Ann. Inst. Statist. Math.
Vol. 48, No. 2, 267–281 (1996)

MINIMUM DISTANCE REGRESSION-TYPE ESTIMATES WITH RATES UNDER WEAK DEPENDENCE*

GEORGE G. ROUSSAS[1] AND YANNIS G. YATRACOS[2]

[1] *Division of Statistics, University of California, Davis, CA 95616-8705, U.S.A.*
[2] *Departement de mathematiques et de statistique, Université Montréal,*
C.P. 6128, succursale A, Montreal, Quebec, Canada H3C 3J7
and University of California, Santa Barbara

(Received June 1, 1992; revised June 5, 1995)

Abstract. Under weak dependence, a minimum distance estimate is obtained for a smooth function and its derivatives in a regression-type framework. The upper bound of the risk depends on the Kolmogorov entropy of the underlying space and the mixing coefficient. It is shown that the proposed estimates have the same rate of convergence, in the L_1-norm sense, as in the independent case.

Key words and phrases: Kolmogorov's entropy, minimum distance estimation, nonparametric regression, φ-mixing, rate of convergence.

1. Introduction

Let $\{Z_j\}$ be a strictly stationary discrete time-parameter time series of real-valued random variables (r.v.'s); no parametric model is stipulated for the time series. Then one of the many statistical problems of interest is that of nonparametrically estimating the conditional expectation of Z_{j+1} on the basis of the immediately u previous observations $Z_{j-u+1}, \dots, Z_j$. This problem may be cast in a slightly different framework as follows. To this effect, set $X_j = (Z_{j-u+1}, \dots, Z_j)$ and $Y_j = Z_{j+1}$, so that (X_j, Y_j) is a strictly stationary sequence of pairs of observations. Then the problem of estimating $\mathcal{E}(Z_{j+1} \mid Z_j, \dots, Z_{j-u+1})$ is identical to that of estimating the regression $\mathcal{E}(Y_j \mid X_j)$. There are other variations of the initial problem, such as, for example, the estimation of the expected value m steps ahead in time in terms of u observations as above; that is, $\mathcal{E}(Z_{j+m} \mid Z_j, \dots, Z_{j-u+1})$. Again, this problem is the same as that of estimating the $\mathcal{E}(Y_j \mid X_j)$, where now $Y_j = Z_{j+m}$ and X_j is as before.

From this point on, consider the observations (X_j, Y_j), $j \geq 1$, where the X_j's are $\mathcal{X}$-valued and the Y_j's, the respective responses, are real-valued. The set

* This work was partially supported by a research grant from the Natural Sciences and Engineering Research Council of Canada.

$\mathcal{X}$ is a compact subset of $\Re^d$, where d is an integer ≥ 1, and, without loss of generality, we may assume that $\mathcal{X} = [0,1]^d$. Let Θ be the space of real-valued continuous functions defined on $\mathcal{X}$ endowed with the sup-norm, and let X and Y be distributed as X_1 and Y_1, respectively. It is assumed that for each $x \in \mathcal{X}$, the conditional distribution of Y, given $X = x$, is dominated by a σ-finite measure, μ_x, and has a probability density function (p.d.f.) of known functional form involving $\theta(x)$, where θ is an element of Θ; that is, $Y \mid X = x \sim f(\cdot \mid x, \theta(x))$. It is to be emphasized, however, that $\theta(x)$ need not be the (conditional) expectation, as is usually the case in the literature. It may be, for instance, the median or a specific quantile or any other characteristic of the conditional p.d.f.

Presently, our aim is that of estimating θ by means of the principle of minimum distance, and calculate the rate of convergence of the proposed estimate to the true parameter in L_1-distance. Actually, this problem has been considered and resolved in Yatracos (1989*a*, 1992), under suitable regularity conditions, provided that, conditionally on $X_1 = x_1, \ldots, X_n = x_n$, the corresponding r.v.'s Y_j, $j = 1, \ldots, n$ are independent. The problem so framed includes as special cases the so-called classical regression problem, where $\theta(x) = \mathcal{E}(Y \mid X = x)$. This latter problem has been discussed by several authors, including Devroye and Wagner (1980), Ibragimov and Khas'minskii (1980), and Stone (1980, 1982). In each case, θ belongs to subsets of Θ consisting of sufficiently "smooth" functions. Relevant is also the reference Yatracos (1985). An early rigorous usage of the principle of minimum distance goes back to Wolfowitz (1957). Beran (1977) employed the Hellinger distance for constructing estimates in parametric models.

The basic difference between the problem discussed here and those resolved in Yatracos (1989*a*, 1992) is that the assumption of independence, which plays a fundamental role in the latter paper, is replaced by φ-mixing (see Definition 2.1(i) below), thus considerably enlarging the range of potential applications. It is known that many stochastic processes satisfy a φ-mixing condition. Such processes include, for example, m-dependent r.v.'s, Markov processes satisfying Doeblin's condition, and Markov processes which are geometrically ergodic. (Details may be found, for instance, in Roussas and Ioannides (1987), Examples 3.2, 3.3 and 3.4.) To be sure, a preferable mode of mixing would be α-mixing (see Definition 2.1 (ii) below), which is weaker than φ-mixing, and is satisfied by a wider class of important stochastic processes. Such issues are discussed in Chanda (1974), Pham and Tran (1985), and Pham (1986); see also Yoshihara (1992). Questions similar to the ones discussed here, but under α-mixing, are currently under investigation; it is hoped we will be able to report on our findings some time in the near future. The case of estimating $\theta(x) = \mathcal{E}(Y \mid X = x)$, under dependence, has been discussed rather extensively. Some references to this and related problems are Robinson (1986), Roussas (1990), and Tran (1989, 1990, 1993). For a general theory of estimation in abstract parameter spaces, the reader is referred to Le Cam (1986).

