Ann. Inst. Statist. Math.
Vol. 56, No. 2, 265–277 (2004)

DEPENDENCE AND THE DIMENSIONALITY REDUCTION PRINCIPLE

YANNIS YATRACOS

Department of Statistics and Applied Probability, The National University of Singapore, 6 Science Drive 2, Singapore 117546, Singapore, e-mail: yatracos@stat.nus.edu.sg

(Received November 27, 2001; revised April 11, 2003)

Abstract. Stone's dimensionality reduction principle has been confirmed on several occasions for independent observations. When dependence is expressed with ϕ-mixing, a minimum distance estimate $\hat{\theta}_n$ is proposed for a smooth projection pursuit regression-type function $\theta \in \Theta$, that is either additive or multiplicative, in the presence of or without interactions. Upper bounds on the L_1-risk and the L_1-error of $\hat{\theta}_n$ are obtained, under restrictions on the order of decay of the mixing coefficient. The bounds show explicitly the additive effect of ϕ-mixing on the error, and confirm the dimensionality reduction principle.

Key words and phrases: Additive and multiplicative regression model, dimensionality reduction, projection pursuit, Kolmogorov's entropy, minimum distance estimation, nonparametric regression, ϕ-mixing, rates of convergence.

1. Introduction

A common problem in statistics is the estimation of a density $f \in \mathcal{F}$, or of a regression function $\theta \in \Theta$ with real values. Different estimates have been proposed for each of these parameters. When $\mathcal{F}$ and Θ consist of q-smooth functions with compact support $\mathcal{X}$ in R^d, Stone (1982) showed that the rate of convergence of the optimal estimates is $n^{-q/(2q+d)}$ in L_v-distance, $1 \leq v < \infty$. This rate is not satisfactory for reasonable values of n when d is large, due to the sparsity of high dimensional samples ("the curse of dimensionality"). It is then tempting to approximate $\theta(\boldsymbol{x})$ and $f(\boldsymbol{x})$ by either the sum or the product of real valued functions with the same smoothness, that are called the functional components of f or θ, having the form either $g_k(\boldsymbol{b}_k^T\boldsymbol{x})$, $k \geq 1$, or $g_j(x_{m_1}, \ldots, x_{m_{r_j}})$, $r_j < d$, $j \geq 1$; $\boldsymbol{b}_k^T\boldsymbol{x}$ denotes scalar product of the vectors $\boldsymbol{b}_k$, $\boldsymbol{x} = (x_1, \ldots, x_d)$. The model dimension r is the largest dimension of the domains of the g's. Since the g's are defined in sub-spaces of $\mathcal{X}$ with smaller dimension, the question arises if the optimal rates of convergence will be affected. Stone (1985) conjectured that, in an r-dimensional model of q-smooth densities or regression functions defined on $\mathcal{X}$, the optimal rate of convergence will be $n^{-q/(2q+r)}$ (*Stone's heuristic dimensionality reduction principle*).

The dimensionality reduction principle was confirmed on several occasions when the observations are independent: for the L_2 distance, in additive regression (Stone (1985)), in generalized additive models (Stone (1986)), in additive projection pursuit regression (Chen (1991)), in generalized regression or densities (Stone (1994)); for the L_1 distance, in additive and multiplicative regression in presence of or without interactions, and $\theta(\boldsymbol{x})$ a regression-type function, namely any parameter of the conditional distribution of

the response variable, not necessarily a conditional mean (Nicoleris and Yatracos (1997), denoted by N&Y in the sequel). In the context of quantile regression, Chaudhuri (1991*a*) pointed out that the dimensionality reduction principle is expected to be confirmed for simple additive regression models in L_p-distance, $1 \leq p \leq \infty$.

Practical considerations dictated the replacement of the assumption of independence of the observations by a suitable mode of dependence. For example, if $\{Z_j\}$ is a strictly stationary discrete time-parameter series of real-valued random variables, a problem of interest is the nonparametric estimation of the conditional expectation of Z_{j+1}, on the basis of the m previous observations $Z_{j-m+1}, \ldots, Z_j$. This is identical to estimating the regression $E(Y_j \mid X_j)$, $Y_j = Z_{j+1}$, $X_j = (Z_{j-m+1}, \ldots, Z_j)$. Recent work in nonparametric estimation of either f or θ under mixing conditions, indicates that the rates of convergence coincide with those under independence, when restrictions are imposed on the mixing coefficients; for example, see Truong and Stone (1992), Tran (1993) and Roussas and Yatracos (1996). A natural question is, whether the dimensionality reduction principle remains valid in this situation. In this work, for a smooth regression-type function θ that follows either an additive or a multiplicative model, and a ϕ-mixing sequence of observations for which the partial sums of $\{\phi(n)\}$ converge, upper bounds on the L_1-risk and the L_1-error (in probability) are obtained, for a minimum distance estimate $\hat{\theta}_n$ of θ. The upper bounds depend on Kolmogorov's entropy of the parameter space and the mixing coefficient, and confirm the dimensionality reduction principle.

2. Motivation, definitions, the models, the tools

In classical nonparametric regression, $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ are a sample of independent pairs, copies of $(\boldsymbol{X}, Y)$; Y is a real valued response, $\boldsymbol{X}$ takes values in a known compact set $\mathcal{X}$ in R^d, $d \geq 1$. Conditionally on $\boldsymbol{X}_i = \boldsymbol{x}_i$, the random variable Y_i has density $f(y \mid \boldsymbol{x}_i, \theta(\boldsymbol{x}_i))$ and $\theta(\boldsymbol{x}_i) = E(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i)$, $\theta \in \Theta$. In a regression-type problem, the regression function θ is not necessarily a conditional mean.

The following questions provided the motivation behind the regression-type problem:

1) Is there an explanation for the coincidence of the optimal rates of convergence of a density or a regression function with the same smoothness?

