The Annals of Statistics
1985, Vol. 13, No. 2, 689–705

ADDITIVE REGRESSION AND OTHER NONPARAMETRIC MODELS[1]

By Charles J. Stone

University of California, Berkeley

Let (X, Y) be a pair of random variables such that $X = (X_1, \cdots, X_J)$ and let f by a function that depends on the joint distribution of (X, Y). A variety of parametric and nonparametric models for f are discussed in relation to flexibility, dimensionality, and interpretability. It is then supposed that each $X_j \in [0, 1]$, that Y is real valued with mean μ and finite variance, and that f is the regression function of Y on X. Let f^*, of the form $f^*(x_1, \cdots, x_J) = \mu + f_1^*(x_1) + \cdots + f_J^*(x_J)$, be chosen subject to the constraints $Ef_j^* = 0$ for $1 \le j \le J$ to minimize $E[(f(X) - f^*(X))^2]$. Then f^* is the closest additive approximation to f, and $f^* = f$ if f itself is additive. Spline estimates of f_j^* and its derivatives are considered based on a random sample from the distribution of (X, Y). Under a common smoothness assumption on f_j^*, $1 \le j \le J$, and some mild auxiliary assumptions, these estimates achieve the same (optimal) rate of convergence for general J as they do for $J = 1$.

1. Introduction.

Let (X, Y) be a pair of random variables, each ranging over some space, and let f be a function that depends on the joint distribution of X and Y. As examples, f could be the regression function of Y on X, the hazard rate of the conditional distribution of Y given X, or the density of X; alternatively, f could be the logarithm or logit of any of these functions, where $\mathrm{logit}(p) = \log(p/(1 - p))$.

Suppose as usual that the joint distribution of X and Y is unknown and consider the problem of estimating f based on a random sample from this joint distribution. The parametric approach starts with the assumption of an a priori model for f that contains only finitely many unknown parameters, while the nonparametric approach eschews such an assumption.

Three fundamental aspects of statistical models are flexibility, dimensionality, and interpretability. *Flexibility* is the ability of the model to provide accurate fits in a wide variety of situations, inaccuracy here leading to bias in estimation. *Dimensionality* can be thought of in terms of the variance in estimation, the "curse of dimensionality" being that the amount of data required to avoid an unacceptably large variance increases rapidly with increasing dimensionality. In practice there is an inevitable trade-off between flexibility and dimensionality or, as usually put, between bias and variance. *Interpretability* lies in the potential for shedding light on the underlying structure.

In Section 2, parametric models for f are discussed in relation to these three aspects. Nonparametric and semiparametric models are similarly discussed in Section 3, where a heuristic dimensionality reduction principle is presented. In

Received April 1983; revised October 1984.

[1] This research was supported in part by National Science Foundation Grant MCS 83-01257.

AMS 1980 *subject classifications.* Primary 62G20; secondary 62G05.

Key words and phrases. Nonparametric model, additivity, spline, rate of convergence.

Section 4, a precise special case of this principle is stated for additive regression functions and, more generally, for additive approximations to a not necessarily additive regression function. The proof is given in Sections 5 and 6.

2. Parametric models. Let $X = (X_1, \cdots, X_J)$, where $X_j \in \mathscr{X}_j$ for $1 \le j \le J$. Suppose each $\mathscr{X}_j$ is either a finite set or an interval (i.e., a nondegenerate subinterval of $\mathbb{R}$). If $\mathscr{X}_j$ is finite, X_j is called a "factor," which can be either quantitative or qualitative; in either case the elements of $\mathscr{X}_j$ are referred to as the possible levels of X_j. If there are any factors present, the remaining X_j's are called "covariates."

Suppose now that the $\mathscr{X}_j$'s are each intervals. Two common parametric models for f are

$$\text{Model 1} \qquad f(x_1, \cdots, x_J) = a + \textstyle\sum_1^J b_j x_j$$

and

$$\text{Model 2} \qquad f(x_1, \cdots, x_J) = a + \textstyle\sum_1^J b_j x_j + \sum_1^J \sum_1^J c_{jk} x_j x_k.$$

If f is the regression function, Model 1 corresponds to linear regression and Model 2 corresponds to quadratic regression; if f is the logit of the regression function, Model 1 corresponds to logistic regression; and if f is the logarithm of the density of X, Model 2 corresponds to the multivariate normal density.

Suppose instead that the $\mathscr{X}_j$'s are each finite sets. Here it is common to consider

$$\text{Model 3} \qquad f(x_1, \cdots, x_J) = \mu + \textstyle\sum_1^J f_j(x_j),$$

which is an additive function of the various factors, or

$$\text{Model 4} \qquad f(x_1, \cdots, x_J) = \mu + \textstyle\sum_1^J f_j(x_j) + \sum\sum_{j<k} f_{jk}(x_j, x_k),$$

which also includes the two-factor interaction terms. If f is the regression function, these models are considered in analysis of variance (see Scheffé, 1959), while if f is the (discrete) density of X, these models arise in discrete multivariate analysis (see Bishop et al., 1975; and Haberman, 1978). In such models the functions f_j and f_{jk} will be referred to as *functional components.*

More generally, suppose that the first L of the $\mathscr{X}_j$'s are finite sets and the remaining $\mathscr{X}_j$'s are intervals, where $1 \le L < J$. (Such situations are especially common in social science.) In this context it is natural to consider models such as those that arise in the analysis of covariance; e.g.,

$$\text{Model 5} \qquad f(x_1, \cdots, x_J) = \mu + \textstyle\sum_1^L f_j(x_j) + \sum_{L+1}^J b_j x_j.$$

These and similar models are the bread-and-butter of applied statistics, even though they are rarely justified on theoretical grounds. The orthodox approach to model selection is by means of hypothesis testing; e.g., Model 1 vs. Model 2, or Model 3 vs. Model 4. But this approach has its limitations. In particular, one does not know in advance that Model 2 or Model 4 is valid; and there may not be enough data to perform such a test (see Scheffé, 1959, for a discussion of the one-degree-of-freedom test for additivity due to Tukey, 1949). Often there is not enough data to fit high-dimensional models such as Model 2 or 4, so lower-dimensional models such as Model 1, 3 or 5 are automatically assumed (as in models involving Latin and Greco-Latin squares and balanced incomplete blocks). In the chapter on Latin squares, John (1971) discusses bias in estimation of treatment effects due to ignored interaction effects. On the other hand, Fisher and McDonald (1978) assert that, when there is enough data to test for the presence of interaction effects, in a large fraction (but by no means all) of the time they are found not to be statistically significant.

Even when there is enough data to fit models such as Model 2 or 4 and when the presence of interaction terms is detected, say, by hypothesis testing, additive models such as Models 1, 3 and 5 may be preferable because of greater interpretability. In particular, in the regression context additive models allow for the following interpretation: if x_j is changed to x_j', then, all other variables remaining the same, the mean of Y is increased by an amount which depends only on x_j and x_j'; under the linear, additive Model 1 the amount of increase is a linear function of $(x_j' - x_j)$. In their discussion of additive regression models involving two factors, Box et al. (1978) state that

> While blind faith in a particular model is foolhardy, refusal to associate data with *any* model is to eschew a powerful tool. As implied earlier, a middle course may be followed. On the one hand, inadequacies in proposed models should be looked for; on the other, if a model appears reasonably appropriate, advantage should be taken of the greater simplicity and clarity of interpretation that it provides.