The paper is organized as follows. In Section 2, the relevant concepts are defined, the assumptions under which the results of the paper are derived are formulated, and two auxiliary results, Lemmas 2.1 and 2.2, are stated. The main results of the paper, Theorem 3.1 followed by a corollary, are stated and proved in Section 3. A reference to Lemma 2.1 is given and the proof of Lemma 2.2

is presented in the Appendix. All limits are taken as $n \to \infty$ unless otherwise explicitly stated.

2. Definition, assumptions, and preliminary results

For $n = 1, 2, \ldots$, let U_n be $\Re^t$-valued r.v.'s defined on a probability space $(\Omega, \mathcal{A}, P)$, and for i, j with $1 \leq i < j \leq \infty$, let $\mathcal{F}_i^j$ be the σ-field induced by the r.v.'s U_n, $n = i, i+1, \ldots, j$.

Definition 2.1. (i) The not necessarily (strictly) stationary sequence $\{U_n\}$, $n \geq 1$, is said to be φ-mixing with mixing coefficient $\varphi(n)$, if, as $n \to \infty$:

$$\sup\left\{\frac{|P(A \cap B) - P(A)P(B)|}{P(A)}; A \in \mathcal{F}_1^k, B \in \mathcal{F}_{k+n}^\infty, k \geq 1\right\} = \varphi(n) \downarrow 0;$$

if the stochastic process is stationary, then the sup over k above is superfluous.

(ii) The process is said to be α-mixing with mixing coefficient $\alpha(n)$, if, as $n \to \infty$:

$$\sup\{|P(A \cap B) - P(A)P(B)|; A \in \mathcal{F}_1^k, B \in \mathcal{F}_{k+n}^\infty, k \geq 1\} = \alpha(n) \downarrow 0;$$

once again, the sup over k is unnecessary, if the process is stationary.

For φ-mixing sequences, the probability inequality stated below holds; this inequality is instrumental in this paper. For its formulation, let $\nu = \nu(n)$ be positive integers tending to ∞, and set $\mu = \mu(n) = [\frac{n}{2\nu}]$, where $[x]$ denotes the integral part of x. Thus, the μ's are the largest integers for which $2\nu\mu \leq n$, $\mu \to \infty$ and $\frac{n}{2\nu\mu} \to 1$.

Lemma 2.1. *Let Z_n, $n \geq 1$, be real-valued r.v.'s centered at their expectations and bounded by $\hat{M}$, and suppose that they are φ-mixing with mixing coefficient $\varphi(n)$ such that $\sum_{n=1}^\infty \varphi(n) \overset{\text{def.}}{=} \varphi^* < \infty$. Set $\bar{S}_n = \frac{1}{n}\sum_{i=1}^n Z_i$, and let $C = 1 + 4\varphi^*$. Then, for all $n \geq 1$:*

$$P(|\bar{S}_n| \geq \varepsilon_n) \leq 6[1 + 2e^{1/2}\varphi(\nu)]^\mu \exp\left(-\frac{n\varepsilon_n^2}{2C\hat{M}^2}\right), \tag{2.1}$$

$$\text{where} \quad 0 < \varepsilon_n \leq \frac{C\hat{M}\mu}{n}.$$

A discussion of such an inequality may be found, for example, in Roussas and Ioannides (1988).

The set $\Theta \subset C(\mathcal{X})$, whose elements θ index the conditional p.d.f. of $Y \mid X = x$, $f(\cdot \mid x, \theta(x))$, is defined as follows.

Definition 2.2. The set Θ is the collection of real-valued functions defined on $[0, 1]^d$, $d \geq 1$ integer, which are uniformly bounded in the sup-norm and whose

p-th order mixed partial derivative $\theta^{(p)}(\cdot)$ satisfies the following Lipschitz condition:

$$(2.2) \qquad |\theta^{(p)}(x) - \theta^{(p)}(y)| \leq L|x-y|^{\alpha}, \qquad 0 < \alpha \leq 1.$$

By setting $q = p+\alpha$, the set Θ is also denoted by $\Theta_{q,d}$ and its elements are referred to as q-"smooth" functions.

The set Θ, as defined above, supplied with the distance D induced by the sup-norm is *totally bounded.* That is to say, for any $a > 0$, there exists a finite number of balls, $N(a)$ say, centered at some points of Θ and having radius a, whose union is Θ. Furthermore, it follows from Kolmogorov and Tikhomirov (1961) that the space (Θ, D) has *Kolmogorov's entropy* $\log_2 N(a_n)$, where $N(a_n) \sim 2^{(1/a_n)^{d/q}}$, $\log_2 N(a_n)$ is log of $N(a_n)$ with base 2, and $N(a_n)$ is the most economic number of balls as described above. The notation $x_n \sim y_n$ signifies that both $x_n = O(y_n)$ and $y_n = O(x_n)$.

Let Θ be as above, recall that $\mathcal{X} = [0,1]^d$, $d \geq 1$ integer, and let $f(\cdot \mid x, \theta(x))$ be the p.d.f. as described in the Introduction. Also, let $X_1, \ldots, X_n$ be the first n observations in the pairs (X_i, Y_i), $i = 1, \ldots, n$, and let $x_1, \ldots, x_n$ be their observed values. In the derivations to follow, it is required that these x_i's are sufficiently dense in $\mathcal{X}$. To make this precise, let $\lambda > 0$, $c > 0$, and let Q^n be the the joint distribution of $X_1, \ldots, X_n$. Define the set $C_{n,d,\lambda}$ as follows:

$$(2.3) \quad C_{n,d,\lambda} = \{(x_1, \ldots, x_n) \in \mathcal{X}^n; \#\{i \text{ for which } |x_i - x| < n^{-\lambda}\} \geq cn^{1-\lambda d} \text{ for every } x \in \mathcal{X}\}.$$

Then proceed to gather together the assumptions under which the results in this paper are derived.

ASSUMPTIONS.

(A1) (i)For $n = 1, 2, \ldots, \{(X_n, Y_n)\}$ is a stationary sequence of observations, where the X_n's are $\mathcal{X}$-valued and the Y_n's are real-valued. It is further assumed that the sequence is φ-mixing with mixing coefficient $\varphi(n)$ of the form: $\varphi(n) = O(\frac{\tau_n}{n})$ for some $0 < \tau_n \to 0$, and $\sum_{n=1}^{\infty} \varphi(n) < \infty$.

