and Auestad (1992a,b), and Cheng and Tong (1992). It has been pointed out in the last reference that order determination has an important role to play beyond its statistical confine. Specifically, in the important science of chaos, the detection of a low dimensional attractor

from (possibly long) experimental data remains one of the great challenges todate. One key question in the detection exercise is an estimate of the dimension of the space in which the attractor lives, namely the so-called embedding dimension. It turns out that the order determination mentioned above has an important contribution to make in this respect and hence, in the detection of chaos.

In many situations, a *univariate* autoregressive model is rather restrictive. For example, there may exist some 'exogenous' time series or 'covariates' in, say, a non-linear transfer function system such as in the so-called TARSO model (see, e.g., Tong (1990, p.101)). Even within the context of univariate autoregressive modeling, we may sometimes wish to identify the 'dominant' lag variables, i.e. an appropriate subset autoregressive model, or to identify the lag structure in the 'diffusion term' (i.e. conditional variance) as in the so-called ARCH models (see, e.g., Tjøstheim and Auestad (1992b)). One framework which encompasses all these cases is the stochastic regression. Lai and Wei (1982) have discussed least squares estimation of parameters of this model, i.e. within a parametric setting.

In this paper, we use the framework of non-linear stochastic regression. Without assuming any specific form for the regression function, our objective is to introduce a cross-validation method based on the kernel estimation of the conditional mean to determine a proper subset of the stochastic regressors to fit the underlying regression model. To justify this approach, we show that under the assumption the observations are strictly stationary and absolutely regular, the selected subset is a consistent estimate of the optimum subset of the stochastic regressors, and, moreover, the fitted model is asymptotically efficient in two senses. To highlight the statistical ideas, without the accompanyment of too much technical detail, we deal only with the case in which it is assumed that the regressors have a compact support. We argue that this assumption is not so restrictive in practice since any real data set which is reasonably stationary could be considered a bounded set. On the other hand, similar to Theorem 6 of Zhang (1991), it can be proved theoretically that by introducing some random weight functions in the residual sums of squares, the cross-validatory selection is still consistent in the case that the regressors are not bounded. Zhang adopted this strategy in studying the cross-validatory method with i.i.d. observations. He also discussed some statistical issues involved in the model selection.

The plan of the paper is as follows. In §1 we introduce the non-linear stochastic regression model and the notion of the optimum subset of the regressors. We show that this model can also be used to investigate possible conditional heteroscedasticity in a general regression model. In §2 we describe the cross-validatory selection procedure and state the main results: Theorem 1 on the consistency of the cross-validation selection, and Theorem 2 on the asymptotic

efficiency of the fitted model. The method is illustrated with the Canadian lynx data, the Wolf's sunspot numbers, and also two simulated models in §3. All lengthy proofs are relegated to the appendices.

1. Model

Suppose that $\{(Y_t, X_t); t = 1, \ldots, N\}$ is a strictly stationary random sequence, $Y_t \in R^1$ and $X_t = (X_{t1}, \ldots, X_{tL}) \in R^L$ $(L \geq 1)$. Consider the regression model

$$Y_t = E(Y_t|X_t) + \epsilon_t \equiv F(X_t) + \epsilon_t, \quad 1 \leq t \leq N, \tag{1.1}$$

where $\epsilon_t = Y_t - E(Y_t|X_t)$. Obviously $E(\epsilon_t|X_t) = 0$. The goal of this paper is to determine, without assuming that F is known, a proper subset $\{X_{ti_1}, \ldots, X_{ti_d}\}$ with d as small as possible which provides (almost) the same information on Y_t as $\{X_{t1}, \ldots, X_{tL}\}$, i.e.

$$E(Y_t \mid X_{ti_1}, \ldots, X_{ti_d}) = E(Y_t|X_t), \qquad \text{a.s.}$$

We now formalize the problem. First, we introduce the following definition, which is based on the *variance function*,

$$\sigma^2(i_1, \ldots, i_k) \equiv E[Y_t - E(Y_t \mid X_{ti_1}, \ldots, X_{ti_k})]^2, \tag{1.2}$$

for $1 \leq k \leq L$, $1 \leq i_1 < \cdots < i_k \leq L$.

Definition. If there exists a subset of $\{1, \ldots, L\}$, say $\{1, \ldots, d\}$, with $d \leq L$, for which

(i) $\sigma^2(1, \ldots, d) = \sigma^2(1, \ldots, L)$,

(ii) for any $\{i_1, \ldots, i_k\} \subset \{1, \ldots, L\}$ with $k \leq d$ and $\{1, \ldots, d\} \neq \{i_1 \ldots, i_k\}$,

$$\sigma^2(i_1, \ldots, i_k) > \sigma^2(1, \ldots, L), \tag{1.3}$$

then $\{X_{t1}, \ldots, X_{td}\}$ is called the *optimum subset* of the regressors of Y_t.

Remark 1. It might be possible that there exists another subset $\{i_1, \ldots, i_d\} \subset \{1, \ldots, L\}$ but $\{i_1, \ldots, i_d\} \neq \{1, \ldots, d\}$ for which the equality $\sigma^2(i_1, \ldots, i_d) = \sigma^2(1, \ldots, L)$ holds. This makes our discussion more complicated. Since it is not a likely case in practice, we agree to discard this case. Note that the notion of an optimum subset obviates any philosophical debate on the existence of a *true model*, which we do not wish to enter in this paper.

From now on, we always make the following assumption.

(M1) The optimum subset of the regressors of Y_t exists. There is no loss of generality to assume that the optimum subset is $\{X_{t1}, \ldots, X_{td}\}$, $(1 \leq d \leq L)$.

It is easy to see that under the assumption (M1),

$$E(Y_t \mid X_{t1}, \ldots, X_{td}) = E(Y_t|X_t) \quad \text{a.s.}, \quad 1 \le t \le N. \tag{1.4}$$

Thus the model (1.1) can be expressed as

$$Y_t = E(Y_t \mid X_{t1}, \ldots, X_{td}) + \epsilon_t \quad \text{a.s.}, \quad 1 \le t \le N.$$

Before ending this section, it is worth mentioning that the above framework can also be used to investigate possible conditional heteroscedasticity in a regression model. To see this, let us assume that in the model (1.1), ϵ_t is of the form

$$\epsilon_t = \lambda(e_t, X_{tj_1}, \ldots, X_{tj_q}), \quad 1 \le j_1 < \cdots < j_q \le L, \ 1 \le q \le L, \tag{1.5}$$

where $\{e_t\}$ is a strictly stationary noise process and e_t is independent of $\{Y_{t-1}, \ldots, Y_1, X_t, \ldots, X_1\}$. Suppose that we regress Y_t^2 on $X_{t1}, \ldots, X_{tL}$. If there is no "redundancy" amongst $\{X_{tj_1}, \ldots, X_{tj_q}\}$ in the expression (1.5), which can be formulated precisely in a similar way as (1.3), the optimum subset of the regressors for Y_t^2 will be the union of $\{X_{tj_1}, \ldots, X_{tj_q}\}$ and $\{X_{t1}, \ldots, X_{td}\}$ (under assumption (M1)). Thus, we suggest that heteroscedasticity should be investigated if the selected subset of the regressors of Y_t^2 contains the selected subset of Y_t as a proper subset. Further study on the heteroscedasticity will be reported elsewhere.

