Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT

Statistics & Probability Letters 65 (2003) 207–216

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Asymptotics for polynomial spline regression under weak conditions

Jianhua Z. Huang

The Wharton School, Department of Statistics, University of Pennsylvania, 467 JMHH, 3760 Walnut Street, Philadelphia, PA 19104-6340, USA

Abstract

We derive rate of convergence for spline regression under weak conditions. The results improve upon those in Huang (Ann. Statist. 26 (1998a) 242) in two ways. Firstly, results are obtained under less stringent conditions on the growing rate of the number of knots. Secondly, L_2 approximation instead of L_∞ approximation by splines are used to bound the rate of convergence, this allows us to obtain results when the regression function or its components in ANOVA decomposition are in broader function classes.

Keywords: Functional ANOVA models; Generalized additive models; Least-squares projection

1. Introduction

Polynomial splines and their tensor products provide useful tools for nonparametric function estimation. Used with functional ANOVA decompositions, they provide convenient building blocks for fitting structural models that are useful in overcoming the "curse of dimensionality" for high-dimensional problems. See Stone et al. (1997) and Huang (2001). Stone (1994) gives the first theoretical treatment on rates of convergence of spline estimation with functional ANOVA decompositions. In the regression context, Huang (1998a) substantially simplifies the treatment of Stone and the new approach has led to many further theoretical development of spline estimation in a variety of statistical contexts, including logistic and other generalized regression, density and conditional density estimation, hazard and conditional hazard estimation. See Huang (1998b), Huang and Stone (1998), Huang et al. (2000), Huang and Stone (2002a, 2003). See also, Huang and Stone (2002b) for a review.

E-mail address: jianhua@wharton.upenn.edu (J.Z. Huang).

0167-7152/$ - see front matter
doi:10.1016/j.spl.2003.09.003

In this paper, we improve the results of Huang (1998a) in two different ways. Firstly, we derive the rate of convergence of spline estimators under less stringent conditions on the growing rate of the number of knots. Secondly, we use L_2 approximation instead of L_∞ approximation by splines to bound the rate of convergence. Using L_2 approximation allows us to obtain rate of convergence of spline estimators when the regression function or its components in ANOVA decomposition are in broader function classes.

The paper is organized as follows. The main result is presented in Section 2 and its application to special cases are given in Section 3, where comparison with results in Huang (1998a) can also be found. The proof of the main result is given in Section 4. Some key lemmas, which are also of independent interest, are given in Section 5.

Notation. Given positive numbers a_n and b_n for $n \geqslant 1$, let $a_n \lesssim b_n$ mean that a_n/b_n is bounded and let $a_n \asymp b_n$ mean that $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2. The main asymptotic result

Consider the following regression problem. Let X represent the predictor variable and Y the real-valued response variable, where X and Y have a joint distribution. We assume that X ranges over a compact subset $\mathscr{X}$ of some Euclidean space. Set $\mu(x)=E(Y|X=x)$ and $\sigma^2(x)=\mathrm{var}(Y|X=x)$. The primary interest is in estimating μ based on a random sample from the distribution of (X,Y).

To facilitate our discussion, we introduce some inner products and norms as in Huang (1998a). For any integrable function f defined on $\mathscr{X}$, set $E_n(f)=(1/n)\sum_{i=1}^n f(X_i)$ and $E(f)=E[f(X)]$. Define the empirical inner product and norm as $\langle f_1, f_2\rangle_n = E_n(f_1 f_2)$ and $\|f_1\|_n^2 = \langle f_1, f_1\rangle_n$ for square-integrable functions f_1 and f_2 on $\mathscr{X}$. The theoretical inner product and norm are given by $\langle f_1, f_2\rangle = E(f_1 f_2)$ and $\|f_1\|^2 = \langle f_1, f_1\rangle$.

Condition 1: $(X_1,Y_1),\ldots,(X_n,Y_n)$ is an i.i.d. sample from the distribution of (X,Y).

Condition 2: $E(Y^2|X=x)$ is bounded, i.e., $\sup_{x\in\mathscr{X}} E(Y^2|X=x)<\infty$.

Condition 2 is equivalent to the requirement that $\mu(\cdot)$ and $\sigma^2(\cdot)$ are bounded on $\mathscr{X}$.

Condition 3: There is a positive constants M such that for any Borel set $B\subset\mathscr{X}$, $P(X\in B)\geqslant M\,\mathrm{Leb}(B)$, where $\mathrm{Leb}(\cdot)$ denote the Lebesgue measure.