2) Would the same optimal rates have been observed if, other things being equal, the regression function were a quantile or another parameter of the conditional density?

In Yatracos (1985), the upper bound on the L_1-rate of convergence of a minimum distance estimate of a density f depends on the size of the parameter space $\mathcal{F}$, measured using its Kolmogorov's entropy (see below). Kolmogorov's entropies of regression and density functions with the same smoothness are of the same order. The missing link to answer both questions is that, a regression problem can be viewed as a combination of several density estimation problems, each occurring at the observed values of the independent variable. This observation is behind the form of the proposed minimum distance estimate in the regression-type problem.

To introduce the functional components of θ, let $\mathcal{X}_m$ be a compact subset in R^m; assume without loss of generality that $\mathcal{X}_m = [0,1]^m$, $m = 1, \ldots, d$, and denote $\mathcal{X}_d$ by $\mathcal{X}$. Let $\Theta_{q,m}$ be a space of q-smooth functions on $\mathcal{X}_m$ with values in a known compact G of the real line. Every $\theta \in \Theta_{q,m}$ is p-times differentiable, with the p-th derivative satisfying a Lipschitz condition with parameters (L, α); that is $|\theta^{(p)}(\boldsymbol{x}) - \theta^{(p)}(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|^{\alpha}$, $\theta^{(p)}(\boldsymbol{x})$ is any p-th order mixed partial derivative of θ evaluated at $\boldsymbol{x}$, $q = p + \alpha$, $0 < \alpha \leq 1$.

Estimates will be constructed for the models that follow, confirming the dimensionality reduction principle under restrictions pertaining the form of dependence. Our main interest is the estimation of θ rather than its functional components. Under the assumptions (A2) and (A4) (see next section), model identifiability holds; if $\theta_1 \neq \theta_2$, the L_1-distance between the joint densities $\|f(\cdot \mid \cdot, \theta_1) - f(\cdot \mid \cdot, \theta_2)\|$ is positive.

The additive super-model.

$$\theta(\boldsymbol{x}) = \sum_{j=1}^{K_1} \theta_{1j}(\boldsymbol{b}_j^T \boldsymbol{x}) + \sum_{j=1}^{K_2} \psi_j(x_{m_1}, \ldots, x_{m_{r_j}})$$

The multiplicative super-model.

$$\theta(\boldsymbol{x}) = \Pi_{j=1}^{K_1} \theta_{1j}(\boldsymbol{b}_j^T \boldsymbol{x}) \Pi_{j=1}^{K_2} \psi_j(x_{m_1}, \ldots, x_{m_{r_j}}).$$

In these models, $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$, $\theta_{1j} \in \Theta_{q,1}$, $\psi_j \in \Theta_{q,r_j}$, $\boldsymbol{b}$ is an element of the unit sphere centered at the origin, $\boldsymbol{b}^T \boldsymbol{x}$ denotes the scalar product of the vectors $\boldsymbol{b}$ and $\boldsymbol{x}$, $(m_1, \ldots, m_{r_j})$ are such that $m_i \neq m_j$ for $i \neq j$. K_1, K_2 are either known or unknown but bounded by the known constants D_1, D_2 respectively, $2 \leq r_j \leq d-1$.

Both models without interactions (the ψ's) and with K_1 not necessarily bounded, appear in Friedman and Stuetzle (1981) and in Huber (1985), and are called projection pursuit regression models (PPR); the model dimension $r = 1$. Special cases of these models appear in Stone (1982, 1985) and Chen (1991). The PPR models bypass the curse of dimensionality when $K_1 \leq D_1$, $K_2 \leq D_2$, as seen in Stone (1985, 1994) and Chen (1991) (but with $D_1 = d$). In the presence of interactions, the model dimension r is the largest dimension of the domains of the ψ's.

The distances used to define optimality of the proposed estimate, and Kolmogorov's entropy of the parameter space follow.

DEFINITION 2.1. For any two functions θ and $\tilde{\theta}$ on $\mathcal{X}$, their $L_1(d\boldsymbol{x})$ and sup-norm distances are respectively given by

$$\|\theta - \tilde{\theta}\| = \int_{\mathcal{X}} |\theta(\boldsymbol{x}) - \tilde{\theta}(\boldsymbol{x})| d\boldsymbol{x} \quad \text{and} \quad \|\theta - \tilde{\theta}\|_\infty = \sup\{|\theta(\boldsymbol{x}) - \tilde{\theta}(\boldsymbol{x})|; \boldsymbol{x} \in \mathcal{X}\}.$$

The notation $z_n \sim w_n$ denotes that $z_n \sim O(w_n)$ and $w_n \sim O(z_n)$. Θ^ϵ is an ϵ-$\tilde{\rho}$-dense subset of a metric space $(\Theta, \tilde{\rho})$, if every point in Θ is at a $\tilde{\rho}$-distance not exceeding ϵ from some point in Θ^ϵ. Kolmogorov and Tikhomirov (1959) have shown that given radius $a_n > 0$, the most economical a_n-$\|\cdot\|_\infty$-dense subset $\Theta_{q,m}^n$ of $\Theta_{q,m}$ has cardinality $N_m(a_n)$, such that $\log_2 N_m(a_n) \sim (1/a_n)^{m/q}$; $\Theta_{q,m}^n$ is *a discretization* of $\Theta_{q,m}$. The quantity $\log_2 N_m(a)$, $a > 0$, is called Kolmogorov's entropy of the space $\Theta_{q,m}$ and measures the size of the parameter space.

With a finite number n of observations we cannot estimate the unknown parameter without error, thus, without much loss, the proposed estimates will take values in a discretization of the parameter space under the model. Le Cam (1973) constructed estimates of a probability measure using discretizations based on Hellinger distance and multiple testing procedures. An extended list of references in nonparametric estimation

of either a density or a regression function under independence may be found in N&Y. The reader may consult Devroye (1987) for the properties and the use of the L_1-distance, and Le Cam and Yang (1990) for questions on estimation in abstract spaces.

