Journal of Multivariate Analysis **67**, 49–71 (1998)
Article No. MV981753

Functional ANOVA Models for Generalized Regression*

Jianhua Z. Huang

University of Pennsylvania

Received April 17, 1996; revised July 29, 1997

The Functional ANOVA model is considered in the context of generalized regression, which includes logistic regression, probit regression, and Poisson regression as special cases. The multivariate predictor function is modeled as a specified sum of a constant term, main effects, and selected interaction terms. Maximum likelihood estimate is used, where the maximization is taken over a suitably chosen approximating space. The approximating space is constructed from virtually arbitrary linear spaces of functions and their tensor products and is compatible with the assumed ANOVA structure on the predictor function. Under mild conditions, the maximum likelihood estimate is consistent and the components of the estimate in an appropriately defined ANOVA decomposition are consistent in estimating the corresponding components of the predictor function. When the predictor function does not satisfy the assumed ANOVA form, the estimate converges to its best approximation of that form relative to the expected log-likelihood. A rate of convergence result is obtained, which reinforces the intuition that low-order ANOVA modeling can achieve dimension reduction and thus overcome the curse of dimensionality. © 1998 Academic Press

AMS 1991 subject classifications: primary 62G05, secondary 62G20.

Key words and phrases: Exponential family; interaction; maximum likelihood estimate; rate of convergence; splines; tensor product.

1. INTRODUCTION

Functional ANOVA modeling provides a useful tool for a variety of multivariate function estimation problems. While it is more flexible than the classical linear and additive modeling, it retains the advantage of good interpretability. In a functional ANOVA model, the (multivariate) function of primary interest is modeled as a specified sum of a constant term, main effects (functions of one variable), and selected interaction terms (functions of two or more variables). When only low-order interaction terms are included in the model, the curse of dimensionality can be overcome. Due to its various advantages, functional ANOVA modeling has gained popularity recently and related literature has been growing steadily; see Stone

* This work was supported in part by the NSF under Grant DMS-9504463.

0047-259X/98 $25.00

et al. (1997) for a comprehensive review. This paper studies the convergence property of the maximum likelihood estimate for a functional ANOVA model in the context of generalized regression, which includes logistic regression, probit regression, and Poisson regression as special cases.

We set up the generalized regression framework following Stone (1994). Consider a pair (X, Y) of random variables, where Y is real-valued and X is possibly real vector-valued; here Y is referred to as a *response* or *dependent variable* and X as the vector of *covariates* or *predictor variables*. The conditional distribution of Y given that $X=x$ is assumed to have the form

$$P(Y \in dy, x; \eta) = \exp\{B(\eta(x))\, y - C(\eta(x))\}\, \rho(dy), \tag{1}$$

where $B(\cdot)$ and $C(\cdot)$ are known functions satisfying some restrictions that will be described in Section 2. The function $\eta = \eta(\cdot)$ specifies how the response depends on the covariates; we refer it as a *predictor function.* Our interest lies in estimating η based on a random sample of size n from the distribution of (X, Y). When $\eta(x) = x^T\beta$, we get a generalized linear model (see McCullagh and Nelder, 1989). If $\eta(\cdot)$ has the form of a specified sum of a constant term, main effects, and selected interaction terms, then we get a functional ANOVA model. See Stone (1994) and Huang (1998) for illustrations of functional ANOVA models.

As discussed in Huang (1998), three fundamental questions regarding the properties of an estimate $\hat{\eta}$ in a functional ANOVA model are the following:

(i) Does $\hat{\eta}$ converge to η when the sample size tends to infinity? If so, what is the rate of convergence?

(ii) How do we define appropriate ANOVA decompositions of η and $\hat{\eta}$, so that the ANOVA components of $\hat{\eta}$ converge to the corresponding components of η?

(iii) How does $\hat{\eta}$ behave when the model is misspecified, that is, when η does not have the assumed ANOVA form?

The main purpose of this paper is to give quite thorough answers to these questions in the context of generalized regression. We employ the maximum (quasi-) likelihood estimate, where maximization is carried out over a suitably chosen approximating space. The approximating space is constructed from virtually arbitrary linear spaces of functions and their tensor products. The linear spaces that serve as building blocks can be any of those commonly used in practice: polynomials, trigonometric polynomials, splines, wavelets, and finite elements.

We shall see that, under mild conditions, the maximum likelihood estimate is consistent, provided that the approximating space is compatible with the assumed ANOVA structure on the predictor function. Moreover, the components of the estimate in an appropriately defined ANOVA decomposition are consistent in estimating the corresponding components of the predictor function. When the predictor function does not satisfy the assumed ANOVA form, the estimate converges to its best approximation of that form relative to the expected log-likelihood. A rate of convergence result is obtained, which reinforces the intuition that low-order ANOVA modeling can achieve dimension reduction and thus overcome the curse of dimensionality.

The results in this paper are similar to those for regression established in Huang (1998). Here, however, the maximum likelihood estimate cannot be viewed simply as an orthogonal projection, due to the nonlinear structure of the problem. A deeper study of the properties of the log-likelihood function is needed to overcome the difficulties. We shall see that the concavity of the log-likelihood and expected log-likelihood functions play a crucial role in our analysis.

The convergence properties of functional ANOVA model for generalized regression have been studied by Stone (1986, 1994) and Hansen (1994) when the approximating spaces are built with polynomial splines and their tensor products. The result in this paper provides a clearer picture of the mathematical structure of the maximum likelihood estimate in fitting a functional ANOVA model. By removing the dependence on splines in the theory developed by Stone and by Hansen, we are able to discern what is essential in getting a consistent estimate in a functional ANOVA model and in getting a consistent estimate of the ANOVA components of the function of interest.

The deep understanding of the structure of the problem enables us to give arguments significantly simpler than those in the previous works by Stone and by Hansen, even though the results here are much more general. Moreover, while a strong assumption on the boundedness of conditional moment generating functions is needed in the previous works, it is relaxed here by only assuming the boundedness of conditional second moments. The result under such a weaker assumption is useful in pseudo-likelihood estimation.

The paper is organized as follows. In Section 2, we state our main results. We describe the model assumptions in Section 2.1; a general theorem on rates of convergence is given in Section 2.2; Section 2.3 studies the functional ANOVA models. We provide some useful preliminary results in Section 3. The proofs of the theorems are deferred to Sections 4 and 5.

In what follows, for any function f on $\mathscr{X}$, set $\|f\|_\infty = \sup_{x \in \mathscr{X}} |f(x)|$. Given positive numbers a_n and b_n for $n \geqslant 1$, let $a_n \asymp b_n$ mean that a_n/b_n is bounded away from zero and infinity. Given random variables W_n for $n \geqslant 1$, let $W_n = O_P(b_n)$ mean that $\lim_{c \to \infty} \limsup_n P(|W_n| \geqslant cb_n) = 0$. Let $\operatorname{range}(h)$ denote the range of a real-valued function h.

2. STATEMENT OF RESULTS

2.1. *Model Assumptions*

Consider an exponential family of distributions on $\mathbb{R}$ of the form $e^{B(\eta)\,y - C(\eta)}\rho(dy)$, where the parameter η ranges over an open subinterval $\mathscr{I}$ of $\mathbb{R}$. Here ρ is a nonzero measure on $\mathbb{R}$ that is not concentrated on a single point and $\int_{\mathbb{R}} e^{B(\eta)\,y - C(\eta)}\rho(dy) = 1$ for $\eta \in \mathscr{I}$. Note that $C(\eta) = \log \int_{\mathbb{R}} e^{B(\eta)\,y}\rho(dy)$.

Assumption 1. $B(\cdot)$ is twice continuously differentiable and its first derivative $B'(\cdot)$ is strictly positive on $\mathscr{I}$.

Under Assumption 1, $B(\cdot)$ is strictly increasing and $C(\cdot)$ is twice continuously differentiable on $\mathscr{I}$. The mean of the distribution is given by $\mu = A(\eta) = C'(\eta)/B'(\eta)$ for $\eta \in \mathscr{I}$. The function $A(\cdot)$ is continuously differentiable and $A'(\eta)$ is strictly positive on $\mathscr{I}$, so $A(\cdot)$ is strictly increasing on $\mathscr{I}$.

Assumption 2. There is a subinterval S of $\mathbb{R}$ such that ρ is concentrated on S and

$$B''(\xi)\,y - C''(\xi) < 0, \qquad \xi \in \mathscr{I}, \tag{2}$$

for all $y \in \mathring{S}$, where $\mathring{S}$ denotes the interior of S. If S is bounded, (2) holds for at least one of its endpoints.