(ii) The p.d.f. $f(\cdot \mid x, \theta(x))$, $x \in \mathcal{X}$, as described in the Introduction, is of known functional form except that it depends on $\theta \in \Theta$.

For any two p.d.f.'s $f(\cdot \mid x, s)$ and $f(\cdot \mid x, t)$, consider the L_1-norm $\|f(\cdot \mid x, s) - f(\cdot \mid x, t)\|$ defined by:

$$(2.4) \qquad \|f(\cdot \mid x, s) - f(\cdot \mid x, t)\| = \int_{\Re} |f(y \mid x, s) - f(y \mid x, t)| d\mu_x.$$

Then:

(A2) The norm defined in (2.4) is $\sim |s-t|$. More precisely, there exist (positive) constants C_1 and C_2, independent of x, such that:

$$(2.5) \qquad C_1|s-t| \leq \|f(\cdot \mid x, s) - f(\cdot \mid x, t)\| \leq C_2|s-t|.$$

(A3) Let $C_{n,d,\lambda}$ be defined by (2.3), and let Q^n be the joint distribution of $X_1, \ldots, X_n$. Then, for all sufficiently small $c > 0$ and a suitable $0 < \lambda < 1/d$:

$$Q^n(C_{n,d,\lambda}) \to 1. \tag{2.6}$$

Remark 2.1. Of the assumptions just made, Assumptions (A2) and (A3) deserve, perhaps, a comment. Assumption (A1) is nothing out of the ordinary. Condition (2.5) is not as strong as it may look at first glance. Concrete examples where such a condition holds, have been worked out in Yatracos (1989*a*), where the interested reader is referred to. There are seven such examples, where the p.d.f. $f(\cdot \mid x, \theta(x))$ ranges from normal to negative exponential to Poisson to geometric to binomial to uniform and negative exponential with only location parameter unknown. Condition (2.6) is, indeed, somewhat unusual although nowhere as strong as it may look. This point is illustrated by the fact that condition (2.6) is implied by familiar and mild conditions on the process. This is the content of the following lemma.

LEMMA 2.2. *Suppose that the mixing coefficient $\varphi(n)$ is of the form $\varphi(n) = O(n^{-(1+\delta)})$ for some $\delta > 0$, and the r.v. X has a p.d.f. which is bounded from below in $\mathcal{X}$ (by $M_1 > 0$, say). Then condition* (2.6) *holds for all sufficiently small $c > 0$ and $\lambda = q/d(2q + d)$.*

The proof of this lemma is presented in the Appendix in order not to disrupt the flow of the main ideas involved.

3. Formulation and proof of main results

Before the main results of this paper are formulated, the minimum distance estimate of θ has got to be defined, and for this purpose, we proceed as follows. The parameter space $\Theta_{q,d}$ is sup-norm totally bounded, and for $a_n > 0$, the most economical a_n-dense subset of it, $\Theta_{n,q,d}$, has cardinality $N_{q,d}(a_n) \sim 2^{(1/a_n)^{d/q}}$ (see Kolmogorov and Tikhomirov (1961)). In all that follows, let us simplify the notation by writing Θ, Θ_n and N_n instead of $\Theta_{q,d}$, $\Theta_{n,q,d}$ and $N_{q,d}(a_n)$, respectively. Let $\Theta_n = \{\theta_{nj}, j = 1, \ldots, N_n\}$, and given $X_i = x_i$, $i = 1, \ldots, n$, set

$$A^n_{k,\ell,i} = \{y \in \Re; f(y \mid x_i, \theta_k(x_i)) > f(y \mid x_i, \theta_\ell(x_i))\}, \qquad 1 \le k < \ell \le N_n. \tag{3.1}$$

Let Y_i be the observation taken at x_i, $i = 1, \ldots, n$, and set

$$\bar{S}_{n;k,\ell,m} = \frac{1}{n}\sum_{i=1}^{n}[I_{A^n_{k,\ell,i}}(Y_i) - P_{x_i,\theta_m(x_i)}(A^n_{k,\ell,i})], \tag{3.2}$$
$$1 \le k < \ell \le N_n, \qquad m = 1, \ldots, N_n,$$

where $P_{x_i,\theta_m(x_i)}$ is the conditional distribution of Y_i, given $X_i = x_i$, calculated under $\theta_m(x_i)$.

Next, maximize $\bar{S}_{n;k,\ell,m}$ over k, ℓ and m, varying as above, and then define the minimum distance estimate $\hat{\theta}_n$ as that θ which minimizes this maximum. More precisely, $\hat{\theta}_n$ is defined by the following relationship:

$$\begin{aligned}
\max\left\{\frac{1}{n}\left|\sum_{i=1}^{n}[I_{A^n_{k,\ell,i}}(Y_i) - P_{x_i,\hat{\theta}_n(x_i)}(A^n_{k,\ell,i})]\right|; 1 \le k < \ell \le N_n\right\} \\
= \min\left[\max\left\{\frac{1}{n}\left|\sum_{i=1}^{n}[I_{A^n_{k,\ell,i}}(Y_i) - P_{x_i,\theta_m(x_i)}(A^n_{k,\ell,i})]\right|;\right.\right. \\
\left.\left. 1 \le k < \ell \le N_n; 1 \le m \le N_n\right\}\right].
\end{aligned} \tag{3.3}$$

We may now formulate the main results of this paper; all logarithms will be with base 2, although this will not be denoted explicitly.