Models 3 and 4 can be fitted by solving the normal equations for the associated dummy variable regression problem. Andrews et al. (1967) described a computer program, Multiple Classification Analysis (MCA), which uses instead a computationally efficient iterative method to fit the additive Model 3. Morgan and Sonquist (1963) proposed a tree structured alternative to Model 4 for detecting and interpreting interaction effects. Their automatic interaction detection technique was implemented in a program called AID (see Sonquist et al., 1971; Fielding, 1977; and the discussion of a similar program, CART, in Breiman et al., 1984). Sonquist (1970) compared AID with MCA and suggested a strategy for using the two programs jointly. Both programs have been very popular among sociologists. They were developed by the Institute for Social Research, University of Michigan, and are currently available in the software package OSIRIS distributed by that organization. (SEARCH, in that package, is an outgrowth of AID.)

3. Nonparametric and semiparametric models. Suppose now that the $\mathscr{X}_j$'s are each intervals. (The following discussion also applies, with obvious modifications, when some of these sets are intervals and the others are finite.) At the opposite end of the spectrum from Models 1 and 2, f can be totally unrestricted (perhaps subject to some smoothness assumptions). Alternatively, one can consider, say, the additive Model 3 with unrestricted functional components. As a further step toward parametric modelling, one could impose parametric restrictions on some of these functional components as in Model 5. For example, if f is the logarithm of the hazard rate of the conditional distribution of the failure time Y given X, then the additive model given by

$$\text{Model 6} \qquad f(x_1, \cdots, x_J, y) = g(y) + \textstyle\sum_1^J b_j x_j$$

is the famous proportional hazards model of Cox (1972). See Lawton et al. (1972), Stützle et al. (1980), Oakes (1981), Begun et al. (1983), Engle et al. (1982), and Wahba (1983) for other such semiparametric models.

An interesting generalization of the additive model is

$$\text{Model 7} \qquad f(x_1, \cdots, x_J) = \textstyle\sum_1^V f_v(\sum_1^J b_{vj} x_j),$$

which arises in projection pursuit regression (PPR) and (log) density estimation (see Friedman and Stuetzle, 1981; Friedman et al., 1984; Friedman et al., 1983; and Huber, 1985). Like Model 4, this model is more flexible than the additive model. Indeed, roughly speaking, the projection pursuit model is so flexible that if V is permitted to be indefinitely large, it can yield an arbitrarily good approximation to any given function. The individual functional components f_v occurring in the projection pursuit model are readily graphed; but when $V > 1$, the model itself is hard to interpret. A less flexible, but more easily interpretable, generalization of the additive model is

$$\text{Model 8} \qquad f(x_1, \cdots, x_J) = g(\textstyle\sum_1^J f_j(x_j)),$$

which was considered by Winsberg and Ramsay (1980) (see also de Leeuw et al., 1976; Young et al., 1976; and Breiman and Friedman, 1985).

Friedman and Stuetzle (1981) also discussed "projection selection," which is PPR restricted to be additive, and they suggested strategies for using projection selection and unrestricted PPR jointly. Tibshirani (1983) discussed the extension of Cox's proportional hazards model (Model 6) to the unrestricted additive model. Hastie (1983) discussed the additive model for the logistic regression function when Y is an indicator variable, and Hastie (1984) pointed out the usefulness in graphical model diagnostics of additive estimates of the regression function or of its logit. Previously, Breiman and Stone (1978) proposed various additive regression estimates as modifications of (nonadditive) tree structured regression. They were motivated both by the successful meteorological application of an ad hoc additive regression technique (Zeldin and Thomas, 1975) and by the realization that unrestricted multivariate nonparametric regression is hard to interpret, as well as being subject to the curse of dimensionality.

It is convenient to think of an arbitrary function of d real variables as being "d-dimensional." Consider a nonparametric model in which f is defined explicitly in terms of other functions, at least one of which is d-dimensional and none of which are more than d-dimensional. Such a model will also be thought of as being d-dimensional. Accordingly, Models 3, 5, 6, 7 (for fixed V) and 8 are one-dimensional, while Model 4 is two-dimensional.

Model dimensionality, as just defined, is relevant to optimal rates of convergence for nonparametric function estimation (see Stone, 1982). In the absence of a restrictive model for f, the optimal rate of convergence in an L^2 norm is typically of the form n^{-2r}; here $r = (p - m)/(2p + J)$, p being a measure of the assumed smoothness of f and m being the order of the derivative of f that is being estimated ($m = 0$ if f itself is being estimated). It is quite plausible that under the additional restriction of a d-dimensional model for f ($d < J$), the optimal rate of convergence is of the same from with J replaced by d in the definition of r. But it is not at all obvious that this *heuristic dimensionality reduction principle* can be established rigorously as a general theorem applying, say, to projection pursuit regression. (Huber, 1985, points out that the sampling theory of PPR is practically nonexistent.) The discussion in Section 4 below suggests more general versions of this principle. The principle is implicit in Question 2 of Stone (1982).

4. Rates of convergence of additive regression. Suppose from now on that $\mathscr{X}_j = [0, 1]$ for $1 \le j \le J$ and that $EY^2 < \infty$. Let f be the regression function of Y on X; so that $f(x) = E(Y \mid X = x)$ for $x \in C = [0, 1]^J$. Set $\mu = EY = Ef(X)$.

If f is additive, it can be written in the form

$$f(x_1, \cdots, x_J) = \mu + \textstyle\sum_1^J f_j(x_j),$$

where $Ef_j(X_j) = 0$ for $1 \le j \le J$. Under the mild Condition 1 below the functional components f_j are uniquely determined up to sets of measure zero (see Lemma 1 in Section 5); and there is at most one continuous version of each such function.

Even if f is not genuinely additive, an additive approximation to f may be sufficiently accurate for a given application as well as being readily interpretable. Let f_j^*, $1 \le j \le J$, be chosen subject to the constraints $Ef_j^*(X_j) = 0$ for $1 \le j \le J$ to minimize $E[(f^*(X) - f(X))^2]$, where

$$f^*(x_1, \cdots, x_J) = \mu + \textstyle\sum_1^J f_j^*(x_j).$$

(These functions exist under Condition 1 below. See Lemma 1 in Section 5.) Then f^* is the closest additive approximation to f in the sense of mean squared error. Again, under Condition 1 below, the functional components f_j^* are uniquely determined up to sets of measure zero and there is at most one continuous version of each such function. If f *is* genuinely additive, of course, then $f^* = f$ and $f_j^* = f_j$ for $1 \le j \le J$.

A smoothness assumption, Condition 3 below, will be imposed on the functions f_j^*, $1 \le j \le J$. Then a special case of the dimensionality reduction principle will be established by determining the rates of convergence for spline estimates of f_j^* and its derivatives and for the corresponding estimates of f^* based on a random sample from the distribution of (X, Y).

Three conditions are required for the statements of the main results.

CONDITION 1. The distribution of X is absolutely continuous and its density g is bounded away from zero and infinity on C.

It follows from Condition 1 that the marginal density g_j of X_j is bounded away from zero and infinity on $[0, 1]$. Set $\operatorname{Var}(Y \mid X = x) = E((Y - f(x))^2 \mid X = x)$.

CONDITION 2. f and $\mathrm{Var}(Y \mid X = \cdot)$ are bounded on C.

Let k be a nonnegative integer, let $\beta \in (0, 1]$ be such that $p = k + \beta > .5$, and let $M \in (0, \infty)$. Let $\mathscr{H}$ denote the collection of functions h on $[0, 1]$ whose kth derivative, $h^{(k)}$, exists and satisfies the Hölder condition with exponent β:

$$|h^{(k)}(t') - h^{(k)}(t)| \le M|t' - t|^{\beta} \quad \text{for} \quad 0 \le t, \quad t' \le 1.$$

CONDITION 3. $f_j^* \in \mathscr{H}$ for $1 \le j \le J$.

Let $(X_1, Y_1), (X_2, Y_2), \cdots$ denote independent random pairs, each having the same distribution as (X, Y) and write X_i as $(X_{i1}, \cdots, X_{ij})$. We will consider additive spline estimates $\hat{f}_n$ of f^* based on the random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ of size n. Set $\bar{Y}_n = (Y_1 + \cdots + Y_n)/n$.

