ROBUST REGRESSION FUNCTION ESTIMATION FROM DEPENDENT OBSERVATIONS

by

Young K. Truong

Department of Biostatistics

University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1834

August 1987

ROBUST REGRESSION FUNCTION ESTIMATION FROM DEPENDENT OBSERVATIONS

By YOUNG K. TRUONG
Department of Biostatistics
University of North Carolina, Chapel Hill

Abstract. The local and global asymptotic properties of kernel methods based on local M-estimators are considered for stationary time series which satisfies weak mixing conditions. For this class of nonparametric estimators, the results presented in this paper constitute a generalization of a result in Härdle and Luckhaus (1984, *Ann. Statist.* **12** 621–623) to time series.

Keywords. Kernel, M-estimators, nonparametric regression, rates of convergence, cumulant, mixing.

1. INTRODUCTION

Let $(\mathbf{X}, Y)$ be a pair of random variables that are respectively d and 1 dimensional; the random variable Y is called the response and the random vector $\mathbf{X}$ is refered to as the predictor variable. The main objective is to construct a function $\theta(\cdot)$ so as to (i) study the relationship between the response and the explanatory variable or (ii) obtain the predictor $\theta(\mathbf{X})$ of Y based on $\mathbf{X}$.

The simplest and most widely used measure of accuracy of $\theta(\mathbf{X})$ as a predictor of Y is the *Mean Square Error*, $E|Y - \theta(\mathbf{X})|^2$. The function $\theta(\cdot)$ which minimizes this measure of accuracy is the regression function of Y on $\mathbf{X}$, defined by $\theta(\mathbf{X}) = E(Y|\mathbf{X})$.

Alternately, if the *Mean Absolute Deviation* $E|Y - \theta(\mathbf{X})|$ is adopted as a measure of accuracy, especially when outliers may be present (Bloomfield and Steiger, 1983); then the optimal function $\theta(\cdot)$ is now defined so that $\theta(\mathbf{X})$ is the conditional median, Median$(Y|\mathbf{X})$, of Y given $\mathbf{X}$. Note that this function is not necessary uniquely defined.

Recently, there has been an increasing interest in adopting $E\rho(Y, \theta(\mathbf{X}))$ as a measure of accuracy, where $\rho(x, y) = \rho(x - y)$ is a positive function defined on $\mathbf{R}^2$. The optimal solution is called the conditional M-predictor. If $\rho(u) = u^2$, then the optimal M-predictor is the conditional mean; while if $\rho(u) = |u|$, it is the conditional median defined above. For other example of the function ρ, see Huber (1981) and Hampel et al. (1986).

In practice, it is necessary to construct estimators of these optimal functions based on a set of observations. Time series prediction is the generic term revolving around the construction of estimators of these predictors based on a realization $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ from the stationary process $(\mathbf{X}_t, Y_t), t = 0, \pm 1, \ldots$.

Parametric Approach vs Nonparametric Approach

To estimate these predictors, the *parametric* approach starts with specific assumptions about the relationship between the response (or future) and the explanatory variables (or past) and about the variation in the response (future) that may or may not be accounted for by the explanatory variables. For instance, the standard regression method (or autore-

gressive method in time series) starts with an a priori model for the regression function $\theta(\cdot)$ which, by assumption or prior knowledge, is a linear function that contains finitely many unknown parameters. Under the assumption that the joint distribution is Gaussian, it is an optimal prediction rule; if the distribution is non-Gaussian, it is not generally posible to determine the funtion $\theta(\cdot)$; so one might settle for the *best* linear predctor. By constrast, in the *nonparametric* approach, the regression function will be estimated directly without assuming such an a priori model for $\theta(\cdot)$. As pointed out in Stone (1985), the nonparametric approach is more *flexible* than the standard regression method; *flexibility* means the ability of the model to provide accurate fits in a wide variety of realistic situations, inaccuracy here leading to *bias* in estimation. In recent years, nonparametric estimation has become an active area in time series analysis because of its flexibility in fitting data (Bierens, 1983; Collomb, 1984; Collomb and Härdle, 1984; Robinson, 1983).