THEOREM 3.1. *Suppose Assumptions* (A1)–(A3) *are fulfilled, and let the parameter space* $\Theta_{q,d}$ *be as in Definition* 2.2. *Then the minimum distance estimate* $\hat{\theta}_n$ *defined by* (3.3) *is a uniformly weakly consistent estimate of* θ *with rate of convergence* a_n (*the same as in the independent case*), *where*

$$a_n = \left(\frac{\log N_{q,d}(a_n)}{n}\right)^{1/2} = n^{-q/(2q+d)};$$

convergence is to be understood in the L_1*-norm sense, where, for any* θ *and* $\tilde{\theta}$ *in* $\Theta_{q,d}$:

$$\|\theta - \tilde{\theta}\| = \int_{\mathcal{X}} |\theta(x) - \tilde{\theta}(x)|dx.$$

COROLLARY 3.1. *For the minimum distance estimate* $\hat{\theta}_n$ *defined in* (3.3), *it holds* (*the same as in the independent case*): $\|\hat{\theta}_n^{(s)} - \theta^{(s)}\| \le C^* n^{-(q-[s])/(2q+d)}$ *in probability* ($C^* > 0$ *constant*).

PROOF OF THEOREM 3.1. Recall that $\mathcal{X} = [0,1]^d$ and split up $\mathcal{X}$ into hypercubes, S_i, $i = 1, \ldots, b_n^{-d}$ with side length b_n. Then

$$\int_{\mathcal{X}} |\hat{\theta}_n(x) - \theta(x)|dx = \sum_{i=1}^{b_n^{-d}} \int_{S_i} |\hat{\theta}_n(x) - \theta(x)|dx. \tag{3.4}$$

Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$. When $\mathbf{X} = \mathbf{x} \in C_{n,d,\lambda}$, let N_i be the number of the coordinates of $\mathbf{x}$ in S_i, and let $M = \min\{N_i; 1 \le i \le b_n^{-d}\}$. Restricting attention to S_i, we approximate $\hat{\theta}_n(x)$ and $\theta(x)$ by their Taylor polynomial of order p around $x_j \in S_i$. The remainder term will be, clearly, bounded in absolute value by $C^* b_n^q$ in both cases; here C^* is a suitable positive constant. The constant C^* will be replaced sequentially by another majorizing constant, however, for the sake of simplicity, we will retain the notation C^* throughout. Retain only the term $|\hat{\theta}_n(x_j) - \theta(x_j)|$ and repeat the same approximation around x for each of the remaining terms in the Taylor expansion. For $s = (s_1, \ldots, s_d)$, let $[s] = s_1 + \cdots + s_d$, and let $\theta^{(s)}(x_0)$ denote the s-th order mixed partial derivative of θ at x_0. Then we obtain:

$$
\begin{aligned}
(3.5)\qquad & \int_{S_i} |\hat{\theta}_n(x) - \theta(x)| dx \\
& \leq C^* \left[b_n^{q+d} + b_n^d |\hat{\theta}_n(x_j) - \theta(x_j)| \right. \\
& \qquad \left. + \sum_{1 \leq [s] \leq p} b_n^{[s]} \int_{S_i} |\hat{\theta}_n^{(s)}(x_j) - \theta^{(s)}(x_j)| dx \right] \\
& \leq C^* \left[b_n^{q+d} + b_n^d |\hat{\theta}_n(x_j) - \theta(x_j)| \right. \\
& \qquad \left. + \sum_{1 \leq [s] \leq p} \sum_{0 \leq [t] \leq p-s]} b_n^{[s+t]} \int_{S_i} |\hat{\theta}_n^{(s+t)}(x) - \theta^{(s+t)}(x)| dx \right] \\
& \leq C^* \left[b_n^{q+d} + b_n^d |\hat{\theta}_n(x_j) - \theta(x_j)| \right. \\
& \qquad \left. + \sum_{1 \leq [s] \leq p} b_n^{[s]} \int_{S_i} |\hat{\theta}_n^{(s)}(x) - \theta^{(s)}(x)| dx \right].
\end{aligned}
$$

For each i with $1 \leq i \leq b_n^{-d}$, repeat (3.5) for M out of the N_i elements in S_i and then add up the relations to obtain:

$$
M \|\hat{\theta}_n - \theta\| \leq C^* \left[M b_n^q + b_n^d \sum_{j=1}^{n} |\hat{\theta}_n(x_j) - \theta(x_j)| + M \sum_{1 \leq [s] \leq p} b_n^{[s]} \|\hat{\theta}_n^{(s)} - \theta^{(s)}\| \right].
$$

Since $\mathbf{x} \in C_{n,d,\lambda}$, it follows (by 2.3) that $M \geq cnb_n^d$.

Thus, we have:

$$
(3.6)\qquad \|\hat{\theta}_n - \theta\| \leq C^* \left[b_n^q + \frac{1}{n} \sum_{j=1}^{n} |\hat{\theta}_n(x_j) - \theta(x_j)| + \sum_{1 \leq [s] \leq p} b_n^{[s]} \|\hat{\theta}_n^{(s)} - \theta^{(s)}\| \right].
$$

By Proposition 2 in Yatracos (1989*b*), it follows that, for $1 \leq [s] \leq p$:

$$
(3.7)\qquad \|\hat{\theta}_n^{(s)} - \theta^{(s)}\| \leq D_1 \gamma_n^{q-[s]} + D_2 \gamma_n^{-[s]} \|\hat{\theta}_n - \theta\|,
$$

where D_1, D_2 and γ_n are positive constants. Take $\gamma_n = \hat{D} b_n$, where $\hat{D}$ is a large enough positive constant, and employ (3.6) and (3.7) to obtain:

$$
(3.8)\qquad \|\hat{\theta}_n - \theta\| \leq C^* \left[a_n + b_n^q + \frac{1}{n} \sum_{j=1}^{n} |\hat{\theta}_n(x_j) - \theta_m(x_j)| \right],
$$

where θ_m is the element of Θ_n closest to θ. Furthermore:

$$\begin{aligned}(3.9)\quad &\frac{1}{n}\sum_{j=1}^{n}|\hat{\theta}_n(x_j)-\theta_m(x_j)| \\ &\leq \tilde{C}_1 a_n + \tilde{C}_2 \max\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{i=1}^{n}[I_{A^n_{k,\ell,i}}(Y_i)-P_{x_i,\theta(x_i)}(A^n_{k,\ell,i})\right]\right|;\right. \\ &\qquad\qquad\qquad\qquad\qquad\qquad \left. 1\leq k<\ell\leq N_n\right\}\end{aligned}$$