Note that Condition 3 does not require that the distribution of X is absolute continuous or equivalently X has a Lebesgue density. If X has a Lebesgue density, then Condition 3 is implied by the requirement that this density is bounded away from zero. Condition 3 implies that, for any measurable function f, $\|f\|^2 \geqslant M\|f\|_2^2$, where $\|f\|_2^2=\int_{\mathscr{X}} f^2(x)\,\mathrm{d}x$. Condition 3 is used mainly in Lemma 2 to give a sufficient condition for Condition 4 below and to ensure the uniqueness of ANOVA decomposition in Corollary 2.

The method of estimation is least-squares projection onto a finite-dimensional linear space G whose dimensionality N_n is allowed to increase with the sample size. More precisely, the least-squares estimate $\hat{\mu}$ is defined as the element $g\in G$ that minimizes $\sum_{i=1}^n [g(X_i)-Y_i]^2$.

In practice, one introduces a working model that requires $\mu\in H$, where H specifies some structure on μ. For statistical estimation, one should construct a fitting space G that incorporate similar structures as in the working model. For example, when H is the space of all additive functions on a set of variables, H specifies an additive model. Correspondingly, G can be taken as a space of

additive splines. Usually, $\mu \notin H$ and the working model is only an approximation. In this case, we think that the least-squares estimate $\hat{\mu}$ is estimating the best approximation μ^* of μ in H, where "best" means minimizing the mean squared error of approximation. It is easily seen that μ^* is the orthogonal projection of μ on H relative to the theoretical inner product. The main result of the paper concerns about the rate of convergence of $\hat{\mu}$ to μ^*.

The following condition, which is crucial for the asymptotic result, says that the empirical and theoretical norms are close uniformly on the spaces of fits with probability tending to one. It is this condition rather than Condition 3 that is directly used in Theorem 1.

Condition 4: $\sup_{g\in G} |\|g\|_n^2/\|g\|^2 - 1| = o_P(1)$.

Huang (1998a) has established results for checking this condition when $G = G_n$ are arbitrary finite-dimensional linear spaces (see Lemma 4 of the cited paper). In this paper, by restricting attention to spaces built by polynomial splines and their tensor products, we are able to verify Condition 4 under weaker conditions on the growing rate of the dimension of G. Details are given in Section 5.

Theorem 1. *Assume Conditions* 1, 2 *and* 4 *hold. Then, for all* $g^* \in G$,

$$\|\hat{\mu} - \mu^*\|_n^2 + \|\hat{\mu} - \mu^*\|^2 = O_P\left(\{1 + \|g^*\|_\infty^2\}\frac{N_n}{n} + \|g^* - \mu^*\|^2\right).$$

Note that the g^* in this theorem is not specified, so one has the freedom to pick an appropriate g^* to obtain useful result, that is, to make the right-hand side of the displayed equation small. See Corollaries 1 and 2 below for examples on how to pick a g^* in special cases. The proof of Theorem 1, which is a modification of the arguments used in Huang (1998a), will be given in Section 4. Note that the right-hand side involves the L_2 approximation to μ^* by functions in G, which is in contrast to the L_∞ approximation used in Huang (1998a). This allows us to consider broader function classes for the regression function or its ANOVA components, which will be detailed in the next section.

3. Application to spline regression

Now we consider the application of Theorem 1 in some special cases. Suppose that $\mathcal{X}$ is the Cartesian product of compact intervals $\mathcal{X}_1, \ldots, \mathcal{X}_L$, which are supposed to be $[0,1]$ without loss of generality. For $1 \leqslant l \leqslant L$, let G_l be a linear space of splines on $\mathcal{X}_l$ with degree $m \geqslant 0$ and J_n interior knots. Suppose the knots have bounded mesh ratio (that is, the ratios of the difference between consecutive knots are bounded away from zero and infinity uniformly in n). Let G be the tensor product of $G_1, \ldots, G_L$.

To measure the smoothness of a function, we define the Besov space as in DeVore and Popov (1988). Let $1 \leqslant p \leqslant \infty$ and let r be a positive integer. Let

$$\omega_r(f,t)_p = \sup_{|h|\leqslant t} \|\Delta_h^r(f,\cdot)\|_p(\mathcal{X}(rh)), \quad t > 0,$$

denote the modulus of smoothness of order r of $f \in L_p(\mathcal{X})$; here $|h|$ is the Euclidean length of the vector h, Δ_h^r is the rth order difference with step $h \in \mathbb{R}^L$, and $\|\cdot\|_p$ is the L_p norm on the set

$\mathcal{X}(rh) = \{x : x, x + rh \in \mathcal{X}\}$. Let $\alpha > 0$ and $1 \leqslant q \leqslant \infty$. We say that f is in the Besov space $B^{\alpha}_{p,q}$ whenever $f \in L_p(\mathcal{X})$ and

$$\left\{ \int_0^{\infty} (t^{-\alpha} \omega_r(f,t)_p)^q \frac{\mathrm{d}t}{t} \right\}^{1/q} < \infty$$

for any integer $r > \alpha$. (When $q = \infty$, the usual change from integral to sup is made.) In particular, if $\omega_r(\mu,t)_2 \lesssim t^{\alpha}$, $r > \alpha$, then $\mu \in B^{\alpha}_{2,\infty}(\mathcal{X})$.