The present approach deals with the asymptotic properties (in terms of rates of convergence) of a class of nonparametric estimators constructed by kernel methods based on local M-estimates. It is hoped that the results obtained here serve as a starting point for further development and understanding of the sampling properties of more complicated nonparametric procedures involving robustification, local polynomial fits, additive regression, and spline approximation.

Some previous work on nonparametric estimation in time series will be surveyed in the next section.

2. DEVELOPMENTS IN TIME SERIES PREDICTION

The theory and practice of linear model fitting has now attained a refined state; see, for example, Brillinger (1981), Hannan (1970), Hannan (1973), Dunsmuir and Hannan (1976) and Priestley, 1979). While the study of non-linear models in time series is still in its early stages, what has been learned so far is sufficient to indicate that this is a very rich and potentially rewarding field. Analysis of particular series have shown that non-linear models can provide better fits to the data (as one would expect) and, more importantly, that the structure underlying the data can not be captured by linear models.

So far, the study of non-linear models has been restricted to a few specific forms. For example, Priestley (1980), Tong and Lim (1980), Nicholls and Quinn (1980), and Haggan and Ozaki (1980, 1981) consider various non-linear filters of, possibly independent, identically distributed Gaussian random variables. In practice it may be difficult to decide a priori, which, if any, of these models is best suited to a given set of data.

Asymptotic results for the conditional expectation has been established by Doukhan and Ghindes (1980), Collomb (1984), Bierens (1983) and Robinson (1983) under various mixing conditions. In Robinson (1983), pointwise consistency and a central limit theorem was obtained for kernel estimators based on local averages under the α-mixing condition. Collomb (1984) and Bierens (1983) considered the uniform consistency and rate of convergence for kernel estimators based on local averages under the ϕ-mixing condition, which is considerably stronger than the α-mixing condition. Collomb and Härdle (1984) considered the uniform rate of convergence (also under ϕ-mixing) for a class of robust nonparametric estimators that did not include local medians. Truong and Stone (1987a) considered the local and global properties of kernel estimators based on local averages and local medians under α-mixing condition; the L^∞ rate of convergence remains unsolved. Under another type of mixing condition defined in terms of cumulants, Truong (1987) considered both local and global rate of convergence (including L^∞) for kernel estimators based on local averages.

In the random sample case, Härdle and Luckhaus (1984) considered the L^∞ rate of convergence for a class of robust nonparametric estimators including an estimator of the conditional median. While Truong and Stone (1987b) considered the L^q $(1 \leq q \leq \infty)$ rates of convergence for estimators of the conditional median in the context of time series. The present approach deals with, for stationay time series, the local and global asymptotic properties of the kernel methods based on local M-estimates. Results are given in Section 3. For this class of nonparametric estimators, the results presented there constitute an answer and an extension to one of the open questions of Stone (1982). Proofs of these results are given in Section 5.

3. NONPARAMETRIC TIME SERIES PREDICTION

Results on the local and global rates of convergence of nonparametric estimators of conditional M-predictor based on a realization of a discrete time stationary time series will be treated in this section. Recall that d is the dimensionality of the explainatory variable $\mathbf{X}$ and let U denote a nonempty bounded open neighborhood of the origin of $\mathbf{R}^d$. Let $\{(\mathbf{X}_i, Y_i), i = 0, \pm 1, \ldots\}$ be an $(d+1)$ vector-valued strictly stationary series and let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ denote a realization of this process. Let $\rho(\cdot)$ denote a positive, even, convex and differentiable function with derivative $\rho'(\cdot) = \psi(\cdot)$. Let $\theta(\mathbf{x})$ denote the zero of the function $t \mapsto E[\,\psi(Y_0 - t)\,|\,\mathbf{X}_0 = \mathbf{x}\,]$.

Assumption 1. *There is a positve constant M_0 such that*

$$|\theta(\mathbf{x}) - \theta(\mathbf{x}')| \leq M_0 \|\mathbf{x} - \mathbf{x}'\| \qquad \text{for } \mathbf{x}, \mathbf{x}' \in U,$$

where $\|\mathbf{x}\| = (x_1^2 + \cdots + x_d^2)^{1/2}$ for $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbf{R}^d$.

