The Annals of Statistics
1997, Vol. 25, No. 6, 2493–2511

RATES OF CONVERGENCE OF ESTIMATES, KOLMOGOROV'S ENTROPY AND THE DIMENSIONALITY REDUCTION PRINCIPLE IN REGRESSION[1]

BY THEODOROS NICOLERIS AND YANNIS G. YATRACOS

Université de Montréal

L_1-optimal minimum distance estimators are provided for a projection pursuit regression type function with smooth functional components that are either additive or multiplicative, in the presence of or without interactions. The obtained rates of convergence of the estimate to the true parameter depend on Kolmogorov's entropy of the assumed model and confirm Stone's heuristic dimensionality reduction principle. Rates of convergence are also obtained for the error in estimating the derivatives of a regression type function.

1. Introduction. Let $(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)$ be a random sample of n independent pairs, copies of $(\mathbf{X}, Y)$ with density $f(\mathbf{x}, y, \theta)$ in a regression setup. The random vector $\mathbf{X}$ belongs to a compact subset $\mathscr{X}$ in R^d, $d \geq 1$, and Y is the corresponding real-valued response. Assume without loss of generality that $\mathscr{X} = [0, 1]^d$ and that conditionally on $\mathbf{X_1} = \mathbf{x_1}, \ldots, \mathbf{X_n} = \mathbf{x_n}$, the random variables $Y_1, \ldots, Y_n$ are independent, each having density $f(y \mid \mathbf{x_i}, \theta(\mathbf{x_i}))$, $i = 1, \ldots, n$. The unknown function θ is an element of $\Theta_{q,d}$, the space of q-smooth real-valued functions in $\mathscr{X}$ (defined in Section 2).

For the classical regression model, when $\theta(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, $\theta \in \Theta_{q,d}$, optimal consistent estimators for θ based on local polynomials have been constructed by Stone (1982) with respect to an L_v-distance, $1 \leq v \leq \infty$. Truong (1989) and Chaudhuri (1991) provided optimal estimators when θ is a local median and a quantile of the conditional density, respectively. Truong and Stone (1994) provided optimal local polynomial estimators for θ for a stationary time series in L_2-distance and pointwise. The rate of convergence of the optimal estimator of θ in all previously mentioned cases is $n^{-q/(2q+d)}$.

In Yatracos (1989a, 1992) it is only assumed that $\theta(\mathbf{x})$ is a parameter of the conditional density without specifying its statistical interpretation, whether for example it is either a mean or a median. Therefore, $\theta(\mathbf{x})$ will hereafter be named a *regression type function* and the corresponding estimation problem *a regression type problem* that may be regarded as a combination of several density estimation problems, each occurring at the observed value of the inde-

Received July 1994; revised March 1997.

[1]Supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

AMS 1991 *subject classifications*. Primary 62J02, 62G20; secondary 62G05, 62G30.

Key words and phrases. Nonparametric regression, optimal rates of convergence, Kolmogorov's entropy, Hoeffding's inequality, dimensionality reduction principle, additive and multiplicative regression, projection pursuit, interactions, model selection.

pendent variable. The space $\Theta_{q,d}$ is discretized and an L_1-optimal estimator is constructed using a minimum distance criterion. Under mild assumptions the estimator converges at the rate $n^{-q/(2q+d)}$. This rate depends on the dimensionality of $\Theta_{q,d}$, expressed via discretization in terms of Kolmogorov's entropy (defined in Section 2). In Roussas and Yatracos (1996) it is assumed that, for $n = 1, 2, \ldots,$ $\{(\mathbf{X_n}, Y_n)\}$ is a stationary sequence of observations that is ϕ-mixing and minimum distance estimators are provided for the regression type function θ. The upper bound on the L_1-error depends on Kolmogorov's entropy and the mixing coefficient ϕ and under suitable conditions on ϕ is of the order $n^{-q/(2q+d)}$ as in the independent case.

As can be noted from these results, the rate of convergence of the optimal estimator in a nonparametric regression problem depends on the dimension d of the space $\mathscr{X}$ and the smoothness q. The lower the dimension of $\mathscr{X}$, the better the rates that are achieved, for the same amount of smoothness. For a model that is restricted to a smaller class of functions the question that arises is whether the rates of convergence will be affected, as far as the dimension is concerned. Such a model occurs, for example, if $\theta(\mathbf{x})$ is the sum of functions with the same smoothness, defined in spaces with lower dimension than $\mathscr{X}$. These functions are called the *functional components* of θ [Stone (1985)]. The *dimension* r of such a model is the *largest dimension of the functional components of* θ. Stone (1985) conjectured that in an r-dimensional model of q-smooth regression functions defined on $\mathscr{X}$ the optimal rate of convergence will be of the form $n^{-q/(2q+r)}$ (*Stone's heuristic dimensionality reduction principle*). Thinking in terms of Kolmogorov's entropy of the parameter space for such a model, one sees clearly the intuition behind the principle.

In the context of classical regression, Stone (1985) examined the L_2-rates of convergence of estimators of a completely additive regression function defined in $\mathscr{X}$. For a pair of random variables $(\mathbf{X}, Y)$ such that $\mathbf{X} = (X_1, \ldots, X_d)$, consider the regression function θ of Y on $\mathbf{X}$. Suppose that Y is real-valued with mean μ and finite variance. Let $\theta^*(\mathbf{x}) = \mu + \theta_1^*(x_1) + \cdots + \theta_d^*(x_d)$ be chosen to minimize $E(\theta(\mathbf{X}) - \theta^*(\mathbf{X}))^2$, subject to the constraints $E\theta_i^*(X_i) = 0,\ i = 1, \ldots, d$. Spline estimates of θ_i^* and of its derivatives were then constructed based on a random sample having the same distribution as $(\mathbf{X}, Y),\ i = 1, \ldots, d$. When $\theta_i^* \in \Theta_{q,1},\ i = 1, \ldots, d$, these estimates achieve the optimal rate of convergence $n^{-q/(2q+1)}$, confirming the dimensionality reduction principle in a special case $(r = 1)$.

Similar results in an additive projection pursuit type model have been obtained by Chen (1991). In this model, the conditional mean $\theta(\mathbf{x})$ of Y given $\mathbf{X} = \mathbf{x}$, is the sum of no more than d smooth functions of $\mathbf{b_i}^T\mathbf{x}$, where $\mathbf{b_i}$ is an element of the unit sphere centered at the origin, $i = 1, \ldots, n$. Based on a sample of size n from the distribution of $(\mathbf{X}, Y)$, estimators were constructed using a least squares polynomial spline and a prediction error criterion. Under some assumptions on θ, including q-smoothness, the estimates achieve the best rate of convergence $n^{-q/(2q+1)}$.

Stone (1994) provided estimates for a smooth function θ and its components; θ is either a regression, a generalized regression, a density or a con-

ditional density. The estimates are sums of tensor products of polynomial splines obtained via least squares or maximum likelihood methods. The function θ follows (or can be approximated by) a *specified* r-dimensional *hierarchical* additive model with interactions. The estimates of θ and its components achieve the L_2-optimal rate of convergence $n^{-q/(2q+r)}$, subject to the restriction $0.5r \le q$ that holds for all but the regression case.

In this paper minimum distance estimators for r-dimensional models of q-smooth functions will be constructed. The model need not be hierarchical and there is no restriction for q and r. The regression type function is of the projection pursuit type, either additive or multiplicative in the presence of or without interactions. The L_1-rate of convergence of the estimator to the true parameter θ depends *on the dimension of the model* via Kolmogorov's entropy, thus confirming Stone's heuristic dimensionality reduction principle.