Note that $A(\eta) \in \mathring{S}$ for $\eta \in \mathscr{I}$. Thus by Assumption 2,

$$B''(\xi)\,A(\eta) - C''(\xi) < 0, \qquad \xi \in \mathscr{I}. \tag{3}$$

The interval S needs to be specified according to the context. It need not be the support of ρ nor the largest set of y such that (2) holds. For example, consider identity link for Poisson regression. We have that $B(\eta) = \log \eta$, $C(\eta) = \eta$, $\mathscr{I} = (0, \infty)$, and $S = [0, \infty)$. Then the support of ρ is $\{0, 1, ...\}$ and the largest set of y such that (2) holds is $(0, \infty)$.

Assumptions 1 and 2 are satisfied by many familiar exponential families, including Normal, Binomial-probit, Binomial-logit, Poisson, gamma, geometric and negative binomial distributions; see Stone (1986). Our setup is more general than that used in Stone (1986). By relaxing the restriction that $\mathscr{I} = \mathbb{R}$, the identity link is allowed for Poisson regression and Binomial regression.

Let X represent the predictor variable and Y the real-valued response variable. We assume that X ranges over a compact subset $\mathscr{X}$ of some Euclidean space and has a positive density. The conditional distribution of

Y given X is connected to the exponential family distribution through the following assumption.

Assumption 3. $P(Y \in S) = 1$ and $E(Y \mid X = x) = A(\eta(x))$ for $x \in \mathscr{X}$.

If the conditional distribution of Y given $X = x$ has the exponential family distribution described at the beginning of this section with parameter $\eta = \eta(x)$, then Assumption 3 is satisfied and the log-likelihood is given by

$$l(h; X, Y) = B(h(X))\, Y - C(h(X)) \tag{4}$$

for any function h on $\mathscr{X}$ that takes values in $\mathscr{I}$. In general, the conditional distribution of Y given $X = x$ does not necessarily belong to the exponential family. We can think of $l(h; X, Y)$ as a pseudo-log-likelihood. For simplicity, we shall still refer to it as a log-likelihood. When it exists, $\Lambda(h) = E[\, l(h; X, Y)]$ is referred to as the expected log-likelihood.

Assumption 4. There is a compact subinterval $\mathscr{K}_0$ of $\mathscr{I}$ such that range$(\eta) \subset \mathscr{K}_0$.

Under Assumption 4, $A(\eta(\cdot))$ ranges over a compact subinterval of $\mathring{S}$. If $\mathscr{I} = \mathbb{R}$, Assumption 4 is equivalent to assuming that η is bounded.

Assumption 5. There is a constant $D > 0$ such that $\sup_{x \in \mathscr{X}} \operatorname{var}(Y \mid X = x) \leqslant D$.

Let $(X_1, Y_1), ..., (X_n, Y_n)$ be an i.i.d. sample of size n from the joint distribution of X and Y. Our goal is to estimate $\eta(\cdot)$. We assume hereafter that Assumptions 1–5 hold. Since the functions $A(\cdot)$, $B(\cdot)$, and $C(\cdot)$ are continuous, $\Lambda(h)$ is well-defined for each h taking values in a compact subinterval of $\mathscr{I}$. Moreover, for each $\eta \in \mathscr{I}$, the function $B(\xi)\, A(\eta) - C(\xi)$, $\xi \in \mathscr{I}$, has a unique maximum at $\xi = \eta$. Thus, the function that maximizes $\Lambda(\cdot)$ is given by the true predictor function η.

2.2. *A General Result.*

For any function f defined on $\mathscr{X}$, set $E_n(f) = 1/n \sum_{i=1}^{n} f(X_i)$ and $E(f) = E[\, f(X)]$. For any two functions f_1 and f_2 on $\mathscr{X}$, define the empirical inner product and norm as $\langle f_1, f_2 \rangle_n = E_n(f_1, f_2)$ and $\| f_1 \|_n^2 = E_n(f_1^2)$. The theoretical versions of these quantities are given by $\langle f_1, f_2 \rangle = E(f_1 f_2)$ and $\| f_1 \|^2 = E(f_1^2)$.

Let H be a linear space of real-valued functions on $\mathscr{X}$. We model η as a member of H and refer to H as the model space. However, the (expected) log-likelihood function need not be defined for all functions in H; hence we need restrict our attention to a subset of H. Set

$$H^* = \{h \in H \text{: range}(h) \subset \mathscr{K} \text{ for a compact subinterval } \mathscr{K} \subset \mathscr{I}\}.$$

Then the log-likelihood function is well-defined on H^*. When $\mathscr{I} = \mathbb{R}$, H^* is just the collection of all bounded functions in H. The model assumptions in the previous subsection imply that the expected log-likelihood $\Lambda(\cdot)$ is strictly concave over functions in H^*. That is, given any two essentially different functions $h_0, h_1 \in H^*$ we have that

$$\Lambda(h_0 + t(h_1 - h_0)) > (1-t)\, \Lambda(h_0) + t\Lambda(h_1), \qquad t \in (0, 1). \tag{5}$$

Here, h_0 and h_1 are said to be essentially different if their difference is nonzero on a set of positive probability relative to the distribution of X.

Condition 1. There exists a function $\eta^* \in H^*$ such that $\Lambda(\eta^*) = \max_{h \in H^*} \Lambda(h)$.

Remark. Condition 1 says that the best approximation of η in H^* relative to the expected log-likelihood exists. We need this condition to handle the case in which the model is misspecified. Since $\Lambda(\cdot)$ is strictly concave on H^*, η^* is essentially uniquely determined if it exists. If the model is correctly specified, that is, $\eta \in H^*$, then $\eta^* = \eta$ exists and this condition is nil. (Here, we identify two functions that are not essentially different.) When $\mathscr{I} = \mathbb{R}$, Lemma 4.1 of Stone (1994) can be used to check Condition 1.

Let $G \subset H$ be a finite-dimensional linear space of bounded functions. The space G may vary with sample size n, but for notational convenience, we suppress the possible dependence on n. We require that the dimension N_n of G be positive for $n \geqslant 1$. We also require that the space G be *theoretically identifiable* in that if $g \in G$ equals zero almost everywhere relative to the measure induced by the distribution of X, then it identically equals zero. Since we hope to choose G such that the functions in H can be well approximated by functions in G, we refer to G as the approximating space. The space G is said to be *empirically identifiable* (relative to $X_1, ..., X_n$) if the only function g in the space such that $g(X_i) = 0$ for $1 \leqslant i \leqslant n$ is the function that identically equals zero. Given a sample $X_1, ..., X_n$, if G is empirically identifiable, then it is a Hilbert space equipped with the empirical inner product.

Set

$$G^* = \{g \in G\colon \text{range}(g) \subset \mathscr{K} \text{ for a compact subinterval } \mathscr{K} \subset \mathscr{I}\}.$$

Given a function $g \in G^*$, let

$$\ell(g) = \frac{1}{n} \sum_{i=1}^{n} [B(g(X_i))\, Y_i - C(g(X_i))]$$

denote the (scaled) log-likelihood function corresponding to the random sample of size n. Then it follows from (2) that $\ell(g)$ is concave on G^*. If

$\hat{\eta} \in G^*$ and $\ell(\hat{\eta}) = \max_{g \in G^*} \ell(g)$, then $\hat{\eta}$ is referred to as a maximum likelihood estimate. As we shall see, under some conditions, $\hat{\eta}$ exists except on an event whose probability tends to zero as $n \to \infty$ (Lemma 4.4).

Let $\bar{\eta} = \arg\max_{g \in G^*} \Lambda(g)$ denote the best approximation in G^* to η. By the strict concavity of $\Lambda(\cdot)$, $\bar{\eta}$ is uniquely defined if it exists. In fact, $\bar{\eta}$ exists for n sufficiently large (Lemma 4.2). We have the decomposition $\hat{\eta} - \eta^* = (\hat{\eta} - \bar{\eta}) + (\bar{\eta} - \eta^*)$. The term $\hat{\eta} - \bar{\eta}$ is referred to as the estimation error and $\bar{\eta} - \eta^*$ as the approximation error.

Set $A_n = \sup_{g \in G} \{ \|g\|_\infty / \|g\| \}$. The constant $A_n \geqslant 1$ is a measure of irregularity of the approximating space G. Since functions in G are bounded and G is theoretically identifiable, A_n is finite for any n.

Set $\rho_n = \inf_{g \in G} \|g - \eta^*\|_\infty$. Under Condition 1, η^* is bounded and thus ρ_n is finite. By a compactness argument, there is a $g^* \in G$ such that $\|g^* - \eta^*\| = \rho_n$. The constant ρ_n characterizes the target function η^* and reflects the approximation property of the space G. For a specific choice of approximating space, a condition on the rate of decay of ρ_n gives a smoothness assumption on η^*. On the other hand, given that the target function falls in a specific function class, the constant ρ_n is a measure of the approximation power of the approximating space in supreme norm.