The notions of optimality in probability and risk optimality follow.

DEFINITION 2.2. A sequence of estimators $\{T_n\}$ is optimal in probability for θ, with respect to the distance $\tilde{\rho}$, if there is a sequence $\{\delta_n\}$, $n = 1, 2, \ldots$, decreasing to zero such that,

$$\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{\theta} P[\tilde{\rho}(T_n, \theta) > C\delta_n] = 0 \tag{2.1}$$

$$\lim_{C\to 0} \liminf_{n\to\infty} \inf_{S_n} \sup_{\theta} P[\tilde{\rho}(S_n, \theta) > C\delta_n] = 1. \tag{2.2}$$

If only (2.1) (resp. (2.2)) holds δ_n is an upper (resp. lower) convergence rate in probability.

The sequence of estimators $\{T_n\}$ is risk-optimal with respect to $\tilde{\rho}$, with rate of convergence $\{\delta_n\}$ decreasing to zero, if there are positive constants C_L, C_U such that

$$C_L\delta_n \leq \inf\{\sup\{E\tilde{\rho}(S_n, \theta); \theta \in \Theta\}; S_n\} \tag{2.3}$$

$$\leq \sup\{E\tilde{\rho}(T_n, \theta); \theta \in \Theta\} \leq C_U\delta_n. \tag{2.4}$$

If only (2.4) (resp. (2.3)) holds, δ_n is a risk upper (resp. lower) convergence rate.

Upper convergence rates are often obtained bounding $\tilde{\rho}(T_n, \theta)$ from above with a finite sum, and an error term that will decrease to zero as the sample size increases. Bounds on the finite sum may be obtained using inequalities like Hoeffding's (1963). Lower convergence rates may be achieved using Fano's Lemma or its extension in regression (Ibragimov and Khas'minski (1981), Le Cam (1986), Yatracos (1988)).

For the regression-type problem determined by the models, upper convergence rates and lower bounds on minimax rates are obtained for the L_1-error and the L_1-risk of the proposed estimate, via the L_1-distance and the Kullback information between conditional distributions of the response variable. The definitions of both distance measures follow.

DEFINITION 2.3. For two probability measures Q, S, defined on the probability space $(\mathcal{W}, \mathcal{A})$, the L_1-distance is defined as $\|Q - S\| = 2\sup\{|Q(A) - S(A)|; A \in \mathcal{A}\}$; the Kullback information is given by $K(Q, S) = E_Q \log(dQ/dS)$, if Q is absolutely continuous with respect to S, and is equal to $+\infty$ otherwise.

The following proposition is a useful tool relating rates of convergence of estimates with those of their derivatives, and explaining why it is easier to estimate a function than its derivatives. Let $\theta^{(s)}$ be an $[s]$-th order mixed partial derivative of θ, not identically 0, $s \in R^d$, $[s] = s_1 + \cdots + s_d$. An upper bound for $\|\hat{\theta}_n^{(s)} - \theta^{(s)}\|$ is provided below.

PROPOSITION 2.1. (Yatracos (1989*b*), Proposition 2) *Let $\tilde{\theta}_n$ be an estimate of the real valued function θ. Both $\tilde{\theta}_n$, θ are defined on a compact set in R^d, have mixed partial derivatives of order p, and the p-th derivative has modulus of continuity $w(z)$, $z > 0$. Then, for $1 \leq [s] \leq p$ and $\|\ \|_v$ the L_v-distance, $1 \leq v \leq \infty$,*

$$\|\tilde{\theta}_n^{(s)} - \theta^{(s)}\|_v \leq c_1 b_n^{p-[s]} w(b_n) + c_2 b_n^{-[s]} \|\tilde{\theta}_n - \theta\|_v.$$

To introduce the notion of ϕ-mixing, let U_n be R^m-valued random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, $n = 1, 2, \ldots$. For i, j with $1 \leq i < j \leq \infty$, let $\mathcal{F}_i^j$ be the σ-field generated by the r.v.'s U_n, $i \leq n \leq j$.

DEFINITION 2.4. The not necessarily (strictly) stationary sequence $\{U_n\}$, $n \geq 1$, is ϕ-mixing with mixing coefficient $\phi(n)$ if as n increases to infinity

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

if the stochastic process is stationary the sup over k is superfluous.

The following inequality is fundamental to calculate rates of convergence of estimates under the assumption of ϕ-mixing. It is used instead of Hoeffding's (1963) inequality to bound the tails of sums of bounded random variables, and can be found in Roussas and Ioannides (1987) and Roussas and Yatracos (1996). Let $\nu = \nu(n)$ and $\mu = \mu(n) = [\frac{n}{2\nu}]$, be both positive integers tending to infinite, where $[x]$ denotes the integral part of x.

PROPOSITION 2.2. *Let Z_n, $n \geq 1$, be real valued r.v.'s centered at their expectation and bounded in absolute value by M, that are ϕ-mixing such that $\sum_{n=1}^{\infty} \phi(n) = \phi^* < \infty$, $L = 1 + 4\phi^*$, and $\bar{Z}_n = \frac{1}{n}\sum_{k=1}^{n} Z_k$. Then, for all $0 < \zeta_n \leq \frac{LM\mu}{n}$ and $n \geq 1$,*

$$P(|\bar{Z}_n| > \zeta_n) \leq 6[1 + 2e^{1/2}\phi(\nu)]^{\mu} e^{-n\zeta_n{}^2/2LM^2}.$$

3. Minimum distance estimation, the discretizations, the assumptions

The minimum distance estimation method was formalized as a principle by Wolfowitz (1957). A lot of work has been devoted ever since to this topic. In particular, under regularity conditions, it is shown that the minimum distance estimator is robust and asymptotically efficient (Beran (1977), Millar (1981), Donoho and Liu (1988*a*)). Pathologies of some minimum distance estimates for the normal model are examined in Donoho and Liu (1988*b*). The proposed minimum distance estimate $\hat{\theta}_n$ of a regression type function θ, motivated by an estimate of a density (Yatracos (1985)), has been used in Yatracos (1989*a*, 1992), Roussas and Yatracos (1996) and N&Y. For the reasons mentioned in the previous section, $\hat{\theta}_n$ takes values in a discretization of the parameter space.

