Statistica Sinica **5**(1995), 303-318

GLOBAL CONVERGENCE RATES OF B-SPLINE M-ESTIMATORS IN NONPARAMETRIC REGRESSION

Peide Shi and Guoying Li

Academia Sinica

Abstract. To compensate for lack of robustness in using regression splines via the least squares principle, a robust data smoothing procedure is proposed for obtaining a robust regression spline estimator of an unknown regression function, g_0, of a one-dimensional measurement variable. This robust regression spline estimator is computed by using the usual M-type iteration procedures proposed for linear models. A simulation study is carried out and numerical examples are given to illustrate the utility of the proposed method. Assume that g_0 is smoothed up to order $r > 1/2$ and denote the derivative of g_0 of order l by $g_0^{(l)}$. Let $\hat{g}_n^{(l)}$ denote an M-type regression spline estimator of $g_0^{(l)}$ based on a training sample of size n. Under appropriate regularity conditions, it is shown that the proposed estimator, $\hat{g}_n^{(l)}$, achieves the optimal rate, $n^{-(r-l)/(2r+1)}$ $(0 \le l < r)$, of convergence of estimators for nonparametric regression when the spline knots are deterministically given.

Key words and phrases: B-spline function, M-estimator, nonparametric regression, optimal rate of convergence.

1. Introduction

Nonparametric regression analysis is an increasingly popular tool for the purpose of data smoothing. Unfortunately, many of the commonly used estimators of nonparametric regression functions including kernel estimators (Gasser and Müller (1979)), smoothing spline estimators (Eubank (1988)), regression spline estimators (Friedman and Silverman (1989) and Friedman (1991)) are nonrobust. To compensate for this defect several authors (Lenth (1977), Huber (1979), Härdle (1984), Härdle and Gasser (1984), Cox (1983), and Cunningham et al. (1991)) have proposed parallels of M-estimators for fitting unknown regression functions.

Suppose that (T_i, Y_i), $1 \le i \le n$, are i.i.d. observations of a two-dimensional random vector (T, Y) with

$$Y_i = g_0(T_i) + u_i, \qquad 1 \le i \le n, \tag{1.1}$$

where the u_i's are random errors which are independent of $T_1, \ldots, T_n$ and g_0 is some unknown regression function. Let λ be a knot set and $B_\lambda(\cdot)$ be a N-vector

of splines associated with λ. The regression spline estimator of g_0 is defined as $\widehat{g}_n(t) = \widehat{g}_{n\lambda}(t) = B_\lambda(t)'\widehat{\theta}$ (Friedman and Silverman (1989)), where $\widehat{\theta}$ is the minimizer of

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{N} B_{j\lambda}(T_i)\theta_j)^2 \tag{1.2}$$

and λ is the trade-off between smoothness and flexibility.

It is well known that outliers in the response values can deteriorate the performance of regression spline estimators based on the least squares (LS) principle. To compensate for this drawback, we robustify criterion (1.2) by minimizing

$$\sum_{i=1}^{n} \rho(Y_i - \sum_{j=1}^{N} B_{j\lambda}(T_i)\theta_j), \tag{1.3}$$

where ρ is a function chosen suitably and $B_{1\lambda}(t), \ldots, B_{N\lambda}(t)$ are B-spline basis functions. An estimator of g_0 obtained by minimizing (1.3) is called an M-type regression spline estimator.

The spline functions used in (1.3) are B-splines of order $m+1$. The special case of $m = 1$ corresponds to piecewise linear smoothing used by Friedman and Silverman (1989) where truncated power functions are taken as an equivalent basis of the spline space. Since the B-spline method is efficient in digital computation and functional approximation, it is widely used for curve and surface fitting. For example, deBoor (1978), Schumaker (1981) and Su and Liu (1989) have convincingly demonstrated the utility of B-splines in curve and surface approximation in nonstochastic settings.

Just as with the selection of the bandwidth of kernel estimators and the penalty parameters of smoothing spline estimators, the choice of the regression spline knots is important. Therefore, we will give a stepwise forward/backward knot placement and deletion strategy via the generalized cross validation criterion (GCV). Related selection schemes of regression splines have been investigated by Stone and Koo (1985), Atilgan (1988), Friedman and Silverman (1989), and Shi (1993).

When the knot set of the B-spline basis is given, the minimization of (1.3) can be performed efficiently with the usual M-type iteration procedures proposed for linear models (Huber (1981)). Three other papers concerning regression splines in related settings are Lenth (1977), Agarwal and Studden (1980) and Friedman (1991).