2. Definitions, the models, the assumptions. Related methods. Let $\mathscr{X}_m$ be a compact subset in R^m and assume without loss of generality that $\mathscr{X}_m = [0,1]^m,\ m = 1, \ldots, d$. Let $\Theta_{q,m}$ be a space of q-smooth functions in $\mathscr{X}_m$ with values in a known compact G of the real line. Every $\theta \in \Theta_{q,m}$ is p-times differentiable, with the pth derivative satisfying a Lipschitz condition with parameters (L, α); that is, $| \theta^{(p)}(\mathbf{x}) - \theta^{(p)}(\mathbf{y}) | \le L\|\mathbf{x} - \mathbf{y}\|^{\alpha}$, $\theta^{(p)}(\mathbf{x})$ is any pth-order mixed partial derivative of θ evaluated at $\mathbf{x}$, $q = p + \alpha$, $0 < \alpha \le 1$.

L_1-optimal estimates will be constructed for the models that follow, reconfirming the dimensionality reduction principle. Models I and II are special cases of the additive supermodel but are considered separately to make easier the presentation of the discretization of the supermodels. In the models $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, $\theta_j \in \Theta_{q,1}$, $\theta_{1j} \in \Theta_{q,1}$, $\psi_j \in \Theta_{q,r_j}$; $\mathbf{b}$ is an element of the unit sphere centered at the origin, $\mathbf{b}^T\mathbf{x}$ denotes the scalar product of the vectors $\mathbf{b}$ and $\mathbf{x}$; $(m_1, \ldots, m_{r_j})$ and $(s_1, \ldots, s_k)$ are such that $m_i \ne m_j$ and $s_i \ne s_j$ for $i \ne j$; k, K, K_1, K_2 are either known or unknown but bounded by the known constants d, D, D_1, D_2, respectively; $2 \le r_j \le d - 1$. The models are:

1. model I, $\theta(\mathbf{x}) = \theta_1(x_{s_1}) + \cdots + \theta_k(x_{s_k}) + \sum_{j=1}^{K} \psi_j(x_{m_1}, \ldots, x_{m_{r_j}})$;
2. model II, $\theta(\mathbf{x}) = \theta_1(\mathbf{b}^T\mathbf{x})$;
3. the additive supermodel, $\theta(\mathbf{x}) = \sum_{j=1}^{K_1} \theta_{1j}(\mathbf{b}_\mathbf{j}^T\mathbf{x}) + \sum_{j=1}^{K_2} \psi_j(x_{m_1}, \ldots, x_{m_{r_j}})$;
4. the multiplicative supermodel, $\theta(\mathbf{x}) = \prod_{j=1}^{K_1} \theta_{1j}(\mathbf{b}_\mathbf{j}^T\mathbf{x}) \prod_{j=1}^{K_2} \psi_j(x_{m_1}, \ldots, x_{m_{r_j}})$.

The additive model I without interactions and the projection pursuit model II both appear in Stone (1982, 1985). The additive supermodel without interactions and K_1 bounded appears in Chen (1991). The supermodels without interactions appear in Friedman and Stuetzle (1981) and in Huber (1985) with K_1 not necessarily bounded and are called *projection pursuit regression* (PPR) *models*. The PPR models *seem to have the potential of overcoming the curse of dimensionality* but this is not always the case; see the discussion in Chen [(1991), pages 143, 144] concerning the additive PPR model. By impos-

ing in the models the restrictions $k \leq d$, $K \leq D$, $K_1 \leq D_1$, $K_2 \leq D_2$ as both Stone (1985, 1994) and Chen (1991) did (but with $D = D_1 = d$), the curse of dimensionality is by-passed.

DEFINITION 2.1. The $L_1(d\mathbf{x})$ and sup-norm distances of any two functions θ and $\tilde{\theta}$ in $\mathscr{X}_d$ are respectively given by

$$\|\theta - \tilde{\theta}\| = \int_{\mathscr{X}_d} | \theta(\mathbf{x}) - \tilde{\theta}(\mathbf{x}) | \, d\mathbf{x}$$

and

$$\|\theta - \tilde{\theta}\|_\infty = \sup\{| \theta(\mathbf{x}) - \tilde{\theta}(\mathbf{x}) |; \mathbf{x} \in \mathscr{X}_d\}.$$

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-ρ-dense subset of a metric space (Θ, ρ) if every point in Θ is at a ρ-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 $\Theta^n_{q,m}$ of $\Theta_{q,m}$ has cardinality $N_m(a_n)$ such that $\log_2 N_m(a_n) \sim (1/a_n)^{m/q}$; $\Theta^n_{q,m}$ 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 dimensionality of the parameter space. For each of the above models a discretization will be presented in Section 4 with cardinality depending on the dimension of the model via Kolmogorov's entropy. The proposed estimates will be chosen in each case from the discretization rather than the parameter space.

Le Cam (1973) was the first to construct estimates of a probability measure using a multiple testing procedure under dimensionality restrictions in Hellinger distance. Developing the procedure proposed by Le Cam (1973), Birgé (1983) showed that the obtained rates of convergence are the best possible in several situations including the estimation of densities restricted to lie in Sobolev balls. Yatracos (1985) constructed minimum distance estimates under the assumption of independent identically distributed observations and related the L_1-rate of convergence of the estimates to Kolmogorov's entropy of the parameter space; when the parameter space consists of q-smooth densities the estimate is L_1-optimal. Roussas and Yatracos (1997) provided minimum distance estimates of the unknown probability measure on the basis of a segment of observations $X_1, \ldots, X_n$ from a ϕ-mixing sequence of random variables and showed that an upper convergence rate of the proposed estimate depends on Kolmogorov's entropy of the parameter space and the mixing coefficient ϕ. When the parameter space consists of q-smooth densities and $\phi(n) = cn^{-(1+\delta)}$, $c > 0$, $\delta > 0$, the rates of convergence coincide with those in the i.i.d. case. All these methods use discretization and are close relatives of Grenander's *method of sieves* that is based on the likelihood of the observations [Grenander (1981)].

The minimum distance method of estimation was formalized as a principle by Wolfowitz (1957). A lot of work has been devoted ever since to this topic. In particular it was shown that under some regularity conditions the mini-

mum distance estimator is robust and asymptotically efficient [Beran (1977), Millar (1982) and Donoho and Liu (1988a)]. Pathologies of some minimum distance estimators for the normal model are examined in Donoho and Liu (1988b). The interested reader should consult Le Cam (1986) and Le Cam and Yang (1990) for a modern approach in abstract estimation theory, Devroye and Györfi (1985) and Devroye (1987) for the use of the L_1-distance and related results in nonparametric estimation problems.

It is of interest to consider alternative potential routes to our results, especially those of Birgé (1983) and Barron, Birgé and Massart (1997). In the latter, which is inspired by Barron and Cover (1991), performance bounds for criteria for model selection are developed based on sieves. The model selection criteria are based on a loss with an added penalty term. The penalized minimum contrast estimator that is obtained is adaptive in the sense that it is minimax-rate optimal for each of the class of models considered. Applications of the method include density estimation and classical nonparametric regression estimation with errors that are additive, centered either at their expectation (for least squares regression) or at their median (for minimum L_1-regression) and subject to boundedness of the moment generating function of their absolute values. It is instructive to recall that the starting point in Yatracos (1988, 1989a, 1992), in Roussas and Yatracos (1996) and in this work was the observation that a regression-type problem can be viewed as a *multiple* density or parameter estimation problem *occurring at each observed value* of the $\mathbf{X}$'s. The minimum distance estimation method that we have used as well as other methods such as, for example, the maximum likelihood method [Fisher (1922, 1925)] and the methods of Le Cam (1973), Birgé (1983), Grenander (1981) and Barron, Birgé and Massart (1997) do not require knowledge of the statistical nature of the estimand when estimating the finite-dimensional parameter of a density. Therefore in principle all these other methods *properly modified* are potential routes for obtaining results similar to ours (via minimum distance) in the estimation of regression-type functions and the confirmation of the dimensionality reduction principle. It remains to be seen if this is indeed feasible and it is expected that the associated proofs and technicalities will not be less involved than those in the above mentioned papers. On the other hand in our setup the smoothness q of the unknown regression type function is assumed to be known, and therefore we do not face a model selection problem as in Barron, Birgé and Massart (1997). Our method could also be applied when the smoothness q is not known as in the additive and multiplicative supermodels when K_1 and K_2 are unknown but bounded by a known constant. A family of discretizations $\Theta_{n,1}, \ldots, \Theta_{n,k_n}$ would be considered and the minimum distance criterion in Definition 5.1 would apply to each separately. From the set of minimum distance estimates so obtained the one with the smallest value in the objective function (of Definition 5.1) would be chosen as our estimate; if two (or more) such estimates exist, the less smooth estimate will be selected. It is expected that the resulting estimate will be adaptive in the sense of Barron, Birgé and Massart (1997) but the details remain to be checked.