The discretization of the parameter space, for the additive and multiplicative super models, is obtained using $a_n - \|\cdot\|_\infty$ discretizations of the spaces of the functional components of θ, and $n^{-1/2} - \|\cdot\|_\infty$ discretizations of the unit sphere for the unknown vector parameters $\boldsymbol{b}_j$ in the projection pursuit model components, $j = 1, \ldots, K_1$; the estimates of the $\boldsymbol{b}_j$'s are included in $\hat{\theta}_n$. The cardinalities of the so obtained $c(a_n + n^{-1/2}) - \|\cdot\|_\infty -$ dense subsets have the same order $N_1^{D_1}(a_n) n^{dD_1/2} \Pi_{j=1}^{D_2} N_{r_j}(a_n)$; for the details see N&Y. In the sequel, denote by Θ^n any of these discretizations and by $N(a_n)$ their cardinalities; with abuse of notation, a_n is used to denote $a_n + n^{-1/2}$.

Given $\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n$, define the sets

$$A_{k,m,i} = \{y : f(y \mid \boldsymbol{x}_i, \theta_k(\boldsymbol{x}_i)) > f(y \mid \boldsymbol{x}_i, \theta_m(\boldsymbol{x}_i))\}, \qquad \theta_k \in \Theta^n,$$
$$\theta_m \in \Theta^n, \qquad 1 \le k < m \le N(a_n), \qquad i = 1, \ldots, n,$$

to be used in the definition of the minimum distance estimate. $I_A(x) = 1$, if $x \in A$, and it is 0 otherwise.

DEFINITION 3.1. Let $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ be observations according to one of the models already described. The minimum distance estimator $\hat{\theta}_n$ of θ satisfies the relation

$$\sup_{1 \le k < m \le N(a_n)} \left\{ \frac{1}{n} \left| \sum_{i=1}^{n} (I_{A_{k,m,i}}(Y_i) - P_{\hat{\theta}_n(\boldsymbol{x}_i)}(A_{k,m,i})) \right| \right\}$$
$$= \inf_{1 \le r \le N(a_n)} \sup_{1 \le k < m \le N(a_n)} \left\{ \frac{1}{n} \left| \sum_{i=1}^{n} (I_{A_{k,m,i}}(Y_i) - P_{\theta_r(\boldsymbol{x}_i)}(A_{k,m,i})) \right| \right\}.$$

As a special case, when $\boldsymbol{x}_i = \boldsymbol{x}$, $i = 1, \ldots, n$, the L_1-minimum distance estimate of a density is obtained (Yatracos (1985)).

The following assumptions are made:

(A1) $\{(\boldsymbol{X}_n, Y_n)\}$ is a stationary sequence of observations that is ϕ-mixing and

$$\sum_{n=1}^{\infty} \phi(n) < \infty.$$

(A2) $c_1|t - s| \le \|f(\cdot \mid \boldsymbol{x}, t) - f(\cdot \mid \boldsymbol{x}, s)\| \le c_2|t - s|$;
c_1, c_2 are constants greater than zero, independent of $\boldsymbol{x}$, $\|\cdot\|$ is the L_1-distance of the conditional densities and t, s take real values in the compact G where the elements of $\Theta_{q,m}$ take values.

(A3) The form of the conditional density $f(y \mid \boldsymbol{x}, \theta(\boldsymbol{x}))$ is known.

(A4) The density $g(\boldsymbol{x})$ of $\boldsymbol{X}$ is bounded below and above, by the positive, finite constants A and B, respectively.

(A5) For every s, t, possible values of $\theta(\boldsymbol{x})$, with P_s denoting the probability measure with density $f(y \mid \boldsymbol{x}, s)$ and $c > 0$, it holds

$$K(P_s, P_t) \le c(s - t)^2.$$

Assumptions (A2)–(A4) are used to construct the proposed minimum distance estimate, and to calculate upper convergence rates. (A2) allows interchanging the distance between parameters with that of the corresponding conditional distributions. Without (A3), we cannot obtain the sets $A_{k,m,i}$ used in the minimum distance criterion. This is the price to be paid in a regression-type problem, since the nature of the parameter θ in the conditional density is unknown, and one cannot determine the functional of the Y's that should be used to estimate θ. Similar assumptions can be found also in classical regression; for example, in Stone (1994), with the conditional densities assumed to be either Bernoulli or Poisson, and in Donoho, Johnstone, Kerkyacharian and Picard (1995), with the errors assumed to be normal. (A4) has been used in Chaudhuri (1991*b*), Chen

(1991), Stone (1982, 1985, 1994) and by several other authors. In the calculations of upper convergence rates, this assumption allows us to pass, without much loss, from the L_1-distance $\|\hat{\theta}_n - \theta\|$ to the expectation $E|\hat{\theta}_n(\boldsymbol{X}) - \theta(\boldsymbol{X})|$. From (A1), convergence of the partial sums of $\{\phi(n)\}$ is used to prove that $E|\hat{\theta}_n(\boldsymbol{X}) - \theta(\boldsymbol{X})|$ can be approximated almost surely by a sum of random variables uniformly in θ, and leads to convergence rates confirming the dimensionality reduction principle. Without (A5), the lower convergence rates obtained for independent samples may not coincide with the upper convergence rates. (A2) and (A5) hold in several of the assumed models for the Y's (Yatracos (1988, 1989*a*)).