Assumption 2. *The distribution of $\mathbf{X}_0$ is absolutely continuous and its density $f(\cdot)$ is bounded away from zero and infinity. That is, there is a positive constant M_1 such that $M_1^{-1} \leq f(\mathbf{x}) \leq M_1$ for $\mathbf{x} \in U$.*

The following technical condition is required for bounding the variance of various terms in the proof.

Assumption 3. *The conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_0$ is absolutely continuous and its density $h(\cdot\,|\mathbf{x})$ is bounded away from zero and infinity. That is $M_1^{-1} \leq h(\mathbf{y}|\mathbf{x}) \leq M_1$ for $\mathbf{x}$ and $\mathbf{y} \in U$.*

Let $(Z_1, \ldots, Z_k)$ denote a vector of k random variables with $E|Z_j|^k < \infty$, for $j = 1, \ldots, k$. The rth order joint cumulant, $\mathrm{cum}(Z_1, \ldots, Z_k)$, of $(Z_1, \ldots, Z_k)$ is given by

$$\mathrm{cum}(Z_1, \ldots, Z_k) = \sum (-1)^{p-1}(p-1)! E\left(\prod_{j \in \nu_1} Z_j\right) \cdots E\left(\prod_{j \in \nu_p} Z_j\right)$$

where the sum extends over all partitions $(\nu_1, \ldots, \nu_p)$, $p = 1, 2, \ldots, k$, of $(1, \ldots, k)$. Equivalently (Brillinger, 1981), $\mathrm{cum}(Z_1, \ldots, Z_k)$ is given by the coefficient of $i^k t_1 \cdots t_k$ in the Taylor series expansion of $\log E[\exp(i \sum_1^k Z_j t_j)]$ about the origin. An important special case of this definition occurs when $Z_j = Z, j = 1, \ldots, k$. This then gives the cumulant of order k of a univriate random variable. See Brillinger (1981) for methods on computing cumulants.

A weak dependence condition on the stationary sequence will now be described. Let $f(\cdot)$ denote a real-valued, measurable functions on $\mathbf{R}^{d+1}$. Set

$$c(i,j) = \sup_{|f| \leq 1} \mathrm{cum}\{f(\mathbf{X}_i, Y_i), f(\mathbf{X}_j, Y_j)\}.$$

and

$$c(t_1, t_2, \ldots, t_k) = \sup_{|f| \leq 1} \mathrm{cum}\{f(\mathbf{X}_{t_1}, Y_{t_1}), f(\mathbf{X}_{t_2}, Y_{t_2}), \ldots, f(\mathbf{X}_{t_k}, Y_{t_k})\}.$$

If the series $(\mathbf{X}_t, Y_t)$, $t = 0, \pm 1, \ldots$ is strictly stationary, then

$$c(t_1, t_2, \ldots, t_k) = c(t_1 + u, t_2 + u, \ldots, t_k + u).$$

for $t_1, \ldots, t_k, u = 0, \pm 1, \ldots$. In this case, we sometimes use the asymmetric notation

$$c(t_1, t_2, \ldots, t_{k-1}) = c(t_1, t_2, \ldots, t_{k-1}, 0).$$

The stationary sequence is said to be mixing if $c(t_1, \ldots, t_k) \to 0$ as $t_1, \ldots, t_k \to \infty$. See Brillinger (1981).

Assumption 4. $\sum_{j \geq N} c(j) = O(N^{-1})$.

The following condition is similar to Assumption 2.6.3 of Brillinger (1981).

Assumption 5. $\sum_k C_k z^k / k! < \infty$ *for* z *in a neighborhood of zero, where*

$$C_k = \sum_{v_1 = -\infty}^{\infty} \cdots \sum_{v_{k-1} = -\infty}^{\infty} |c(v_1, \ldots, v_{k-1})|.$$

A condition on the function $\psi(\cdot)$ is required to guarantee the uniqueness of the conditional M-predictor (uniqueness will ensure consistency) and also the achievability of the desired rate of conververgence. (The same condition is required in order to obtain the usual asymptotic result about the M-estimate in the univariate case.)

Assumption 6. *The function $\psi(\cdot)$ is bounded and increasing.*

For the motivation of the following condition, see Härdle and Luckhaus (1984).