DEFINITION 2.2. The L_1-distance of two probability measures Q, S, defined on the probability space $(\mathscr{W}, \mathscr{A})$, is defined as

$$\|Q - S\| = 2\sup\{|Q(A) - S(A)|;\ A \in \mathscr{A}\};$$

their 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.

DEFINITION 2.3. A sequence of estimators $\{T_n\}$ is *optimal in probability* for θ, with respect to a distance ρ, if there is a sequence $\{\delta_n\},\ n = 1, 2, \ldots,$ decreasing to zero such that

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

and

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

If only (2.1) holds, δ_n is an *upper convergence rate* in probability.

The sequence of estimators $\{T_n\}$ is *risk optimal* with respect to ρ with rate of convergence δ_n, $n = 1, 2, \ldots,$ if there are positive constants C_L, C_U such that

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

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

If only (2.4) holds, δ_n is a *risk upper convergence rate*.

The following distributional assumptions are made:

(A1) $c_1 \mid t - s \mid \le \|f(\cdot \mid \mathbf{x}, t) - f(\cdot \mid \mathbf{x}, s)\| \le c_2 \mid t - s \mid$; c_1, c_2 are constants greater than zero, independent of $\mathbf{x}$, $\|\cdot\|$ 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.

(A2) The form of the conditional density $f(y|\mathbf{x}, \theta(\mathbf{x}))$ is known.

(A3) The density $g(\mathbf{x})$ of $\mathbf{X}$ is bounded below and above, by the positive, finite constants A and B, respectively.

(A4) $K(P_s, P_t) \le c(s - t)^2$, for every s, t, possible values of $\theta(\mathbf{x})$; c is a positive constant, P_s denotes the probability measure with density $f(y \mid \mathbf{x}, s)$.

Assumptions (A1)–(A3) are used to construct the proposed minimum distance estimate and calculate upper convergence rates. Assumption (A1) holds in most of the commonly assumed models for the Y's; see Examples 1–7 in Yatracos [(1989a), pages 1600, 1601]. Without assumption (A2) we cannot obtain the *separating sets* $A_{k,m,i}$ used in the minimum distance criterion; see Definition 5.1. A similar and more specific assumption has been used in Stone [(1994), pages 119, 120], where the conditional densities are assumed to be either Bernoulli or Poisson, and also in Donoho, Johnstone, Kerkyacharian and Picard (1995), where in addition to normal errors the parameter of interest

is known to be a conditional mean [Donoho, Johnstone, Kerkyacharian and Picard (1995), Section 2, page 302, Section 3.1, page 308] and the obtained estimates are nearly optimal [Donoho, Johnstone, Kerkyacharian and Picard (1995), Section 3.2, page 312]. Neither Stone (1982, 1985) nor Chen (1991) nor Chaudhuri (1991) used (A2) because the nature of the estimand in the conditional density was known to be either a mean, a median or another quantile. It is the price to pay in a *regression type problem* where the nature of the parameter θ in the conditional density is unknown. Therefore, one cannot determine the functional of the Y's that should be used to estimate θ. Assumption (A3) has been used by Chaudhuri (1991), Chen (1991), Stone (1982, 1985, 1994), Truong and Stone (1994) and several other authors. It allows in our calculations the passage, without much loss, from the L_1-distance $\|\hat{\theta}_n - \theta\|$ to the expectation $E|\hat{\theta}_n(\mathbf{X}) - \theta(\mathbf{X})|$ that is used in proving Proposition 3.1 and leads to the dimensionality reduction principle. It also confirms indirectly that the sample $X_1, \ldots, X_n$ provides enough information about θ over all its domain.

Assumption (A4) is satisfied in several models for the Y's [see Yatracos (1988), page 1186, Example 1]. Without (A4) the proposed estimate may not be L_1-optimal; the lower convergence rates may not coincide with the upper convergence rates. This occurs when the conditional density is uniform but also for other models; see Yatracos [(1988), Example 2, page 1186] for such situations and for the use of Fano's lemma and lower convergence rates in regression type problems. In the same paper [Yatracos (1988), Proposition 1, page 1184] it is also shown that Stone's (1982) assumptions imply (A4).

Our main interest is the estimation of θ rather than its components. Under assumptions (A1) and (A3) the elements of the models considered are identifiable; if $\theta_1 \neq \theta_2$ the L_1-distance between the joint densities $\|f(\cdot, \cdot, \theta_1) - f(\cdot, \cdot, \theta_2)\|$ is positive.

The proofs of Proposition 3.1 and 5.1 are based on the following.

HOEFFDING'S INEQUALITIES [Hoeffding (1963)]. (i) *Let $U_1, \ldots, U_n$ be independent random variables such that $0 \leq U_j \leq 1$, $j = 1, \ldots, n$. Let $S_n = U_1 + U_2 + \cdots + U_n$, $ES_n = np$. Then*

$$P[S_n \geq np + k] \leq \exp\{-k^2/2(np + k)\}$$

and

$$P[S_n \leq np - k] \leq \exp\{-k^2/2np(1 - p)\}.$$

(ii) *Let $W_1, \ldots, W_n$ be independent random variables, $0 \leq W_j \leq b$, $j = 1, \ldots, n$, $\bar{W} = (W_1 + W_2 + \cdots + W_n)/n$, $E\bar{W} = \mu$. Then*

$$P[\bar{W} - \mu \geq t] \leq \exp\{-2nt^2/b^2\}, \qquad 0 < t < b - \mu$$

and

$$P[\bar{W} - \mu < -t] \leq \exp\{-2nt^2/b^2\}, \qquad 0 < t < \mu.$$

3. Some results for a general parameter space. Let Θ be the parameter space for a regression type problem, Θ_n a finite subset of it with cardinality $N(a_n)$, the sequence $\{a_n\}$ decreasing to 0. Let $\tilde{\theta}_n$ be an estimate of the unknown regression type function θ, with values in Θ_n; let θ_k, θ_m be elements of Θ_n. Let $P_{\theta(\mathbf{x_i})}$ be a probability measure with density $f(y|\mathbf{x_i},\ \theta(\mathbf{x_i}))$, $i = 1, \ldots, n$; Q is the distribution of any of the $\mathbf{X}$' s. Let Q^n, Q^∞, P^n denote, respectively, the joint distribution of $(\mathbf{X_1} = \mathbf{x_1}, \ldots, \mathbf{X_n} = \mathbf{x_n})$, the distribution of the infinite vector of the $\mathbf{X}$'s, and the product measure of the Y's conditionally on the $\mathbf{X}$'s. Define the following quantities:

$$E \mid \theta_k - \theta_m \mid = E \mid \theta_k(\mathbf{X}) - \theta_m(\mathbf{X}) \mid;$$

$$E_n \mid \theta_k - \theta_m \mid = \frac{1}{n} \sum_{i=1}^{n} \mid \theta_k(\mathbf{X_i}) - \theta_m(\mathbf{X_i}) \mid;$$

$$\Delta_n(\theta_k, \theta_m) = \left| E \mid \theta_k(\mathbf{X}) - \theta_m(\mathbf{X}) \mid - \frac{1}{n} \sum_{i=1}^{n} \mid \theta_k(\mathbf{X_i}) - \theta_m(\mathbf{X_i}) \mid \right|;$$

$$A_n(\varepsilon_n, m) = \bigcup_{k=1}^{N(a_n)} \left\{(\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})\colon \Delta_n(\theta_k, \theta_m) > c\varepsilon_n\right\};$$

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

The following proposition is fundamental in proving the dimensionality reduction principle holds for both models. All the proofs follow in the Appendix. From now on, the letters $c, c_1, c_2 \ldots$ will denote generic, positive constants, independent of n; $I_A(x) = 1$, if $x \in A$; it is 0 otherwise.