(see relations (5) in Yatracos (1989*a*)). By means of (3.6)–(3.9), one has:

$$\begin{aligned}(3.10)\qquad \|\hat{\theta}_n-\theta\| \leq C^*\Bigg[a_n + b_n^q \\ +\max\left\{\left|\frac{1}{n}\sum_{j=1}^{n}[I_{A^n_{k,\ell,i}}(Y_i)-P_{x_i,\theta(x_i)}(A^n_{k,\ell,i})]\right|;\right. \\ \left. 1\leq k<\ell\leq N_n\right\}\Bigg].\end{aligned}$$

It is at this point, where the mixing assumption on the observables enters the picture. To this effect, let $P^n_{\mathbf{x},\theta(\mathbf{x})}$, to be shortened to P^n_θ, stand for the conditional joint distribution of the Y_i's, $i=1,\ldots,n$, given $\mathbf{X}=\mathbf{x}\in C_{n,d,\lambda}$, and apply inequality (2.1) to the third term on the right-hand side of (3.10) with $\hat{M}=1$ (see Lemma 2.1), to obtain:

$$\begin{aligned}(3.11)\qquad & P^n_\theta(\|\hat{\theta}_n-\theta\|\geq C^*\varepsilon_n) \\ &\leq P^n_\theta\left[\max\left\{\left|\frac{1}{n}\sum_{i=1}^{n}[I_{A^n_{k,\ell,i}}(Y_i)-P_{x_i,\theta(x_i)}(A^n_{k,\ell,i})]\right|;\right.\right. \\ &\qquad\qquad \left.\left. 1\leq k<\ell\leq N_n\right\}\geq \varepsilon_n-a_n-b_n^q\right] \\ &\leq 6N_n^2[1+2e^{1/2}\varphi(\nu)]^{\mu}\exp\left[-\frac{n}{2C}(\varepsilon_n-a_n-b_n^q)^2\right] \\ &\leq 6N_n^2[1+2e^{1/2}\varphi(\nu)]^{n/2\nu}\exp\left[-\frac{n}{2C}(\varepsilon_n-a_n-b_n^q)^2\right],\end{aligned}$$

provided

$$(3.12)\qquad (0<)\varepsilon_n-a_n-b_n^q\leq C/3\nu.$$

The specification of the quantities μ and ν above is given in the paragraph just prior to Lemma 2.1, and inequality (3.12) implies the inequality $\varepsilon_n\leq C\mu/n$ in

(2.1) for all sufficiently large n. Also, the last inequality on the right-hand side in (3.11) follows by the fact that $\mu \leq n/2\nu$.

By the fact that the expression on the right-hand side of (3.11) is independent of θ, it suffices to show that this expression tends to 0. For simplicity, set $2e^{1/2} = C_1$, $\frac{1}{2C} = C_2$. From (3.11), we have then to determine b_n, a_n and ε_n to satisfy (3.12) and also the convergence

$$N_n^2[1 + C_1\varphi(\nu)]^{n/2\nu} \exp[-C_2 n(\varepsilon_n - a_n - b_n^q)^2] \to 0,$$

or, equivalently,

$$\hat{C}_2 n(\varepsilon_n - a_n - b_n^q)^2 - 2\log N_n - \frac{n}{2\nu}\log[1 + C_1\varphi(\nu)] \to \infty, \tag{3.13}$$

where $\hat{C}_2 = C_2 \log e = \log e/2C$. Take $b_n = a_n^{1/q}$, so that $b_n^q = a_n$. Then (3.13) becomes:

$$\hat{C}_2 n(\varepsilon_n - 2a_n)^2 - 2\log N_n - \frac{n}{2\nu}\log[1 + C_1\varphi(\nu)] \to \infty. \tag{3.14}$$

From $1 + t \leq e^t (t \geq 0)$, we get $\log(1 + t) \leq t \log e$. Apply this inequality for $t = C_1\varphi(\nu)$ to obtain: $\log[1+C_1\varphi(\nu)] \leq \hat{C}_1\varphi(\nu)$ where $\hat{C}_1 = C_1 \log e = 2e^{1/2}\log e$, and define ν by:

$$\nu = \left[\left(k \cdot \frac{n}{\log N_n}\right)^{1/2}\right], \tag{3.15}$$

so that

$$\frac{1}{2}\left(k \cdot \frac{n}{\log N_n}\right)^{1/2} \leq \nu \leq \left(k \cdot \frac{n}{\log N_n}\right)^{1/2}$$

for all sufficiently large n; k is a constant to be specified below (see (3.20)). By means of (3.15) and the form of $\varphi(n)$, we have then:

$$\frac{n}{2\nu}\log[1 + C_1\varphi(\nu)] \leq 0.5 \log N_n \text{ for all sufficiently large } n.$$

Then (3.14) is implied by:

$$\hat{C}_2 n(\varepsilon_n - 2a_n)^2 - 2.5 \log N_n \to \infty. \tag{3.16}$$

At this point, take a_n and ε_n as follows:

$$\begin{aligned} a_n &= \left(\frac{\log N_n}{n}\right)^{1/2} = n^{-\frac{q}{2q+d}}, \\ \varepsilon_n &= 2a_n + \rho\left(\frac{\log N_n}{n}\right)^{1/2} = (2 + \rho)\left(\frac{\log N_n}{n}\right)^{1/2}, \end{aligned} \tag{3.17}$$

where $\rho > 0$ is to be determined below (see (3.18)). For a_n and ε_n as above, the convergence in (3.16) becomes:

$$(\hat{C}_2\rho^2 - 2.5)\log N_n = \left(\frac{\log e}{2C}\rho^2 - 2.5\right)\log N_n \to \infty,$$