Assumption 7. $|E[\psi(Y - \theta(\mathbf{x}) + t) \,|\, \mathbf{X} = \mathbf{x}]| > M_1|t|$ *for* $|t| < M_1^{-1}$.

The kernel estimators of conditional M-predictor will now be described. For each $n \geq 1$, let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a realization of the (strictly) stationary time series. Let δ_n, $n \geq 1$, be positive numbers that tend to zero as n tends to infinity. Set $I_n(\mathbf{x}) = \{i : 1 \leq i \leq n \text{ and } \|\mathbf{X}_i - \mathbf{x}\| \leq \delta_n\}$ and $N_n(\mathbf{x}) = \#(I_n(\mathbf{x}))$. Let $\hat{\theta}_n(\mathbf{x})$ denote the zero of $t \mapsto N_n^{-1} \sum_{I_n} \psi(Y_i - t)$.

Given positive numbers a_n and b_n, $n \geq 1$, let $a_n \sim b_n$ mean that a_n/b_n is bounded away from zero and infinity. Given random variables V_n, $n \geq 1$, let $V_n = O_{pr}(b_n)$ mean that the random variables $b_n^{-1}V_n$, $n \geq 1$ are bounded in probability or, equivalently, that

$$\lim_{c \to \infty} \limsup_n P(|V_n| > cb_n) = 0.$$

Set $r = (2 + d)^{-1}$.

Theorem 1. *Suppose that Assumptions 1–4, 6 and 7 hold and that* $\delta_n \sim n^{-r}$. *Then*

$$|\hat{\theta}_n(\mathbf{0}) - \theta(\mathbf{0})| = O_{pr}(n^{-r}).$$

The proof of this theorem, which will be given in Section 5, is basically a refinement of the corresponding one given in Truong and Stone (1987b).

Let C be a fixed compact subset of U having a nonempty interior and let $g(\cdot)$ be a real-valued function on $\mathbf{R}^d$. Set

$$\|g\|_q = \left\{ \int_C |g(\mathbf{x})|^q \, d\mathbf{x} \right\}^{\frac{1}{q}}, \qquad 1 \leq q < \infty;$$
$$\|g\|_\infty = \sup_{\mathbf{x} \in C} |g(\mathbf{x})|.$$

Theorem 2. *Suppose that Assumptions 1–7 hold and that* $\delta_n \sim (n^{-1} \log n)^r$. *Then there exists a* $c > 0$ *such that*

$$\lim_n P\left(\|\hat{\theta}_n - \theta\|_\infty \geq c(n^{-1} \log n)^r \right) = 0.$$

The proof of this theorem, which can be found in Section 5, is simpler and more intuitive than the corresponding proof given by Härdle and Luckhaus (1984). In the context of time series, this result also generalizes that of Collomb and Härdle (1984) under weaker mixing condition and non-differentiability of ψ.

4. DISCUSSION

For $n \geq 1$, let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a random sample of size n from the distribution of $(\mathbf{X}, Y)$ and let k denote a non-negative integer. Let $\theta(\cdot)$ be the regression function, $E(Y \mid \mathbf{X} = \cdot)$, of Y on $\mathbf{X}$ and suppose that $\theta(\cdot)$ has bounded $(k+1)$th derivative. Set $r = p/(2p+d)$ where $p = k+1$. Stone (1980, 1982) showed that if $1 \leq q < \infty$, then n^{-r} is the optimal rate of convergence in both pointwise and L^q norms; while $(n^{-1}\log n)^{-r}$ is the optimal rate of convergence in L^∞ norm. To find an estimator of $\theta(\cdot)$ that achieves these optimal rates of concergence, given $\mathbf{x}$, let $\hat{P}_n(\cdot;\mathbf{x})$ be the polynomial on $\mathbf{R}^d$ of degree k that minimizes

$$\sum\nolimits_{I_n(\mathbf{x})} [\, Y_i - \hat{P}_n(\mathbf{X}_i;\mathbf{x}) \,]^2$$

and set $\hat{\theta}_n(\mathbf{x}) = \hat{P}_n(\mathbf{x};\mathbf{x})$ (if $q = \infty$, define $\hat{\theta}_n$ as above over a finite subset of C and then extend it to all of C by suitable interpolation). Note that this estimator can be easily obtained by solving the corresponding normal equation.