In this paper, we prove, under some regularity conditions, that the M-type regression spline estimator and the derivatives of it all achieve the optimal rates of convergence of estimators for nonparametric regression when the spline knots are deterministically given. These results are shown in Section 3. The computational

aspects of the estimator are discussed in Section 2. In Section 4, examples are given to illustrate the utility of the proposed methodology and a Monte Carlo study is carried out. The proofs of the main results are given in Section 5. Our simulation results show that when the random errors are normally distributed the M-estimators are as good as LS estimators; however, when the random errors are drawn from a symmetrically contaminated normal distribution the M-estimators are superior to LS estimators; and when the random errors are distributed as Cauchy distribution the M-estimators seem acceptable but the LS estimators behave poorly.

The asymptotics of M-estimators of nonparametric regression functions has been investigated by several authors, e.g. Cox (1983), Härdle and Luckhaus (1984), Härdle (1984), Härdle and Tsybakov (1988), Härdle (1990), and Cunningham et al. (1991).

2. Computation of the Estimator

First, some notation is needed. Let $m \geq 0$ and $k_n > 0$ be integers, $N = k_n + m$, and $t_0, t_1, \ldots, t_{k_n}$ be a D_0-quasi-uniform sequence of partitions of [0, 1] (Schumaker (1981, p.216))

$$\max_{1\leq i\leq k_n}(t_i - t_{i-1})/\min_{1\leq k\leq k_n}(t_k - t_{k-1}) \leq D_0$$

uniformly in n, where $D_0 > 0$ is a constant. Let $\lambda = \{t_i, 0 \leq i \leq k_n\}$ and $B_\lambda(t) \widehat{=} (B_{1\lambda}(t), \ldots, B_{N\lambda}(t))'$ be a vector of normalized B-splines (of order $m+1$) associated with an extended partition of $[0, 1]$ determined by $\{t_i\}_1^{k_n}$ (Schumaker (1981, p.224)). Note that the B-splines $B_{1\lambda}(t), \ldots, B_{N\lambda}(t)$ are $m-1$ times continuously differentiable in (0, 1). In the sequel, $B_\lambda(\cdot), B_{1\lambda}(\cdot), \ldots, B_{N\lambda}(\cdot)$ will be abbreviated as $B(\cdot), B_1(\cdot), \ldots, B_N(\cdot)$ respectively when it is necessary.

From Corollary 6.21 in Schumaker (1981), $g_0(t)$ can be reasonably approximated with a B-spline function $B(t)'\theta$. Therefore, we define

$$\widehat{g}_n(t) = \widehat{g}_{n\lambda}(t) \widehat{=} B_\lambda(t)'\widehat{\theta}$$

as the M-type regression spline estimator of $g_0(t)$, where $\widehat{\theta}$ minimizes

$$\sum_{i=1}^{n} \rho(Y_i - B_\lambda(T_i)'\theta) \tag{2.1}$$

or satisfies

$$\sum_{i=1}^{n} \Psi(Y_i - B_\lambda(T_i)'\widehat{\theta})B(T_i) = 0, \tag{2.2}$$

where $\Psi(s) = \rho'(s)$.

For a given knot set and an initial vector $\theta^{(1)}$, we adopt Huber's Iteration Procedure (see Huber (1981, p.182)) with unit scale

$$\theta^{(l+1)} = \theta^{(l)} + \left(\sum_{j=1}^{n} B(T_j)B(T_j)'\right)^{-1} \sum_{i=1}^{n} \Psi(Y_i - B(T_i)'\theta^{(l)})B(T_i) \tag{2.3}$$

for $l = 1, 2, \ldots$. If at some step l_0, $|\theta^{(l_0+1)} - \theta^{(l_0)}| < 10^{-3}$, then the above iteration procedure is terminated.

As in the linear model case, an alternative to procedure (2.3) is a reweighted least square iteration procedure based on the equivalence between (2.1) and

$$\sum_{i=1}^{n} w_i B(T_i)B(T_i)'\widehat{\theta} = \sum_{i=1}^{n} w_i Y_i B(T_i),$$

where $w_i = \Psi\left(Y_i - B(T_i)'\widehat{\theta}\right)/(Y_i - B(T_i)'\widehat{\theta})$ (Huber (1981, p.184)). But in this way we incur a little more computational expense in each iteration.