and this, actually, holds, provided

$$\rho > (5C/\log e)^{1/2}. \tag{3.18}$$

On the other hand, with the above choices of b_n^q, a_n and ε_n, inequality (3.12) becomes: $\rho(\log N_n/n)^{1/2} \le C/3\nu$, and by way of (3.15), this inequality is implied by: $\rho(\frac{\log N_n}{n})^{1/2} \le \frac{C}{3k^{1/2}}(\frac{\log N_n}{n})^{1/2}$, or

$$\rho \le \frac{C}{3k^{1/2}}. \tag{3.19}$$

Relations (3.18) and (3.19) are consistent, provided $(5C/\log e)^{1/2} < \frac{C}{3k^{1/2}}$, or

$$k < \frac{C\log e}{45}. \tag{3.20}$$

To summarize: the quantities ν, and a_n, ε_n, given by relations (3.15) and (3.17), respectively, satisfy inequality (3.12), and also cause the expression on the right-hand side in (3.11) to tend to 0. Then the proof is completed by writing

$$\begin{aligned} P(\|\hat{\theta}_n - \theta\| \ge C^*\varepsilon_n) &= \mathcal{E}_{Q^n} P_\theta^n(\|\hat{\theta}_n - \theta\| \ge C^*\varepsilon_n | \mathbf{X} = \mathbf{x}) I(\mathbf{x} \in C_{n,d,\lambda}) \\ &\quad + \mathcal{E}_{Q^n} P_\theta^n(\|\hat{\theta}_n - \theta\| \ge C^*\varepsilon_n | \mathbf{X} = \mathbf{x}) I(\mathbf{x} \in C_{n,d,\lambda}^c), \end{aligned}$$

and using assumption (A3). □

Proof of Corollary 3.1. For $[s]$ with $1 \le [s] \le p$, we have as in (3.7):

$$\|\hat{\theta}_n^{(s)} - \theta^{(s)}\| \le D_1\gamma_n^{q-[s]} + D_2\gamma_n^{-[s]}\|\hat{\theta}_n - \theta\|,$$

where D_1, D_2 and γ_n are positive constants. But $\|\hat{\theta}_n - \theta\| \le C^* n^{-q/(2q+d)}$ in probability for some $C^* > 0$, and γ_n may be chosen to be: $\gamma_n = n^{-1/(2q+d)}$. Then, retaining the same notation C^* for a majorizing constant, we obtain:

$$\|\hat{\theta}_n^{(s)} - \theta^{(s)}\| \le C^* n^{-(q-[s])/(2q+d)} \quad \text{in probability.} \qquad \square$$

Acknowledgements

The authors are indebted to an associate editor and an anonymous referee for their constructive comments which helped improve an earlier version of this paper.

Appendix

The purpose of this appendix is to establish Lemma 2.2 which plays a central role in the proof of the main result of the paper. The proof of the lemma is given for $d = 1$; its proof for $d > 1$ is quite analogous. Thus, let $d = 1$, so that $\mathcal{X} = [0, 1]$, and for each $x \in \mathcal{X}$, set:

$$\text{(A.1)} \qquad K_{ni}(x) = I_{(x-\delta_n, x+\delta_n)}(x_i), \qquad i = 1, \dots, n, \qquad K_n(x) = \sum_{i=1}^{n} K_{ni}(x);$$

here $\delta_n = n^{-\lambda}$, and recall that x_i is the observed value of X_i, $i = 1, \dots, n$. With the above notation, observe that (2.5) may be rewritten as follows:

$$\begin{aligned}\text{(A.2)} \quad C_{n,1,\lambda} = C_n &= \{(x_1, \dots, x_n) \in \mathcal{X}^n; K_n(x) \geq cn\delta_n \text{ for every } x \in \mathcal{X}\} \\ &= \bigcap_{x \in \mathcal{X}} \{(x_1, \dots, x_n) \in \mathcal{X}^n; K_n(x) \geq cn\delta_n\}.\end{aligned}$$

The collection of the sets $\{(x_1, \dots, x_n) \in \mathcal{X}^n; K_n(x) \geq cn\delta_n\}$, $x \in \mathcal{X}$, is discretized as follows: Let $\varepsilon > 0$ to be specified below, and divide the unit interval $[0, 1]$ into the intervals J_{nj}, $j = 1, \dots, N$ defined by:

$$\text{(A.3)} \quad J_{nj} = [(j-1)\varepsilon\delta_n, j\varepsilon\delta_n], \qquad j = 1, \dots, N-1, \qquad J_{nN} = [(N-1)\varepsilon\delta_n, 1];$$

here $N = 1/\varepsilon\delta_n$, if this expression is an integer, or $N = [1/\varepsilon\delta_n] + 1$ otherwise. At any rate, $N < \frac{1}{\varepsilon\delta_n} + 1 \leq \frac{\varepsilon+1}{\varepsilon} n^{\lambda}$. Next, set:

$$\text{(A.4)} \qquad L_{nji} = I_{J_{nj}}(x_i), \qquad i = 1, \dots, n, \ j = 1, \dots, N, \qquad L_{nj} = \sum_{i=1}^{n} L_{nji}.$$

Then the following relationship holds true.

LEMMA A.1. *Let* $K_n(x)$, δ_n, N *and* L_{nj} *be defined by* (A.1), (A.3) *and* (A.4), *respectively. Then*:

$$\bigcap_{j=1}^{N} (L_{nj} \geq cn\delta_n) \subseteq \bigcap_{x \in \mathcal{X}} [K_n(x) \geq cn\delta_n].$$

PROOF. The intervals J_{nj}, $j = 1, \dots, N$ have common length $\varepsilon\delta_n$ except, perhaps, for the interval J_{nN} whose length may be $< \varepsilon\delta_n$. To each $x \in \mathcal{X}$, assign the interval $(x - \delta_n, x + \delta_n)$ of length $2\delta_n$, and let ε be less than one, $(0 <)\varepsilon < 1$, so that $2\delta_n > \varepsilon\delta_n$. From the definition of $K_n(x)$ and L_{nj} by (A.1) and (A.4), respectively, we have that, for each fixed $j = 1, \ldots, N$ and every $x \in J_{nj}, L_{nj}$ is the number of x_i's, $i = 1, \ldots, n$ which are in J_{nj}, and $K_n(x)$ is the number of same which are in $(x - \delta_n, x + \delta_n)$. By the restrictions that $x \in J_{nj}$ and $\varepsilon < 1$, it follows that $L_{nj} \leq K_n(x)$. Therefore:

$$(L_{nj} \geq cn\delta_n) \subseteq [K_n(x) \geq cn\delta_n], \qquad x \in J_{nj}, \ j = 1, \ldots, N.$$