THEOREM 1. *Suppose Condition* 1 *holds and that* $\lim_n A_n^2 N_n / n = 0$ *and* $\lim_n A_n \rho_n = 0$. *Then*

$$\|\hat{\eta} - \bar{\eta}\|^2 = O_P(N_n/n), \qquad \|\hat{\eta} - \bar{\eta}\|_n^2 = O_P(N_n/n);$$

$$\|\bar{\eta} - \eta^*\|^2 = O_P(\rho_n^2), \qquad \|\bar{\eta} - \eta^*\|_n^2 = O_P(\rho_n^2).$$

Consequently,

$$\|\hat{\eta} - \eta^*\|^2 = O_P(N_n/n + \rho_n^2) \qquad \textit{and} \qquad \|\hat{\eta} - \eta^*\|_n^2 = O_P(N_n/n + \rho_n^2).$$

Remarks. 1. The condition that $\lim_n A_n \rho_n = 0$ is required in the proof of Lemma 4.2. If $\mathscr{I} = \mathbb{R}$, then this condition can be weakened to $\limsup_n A_n \rho_n < \infty$.

2. Theorem 1 gives a unified treatment of the rate of convergence for the maximum likelihood estimate in a finite-dimensional linear space. We need only find upper bounds of A_n and ρ_n. The results in approximation theory literature can be used to find these upper bounds for various commonly used approximating spaces, including polynomials, trigonometric polynomials, splines, wavelets, and finite elements; see Huang (1998).

3. Note that we do not require that the dimension of G go to infinity with the sample size. When H is a finite-dimensional linear space of bounded functions, we can choose $G = H$, which does not depend on the sample

size. Then A_n is finite and independent of n and $\rho_n = 0$. If $\eta \in H^*$, then the integrated squared error of $\hat{\eta}$ to η converges to zero at the parametric rate $1/n$.

2.3. *Functional ANOVA Models*

In this section, we introduce the ANOVA model for functions and establish the rates of convergence for the maximum likelihood estimate and its components. Our terminology and notation follow closely those in Huang (1998); see also Stone (1994). The next condition is needed to prevent confounding in the ANOVA decomposition (see Lemma 3.2).

Condition 2. The distribution of X is absolutely continuous and its density function $f_X(\cdot)$ is bounded away from zero and infinity on $\mathscr{X}$.

Model space. Suppose $\mathscr{X}$ is the Cartesian product of some compact sets $\mathscr{X}_1, ..., \mathscr{X}_L$, where $\mathscr{X}_l \subset \mathbb{R}^{d_l}$ with $d_l \geqslant 1$. Let $\mathscr{S}$ be a fixed hierarchical collection of subsets of $\{1, ..., L\}$, where hierarchical means that if $s \in \mathscr{S}$ and $r \subset s$, then $r \in \mathscr{S}$. Let $H_\varnothing$ denote the space of constant functions on $\mathscr{X}$. Given a nonempty subset $s \in \mathscr{S}$, let H_s denote the space of square-integrable functions on $\mathscr{X}$ that depend only on the variables x_l, $l \in s$. Set

$$H = \left\{ \sum_{s \in \mathscr{S}} h_s : h_s \in H_s \right\}.$$

Let H_s^0 denote the space of all functions in H_s that are theoretically orthogonal to each function in H_r for every proper subset r of s. Under Condition 4, every function $h \in H$ can be written in an essentially unique manner as $\sum_{s \in \mathscr{S}} h_s$, where $h_s \in H_s^0$ for $s \in \mathscr{S}$ (see Lemma 3.2). We refer to $\sum_{s \in \mathscr{S}} h_s$ as the *theoretical ANOVA decomposition* of h, and we refer to $h_s \in H_s^0$, $s \in \mathscr{S}$, as the components of h. The component $h_s \in H_s^0$ is referred to as the constant component if $\#(s) = 0$, as a main effect component if $\#(s) = 1$, and as an interaction component if $\#(s) \geqslant 2$; here $\#(s)$ is the number of elements of s.

Since each function in the model space H has a unique ANOVA decomposition, we refer to the model specified by H as a *functional ANOVA model*. In particular, $\mathscr{S}$ specifies the main effects and interaction terms that are in the model. As special cases, if $\max_{s \in \mathscr{S}} \#(s) = L$, then all interaction terms are included and we get a saturated model; if $\max_{s \in \mathscr{S}} \#(s) = 1$, we get an additive model. As in Section 2.2, let H^* consist of those functions in H whose range is contained in a compact subinterval of $\mathscr{I}$. We need to restrict our attention to H^* where the log-likelihood is well-defined.

Approximating space. Let $G_\varnothing$ denote the space of constant functions on $\mathscr{X}$, which has dimension $N_\varnothing = 1$. Given $1 \leqslant l \leqslant L$, let $G_l \supset G_\varnothing$ denote a linear space of bounded, real-valued functions on $\mathscr{X}_l$ which varies with sample size and has finite, positive dimension N_l. Given a nonempty subset $s=\{s_1, ..., s_k\}$ of $\{1, ..., L\}$, let G_s be the tensor product of $G_{s_1}, ..., G_{s_k}$, which is the space of functions on $\mathscr{X}$ spanned by the functions of the form $\prod_{i=1}^k g_{s_i}(x_{s_i})$, where $g_{s_i} \in G_{s_i}$ for $1 \leqslant i \leqslant k$. Then the dimension of G_s is given by $N_s = \prod_{i=1}^k N_{s_i}$. Set

$$G = \left\{ \sum_{s \in \mathscr{S}} g_s \colon g_s \in G_s \right\}.$$

The dimension N_n of G satisfies $\max_{s \in \mathscr{S}} N_s \leqslant N_n \leqslant \sum_{s \in \mathscr{S}} N_s \leqslant \#(\mathscr{S}) \max_{s \in \mathscr{S}} N_s$. Let G_s^0 denote the space of all functions in G_s that are empirically orthogonal to each function in G_r for every proper subset r of s. If the space G is empirically identifiable, then each function $g \in G$ can be written uniquely in the form $\sum_{s \in \mathscr{S}} g_s$, where $g_s \in G_s^0$ for $s \in \mathscr{S}$ (see Lemma 3.3). If so, we refer to $\sum_{s \in \mathscr{S}} g_s$ as the *empirical ANOVA decomposition* of g, and we refer to $g_s \in G_s^0$, $s \in \mathscr{S}$, as the components of g. As in Section 2.2, let G^* consist of the functions in G whose range is contained in a compact subinterval of $\mathscr{I}$. Then the log-likelihood is well-defined on G^*.

We now define some constants that are analogs of the constants A_n and ρ_n in Section 2.2. These constants are more straightforward to determine than the constants A_n and ρ_n themselves. Set

$$A_s = A_{sn}(G_s) = \sup_{g \in G_s} \frac{\|g\|_\infty}{\|g\|}, \qquad s \in \mathscr{S}.$$

Recall that η^* is the best approximation in H^* to η and its ANOVA decomposition has the form $\eta^* = \sum_{s \in \mathscr{S}} \eta_s^*$, where $\eta_s^* \in H_s^0$ for $s \in \mathscr{S}$. Set

$$\rho_s = \rho_{sn}(\eta_s^*, G_s) = \inf_{g \in G_s} \|g - \eta_s^*\|_\infty, \qquad s \in \mathscr{S}.$$

THEOREM 2. *Suppose Conditions* 1 *and* 2 *hold and that* $\lim_n A_s \rho_{s'} = 0$ *and* $\lim_n A_s^2 N_{s'}/n = 0$ *for* $s, s' \in \mathscr{S}$. *Then the results of Theorem* 1 *hold with* N_n *and* ρ_n *replaced by* $\sum_{s \in \mathscr{S}} N_s$ *and* $\sum_{s \in \mathscr{S}} \rho_s$.

Proof. We have that $A_n \leqslant [\varepsilon_1^{1-\#(\mathscr{S})} \sum_{s \in \mathscr{S}} A_s^2]^{1/2}$, $N_n \leqslant \sum_{s \in \mathscr{S}} N_s$, and $\rho_n \leqslant \sum_{s \in \mathscr{S}} \rho_s$ (see Huang, 1998). The conditions of Theorem 1 now follow from the conditions of this theorem. ▌

The next result demonstrates that the components of the ANOVA decomposition of $\hat{\eta}$ provide consistent estimates of the corresponding components of η^*. Recall that $\bar{\eta}$ is the best approximation in G^* to η. The ANOVA decompositions of $\hat{\eta}$ and $\bar{\eta}$ are given by $\hat{\eta} = \sum_{s \in \mathscr{S}} \hat{\eta}_s$ and

$\bar{\eta} = \sum_{s \in \mathcal{S}} \bar{\eta}_s$, where $\hat{\eta}_s, \bar{\eta}_s \in G_s^0$ for $s \in \mathcal{S}$. We have an identity involving the ANOVA components: $\hat{\eta}_s - \eta_s^* = (\hat{\eta}_s - \bar{\eta}_s) + (\bar{\eta}_s - \eta_s^*)$.