4. Rates of convergence

The steps to bound $\|\hat{\theta}_n - \theta\|$ with a finite sum have been described in the previous section; Proposition 4.1 is crucial for the approximation.

Let Θ be the parameter space for the regression-type problem, and Θ^n the $c(a_n + n^{-1/2}) - \|\cdot\|_\infty-$ dense subset with cardinality $N(a_n)$ defined in the previous section, with the sequence $\{a_n\}$ decreasing to 0 and $c > 0$; a_n to be determined such that $\hat{\theta}_n$ is optimal. Let $\tilde{\theta}_n$ be an estimate of the unknown regression-type function θ, with values in Θ^n, and θ_k, θ_m be elements of Θ^n. Let $P_{\theta(\boldsymbol{x}_i)}$ be a probability measure with density $f(y \mid \boldsymbol{x}_i, \theta(\boldsymbol{x}_i))$, $i = 1, \ldots, n$; Q is the distribution of any of the $\boldsymbol{X}$'s; Q^n and Q^∞ denote respectively the joint distribution of $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ and the distribution of the infinite vector of the $\boldsymbol{X}$'s. Define the quantities:

$$E|\theta_k - \theta_m| = E|\theta_k(\boldsymbol{X}) - \theta_m(\boldsymbol{X})|, \qquad E_n|\theta_k - \theta_m| = \frac{1}{n}\sum_{i=1}^{n} |\theta_k(\boldsymbol{X}_i) - \theta_m(\boldsymbol{X}_i)|,$$

$$\Delta_n(\theta_k, \theta_m) = \left| E|\theta_k(\boldsymbol{X}) - \theta_m(\boldsymbol{X})| - \frac{1}{n}\sum_{i=1}^{n} |\theta_k(\boldsymbol{X}_i) - \theta_m(\boldsymbol{X}_i)| \right|,$$

$$A_n(\varepsilon_n, m) = \bigcup_{k=1}^{N(a_n)} \{(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n) : \Delta_n(\theta_k, \theta_m) > c\varepsilon_n\},$$

$$A_n(\varepsilon_n) = \bigcup_{m=1}^{N(a_n)} A_n(\varepsilon_n, m).$$

From now on, the letters $C, C_1, C_2 \ldots$ will denote generic, positive constants, independent of n.

PROPOSITION 4.1. *Let $\tilde{\theta}_n$, $A_n(\varepsilon_n)$, $A_n(\varepsilon_n, m)$, $\Delta_n(\theta_k, \theta_m)$ be defined as above, for a regression-type problem following either the additive or the multiplicative super-model, and with observations satisfying* (A1),

$$c\varepsilon_n = \left(2LM^2c^* \frac{\log N(a_n)}{n}\right)^{1/2} \downarrow 0, \qquad c^* > 0, \qquad \textit{and}$$

$$\nu = \left[\left(w\frac{n}{\log N(a_n)}\right)^{1/2}\right],$$

with $w \le \frac{L}{18c^*}$ and L, M, ν all defined in Proposition 2.2.

Then,

a) $Q^n(A_n(\varepsilon_n, m)) \leq 6N(a_n)^{1.5-c^*}$,

b) $P[\{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) : \Delta_n(\tilde{\theta}_n, \theta_m) > c\varepsilon_n\}] \leq Q^n(A_n(\varepsilon_n, m))$,

c) *For* $c^* > 1.5$ *it holds* $\sum_{n=1}^{\infty} N(a_n)^{1.5-c^*} < \infty$, *and therefore,*

$$P[\liminf\{\Delta_n(\tilde{\theta}_n, \theta_m) \leq c\varepsilon_n\}] = 1.$$

Thus, there is a set of probability 1 *such that* $\Delta_n(\tilde{\theta}_n, \theta_m) \leq c\varepsilon_n$ *almost surely.*

d) $Q^n(A_n(\varepsilon_n)) \leq 6N(a_n)^{2.5-c^*}$,

e) $P[\bigcup_{m=1}^{N(a_n)}\{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) : \Delta_n(\tilde{\theta}_n, \theta_m) > c\varepsilon_n\}] \leq Q^n(A_n(\varepsilon_n))$,

f) *For* $c^* > 2.5$ *it holds* $\sum_{n=1}^{\infty} N(a_n)^{2.5-c^*} < \infty$, *and therefore*

$$P\left[\liminf\left(\bigcup_{m=1}^{N(a_n)}\{\Delta_n(\tilde{\theta}_n, \theta_m) > c\varepsilon_n\}\right)^c\right] = 1.$$

Thus, there is a set of probability 1 *such that for every* θ *in* Θ *and its nearest neighbor* θ_m *it holds* $\Delta_n(\tilde{\theta}_n, \theta_m) \leq c\varepsilon_n$.

The proposition which shows explicitly the additive effect of the dependence on the error $\|\hat{\theta}_n - \theta\|$ follows.

PROPOSITION 4.2. *Let* $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ *be a sample in a regression-type problem for which assumptions* (A1)–(A4) *hold, and* θ *follows either the additive or the multiplicative super-model. The random vectors* $\boldsymbol{X}_i$ *take values in* $[0,1]^d$, $d \geq 1$, *and* Y_i *are the corresponding real valued responses,* $i = 1, \ldots, n$. *Then, the minimum distance estimate* $\hat{\theta}_n$ *is uniformly consistent with upper rate* δ_n, *in* L_1*-distance, such that*

$$\delta_n \sim a_n + \left(\frac{\log N(a_n)}{n}\right)^{1/2} + \left(\frac{\log[1 + C_1\phi(\nu)]}{\nu}\right)^{1/2},$$

where $C_1 = 2e^{1/2}$, $\nu = [(w\frac{n}{\log N(a_n)})^{1/2}]$ *and* $w \leq \frac{L}{72}$; L *appears in Proposition* 2.2.