Corollary 1. *Suppose Conditions* 1–3 *hold. Assume that* $\mu \in B^{\alpha}_{2,\infty}(\mathcal{X})$. *If* $m \geqslant \alpha - 1$ *and* $\lim_n J_n^L \log(J_n)/n = 0$, *then*

$$\|\hat{\mu} - \mu\|_n^2 + \|\hat{\mu} - \mu\|^2 = \mathrm{O_P}\left(\frac{J_n^L}{n} + J_n^{-2\alpha} \right).$$

In particular, if $\alpha > 0$, $m \geqslant \alpha - 1$ *and* $J_n \asymp n^{1/(2\alpha+L)}$, *then*

$$\|\hat{\mu} - \mu\|_n^2 + \|\hat{\mu} - \mu\|^2 = \mathrm{O_P}(n^{-2\alpha/(2\alpha+L)}).$$

Proof. Condition 4 is satisfied by Corollary 3 in Section 5.1. Note $N_n \asymp J_n^L$. Since $\mu \in B^{\alpha}_{2,\infty}(\mathcal{X})$, it follows from Theorem 4.2 of Dahmen et al. (1980) that there is a $g^* \in G$ such that $\|g^* - \mu\| \lesssim J_n^{-\alpha}$; and by Lemma 4.1 of the cited paper, this g^* can be chosen to satisfy $\|g^*\|_{\infty} \leqslant C$. Application of Theorem 1 yields the desired result. □

The rate $n^{-2\alpha/(2\alpha+L)}$ is the optimal rate of convergence (Stone, 1982). Application of Theorem 1 of Huang (1998a) can give the same rate of convergence as in Corollary 1, but the more stringent condition $\lim_n J_n^{2L}/n = 0$ is required there and in particular, $\alpha > L/2$ is required to obtain the optimal rate of convergence. Another limitation of Theorem 1 of Huang (1998a) is that it uses best L_{∞} approximation rate to bound the rate of convergence and thus it is only applicable when μ is in the function class $B^{\alpha}_{\infty,\infty}$ which is smaller than the function class $B^{\alpha}_{2,\infty}$ considered in Corollary 1.

The condition $\lim_n J_n^{2L}/n = 0$ and correspondingly $\alpha > L/2$ is commonly used in the literature on asymptotic analysis of spline-based estimators. See, for example, Chen (1991) and Zhou et al. (1998) among many others. He and Shi (1996, p. 170) gave an explanation why such a condition might be necessary. In Corollary 1, however, we show that this condition is purely technical and can be weakened. The main technical reason is that Condition 4 can be established under weaker conditions by using the special properties of spline spaces.

The weaker condition $\lim_n J_n^L \log(J_n)/n = 0$ is comparable to the typical condition $nh_n^L \to \infty$ required for consistency of local polynomial kernel estimators; here h_n is the bandwidth, which plays a similar role as J_n^{-1}. We think the $\log n$ term cannot be dropped, since the spline method is a global smoothing method and deals with all points in the design space at the same time, while the local method treats one point at a time.

Note that the dependence of the optimal rate of convergence on L reflects the "curse of dimensionality"—the larger the L, the slower the rate of convergence. This motivates introducing structural models to reduce the dimensionality. Next, we consider structural models specified by functional ANOVA decompositions. Our development follows closely Section 3 of Huang (1998a).

As above, suppose $\mathcal{X}$ is the Cartesian product of some compact intervals $\mathcal{X}_1, \ldots, \mathcal{X}_L$, which are assumed to be $[0,1]$ without loss of generality. Let $\mathcal{S}$ be a fixed hierarchical collection of subsets of

$\{1,\dots,L\}$, where *hierarchical* means that if s is a member of $\mathscr{S}$ and r is a subset of s, then r is a member of $\mathscr{S}$. Clearly, if $\mathscr{S}$ is hierarchical, then $\emptyset \in \mathscr{S}$. Let $H_\emptyset$ 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, $l \in s$. Let the model space be given by $H = \{\sum_{s\in\mathscr{S}} h_s : h_s \in H_s\}$. For $s \in \mathscr{S}$, let H_s^0 denote the space of all functions in H_s that are orthogonal (relative to the theoretical inner product) to each function in H_r for every proper subset r of s. Every function $h \in H$ can be written in an essentially unique manner as $h = \sum_{s\in\mathscr{S}} h_s$, $h_s \in H_s^0$, $s \in \mathscr{S}$. (The uniqueness follows from the argument in the proof of Lemma 3.2 of Stone (1994) under Condition 3 of this paper.) We refer to such a decomposition as the ANOVA decomposition of h and h_s as the ANOVA components of h. The component h_s 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.