THEOREM 3. *Suppose Conditions* 1 *and* 2 *hold and that* $\lim_n A_s \rho_{s'} = 0$ *and* $\lim_n A_s^2 N_{s'}/n = 0$ *for* $s, s' \in \mathcal{S}$. *Then, for each* $s \in \mathcal{S}$,

$$\|\hat{\eta}_s - \bar{\eta}_s\|^2 = O_P\left(\sum_{s \in \mathcal{S}} N_s/n\right), \qquad \|\hat{\eta}_s - \bar{\eta}_s\|_n^2 = O_P\left(\sum_{s \in \mathcal{S}} N_s/n\right);$$

$$\|\bar{\eta}_s - \eta_s^*\|^2 = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2\right),$$

$$\|\bar{\eta}_s - \eta_s^*\|_n^2 = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2\right).$$

Consequently,

$$\|\hat{\eta}_s - \eta_s^*\|^2 = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2\right)$$

and

$$\|\hat{\eta}_s - \eta_s^*\|_n = O_P\left(\sum_{s \in \mathcal{S}} \frac{N_s}{n} + \sum_{s \in \mathcal{S}} \rho_s^2\right).$$

EXAMPLE (Univariate Splines). Throughout this example, we assume that $\mathcal{X}$ is the Cartesian product of compact intervals $\mathcal{X}_1, ..., \mathcal{X}_L$. Without loss of generality, it is assumed that each of these intervals equals $[0, 1]$. Let m be a nonnegative integer and set $p = m + \beta$. Following Stone (1994), we say a function on $\mathcal{X}$ is *p-smooth* if it is m times continuously differentiable on $\mathcal{X}$ and D^α satisfies a Hölder condition with exponent β for all α with $[\alpha] = m$. Suppose that Conditions 1 and 2 hold. In addition, assume that η_s^* is p-smooth for each $s \in \mathcal{S}$. (This is a natural assumption when the model is correct, that is, when $\eta^* = \eta$. However, it is hard to check in general.) Set $d = \max_{s \in \mathcal{S}} \#(s)$.

Let J be a positive integer, and let $t_0, t_1, ..., t_J, t_{J+1}$ be real numbers with $0 = t_0 < t_1 < \cdots < t_J < t_{J+1} = 1$. Partition $[0, 1]$ into $J+1$ subintervals $I_j = [t_j, t_{j+1})$, $j = 0, ..., J-1$, and $I_J = [t_J, t_{J+1}]$. Let m be a nonnegative integer. A function on $[0, 1]$ is a spline of degree m with knots $t_1, ..., t_J$ if the following hold: (i) it is a polynomial of degree m or less on each interval I_j, $j = 0, ..., J$; and (ii) (for $m \geqslant 1$) it is $(m-1)$-times continuously differentiable on $[0, 1]$. Such spline functions constitute a linear space of dimension $K = J + m + 1$. For detailed discussions of univariate splines, see de Boor (1978) and Schumaker (1981).

Let G_l be the space of splines of degree m for $l = 1, ..., L$, where m is fixed. We allow J, $(t_j)_1^J$ and thus G_l to vary with the sample size. Suppose that

$$\frac{\max_{0 \leqslant j \leqslant J} (t_{j+1} - t_j)}{\min_{0 \leqslant j \leqslant J} (t_{j+1} - t_j)} \leqslant \gamma$$

for some positive constant γ. Then $A_s \asymp J^{\#(s)}$, $N_s \asymp J^{\#(s)}$, and $\rho_s \asymp J^{-p}$; see Huang (1998). Suppose $p > d/2$ and $J^{2d} = o(n)$. Then the condition in Theorems 2 and 3 are satisfied. Hence, $\|\hat{\eta}_s - \eta_s^*\|^2 = O_P(J^d/n + J^{-2p})$ for $s \in \mathscr{S}$ and $\|\hat{\eta} - \eta^*\|^2 = O_P(J^d/n + J^{-2p})$. Taking $J \asymp n^{1/(2p+d)}$, we get that $\|\hat{\eta}_s - \eta_s^*\|^2 = O_P(n^{-2p/(2p+d)})$ for $s \in \mathscr{S}$ and $\|\hat{\eta} - \eta^*\|^2 = O_P(n^{-2p/(2p+d)})$. The rate $n^{-2p/(2p+d)}$ is the optimal rate for estimating a p-smooth, d-dimensional function; see Stone (1982).

Remarks. 1. Theorem 3 is parallel to Theorem 3 of Huang (1998) for regression. We can obtain similar rate of convergence results when polynomials, trigonometric polynomials, or wavelets and their tensor products are used as building blocks for the approximating spaces. The same arguments as those in Huang (1998) can be used to check the conditions in Theorems 2 and 3. As a consequence, we can get the same message as that for regression: the structure of the approximating space is critical; as long as we construct the approximating space to have the same structure as the model space, under mild conditions, we can always get consistent estimates of the regression function and its ANOVA components.

2. The result from the previous example tells us that the rates of convergence are determined by the smoothness of the ANOVA components of η^* and the highest order of interactions included in the model. It also demonstrates that, by using models with only low-order interactions, we can ameliorate the curse of dimensionality that the saturated model suffers. For example, by considering additive models ($d = 1$) or by allowing interactions involving only two factors ($d = 2$), we can get faster rates of convergence than by using the saturated model ($d = L$).

3. In the example, the rate of convergence for each η_s^* is $n^{-2p/(2p+d)}$, which is the optimal rate only for $\#(s) = d = \max_{r \in \mathscr{S}} \#(r)$. If $\#(s) < d$, it is unclear whether $\hat{\eta}_s$ can achieve the optimal rate $n^{-2p/(2p+\#(s))}$ for η_s^*. In light of the work by Fan, Härdle, and Mammen (1996), if we assume the correct model ($\eta^* = \eta$) and focus on a specific ANOVA component, then it is possible to design an estimate to achieve the optimal rate for that component. But the message here is somewhat different. Using suitably defined ANOVA components of a single estimate, we can consistently estimate all ANOVA components of the target function. Moreover, when the model is misspecified, the estimate still converges to a reasonable target, that is, the best approximation.

4. Using univariate functions and their tensor products to model η^* restricts the domain of η^* to be a hyperrectangle. By allowing bivariate or multivariate functions and their tensor products to model η^*, we gain flexibility, especially when some predictor variable is of spatial type. Our theorems also apply to these cases, where the approximating spaces are built with multivariate splines and their tensor products or more general, finite-element spaces and their tensor products. The techniques in Huang (1998) can be employed to check the conditions of the theorems.

5. Functional ANOVA modeling is by no means the only way to achieve dimensionality reduction. For example, an effective alternative approach is the generalized linear model with unknown link; see Weisberg and Welsh (1994).

3. PRELIMINARIES

In this section, we collect some useful facts. Lemma 3.1 states that the empirical norm on G is equivalent to its theoretical counterpart. Corollary 3.1 gives us a sufficient condition for the empirical identifiability of G. Lemma 3.2 reveals that the theoretical components of H are not too confounded. Lemma 3.3 tells us that each function in G can be represented uniquely as a sum of the components in the empirical ANOVA decomposition. Lemma 3.4 states that the components of G are not too confounded, either empirically or theoretically.

The following lemma and corollary are borrowed from Huang (1998).

LEMMA 3.1. *Suppose* $\lim_n A_n^2 N_n/n = 0$ *and let* $t > 0$. *Then, except on an event whose probability tends to zero as* $n \to \infty$,

$$|\langle f, g\rangle_n - \langle f, g\rangle| \leqslant t\, \|f\|\, \|g\|, \qquad f, g \in G.$$

Consequently, except on an event whose probability tends to zero with n,

$$\tfrac{1}{2}\, \|g\|^2 \leqslant \|g\|_n^2 \leqslant 2\, \|g\|^2, \qquad g \in G.$$

COROLLARY 3.1. *Suppose* $\lim_n A_n^2 N_n/n = 0$. *Then, except on an event whose probability tends to zero as* $n \to \infty$, *G is empirically identifiable.*

Let $|\mathscr{X}|$ denote the Lebesgue measure of $\mathscr{X}$. Under Condition 2, let M_1 and M_2 be positive numbers such that

$$\frac{M_1^{-1}}{|\mathscr{X}|} \leqslant f_X(x) \leqslant \frac{M_2}{|\mathscr{X}|}, \qquad x \in \mathscr{X}.$$

Then $M_1, M_2 \geqslant 1$. The following two fundamental lemmas were essentially established in Stone (1994, Lemmas 3.1, 3.2).