If the unknown parameter θ does not follow exactly one of the two models, let θ^* be its closest approximation in the chosen model such that $\|\theta - \theta^*\|_\infty < \epsilon$. Following Proposition 2 in Yatracos (1985) and (A2), it is easy to see that $\|\hat{\theta}_n - \theta\| \leq c_1\epsilon + c_2\delta_n$, with δ_n as in Proposition 4.2.

The corollaries below, confirm the dimensionality reduction principle for the L_1-error and the L_1-risk in estimating θ by $\hat{\theta}_n$, and $\theta^{(s)}$ by $\hat{\theta}_n^{(s)}$.

COROLLARY 4.1. *Under the assumptions in Proposition* 4.2 *and* (A5), *the* L_1*-rate of convergence in probability* δ_n *is no slower than the optimal rate under independence:*

a) *for the additive and the multiplicative super model with interactions*

$$\delta_n \sim n^{-q/(2q+r)},$$

where r *is the model dimension*;

b) *for the additive and multiplicative super model without interactions*

$$\delta_n \sim n^{-q/(2q+1)}.$$

In both a) *and* b), *the rate of convergence of* $\hat{\theta}_n^{(s)}$ *to* $\theta^{(s)}$ *is no slower than* $n^{-(q-[s])/(2q+r)}$, $r \geq 1$.

COROLLARY 4.2. *Under the assumptions in Proposition* 4.2 *and* (A5), *the risk of* $\hat{\theta}_n$ *converges to* 0 *at the optimal rate obtained for the same model under independence; the upper convergence rates in probability hold almost surely.*

5. Concluding remarks

The dimensionality reduction principle has been confirmed for regression-type functions that follow the additive and the multiplicative super-models, when dependence is expressed in terms of ϕ-mixing. The principle is expected to hold for the minimum distance estimate $\hat{\theta}_n$ under other forms of dependence, as for example α-mixing, if an appropriate exponential bound becomes available for the tails of the sums of random variables with the assumed type of dependence.

Acknowledgements

Many thanks are due to anonymous referees and an Associate Editor for various useful suggestions that helped improving the presentation of this work.

Appendix

PROOF OF PROPOSITION 4.1. a) For $\varepsilon_n \downarrow 0$ it holds,

$$
\begin{aligned}
(\text{A.1}) \qquad Q^n(A_n(\varepsilon_n, m)) &= Q^n\left(\bigcup_{k=1}^{N(a_n)} \{(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : \Delta_n(\theta_k, \theta_m) > c\varepsilon_n\}\right) \\
&\leq \sum_{k=1}^{N(a_n)} Q^n[(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : \Delta_n(\theta_k, \theta_m) > c\varepsilon_n] \\
&\leq N(a_n) \sup_{1 \leq k \leq N(a_n)} Q^n[(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) : \Delta_n(\theta_k, \theta_m) > c\varepsilon_n] \\
&\leq 6N(a_n)[1 + 2e^{1/2}\phi(\nu)]^{n/2\nu} e^{-nc^2\epsilon_n^2/2LM^2},
\end{aligned}
$$

provided $0 < c\epsilon_n \leq \frac{LM\mu}{n}$.

The last inequality in (A.1) was obtained using Proposition 2.2; M denotes the bound of the difference of functions in $\Theta_{q,d}$. Note that μ is the largest integer such that $2\nu\mu \leq n$; for large n, $3\nu\mu = 2\nu\frac{3}{2}\mu > n$ or $\frac{\mu}{n} > \frac{1}{3\nu}$. It will be enough then to have $0 < c\epsilon_n \leq \frac{LM}{3\nu}$.

Let $C_1 = 2e^{1/2}$ and take $\nu = [(w\frac{n}{\log N(a_n)})^{1/2}]$, with w satisfying (A.4). We have then for large n,

$$
\frac{1}{2}\left(w\frac{n}{\log N(a_n)}\right)^{1/2} \leq \nu \leq \left(w\frac{n}{\log N(a_n)}\right)^{1/2}
$$

or

$$
(\text{A.2}) \qquad \frac{1}{\nu^2} \leq \frac{4}{w}\frac{\log N(a_n)}{n} \quad \text{and} \quad \left(\frac{\log N(a_n)}{wn}\right)^{1/2} \leq \frac{1}{\nu}.
$$

It follows that

$$\frac{n}{2\nu}\log[1+C_1\phi(\nu)] \le \frac{n}{2\nu^2}C_1\nu\phi(\nu).$$

Using (A.2) and since from (A1) $\phi(\nu)=o(\nu^{-1})$, for sufficiently large n it holds

$$\frac{n}{2\nu}\log[1+C_1\phi(\nu)] \le .5\log N(a_n). \tag{A.3}$$

Choosing in (A.1) $c\varepsilon_n = (2LM^2c^*\frac{\log N(a_n)}{n})^{1/2}$, and using (A.3) we obtain

$$Q^n(A_n(\varepsilon_n,m)) \le 6N(a_n)^{1.5-c^*};$$

the constant c^* may be chosen to be greater than 1.5.