Let N_n denote a positive integer and let $I_{n\nu}$, $1 \le \nu \le N_n$, denote the subintervals of $[0, 1]$ defined by $I_{n\nu} = [(\nu - 1)/N_n, \nu/N_n)$ for $1 \le \nu < N_n$ and $I_{nN_n} = [1 - N_n^{-1}, 1]$. Let ℓ and ℓ' be integers such that $\ell \ge k$ and $\ell > \ell'$. Let $\mathscr{S}_n$ denote the collection of functions s on $[0, 1]$ such that

(i) the restriction of s to $I_{n\nu}$ is a polynomial of degree ℓ (or less) for $1 \le \nu \le N_n$;

and, if $\ell' \ge 0$,

(ii) s is ℓ'-times continuously differentiable on $[0, 1]$.

A function satisfying (i) is called a piecewise polynomial; if $\ell = 0$, it is piecewise constant. A function satisfying (i) and (ii) is called a spline. Typically, splines are considered with $\ell' = \ell - 1$ and then called linear, quadratic or cubic splines accordingly as $\ell = 1, 2$ or 3. Wegman and Wright (1983) give a broad survey of works on splines in statistics.

Let $\hat{f}_n$, of the form

$$\hat{f}_n(x_1, \cdots, x_J) = \bar{Y}_n + \textstyle\sum_1^J \hat{f}_{nj}(x_j),$$

be chosen subject to the constraints that $\hat{f}_{nj} \in \mathscr{S}_n$ and $\sum_1^n \hat{f}_{nj}(X_{ij}) = 0$ for $1 \le j \le J$ to minimize the residual sum of squares $\sum_1^n (Y_i - \hat{f}_n(X_i))^2$. We call the estimate $\hat{f}_n$ of f^* of f^* an *additive spline*. If $\ell = 0$, we borrow from Tukey (1961) and call $\hat{f}_n$ an *additive regressogram*. The numerical minimization is readily solved by using B-splines (see de Boor, 1978; or Powell, 1981); either the normal equations can be solved or iterative techniques can be employed as in MCA. It follows easily from Lemma 3 in Section 5 that under Condition 1 and the condition on N_n in Theorem 1 below, the solutions $\hat{f}_{nj}$, $1 \le j \le J$, to the constrained minimization problem are uniquely determined except on an event whose probability tends to zero with n.

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

$$\lim_{C \to \infty} \lim\, \sup_n \Pr(Z_n| > cb_n) = 0.$$

Let $\| \varphi \|$ denote the L^2 norm of a function φ on C, defined by $\| \varphi \|^2 = E\varphi^2(X) = \int_C \varphi^2(x)g(x)\,dx$. For $1 \le j \le J$ let $\| h \|_j^2$ denote the L^2 norm of a function h on $[0, 1]$, defined by $\| h \|_j^2 = Eh^2(X_j) = \int_0^1 h^2(x_j)g_j(x_j)\,dx_j$. Set $\gamma = 1/(2p + 1)$ and $r = p/(2p + 1)$. Let m be a nonnegative integer such that $m \le k$ and set $r_m = (p - m)/(2p + 1)$.

THEOREM 1. *Suppose that Conditions* 1–3 *hold and that $N_n \sim n^\gamma$. Then*

$$(1) \qquad E(\| \hat{f}_{nj}^{(m)} - (f_j^*)^{(m)} \|_j^2 \,|\, X_1, \cdots, X_n) = O_{\mathrm{pr}}(n^{-2r_m}) \quad \textit{for} \quad 1 \le j \le J.$$

Since Y has mean μ and finite second moment,

$$(2) \qquad E((\bar{Y}_n - \mu)^2 \,|\, X_1, \cdots, X_n) = O_{\mathrm{pr}}(n^{-1}) \mp 0_{\mathrm{pr}}(n^{-2r}).$$

Thus Theorem 1 has the following consequence.

COROLLARY 1. *Suppose that Conditions* 1–3 *hold and that $N_n \sim n^\gamma$. Then*

$$(3) \qquad E(\| \hat{f}_{nj} - f_j^* \|_j^2 \,|\, X_1, \cdots, X_n) = O_{\mathrm{pr}}(n^{-2r}) \quad \textit{for} \quad 1 \le j \le J$$

and

$$(4) \qquad E(\| \hat{f}_n - f^* \|^2 \,|\, X_1, \cdots, X_n) = O_{\mathrm{pr}}(n^{-2r}).$$

The rates of convergence in (1), (3) and (4) do not depend on J. It is clear from the results in Stone (1982) for $J = 1$ that these rates are optimal.

5. Proof of Theorem 1. Throughout the remainder of the paper, it will be assumed that Condition 1 is satisfied. Let b and B be positive numbers such that

$$(5) \qquad b \le g \le B \quad \text{on} \quad C.$$

Set $\delta = (1 - bB^{-1})^{1/2} < 1$. Let $\mathrm{SD}(Z)$ denote the standard deviation of a random variable Z.

LEMMA 1. *Let $V_j = h_j(X_j)$ be random variables such that $\sum_1^J V_j$ has finite second moment. Then each V_j has finite second moment. Also*

$$\mathrm{SD}(V_1 + \cdots + V_j) \ge ((1 - \delta)/2)^{(j-1)/2}(\mathrm{SD}(V_1) + \cdots + \mathrm{SD}(V_j)) \quad \text{for} \quad 1 \le j \le J.$$

PROOF. By Condition 1, the first conclusion reduces to that for independent V_j's, which follows from Theorem 6.4.1 of Chung (1974). In proving the second conclusion, it can be assumed that $EV_j = 0$ for $1 \le j \le J$. Set $\sigma_j = \mathrm{SD}(V_j)$ and $\tau_j = \mathrm{SD}(V_1 + \cdots + V_j)$. The desired result is trivially true for $j = 1$. Suppose it is true for j, where $1 \le j < J$. It will now be shown to hold for $j + 1$. If $\tau_j = 0$, then $\sigma_1 = \cdots = \sigma_j = 0$ and the desired conclusion reduces to the obvious inequality $0 \le (1 - \delta)/2 \le 1$. If $\sigma_{j+1} = 0$, the desired conclusion follows from that for j. Suppose instead that $\tau_j > 0$ and $\sigma_{j+1} > 0$, and let ρ denote the correlation between $V_1 + \cdots + V_j$ and V_{j+1}. Set $W = (X_1, \cdots, X_j)$ and $Z = X_{j+1}$. Let $g_{W,Z}$ and g_Z denote, respectively, the joint density of W, Z and the marginal density of Z. Then $g_{W,Z} \geq b$ on $[0, 1]^{j+1}$ and $g_Z \leq B$ on $[0, 1]$. Write $V_1 + \cdots + V_j$ as $\phi(W)$ and V_{j+1} as $\psi(Z)$. Then

$$\begin{aligned}(1 - \rho^2)\sigma_{j+1}^2 &= \min_\beta E((\psi(Z) - \beta\phi(W))^2) \\ &= \min_\beta \int \int (\psi(z) - \beta\phi(w))^2 g_{W,Z}(w, z)\, dw\, dz \\ &> bB^{-1}\min_\beta \int \left[\int (\psi(z) - \beta\phi(w))^2 g_Z(z)\, dz\right] dw \geq bB^{-1}\sigma_{j+1}^2,\end{aligned}$$

so that $\rho^2 \leq 1 - bB^{-1}$ and hence $\rho \geq -\delta$. Consequently,

$$\begin{aligned}\tau_{j+1}^2 &= \tau_j^2 + 2\rho\tau_j\sigma_{j+1} + \sigma_{j+1}^2 \\ &\geq ((1 + \rho)/2)(\tau_j + \sigma_{j+1})^2 \\ &\geq ((1 - \delta)/2)(((1 - \delta)/2)^{(j-1)/2}(\sigma_1 + \cdots + \sigma_j) + \sigma_{j+1})^2 \\ &\geq ((1 - \delta)/2)^j(\sigma_1 + \cdots + \sigma_{j+1})^2.\end{aligned}$$

Therefore, the desired result holds for $j + 1$. By induction it holds for $1 \leq j \leq J$ as desired.