Note that requiring $\mu \in H$ introduces a particular structure on the regression function, the resulting model is called functional ANOVA model in Huang (1998a). In particular, $\mathscr{S}$ specifies the main effects and interaction terms that are in the model. Let $d = \max\{\#(s), s \in \mathscr{S}\}$. If $d = L$, then all interaction terms are included and we get a saturated model; if $d = 1$, we get an additive model.

For statistical estimation, we construct G to have the same structure as H. Let $G_\emptyset$ denote the space of constant functions on $\mathscr{X}$. Given $1 \leqslant l \leqslant L$, let $G_l \supset G_\emptyset$ denote a linear space of spline functions as used in Corollary 1. Given any nonempty subset $s = \{s_1,\dots,s_k\}$ of $\{1,\dots,L\}$, let G_s be the tensor product of $G_{s_1},\dots,G_{s_k}$. Set $G = \{\sum_{s\in\mathscr{S}} g_s : g_s \in G_s\}$. The dimension N_n of G satisfies $N_n \asymp J_n^d$.

Recall that μ^* is the best approximation μ in H. The next result is concerned about the rate of convergence of $\hat{\mu}$ to μ^* when G takes the form specified in the previous paragraph. Write the ANOVA decomposition of μ^* as $\mu^* = \sum_{s\in\mathscr{S}} \mu_s^*$.

Corollary 2. *Suppose Conditions* 1–3 *hold. Assume that* $\mu_s^* \in B_{2,\infty}^\alpha(\mathscr{X}_s)$, $s \in \mathscr{S}$. *If* $m \geqslant \alpha - 1$ *and* $\lim_n J_n^{d'} \log(J_n)/n = 0$, *where* $d' = \max\{\#(s \cup s'), s, s' \in \mathscr{S}\}$, *then*

$$\|\hat{\mu} - \mu^*\|_n^2 + \|\hat{\mu} - \mu^*\|^2 = \mathrm{O_P}\left(\frac{J_n^d}{n} + J_n^{-2\alpha}\right).$$

In particular, if $\alpha > (d' - d)/2$, $m \geqslant \alpha - 1$ *and* $J_n \asymp n^{1/(2\alpha+d)}$, *then*

$$\|\hat{\mu} - \mu^*\|_n^2 + \|\hat{\mu} - \mu^*\|^2 = \mathrm{O_P}(n^{-2\alpha/(2\alpha+d)}).$$

Proof. Condition 4 is satisfied by Lemma 3 in Section 5.2. Note $N_n \asymp J_n^d$. Since $\mu_s^* \in B_{2,\infty}^\alpha(\mathscr{X})$, it follows from Theorem 4.2 of Dahmen et al. (1980) that there is a $g_s^* \in G_s$ such that $\|g_s^* - \mu_s^*\| \lesssim J_n^{-\alpha}$; and by Lemma 4.1 of the cited paper, this g_s^* can be chosen to satisfy $\|g_s^*\|_\infty \leqslant C$ for some constant C. Let $g^* = \sum_{s\in\mathscr{S}} g_s^*$. Application of Theorem 1 yields the desired result. □

Application of Theorem 5 of Huang (1998a) can give the same rate of convergence as in Corollary 2, but it requires that $\lim_n J_n^{2d}/n = 0$, which is usually more stringent than $\lim_n J_n^{d'} \log(J_n)/n = 0$ when $d' < 2d$ (note that the log term is ignorable in the comparison when J_n is some power of n). Suppose, for example, that $L = 4$. If $\mathscr{S} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{1,3\}\}$, then $d' = 3 < 2d = 4$.

Another requirement for application of Theorem 5 of Huang (1998a) is that the model is correctly specified, that is, $\mu^* = \mu$, which is relaxed in Corollary 2 here.

Hansen (1994) has obtained rate of convergence under a similar condition on J_n as in Corollary 2. His result applies only to $\mu_s^* \in B^p_{\infty,\infty}(\mathscr{X}_s)$, a function class smaller than $B^\alpha_{2,\infty}$, since the best L_∞ approximation is used to characterize the rate of convergence.

4. Proof of Theorem 1

Let Q denote the empirical orthogonal projection onto G, P the theoretical orthogonal projection onto G, and P^* the theoretical orthogonal projection onto H. Let $Y = Y(\cdot)$ denote a function interpolating the data points. We observe that $\hat{\mu} = QY$ and $\mu^* = P^*\mu$. Set $\tilde{\mu} = Q\mu$ and $\bar{\mu} = P\mu$. Consider the following decomposition:

$$\hat{\mu} - \mu^* = (\hat{\mu} - \tilde{\mu}) + (\tilde{\mu} - \bar{\mu}) + (\bar{\mu} - \mu^*), \tag{1}$$

where $\hat{\mu} - \tilde{\mu}$, $\tilde{\mu} - \bar{\mu}$ and $\bar{\mu} - \mu^*$ are referred to as the variance component, the estimation bias and the approximation error, respectively. We can treat these three components separately in studying the rate of convergence of $\hat{\mu} - \mu^*$.