2. Cross-Validatory Selection

We now propose to use the cross-validation approach, based on the kernel estimate of the regression function, for the selection of the optimum subset of regressors, and make the following assumptions on the model. First, note that c always denotes some finite positive constant; it may be different in different places.

(M2) X_t has probability density function f, and $G \equiv \{x : f(x) > 0\}$ is a compact subset of R^L.

(M3) f satisfies a Lipschitz condition, i.e.

$$|f(x_1) - f(x_2)| \le c\|x_1 - x_2\|, \quad \forall\, x_1, x_2 \in G$$

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

(M4) For $t \ge 1$, $E(Y_t \mid X_t, \ldots, X_1, Y_{t-1}, \ldots, Y_1) = E(Y_t|X_t)$.

(M5) $E|Y_t|^6 < \infty$.

(M6) For $F(\cdot)$ given in (1.1), $|F(x_1) - F(x_2)| \le c\|x_1 - x_2\|^\mu$ for all $x_1, x_2 \in G$, where $\mu > 0$ is a constant.

(M7) Let $\beta_n = \sup_{k\geq 1} E[\sup_{A\in\mathcal{F}_{k+n}^{\infty}} |P(A|\mathcal{F}_1^k) - P(A)|]$, where $\mathcal{F}_k^n$ is the sigma field generated by $\{(Y_t, X_t); t = k, k+1, \ldots, n\}$. Then $\beta_n = O(n^{-(2+\delta)/\delta})$ as $n \to \infty$, where δ is a constant in $(0, 2/5)$. Furthermore, there exists a positive integer $N_1 = N_1(N)$ such that for $N_2 \equiv [N/(2N_1)] > 0$, $\limsup_{N\to\infty}(1 + 6e^{1/2}\beta_{N_1}^{1/(1+N_2)})^{N_2} < \infty$.

Assumption (M2) implies that X_t is bounded, which avoids the 'infinite integration problem' in the asymptotic expansion encountered by Auestad and Tjøstheim (1990). The condition that $\{x : f(x) > 0\}$ is closed guarantees the uniform convergence of the kernel estimate for the regression functions (cf. Györfi et al. (1989, §3.3.2)). (M3)-(M6) are self-explanatory. Assumption (M7) implies that the process $\{Y_t, X_t\}$ is absolutely regular. Further, the assumptions on the rate of convergence of β_n allow us to apply the results of Yoshihara (1976) and Roussas (1988). In fact, the condition $\beta_n = O(n^{-(2+\delta)/\delta})$ is for technical convenience, and is not the weakest possible. Further, the assumption that the process is absolutely regular could be replaced by the weaker mixing conditions (cf. Györfi et al. (1989)), in which case the proof would then contain more technical details.

We now state the kernel estimate of the regression function, based on which the cross-validatory residual sum of squares will be formed. For analytical convenience, the Nadaraya-Watson method will be used (cf. Nadaraya (1965) and Watson (1964)).

For any $1 \leq i_1 < \cdots < i_k \leq L$, with $1 \leq k \leq L$, let $p_k(\cdot)$ be a probability density function on R^k, and $h(i_1, \ldots, i_k, N) \equiv N^{-\lambda(i_1,\ldots,i_k)}$ be the bandwidth, where $\lambda(\cdot)$ is a positive function. Simply, write it as

$$h \equiv h(k, N) \equiv N^{-\lambda(k)}, \tag{2.1}$$

though h can be different for different regressors. Note that $\lambda(k)$ stands for the function $\lambda(i_1, \ldots, i_k)$.

The kernel estimate for the density of $(X_{ti_1}, \ldots, X_{ti_k})$ is

$$\hat{f}_{i_1,\ldots,i_k}(x_1, \ldots, x_k) = (Nh^k)^{-1} \sum_{t=1}^{N} p_k\left(\frac{x_1 - X_{ti_1}}{h}, \ldots, \frac{x_k - X_{ti_k}}{h}\right). \tag{2.2}$$

The cross-validatory approach leaves one out each time, i.e. for $s = 1, \ldots, N$, define

$$\hat{f}^{(s)}_{i_1,\ldots,i_k}(x_1, \ldots, x_k) = \frac{h^{-k}}{N-1} \sum_{t\neq s} p_k\left(\frac{x_1 - X_{ti_1}}{h}, \ldots, \frac{x_k - X_{ti_k}}{h}\right), \tag{2.3}$$

and estimate the regression function $E(Y_t \mid X_{ti_1}, \ldots, X_{ti_k})$ by

$$\hat{F}^{(s)}_{i_1,\ldots,i_k}(x_1, \ldots, x_k)$$

$$= \frac{h^{-k}}{N-1}\sum_{t\neq s} Y_t p_k\left(\frac{x_1 - X_{ti_1}}{h}, \ldots, \frac{x_k - X_{ti_k}}{h}\right) \Big/ \hat{f}^{(s)}_{i_1,\ldots,i_k}(x_1, \ldots, x_k), \quad (2.4)$$

and interpret "0/0" as 0. Based on this, the cross-validatory residual sum of squares is defined as

$$\mathrm{CV}(i_1, \ldots, i_k) = \frac{1}{N}\sum_{s=1}^{N}\left\{Y_s - \hat{F}^{(s)}_{i_1,\ldots,i_k}(X_{si_1}, \ldots, X_{si_k})\right\}^2, \quad (2.5)$$

for all $1 \leq i_1 < \cdots < i_k \leq L$, $k = 1, \ldots, L$.

The cross-validatory criterion. Choose that subset S_{cv} of $\{1, \ldots, L\}$ which minimizes $\mathrm{CV}(i_1, \ldots, i_k)$ over all $1 \leq i_1 < \cdots < i_k \leq L$ with $1 \leq k \leq L$.

Under the following assumptions on the kernel densities and the bandwidths, Theorem 1 below shows that the probability of the event that $\{X_{ti} : i \in S_{cv}\}$ coincides with the optimum subset of regressors of Y_t tends to 1. To simplify the discussion, we prove the theorem only with deterministic bandwidths, though we strongly believe that the same conclusion holds for the more relevant method with data-driven bandwidths (cf. Zhang (1991), also Section 3).

(K1) For $1 \leq k \leq L$, $p_k(\cdot)$ is bounded, and

$$|p_k(x_1) - p_k(x_2)| \leq c\|x_1 - x_2\|, \quad \text{for all } x_1, x_2 \in R^k.$$

(K2) For $1 \leq k \leq L$, $0 < k\lambda(k) < 1/2$.

(K3) For $N_1 = N_1(N)$ given in (M7), $\limsup_{N\to\infty} N_2/N^{\lambda(k)} < \infty$ for all $1 \leq k \leq L$.

(K4) For μ given in (M6), $(k+\mu)\lambda(k) > 1/2$ for all $1 \leq k \leq L$.

(K5) $k\lambda(k)$ is a strictly increasing function of k in the range of $1 \leq k \leq L$.

Remark 2. We can take $\lambda(k) = 1/(2k+\mu)$, which satisfies conditions (K2), (K4) and (K5).