Let $\Pr_n$ correspond to the empirical distribution of $X_1, \cdots, X_n$; so that $\Pr_n(X_j \in I_{n\nu})$ is n^{-1} times the cardinality of $\{i: 1 \leq i \leq n \text{ and } X_{ij} \in I_{n\nu}\}$. Recall that $\gamma = 1/(2p + 1) < .5$ since $p > .5$. Thus if $N_n \sim n^\gamma$, then

$$\lim_n n^{a-1}N_n^2 = 0 \quad \text{for some} \quad a > 0. \tag{6}$$

The next result follows easily from (5) and Bernstein's inequality (see Theorem 3 of Hoeffding, 1963) applied to the Binomial distribution.

LEMMA 2. *Suppose* (6) *holds and let* $\varepsilon > 0$. *Then, except on an event whose probability tends to zero with n,*

$$|\Pr_n(X_j \in I_{n\nu}) - \Pr(X_j \in I_{n\nu})| \leq \varepsilon \Pr(X_j \in I_{n\nu}) \quad \textit{for} \quad 1 \leq j \leq J \quad \textit{and} \quad 1 \leq \nu \leq N_n$$

and

$$\begin{aligned}&|\Pr_n(X_{j_1} \in I_{n\nu_1}, X_{j_2} \in I_{n\nu_2}) - \Pr(X_{j_1} \in I_{n\nu_1}, X_{j_2} \in I_{n\nu_2})| \\ &\qquad \leq \varepsilon \Pr(X_{j_1} \in I_{n\nu_1}, X_{j_2} \in I_{n\nu_2}) \\ &\qquad \textit{for} \quad 1 \leq j_1, \quad j_2 \leq J, \quad j_1 \neq j_2, \quad \textit{and} \quad 1 \leq \nu_1, \nu_2 \leq N_n.\end{aligned}$$

For $1 \leq j \leq J$ and s a function on $[0, 1]$, let $V_j(s)$ be the function on C defined by

$$V_j(s)(x_1, \cdots, x_J) = s(x_j).$$

Let SD_n denote the standard deviation corresponding to the empirical distribution of $X_1, \cdots, X_n$. The proof of the next result will be given in Section 6.

LEMMA 3. *Suppose* (6) *holds and let* $\delta_1 \in (\delta, 1)$. *Then, except on an event whose probability tends to zero with n, the following statement is valid for all choices of* $s_1, \cdots, s_J \in S_n$:

$$\mathrm{SD}_n(\textstyle\sum_1^J V_j(s_j)) \geq ((1-\delta_1)/2)^{(J-1)/2} \sum_1^J \mathrm{SD}_n(V_j(s_j)).$$

If the constrained minimization problem arising in the definition of $\hat{f}_n$ has a unique solution, then

$$\hat{f}_{nj}(x_j) = \textstyle\sum_{i=1}^n W_{nji}(x_j) Y_i,$$

where the functions W_{nji} on [0, 1] are uniquely determined. Set

$$| W_{nj}(x_j) |^2 = \textstyle\sum_{i=1}^n W_{nji}^2(x_j).$$

LEMMA 4. *Suppose* (6) *holds. Then the constrained minimization problem has a unique solution, except on an event whose probability tends to zero with n. Moreover,*

$$\sup_{0 \leq x_j \leq 1} | W_{nj}(x_j) |^2 = O_{\mathrm{pr}}(n^{-1} N_n) \quad \textit{for} \quad 1 \leq j \leq J.$$

PROOF. Think of $X_1, \cdots, X_n$ as fixed. Consider an alternative experiment having n^J cases $(\tilde{X}_\iota, \tilde{Y}_\iota)$, $\iota = (\iota_1, \cdots, \iota_J)$, where $1 \leq \iota_1, \cdots, \iota_J \leq n$ and $\tilde{X}_\iota \in \mathbb{R}^J$ is defined by $\tilde{X}_{\iota j} = X_{\iota_j j}$ for $1 \leq j \leq J$; so that the jth coordinate of $\tilde{X}_\iota$ is the same as the jth coordinate of X_{ι_j}. Note that, for $1 \leq j \leq J$ and a function h defined on [0, 1],

$$\textstyle\sum_\iota h(\tilde{X}_{\iota j}) = \sum_\iota h(X_{\iota_j j}) = n^{J-1} \sum_{i=1}^n h(X_{ij}).$$

Set $\bar{Y}_n = n^{-J} \sum_\iota \tilde{Y}_\iota$ and

$$\bar{Y}_{ijn} = n^{1-J} \textstyle\sum_{\iota:\iota_j = i} \tilde{Y}_\iota \quad \text{for} \quad 1 \leq j \leq J \quad \text{and} \quad 1 \leq i \leq n.$$

Then

$$\textstyle\sum_{i=1}^n (\bar{Y}_{ijn} - \bar{Y}_n) = 0 \quad \text{for} \quad 1 \leq j \leq J.$$

Let $\hat{f}_n$, of the form

$$\hat{f}_n(x_1, \cdots, x_J) = \bar{Y}_n + \textstyle\sum_1^J \hat{f}_{nj}(x_j),$$

be chosen subject to the constraints that $\hat{f}_{nj} \in S_n$ and $\sum_\iota \hat{f}_{nj}(\tilde{X}_{\iota j}) = 0$ for $1 \leq j \leq J$ to minimize the residual sum of squares $S^2 = \sum_\iota (\tilde{Y}_\iota - \hat{f}_n(\tilde{X}_\iota))^2$. Observe that the constraint $\sum_\iota \hat{f}_{nj}(\tilde{X}_{\iota j}) = 0$ is equivalent to the constraint $\sum_1^n \hat{f}_{nj}(X_{ij}) = 0$; observe also that

$$S^2 = n^{J-1} \textstyle\sum_1^J \sum_1^n (\bar{Y}_{ijn} - \bar{Y}_n - \hat{f}_{nj}(X_{ij}))^2 + \sum_\iota (\tilde{Y}_\iota - \bar{Y}_n - \sum_1^J (\bar{Y}_{\iota_n jn} - \bar{Y}_n))^2,$$

since the omitted cross-product terms equal zero. Thus $\hat{f}_{nj}$ minimizes

$$S_j^2 = \textstyle\sum_1^n (\bar{Y}_{ijn} - \bar{Y}_n - \hat{f}_{nj}(X_{ij}))^2$$

subject to the constraints that $\hat{f}_{nj} \in \mathcal{S}_n$ and $\sum_1^n \hat{f}_{nj}(X_{ij}) = 0$; equivalently, $\hat{f}_{nj}$ minimizes S_j^2 subject to the constraint that $\hat{f}_{nj} \in \mathcal{S}_n$.