Variance component: Arguing as in Huang, we obtain that $\|\hat{\mu} - \tilde{\mu}\|_n^2 = \mathrm{O_P}(N_n/n)$ and thus by Condition 4 we get that $\|\hat{\mu} - \tilde{\mu}\|^2 = \mathrm{O_P}(N_n/n)$.

The treatment of estimation bias and approximation error in Huang (1998a) can be simplified slightly as follows.

Estimation bias: Denote $\eta_n = \|g^* - \mu^*\|$. Since $\bar{\mu} = P\mu^*$, $\|g^* - \bar{\mu}\| \leqslant \|g^* - \mu^*\| = \eta_n$. Observe that

$$\|\tilde{\mu} - \bar{\mu}\|_n = \sup_{g\in G} \frac{|\langle \tilde{\mu} - \bar{\mu}, g\rangle_n|}{\|g\|_n} = \sup_{g\in G} \frac{|\langle \mu - \bar{\mu}, g\rangle_n - \langle \mu - \bar{\mu}, g\rangle|}{\|g\|_n}.$$

Hence, by Condition 4,

$$\begin{aligned}\|\tilde{\mu} - \bar{\mu}\|_n &\leqslant \sup_{g\in G} \frac{|\langle \mu - g^*, g\rangle_n - \langle \mu - g^*, g\rangle|}{\|g\|_n} + \sup_{g\in G} \frac{|\langle g^* - \bar{\mu}, g\rangle_n - \langle g^* - \bar{\mu}, g\rangle|}{\|g^* - \bar{\mu}\|\|g\|} \frac{\|g\|}{\|g\|_n} \|g^* - \bar{\mu}\| \\ &= \mathrm{O_P}\left(\|\mu - g^*\|_\infty \sqrt{\frac{N_n}{n}}\right) + \mathrm{o_P}(1)\mathrm{O_P}(\eta_n) = \mathrm{O_P}\left(\{1 + \|g^*\|_\infty\}\sqrt{\frac{N_n}{n}}\right) + \mathrm{o_P}(\eta_n).\end{aligned}$$

A bound on $\|\tilde{\mu} - \bar{\mu}\|$ can be obtained by using Condition 4 again.

Approximation error: Since $\bar{\mu} = P\mu^*$, $\|\bar{\mu} - \mu^*\| \leqslant \|g^* - \mu^*\|$. Note that $E(\|\bar{\mu} - \mu^*\|_n^2) = \|\bar{\mu} - \mu^*\|^2$. It follows from the Markov inequality that $\|\bar{\mu} - \mu^*\|_n = \mathrm{O_P}(\|g^* - \mu^*\|)$.

The difference between the arguments used here and those in Huang (1998a) is that the L_∞ approximation rate is avoided in handling the estimation bias and the approximation error.

5. Equivalence of the empirical and theoretical inner products

In this section we show that, under the spaces of fits built by polynomial splines, the empirical inner product is uniformly close to the theoretical inner product. Such results are useful in checking

Condition 4. They are presented in more general form than what is needed in Section 3. These results are also of independent interest. Corollary 3 is cited in Huang (2003).

The result in Section 5.1 extends those in Buja (1994) to a more general setting that allows multivariate splines over triangulations. The result in Section 5.2 can also be established by combining the arguments of Buja (1994) and Stone (1994), but the argument used here is much simpler.

5.1. Saturated model

In this subsection, we provide sufficient conditions for Condition 4 when G is a polynomial spline on a domain $\mathcal{X}$. Here, a polynomial spline is referred to broadly as any possibly smooth, piecewise polynomial function. Let $\mathcal{X}$ be a closed, bounded subset of $\mathbb{R}^d$. Suppose that $\mathcal{X}$ is polyhedral, that is, representable by a finite partition into non-degenerating simplices. Consider a sequence of partitions $\Delta_n = \{\delta : \delta \subset \mathcal{X}\}$ of $\mathcal{X}$. We require that each $\delta \in \Delta_n$ be polyhedral. In particular, an element of Δ_n can be an interval, a two-dimensional triangle or rectangle, or a high-dimensional simplex or hyper-rectangle for $d = 1$, $d = 2$ and $d \geqslant 3$, respectively. As n grows, the elements in Δ_n are required to be shrinking in size and increasing in number. Consider a space G of piecewise polynomials (polynomial splines) over the partition Δ_n of $\mathcal{X}$. Let m be a fixed integer. Every $g \in G$ is a polynomial of degree m or less when restricted to each $\delta \in \Delta_n$. This setup is very general, containing as special cases univariate splines, tensor product splines, and bivariate or multivariate splines on triangulations.