Assumptions (K1)–(K3) were introduced by Roussas (1988). Assumption (K4) is a standard condition in non-parametric inference. Assumption (K5) is essential for the proof of asymptotically negligible over-fitting. Obviously, the above assumptions do not offer an explicit construction of the bandwidth. In practice, a frequently used bandwidth selection technique is the cross-validation method. Our experience suggests that the bandwidths selected in this data-driven way seem to satisfy the assumption (K5) with minor modification (cf. Examples 3 and 4 in Section 3).

Theorem 1. *Under assumptions* (M1)–(M7) *and* (K1)–(K5),

$$\lim_{N\to\infty} P(S_{cv} = \{1, \ldots, d\}) = 1.$$

The proof of Theorem 1 is based on the following Lemma 1, the proof of which is postponed to Appendix A.

Lemma 1. *Suppose assumptions* (M2)–(M7) *and* (K1)–(K4) *hold.*
(i) *For any* $1 \le i_1 < \cdots < i_k \le L$, $1 \le k \le L$, *as* $N \to \infty$,

$$\mathrm{CV}(i_1, \ldots, i_k) \xrightarrow{P} \sigma^2(i_1, \ldots, i_k),$$

where $\sigma^2(i_1, \ldots, i_k)$ *is given in* (1.2).
(ii) *For some* $1 \le i_1 < \cdots < i_k \le L$, $1 \le k \le L$, *let* $\epsilon_t^{(i_1,\ldots,i_k)} = Y_t - E(Y_t \mid X_{ti_1}, \ldots, X_{ti_k})$. *If*

$$\epsilon_s^{(i_1,\ldots,i_k)} = \epsilon_s, \quad \text{a.s. for all} \quad s = 1, \ldots, N, \tag{2.6}$$

where $\{\epsilon_s, s = 1, \ldots, N\}$ *is as given in* (1.1), *then, as* $N \to \infty$,

$$\mathrm{CV}(i_1, \ldots, i_k) = \sigma_N^2 + \gamma_{i_1,\ldots,i_k}/(Nh^k(k,N)) + o_p(N^{-1}h^{-k}(k,N)),$$

where $\sigma_N^2 = N^{-1}\sum_{t=1}^N \epsilon_t^2$, *and*

$$\gamma_{i_1,\ldots,i_k} = E\{\epsilon_t^2 / f_{i_1,\ldots,i_k}(X_{ti_1}, \ldots, X_{ti_k})\} \cdot \int p_k^2(x)dx. \tag{2.7}$$

Proof of Theorem 1. For any $1 \le i_1 < \cdots < i_k \le L$, $1 \le k \le L$, if

$$\sigma^2(i_1, \ldots, i_k) > \sigma^2(1, \ldots, L) = \sigma^2(1, \ldots, d),$$

Lemma 1 (i) implies that for such $(i_1, \ldots, i_k)$

$$P\{\mathrm{CV}(1, \ldots, d) < \mathrm{CV}(i_1, \ldots, i_k)\} \to 1.$$

If $\sigma^2(i_1, \ldots, i_k) = \sigma^2(1, \ldots, d)$, it is easy to see that the relation (2.6) holds in this case. From the definition of the optimum subset of regressors, k must be larger than d. Assumption (K5) implies that

$$h^d(d,N)/h^k(k,N) = N^{k\lambda(k)-d\lambda(d)} \to \infty, \quad \text{as} \quad N \to \infty. \tag{2.8}$$

Hence, by Lemma 1 (ii),

$$\begin{aligned} & P\Big\{\mathrm{CV}(i_1, \ldots, i_k) - \mathrm{CV}(1, \ldots, d) > 0\Big\} \\ = \; & P\Big\{Nh^d(d,N)[\mathrm{CV}(i_1, \ldots, i_k) - \mathrm{CV}(1, \ldots, d)] > 0\Big\} \\ = \; & P\left\{\gamma_{i_1,\ldots,i_k}\frac{h^d(d,N)}{h^k(k,N)} - \gamma_{1,\ldots,d} + o_p\left(\frac{h^d(d,N)}{h^k(k,N)}\right) > 0\right\} \to 1. \end{aligned} \tag{2.9}$$

Consequently, $P(S_{cv} = \{1, \ldots, d\}) \to 1$. This completes the proof.

Based on $S_{cv} = \{\tau_1, \ldots, \tau_{\hat{d}}\}$ say, a natural estimate of the regression function $F(X_t) = E(Y_t|X_t)$ can be defined as follows.

$$\hat{F}(x) = N^{-1}[h(\hat{d}, N)]^{-\hat{d}} \sum_{t=1}^{N} Y_t p_{\hat{d}}\left(\frac{x_{\tau_1} - X_{t\tau_1}}{h(\hat{d}, N)}, \ldots, \frac{x_{\tau_{\hat{d}}} - X_{t\tau_{\hat{d}}}}{h(\hat{d}, N)}\right) \times \hat{f}^{-1}_{\tau_1, \ldots, \tau_{\hat{d}}}(x_{\tau_1}, \ldots, x_{\tau_{\hat{d}}}), \tag{2.10}$$

where $x = (x_1, \ldots, x_L) \in R^L$, and h and $\hat{f}$ are given in (2.1) and (2.2) respectively. Obviously $\hat{d}$, $\tau_1, \ldots, \tau_{\hat{d}}$ are random functions defined on the whole sample $\{(Y_t, X_t),\ 1 \le t \le N\}$. The statements (i) and (ii) in the following theorem represent two kinds of asymptotic efficiency of the estimate $\hat{F}$. The proof is given in Appendix B.

Theorem 2. *Suppose that assumptions* (M1)–(M7) *and* (K1)–(K5) *hold. Then as* $N \to \infty$, *the following limits hold if either* $E|Y_t|^k < \infty$ *for any* $k \ge 1$, *or* $k\lambda(k) < 2/5$ *for* $1 \le k \le L$.

(i) *For any random vector* (Y, X) *which is independent of* $\{(Y_t, X_t), 1 \le t \le N\}$, *and identically distributed as* (Y_1, X_1), $E[Y - \hat{F}(X)]^2 \to \sigma^2(1, \ldots, L)$;

(ii) $N^{-1} \sum_{t=1}^{N} [Y_t - \hat{F}(X_t)]^2 \xrightarrow{P} \sigma^2(1, \ldots, L)$,
where $\sigma^2(\cdot)$ *is defined as in* (1.2).

3. Examples

To get insight into the finite-sample behaviour of the cross-validatory selection method, we use two simulated examples and two real data sets as illustrations. In the following examples, we always use the Gaussian kernel. Our experience suggests that the choice of the kernel is much less critical than the choice of the bandwidth. The bandwidth is chosen among 100 values by the cross-validatory approach. It turns out that the data-driven bandwidths satisfy the monotone assumption (K5) to a high degree of approximation, even without any modification. (See Examples 3 and 4 below.) In fact, in each of Examples 3 and 4, minor modification will readily furnish a sequence of bandwidths which satisfies assumption (K5) fully (leading to the same optimal subset), if rigid adherence to (K5) is deemed necessary.