Based on results presented in the previous sections, the following generalization to the case of conditional M-estimates seems plausible. Suppose that the conditional M-predictor $\theta(\cdot)$ has bounded pth derivative. To find an estimator that achieves the above L^q $(1 \leq q \leq \infty)$ rates of convergence, given $\mathbf{x}$, let $\hat{P}_n(\cdot;\mathbf{x})$ be a polynomial on $\mathbf{R}^d$ of degree k which minimizes

$$\sum\nolimits_{I_n(\mathbf{x})} \rho(Y_i - \hat{P}_n(\mathbf{X}_i;\mathbf{x}))$$

and set $\hat{\theta}_n(\mathbf{x}) = \hat{P}_n(\mathbf{x};\mathbf{x})$. The corresponding generalization to time series is straightforward.

One drawback that the nonparametric approach has is the high *dimensionality*, which can be thought of in terms of the *variance* in estimation. In other words: A *huge* data

set may be required for nonparametric estimation of a function of many variables; otherwise the variance of the estimator may be unacceptably large. This drawback is serious especially in time series analysis where the future usually depends on much of the past.

A possible solution would be to use *additivity* as in Stone (1985) to alleviate *curse of dimensionality*. More formally, let $\theta(\cdot)$ be the regression function defined on $\mathbf{R}^d$ and suppose that θ is additive; that is, that there is smooth functions $\theta_1(\cdot),\ldots,\theta_d(\cdot)$ defined on $\mathbf{R}^1$ such that

$$\theta(x_1,\ldots,x_d) = \mu + \theta_1(x_1) + \cdots + \theta_d(x_d)\,,$$

where $\mu = E(Y)$. Using B-splines, an estimator of $\theta(\cdot)$ can be constructed to achieve the optimal rates of convergence n^{-r}, where r now is equal to $p/(2p+1)$. The rates of convergence here do not depend on the dimensional parameter d. Another nice feature about this estimator is that it is smoother and is as flexible as ordinary nonparametric procedures constructed by the kernel method.

The corresponding methodology is generalized immediately to time series, and it is an interesting open problem to determine whether the asymptotic properties described above (with r independent of d) also hold in this context.

5. PROOF OF THEOREMS

Proof of Theorem 1. By symmetry, it suffices to show that

$$\lim_n P(\hat{\theta}_n(\mathbf{0}) \geq \theta(\mathbf{0}) + c\delta_n) = 0.$$

Let B_{ni} be the event that $\|\mathbf{X}_i\| \leq \delta_n$. By Assumption 1, 6 and 7,

$$\begin{aligned} E[\,\psi(Y_i - \theta(\mathbf{0}) - c\delta_n)\,|\,B_{ni}\,] &\leq E[\,\psi(Y_i - \theta(\mathbf{X}_i) - (c - M_0)\delta_n)\,|\,B_{ni}\,] \\ &\leq -(c - M_0)M_1\delta_n \qquad \text{for } c > M_0. \end{aligned} \tag{5.1}$$

Set $K_i = 1_{\{\|\mathbf{X}_i\| \leq \delta_n\}}$ and $Z_i = \psi(Y_i - \theta(\mathbf{0}) - c\delta_n) - E[\,\psi(Y_i - \theta(\mathbf{0}) - c\delta_n)\,|\,\mathbf{X}_i\,]$. Then $E[\,Z_i\,] = 0$ and, by the first argument in the proof of Lemma 7 of Truong (1987), (since Z_i's are bounded)

$$\mathrm{Var}(\textstyle\sum_i K_i Z_i) = \mathrm{Var}(\sum_{I_n} Z_i) = O(n\delta_n^d).$$

According to (5.1)

$$N_n^{-1}\sum_i K_i E[\,\psi(Y_i-\theta(0)-c\delta_n)\,|\,\mathbf{X}_i\,] = N_n^{-1}\sum_i E[\,\psi(Y_i-\theta(\mathbf{X}_i)-(c-M_0)\delta_n)\,|\,B_{ni}\,]$$
$$\leq -(c-M_0)M_1\delta_n \qquad \text{for } c > M_0.$$