Condition 5: (i) The ratio of the sizes of inscribed and circumscribed balls of each $\delta \in \Delta_n$ is bounded away from zero (uniformly in n). (ii) the ratios of diameters h_δ of $\delta \in \Delta_n$ are bounded above and below, that is, $h_\delta \asymp h$ for some $h = h_n$.

For $\delta \in \Delta_n$, set $G_\delta = \{g(x)|_{x\in\delta}, g \in G\}$ and $k_\delta = \dim(G_\delta)$. Let $I_\delta(x)$ denote the indicator function that takes value 1 when $x \in \delta$ and 0 otherwise. Define an inner product on G_δ by $\langle g_1, g_2\rangle_\delta = E[g_1(X)g_2(X)I_\delta(X)]$ and the corresponding norm is given by $\|g\|_\delta^2 = E[g^2(X)I_\delta(X)]$. Note that $\|g\|_\delta$ is also well-defined for $g \in G$. Define the supreme norm as $\|g\|_{\infty,\delta} = \sup_{x\in\delta} |g(x)|$. Set

$$A_{n,\delta} = \sup_{g\in G} \frac{\|g\|_{\infty,\delta}}{\|g\|_\delta}.$$

Lemma 1. *Under Condition* 1,

$$P\left\{\sup_{f,g\in G} \frac{|(E_n - E)(fg)|}{\|f\|\|g\|} > t\right\} \leqslant P\left\{\sup_\delta \sup_{f,g\in G} \frac{|(E_n - E)(fgI_\delta)|}{\|f\|_\delta\|g\|_\delta} > t\right\}$$
$$\leqslant 2\sum_\delta k_\delta^2 \exp\left\{-\frac{1}{2}\frac{(nt)^2}{nA_{n,\delta}^2 + 2A_{n,\delta}^2 nt/3}\right\}.$$

Proof. If $|(E_n - E)(fgI_\delta)| \leqslant t\|f\|_\delta\|g\|_\delta$ for $\delta \in \Delta_n$ and $f, g \in G$, then

$$|(E_n - E)(fg)| \leqslant \sum_\delta |(E_n - E)(fgI_\delta)| \leqslant \sum_\delta t\|f\|_\delta\|g\|_\delta \leqslant t\|f\|\|g\|,$$

where for the last "$\leqslant$", we use the Cauchy–Schwarz inequality and $\|f\|^2 = \sum_\delta \|f\|_\delta^2$. Thus,

$$\sup_{f,g\in G} \frac{|(E_n - E)(fg)|}{\|f\|\|g\|} \leqslant \sup_\delta \sup_{f,g\in G} \frac{|(E_n - E)(fgI_\delta)|}{\|f\|_\delta \|g\|_\delta},$$

which shows the validity of the first "$\leqslant$" in the lemma.

Let $\phi_j, j = 1,\dots,k_\delta$ be an orthonormal basis of G_δ relative to the inner product $\langle\cdot,\cdot\rangle_\delta$. Note that $E[\{\phi_j(X)\phi_{j'}(X)I_\delta(X)\}^2] \leqslant A_{n,\delta}^2$ and $|\phi_j(x)\phi_{j'}(x)| \leqslant A_{n,\delta}^2$ for $x \in \delta$. By Bernstein's inequality (Pollard, 1984, p. 193),

$$P(|(E_n - E)(\phi_j\phi_{j'})| > t) \leqslant 2\exp\left\{-\frac{1}{2}\frac{(nt)^2}{nA_{n,\delta}^2 + 2A_{n,\delta}^2 nt/3}\right\}.$$

For $f, g \in G_\delta$, write $f = \sum_j \beta_j\phi_j$ and $g = \sum_j \gamma_j\phi_j$. Then $\|f\|_\delta^2 = \sum_j \beta_j^2$ and $\|g\|_\delta^2 = \sum_j \gamma_j^2$. If $|(E_n - E)(\phi_j\phi_{j'})| \leqslant t/k_\delta$, $j, j' = 1,\dots,k_\delta$, then

$$|(E_n - E)(fg)| = \left|\sum_j\sum_{j'} \beta_j\gamma_{j'}(E_n - E)\phi_j\phi_{j'}\right| \leqslant \sum_j\sum_{j'} |\beta_j\|\gamma_{j'}|\,\frac{t}{k_\delta} \leqslant t\|f\|_\delta\|g\|_\delta.$$

Consequently,

$$P(|(E_n - E)(fgI_\delta)| > t\|f\|_\delta\|g\|_\delta \text{ for all } f, g \in G_\delta) \leqslant 2k_\delta^2 \exp\left\{-\frac{1}{2}\frac{(nt)^2}{nA_{n,\delta}^2 + 2A_{n,\delta}^2 nt/3}\right\} \qquad (2)$$

and the second "$\leqslant$" in the lemma follows. □

Lemma 2. *Under Conditions* 3 *and* 5, *for some* A_n *with* $A_n \asymp h^{-d/2}$, $A_{n,\delta} \leqslant A_n$ *for all* $\delta \in \Delta_n$.