It is generally accepted in non-parametric estimation that the sample size should increase exponentially as the dimensionality of regressors increases. Notice that the convergence in (2.8) is slow for small values of $(k - d)$. Furthermore, for commonly used kernels, e.g. Gaussian and triangular, $\int p_k^2(x)dx$ decreases rapidly as k increases. Therefore, a large N seems necessary to obviate the over-fitting

(cf. (2.9) and (2.7)). However, the following examples show that the proposed method works quite well even for moderate sample sizes despite the asymptotic theory. Other examples and simulation results are available in Cheng and Tong (1992), and Tong (1992). This suggests that there might be some nice features beyond our present understanding of the kernel method in subset selection (or order determination) problems. Although the kernel method suffers from the burden of dimensionality for the estimation of the functional form of F, the sample size requirement for subset selection seems not as severe.

Example 1. We begin with the simple model

$$Y_t = 0.6X_{t-2}^2 + \epsilon_t, \quad t = 1, \ldots, N,$$

where $\{X_t\}$ is an AR(1) process given by $X_t = 0.5X_{t-1} + \eta_t$, and ϵ_t, η_t, $t = 1, 2, \ldots$, are independent random variables with the same distribution as the random variable η, and η is equal to the sum of 48 independent random variables each uniformly distributed on $[-0.25, 0.25]$. According to the central limit theorem, we can treat η as being nearly a standard normal random variable. However it has bounded support, namely $[-12, 12]$. Note that the standard linear methods based on cross-spectral analysis or cross-correlation analysis will fail in estimating the *delay* between the input X and the output Y. Set X_t, X_{t-1}, X_{t-2}, X_{t-3}, Y_{t-1}, and Y_{t-2} as the candidates of the regressors. The cross-validatory subset selection is performed on the simulated data with $N = 200$. Out of 100 replications, $\{X_{t-2}\}$ is selected 97 times as the regressor; the other three choices are $\{X_{t-1}, X_{t-2}\}$, $\{X_{t-2}, X_{t-3}\}$, and $\{X_t, X_{t-2}, X_{t-3}\}$. The above all-subset search took about 100 CPU hours on a SUN4 workstation to produce results for the 100 replications.

Example 2. Let

$$Y_t = 0.3Y_{t-1}e^{X_{t-1}} + \sin X_{t-1} + \epsilon_t, \quad t = 1, \ldots, N,$$

where $\{X_t\}$ is an AR(2) process given by $X_t = 0.1X_{t-1} - 0.56X_{t-2} + \eta_t$, and ϵ_t, η_t, $t = 1, 2, \ldots$, are independent random variables with the same distribution as the random variable 0.6η, and η is the same as in Example 1. Set X_t, X_{t-1}, X_{t-2}, Y_{t-1}, Y_{t-2}, and Y_{t-3} as the candidates of the regressors. Table 1 reports the results of the simulation for $N = 200$, and 500, each with 100 replications. The complete calculation took about 130 CPU hours in the distributed array processor AMT DAP 500.

Example 3. Let $\{Y_t, 1 \leq t \leq N\}$ denote the Canadian lynx data for 1821-1934 (listed in Tong (1990)). Now $N = 114$. Set $Y_{t-1}, \ldots, Y_{t-6}$ as the candidates of the regressors. On applying the cross-validatory subset selection with the Gaussian kernel for Y_t and Y_t^2, the results are respectively reported in Table 2 and Table 3.

Table 1. Frequencies of selected regressors in 100 replications for Example 2

Selected regressors	$N = 200$	$N = 500$
$\{X_{t-1}, Y_{t-1}\}$	80	95
$\{X_{t-1}, Y_{t-1}, X_{t-2}\}$	6	1
$\{X_{t-1}, Y_{t-1}, Y_{t-2}\}$	5	–
$\{X_{t-1}, Y_{t-1}, X_t\}$	2	3
$\{X_{t-1}, Y_{t-1}, X_{t-2}, Y_{t-3}\}$	2	–
$\{X_{t-1}, Y_{t-1}, Y_{t-3}\}$	1	1
$\{X_{t-1}, Y_{t-1}, X_t, Y_{t-3}\}$	1	–
$\{X_{t-1}, Y_{t-1}, X_t, X_{t-2}, Y_{t-3}\}$	1	–
$\{X_{t-1}, X_{t-2}\}$	1	–
$\{X_{t-1}, Y_{t-2}\}$	1	–

In both cases, the global minimum is attained at the subset $\{Y_{t-1}, Y_{t-3}, Y_{t-6}\}$. As expected, there is no evidence of conditional heteroscedasticity in this data set (cf. Tong (1990)). The above calculation took about 35 minutes on a SUN4 workstation.

In the last column of each table, we list the values of $k\lambda(k)$ $(= -k \log h / \log N$, cf. (2.1)), for the selected bandwidth h using a data-driven method. The assumption (K5) is fulfilled except for $k = 4$ in Table 2 and $k = 5$ in Table 3. In fact, if we use $h = 0.326$ instead of 0.361 for the case $k = 4$ in Table 2, (thereby increasing the corresponding CV-value by 0.0010), the CV-selected lag variables are unchanged. The modified results are reported in parentheses. However, the value of $4\lambda(4)$ becomes 0.944. Therefore, now $k\lambda(k)$ is strictly increasing as k increases. Of course, the global minimum is unchanged. The same adaptation can be applied to Table 3.

Table 2. Subset regression of Y_t in Example 3

Typically row 3 reads: amongst all subsets containing three regressors, the minimum CV is attained at the subset $\{Y_{t-1}, Y_{t-3}, Y_{t-6}\}$ with a CV-value = 0.2002 and the bandwidth = 0.256; further, for this bandwidth $3\lambda(3)$ is equal to 0.86.

k	lags	CV-value	bandwidth	$k\lambda(k)$
1	$\{1\}$	0.4136	0.326	0.24
2	$\{1,2\}$	0.2034	0.221	0.64
3	$\{1,3,6\}$	0.2002	0.256	0.86
4	$\{1,2,3,6\}$	0.2099 (0.2109)	0.361 (0.326)	0.86 (0.94)
5	$\{1,2,3,5,6\}$	0.2200	0.384	1.01
6	$\{1,\ldots,6\}$	0.2268	0.407	1.14

Table 3. Subset regression of Y_t^2 in Example 3

Some convention adopted as in Table 2.

k	lags	CV-value	bandwidth	$k\lambda(k)$
1	$\{1\}$	42.83	0.209	0.33
2	$\{1,2\}$	20.32	0.209	0.66
3	$\{1,3,6\}$	19.66	0.244	0.89
4	$\{1,3,5,6\}$	21.34	0.302	1.01
5	$\{1,2,3,5,6\}$	22.57 (22.59)	0.407 (0.361)	0.95 (1.08)
6	$\{1,\ldots,6\}$	23.12	0.407	1.14

Example 4. Finally we illustrate the method with Wolf's annual sunspot numbers (1700-1988) listed in Tong (1990). First, normalize these data by division by the sample standard deviation. For $t = 1, 2, \ldots, 289$, let Y_t = the normalized sunspot number in the year $(1699 + t)$, and let X_{ti} = the normalized sunspot number in the year $(1699 + t - i)$. Set $L = 10$. Table 4 below shows that the global minimum of CV with respect to all possible subsets and the bandwidths of a Gaussian kernel is attained at the subset $\{1, 2, 4\}$ i.e. $\{X_{t-1}, X_{t-2}, X_{t-4}\}$ with a CV-value $= 0.1462$. Note that the subset $\{1, 2, 4, 7\}$ i.e. $\{X_{t-1}, X_{t-2}, X_{t-4}, X_{t-7}\}$ with a CV-value $= 0.1465$ is almost just as optimal. Therefore, we argue that any reasonable 'confidence set' of the optimal subsets must include both of the above subsets. The above optimization over *all* possible subsets and over the bandwidths took about 50 hours on a SUN4 workstation.