Suppose now that $\ell' = -1$ in the definition of $\mathcal{S}_n$. Let $1 \le j \le J$ and $1 \le \nu \le N_{n\nu}$ and set

$$\mathcal{I}_{nj\nu} = \{i\colon 1 \le i \le n \text{ and } X_{ij} \in I_{n\nu}\}.$$

Then $\hat{f}_{nj}$ reduces to a polynomial $\hat{P}_{nj\nu}$ of degree ℓ (or less) on $I_{nj\nu}$; $\hat{P}_{nj\nu}$ minimizes

$$\textstyle\sum_{\mathcal{I}_{nj\nu}} (\bar{Y}_{ijn} - \bar{Y}_n - P(X_{ij}))^2$$

among all such polynomials. Suppose this minimization problem has a unique solution. Then

$$\hat{f}_{nj}(x_j) = \hat{P}_{nj\nu}(x_j) = \textstyle\sum_{\mathcal{I}_{nj\nu}} \tilde{V}_{nji}(x_j)(\bar{Y}_{ijn} - \bar{Y}_n), \quad x_j \in I_{n\nu},$$

where the functions $\tilde{V}_{nji}$ on $I_{n\nu}$ are uniquely determined. Set

$$|\tilde{V}_{nj}(x_j)|^2 = \textstyle\sum_{\mathcal{I}_{nj\nu}} \tilde{V}_{nji}^2(x_j), \quad x_j \in I_{n\nu}.$$

It follows from (5), (6), Hoeffding's inequality (Theorem 1 of Hoeffding, 1963) and the argument on pages 1354–1355 of Stone (1980) that, except on an event whose probability tends to zero with n, these minimization problems all have a unique solution; and

$$\sup_{0 \le x_j \le 1} |\tilde{V}_{nj}(x_j)|^2 = O_{\mathrm{pr}}(n^{-1}N_n) \quad \text{for} \quad 1 \le j \le J. \tag{7}$$

Write

$$\hat{f}_{nj}(x_j) = \textstyle\sum_{\mathbf{i}} \tilde{W}_{nj\mathbf{i}}(x_j)\tilde{Y}_{\mathbf{i}}$$

and set

$$|\tilde{W}_{nj}(x_j)|^2 = \textstyle\sum_{\mathbf{i}} \tilde{W}_{nj\mathbf{i}}^2(x_j) \quad \text{for} \quad 0 \le x_j \le 1.$$

It follows from (7) that

$$\sup_{0 \le x_j \le 1} |\tilde{W}_{nj}(x_j)|^2 = O_{\mathrm{pr}}(n^{-J}N_n) \quad \text{for} \quad 1 \le j \le J. \tag{8}$$

If ℓ' is changed from -1 to a nonnegative number, then additional constraints are imposed on the overall minimization problem. Thus the probability of nonuniqueness and the quantity $|\tilde{W}_{nj}(x_j)|^2$ under uniqueness are both reduced or unaltered. Consequently, even without the supposition that $\ell' = -1$ it is true that, except on an event whose probability tends to zero with n, the overall constrained minimization problem has a unique solution; and it is true that (8) holds.

Consider a function ϕ on C of the form

$$\phi(x_1, \cdots, x_J) = u + \textstyle\sum_1^J s_j(x_j),$$

where u is any real number and

$$s_j \in \mathcal{S}_n \quad \text{and} \quad \textstyle\sum_{i=1}^n s_j(X_{ij}) = 0 \quad \text{for} \quad 1 \le j \le J.$$

Observe that

$$\textstyle\sum_{\mathbf{i}} \phi^2(\tilde{X}_{\mathbf{i}}) = n^J u^2 + n^{J-1} \sum_{j=1}^J \sum_{i=1}^n s_j^2(X_{ij})$$

and

$$\sum_{i=1}^{n} \phi^2(X_i) = nu^2 + \sum_{i=1}^{n} (\sum_{j=1}^{J} s_j(X_{ij}))^2.$$

Thus it follows from Lemma 3 that there is a $\gamma \in (0, 1)$ such that, except on an event whose probability tends to zero with n,

$$\sum_{i=1}^{n} \phi^2(X_i) \geq \gamma n^{1-J} \sum_{i'} \phi^2(\tilde{X}_i)$$

for every real number u and, for $1 \leq j \leq J$, for all choices of $s_j \in \mathscr{S}_n$ such that $\sum_1^n s_j(X_{ij}) = 0$. The conclusion of the lemma now follows from (8) and Lemma 2.1 of Ehrenfeld (1956), a result involving the comparison of two experiments.

Let $\| h \|_\infty = \sup_{0 \leq t \leq 1} | h(t) |$ denote the supnorm of a function h on [0, 1].

LEMMA 5. *For each $h \in \mathscr{H}$ and $n \geq 1$ there is an $s \in \mathscr{S}_n$ with $\| s - h \|_\infty \leq M_1 N_n^{-p}$; here M_1 is some fixed positive constant.*

This result is due to de Boor (1968); see also de Boor (1978) or Powell (1981).

LEMMA 6. *Consider a Hilbert space with norm $\| \;\|$. Let P_j denote the orthogonal projection onto a subspace V_j for $0 \leq j \leq J$. Suppose $v = \sum_0^J v_j$, where $v_j \in V_j$ for $0 \leq j \leq J$. Then*

$$\| v \|^2 \leq (\max_{0 \leq j \leq J} \| v_j \|)(\sum_0^J \| P_j v \|).$$

PROOF. Let "$\cdot$" denote the Hilbert space inner product. Observe that $v_j \cdot (v - P_j v) = 0$ for $0 \leq j \leq J$. Thus

$$\| v \|^2 = \sum_j v_j \cdot v = \sum_j v_j \cdot P_j v \leq \sum_j \| v_j \| \, \| P_j v \|,$$

which yields the desired result.

Set $\bar{\mu}_n = n^{-1} \sum_1^n f(X_i)$. Since $f(X)$ has mean μ and finite second moment,

$$(\bar{\mu}_n - \mu)^2 = O_{\mathrm{pr}}(n^{-1}) = O_{\mathrm{pr}}(n^{-2r}). \qquad (9)$$

Suppose (6) holds. Let $\bar{f}_n$, of the form

$$\bar{f}_n(x_1, \cdots, x_J) = \bar{\mu}_n + \sum_1^J \bar{f}_{nj}(x_j),$$

be a solution to the constrained minimization problem with Y_i replaced by $f(X_i)$; by the first conclusion to Lemma 4, this problem has a unique solution except on an event whose probability tends to zero with n.

Let $\| \phi \|_n$ denote the L^2 norm of a function ϕ on C with respect to the empirical distribution of $X_1, \cdots, X_n$; so that $\| \phi \|_n^2 = n^{-1} \sum_1^n \phi^2(X_i)$. Similarly, for $1 \leq j \leq J$, let $\| h \|_{nj}$ denote the L^2 norm of a function h on [0, 1] with respect to the empirical distribution of $X_{1j}, \cdots, X_{nj}$.

LEMMA 7. *Suppose that Conditions 1 and 3 hold and that $N_n \sim n^\gamma$. Then*

$$\| \bar{f}_{nj} - f_j^* \|_{nj}^2 = O_{\mathrm{pr}}(n^{-2r}) \quad \text{for} \quad 1 \leq j \leq J.$$

PROOF. Since $f = f^* + f - f^*$, $(f^*)^* = f^*$ and $(f - f^*)^* = 0$, it suffices to verify the lemma when $f = f^*$ and when $f^* = 0$. Suppose first that $f = f^*$ and set $f_j = f_j^*$.