PROPOSITION 3.1. *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. Then, for* $\varepsilon_n \sim \{(\log N(a_n))/n\}^{1/2} \downarrow 0$:

(a) $Q^n(A_n(\varepsilon_n, m)) \leq N(a_n)^{1-c_2(c)}$; *the constant c may be chosen large enough such that* $1 < c_2(c)$;

(b) $P[\{(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)\colon \Delta_n(\tilde{\theta}_n, \theta_m) > c\varepsilon_n\}] \leq Q^n(A_n(\varepsilon_n, m))$;

(c) *For sequences* $\{N(a_n)\}$, *such that* $\sum_{n=1}^{\infty} N(a_n)^{1-c_2(c)} < \infty$, *then*

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

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

(d) $Q^n(A_n(\varepsilon_n)) \leq N(a_n)^{2-c_2(c)}$; *the constant c may be chosen large enough that* $2 < c_2(c)$;

(e) $P[\bigcup_{m=1}^{N(a_n)}\{(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)\colon \Delta_n(\tilde{\theta}_n, \theta_m) > c\varepsilon_n\}] \leq Q^n(A_n(\varepsilon_n))$;

(f) *For sequences* $\{N(a_n)\}$, *such that* $\sum_{n=1}^{\infty} N(a_n)^{2-c_2(c)} < \infty$,

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

A different version of the next lemma appears in Yatracos (1989a, 1992).

LEMMA 3.1. *Let* $\Phi_n = \{A_{k,m,i}: 1 \le k < m \le M_n,\ i = 1, \ldots, n\}$ *be a family of sets. Conditionally on* $\mathbf{X_1} = \mathbf{x_1}, \ldots, \mathbf{X_n} = \mathbf{x_n}$, $A_{k,m,i}$ *is an element of the* σ*-field where* $P_{\theta(\mathbf{x_i})}$ *is defined,* $\theta \in \Theta$, $i = 1, \ldots, n$. *Let* $\{\gamma_n\}$ *be a sequence such that* $\{\gamma_n\} \sim \{(\log M_n)/n\}^{1/2} \downarrow 0$. *Then*

$$\lim_{n\to\infty} P^n\left[n^{-1} \sup_{1\le k<m\le M_n} \left| \sum_{i=1}^{n} (I_{A_{k,m,i}}(Y_i) - P_{\theta(\mathbf{x_i})}(A_{k,m,i})) \right| > \gamma_n\right] = 0.$$

4. Discretization. To construct the proposed estimates a discretization of the parameter space Θ is needed for all models. For convenience, the definition of the models will be repeated.

MODEL I. We have $\theta(\mathbf{x}) = \theta_1(x_{s_1}) + \cdots + \theta_k(x_{s_k}) + \sum_{j=1}^{K} \psi_j(x_{m_1}, \ldots, x_{m_{r_j}})$, $\theta_i \in \Theta_{q,1}$, $1 \le i \le k$, $\psi_j \in \Theta_{q,r}$, $1 \le j \le K$; $(s_1, \ldots, s_k)$ and $(m_1, \ldots, m_r)$ are such that $s_i \ne s_j$ and $m_i \ne m_j$ for $i \ne j$, $1 \le k \le d$, $2 \le r_j \le d-1$, K is either known or unknown but bounded by a known constant D.

When K is unknown but bounded by a known constant D it is viewed as an additional finite-dimensional parameter of the model, with possible values in $\{1, 2, \ldots, D\}$. It is selected at the same time as the other parameters using the minimum distance criterion. The same approach is taken when k is unknown but also in the other models of Section 2 when only the bounds D_1 and D_2 are known.

To discretize the parameter space that corresponds to model I assume for now that k is known and that there is only one interaction, that is, $\theta(\mathbf{x}) = \theta_1(x_{s_1}) + \cdots + \theta_k(x_{s_k}) + \psi(x_{m_1}, \ldots, x_{m_r})$. Let $\Theta_{q,1}^n$ be an a_n-$\|\cdot\|_\infty$-dense subset of $\Theta_{q,1}$ with cardinality $N_1(a_n)$, and let $\Theta_{q,r}^n$ be an a_n-$\|\cdot\|_\infty$-dense subset of $\Theta_{q,r}$ with cardinality $N_r(a_n)$. Define the set Θ_I^n with elements

$$\theta_i(\mathbf{x}) = \theta_{i_1}(x_{s_1}) + \cdots + \theta_{i_k}(x_{s_k}) + \psi_{i_{k+1}}(x_{m_1}, \ldots, x_{m_r}),$$

where $\theta_{i_j} \in \Theta_{q,1}^n$, $\psi_{i_{k+1}} \in \Theta_{q,r}^n$, $1 \le i_j \le N_1(a_n)$, $1 \le j \le k$, $1 \le i_{k+1} \le N_r(a_n)$, $1 \le i \le N_I(a_n)$.

Therefore, the cardinality $N_I(a_n)$ of Θ_I^n satisfies the relation $N_I(a_n) \sim \{N_1(a_n)\}^k N_r(a_n)$. In the presence of K interactions

$$N_I(a_n) \sim \{N_1(a_n)\}^k \prod_{j=1}^{K} N_{r_j}(a_n);$$

if k and K are only known to be bounded, respectively, by d and D, replace both by their bounds in the expression for the cardinality $N_I(a_n)$.

LEMMA 4.1. *The set* Θ_I^n *is a* ca_n-$\|\cdot\|_\infty$*-dense subset for model I.*

MODEL II. We have $\theta(\mathbf{x}) = \theta_1(\mathbf{b}^T\mathbf{x})$, $\theta_1 \in \Theta_{q,1}$, $q \ge 1$, $\mathbf{b}$ is an element of the unit sphere centered at the origin, $\mathbf{b}^T\mathbf{x}$ denotes the scalar product of the vectors $\mathbf{b}$ and $\mathbf{x}$.

To discretize the parameter space that corresponds to this model let $\Theta^n_{q,1}$ be an a_n-$\|\cdot\|_\infty$-dense subset of $\Theta_{q,1}$ with cardinality $N_1(a_n)$. Let b_i, $1 \le i \le h_n$, be the $n^{-1/2}$-discretisation of $[0,1]$, that is, the centers of successive intervals of length $n^{-1/2}$ needed to cover $[0,1]$. The quantity h_n is defined as the integer part of $n^{1/2}$ augmented by 1, if $n^{1/2}$ is not an integer, and as $n^{1/2}$ otherwise. Since h_n is of the same order of magnitude as $n^{1/2}$, without loss of generality it will be replaced from now on by $n^{1/2}$. Define the set F_n, a $dn^{-1/2}$-discretization of $[0,1]^d$, with elements $\mathbf{b_i} = (b_{i_1}, \ldots, b_{i_d})$, $1 \le i_j \le n^{1/2}$, $1 \le j \le d$, $1 \le i \le n^{d/2}$, b_{i_j} is an element of the $n^{-1/2}$-discretization of $[0,1]$ already described. The cardinality N_{F_n} of F_n is of the order $n^{d/2}$. Define the set Θ^n_{II} with elements

$$\theta_k(\mathbf{x}) = \theta_{i_j}(\mathbf{b}^T_{\mathbf{i_m}}\mathbf{x}), \qquad \theta_{i_j} \in \Theta^n_{q,1},\ 1 \le i_j \le N_1(a_n),\ \mathbf{b_{i_m}} \in F_n,\ 1 \le i_m \le n^{d/2}.$$

It follows that the cardinality $N_{\mathrm{II}}(a_n)$ of Θ^n_{II} satisfies the relation $N_{\mathrm{II}}(a_n) \sim n^{d/2}N_1(a_n)$.