Table 4. Subset regression of Y_t in Example 4

Same convention adopted as in Table 2.

k	lags	CV-value	bandwidth	$k\lambda(k)$
1	$\{1\}$	0.3511	0.103	0.40
2	$\{1,3\}$	0.1630	0.154	0.66
3	$\{1,2,4\}$	0.1462	0.180	0.91
4	$\{1,2,4,7\}$	0.1465 (0.1468)	0.283 (0.261)	0.89 (0.95)
5	$\{1,2,5,6,10\}$	0.1492	0.309	1.04
6	$\{1,2,3,4,6,10\}$	0.1558	0.309	1.24
7	$\{1,2,3,4,5,6,10\}$	0.1594	0.335	1.35
8	$\{1,2,3,4,5,6,7,8\}$	0.1803	0.361	1.45
9	$\{1,2,3,4,5,6,7,9,10\}$	0.1994 (0.2021)	0.438 (0.386)	1.31 (1.51)
10	$\{1,\ldots,10\}$	0.2227 (0.2301)	0.541 (0.412)	1.13 (1.57)

Table 5 below shows the results of the CV selection for the case of Y_t^2 on $X_{t1}, \ldots, X_{t10}$. Here the global minimum of CV is attained at the subset $\{1, 2, 5, 7\}$ with a CV-value of 2.7431. The subset $\{1, 2, 4, 7\}$ has a CV-value of 2.7692 (at the bandwidth 0.283), which is the next smallest and is only marginally greater than 2.7431. Again, we argue that any reasonable 'confidence' set of optimal subsets must include these two subsets. Together with the results of Y_t on X_t, and also some plots on the residuals which are not reported here, we would conclude that there is not strong evidence of conditional heteroscedasticity in the sunspot data.

Our choice may be compared with (i) $\{1, 2, 9\}$, the best subset linear autoregressive based on AIC (Subba Rao and Gabr (1984)); (ii) $\{1, 2, 3, 4, 5, 9\}$, the ASTAR model of Lewis and Stevens (1991); and (iii) $\{1, 2, 3, 7, 9\}$ as reported by Tjøstheim in his discussion of Cheng and Tong (1992).

Similar to Example 3, we modified a few bandwidths subjectively to satisfy condition (K5) rigidly. The modified results are reported in parentheses. Obviously, the modifications do not change the overall conclusions.

Table 5. Subset regression of Y_t^2 in Example 4

Same convention adopted as in Table 2.

k	lags	CV-value	bandwidth	$k\lambda(k)$
1	$\{1\}$	5.8179	0.309	0.21
2	$\{1,3\}$	3.3179	0.231	0.26
3	$\{1,3,5\}$	2.8987	0.231	0.39
4	$\{1,2,5,7\}(\{1,2,4,7\})$	2.7431 (2.7692)	0.309 (0.283)	0.83 (0.89)
5	$\{1,2,5,6,10\}$	2.8154	0.335	0.96
6	$\{1,2,4,5,6,10\}$	2.9436	0.361	1.09
7	$\{1,2,3,4,5,6,10\}$	3.0036	0.386	1.18
8	$\{1,2,3,4,5,6,7,9\}$	3.1546	0.361	1.45
9	$\{1,2,3,4,5,6,7,8,9\}$	3.5037 (3.5612)	0.489 (0.401)	1.14 (1.45)
10	$\{1,\ldots,10\}$	3.9869 (4.0291)	0.489 (0.412)	1.26 (1.56)

Appendix A: Proof of Lemma 1

We use the same notation as Section 2.

For $1 \le i_1 < \cdots < i_k \le L$, $1 \le k \le L$, let

$$F_{i_1,\ldots,i_k}(x_1, \ldots, x_k) = E(Y_t \mid X_{ti_1} = x_1, \ldots, X_{ti_k} = x_k).$$

Then the expression (2.5) can be written as follows.

$$\mathrm{CV}(i_1, \ldots, i_k) = \frac{1}{N} \sum_{s=1}^{N} \left[F_{i_1,\ldots,i_k}(X_{si_1}, \ldots, X_{si_k}) - \hat{F}^{(s)}_{i_1,\ldots,i_k}(X_{si_1}, \ldots, X_{si_k}) \right]^2$$

$$+\frac{1}{N}\sum_{s=1}^{N}(\epsilon_s^{(i_1,\ldots,i_k)})^2+\frac{2}{N}\sum_{s=1}^{N}\epsilon_s^{(i_1,\ldots,i_k)}$$
$$\cdot\Big[F_{i_1,\ldots,i_k}(X_{si_1},\ldots,X_{si_k})-\hat{F}^{(s)}_{i_1,\ldots,i_k}(X_{si_1},\ldots,X_{si_k})\Big]. \quad \text{(A.1)}$$

It follows from the standard ergodic theorem that the second term on the right hand side of (A.1) converges to $\sigma^2(i_1,\ldots,i_k)$. From Lemma 2 (i) and (ii) below, the other two terms on the right hand side of (A.1) converges 0. Hence, Lemma 1 (i) holds. Note that under the condition (2.6), the second term on the right hand side of (A.1) is equal to σ_N^2. Consequently, Lemma 1 (ii) follows immediately from Lemma 2 (iii) and (iv) below. The proof of Lemma 1 is now complete.

Lemma 2. *Suppose that conditions* (M2)–(M7) *and* (K1)–(K4) *hold. Then, for any* $1\le i_1<\cdots<i_k\le L, 1\le k\le L$, *as* $N\to\infty$,

(i) $\frac{1}{N}\sum_{s=1}^{N}\epsilon_s^{(i_1,\ldots,i_k)}[F_{i_1,\ldots,i_k}(X_{si_1},\ldots,X_{si_k})-\hat{F}^{(s)}_{i_1,\ldots,i_k}(X_{si_1},\ldots,X_{si_k})]\xrightarrow{P}0$;

(ii) $\frac{1}{N}\sum_{s=1}^{N}[F_{i_1,\ldots,i_k}(X_{si_1},\ldots,X_{si_k})-\hat{F}^{(s)}_{i_1,\ldots,i_k}(X_{si_1},\ldots,X_{si_k})]^2\xrightarrow{P}0$.

Furthermore, if the relation (2.6) *holds, then*

(iii) $\frac{1}{N}\sum_{s=1}^{N}\epsilon_s^{(i_1,\ldots,i_k)}[F_{i_1,\ldots,i_k}(X_{si_1},\ldots,X_{si_k})-\hat{F}^{(s)}_{i_1,\ldots,i_k}(X_{si_1},\ldots,X_{si_k})]$
$=o_p([Nh^k(k,N)]^{-1})$;

(iv) $\frac{1}{N}\sum_{s=1}^{N}[F_{i_1,\ldots,i_k}(X_{si_1},\ldots,X_{si_k})-\hat{F}^{(s)}_{i_1,\ldots,i_k}(X_{si_1},\ldots,X_{si_k})]^2$
$=\gamma_{i_1,\ldots,i_k}[Nh^k(k,N)]^{-1}(1+o_p(1))$,

where $\gamma_{i_1,\ldots,i_k}$ *is the same as* (2.7).