Then $f(x_1, \cdots, x_J) = \mu + \sum_1^J f_j(x_j)$. For $1 \le j \le J$ choose $s_{nj} \in \mathscr{S}_n$ such that $\| f_j - s_{nj} \|_\infty \le M_1 n^{-r}$, which is possible by Lemma 5. Define f_n by $f_n(x_1, \cdots, x_J) = \mu + \sum_1^J s_{nj}(x_j)$. Then $\| f_n - f \|_\infty = O(n^{-r})$ and hence $\| f_n - f \|_n^2 = O(n^{-2r})$. Now $\| \bar{f}_n - f \|_n \le \| f_n - f \|_n$, so that $\| \bar{f}_n - f \|_n^2 = O_{\rm pr}(n^{-2r})$ and hence

$$(10) \qquad \| \bar{f}_n - f_n \|_n^2 = O_{\rm pr}(n^{-2r}).$$

For $1 \le j \le J$ set $\bar{\mu}_{nj} = n^{-1} \sum_1^n f_j(X_{ij})$ and observe that f_j is bounded by Condition 3 and hence that $\bar{\mu}_{nj}^2 = O_{\rm pr}(n^{-1}) = O_{\rm pr}(n^{-2r})$; also set $\bar{\nu}_{nj} = n^{-1} \sum_1^n s_{nj}(X_{ij})$ and observe that $\bar{\nu}_{nj}^2 = O_{\rm pr}(n^{-2r})$. It now follows from (10) and Lemma 3 that, for $1 \le j \le J$, $\| \bar{f}_{nj} - s_{nj} - \bar{\nu}_{nj} \|_{nj}^2 = O_{\rm pr}(n^{-2r})$ and hence $\| \bar{f}_{nj} - s_{nj} \|_{nj}^2 = O_{\rm pr}(n^{-2r})$; consequently, $\| \bar{f}_{nj} - f_j^* \|_{nj}^2 = \| \bar{f}_{nj} - f_j \|_{nj}^2 = O_{\rm pr}(n^{-2r})$ as desired.

Suppose next that $f^* = 0$ or, equivalently, that $E(f(X) \mid X_j) = 0$ for $1 \le j \le J$ and hence that $\mu = 0$. Then $\bar{\mu}_n = O_{\rm pr}(n^{-1}) = O_{\rm pr}(n^{-2r})$ by (9). For $1 \le j \le J$ let $\mathring{f}_{nj}$, of the form $\mathring{f}_{nj}(x_1, \cdots, x_J) = s_{nj}(x_j)$, be chosen subject to the constraint that $s_{nj} \in \mathscr{S}_n$ and $n^{-1} \sum_1^n s_{nj}(X_{ij}) = 0$ to minimize $\| \mathring{f}_{nj} - f \|_n^2$. It follows from the assumption that $f^* = 0$, the boundness of f and Lemma 4 (with $J = 1$) that

$$(11) \qquad \| \mathring{f}_{nj} \|_n^2 = O_{\rm pr}(n^{-2r}) \quad \text{for} \quad 1 \le j \le J.$$

Temporarily, think of $X_1, \cdots, X_n$ as fixed. Consider the Hilbert space of functions defined on the range of $X_1, \cdots, X_n$ with the empirical norm $\| \ \|_n$ defined by $\| \phi \|_n^2 = n^{-1} \sum_1^n \phi^2(X_i)$. Let V_0 denote the space of constant functions and, for $1 \le j \le J$, let V_j denote the space of functions of the form $\phi(x_1, \cdots, x_J) = s(x_j)$, where $s \in \mathscr{S}_n$ and $\sum_1^n s(X_{ij}) = 0$. Let P_j denote the orthogonal projection onto V_j for $0 \le j \le J$. Let v_0 denote the constant function $\bar{\mu}_n$ and, for $1 \le j \le J$, let v_j denote the function defined by $v_j(x_1, \cdots, x_J) = \bar{f}_{nj}(x_j)$. Set $v = \bar{f} = \sum_0^J v_j$. It follows from (9) with $\mu = 0$ and (11) that

$$(12) \qquad \| P_j v \|_n^2 = O_{\rm pr}(n^{-2r}) \quad \text{for} \quad 0 \le j \le J.$$

According to Lemma 3 there is a constant $\gamma \in (0, 1]$ such that, except on an event whose probability tends to zero with n,

$$\| \textstyle\sum_1^J v_j \|_n^2 \ge \gamma \sum_1^J \| v_j \|_n^2$$

and hence

$$\| v \|_n^2 = \| \textstyle\sum_0^J v_j \|_n^2 = \| v_0 \|_n^2 + \| \sum_1^J v_j \|_n^2 \ge \gamma \max_{0 \le j \le J} \| v_j \|_n^2.$$

Thus, by Lemma 6, $\| v \|_n^2 \le \gamma^{-1} (\sum_0^J \| P_j v \|_n)^2$; so, by (12), $\| v \|_n^2 = O_{\rm pr}(n^{-2r})$. It now follows from Lemma 3 applied to the norm $\| \ \|_n$ that $\| v_j \|_n^2 = O_{\rm pr}(n^{-2r})$ for $1 \le j \le J$, which is equivalent to the conclusion of the lemma when $f^* = 0$.

The proofs of the next two results will be given in Section 6.

LEMMA 8. *There is an $M_2 \in (0, \infty)$ such that*

$$\| s^{(m)} - h^{(m)} \|_j^2 \le M_2 (N_n^{2(m-p)} + N_n^{2m} \| s - h \|_j^2)$$

$$\textit{for} \quad 1 \le j \le J, \quad h \in \mathscr{H}, \quad n \ge 1 \quad \textit{and} \quad s \in \mathscr{S}_n.$$

LEMMA 9. *Suppose that*

$$\lim_n n^{a-1} N_n = 0 \quad \textit{for some} \quad a > 0. \tag{13}$$

Then there is an $M_3 \in (0, \infty)$ such that, except on an event whose probability tends to zero with n,

$$\| s - h \|_j^2 \le M_3 (N_n^{-2p} + \| s - h \|_{nj}^2) \quad \textit{for} \quad 1 \le j \le J, \quad h \in \mathscr{H} \quad \textit{and} \quad s \in \mathscr{S}_n.$$

The next result follows from Lemmas 7–9.

LEMMA 10. *Suppose that Conditions* 1 *and* 3 *hold and that $N_n \sim n^\gamma$. Then*

$$\| \bar{f}_{nj}^{(m)} - (f_j^*)^{(m)} \|_j^2 = O_{\mathrm{pr}}(n^{-2r_m}) \quad \textit{for} \quad 1 \le j \le J.$$

PROOF OF THEOREM 1. Suppose that Conditions 1–3 hold and that $N_n \sim n^\gamma$. Now

$$\hat{f}_{nj}^{(m)} - (f_j^*)^{(m)} = \hat{f}_{nj}^{(m)} - \bar{f}_{nj}^{(m)} + \bar{f}_{nj}^{(m)} - (f_j^*)^{(m)}. \tag{14}$$

It follows from Lemma 4 that

$$E(\| \hat{f}_{nj} - \bar{f}_{nj} \|_j^2 \,|\, X_1, \cdots, X_n) = O_{\mathrm{pr}}(n^{-2r}).$$

Thus, by Lemma 8,

$$E(\| \hat{f}_{nj}^{(m)} - \bar{f}_{nj}^{(m)} \|_j^2 \,|\, X_1, \cdots, X_n) = O_{\mathrm{pr}}(n^{-2r_m}). \tag{15}$$

The desired conclusion follows from (14), (15) and Lemma 10.

6. Proofs of Lemmas 3, 8 and 9.

The next result follows easily from a compactness argument and change of variables.

LEMMA 11. *If Q is a polynomial of degree k (or less), I is an interval of finite positive length τ, and $t_0 \in I$, then*

$$\left(\Sigma_0^k \frac{\tau^m}{m!} \,|\, Q^{(m)}(t_0) \,|\, \right)^2 \le c_k^{2\tau} \tau^{-1} \int_I Q^2(t)\, dt.$$

Here c_k is some positive constant depending only on k.