To determine the spline knots, one may choose the knot number k_n and use equispaced knots. Another way, which is more reasonable in practical applications, is to select spline knots $0 = t_0 < t_1 < \cdots < t_{k_n} = 1$ with a data driven method. For simplicity, we select the knots and the knot number by minimizing the well known GCV criterion $\mathrm{GCV}(\lambda) = n^{-1}\sum_{i=1}^{n}(Y_i - \widehat{g}_{n\lambda}(T_i))^2/(1 - (k_n + m)/n)^2$. Alternative criteria are cross validation (CV), modified GCV (Friedman and Silverman (1989)), AIC (Akaike (1973)), the generalized version of the corrected Akaike information criterion (GAICC) (see Shi (1993))

$$\mathrm{GAICC}(\lambda) = \widehat{\sigma}_n^2 \exp\left(\frac{2N}{n} + \frac{2(N+1)(N(1-c)+2)}{n(n-N(1-c)-2)}\right),$$

where $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \widehat{g}_{n\lambda}(T_i))^2$ and c is a constant in $(0, 1)$ (e.g. $c = 7/8$), and the M-type Akaike information criterion (MAIC) (Shi (1992)) $\mathrm{MAIC}(\lambda) = \sum_{i=1}^{n}\rho(Y_i - \widehat{g}_{n\lambda}(T_i)) + 2(k_n + m)$, which is the robustified AIC.

To implement our estimating method, we adopt a forward/backward stepwise knot placement and deletion strategy similar to that of Friedman and Silverman (1989). First, the knot t_1 is placed at the position for which the following equation is satisfied:

$$\mathrm{GCV}(\{t_1\}) = \inf_{s\in(0,1)} \mathrm{GCV}(\{s\}).$$

Suppose that $t_1, \ldots, t_{k-1}$ have already been found, $\lambda_1 = \{t_1, \ldots, t_{k-1}\}$ and $\lambda_1 \cup \{t_k\} = \lambda$. The additional knot t_k is placed at the position satisfying

$$\mathrm{GCV}(\{\lambda\}) = \inf_{s\in(0,1)} \mathrm{GCV}(\lambda_1 \cup \{s\}) \quad \text{and} \quad \mathrm{GCV}(\lambda) < \mathrm{GCV}(\lambda_1). \tag{2.4}$$

If there is no such point that satisfies (2.4), the knot placement process is terminated. The current knot set λ selected is taken to be the input of the backward stepwise deletion strategy. Each of its elements is in turn deleted and the corresponding leave-one-knot-out model is fitted. The fit with the smallest GCV value is found and the corresponding knot is permanently deleted unless this fit results in a significantly improved GCV. The above knot deletion procedure is repeated for the knots left. Sometimes, no selected knots are deleted or left in the final knot set. Generally, more knots selected corresponds to fitting a more complicated curve unless the signal-to-noise ratio (standard deviation of the function divided by the standard deviation of the noise) is too low.

3. Asymptotic Theory

In this section, we examine the large sample properties of $\widehat{g}_{n\lambda}$ when the knot set is deterministically given. Throughout, it will be assumed that the function $\rho(\cdot)$ is continuously differentiable everywhere.

Let $\gamma \in (0, 1]$ be a number such that $m + \gamma > 1/2$ and let $M_0 \in (0, \infty)$. Let $\mathcal{H}$ stand for the collection of functions on $[0, 1]$ such that, for every $h \in \mathcal{H}$, the mth derivative of h, denoted by $h^{(m)}$, exists and satisfies a Hölder condition of order γ :

$$|h^{(m)}(t) - h^{(m)}(t')| \le M_0 |t - t'|^{\gamma}, \qquad \text{for } 0 \le t,\ t' \le 1.$$

The following four conditions are sufficient for the statement of the theoretical results.

Condition 1. The distribution of T is absolutely continuous with density f and there are two constants b and B such that $0 < b \le f(t) \le B < \infty$ for all $t \in [0, 1]$.

Condition 2. $g_0 \in \mathcal{H}$.

Condition 3. $\Psi(\cdot) = \rho'(\cdot)$ is continuously nondecreasing on R, $E\,\Psi(u_1) = 0$ and $E\,\Psi^2(u_1) = v_0 < \infty$, where u_1 is a random error in (1.1).

Condition 4. There exist six positive constants c_1, c_2, c_3, d_1, d_2, and d_3 ($d_3 \ge d_1$) such that

$$\begin{aligned} &q \mathrel{\widehat{=}} \mathcal{P}(|u_1| \le c_1) > 0, \\ &D(s, t) \ge d_1 \text{ when } |s| \le c_1 \text{ and } |t| \le c_2, \text{ and} \\ &D(s, t) \le d_2 \text{ when } |t| \le c_3, \end{aligned}$$

where

$$D(s, t) = \begin{cases} (\Psi(s + t) - \Psi(s))/t, & t \ne 0, \\ d_3, & t = 0. \end{cases}$$

Remark 3.1. It is well known that if the function $\rho(\cdot)$ is convex and continuously

differentiable everywhere with derivative $\Psi(\cdot)$, then (2.1) and (2.2) are equivalent. By Theorem 27.2 of Rockafellar (1970), the convexity of $\rho(\cdot)$ and the equivalence between (2.1) and (2.2), if $\lim_{t\to\infty}\rho(t) = \infty$, we can easily conclude that the solution set of (2.2) is non-empty for each given sample $\{(T_1, Y_1), \ldots, (T_n, Y_n)\}$. From Conditions 3 and 4, it follows that $\lim_{t\to\infty}\rho(t) = \infty$.