To prove Lemma 2, we adopt the technique which was used in §3 and §4 of Cheng and Tong (1993). To proceed, we need the following Lemma 3 which follows Theorem 3.1 of Roussas (1988) immediately.

Lemma 3. *Suppose that conditions* (M2), (M3), (M7) *and* (K1)–(K3) *hold. Then for any* $1\le i_1<\cdots<i_k\le L$, $1\le k\le L$, *as* $N\to\infty$,

$$\sup|\hat{f}_{i_1,\ldots,i_k}(x)-f_{i_1,\ldots,i_k}(x)|\to 0,$$

where $f_{i_1,\ldots,i_k}$ *is the marginal probability density function of* $(X_{ti_1},\ldots,X_{ti_k})$, $\hat{f}_{i_1,\ldots,i_k}$ *is as given in* (2.2), *and the supremum is taken over all values of* x *such that* $f_{i_1,\ldots,i_k}(x)>0$.

Proof of Lemma 2. First we prove (i).

For given $1\le i_1<\cdots<i_k\le L$, let $Z_s=(X_{si_1},\ldots,X_{si_k}), s=1,\ldots,N$. It follows from Lemma 3 that

$$F_{i_1,\ldots,i_k}(Z_s)-\hat{F}^{(s)}_{i_1,\ldots,i_k}(Z_s)=\Big[F_{i_1,\ldots,i_k}(Z_s)-\hat{F}^{(s)}_{i_1,\ldots,i_k}(Z_s)\Big]\frac{\hat{f}_{i_1,\ldots,i_k}(Z_s)}{f_{i_1,\ldots,i_k}(Z_s)}(1+o_p(1))\ \text{a.s.},$$

where the term $o_p(1)$ does not depend on s. Note that the relations (2.2)–(2.4) imply that

$$\begin{aligned}&\Big[F_{i_1,\ldots,i_k}(Z_s) - \hat{F}^{(s)}_{i_1,\ldots,i_k}(Z_s)\Big]\hat{f}_{i_1,\ldots,i_k}(Z_s)\\ =\ & \frac{h^{-k}}{N}\left(\sum_{t=1}^{N} C_{s,t} - \sum_{t\neq s}\epsilon_t^{(i_1,\ldots,i_k)} d_{s,t} + F_{i_1,\ldots,i_k}(Z_s)p_k(0)\right),\end{aligned}$$

where

$$d_{s,t} = p_k\left(\frac{Z_s - Z_t}{h}\right), \qquad C_{s,t} = d_{s,t}\Big[F_{i_1,\ldots,i_k}(Z_s) - F_{i_1,\ldots,i_k}(Z_t)\Big].$$

Hence, the sum in (i) can be expressed as follows

$$\begin{aligned}&N^{-2}h^{-k}\Bigg\{\sum_{s=1}^{N}\sum_{t\neq s}\epsilon_s^{(i_1,\ldots,i_k)}\Big[C_{s,t} - \epsilon_t^{(i_1,\ldots,i_k)} d_{s,t}\Big] f^{-1}_{i_1,\ldots,i_k}(Z_s)\\ &+ p_k(0)\sum_{s=1}^{N}\epsilon_s^{(i_1,\ldots,i_k)} F_{i_1,\ldots,i_k}(Z_s) f^{-1}_{i_1,\ldots,i_k}(Z_s)\Bigg\}(1 + o_p(1))\\ =\ & R_1 + R_2 + o_p(R), \quad \text{say},\end{aligned}$$

where $R = R_1 + R_2$. By the standard ergodic theorem, $R_2 \xrightarrow{P} 0$. To calculate R_1, we write $e_t = (Z_t, \epsilon_t^{(i_1,\ldots,i_k)})$. Moreover, define

$$\begin{aligned}H(e_t, e_s) \ =\ & \epsilon_s^{(i_1,\ldots,i_k)}\Big[C_{s,t} - \epsilon_t^{(i_1,\ldots,i_k)} d_{s,t}\Big] f^{-1}_{i_1,\ldots,i_k}(Z_s)\\ &+ \epsilon_t^{(i_1,\ldots,i_k)}\Big[C_{t,s} - \epsilon_s^{(i_1,\ldots,i_k)} d_{t,s}\Big] f^{-1}_{i_1,\ldots,i_k}(Z_t),\end{aligned}$$

and

$$H(e_t) = \int H(e_t, e_s)dP(e_s) = \epsilon_t^{(i_1,\ldots,i_k)} f^{-1}_{i_1,\ldots,i_k}(Z_t)\int C_{t,s}dP(Z_s),$$

where $P(\xi)$ denotes the probability distribution of the random vector ξ. It is easy to see that R_1 can be expressed as follows

$$R_1 = \frac{1}{2N^2h^k}\sum_{t\neq s}[H(e_t, e_s) - H(e_t) - H(e_s)] + \frac{N-1}{N^2h^k}\sum_{t=1}^{N} H(e_t), \tag{A.2}$$

and the first term on the right hand side is symmetric in t and s, which can be interpreted as the remainder in Hoeffding's projection decomposition of the U-statistic generated by $H(e_t, e_s)$ since $\int H(e_t)dP(e_t) = 0$ (cf. Yoshihara (1976), (2.11)). With assumption (M2), the following inequality holds

$$|H(e_t, e_s)| \le c_1\left(|\epsilon_s^{(i_1,\ldots,i_k)}| + |\epsilon_t^{(i_1,\ldots,i_k)}|\right) + c_2|\epsilon_s^{(i_1,\ldots,i_k)}|\cdot|\epsilon_t^{(i_1,\ldots,i_k)}| \quad \text{a.s.},$$

where c_1, c_2 are some positive constants. Hence, by assumption (M5),

$$\int |H(e_t, e_s)|^3 dP(e_t) dP(e_s) < \infty.$$

Since under assumption (M7), $\{e_t, t \geq 1\}$ is an absolutely regular process, it follows from Lemma 1 of Yoshihara (1976) that $\sup_{s<t} E|H(e_t, e_s)|^3 < \infty$, which, together with Lemma 2 of Yoshihara (1976), implies that

$$E\left\{\frac{1}{N^2}\sum_{t\neq s}[H(e_t, e_s) - H(e_t) - H(e_s)]\right\}^2 = o(N^{-2}).$$

Consequently,

$$\frac{1}{2N^2h^k}\sum_{t\neq s}[H(e_t, e_s) - H(e_t) - H(e_s)] = o_p(N^{-1}h^{-k}). \tag{A.3}$$

On the other hand, it is easy to show that under assumption (M6),

$$\left|F_{i_1,\ldots,i_k}(x_1) - F_{i_1,\ldots,i_k}(x_2)\right| \leq c\|x_1 - x_2\|^\mu \quad \text{for all} \quad x_1, x_2 \in R^k. \tag{A.4}$$

Since $|H(e_t)| \leq ch^{k+\mu}|\epsilon_t^{(i_1,\ldots,i_k)}|$ a.s., it follows from the standard ergodic theorem,

$$\left|\frac{N-1}{N^2h^k}\sum_{t=1}^{N} H(e_t)\right| \leq ch^\mu \frac{1}{N}\sum_{t=1}^{N}\left|\epsilon_t^{(i_1,\ldots,i_k)}\right| \to 0, \quad \text{a.s.}$$

which, together with (A.3), implies that R_1 tends to 0 in probability. The proof of (i) is now complete.