LEMMA 4.2. *Let θ_r be the nearest neighbor of θ_1 in $\Theta^n_{q,1}$, and let $\mathbf{b_r}$ be the nearest neighbor of $\mathbf{b}$ in F_n. Then*

$$\left|\, \theta(\mathbf{x}) - \theta_r(\mathbf{b}^T_{\mathbf{r}}\mathbf{x}) \,\right| \le a_n + cn^{-1/2}.$$

It follows that the set Θ^{II}_n is a $(a_n + cdn^{-1/2})$-$\|\cdot\|_\infty$-dense subset for model II.

THE ADDITIVE SUPERMODEL. We have

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

$\theta_{1,j} \in \Theta_{q,1}$, $\psi_j \in \Theta_{q,r_j}$, $q \ge 1$, $\mathbf{b}^T_{\mathbf{j}}\mathbf{x}$ is the scalar product of $\mathbf{b_j}$ and $\mathbf{x}$, $\mathbf{b_j}$ is an element of the unit sphere centered at the origin, $(m_1, \ldots, m_{r_j})$ are such that $m_i \ne m_j$ for $i \ne j$, K_1, K_2 are either known or unknown but bounded by known constants D_1, D_2, respectively, $2 \le r_j \le d-1$.

When K_1, K_2 are known, a $c(a_n + n^{-1/2})$-$\|\ \|_\infty$-dense subset for the corresponding model has cardinality of the order of $N_1^{K_1}(a_n)n^{K_1 d/2}\prod_{j=1}^{K_2} N_{r_j}(a_n)$. Such a dense subset may be obtained by a straightforward generalization of Lemma 4.2; every θ_{1j} will be approximated by an element of Θ^n_{II}, every ψ_j will be approximated by an element of Θ^n_{q,r_j} and $\theta(\mathbf{x})$ will be approximated by the sum of such elements of Θ^n_{II} and Θ^n_{q,r_j}. For a discretization of the supermodel for unknown K_1, K_2 we will put together all discretizations for $K_1 = 1, \ldots, D_1$, $K_2 = 1, \ldots, D_2$. The cardinality $N_A(a_n)$ of this superdiscretization is of the order of $N_1^{D_1}(a_n)n^{dD_1/2}\prod_{j=1}^{D_2} N_{r_j}(a_n)$. Index the elements of the superdiscretization with k, $1 \le k \le N_A(a_n)$. Note that the results of Proposition 3.1 and of Lemma 3.1 hold, with Θ^n_A denoting the discretization of the additive supermodel.

THE MULTIPLICATIVE SUPERMODEL. We have

$$\theta(\mathbf{x}) = \prod_{j=1}^{K_1} \theta_{1j}(\mathbf{b_j}^T\mathbf{x}) \prod_{j=1}^{K_2} \psi_j(x_{m_1}, \ldots, x_{m_{r_j}}).$$

(See the previous model for details about θ_{1j}, $\mathbf{b_j}$, ψ_j, K_1, K_2.)

To discretize the parameter space that corresponds to this model assume for now that there is only one interaction, $\theta(\mathbf{x}) = \prod_{j=1}^{K_1} \theta_{1j}(\mathbf{b_j}^T\mathbf{x})\psi(x_{m_1}, \ldots, x_{m_r})$. When K_1 is known a $c(a_n + n^{-1/2})$-$\| \ \|_\infty$-dense subset for the corresponding model has cardinality of the order of $N_1^{K_1}(a_n)n^{K_1 d/2}N_r(a_n)$. Every θ_{1j} is approximated by an element of Θ_{II}^n; $\psi(x_{m_1}, \ldots, x_{m_r})$ is approximated by an element of $\Theta_{q,r}$; $\theta(\mathbf{x})$ is approximated by the product of such elements. A bound on the approximation of θ is obtained using the relation $|\prod_{j=1}^K f_j(\mathbf{x}) - \prod_{j=1}^K g_j(\mathbf{x})| \leq \sum_{j=1}^K c_j|f_j(\mathbf{x}) - g_j(\mathbf{x})|$ that holds when the f's and the g's are uniformly bounded, $c_1, \ldots, c_K$ are known constants. When K_1 is not known we consider all possible discretizations for $K_1 = 1, \ldots, D_1$. The cardinality $N_M(a_n)$ of this superdiscretization is of the order of $N_1^{D_1}(a_n)n^{D_1 d/2}N_r(a_n)$. In the presence of K_2 interactions $N_M(a_n) \sim N_1^{D_1}(a_n)n^{D_1 d/2}\prod_{j=1}^{K_2} N_{r_j}(a_n)$ and if it is only known that K_2 is bounded by D_2, then $N_M(a_n) \sim N_1^{D_1}(a_n)n^{D_1 d/2}\prod_{j=1}^{D_2} N_{r_j}(a_n)$. Index the elements of the superdiscretization with k, $1 \leq k \leq N_M(a_n)$. Note that the results of Proposition 3.1 and of Lemma 3.1 hold, with Θ_M^n denoting the discretization of the multiplicative supermodel.

REMARK 4.1. For the implementation of the proposed estimation method the dense subsets used in discretization should become available. This could be done along the lines of Devroye (1987) or Kolmogorov and Tikhomirov (1959) but will represent a major task. Nevertheless, this approach remains a useful tool providing simple solutions in estimation problems using only Hoeffding's inequalities.

5. Estimates and rates of convergence. For the regression type problem let the parameter space Θ follow one of the models already defined. Let $\{a_n\}$ be a sequence decreasing to 0, and let Θ_n be a $c(a_n + \zeta n^{-1/2})$-$\| \cdot \|_\infty$-dense subset of Θ with cardinality $N(a_n)$; Θ_n could be any of the dense subsets $\Theta_{\mathrm{I}}^n, \Theta_{\mathrm{II}}^n, \Theta_A^n, \Theta_M^n$ of the models that were described in the previous section; ζ takes the values 0 or 1. Given $\mathbf{X_1} = \mathbf{x_1}, \ldots, \mathbf{X_n} = \mathbf{x_n}$, define the sets

$$A_{k,m,i} = \{y\colon f(y \mid \mathbf{x_i}, \theta_k(\mathbf{x_i})) > f(y \mid \mathbf{x_i}, \theta_m(\mathbf{x_i}))\},$$
$$\theta_k \in \Theta_n, \ \theta_m \in \Theta_n, \ 1 \leq k < m \leq N(a_n), \ i = 1, \ldots, n.$$

DEFINITION 5.1. Let $(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)$ be a random sample of n independent pairs in the setup of the regression type problem already described.

The minimum distance estimator $\hat{\theta}_n$ of θ is defined, such that

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

When all the $\mathbf{x}$'s are the same the obtained estimate is the minimum distance estimate of the unknown probability measure of the Y's that belongs to an L_1-totally bounded space of measures [Yatracos (1985)]. Note that, in the case of model II, $\hat{\theta}_n(\mathbf{x}) = \hat{\theta}_{1,n}(\hat{\mathbf{b}}_{\mathbf{n}}^T\mathbf{x})$; $\hat{\theta}_{1,n}$ and $\hat{\mathbf{b}}_{\mathbf{n}}$ are estimates of θ_1 and $\mathbf{b}$, respectively. For the supermodel and the multiplicative model without interactions $\hat{\theta}_n(\mathbf{x})$ will be the sum and the product, respectively, of such estimates.