PROOF OF LEMMA 8. Let $t_0 \in I = I_{n\nu} \subseteq [0, 1]$, where I has length $\tau = N_n^{-1}$. Let Q be the Taylor polynomial of degree k about t_0 corresponding to h. Since $h \in \mathscr{H}$, $| h - Q | \le M\tau^p/k!$ on I and hence $\int_I (h - Q)^2 \le M^2 \tau^{2p+1}/(k!)^2$. Thus by Lemma 11

$$\begin{aligned}(s^{(m)}(t_0) - h^{(m)}(t_0))^2 &= (s^{(m)}(t_0) - Q^{(m)}(t_0))^2 \\ &\le M_2 \tau^{-(2m+1)} \left(\tau^{2p+1} + \int_I (s - h)^2 \right)\end{aligned}$$

for some $M_2 \in (0, \infty)$. Consequently

$$\int_I (s^{(m)} - h^{(m)})^2 \le M_2 \tau^{-2m}\left(\tau^{2p+1} + \int_I (s-h)^2\right)$$

and hence

$$\int_0^1 (s^{(m)} - h^{(m)})^2 \le M_2\left(\tau^{2(p-m)} + \tau^{-2m}\int_0^1 (s-h)^2\right).$$

Since each g_j is bounded away from 0 and ∞ by Condition 1, the desired result holds.

Let I denote an interval of finite positive length τ and let (T, U) denote a pair of I-valued random variables each having an absolutely continuous distribution. Let E' be the expectation operator corresponding to the empirical distribution based on a random sample of size n_0 from the distribution of (T, U). Similarly, let Cov$'$ denote the covariance operator corresponding to this empirical distribution.

LEMMA 12. *Suppose that the marginal densities of T and U are each bounded below by β/τ on I, where $\beta > 0$. Let $t > 0$. Then, except on an event having probability at most $d_k\exp(-2n_0t^2)$, the following inequalities hold simultaneously for all polynomials Q, R of degree k:*

$$|E'Q(T) - EQ(T)| \le t\beta^{-1/2}c_k\mathrm{SD}(Q(T));$$

$$|E'Q^2(T) - EQ^2(T)| \le t\beta^{-1}c_k^2\mathrm{Var}(Q(T));$$

$$|E'R(U) - ER(U)| \le t\beta^{-1/2}c_k\mathrm{SD}(R(U));$$

and

$$|\mathrm{Cov}'(Q(T), R(U)) - \mathrm{Cov}(Q(T), R(U))| \le t(t+3)\beta^{-1}c_k^2\mathrm{SD}(Q(T))\mathrm{SD}(R(U)).$$

Here d_k is some positive constant depending only on k.

PROOF. Let t_0 denote the left endpoint of I. It follows from Hoeffding's inequality that, for $m \ge 1$,

$$\Pr\left(\left| E'\left[\left(\frac{T-t_0}{\tau}\right)^m\right] - E\left[\left(\frac{T-t_0}{\tau}\right)^m\right]\right| > t\right) \le \exp(-2n_0t^2).$$

Thus by Lemma 11, except on an event having probability at most $k\exp(-2n_0t^2)$,

$$\begin{aligned}|E'Q(T) - EQ(T)| &= \left|\textstyle\sum_0^k \frac{Q^{(m)}(t_0)}{m!}\tau^m\left(E'\left[\left(\frac{T-t_0}{\tau}\right)^m\right] - E\left[\left(\frac{T-t_0}{\tau}\right)^m\right]\right)\right| \\ &\le t\textstyle\sum_0^k \frac{\tau^m}{m!}|Q^{(m)}(t_0)| \le tc_k\left(\tau^{-1}\int_I Q^2(t)\,dt\right)^{1/2} \\ &\le t\beta^{-1/2}(EQ^2(T))^{1/2} = t\beta^{-1/2}SD(Q(T))\end{aligned}$$

for all polynomials Q of degree k such that $EQ(T) = 0$; and hence the first inequality of the lemma holds for all polynomials Q of degree k. The proofs of the remaining inequalities are similar.

PROOF OF LEMMA 9. Let T_i denote the jth coordinate of X_i. Observe that $I_\nu = I_{n\nu}$ has length $\tau = N_n^{-1}$. Set $\mathcal{J}_\nu = \{i\colon 1 \le i \le n$ and $T_i \in I_\nu\}$ and let $|\mathcal{J}_\nu|$ denote the number of elements in $\mathcal{J}_\nu$. It follows from (5), (13) and Lemma 2 that, except on an event whose probability tends to zero with n, for n sufficiently large

$$|\mathcal{J}_\nu| \ge b\tau n/2 \ge n^a \quad \text{for all} \quad \nu;$$

here a is some positive constant. Choose $t_\nu \in I_\nu$ and let Q_ν be the Taylor polynomial of degree k about t_ν corresponding to h. Then $|h - Q_\nu| \le M\tau^p/k!$ on I_ν. It follows from Lemma 12 (the second inequality) that, except on an event whose probability tends to zero with n, for n sufficiently large and for all ν

$$\begin{aligned}\frac{1}{2\int_{I_\nu} g_j}\int_{I_\nu}(s - Q_\nu)^2 g_j &\le \frac{1}{|\mathcal{J}_\nu|}\sum\nolimits_{\mathcal{J}_\nu}(s(T_i) - Q_\nu(T_i))^2\\ &\le 2\left(\left(\frac{M\tau^p}{k!}\right)^2 + \frac{1}{|\mathcal{J}_\nu|}\sum\nolimits_{\mathcal{J}_\nu}(s(T_i) - h(T_i))^2\right)\end{aligned}$$

and hence

$$\frac{1}{2\int_{I_\nu} g_j}\int_{I_\nu}(s - h)^2 g_j \le 6\left(\left(\frac{M\tau^p}{k!}\right)^2 + \frac{1}{|\mathcal{J}_\nu|}\sum\nolimits_{\mathcal{J}_\nu}(s(T_i) - h(T_i))^2\right).$$

Consequently by (5),

$$\int_{I_\nu}(s - h)^2 g_j \le 12B\tau\left(\left(\frac{M\tau^p}{k!}\right)^2 + \frac{1}{|\mathcal{J}_\nu|}\sum\nolimits_{\mathcal{J}_\nu}(s(T_i) - h(T_i))^2\right).$$

The conclusion to Lemma 9 now follows easily by summing on ν.

Let (Ω, P) be a probability space. Let Λ be a positive integer and let $A_1, \cdots, A_\Lambda$ be a finite partition of Ω which is nondegenerate in the sense that $p_\lambda = P(A_\lambda) > 0$ for $1 \le \lambda \le \Lambda$. Let V_1, V_2 be a pair of random variables having finite second moment. Set $\mu_j = EV_j$, $\sigma_{j\ell} = \operatorname{Cov}(V_j, V_\ell)$ and $\sigma_j^2 = \sigma_{jj} = \operatorname{Var}(V_j)$. Also set $\mu_{j|\lambda} = E(V_j \mid A_\lambda)$, $\sigma_{j\ell|\lambda} = \operatorname{Cov}(V_j, V_\ell \mid A_\lambda) = E((V_j - \mu_{j|\lambda})(V_\ell - \mu_{\ell|\lambda}) \mid A_\lambda)$ and $\sigma_{j|\lambda} = \sqrt{\sigma_{jj|\lambda}}$. Let (Ω', P') be a second probability space having a nondegenerate partition $A_1', \cdots, A_\Lambda'$ and a pair V_1', V_2' of random variables each having finite second moment. Define p_λ', μ_j', etc. as before.