Lemma 3.2. *Suppose Condition* 2 *holds. Set* $\varepsilon_1 = 1 - \sqrt{-1 - M_1^{-1} M_2^{-2}} \in (0, 1]$. *Then* $\|h\|^2 \geqslant \varepsilon_1^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \|h_s\|^2$ *for all* $h = \sum_s h_s$, *where* $h_s \in H_s^0$ *for* $s \in \mathcal{S}$.

Lemma 3.3. *Suppose* G *is empirically identifiable. Let* $g = \sum_{s \in \mathcal{S}} g_s$, *where* $g_s \in G_s^0$ *for* $s \in \mathcal{S}$. *If* $g = 0$, *then* $g_s = 0$ *for* $s \in \mathcal{S}$.

As a consequence of Lemma 3.2, each function in H can be represented uniquely as a sum of the components in the theoretical ANOVA decomposition. Since H_s, $s \in \mathcal{S}$, are Hilbert spaces equipped with the theoretical inner product, it is easily, shown by using Lemma 3.2 that, under Condition 2, H is a complete subspace of the space of all square-integrable functions on $\mathcal{X}$ equipped with the theoretical inner product. Lemma 3.3 tells us that each function $g \in G$ can be represented uniquely as a sum of the components in the empirical ANOVA decomposition.

According to next result, the components G_s^0, $s \in \mathcal{S}$, of g are not too confounded, either empirically or theoretically. This result was established in Huang (1998).

Lemma 3.4. *Suppose Condition* 2 *holds and that* $\lim_n A_s^2 N_{s'}/n = 0$ *for* $s, s' \in \mathcal{S}$. *Let* $0 < \varepsilon_2 < \varepsilon_1$. *Then, except on an event whose probability tends to zero as* $n \to \infty$, $\|g\|^2 \geqslant \varepsilon_2^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \|g_s\|^2$ *and* $\|g\|_n^2 \geqslant \varepsilon_2^{\#(\mathcal{S})-1} \sum_{s \in \mathcal{S}} \|g_s\|_n^2$ *for all* $g = \sum_{s \in \mathcal{S}} g_s$, *where* $g_s \in G_s^0$ *for* $s \in \mathcal{S}$.

4. PROOF OF THEOREM 1

The proof of Theorem 1 is divided into two parts. The approximation and estimation errors are handled separately.

4.1. *Approximation Error*

Lemma 4.1. *Suppose Condition* 1 *holds. Let* $\mathcal{K}$ *be a compact subinterval of* $\mathcal{I}$ *such that* range$(\eta^*) \subset \mathcal{K}$. *Then there are positive numbers* M_3 *and* M_4 *such that*

$$-M_3 \|h - \eta^*\|^2 \leqslant \Lambda(h) - \Lambda(\eta^*) \leqslant -M_4 \|h - \eta^*\|^2$$

for all $h \in H$ *with* range$(h) \subset \mathcal{K}$.

Proof. Given $h \in H$ with range$(h) \subset \mathscr{K}$, set $h^{(t)} = (1-t)\,\eta^* + th$. Then

$$\frac{d}{dt}\,\Lambda(h^{(t)})\bigg|_{t=0} = 0$$

and hence

$$\Lambda(h) - \Lambda(\eta^*) = \int_0^1 (1-t)\,\frac{d^2}{dt^2}\,\Lambda(h^{(t)})\,dt$$

(integrate by parts). Observe that

$$\frac{d^2}{dt^2}\,\Lambda(h^{(t)}) = E\{(h(X) - \eta^*(X))^2\,[B''(h^{(t)}(X))\,A(\eta(X)) - C''(h^{(t)}(X))]\}.$$

By (2) and the continuity of the functions $A(\cdot)$, $B''(\cdot)$, and $C''(\cdot)$,

$$\inf_{\substack{\xi \in \mathscr{K} \\ \eta \in \mathscr{K}_0}} [B''(\xi)\,A(\eta) - C''(\xi)] := -2M_3 < 0$$

and

$$\sup_{\substack{\xi \in \mathscr{K} \\ \eta \in \mathscr{K}_0}} [B''(\xi)\,A(\eta) - C''(\xi)] := -2M_4 < 0.$$

The desired result now follows. ▌

Recall that $\bar{\eta}$ is the best approximation in G^* to η.

Lemma 4.2. *Suppose Condition* 1 *holds and that* $\lim_n A_n \rho_n = 0$. *Then* $\bar{\eta}$ *exists for n sufficiently large and satisfies* $\|\bar{\eta} - \eta^*\|^2 = O(\rho_n^2)$ *and* $\|\bar{\eta} - \eta^*\|_n^2 = O_P(\rho_n^2)$.

Proof. Let $g^* \in G$ be such that $\|g^* - \eta^*\|_\infty = \rho_n$. Let a denote a positive constant (to be determined later). Choose $g \in G$ with $\|g - \eta^*\| \leqslant a\rho_n$. Then

$$\|g - g^*\|_\infty \leqslant A_n\,\|g - g^*\| \leqslant A_n(\|g - \eta^*\| + \|\eta^* - g^*\|) \leqslant A_n(a\rho_n + \rho_n).$$

Note that η^* takes values in a compact subinterval of $\mathscr{I}$. Since $\lim_n \rho_n = 0$ and $\lim_n A_n \rho_n = 0$, for n sufficiently large there is a compact subinterval $\mathscr{K}$ of $\mathscr{I}$ such that range$(g^*) \subset \mathscr{K}$ and range$(g) \subset \mathscr{K}$ for all $g \in G$ with

$\|g-\eta^*\| \leqslant a\rho_n$. Thus, it follows from Lemma 4.1 that, for n sufficiently large,

$$\Lambda(g)-\Lambda(\eta^*) \leqslant -M_4 a^2 \rho_n^2 \qquad \text{for all} \quad g \in G \qquad \text{with} \quad \|g-\eta^*\| = a\rho_n \tag{6}$$

and

$$\Lambda(g^*)-\Lambda(\eta^*) \geqslant -M_3 \rho_n^2. \tag{7}$$

Let a be chosen such that $a > \sqrt{(M_3/M_4)}$. Then $\|g^*-\eta^*\| < a\rho_n$, and it follows from (6) and (7) that, for n sufficiently large,

$$\Lambda(g) < \Lambda(g^*) \qquad \text{for all} \quad g \in G \qquad \text{with} \quad \|g-\eta^*\| = a\rho_n.$$

Note that, for n sufficiently large, $g^* \in G^*$ and $g \in G^*$ for all $g \in G$ with $\|g-\eta^*\| \leqslant a\rho_n$. Therefore, by the definition of $\bar{\eta}$ and the concavity of $\Lambda(g)$ as a function of g, $\bar{\eta}$ exists and satisfies $\|\bar{\eta}-\eta^*\| < a\rho_n$ for n sufficiently large. Hence $\|\bar{\eta}-\eta^*\|^2 = O(\rho_n^2)$. To prove that $\|\bar{\eta}-\eta^*\|_n^2 = O_P(\rho_n^2)$, by the triangle inequality and Lemma 3.1, we have that

$$\begin{aligned} \|\bar{\eta}-\eta^*\|_n &\leqslant \|\bar{\eta}-g^*\|_n + \|g^*-\eta^*\|_n \\ &\leqslant 2\,\|\bar{\eta}-g^*\| + \|g^*-\eta^*\|_\infty \leqslant 2\,\|\bar{\eta}-\eta^*\| + 3\,\|g^*-\eta^*\|_\infty, \end{aligned}$$

except on an event whose probability tends to zero as $n \to \infty$. The desired result now follows. ▌

4.2. *Estimation Error*

Let $\{\phi_j, 1 \leqslant j \leqslant N_n\}$ be an orthonormal basis of G relative to the theoretical inner product. Then each $g \in G$ can be represented uniquely as $g = \sum_j \beta_j \phi_j$, where $\beta_j = \langle g, \phi_j \rangle$ for $j = 1, ..., N_n$. Let $\boldsymbol{\beta}$ denote the N_n-dimensional vector with entries β_j. To indicate the dependence of g on $\boldsymbol{\beta}$, we write $g(\cdot) = g(\cdot; \boldsymbol{\beta})$. Let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^{N_n}$. Then $\|g(\cdot; \boldsymbol{\beta})\| = |\boldsymbol{\beta}|$.