LEMMA 5.1. *Let $\hat{\theta}_n(\mathbf{x})$ be the minimum distance estimator of θ, θ_r the nearest neighbor of θ in Θ_n. Then, under assumption* (A1), *given $\mathbf{X_1}=\mathbf{x_1},\ldots,\mathbf{X_n}=\mathbf{x_n}$,*

$$n^{-1}\sum_{i=1}^{n}\mid\hat{\theta}_n(\mathbf{x_i})-\theta_r(\mathbf{x_i})\mid$$
$$\le c_1(a_n+\zeta n^{-1/2})$$
$$+c_2n^{-1}\sup_{1\le k<m\le N(a_n)}\left\{\left|\sum_{i=1}^{n}(I_{A_{k,m,i}}(Y_i)-P_{\theta(\mathbf{x_i})}(A_{k,m,i}))\right|\right\}.$$

The theorem confirming Stone's dimensionality reduction principle for all the models follows.

THEOREM 5.1. *Let $(\mathbf{X_1},Y_1),\ldots,(\mathbf{X_n},Y_n)$ be a random sample of n independent pairs, in a regression type problem. The random vectors $\mathbf{X_i}$ take values in $[0,1]^d$, $d\ge 1$, and Y_i are the corresponding real-valued responses, $i=1,\ldots,n$. Conditionally on $\mathbf{X_1}=\mathbf{x_1},\ldots,\mathbf{X_n}=\mathbf{x_n}$, the random variables $Y_1,\ldots,Y_n$ are independent, each having as density $f(y|\mathbf{x_i},\theta(\mathbf{x_i}))$, $i=1,\ldots,n$; the unknown function θ follows one of the models already defined. If assumptions* (A1)–(A3) *are satisfied and $n^{-1/2}=o(a_n)$ the minimum distance estimator is uniformly consistent with upper rate of convergence in probability a_n, in L_1-distance, such that*

$$a_n\sim\left\{\frac{\log N(a_n)}{n}\right\}^{1/2}.$$

COROLLARY 5.1. *Under the assumptions of Theorem* 5.1 *and* (A4) *the minimum distance estimator is optimal in probability with rates of con-*

vergence a_n:

(a) *for model I and the additive and the multiplicative supermodels with interactions* $a_n \sim \{(\log N_r(a_n))/n\}^{1/2} \sim n^{-q/(2q+r)}$, r *is the dimension of the model*;

(b) *for model II and the additive and the multiplicative supermodels without interactions* $a_n \sim \{(\log N_1(a_n))/n\}^{1/2} \sim n^{-q/(2q+1)}$.

COROLLARY 5.2. *Under the assumptions of Theorem* 5.1 *and* (A4) *the minimum distance estimator is risk-optimal with rates of convergence as in Corollary* 5.1. *Moreover*, *these rates of convergence hold almost surely*.

REMARK 5.1. If the unknown parameter θ does not follow exactly one of the models in Section 2, let θ^* be its closest approximation in the chosen model such that $\|\theta - \theta^*\|_\infty < \varepsilon$. Following Proposition 2 in Yatracos (1985) and (A1) it is easy to see that $\|\hat{\theta}_n - \theta\| \le c_1\varepsilon + c_2 a_n$, with a_n satisfying the relation of Theorem 5.1. The optimal rates of convergence also hold when the constant L in the Lipschitz condition is bounded by a known constant M. Otherwise we may consider a sequence (of bounds) M_n going to infinity; we discretize for every M_n and construct the minimum distance estimate. Finally, L will be smaller than one of the M_n's and the parameter θ will be an element of the parameter space from then on. The rate will be as close as we like to the optimal depending on how fast the sequence M_n increases. The situation is different if α is unknown even if it is bounded by 1. The reason is that α appears through the total smoothness q as exponent of the radius a_n in the entropy of the space Θ for the model considered.

Let $\theta^{(\mathbf{s})}$ be an [$\mathbf{s}$]th-order mixed partial derivative of θ, not identically 0, $[\mathbf{s}] = s_1 + \cdots + s_d$. An upper bound in probability will be computed for $\|\hat{\theta}_n^{(\mathbf{s})} - \theta^{(\mathbf{s})}\|$ with the help of the following proposition, showing it is easier to estimate θ than its derivative $\theta^{(\mathbf{s})}$.

PROPOSITION 5.1 [Yatracos (1989b), Proposition 2]. *Let* θ *be a real-valued function*, $\tilde{\theta}_n$ *an estimator of* θ, *both defined on a compact set in* R^d, *having mixed partial derivatives of order* p, *the* p*th derivative having a modulus of continuity* $w(z)$, $z > 0$.

Then, *for* $1 \le [\mathbf{s}] \le p$,

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

$\|\cdot\|_v$ *is an* L_v*-distance*, $1 \le v \le \infty$.

If $\|\tilde{\theta}_n - \theta\|_v \sim a_n$ *in probability*, *a value of* b_n *that gives an upper convergence rate in probability satisfies the relation* $b_n^p w(b_n) \sim a_n$.

In all assumed models the rate of convergence $a_n \sim n^{-q/(2q+r)}, r \ge 1$, $w(b_n) \sim b_n^\alpha$. Thus, choosing $b_n \sim a_n^{1/q}$, since $\|\hat{\theta}_n - \theta\| \sim a_n$, it holds that $\|\hat{\theta}_n^{(\mathbf{s})} - \theta^{(\mathbf{s})}\| \le c n^{-(q-[\mathbf{s}])/(2q+r)}$, in probability.

APPENDIX

PROOF OF PROPOSITION 3.1. (a) For $\varepsilon_n \downarrow 0$ it holds that

$$\begin{aligned} Q^n(A_n(\varepsilon_n, m)) &= Q^n\Big(\bigcup_{k=1}^{N(a_n)} \{(\mathbf{X_1}, \ldots, \mathbf{X_n})\colon \Delta_n(\theta_k, \theta_m) > c\varepsilon\}\Big) \\ &\le \sum_{k=1}^{N(a_n)} Q^n[(\mathbf{X_1}, \ldots, \mathbf{X_n})\colon \Delta_n(\theta_k, \theta_m) > c\varepsilon] \\ &\le N(a_n) \sup_{1\le k\le N(a_n)} Q^n[(\mathbf{X_1}, \ldots, \mathbf{X_n})\colon \Delta_n(\theta_k, \theta_m) > c\varepsilon_n] \\ &\le N(a_n) \exp\{-c_1 n\varepsilon_n^2\}. \end{aligned}$$

The last inequality was obtained using Hoeffding's inequality since the functions $\theta \in \Theta_{q,d}$ are uniformly bounded. Choosing $\varepsilon_n \sim \{(\log N(a_n))/n\}^{1/2}$ we obtain $Q^n(A_n(\varepsilon_n, m)) \le N(a_n)^{1-c_2(c)}$. The constant c may be chosen large enough that $c_2(c) > 1$.