The proof of (ii) is more detailed and tedious than the above proof of (i) but involves no fundamentally new idea. It is therefore omitted.

To prove (iii), note the fact that (2.6) and (M4) imply that for any $s \neq t$, $E[H(e_t)H(e_s)] = 0$. Consequently, by assumption (M2) and inequality (A.4),

$$E\left[\frac{1}{N}\sum_{t=1}^{N} H(e_t)\right]^2 = \frac{1}{N}EH^2(e_t) \leq ch^{4k+2\mu}/N.$$

Hence, by assumption (K4),

$$\frac{N-1}{N^2h^k}\sum_{t=1}^{N} H(e_t) = O_p(h^{k+\mu}/\sqrt{N}) = o_p(N^{-1}h^{-k}).$$

With inequality (A.4), we can show that $R_{12} = o_p(N^{-1}h^{-k})$ in the same way. Together with (A.3), we have now proved (iii).

Similarly, we can show that the sum in (iv) is equal to

$$h^{-2k}N^{-3}\sum_{\substack{s,t\\ s\neq t}}\epsilon_t^2 d_{s,t}^2 f_{i_1,\ldots,i_k}^{-2}(Z_s)+o_p(N^{-1}h^{-k}).$$

On writing the sum in above expression as a U-statistic by symmetrization, and performing the Hoeffding's projection decomposition on the U-statistic (cf. (A.2)), it follows from Lemma 2 of Yoshihara (1976) that the primary term in the decomposition is the integral

$$N^{-1}h^{-2k}\int \epsilon_t^2 d_{s,t}^2 f_{i_1,\ldots,i_k}^{-2}(Z_s)dP(e_s)dP(e_t),$$

where (e_s, e_t) is the same as in (A.2). The proof is completed by the fact that the above expression is asymptotically equivalent to $\gamma_{i_1,\ldots,i_k}/(Nh^k)$.

Appendix B: The proof of Theorem 2

For any $1\le i_1<\cdots<i_k\le L$, $1\le k\le L$, and $x=(x_1,\ldots,x_L)\in R^L$, define

$$\begin{aligned}\hat{F}_{i_1,\ldots,i_k}(x) &= N^{-1}[h(k,N)]^{-k}\sum_{t=1}^{N}Y_t p_k\left(\frac{x_{i_1}-X_{ti_1}}{h(k,N)},\ldots,\frac{x_{i_k}-X_{ti_k}}{h(k,N)}\right)\\ &\quad\times\hat{f}_{i_1,\ldots,i_k}(x_{i_1},\ldots,x_{i_k}).\end{aligned}$$

To prove Theorem 2, we need the following lemma, which shows that $\hat{F}_{1,\ldots,d}$ converges to F uniformly on $G=\{f>0\}$. Similar results have been proved by Györfi et al. (1989) under different assumptions.

Lemma 4. *Suppose that the assumptions* (M1)–(M7) *hold, and also the assumptions* (K1)–(K4) *hold specifically only for* $k=d$. *Then as* $N\to\infty$, *the following limits hold if either* $E|Y_t|^k<\infty$ *for any* $k\ge 1$, *or* $d\lambda(d)<2/5$.

$$\sup_{x\in G}\left|F(x)-\hat{F}_{1,\ldots,d}(x)\right|\xrightarrow{P}0.$$

Proof. It follows from Lemma 2 that for any $x\in G$,

$$\begin{aligned}\hat{F}_{1,\ldots,d}(x) &= N^{-1}h^{-d}f_{1,\ldots,d}^{-1}(x_1,\ldots,x_d)\\ &\quad\cdot\sum_{t=1}^{N}Y_t p_d\left(\frac{x_1-X_{t1}}{h},\ldots,\frac{x_d-X_{td}}{h}\right)(1+o_p(1)),\end{aligned}\tag{B.1}$$

where the term $o_p(1)$ does not depend on x. Thus

$$\sup_{x\in G}E\{\hat{F}_{1,\ldots,d}(x)\}$$

$$
\begin{aligned}
&= \sup_{x\in G} h^{-d} f_{1,\ldots,d}^{-1}(x_1,\ldots,x_d) E\Big\{ E(Y_t|X_{t1},\ldots,X_{td}) \\
&\quad \times p_d\Big(\frac{x_1 - X_{t1}}{h},\ldots,\frac{x_d - X_{td}}{h}\Big)\Big\} + o(1) \\
&= \sup_{x\in G} f_{1,\ldots,d}^{-1}(x_1,\ldots,x_d) \int E(Y_t \mid X_{t1} = x_1 - hz_1,\ldots,X_{td} = x_d - hz_d) \\
&\quad \times p_d(z_1,\ldots,z_d) f_{1,\ldots,d}(x_1 - hz_1,\ldots,x_d - hz_d) dz_1 \cdots dz_d + o(1) \\
&\to \sup_{x\in G} E(Y_t \mid X_{t1} = x_1,\ldots,X_{td} = x_d) = \sup_{x\in G} F(x). \qquad \text{(B.2)}
\end{aligned}
$$

The above limit follows from the assumptions (M3) and (M6), and the last equality follows from (1.4).

For $M = M(N) > 0$, which will be specified later, we define

$$
\begin{aligned}
R_1(x) &= N^{-1}h^{-d}\sum_{t=1}^{N} Y_t\, p_d\Big(\frac{x_1 - X_{t1}}{h},\ldots,\frac{x_d - X_{td}}{h}\Big) I_{\{Y_t \le M\}}, \\
R_2(x) &= N^{-1}h^{-d}\sum_{t=1}^{N} Y_t\, p_d\Big(\frac{x_1 - X_{t1}}{h},\ldots,\frac{x_d - X_{td}}{h}\Big) I_{\{Y_t \ge M\}}.
\end{aligned}
$$

By Schwarz' inequality and Chebyshev's inequality, we have

$$
P\Big(\sup_{x\in G} |R_2(x) - ER_2(x)| > \epsilon\Big) \le c\epsilon^{-1}h^{-1}M^{-k/2} \qquad \text{(B.3)}
$$

for any $\epsilon > 0$ and $k \ge 1$ such that $E|Y_t|^k < \infty$.

It follows from Theorem 3.1 of Roussas and Ioannides (1988) that

$$
P(|R_1(x) - ER_1(x)| > \epsilon) \le c_1 \exp\Big(-c_2\epsilon^2 N h^{-2d}/M^2\Big)
$$

for some positive constants c_1 and c_2 independent of x. This leads to

$$
\sup_{x\in G} P(|R_1(x) - ER_1(x)| > \epsilon) \le c_1 \exp\Big(-c_2\epsilon^2 N h^{-2d}/M^2\Big).
$$

Using the same arguments as in p.30-31 of Györfi et al. (1989), it can be shown that for any $\epsilon > 0$, the following inequality holds for all N greater than an integer N_ϵ.

$$
P\Big(\sup_{x\in G} |R_1(x) - ER_1(x)| > \epsilon\Big) \le ch^{-p} \exp\Big(-c_2\epsilon^2 N h^{-2d}/M^2\Big), \qquad \text{(B.4)}
$$

where $p > 0$ is an integer, which is independent of N.