LEMMA 13. *Given $\varepsilon > 0$, there is a $\delta > 0$ such that the following statement is valid: If $|p_\lambda' - p_\lambda| \le \delta p_\lambda$, $|\mu_{j|\lambda}' - \mu_{j|\lambda}| \le \delta\sigma_{j|\lambda}$ and $|\sigma_{j\ell|\lambda}' - \sigma_{j\ell|\lambda}| \le \delta\sigma_{j|\lambda}\sigma_{\ell|\lambda}$ for $1 \le j, \ell \le 2$ and $1 \le \lambda \le \Lambda$, then*

$$|\sigma_{j\ell}' - \sigma_{j\ell}| \le \varepsilon\sigma_j\sigma_\ell \quad \text{for} \quad 1 \le j, \ell \le 2.$$

PROOF. Observe that

$$\mu_j = \sum p_\lambda \mu_{j|\lambda}, \; \sigma_j^2 = \sum p_\lambda(\sigma_{j|\lambda}^2 + (\mu_{j|\lambda} - \mu_j)^2)$$

and

$$\sigma_{j\ell} = \sum p_\lambda(\sigma_{j\ell|\lambda} + (\mu_{j|\lambda} - \mu_j)(\mu_{\ell|\lambda} - \mu_\ell))$$

and that similar formulas hold for the second probability space. The desired conclusion now follows in a straightforward manner (Schwarz's inequality is used several times).

PROOF OF LEMMA 3. Choose ε with $0 < \varepsilon < ((1 - \delta)/2)^{J-1}$. It follows easily from Lemmas 2, 12 and 13 that, except on an event whose probability tends to zero with n, the following inequality holds for all choices of $s_1, \cdots, s_J \in \mathscr{S}_n$ and the corresponding choices of $V_j = V_j(s_j)$:

$$|\operatorname{Cov}_n(V_j, V_\ell) - \operatorname{Cov}(V_j, V_\ell)| \le \varepsilon\, \mathrm{SD}(V_j)\mathrm{SD}(V_\ell) \quad \text{for} \quad 1 \le j, \quad \ell \le J;$$

in particular,

$$\mathrm{SD}_n(V_j) \le (1 + \varepsilon)^{1/2}\mathrm{SD}(V_j) \quad \text{for} \quad 1 \le j \le J$$

and hence by Lemma 1

$$\begin{aligned}\operatorname{Var}_n(V_1 + \cdots + V_J) &\ge \operatorname{Var}(V_1 + \cdots + V_J) - \varepsilon\, (\textstyle\sum_1^J \mathrm{SD}(V_j))^2\\ &\ge (((1 - \delta)/2))^{J-1} - \varepsilon)(\textstyle\sum_1^J \mathrm{SD}(V_j))^2\\ &\ge (((1 - \delta)/2)^{J-1} - \varepsilon)/(1 + \varepsilon))(\textstyle\sum_1^J \mathrm{SD}_n(V_j))^2.\end{aligned}$$

Since ε can be made arbitrarily small, the desired result holds.

REFERENCES

ANDREWS, F. M., MORGAN, J. N. and SONQUIST, J. A. (1967). *Multiple Classification Analysis.* Institute for Social Research, Univ. of Michigan, Ann Arbor.

BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis.* MIT Press, Cambridge, Mass.

DE BOOR, C. (1968). On uniform approximation by splines. *J. Approx. Theory* **1** 219–235.

DE BOOR, C. (1978). *A Practical Guide to Splines.* Springer, New York.

BOX, G. E. P., HUNTER, W. G. and HUNTER, J. S. (1978). *Statistics for Experimenters.* Wiley, New York.

BREIMAN, L. and FRIEDMAN, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** (to appear).

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., and STONE, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont.

BREIMAN, L. and STONE, C. J. (1978). Nonlinear additive regression, note.

CHUNG, K. L. (1974). *A Course in Probability Theory,* 2nd ed. Academic, New York.

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

DE LEEUW, J., YOUNG, F. W., and TAKANE, Y. (1976). Additive structure in qualitative data: an alternating least squares method with optimal scaling features. *Psychometrika* **41** 471–503.

EHRENFELD, S. (1956). Complete class theorems in experimental design. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **I**, 57–67. Univ. of California Press, Berkeley.

ENGLE, R. F., GRANGER, C. W. J., RICE, J., and WEISS A. (1982). Nonparametric estimates of the relation between weather and electricity demand. Discussion paper 83-17, Dept. of Economics, Univ. of California, San Diego.

FIELDING, A. (1977). Binary segmentation: The automatic interaction detector and related techniques for exploring data structure. In *Exploring Data Structures.* (C. A. O'Muircheartaigh and C. Payne, ed.) 221–257. Wiley, Chichester.

FISHER, L. and MCDONALD, J. (1978). *Fixed Effects Analysis of Variance.* Academic, New York.

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

FRIEDMAN, J. H., GROSSE, E., and STEUTZLE, W. (1983). Multidimensional additive spline approximation. *SIAM J. Sci. Statist. Comput.* **4** 291–301.

FRIEDMAN, J. H., STUETZLE, W., and SCHROEDER, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79** 599–608.

HABERMAN, S. J. (1978). *Analysis of Qualitative Data,* 2 volumes. Academic, New York.

HASTIE, T. J. (1983). Non-parametric logistic regression. Technical report, Dept. of Statistics, Stanford University.

HASTIE, T. J. (1984). Comment (on pages 77–78) to Graphical methods for assessing logistic regression models, by J. M. Landwehr, D. Pregibon, and A. C. Shoemaker. *J. Amer. Statist. Assoc.* **79** 61–83.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random varaibles. *J. Amer. Statist. Assoc.* **58** 13–30.

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

JOHN, P. W. M. (1971). *Statistical Design and Analysis of Experiments.* Macmillan, New York.

LAWTON, W. H., SYLVESTRE, E. A., and MAGGION, M. S. (1972). Self modelling nonlinear regression. *Technometrics* **14** 513–532.

MORGAN, J. N. and SONQUIST, J. A. (1963). Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc.* **58** 415–435.

OAKES, P. (1981). Survival times: aspects of partial likelihood. *Internat. Statist. Rev.* **49** 235–264.

POWELL, M. J. D. (1981). *Approximation Theory and Methods.* Cambridge University Press, Cambridge, Mass.

SCHEFFÉ, H. (1959). *The Analysis of Variance.* Wiley, New York.

SONQUIST, J. A. (1970). *Multivariate Model Building.* Institute for Social Research, Univ. of Michigan, Ann Arbor.

SONQUIST, J. A., BAKER, E. L., and MORGAN, J. N. (1971). *Searching for Structure.* Institute for Social Research, Univ. of Michigan, Ann Arbor.

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

STÜTZLE, W., GASSER, T., MOLINARI, L., LARGO, R. H., PRADER, A., and HUBER, P. J. (1980). Self-invariant modelling of human growth. *Ann. Human Biology* **7** 507–528.

TIBSHIRANI, R. (1983). Non-parametric estimation of relative risk. Technical report, Dept. of Statistics, Stanford University.

TUKEY, J. (1949). One degree of freedom for non-additivity. *Biometrics* **5** 232–243.

TUKEY, J. (1961). Curves as parameters, and touch estimation. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **I**, 681–694. Univ. of California Press, Berkeley.

WAHBA, G. (1984). Cross validated spline methods for the estimation of multivariate functions from data on functionals. In *Statistics: An Appraisal, Proc. 50th Anniversary Conference Iowa State Statistical Laboratory* (H. A. David and H. T. David, eds.) 205–235, Iowa State University Press, Ames, Iowa.

WEGMAN, E. J. and WRIGHT, I. W. (1983). Splines in Statistics. *J. Amer. Statist. Assoc.* **78** 351–365.

WINSBERG, S. and RAMSAY, J. O. (1980). Monotonic transformations to additivity using splines. *Biometrika* **67** 669–674.

YOUNG, F. W., DE LEEUW, J. and TAKANE, Y. (1976). Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. *Psychometrika* **41** 505–529.

ZELDIN, M. D. and THOMAS, D. M. (1975). Ozone trends in the Eastern Los Angeles basin corrected for meteorological variations. *Proc. Internat. Conf. Environmental Sensing and Assessment,* **2,** held September 14–19, 1975, in Las Vegas, Nevada.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720