(b)
$$\begin{aligned} &P[(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)\colon \Delta_n(\tilde\theta_n, \theta_m) > c\varepsilon_n] \\ &\quad = P\Big[\bigcup_{k=1}^{N(a_n)} \{(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)\colon \tilde\theta_n = \theta_k, \Delta_n(\tilde\theta_n, \theta_m) > c\varepsilon_n\}\Big] \\ &\quad = P\Big[\bigcup_{k=1}^{N(a_n)} \{(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)\colon \tilde\theta_n = \theta_k, \Delta_n(\theta_k, \theta_m) > c\varepsilon_n\}\Big] \\ &\quad \le P\Big[\bigcup_{k=1}^{N(a_n)} \{(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)\colon \Delta_n(\theta_k, \theta_m) > c\varepsilon_n\}\Big] \\ &\quad = Q^n(A_n(\varepsilon_n, m)), \end{aligned}$$

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

(c)
$$\begin{aligned} &P(\Delta_n(\tilde\theta_n, \theta_m) > c\varepsilon_n, \text{ infinitely often } n) \\ &\quad \le \lim_{k\to\infty} \sum_{n=k}^{\infty} P[(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)\colon \Delta_n(\tilde\theta_n, \theta_m) > c\varepsilon_n] \\ &\quad \le \lim_{k\to\infty} \sum_{n=k}^{\infty} Q^n(A_n(\varepsilon_n, m)) \\ &\quad \le \lim_{k\to\infty} \sum_{n=k}^{\infty} N(a_n)^{1-c_2(c)} = 0. \end{aligned}$$

The proofs of (d), (e) and (f) are analogous to those of (a), (b) and (c). □

PROOF OF LEMMA 3.1. Let

$$S_n = \sum_{i=1}^{n} I_{A_{k,m,i}}(Y_i), \qquad np = \sum_{i=1}^{n} P_{\theta(\mathbf{x_i})}(A_{k,m,i});$$

assume without loss of generality that $p \le 0.5$. Using Hoeffding's inequality we obtain

$$\begin{aligned}
&P^n\left[n^{-1} \sup_{1\le k<m\le M_n} \left\{ \left| \sum_{i=1}^{n} (I_{A_{k,m,i}}(Y_i) - P_{\theta(\mathbf{x_i})}(A_{k,m,i})) \right| \right\} > \gamma_n \right] \\
&\quad \le \sum_{1\le k<m\le M_n} P^n\left[n^{-1} \left| \sum_{i=1}^{n} (I_{A_{k,m,i}}(Y_i) - P_{\theta(\mathbf{x_i})}(A_{k,m,i})) \right| > \gamma_n \right] \\
&\quad \le M_n^2 \left\{ \exp\left(-\frac{n\gamma_n^2}{2(p+\gamma_n)} \right) + \exp\left(-\frac{n\gamma_n^2}{2p(1-p)} \right) \right\} \\
&\quad \le 2M_n^2 \exp\left\{ -\frac{n\gamma_n^2}{1+2\gamma_n} \right\} \\
&\quad \le 2M_n^2 \exp\left\{ -\frac{n\gamma_n^2}{2} \right\}
\end{aligned}$$

for large n. Choosing $\gamma_n \sim c\{(\log M_n)/n\}^{1/2}$ and an appropriate constant c we get convergence to 0. □

Lemma 4.1 has a simple proof that is omitted.

PROOF OF LEMMA 4.2. Let θ_r be the nearest neighbor of θ_1 in $\Theta_{q,1}^n$ and $\mathbf{b_r}$ be the nearest neighbor of $\mathbf{b}$ in F_n. Then, we have

$$\begin{aligned}
|\theta(\mathbf{x}) - \theta_r(\mathbf{b_r}^T\mathbf{x})| &= |\theta_1(\mathbf{b}^T\mathbf{x}) - \theta_r(\mathbf{b_r}^T\mathbf{x})| \\
&\le |\, \theta_1(\mathbf{b}^T\mathbf{x}) - \theta_r(\mathbf{b}^T\mathbf{x}) \,| \\
&\quad + |\, \theta_r(\mathbf{b}^T\mathbf{x}) - \theta_r(\mathbf{b_r}^T\mathbf{x}) \,| \\
&\le a_n + cdn^{-1/2} \le a_n + cn^{-1/2}.
\end{aligned}$$

□

PROOF OF LEMMA 5.1. We have

$$\begin{aligned}
\sum_{i=1}^{n} \| P_{\hat\theta_n(\mathbf{x_i})} - P_{\theta_r(\mathbf{x_i})} \| &\le 2 \sup_{1\le k<m\le N(a_n)} \left| \sum_{i=1}^{n} (P_{\hat\theta_n(\mathbf{x_i})}(A_{k,m,i}) - P_{\theta_r(\mathbf{x_i})}(A_{k,m,i})) \right| \\
&\le c_1 \sup_{1\le k<m\le N(a_n)} \left| \sum_{i=1}^{n} (P_{\hat\theta_n(\mathbf{x_i})}(A_{k,m,i}) - I_{A_{k,m,i}}(Y_i)) \right| \\
&\quad + c_2 \sup_{1\le k<m\le N(a_n)} \left| \sum_{i=1}^{n} (P_{\theta_r(\mathbf{x_i})}(A_{k,m,i}) - I_{A_{k,m,i}}(Y_i)) \right| \\
&\le c \sup_{1\le k<m\le N(a_n)} \left| \sum_{i=1}^{n} (P_{\theta_r(\mathbf{x_i})}(A_{k,m,i}) - I_{A_{k,m,i}}(Y_i)) \right|
\end{aligned}$$

$$\leq c_1 n(a_n + \zeta n^{-1/2}) + c_2 \sup_{1\leq k<m\leq N(a_n)} \left\{ \left| \sum_{i=1}^{n} (I_{A_{k,m,i}}(Y_i) - P_{\theta(\mathbf{x_i})}(A_{k,m,i})) \right| \right\}.$$

The result follows using assumption (A1). □

PROOF OF THEOREM 5.1. Let $a_n > 0$ to be determined later in order to achieve optimality of the proposed estimate. Let Θ_n be any of the dense subsets constructed in Section 4. Let $\hat{\theta}_n$ be the minimum distance estimator of θ and let θ_r be the nearest neighbor of θ in Θ_n. Define the quantities

$$E \mid \theta_k - \theta_m \mid = E \mid \theta_k(\mathbf{X}) - \theta_m(\mathbf{X}) \mid,$$

$$E_n \mid \theta_k - \theta_m \mid = \frac{1}{n} \sum_{i=1}^{n} \mid \theta_k(\mathbf{X_i}) - \theta_m(\mathbf{X_i}) \mid.$$

Using assumption (A3) and the condition $n^{-1/2} = o(a_n)$ we have

$$\begin{aligned} \|\hat{\theta}_n - \theta\| &= \int_{[0,1]^d} \mid \hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x}) \mid d\mathbf{x} \\ &\leq c_1(a_n + \zeta n^{-1/2}) \\ &\quad + \int_{[0,1]^d} \mid \hat{\theta}_n(\mathbf{x}) - \theta_r(\mathbf{x}) \mid d\mathbf{x} \\ &\leq c_1 a_n + A^{-1} E \mid \hat{\theta}_n - \theta_r \mid. \end{aligned} \tag{5.1}$$

From Proposition 3.1 with $\varepsilon_n \sim \{(\log N(a_n))/n\}^{1/2} \downarrow 0$ we obtain almost surely

$$E \mid \hat{\theta}_n - \theta_r \mid \leq c_2 \varepsilon_n + E_n \mid \hat{\theta}_n - \theta_r \mid. \tag{5.2}$$

Thus (5.1) becomes

$$\|\hat{\theta}_n - \theta\| \leq c_1 a_n + c_2 \varepsilon_n + c_3 E_n \mid \hat{\theta}_n - \theta_r \mid. \tag{5.3}$$

A sequence $\{\delta_n\}$ will be determined such that

$$P[\|\hat{\theta}_n - \theta\| > c\delta_n] \to 0 \quad \text{as } n \to \infty. \tag{5.4}$$