We write $\ell(g(\cdot; \boldsymbol{\beta}))$ as $\ell(\boldsymbol{\beta})$. Let $\mathbf{S}(\boldsymbol{\beta}) = (\partial/\partial\boldsymbol{\beta})\, \ell(\boldsymbol{\beta})$ denote the score at $\boldsymbol{\beta}$, that is, the N_n-dimensional vector having entries

$$\frac{\partial}{\partial \beta_j} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_i \phi_j(X_i)[B'(g(X_i; \boldsymbol{\beta}))\ Y_i - C'(g(X_i; \boldsymbol{\beta}))],$$

and let $\mathbf{D}(\boldsymbol{\beta}) = (\partial^2/\partial\boldsymbol{\beta}\, \partial\boldsymbol{\beta}^T)\, \ell(\boldsymbol{\beta})$ be the $N_n \times N_n$ Hessian matrix, which has entries

$$\frac{\partial^2}{\partial \beta_{j_1}\, \partial \beta_{j_2}} \ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_i \phi_{j_1}(X_i)\, \phi_{j_2}(X_i)[B''(g(X_i; \boldsymbol{\beta}))\ Y_i - C''(g(X_i; \boldsymbol{\beta}))].$$

LEMMA 4.3. *Suppose* $\lim_n A_n^2 N_n/n = 0$. *Let* $\mathcal{K}$ *be a compact subinterval of* $\mathcal{I}$. *Then, there is a positive constant* M_5 *such that, except on an event whose probability tends to zero as* $n \to \infty$,

$$\frac{d^2}{dt^2}\ell(g_0 + t(g_1 - g_0)) \leqslant -M_5 \|g_1 - g_0\|^2$$

for $0 < t < 1$ *and all* $g_0, g_1 \in G$ *with* $\operatorname{range}(g_0), \operatorname{range}(g_1) \subset \mathcal{K}$.

Proof. Let $\boldsymbol{\beta}_0 = (\beta_{0j})$ and $\boldsymbol{\beta}_1 = (\beta_{1j})$ be given by the equations $g_0 = \sum_j \beta_{0j}\phi_j$ and $g_1 = \sum_j \beta_{1j}\phi_j$. Then $\|g_1 - g_0\|^2 = |\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0|^2$. Moreover,

$$\begin{aligned}\frac{d^2}{dt^2}\ell(g_0 + t(g_1 - g_0)) &= \frac{d^2}{dt^2}\ell(\boldsymbol{\beta}_0 + t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)) \\ &= (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^T \mathbf{D}(\boldsymbol{\beta}_0 + t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0))(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \qquad (8)\end{aligned}$$

for $0 < t < 1$. We need the following result (to be proved later):

CLAIM 1. *There exist a positive constant* δ_1 *and a compact subinterval* S_0 *of* S *such that* $P(Y \in S_0 \mid X = x) \geqslant \delta_1$ *for* $x \in \mathcal{X}$ *and* $B''(\xi)\, y - C''(\xi) < 0$ *for* $\xi \in \mathcal{I}$ *and* $y \in S_0$.

By Claim 1 and the continuity of B'' and C'', there is a positive constant δ_2 such that

$$B''(\xi)\, y - C''(\xi) \leqslant -\delta_2, \qquad \xi \in \mathcal{K} \qquad \text{and} \qquad y \in S_0. \qquad (9)$$

Set $\mathcal{I}_n = \{i\colon 1 \leqslant i \leqslant n \text{ and } Y_i \in S_0\}$. By (2) and (9), except on an event whose probability tends to zero as $n \to \infty$,

$$\begin{aligned}&(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^T \mathbf{D}(\boldsymbol{\beta}_0 + t(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0))(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \\ &\quad = \frac{1}{n}\sum_i \{[g_1(X_i) - g_0(X_i)]^2 \\ &\qquad \times [B''([g_0 + t(g_1 - g_0)](X_i))\, Y_i - C''([g_0 + t(g_1 - g_0)](X_i))]\} \\ &\quad \leqslant -\frac{\delta_2}{n}\sum_{i \in \mathcal{I}_n} [g_1(X_i) - g_0(X_i)]^2 \qquad (10)\end{aligned}$$

for all $g_0, g_1 \in G$ with $\operatorname{range}(g_0), \operatorname{range}(g_1) \subset \mathcal{K}$. Set $I_n = \#(\mathcal{I}_n)$. Then $\lim_n P(I_n \geqslant \delta_1 n/2) = 1$. Observe that, given $\mathcal{I}_n = \{i_1, ..., i_{I_n}\}$, the covariates X_j, $j \in \mathcal{I}_n$, are independent and have the common density

$$f(x \mid Y \in S_0) = \frac{f_X(x)\, P(Y \in S_0 \mid X = x)}{P(Y \in S_0)}.$$

Note that $\delta_1 f_X(x) \leqslant f(x \mid Y \in S_0) \leqslant (1/\delta_1)\, f_X(x)$. Therefore, it follows from Lemma 3.1 that

$$\frac{\delta_2}{n} \sum_{i \in \mathscr{I}_n} [g_1(X_i) - g_0(X_i)]^2 \geqslant M_5 \| g_1 - g_0 \|^2 \tag{11}$$

for all g_0, $g_1 \in G$ with range(g_0), range(g_1) $\subset \mathscr{K}$, except on an event whose probability tends to zero as $n \to \infty$. Lemma 4.3 now follows from (10) and (11).

Proof of Claim 1. By Assumptions 3 and 4, $P(Y \in S) = 1$ and $\eta(\cdot)$ takes values in a compact subinterval $\mathscr{K}_0$ of $\mathscr{I}$. Since $A(\cdot)$ is continuous and increasing, $E(Y \mid X = x) = A(\eta(x))$ ranges over a compact subinterval $S_1 = [c_1, c_2]$ of $\mathring{S}$. We consider three cases.

Case I. $S = \mathbb{R}$. By Chebyshev's inequality and Assumption 5,

$$P(|Y - E(Y \mid X = x)| \leqslant \sqrt{2D} \mid X = x) \geqslant 1 - \frac{\operatorname{var}(Y \mid X = x)}{2D} \geqslant \frac{1}{2}, \qquad x \in \mathscr{X}.$$

Therefore, Claim 1 holds with $S_0 = [c_1 - \sqrt{2D}, c_2 + \sqrt{2D}]$ and $\delta_1 = 1/2$.

Case II. $\mathring{S} = (-\infty, a)$ or (a, ∞) for some $a \in \mathbb{R}$. Without loss of generality, suppose that $\mathring{S} = (0, \infty)$. Otherwise, we can replace Y by $-Y + a$ or $Y - a$. Thus $0 < c_1 < c_2$. By Assumption 5,

$$E(Y^2 \mid X = x) = \operatorname{var}(Y \mid X = x) + [E(Y \mid X = x)]^2 \leqslant D + c_2^2.$$

By an obvious modification of Markov's inequality, for any $M > 0$,

$$E[Y \operatorname{ind}(Y > M) \mid X = x] \leqslant \frac{E(Y^2 \mid X = x)}{M} \leqslant \frac{D + c_2^2}{M};$$

here ind(A) denotes the indicator function of the set A. Hence, for any δ, $M \in \mathbb{R}$ with $M > \delta > 0$,

$$\begin{aligned} c_1 &\leqslant E(Y \mid X = x) \\ &= E(Y \operatorname{ind}(Y < \delta) \mid X = x) \\ &\quad + E(Y \operatorname{ind}(\delta \leqslant Y \leqslant M) \mid X = x) + E(Y \operatorname{ind}(Y > M) \mid X = x) \\ &\leqslant \delta + MP(\delta \leqslant Y \leqslant M \mid X = x) + \frac{D + c_2^2}{M}. \end{aligned}$$

This implies that

$$P(\delta \leqslant Y \leqslant M \mid X=x) \geqslant \frac{c_1 - \delta - (D + c_2^2)/M}{M}.$$

Letting $\delta = c_1/3$ and $M = 3(D + c_2^2)/c_1$, we get that

$$P(\delta \leqslant Y \leqslant M \mid X=x) \geqslant \frac{c_1^2}{9(D + c_2^2)} > 0.$$

Therefore, Claim 1 holds with $S_0 = [c_1/3, 3(D + c_2^2)/c_1]$ and $\delta_1 = c_1^2/(9(D + c_2^2))$.