From (A.2), the condition $0 < c\epsilon_n \le \frac{LM}{3\nu}$ holds if

$$0 \le \left(2LM^2c^*\frac{\log N(a_n)}{n}\right)^{1/2} \le \frac{LM}{3w^{1/2}}\left(\frac{\log N(a_n)}{n}\right)^{1/2}$$

or

$$w \le \frac{L}{18c^*}. \tag{A.4}$$

b) $P[(\boldsymbol{X}_1,Y_1),\dots,(\boldsymbol{X}_n,Y_n):\Delta_n(\tilde\theta_n,\theta_m)>c\varepsilon_n]$

$$= P\left[\bigcup_{k=1}^{N(a_n)}\{(\boldsymbol{X}_1,Y_1),\dots,(\boldsymbol{X}_n,Y_n):\tilde\theta_n=\theta_k,\Delta_n(\tilde\theta_n,\theta_m)>c\varepsilon_n\}\right]$$

$$= P\left[\bigcup_{k=1}^{N(a_n)}\{(\boldsymbol{X}_1,Y_1),\dots,(\boldsymbol{X}_n,Y_n):\tilde\theta_n=\theta_k,\Delta_n(\theta_k,\theta_m)>c\varepsilon_n\}\right]$$

$$\le P\left[\bigcup_{k=1}^{N(a_n)}\{(\boldsymbol{X}_1,Y_1),\dots,(\boldsymbol{X}_n,Y_n):\Delta_n(\theta_k,\theta_m)>c\varepsilon_n\}\right] = Q^n(A_n(\varepsilon_n,m)),$$

since $\Delta_n(\theta_k,\theta_m)$ does not depend on the Y's.

c) $P[\Delta_n(\tilde\theta_n,\theta_m)>c\varepsilon_n,\ \text{infinitely often } n]$

$$\le \lim_{k\to\infty}\sum_{n=k}^{\infty}P[(\boldsymbol{X}_1,Y_1),\dots,(\boldsymbol{X}_n,Y_n):\Delta_n(\tilde\theta_n,\theta_m)>c\varepsilon_n]$$

$$\le \lim_{k\to\infty}\sum_{n=k}^{\infty}Q^n(A_n(\varepsilon_n,m)) \le 6\lim_{k\to\infty}\sum_{n=k}^{\infty}N(a_n)^{1.5-c^*}=0.$$

The proofs of (d), (e) and (f) are analogous to those of (a), (b), (c), with $N^2(a_n)$ in (d) instead of $N(a_n)$ as an upper bound in an inequality derived as (A.1).

PROOF OF PROPOSITION 4.2. The goal is to derive an upper bound for $\|\hat\theta_n-\theta\|$. Let θ_r be the nearest neighbor of θ in Θ^n, and $\hat\theta_n$ be the minimum distance estimate of θ. Then we have:

$$\int_{\mathcal{X}}|\hat\theta_n(\boldsymbol{x})-\theta(\boldsymbol{x})|d\boldsymbol{x} \le C_2(a_n+n^{-1/2})+\int_{\mathcal{X}}|\hat\theta_n(\boldsymbol{x})-\theta_r(\boldsymbol{x})|d\boldsymbol{x},$$

and using assumption (A4)

$$\leq C_2(a_n + n^{-1/2}) + A^{-1}E|\hat{\theta}_n - \theta_r| \leq C_2(a_n + n^{-1/2}) + C_3\epsilon_n + E_n|\hat{\theta}_n - \theta_r| \quad \text{a.s.}$$

by Proposition 4.1 (f), with $\varepsilon_n \sim (\frac{\log N(a_n)}{n})^{1/2}$; from the definition of $\hat{\theta}_n$

$$\leq C_2(a_n + n^{-1/2}) + C_3\epsilon_n + C_4 \sup_{1\leq k<m\leq N(a_n)} \left\{ \frac{1}{n} \left| \sum_{i=1}^{n} (I_{A_{k,m,i}}(Y_i) - P_{\theta_r(x_i)}(A_{k,m,i})) \right| \right\}$$

and using assumption (A2)

$$\text{(A.5)} \qquad \leq C_5 a_n + C_2 n^{-1/2} + C_3\epsilon_n + C_6 \sup_{1\leq k<m\leq N(a_n)} \left\{ \frac{1}{n} \left| \sum_{i=1}^{n} (I_{A_{k,m,i}}(Y_i) - P_{\theta(x_i)}(A_{k,m,i})) \right| \right\}.$$

Let $\gamma_n = (\delta_n - C_5 a_n - C_2 n^{-1/2} - C_3\epsilon_n)/C_6$, and denote by P_θ the conditional probability of the Y's given the $\boldsymbol{X}$'s. Using Proposition 2.2 it holds:

$$\text{(A.6)} \quad P[\|\hat{\theta}_n - \theta\| > \delta_n] \leq E_{Q^n} P_\theta \left[\sup_{1\leq k<m\leq N(a_n)} \left\{ \frac{1}{n} \left| \sum_{i=1}^{n} (I_{A_{k,m,i}}(Y_i) - P_{\theta(x_i)}(A_{k,m,i})) \right| \right\} > \gamma_n \right]$$
$$\leq 6N(a_n)^2[1 + C_1\phi(\nu)]^{n/2\nu} e^{-n\gamma_n^2/2L}$$

for $0 < \gamma_n \leq \frac{L\mu}{n}$ (the bound M of Proposition 2.2 in this case is 1).

For (A.6) to converge to 0 as n increases, it is enough that

$$\text{(A.7)} \qquad \frac{n\gamma_n^2}{2L} - 2\log N(a_n) - \frac{n}{2\nu}\log[1 + C_1\phi(\nu)] \to \infty.$$

Taking $\delta_n = \rho\sqrt{2L}[\frac{\log N(a_n)}{n} + \frac{\log[1+C_1\phi(\nu)]}{2\nu}]^{1/2} + C_5 a_n + C_3\epsilon_n + C_2 n^{-1/2}$ we obtain $\gamma_n = \rho\sqrt{2L}[\frac{\log N(a_n)}{n} + \frac{\log[1+C_1\phi(\nu)]}{2\nu}]^{1/2}$, and for (A.7) to hold it is enough that

$$\rho^2 \left[\log N(a_n) + \frac{n\log[1 + C_1\phi(\nu)]}{2\nu} \right] - 2\log N(a_n) - \frac{n\log[1 + C_1\phi(\nu)]}{2\nu} \to \infty,$$

or $\rho^2 > 2$, $\rho > 0$.