Consequently, by Assumption 6, Lemma 6 (Truong, 1987) and Tchebychev's inequality

$$\begin{aligned}
P(\hat{\theta}_n(\mathbf{0}) \geq \theta(\mathbf{0}) + c\delta_n) &= P\left(N_n^{-1}\sum_{I_n}\psi(Y_i-\theta(\mathbf{0})-c\delta_n)\geq 0\right)\\
&\leq P\left(N_n^{-1}\sum_{I_n} Z_i \geq -\sum_{I_n} E[\,\psi(Y_i-\theta(\mathbf{0})-c\delta_n)\,|\,B_{ni}\,]\right)\\
&\leq P\left(N_n^{-1}\sum_{I_n} Z_i \geq (c-M_0)M_1\delta_n\right)\\
&\leq P\left(N_n^{-1}\sum_{I_n} Z_i \geq (c-M_0)M_1\delta_n; N_n \geq \tfrac{1}{2}n\delta_n^d\right) + P(N_n < \tfrac{1}{2}n\delta_n^d)\\
&= \frac{O(1)}{(c-M_0)^2}\frac{n\delta_n^d}{(n\delta_n^d\delta_n)^2} + o(1) = o(1) \qquad \text{as } n, c \to \infty,
\end{aligned}$$

since δ_n is chosen so that $n\delta_n^d\delta_n^2 = 1$, or equivalently, $\delta_n = n^{-r}$. This completes the proof of Theorem 1.

Proof of Theorem 2. Without loss of generality it can be assumed that $C = [\,-\frac{1}{2}, \frac{1}{2}\,]^d$. Set $L_n = [n^{2r}]$. Let W_n be the collection of $(2L_n+1)^d$ points in C each of whose coordinates is of the form $j/(2L_n)$ for some integer j such that $|j| \leq L_n$. Then C can be written as the union of $(2L_n)^d$ subcubes, each having length $2\lambda_n = (2L_n)^{-1}$ and all of its vertices in W_n. For each $\mathbf{x} \in C$ there is a subcube Q_w with center w such that $\mathbf{x} \in Q_w$. Let C_n denote the collection of the centers of these subcubes. Then

$$P\left(\sup_{\mathbf{x}\in C}|\hat{\theta}_n(\mathbf{x})-\theta(\mathbf{x})| \geq \delta_n\right) = P\left(\max_{C_n}\sup_{\mathbf{x}\in Q_w}|\hat{\theta}_n(\mathbf{x})-\theta(\mathbf{x})| \geq \delta_n\right).$$

It follows from $\lambda_n \sim n^{-2r}$ and Assumption 1 that $|\theta(\mathbf{x})-\theta(w)| \leq M_0\|\mathbf{x}-w\| \leq M_0\delta_n$ for $\mathbf{x} \in Q_w$, $w \in C_n$ (for n sufficiently large). Therefore, to prove the lemma, it is sufficient to show that there is a positive constant c such that

$$\lim_n P\left(\max_{w\in C_n}\sup_{\mathbf{x}\in Q_w}|\hat{\theta}_n(\mathbf{x})-\theta(w)| \geq c\delta_n\right) = 0. \tag{5.2}$$

To prove (5.2), let $\eta \equiv \sqrt{d}$, $\mathbf{x} \in Q_w$ and let $I_n^* \equiv I_n^*(w) = \{i : \|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n\}$, $N_n^* \equiv N_n^*(w) = \#I_n^*$. Then $\{\hat{\theta}_n(\mathbf{x}) - \theta(w) \ge c\delta_n\} \subseteq \{N_n^{-1} \sum_{I_n} \psi(Y_i - \theta(w) - c\delta_n) \ge 0\} \subseteq \{N_n^{-1} \sum_{I_n^*} \psi(Y_i - \theta(w) - c\delta_n) \ge 0\}\}$. Thus

$$\cup_{Q_w} \left\{\hat{\theta}_n(\mathbf{x}) - \theta(w) \ge c\delta_n\right\} \subseteq \left\{\sum_{I_n^*} \psi(Y_i - \theta(w) - c\delta_n) \ge 0\right\}. \tag{5.3}$$