Hence:

$$(L_{nj} \geq cn\delta_n) \subseteq \bigcap_{x \in J_{nj}} [K_n(x) \geq cn\delta_n], \qquad j = 1, \ldots, N,$$

which implies that:

$$\text{(A.5)} \qquad \bigcap_{j=1}^{N} (L_{nj} \geq cn\delta_n) \subseteq \bigcap_{j=1}^{N} \bigcap_{x \in J_{nj}} [K_n(x) \geq cn\delta_n].$$

On the other hand, clearly,

$$\text{(A.6)} \qquad \bigcap_{j=1}^{N} \bigcap_{x \in J_{nj}} [K_n(x) \geq cn\delta_n] = \bigcap_{x \in \mathcal{X}} [K_n(x) \geq cn\delta_n].$$

Relations (A.5) and (A.6) complete the proof. □

PROOF OF LEMMA 2.2 for $d = 1$. By (A.2) and Lemma A.1, it suffices to show that:

$$Q^n \left[\bigcap_{j=1}^{N} (L_{nj} \geq cn\delta_n) \right] \to 1 \quad \text{or, equivalently,}$$

$$\text{(A.7)} \qquad Q^n \left[\bigcup_{j=1}^{N} (L_{nj} < cn\delta_n) \right] \to 0.$$

But:

$$\text{(A.8)} \qquad Q^n \left[\bigcup_{j=1}^{N} (L_{nj} < cn\delta_n) \right] \leq \sum_{j=1}^{N} Q^n(L_{nj} < cn\delta_n)$$
$$= \sum_{j=1}^{N} Q^n(L_{nj} - \mathcal{E}L_{nj} < cn\delta_n - \mathcal{E}L_{nj}),$$

and, for $j = 1, \ldots, N - 1$, and by means of the boundedness from below of the p.d.f. f_X:

$$\mathcal{E}L_{nj} = n\mathcal{E}L_{nj_1} = nP(X \in Jnj) = n\int_{J_{nj}} f_X(t)dt \geq nM_1\varepsilon\delta_n = M_1\varepsilon n\delta_n.$$

Then:

$$cn\delta_n - \mathcal{E}L_{nj} < cn\delta_n - M_1\varepsilon n\delta_n = -n\delta_n(M_1\varepsilon - c) = -\beta n\delta_n, \quad \text{where} \quad \beta = M_1\varepsilon - c > 0$$

by choosing $c < \frac{M_1\varepsilon}{2}$.

The same arguments hold for $j = N$. Therefore (A.8) yields:

$$\text{(A.9)} \qquad Q^n\left[\bigcup_{j=1}^{N}(L_{nj} < cn\delta_n)\right] \leq \sum_{j=1}^{N} Q^n(L_{nj} - \mathcal{E}L_{nj} < -\beta n\delta_n) \leq \sum_{j=1}^{N} Q^n(|L_{nj} - \mathcal{E}L_{nj}| > \beta n\delta_n).$$

However:

$$Q^n(|L_{nj} - \mathcal{E}L_{nj}| > \beta n\delta_n) = Q^n\left[\left|\frac{1}{n}\sum_{i=1}^{n}(L_{nji} - \mathcal{E}L_{nji})\right| > \beta\delta_n\right] = P(|\bar{S}_n| > \beta\delta_n),$$

where:

$$\text{(A.10)} \qquad \bar{S}_n = \frac{1}{n}\sum_{i=1}^{n}[I_{J_{nj}}(X_i) - \mathcal{E}I_{J_{nj}}(X_i)].$$

By Lemma 2.1, applied with $\varepsilon_n = \beta\delta_n$ and $\hat{M} = 1$, we get:

$$\text{(A.11)} \qquad P(|\bar{S}_n| > \beta\delta_n) \leq 6[1 + 2e^{1/2}\varphi(\nu)]^{\mu} \exp\left(-\frac{\beta^2 n\delta_n^2}{2C}\right) = 6[1 + 2e^{1/2}\varphi(\nu)]^{\mu} \exp(-C_0 n^{1-2\lambda}),$$

where $C_0 = \beta^2/2C$, provided $n^{-\lambda} \leq C\mu/\beta n$; μ and ν are as specified in the paragraph prior to Lemma 2.1. Relation (A.9) becomes, by means of (A.10), (A.11), the observation that $N < \frac{\varepsilon+1}{\varepsilon}n^{\lambda}$, and the proviso that $\mu \geq \frac{\beta}{C}n^{1-\lambda}$:

$$\text{(A.12)} \qquad Q^n\left[\bigcup_{j=1}^{N}(L_{nj} < cn\delta_n)\right] \leq \frac{6(\varepsilon+1)}{\varepsilon}[1 + 2e^{1/2}\varphi(\nu)]^{\mu} \cdot n^{\lambda}\exp(-C_0 n^{1-2\lambda}).$$

It is shown below that the right-hand side in (A.12) tends to 0, for a suitable choice of λ, subject to the requirement that $\mu \geq \frac{\beta}{C}n^{1-\lambda}$. To this end, recall that the only requirements on ε, as used here, are that $0 < \varepsilon < 1$. Next, in reference to the constant c in (2.3), choose $c < \frac{M_1\varepsilon}{2}$, for a fixed ε as above, and set $\beta = M_1\varepsilon - c(> 0)$. Also, take $\lambda = q/(2q+1)$. Then the desired convergence is equivalent to:

$$\begin{aligned} &C_3 n^{1/(2q+1)} - \lambda \log n - \mu \log[1 + C_1\varphi(\nu)] \to \infty, \\ &C_3 = C_0 \log e = \beta^2 \log e/2C. \end{aligned} \tag{A.13}$$