For $0 < \gamma_n \leq \frac{L\mu}{n}$ it suffices that $0 < \gamma_n < \frac{L}{3\nu}$ or by (A.2), (A.3) that

$$\text{(A.8)} \qquad w \leq \frac{L}{36\rho^2}.$$

Therefore, for the proposition to hold, it is enough from (A.4) and (A.8) to have:

$$\rho > 0, \qquad \rho^2 > 2, \qquad c^* > 2.5, \qquad w \leq \min\left\{ \frac{L}{18c^*}, \frac{L}{36\rho^2} \right\} < \frac{L}{72}.$$

Using the inequality

$$\frac{a^{1/2} + b^{1/2}}{\sqrt{2}} \leq (a + b)^{1/2} \leq a^{1/2} + b^{1/2},$$

it holds that

$$\delta_n \sim a_n + \left(\frac{\log N(a_n)}{n}\right)^{1/2} + \left(\frac{\log[1 + C_1\phi(\nu)]}{\nu}\right)^{1/2}.$$

PROOF OF COROLLARY 4.1. When the observations are independent, lower convergence rates for a regression-type problem are obtained in Yatracos (1998). Both a) and b) follow from (A.3), choosing $a_n = (\frac{\log N(a_n)}{n})^{1/2}$.

The convergence rates for the derivatives follow from the rate of $\|\hat{\theta}_n - \theta\|$ and Proposition 2.1, with $b_n \sim a_n^{1/q}$.

PROOF OF COROLLARY 4.2. With the chosen value for δ_n and (A.3) it holds

$$\begin{aligned} P[\|\hat{\theta}_n - \theta\| > \delta_n] &\leq 6[N(a_n)]^{2-\rho^2}[1 + C_1\phi(\nu)]^{(1-\rho^2)(n/2\nu)} \\ &\leq 6[N(a_n)]^{2.5-1.5\rho^2}. \end{aligned} \tag{A.9}$$

For the L_1-risk, truncating at δ_n we obtain for ρ large enough

$$E\|\hat{\theta}_n - \theta\| \leq \delta_n + C_7 P[\|\hat{\theta}_n - \theta\| > \delta_n] \leq C_8\delta_n.$$

From (A.9) and ρ large enough, it also holds that $\sum_{n=1}^{\infty}[N(a_n)]^{2.5-1.5\rho^2} < \infty$. Thus, the rate of convergence δ_n of the L_1-error holds almost surely.

REFERENCES

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

Chaudhuri, P. (1991*a*). Global nonparametric estimation of conditional quantile functions and their derivatives, *Journal of Multivariate Analysis*, **39**, 246–269.

Chaudhuri, P. (1991*b*). Nonparametric estimates of regression quantiles and their local Bahadur representation, *Annals of Statistics*, **19**, 760–777.

Chen, H. (1991). Estimation of a projection-pursuit regression model, *Annals of Statistics*, **19**, 142–157.

Devroye, L. P. (1987). *A Course in Density Estimation*, Birkhauser, Boston.

Donoho, D. L. and Liu, R. C. (1988*a*). The "automatic" robustness of minimum distance functionals, *Annals of Statistics*, **16**, 552–586.

Donoho, D. L. and Liu, R. C. (1988*b*). Pathologies of some minimum distance estimators, *Annals of Statistics*, **16**, 587–608.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia?, *Journal of the Royal Statistical Society, Series B*, **57**, 301–369.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association*, **76**, 817–823.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association*, **58**, 13–31.

Huber, P. J. (1985). Projection pursuit, *Annals of Statistics*, **13**, 435–475.

Ibragimov, I. A. and Khas'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*, Springer, New York.

Kolmogorov, A. N. and Tikhomirov, V. M. (1959). ϵ-entropy and ϵ-capacity of sets in function spaces, *Uspekhi Matematicheskikh Nauk*, **14**(2), 3–86 (in Russian) ((1961). *American Mathematical Society Translations (2)*, **17**, 277–364).

Le Cam, L. M. (1973). Convergence of estimates under under dimensionality restrictions, *Annals of Statistics*, **1**, 38–53.

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

Le Cam, L. M. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts*, Springer, New York.

Millar, P. W. (1981). Robust estimation via minimum distance methods, *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **55**, 73–89.

Nicoleris, T. and Yatracos, Y. G. (1997). Rates of convergence of estimates, Kolmogorov's entropy and the dimensionality reduction principle in regression, *Annals of Statistics*, **25**, 2493–2511.

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

Roussas, G. G. and Yatracos, Y. G. (1996). Minimum distance regression-type estimates with rates under weak dependence, *Annals of the Institute of Statistical Mathematics*, **48**, 267–281.

Stone, C. J. (1982). Optimal global rates of convergence in nonparametric regression, *Annals of Statistics*, **10**, 1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models, *Annals of Statistics*, **13**, 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models, *Annals of Statistics*, **14**, 590–606.

Stone, C. J. (1994). The use of polynomial splines and their tensor product in multivariate function estimation, *Annals of Statistics*, **22**, 118–184.

Tran, L. T. (1993). Nonparametric function estimation for time series by local average estimators, *Annals of Statistics*, **21**, 1040–1057.

Truong, Y. K. and Stone, C. J. (1992). Nonparametric function estimation involving time series, *Annals of Statistics*, **20**, 77–97.

Wolfowitz, J. (1957). The minimum distance method, *Annals of Mathematical Statistics*, **28**, 75–88.

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

Yatracos, Y. G. (1988). A lower bound on the error in nonparametric regression type problems, *Annals of Statistics*, **16**, 1180–1187.

Yatracos, Y. G. (1989*a*). A regression type problem, *Annals of Statistics*, **17**, 1597–1607.

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

Yatracos, Y. G. (1992). L_1-optimal estimates for a regression type function in R^d, *Journal of Multivariate Analysis*, **40**, 213–220.