Let $B_{ni} = B_{ni}(w)$ denote the event $\|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n$ for $i = 1, \ldots, n$. According to Assumption 1, $\theta(\mathbf{X}_i) \le \theta(w) + M_0(\delta_n + \eta\lambda_n)$ whenever $\|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n$. Thus, by Assumption 6 and 7, there is a positive constant M_2 such that

$$\begin{aligned} E[\,\psi(Y_i - \theta(w) - c\delta_n) \,|\, B_{ni}\,] &\le E[\,\psi(Y_i - \theta(\mathbf{X}_i) - c\delta_n + M_0(\delta_n + \eta\lambda_n)) \,|\, B_{ni}\,] \\ &\le -cM_2\delta_n \qquad \text{for } n \text{ sufficient large.} \end{aligned} \tag{5.4}$$

Set $K_i = 1_{\{\|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n\}}$ and $Z_i = \psi(Y_i - \theta(w) - c\delta_n) - E[\,\psi(Y_i - \theta(w) - c\delta_n) \,|\, \mathbf{X}_i\,]$. Then $E(Z_i) = 0$ and by (5.4)

$$N_n^{*-1} \sum_{I_n^*} E[\psi(Y_i - \theta(w) - c\delta_n) \,|\, \mathbf{X}_i\,] \le -cM_2\delta_n. \tag{5.5}$$

Set $\Psi_n = \left\{N_n^*(w) \ge k_1 n\delta_n^d \quad \text{for all } w \in C_n\right\}$. It follows from (5.3) and (5.5) that

$$\begin{aligned} P\left(\max_{C_n} \sup_{x \in Q_w} [\,\hat{\theta}_n(\mathbf{x}) - \theta(w)\,] \ge c\delta_n\right) &\le P\left(\cup_{C_n} \cup_{Q_w} \left\{\hat{\theta}_n(\mathbf{x}) - \theta(w) \ge c\delta_n\right\}\right) \\ &\le P\left(\cup_{C_n} \left\{N_n^{*-1} \sum_{I_n^*} \psi(Y_i - \theta(w) - c\delta_n) \ge 0\right\}\right) \\ &\le P\left(\cup_{C_n} \left\{N_n^{*-1} \sum_{I_n^*} Z_i \ge cM_2\delta_n\right\}\right) \\ &\le P(\Psi_n^c) \\ &\quad + [n^{2r}]^d \max_{C_n} P\left(\sum_{I_n^*} Z_i \ge ck_1 n\delta_n^{d+1}\right). \end{aligned} \tag{5.6}$$

Note that $\sum_{I_n^*} Z_i = \sum_i K_i Z_i$, $E(K_i Z_i) = 0$, $E|K_i Z_i| = O(\delta_n^d)$ and $\mathrm{Var}(\sum_i K_i Z_i) = O(n\delta_n^d)$ (by Lemma 7 of Truong (1987)). By Lemma 8 (Truong, 1987), there are positive constants M_3 and M_4 such that

$$P\left(\sum_{I_n^*} Z_i \ge ck_1 n\delta_n^{d+1}\right) \le M_3 \exp(-c^2 M_4 n\delta_n^{d+2}) \qquad \text{for } w \in C_n.$$

Since δ_n is chosen so that $n\delta_n^{d+2} \sim \log n$, thus there is a positive integer c such that

$$[n^{2r}]^d \max_{C_n} P\left(\sum_{I_n^*} Z_i \geq ck_1 n\delta_n^{d+1}\right) \to 0 \qquad \text{as } n \to \infty. \tag{5.7}$$

Set $p_n = p_n(w) = P(\|\mathbf{X}_i - w\| \leq \delta_n + \eta\lambda_n)$ (by stationary, p_n does not depend on i). Then $p_n \sim \delta_n^d$. By Lemma 9 of Truong (1987), there are positive constants M_5 and M_6 such that

$$P(N_n^* \leq \tfrac{1}{2}np_n) \leq M_5 \exp(-M_6 n\delta_n^d).$$

Thus

$$[n^{2r}]^d \max_{C_n} P(N_n^* \leq \tfrac{1}{2}np_n) \to 0 \qquad \text{as } n \to \infty.$$