Thus, we have

$$\begin{aligned} &P\big[\|\hat{\theta}_n - \theta\| > c\delta_n\big] \\ &\quad \leq P\big[c_1 a_n + c_2 \varepsilon_n + c_3 E_n \mid \hat{\theta}_n - \theta_r \mid > c\delta_n\big] \\ &\quad = E_{Q^n}\big\{P^n\big[c_1 a_n + c_2 \varepsilon_n + c_3 E_n \mid \hat{\theta}_n - \theta_r \mid > c\delta_n\big]\big\} \\ &\quad \leq E_{Q^n}\left\{P^n\left[n^{-1} \sup_{1\leq k<m\leq N_I(a_n)} \left\{ \left| \sum_{i=1}^{n} (P_{\theta(\mathbf{x_i})}(A_{k,m,i}) - I_{A_{k,m,i}}(Y_i)) \right| \right\} \right.\right. \\ &\qquad\qquad \left.\left. > c\delta_n - c_1 a_n - c_2 \varepsilon_n \right]\right\}. \end{aligned}$$

The last inequality was obtained using Lemma 5.1. Applying Lemma 3.1 with the family Φ_n consisting of the sets $A_{k,m,i}$ used to define the minimum distance estimator $\hat{\theta}_n$, $1 \le k < m \le N(a_n)$, $1 \le i \le n$, with $\gamma_n = c\delta_n - c_1 a_n - c_2\varepsilon_n$, $a_n \sim \varepsilon_n \sim \delta_n \sim \{(\log N(a_n))/n\}^{1/2}$ and an appropriate constant c such that $\gamma_n > 0$, (5.4) is obtained using the bounded convergence theorem. Therefore,

$$\gamma_n \sim a_n \sim \left\{\frac{\log N_r(a_n)}{n}\right\}^{1/2}. \qquad \square$$

PROOF OF COROLLARY 5.1. The discretization results of Section 4 will be used.

For model I the cardinality $N_{\mathrm{I}}(a_n)$ of the largest discretization set Θ_{I}^n with K unknown but bounded by D is of the order of $\{N_1(a_n)\}^d \prod_{j=1}^{D} N_{r_j}(a_n)$ and the resulting upper rate of convergence is $a_n \sim n^{-q/(2q+r)}$; r is the largest dimension of the functional components in the model.

For model II the cardinality $N_{\mathrm{II}}(a_n)$ of the set Θ_{II}^n is of the order of $n^{d/2}N_1(a_n)$ and the resulting upper rate of convergence is $a_n \sim n^{-q/(2q+1)}$.

For both the additive and the multiplicative supermodels with interactions the cardinality of the largest discretization set with K_1, K_2 unknown but bounded by D_1, D_2 is of the order of $N_1^{D_1}(a_n)n^{D_1 d/2}\prod_{j=1}^{D_2} N_{r_j}(a_n)$ and the upper rate of convergence is $a_n \sim n^{-q/(2q+r)}$; r is the largest dimension of the functional components in the model. When there are no interactions $a_n \sim n^{-q/(2q+1)}$.

Assumption (A4) and the results of Yatracos (1988) on lower bounds for convergence rates show that this rate is optimal. Note that the condition $n^{-1/2} = o(a_n)$ is satisfied for all models. $\square$

PROOF OF COROLLARY 5.2. The proof of Proposition 2 in Yatracos (1985) is followed. Recall that the parameter space Θ (of the chosen model) is uniformly bounded. Let $c_1 > 0$ to be determined. Then

$$\begin{aligned} E\|\hat{\theta}_n - \theta\| &= E\|\hat{\theta}_n - \theta\| I(\|\hat{\theta}_n - \theta\| > c_1 a_n) + E\|\hat{\theta}_n - \theta\| I(\|\hat{\theta}_n - \theta\| \le c_1 a_n) \\ &\le c_1 a_n + c_2 P[\|\hat{\theta}_n - \theta\| > c_1 a_n] \le c_1 a_n + c_2 N(a_n)^{-c_3(c_1)} \le c a_n, \end{aligned}$$

for an appropriate choice of c_1 that makes $N(a_n)^{-c_3(c_1)}$ converge to 0 faster than a_n. Assumption (A4) and the results of Yatracos (1988) on lower bounds for convergence rates show that this rate is risk-optimal.

From the relation $P[\|\hat{\theta}_n - \theta\| > c_1 a_n] \le N(a_n)^{-c_3(c_1)}$ and for an appropriate choice of c_1 it holds that $\sum_{n=1}^{\infty} N(a_n)^{-c_3(c_1)} < \infty$. Therefore, by Borel–Cantelli the rates of convergence hold almost surely. $\square$

Acknowledgments. Part of this work appears in T. Nicoleris's Ph.D. dissertation [Nicoleris (1994)]. The authors are indebted to a referee and his student for their constructive comments that helped improving an earlier version of this paper. Many thanks are due to Professor Lucien Le Cam for his

helpful consultation on a technical issue. Thanks are also due to co-Editor Professor John Rice for expediting the last part of the review process.

REFERENCES

BARRON, A., BIRGÉ, L. and MASSART, P. (1997). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*. To appear.

BARRON, A. and COVER, T. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37** 1034–1054.

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

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237.

CHAUDHURI, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Ann. Statist.* **19** 760–777.

CHEN, H. (1991). Estimation of a projection-pursuit regression model. *Ann. Statist.* **19** 1142–1157.

DEVROYE, L. P. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.

DEVROYE, L. P. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1-View.* Wiley, New York.

DONOHO, D. L., JOHNSTONE I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia (with discussion)? *J. Roy. Statist. Soc. Ser. B* **57** 301–369.

DONOHO, D. L. and LIU, R. C. (1988a). The "automatic" robustness of minimum distance functionals. *Ann. Statist.* **16** 552–586.

DONOHO, D. L. and LIU, R. C. (1988b). Pathologies of some minimum distance estimators. *Ann. Statist.* **16** 587–608.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. A* **222** 309–368.

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 700–725.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GRENANDER, U. (1981). *Abstract Inference.* Wiley, New York.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–31.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

KOLMOGOROV, A. N. and TIKHOMIROV, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk.* **14** 3–86. (In Russian.) [Published in English in (1961) *Amer. Math. Soc. Transl. (2)* **17** 277–364.]

LE CAM, L. M. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **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. (1982). Robust estimation via minimum distance methods. *Z. Warsch. Verw. Gebiete* **55** 73–89.

NICOLERIS, T. (1994). Selected topics in estimation. Ph.D. dissertation, Univ. Montréal.

ROUSSAS, G. G. and YATRACOS, Y. G. (1996). Minimum distance regression-type estimates with rates under weak dependence. *Ann. Inst. Statist. Math.* **48** 267–281.

ROUSSAS, G. G. and YATRACOS, Y. G. (1997). Minimum distance estimates with rates under ϕ-mixing. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* (D. Pollard, E. Torgersen and G. L. Yang, eds.) 337–345. Springer, New York.

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

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

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

TRUONG, Y. K. (1989). Asymptotic properties of kernel estimators based on local medians. *Ann. Statist.* **17** 606–617.

TRUONG, Y. K. and STONE, C. J. (1994). Semi-parametric time series regression. *J. Time Ser. Anal.* **15** 405–428.

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. (1988). A lower bound on the error in nonparametric regression type problems. *Ann. Statist.* **16** 1180–1187.

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

YATRACOS, Y. G. (1989b). On the estimation of the derivatives of a function via 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 R^d. *J. Multivariate Anal.* **40** 213–221.

29 AGAPIS ST.
VARKIZA 166 72
ATTIKI
GREECE
E-MAIL: spinik@hol.gr

DÉPARTEMENT DE MATHÉMATIQUES
ET DE STATISTIQUE
UNIVERSITÉ DE MONTRÉAL
CP 6128, SUC. CENTRE VILLE
MONTRÉAL
H3C 3J7 CANADA
E-MAIL: yatracos@dms.umontreal.ca