Let $M = N^{-\xi}$ with $\xi < 0.5 - d\lambda(d)$. Then the right hand side of (B.4) tends to 0 as $N \to \infty$. In case $E|Y_t|^k < \infty$ for any $k \ge 1$, the right hand side of (B.3) tends

to 0 for sufficiently large k. In case $d\lambda(d) < 2/5$, the limit still holds by choosing $\xi > d\lambda(d)/4$. Therefore, we have proved that for $i = 1, 2, |R_i(x) - ER_i(x)|$ converges to 0 in probability. It follows from (B.1) and the assumption (M2) that

$$\begin{aligned} &\sup_{x\in G}\left|\hat{F}_{1,\ldots,d}(x) - E\hat{F}_{1,\ldots,d}(x)\right| \\ \leq\ & c\Big(\sup_{x\in G}|R_1(x) - ER_1(x)| + \sup_{x\in G}|R_2(x) - ER_2(x)|\Big) + o_p(1) \xrightarrow{P} 0. \quad \text{(B.5)} \end{aligned}$$

The lemma follows immediately from (B.2) and (B.5).

Proof of Theorem 2. Since X_t is bounded, it is easy to show that for any $1 \leq i_1 < \cdots < i_k \leq L, 1 \leq k \leq L$, and any compact set $B \subset R^L$

$$\sup_{x\in B}|\hat{F}_{i_1,\ldots,i_k}(x)|^2 \leq c\left(\frac{1}{N}\sum_{t=1}^{N}|Y_t|\right)^2 \leq \frac{c}{N}\sum_{t=1}^{N}Y_t^2.$$

It follows from assumption (M5) that $\{Y_t^2, t \geq 1\}$ and hence also $\{N^{-1}\sum_{t=1}^{N}Y_t^2, N \geq 1\}$ are uniformly integrable. Thus, $\{\sup_{x\in B}|\hat{F}_{i_1,\ldots,i_k}(x)|^2, N \geq 1\}$ is uniformly integrable. Take $B = G$, which is compact in R^L by assumption (M2). Then,

$$E[\hat{F}(X)]^2 \leq E\left[\sup_{x\in G}\hat{F}(x)\right]^2 \leq c\sum E\left[\sup_{x\in G}\hat{F}_{i_1,\ldots,i_k}(x)I_{(S_{cv}=\{i_1,\ldots,i_k\})}\right]^2 < \infty, \quad \text{(B.6)}$$

where the sum is taken over all $1 \leq i_1 < \cdots < i_k \leq L$, $1 \leq k \leq L$, and I_A denotes the indicator function of set A.

It follows from Theorem 1 that for any $\epsilon > 0$, there exists a positive integer N_ϵ such that when $N \geq N_\epsilon$, S_{cv} coincides with $\{1, \ldots, d\}$ on a subset with probability greater than $1 - \epsilon$. Hence, by (B.6)

$$E[F(X) - \hat{F}(X)]^2 \leq c\epsilon + E[F(X) - \hat{F}_{1,\ldots,d}(X)]^2. \quad \text{(B.7)}$$

It follows from Lemma 4 that $\sup_{x\in G}|F(x) - \hat{F}_{1,\ldots,d}(x)| \xrightarrow{P} 0$. By the standard mean convergence theorem, $E[F(X) - \hat{F}_{1,\ldots,d}(X)]^2 \to 0$. Together with (B.7), this gives

$$E[F(X) - \hat{F}(X)]^2 \to 0.$$

Consequently, by the Schwarz inequality,

$$|E\{[Y - F(X)][F(X) - \hat{F}(X)]\}| \leq \{EY^2 \cdot E[F(X) - \hat{F}(X)]^2\}^{\frac{1}{2}} \to 0.$$

Finally, we have

$$\begin{aligned} E[Y - \hat{F}(X)]^2 &= E[Y - F(X)]^2 + E[F(X) - \hat{F}(X)]^2 \\ &\quad + 2E\{[Y - F(X)][F(X) - \hat{F}(X)]\} \\ &\rightarrow E[Y - F(X)]^2 = \sigma^2(1, \ldots, L), \end{aligned}$$

which competes the proof of (i).

(ii) can be proved in a similar but simpler way, which is omitted here.

Acknowledgement

The authors wish to thank two referees for their helpful comments. Thanks also go to Ms. J. Leng for her computational assistance. This research was partially supported by the Science and Engineering Research Council (U.K.).

References

Auestad, B. and Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika* **77**, 669–687.

Cheng, B. and Tong, H. (1992). On consistent nonparametric order determination and chaos (with discussion). *J. Roy. Statist. Soc. Ser.B* **54**, 427–474.

Cheng, B. and Tong, H. (1993). On residual sums of squares in non-parametric autoregression. *Stochastic Process. Appl.* **48**, 154–174.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989). *Non-parametric Curve Estimation from Time Series.* Springer-Verlag, Berlin.

Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10**, 154–166.

Lewis, P. A. W. and Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). *J. Amer. Statist. Assoc.* **86**, 864–877.

Masry, E. (1989). Nonparametric estimation of conditional probability densities and expectations of stationary processes: Strong consistency and rates. *Stochastic Process. Appl.* **32**, 109–127.

Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. *Theory Probab. Appl.* **10**, 186–190.

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

Roussas, G. G. and Ioannides, D. (1988). Probability bounds for sums in triangular arrays of random variables under mixing conditions. *Statistical Theory and Data Analysis II* (Edited by K. Matusita), 293–308, North Holland.

Subba Rao, T. and Gabr, M. M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models. Lecture Notes in Statist.* **24**, Springer-Verlag, New York.

Tjøstheim, D. and Auestad, B. H. (1992a). Non-parametric identification of time series: Projection. Technical Report, Dept. of Math., Univ. of Bergen, 5007 Bergen, Norway.

Tjøstheim, D. and Auestad, B. H. (1992b). Non-parametric identification of time series: Selecting significant lags. Technical Report, Dept. of Math., Univ. of Bergen, 5007 Bergen, Norway.

Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach.* Oxford University Press.

Tong, H. (1992). Akaike's approach can yield consistent order determination. *The First US/ Japan Conference on The Frontiers of Statistical Modeling: An Informational Approach*, Kluwer Academic Publishers. (To appear.)

Truong, Y. K. and Stone, C. (1990). Non-parametric function estimation involving time series. Preprint, University of North Carolina, Chapel Hill, U.S.A.

Watson, G. S. (1964). Smooth regression analysis. *Sankyā Ser.A* **26**, 359–372.

Yoshihara, K.-I. (1976). Limiting behavior of U-statistics for stationary, absolutely regular process. *Z. Wahrsch. verw. Gebiete* **35**, 237–252.

Zhang, P. (1991). Variable selection in nonparametric regression with continuous covariates. *Ann. Statist.* **19**, 1869–1882.

Department of Mathematics, Southeast University, Nanjing 210018.
Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent CT2 7NF, U.K.

(Received June 1992; accepted September 1993)