Consequently,

$$\lim_n P(\Psi_n^c) = 0 \qquad \text{as } n \to \infty. \tag{5.8}$$

Hence by (5.7)–(5.8), there is a positive constant c such that

$$\lim_n P\left(\max_{C_n} \sup_{\mathbf{x}\in Q_w} [\hat{\theta}_n(\mathbf{x}) - \theta(w)] \geq c\delta_n\right) = 0. \tag{5.9}$$

Similarly,

$$\lim_n P\left(\max_{C_n} \sup_{\mathbf{x}\in Q_w} [\hat{\theta}_n(\mathbf{x}) - \theta(w)] \leq -c\delta_n\right) = 0. \tag{5.10}$$

It follows from (5.9) and (5.10) that (5.2) is valid. This completes the proof of Theorem 2.

REFERENCES

BIERENS, H. J. (1983) Uniform consistency of kernel estimators of a regression function under generalized conditions. *J. Amer. Statist. Assoc.* **78**, 699–707.

BLOOMFIELD, P. and STEIGER, (1983) *Least Absolute Deviations.* Birkhauser, Boston.

BRILLINGER, D. R. (1981) *Time series: Data analysis and theory.* Holden-Day, San Francisco.

COLLOMB, B. G. (1984) Propriétés de convergence presque complète du prédicteur à noyau. *Z. Wahrsch. verw. Gebiete* **66**, 441–460.

COLLOMB, B. G. and HÄRDLE, W. (1984) Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations. *Preprint.*

DOUKHAN, P. and GHINDES, M. (1980) Estimations dans le processus $X_{n+1} = f(X_n) + \epsilon_n$. *C. R. de l'Academie des Sciences, Paris.* **291**, 61–64.

DUNSMUIR, W. and HANNAN, E. J. (1976) Vector linear time series models. *Adv. Appl. Prob.* **8**, 339–364.

HAGGAN, V. and OZAKI, T. (1980) Amplitude-dependent exponential AR model fitting for non-linear random vibrations. In *Time Series*, edited by O. D. Anderson. North-Holland: Amsterdam.

HAGGAN, V. and OZAKI, T. (1981) Modelling non-linear random vibrations using an amplitude-dependent autoregressive model. *Biometrika* **68**, 189–196.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986) *Robust statistics: The approach based on influence functions.* Wiley, New York.

HANNAN, E. J. (1970) *Multiple Time series.* Wiley, New York.

HANNAN, E. J. (1973) The asymptotic theory of linear time series models. *J. Appl. Prob.* **10**, 130–145.

HÄRDLE, W. and LUCKHAUS, S. (1984) Uniform consistency of a class of regression function estimators. *Ann. Statist.* **12**, 612–623.

HUBER, P. J. (1981) *Robust statistics.* Wiley, New York.

NICHOLLS, D. F. and QUINN, B. G. (1980) The estimation of random coefficient autoregressive models. I. *J. Time Series Anal.* **1**, 37–46.

PRIESTLEY, M. B. (1979) *Time series and spectral analysis.* Academic Press, New York.

PRIESTLEY, M. B. (1980) State-dependent models: A general approach to non-linear time series analysis. *J. Time Series Anal.* **1**, 47–71.

ROBINSON, P. M. (1983) Nonparametric estimators for time series. *J. Time Series Anal.* **4**, 185–207.

STONE, C. J. (1980) Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348–1360.

STONE, C. J. (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053.

STONE, C. J. (1985) Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689–705.

TONG, H. and LIM, K. S. (1980) Threshold Autoregression, limit cycles and cyclical data. *J. Roy. Statist. Soc. B* **42**, 245–292.

TRUONG, K. Y. (1987) Nonparametric time series prediction: Kernel estimators based on local averages. **Institute of Statistics Mimeo Series No. 1832.**

TRUONG, K. Y. and STONE, C. S. (1987a) Asymptotic Properties of nonparametric time series prediction. **Institute of Statistics Mimeo Series No. 1830.**

TRUONG, K. Y. and STONE, C. S. (1987b) Nonparametric time series prediction. II. Kernel estimators based on local medians. **Institute of Statistics Mimeo Series No. 1833.**