Case III. $\mathring{S} = (a_1, a_2)$ for $a_1, a_2 \in \mathbb{R}$ and (2) holds at $y = a_1$ or $y = a_2$. Without loss of generality, suppose that $\mathring{S} = (0, 1)$ and (2) holds at $y = 1$. Otherwise, we can replace Y by $(Y - a_1)/(a_2 - a_1)$ or $(-Y + a_2)/(a_2 - a_1)$. Thus $Y \leqslant 1$ and $c_1 > 0$. Note that, for $\delta > 0$,

$$c_1 \leqslant E(Y \mid X=x) \leqslant \delta + P(Y \geqslant \delta \mid X=x), \qquad x \in \mathcal{X}.$$

Let $\delta = c_1/2$. Then $P(Y \geqslant c_1/2 \mid X=x) \geqslant c_1/2$ for $x \in \mathcal{X}$. Therefore, Claim 1 holds with $S_0 = [c_1/2, 1]$ and $\delta_1 = c_1/2$.

The proof of Lemma 4.3 is complete. ▮

Corollary 4.1. *Suppose* $\lim_n A_n^2 N_n/n = 0$. *Then the log-likelihood* $\ell(g)$ *is strictly concave on* G^* *except on an event whose probability tends to zero as* $n \to \infty$.

Lemma 4.4. *Suppose Condition* 1 *holds and that* $\lim_n A_n^2 N_n/n = 0$ *and* $\lim_n A_n \rho_n = 0$. *Then* $\hat{\eta}$ *exists except on an event whose probability tends to zero as* $n \to \infty$. *Moreover,* $\|\hat{\eta} - \bar{\eta}\|^2 = O_P(N_n/n)$.

Proof. Recall that $\hat{\eta}$ is the maximum likelihood estimate and $\bar{\eta}$ is the best approximation in G^* to η. By Lemma 4.2, $\bar{\eta}$ exists for n sufficiently large. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_j)$ and $\bar{\boldsymbol{\beta}} = (\bar{\beta}_j)$ be given by the equations $\hat{\eta} = \sum_j \hat{\beta}_j \phi_j$ and $\bar{\eta} = \sum_j \bar{\beta}_j \phi_j$. Then $\|\hat{\eta} - \bar{\eta}\|^2 = |\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}|^2$ and $\|g - \bar{\eta}\|^2 = |\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}|^2$ for $g = g(\cdot, \boldsymbol{\beta})$. Moreover, the following identity holds:

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \ell(\bar{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T \mathbf{S}(\bar{\boldsymbol{\beta}}) \\ &\quad + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T \left[\int_0^1 (1-t)\, \mathbf{D}(\bar{\boldsymbol{\beta}} + t(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}))\, dt \right] (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}). \end{aligned} \tag{12}$$

To complete the proof of the lemma, we need the following two results (to be proved later):

CLAIM 2. *For any positive constant M,*

$$\lim_{a\to\infty}\limsup_{n\to\infty} P\left(|\mathbf{S}(\bar{\boldsymbol\beta})| \geqslant Ma\left(\frac{N_n}{n}\right)^{1/2}\right)=0.$$

CLAIM 3. *There is a positive constant M_6 such that, for any fixed positive constant a,*

$$(\boldsymbol\beta-\bar{\boldsymbol\beta})^T\left[\int_0^1 (1-t)\,\mathbf{D}(\bar{\boldsymbol\beta}+t(\boldsymbol\beta-\bar{\boldsymbol\beta}))\,dt\right](\boldsymbol\beta-\bar{\boldsymbol\beta})$$
$$\leqslant -M_6\,|\boldsymbol\beta-\bar{\boldsymbol\beta}|^2 \qquad \textit{for all}\quad \boldsymbol\beta\in\mathbb{R}^{N_n} \qquad \textit{with}\quad |\boldsymbol\beta-\bar{\boldsymbol\beta}|=a\left(\frac{N_n}{n}\right)^{1/2}$$

on an event $\Omega_n(a)$ with $\lim_n P(\Omega_n(a))=1$.

Choose $\boldsymbol\beta\in\mathbb{R}^{n_n}$ such that $|\boldsymbol\beta-\bar{\boldsymbol\beta}|=a(N_n/n)^{1/2}$. Then by Condition 2, we have that $\|g(\cdot;\boldsymbol\beta)-\bar\eta\|_\infty\leqslant A_n\,\|g(\cdot;\boldsymbol\beta)-\bar\eta\|=a(A_n^2N_n/n)^{1/2}=o(1)$. Note that $\bar\eta\in G^*$. Thus $g(\cdot;\boldsymbol\beta)\in G^*$ for n sufficiently large. Fix $\varepsilon>0$. By Claim 2, we can choose a sufficiently large such that $|\mathbf{S}(\bar{\boldsymbol\beta})|<M_6a(N_n/n)^{1/2}$ except on an event whose probability is less than ε. On the nonexceptional event,

$$|(\boldsymbol\beta-\bar{\boldsymbol\beta})^T\,\mathbf{S}(\bar{\boldsymbol\beta})|<M_6a^2\left(\frac{N_n}{n}\right) \qquad \text{for all}\quad \boldsymbol\beta\in\mathbb{R}^{N_n}$$
$$\text{with}\quad |\boldsymbol\beta-\bar{\boldsymbol\beta}|=a\left(\frac{N_n}{n}\right)^{1/2}. \qquad (13)$$

Moreover, it follows from Claim 3 that, except on an event whose probability tends to zero as $n\to\infty$,

$$(\boldsymbol\beta-\bar{\boldsymbol\beta})^T\left[\int_0^1 (1-t)\,\mathbf{D}(\bar{\boldsymbol\beta}+t(\boldsymbol\beta-\bar{\boldsymbol\beta}))\,dt\right](\boldsymbol\beta-\bar{\boldsymbol\beta})$$
$$\leqslant -M_6a^2\left(\frac{N_n}{n}\right) \qquad \text{for all}\quad \boldsymbol\beta\in\mathbb{R}^{N_n} \qquad \text{with}\quad |\boldsymbol\beta-\bar{\boldsymbol\beta}|=a\left(\frac{N_n}{n}\right)^{1/2}. \qquad (14)$$

Suppose (13) and (14) hold. Then, by (12), $\ell(\boldsymbol\beta)<\ell(\bar{\boldsymbol\beta})$ for all $\boldsymbol\beta\in\mathbb{R}^{N_n}$ with $|\boldsymbol\beta-\bar{\boldsymbol\beta}|=a(N_n/n)^{1/2}$. Hence by the concavity of $\ell(\boldsymbol\beta)$ as a function of $\boldsymbol\beta$, $\hat\eta=g(\cdot,\hat{\boldsymbol\beta})$ exists and satisfies $\|\hat\eta-\bar\eta\|\leqslant a(N_n/n)^{1/2}$. Since ε is arbitrary, the conclusion of the lemma follows. ▮

Proof of Claim 2. Since $\bar{\boldsymbol\beta}$ maximizes

$$\Lambda(g(\cdot;\boldsymbol\beta))=E[B(g(X;\boldsymbol\beta))\;Y-C(g(X;\boldsymbol\beta))],$$

we have that

$$\frac{d}{d\boldsymbol{\beta}} \Lambda(g(\cdot;\boldsymbol{\beta}))\bigg|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} = 0.$$

This implies that

$$E[\phi_j(X)(B'(\bar{\eta}(X))\, Y - C'(\bar{\eta}(X)))] = 0, \qquad 1 \leqslant j \leqslant N_n.$$

Thus

$$E(|\mathbf{S}(\bar{\boldsymbol{\beta}})|^2) = \sum_j E\left[\frac{\partial}{\partial \beta_j} \ell(\bar{\boldsymbol{\beta}})\right]^2 = \frac{1}{n} \sum_j \operatorname{var}[\phi_j(X)(B'(\bar{\eta}(X))\, Y - C'(\bar{\eta}(X)))].$$

Note that

$$\begin{aligned}
&\operatorname{var}[\phi_j(X)(B'(\bar{\eta}(X))\, Y - C'(\bar{\eta}(X)))] \\
&\quad = E[\operatorname{var}[\phi_j(X)(B'(\bar{\eta}(X))\, Y - C'(\bar{\eta}(X))) \mid X]] \\
&\qquad + \operatorname{var}[E[\phi_j(X)(B'(\bar{\eta}(X))\, Y - C'(\bar{\eta}(X))) \mid X]] \\
&\quad = E[\phi_j^2(X)(B'(\bar{\eta}(X)))^2\, \sigma^2(X)] \\
&\qquad + \operatorname{var}[\phi_j(X)(B'(\bar{\eta}(X))\, A(\eta(X)) - C'(\bar{\eta}(X)))] \\
&\quad \leqslant M E[\phi_j^2(X_i)]
\end{aligned}$$

for some positive constant M by Lemma 4.2, Assumption 5, and the continuity of $B'(\cdot)$, $C'(\cdot)$, and $A(\cdot)$. Consequently,

$$E(|\mathbf{S}(\bar{\boldsymbol{\beta}})|^2) \leqslant \frac{M}{n} \sum_j E[\phi_j^2(X_i)] = \frac{M}{n} \sum_j \|\phi_j\|^2 = M \frac{N_n}{n},$$

which yields Claim 2.