Proof. Fix $\delta \in \Delta_n$. Note that each $g \in G_\delta$ is a polynomial. In the proof of this lemma, we also use g to denote its natural extension to $\mathbb{R}^d$ (note that the values of a polynomial on δ determines its values on $\mathbb{R}^d$). We can find two balls B_1 and B_2 with the same center and satisfy $B_1 \subset \delta \subset B_2$. Without loss of generality, we assume the common center of B_1 and B_2 is the origin. Denote the radii of B_1 and B_2 as r_1 and r_2. By Condition 5, we can make r_1 and r_2 to be universal (i.e., applicable to all $\delta \in \Delta_n$) and satisfy $r_1 \asymp h \asymp r_2$. By Condition 3,

$$\|g\|_\delta^2 \geqslant M\int_\delta g^2(x)\,\mathrm{d}x \geqslant M\int_{B_1} g^2(x)\,\mathrm{d}x.$$

Let B_0 denote the unit ball on $\mathbb{R}^d$. By a change of variable, $\|g\|_\delta^2 \geqslant Mr_1^d\int_{B_0} g^2(r_1 y)\,\mathrm{d}y$. On the other hand, $\|g\|_{\infty,\delta} \leqslant \|g\|_{\infty,B_2} = \sup_{y\in B_0}|g(r_2 y)|$. Consequently,

$$\frac{\|g\|_{\infty,\delta}}{\|g\|_\delta} \leqslant \frac{1}{\sqrt{Mr_1^d}}\,\frac{\sup_{y\in B_0}|g(r_2 y)|}{\{\int_{B_0} g^2(r_1 y)\,\mathrm{d}y\}^{1/2}}.$$

Since $\sup_{y\in B_0}|g(r_2 y)|$ and $\{\int_{B_0} g^2(r_1 y)\,\mathrm{d}y\}^{1/2}$ are just two norms on a finite-dimensional linear space (i.e., the space of polynomials on $\mathbb{R}^d$ of certain degree), they are equivalent. The desired result then follows. □

Corollary 3. *Under Conditions* 1, 3 *and* 5,

$$\sup_{f,g\in G}\frac{|(E_n-E)(fg)|}{\|f\|\|g\|}\leqslant\sup_{\delta}\sup_{f,g\in G}\frac{|(E_n-E)(fgI_\delta)|}{\|f\|_\delta\|g\|_\delta}=\mathrm{O_P}\left(\sqrt{\frac{\log(h^{-1})}{nh^d}}\right).$$

In particular, if $\lim_n h=0$ *and* $\lim_n \log(h^{-1})/(nh^d)=0$, *then* $\sup_{g\in G}|\|g\|_n^2/\|g\|^2-1|=\mathrm{o_P}(1)$.

Proof. We only need prove the result when $\lim_n \log(h^{-1})/(nh^d)=0$, since otherwise the result is trivially true. Under Condition 5, $\#\{\delta:\delta\in\Delta_n\}\lesssim h^{-d}$. Let $c_1,c_2,\ldots,$ denote generic constants. By Lemmas 1 and 2,

$$P\left\{\sup_{\delta}\sup_{f,g\in G}\frac{|(E_n-E)(fgI_\delta)|}{\|f\|_\delta\|g\|_\delta}>t\right\}\lesssim h^{-d}\exp\left\{-\frac{(nt)^2}{c_1nh^{-d}+c_2nth^{-d}}\right\}$$

(note k_δ is bounded). Taking $t=c_3\{\log(h^{-1})/(nh^d)\}^{1/2}$ for some large constant c_3, the right-hand side of the above inequality is bounded above by $\exp\{-c_4\log(h^{-1})\}\to 0$. □

5.2. ANOVA model

Now, we consider the fitting space G for the ANOVA model on a tensor product domain $\mathscr{X}$ as in Corollary 2. Let $h=h_n$ denote the smallest distance between two consecutive knots for the splines that build G. Note that $h_n\asymp J_n^{-1}$ where J_n denote the number of knots for each space G_l.

Lemma 3. *Under Conditions* 1, 3 *and* 5, *if* $\lim_n h=0$ *and* $\max_{s,s'\in\mathscr{S}}\log(h^{-1})/(nh^{\#(s\cup s')})\to 0$, *then* $\sup_{g\in G}|\|g\|_n^2/\|g\|^2-1|=\mathrm{o_P}(1)$.