At this point, choose $\nu = [kn^{q/(2q+1)}]$, from some $k > 0$ (see (A.16) below), so that $\frac{1}{\nu} \le \frac{2}{k} n^{-q/(2q+1)}$ for all sufficiently large n. Also,

$$\log[1 + C_1\varphi(\nu)] \le (C_1 \log e)\varphi(\nu) = (2e^{1/2} \log e)\varphi(\nu),$$

and, by assumption and for some $C_4 > 0$:

$$\varphi(\nu) \le C_4 \nu^{-1-\delta} \le C_4 \left(\frac{2}{k}\right)^{1+\delta} n^{-q(1+\delta)/(2q+1)}.$$

Therefore (A.13) is implied by:

$$\begin{aligned} &C_3 n^{1/(2q+1)} - \lambda \log n - C_5 \mu n^{-q(1+\delta)/(2q+1)} \to \infty, \\ &C_5 = C_4(2e^{1/2} \log e) \left(\frac{2}{k}\right)^{1+\delta}. \end{aligned} \tag{A.14}$$

From $\mu = [\frac{n}{2\nu}]$, we have $\mu = \mu_n^* \cdot \frac{n}{2\nu}$ for some $0 < \mu_n^* \to 1$, and hence $\mu \ge \frac{\mu_n^*}{2k} n^{(q+1)/(2q+1)}$. Then (A.14) is implied by:

$$C_3 n^{1/(2q+1)} - \lambda \log n - C_6 \mu_n^* n^{(1-\delta q)/(2q+1)} \to \infty, \qquad C_6 = C_5/2k,$$

or

$$n^{1/(2q+1)}[C_3 - \lambda(\log n) n^{-1/(2q+1)} - C_6 \mu_n^* n^{-\delta q/(2q+1)}] \to \infty,$$

which is true. It remains to specify the range of k for which the inequality $\mu \ge \frac{\beta}{C} n^{1-\lambda}$ is satisfied for the choices of λ, ν and μ made above. From the definition of ν, we have $\nu = \nu_n^* k n^{q/(2q+1)}$ for some $0 < \nu_n^* \to 1$. Therefore $\mu = (\mu_n^*/2k\nu_n^*) n^{(q+1)/(2q+1)}$, and the inequality $\mu \ge \frac{\beta}{C} n^{1-\lambda}$ becomes:

$$\frac{\mu_n^*}{2k\nu_n^*} n^{(q+1)/(2q+1)} \ge \frac{\beta}{C} n^{1-\lambda}. \tag{A.15}$$

This inequality is satisfied by taking $\lambda = q/(2q+1)$. For this choice of λ, let $n \to \infty$ in (A.15) in order to obtain the following requirement for k, namely:

$$k \le C/2\beta. \tag{A.16}$$

To summarize: for sufficiently small $c > 0$, $\lambda = q/(2q+1)$, and ν and μ as chosen above, the right-hand side in (A.12) converges to 0, which implies (A.7). The proof of the lemma is completed. □

References

Beran, R. J. (1977). Minimum Hellinger distance estimates for parametric models, *Ann. Statist.*, **5**, 445–463.

Chanda, K. C. (1974). Strong mixing properties of linear stochastic processes, *J. Appl. Probab.*, **14**, 67–77.

Devroye, L. P. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation, *Ann. Statist.*, **8**, 231–239.

Ibragimov, I. A. and Khas'minskii, R. Z. (1980). On nonparametric estimation of regression, *Soviet Math. Dokl.*, **21**, 130–131.

Kolmogorov, A. N. and Tikhomirov, V. M. (1961). ε-entropy and ε-capacity of sets of function spaces, *Amer. Math. Soc. Transl.*, **17**, 277–364 (translation).

Le Cam, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory*, Springer, New York.

Pham, D. T. (1986). The mixing property of linear and generalized random coefficient autoregessive models, *Stochastic Process. Appl.*, **23**, 291–300.

Pham, D. T. and Tran, L. T. (1985). Some strong mixing properties of time series models, *Stochastic Process. Appl.*, **19**, 297–303.

Robinson, P. M. (1986). On the consistency and finite-sample properties of nonparametric kernel time series regression, autoregression and density estimators, *Ann. Inst. Statist. Math.*, **38**, 539–549.

Roussas, G. G. (1990). Nonparametric regression estimation under mixing conditions, *Stochastic Process. Appl.*, **36**, 107–116.

Roussas, G. G. and Ioannides, D. (1987). Moment inequalities for mixing sequences of random variables, *Stochastic Anal. Appl.*, **5**, 61–120.

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* (ed. K. Matusita), Elsevier Science (North Holland), New York.

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 estimators, *Ann. Statist.*, **10**, 1040–1053.

Tran, L. T. (1989). The L_1-convergence of kernel density estimates under dependence, *Canad. J. Statist.*, **17**, 197–208.

Tran, L. T. (1990). Kernel density estimation under dependence, *Statist. Probab. Lett.*, **10**, 193–201.

Tran, L.T. (1993). Nonparametric functional estimation for time series by local average estimators, *Ann. Statist.*, **40**, 1040–1057.

Wolfowitz, J. (1957). The minimum distance method, *Ann. Math. Statist.*, **28**, 75–88.

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy, *Ann. Statist.*, **13**, 768–774.

Yatracos, Y. G. (1989*a*). A regression type problem, *Ann. Statist.*, **17**, 1597–1607.

Yatracos, Y. G. (1989*b*). On the estimation of the derivatives of a function with the derivatives of an estimate, *J. Multivariate Anal.*, **28**, 172–175.

Yatracos, Y. G. (1992). L_1-optimal estimates for a regression type function in $\Re^d$, *J. Multivariate Anal.*, **40**, 213–220.

Yoshihara, Ken-ichi (1992). *Weakly Dependent Stochastic Sequences and Their Applications*, Vol. 1, Sanseido, Tokyo.