Proof of Claim 3. Choose $g \in G$ such that $\|g - \bar{\eta}\| = a(N_n/n)^{1/2}$. Then $\|g - \bar{\eta}\|_\infty \leqslant A_n \|g - \bar{\eta}\| = A_n a(N_n/n)^{1/2} = o(1)$. Thus by Lemma 4.2, for n sufficiently large, there is a compact subinterval $\mathscr{K}$ of $\mathscr{I}$ such that $\operatorname{range}(\bar{\eta}) \subset \mathscr{K}$ and $\operatorname{range}(g) \subset \mathscr{K}$ for all $g \in G$ with $\|g - \bar{\eta}\| = a(N_n/n)^{1/2}$. Hence it follows from Lemma 4.3 that, except on an event whose probability tends to zero as $n \to \infty$,

$$\frac{d^2}{dt^2} \ell(\bar{\eta} + t(g - \bar{\eta})) \leqslant -M_5 \|g - \bar{\eta}\|^2$$

for $0 < t < 1$ and all $g \in G$ with $\|g - \bar{\eta}\| = a(N_n/n)^{1/2}$. Equivalently, by (8),

$$(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T \mathbf{D}(\bar{\boldsymbol{\beta}} + t(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}))(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \leqslant -M_5 |\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}|^2$$

for $0<t<1$ and all $\boldsymbol{\beta}\in\mathbb{R}^{N_n}$ with $|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}|=a(N_n/n)^{1/2}$ on an event $\Omega_n(a)$ with $\lim_n P(\Omega_n(a))=1$. Claim 3 now follows with $M_6=M_5/\int_0^1(1-t)\,dt=M_5/2$.

The proof of Lemma 4.4 is complete. ▮

Theorem 1 follows from Lemmas 3.1, 4.2, and 4.4.

5. PROOF OF THEOREM 3

Recall that the estimation error has the ANOVA decomposition $\hat{\eta}-\bar{\eta}=\sum_{s\in\mathscr{S}}(\hat{\eta}_s-\bar{\eta}_s)$, where $\hat{\eta}_s, \bar{\eta}_s\in G_s^0$. The following lemma gives the rates of convergence of the various components of $\hat{\eta}-\bar{\eta}$.

Lemma 5.1. *Suppose Conditions* 1 *and* 2 *hold and that* $\lim_n A_s^2 N_{s'}/n=0$ *and* $\lim_n A_s\rho_{s'}=0$ *for* $s, s'\in\mathscr{S}$. *Then, four each* $s\in\mathscr{S}$,

$$\|\hat{\eta}_s-\bar{\eta}_s\|^2=O_P\left(\sum_{s\in\mathscr{S}}N_s/n\right)\qquad \textit{and}\qquad \|\hat{\eta}_s-\bar{\eta}_s\|_n^2=O_P\left(\sum_{s\in\mathscr{S}}N_s/n\right).$$

Proof. The desired results follow from Theorem 2 and Lemmas 3.4 and 4.4. ▮

Recall that $\eta_s^*\in H_s^0$, $s\in\mathscr{S}$, are the components in the ANOVA decomposition of η^*. The following lemma, which is borrowed from Huang (1998), tells us how well η_s^* can be approximated by functions in G_s^0.

Lemma 5.2. *Suppose Conditions* 1 *and* 2 *hold and that* $\lim_n A_s^2 N_{s'}/n=0$ *for* $s, s'\in\mathscr{S}$. *In addition, assume that* η_s^* *is bounded for* $s\in\mathscr{S}$. *Then, for each* $s\in\mathscr{S}$, *there exist functions* $g_s\in G_s^0$ *such that*

$$\|\eta_s^*-g_s\|^2=O_P\left(\sum_{r\subset s,\,r\neq s}\frac{N_r}{n}+\rho_s^2\right)\tag{15}$$

and

$$\|\eta_s^*-g_s\|_n^2=O_P\left(\sum_{r\subset s,\,r\neq s}\frac{N_r}{n}+\rho_s^2\right).\tag{16}$$

Recall that $\bar{\eta}-\eta^*$ is the approximation error. We have the ANOVA decompositions $\bar{\eta}=\sum_{s\in\mathscr{S}}\bar{\eta}_s$ and $\eta^*=\sum_{s\in\mathscr{S}}\eta_s^*$, where $\bar{\eta}_s\in G_s^0$ and $\eta_s^*\in H_s^0$ for $s\in\mathscr{S}$. The next lemma gives the rates of convergence of the various components of $\bar{\eta}-\eta^*$.

Lemma 5.3. *Suppose Conditions* 1 *and* 2 *hold and that* $\lim_n A_s^2 N_{s'}/n = 0$ *and* $\lim_n A_s \rho_{s'} = 0$ for $s, s' \in \mathscr{S}$. *Then, for each* $s \in \mathscr{S}$,

$$\|\bar{\eta}_s - \eta_s^*\|^2 = O_P\left(\sum_{s \in \mathscr{S}} \frac{N_s}{n} + \sum_{s \in \mathscr{S}} \rho_s^2\right)$$

and

$$\|\bar{\eta}_s - \eta_s^*\|_n^2 = O_P\left(\sum_{s \in \mathscr{S}} \frac{N_s}{n} + \sum_{s \in \mathscr{S}} \rho_s^2\right).$$

Proof. By Lemma 5.2, for each $s \in \mathscr{S}$, there are functions $g_s \in G_s^0$ such that (15) and (16) hold. Write $g = \sum_{s \in \mathscr{S}} g_s$. Then $\|g - \eta^*\|^2 = O_P(\sum_{s \in \mathscr{S}} N_s/n + \sum_{s \in \mathscr{S}} \rho_s^2)$. Thus, by Lemma 4.2,

$$\|g - \bar{\eta}\|^2 \leqslant 2\,\|g - \eta^*\|^2 + 2\,\|\bar{\eta} - \eta^*\|^2 = O_P\left(\sum_{s \in \mathscr{S}} \frac{N_s}{n} + \sum_{s \in \mathscr{S}} \rho_s^2\right).$$

Therefore, by Lemma 3.4, except on an event whose probability tends to zero as $n \to \infty$,

$$\|g_s - \bar{\eta}_s\|^2 \leqslant \varepsilon_2^{1 - \#(s)}\,\|g - \bar{\eta}\|^2 = O_P\left(\sum_{s \in \mathscr{S}} \frac{N_s}{n} + \sum_{s \in \mathscr{S}} \rho_s^2\right).$$

Hence, the desired results follow from (15), (16), the triangle inequality, and Lemma 3.1. ▮

Theorem 3 follows from Lemmas 5.1 and 5.3.

ACKNOWLEDGMENTS

This work is part of the author's Ph.D. dissertation at the University of California, Berkeley, written under the supervision of Professor Charles J. Stone, whose generous guidance and suggestions are gratefully appreciated. The author thanks a referee for helpful comments.

REFERENCES

1. C. de Boor, "A Practical Guide to Splines," Springer-Verlag, New York, 1978.
2. J. Fan, W. Härdle, and E. Mammen, Direct estimation of low dimensional components in additive models, manuscript, 1996.
3. M. Hansen, Extended linear models, multivariate splines, and ANOVA, Ph.D. dissertation, University of California at Berkeley, 1994.

4. J. Z. Huang, Projection estimation in multiple regression with application to functional ANOVA models, *Ann. Statist.* **26** (1998), 242–272.
5. P. McCullagh and J. A. Nelder, “Generalized Linear Models,” 2nd ed., Chapman & Hall, London, 1989.
6. L. L. Schumaker, “Spline Functions: Basic Theory,” Wiley, New York, 1981.
7. C. J. Stone, Optimal global rates of convergence for nonparametric regression, *Ann. Statist.* **8** (1982), 1348–1360.
8. C. J. Stone, The dimensionality reduction principle for generalized additive models, *Ann. Statist.* **14** (1986), 590–606.
9. C. J. Stone, The use of polynomial splines and their tensor products in multivariate function estimation (with discussion), *Ann. Statist.* **22** (1994), 118–171.
10. C. J. Stone, M. Hansen, C. Kooperberg, and Y. Truong, Polynomial splines and their tensor products in extended linear modeling, *Ann. Statist.* **25** (1997), 1371–1470.
11. S. Weisberg and A. H. Welsh, Adapting for the missing link, *Ann. Statist.* **22** (1994), 1674–1700.