Proof. Write $f=\sum_s f_s$ and $g=\sum_s g_s$ where $f_s,g_s\perp G_{s'}$ for $s'\subset s$, $s'\neq s$. If $|\langle f_s,g_{s'}\rangle_n-\langle f_s,g_{s'}\rangle|\leqslant t\|f_s\|\|g_{s'}\|$ for all $f_s\in G_s$, $g_{s'}\in G_{s'}$, $s,s'\in\mathscr{S}$, then

$$\begin{aligned}|\langle f,g\rangle_n-\langle f,g\rangle|&\leqslant\sum_s\sum_{s'}|\langle f_s,g_{s'}\rangle_n-\langle f_s,g_{s'}\rangle|\\&\leqslant t\sum_s\sum_{s'}\|f_s\|\|g_{s'}\|\leqslant t\|f\|\|g\|\{\#(\mathscr{S})\}.\end{aligned}$$

Hence, by Lemma 1,

$$\begin{aligned}P\left\{\sup_{f,g\in G}\frac{|\langle f,g\rangle_n-\langle f,g\rangle|}{\|f\|\|g\|}>\#(\mathscr{S})t\right\}&\leqslant\sum_s\sum_{s'}P\left\{\sup_{f_s\in G_s,g_s'\in G_{s'}}\frac{|\langle f_s,g_{s'}\rangle_n-\langle f_s,g_{s'}\rangle|}{\|f_s\|\|g_{s'}\|}>t\right\}\\&\lesssim\sum_s\sum_{s'}\exp\left\{-\frac{n}{ch^{-\#(s\cup s')}}\right\}h^{-\#(s\cup s')}\to 0.\end{aligned}$$

The desired result follows. □

Acknowledgements

I would like to thank Chuck Stone and a referee for helpful comments. This work is partially supported by National Science Foundation Grant DMS-0204556.

References

Buja, A., 1994. Discussion to "The use of polynomial splines and their tensor products in multivariate function estimation" by C.J. Stone. Ann. Statist. 22, 118–171.

Chen, H., 1991. Polynomial splines and nonparametric regression. J. Nonparametric Statist. 1, 143–156.

Dahmen, W., DeVore, R., Scherer, K., 1980. Multi-dimensional spline approximation. SIAM J. Numer. Anal. 17, 380–402.

DeVore, R.A., Popov, V., 1988. Interpolation of Besov spaces. Trans. Amer. Math. Soc. 305, 397–414.

Hansen, M., 1994. Extended linear models, multivariate splines, and ANOVA. Ph.D. Dissertation, University of California, Berkeley.

He, X., Shi, P., 1996. Bivariate tensor-product B-splines in a partly linear model. J. Multivariate Anal. 58, 162–181.

Huang, J.Z., 1998a. Projection estimation in multiple regression with application to functional ANOVA models. Ann. Statist. 26, 242–272.

Huang, J.Z., 1998b. Functional ANOVA models for generalized regression. J. Multivariate Anal. 67, 49–71.

Huang, J.Z., 2001. Concave extended linear modeling: a theoretical synthesis. Statistica Sinica 11, 173–197.

Huang, J.Z., 2003. Local asymptotics for polynomial spline regression. Ann. Statist. 31, 1600–1635.

Huang, J.Z., Stone, C.J., 1998. The L_2 rate of convergence for event history regression with time-dependent covariates. Scand. J. Statist. 25, 603–620.

Huang, J.Z., Kooperberg, C., Stone, C.J., Truong, Y.K., 2000. Functional ANOVA modeling for proportional hazards regression. Ann. Statist. 28, 960–999.

Huang, J.Z., Stone, C.J., 2002a. Free knot splines in concave extended linear modeling. J. Statist. Plann. Inference 108, 219–253.

Huang, J.Z., Stone, C.J., 2002b. In: Dension, D.D., Hansen, M.H., Holmes, C.C., Malick, B., Yu, B. (Eds.), Nonlinear Estimation and Classification. Springer, New York, pp. 213–234.

Huang, J.Z., Stone, C.J., 2003. Statistical modeling of diffusion processes with free knot splines. J. Statist. Plann. Inference 116, 451–474.

Stone, C.J., 1982. Optimal global rates of convergence for nonparametric regression. Ann. Statist. 8, 1348–1360.

Stone, C.J., 1994. The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). Ann. Statist. 22, 118–171.

Stone, C.J., Hansen, M., Kooperberg, C., Truong, Y.K., 1997. Polynomial splines and their tensor products in extended linear modeling (with discussion). Ann. Statist. 25, 1371–1470.

Zhou, S., Shen, X., Wolfe, D.A., 1998. Local asymptotics for regression splines and confidence regions. Ann. Statist. 26, 1760–1782.