Let $\|h\|_{\mathcal{L}^2}$ denote the $\mathcal{L}^2$ norm, defined by $\|h\|^2_{\mathcal{L}^2} \widehat{=} \int_0^1 h^2(t)f(t)dt$. Let $|\cdot|$ denote either the Euclidean norm of a vector or the absolute value of a real number according to the context. For positive numbers a_n and b_n, $n \geq 1$, let $a_n \sim b_n$ denote that a_n/b_n is bounded away from zero and infinity.

Then we have the following theorem.

Theorem 1. *Suppose that $\Psi(\cdot) = \rho'(\cdot)$, $\widehat{\theta}$ is a solution of (2.2), $\widehat{g}_n(t) = B(t)'\widehat{\theta}$ is the regression spline M-estimator of $g_0(t)$ and $k_n \sim n^{1/[2(m+\gamma)+1]}$. If Conditions $1-4$ are all satisfied, then for $l = 0, 1, \ldots, m$*

$$\|\widehat{g}_n^{(l)} - g_0^{(l)}\|_{\mathcal{L}^2} = O_P(n^{-(m+\gamma-l)/[2(m+\gamma)+1]}).$$

Remark 3.2. According to Stone (1980, 1982), the convergence rates of $\widehat{g}_n$ are the optimal global convergence rates of estimators for nonparametric regression.

Remark 3.3. Our proof of Theorem 1 with the quasi-uniform knots can not be extended to the case of data-dependent knots described in Section 2 for technique reasons.

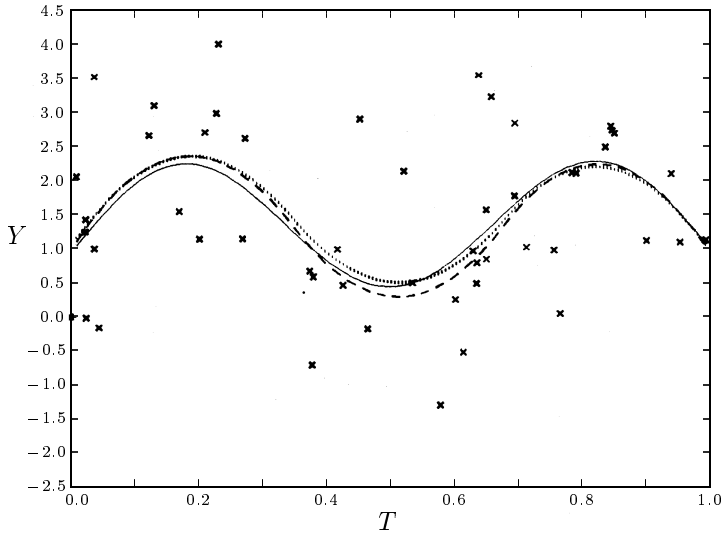

Figure 1. The example of simulated data with sample size $n = 50$. Error distribution is $N(0, 1)$. The solid line is the true curve; the dash line is the regression spline M-fit; the closely spaced dot line is the regression spline LS-fit.

4. Numerical Results

We have presented the procedure, which will be illustrated numerically, for estimating the unknown regression function. The measure for goodness of fit of the estimator $\widehat{g}_n$ of g_0 is called MSE and defined by $\mathrm{MSE}(\widehat{g}_n, g_0) = n^{-1}\sum_1^n(\widehat{g}_n(T_i) - g_0(T_i))^2$.

4.1. Simulated examples

We select a pseudo-sample of size 50, $\{(T_1, Y_1), \ldots, (T_{50}, Y_{50})\}$, such that

$$Y_i = \sin(2\pi T_i) + \exp\{-3(T_i - 0.5)^2\} + 0.4 + u_i, \quad 1 \le i \le 50, \tag{4.1}$$

where the u_i's are independently drawn from the standard normal distribution $N(0,1)$ and $T_1, \ldots, T_{50}$ are independent and identically distributed as the uniform distribution $U(0,1)$ on $[0,1]$. The M-type regression spline curve and the least squares curve are obtained. The corresponding MSEs are 0.018 and 0.011 respectively. The fitted curves are shown in Figure 1. Another pseudo-sample of size 50 is drawn in the same way as above except that the error distribution is the symmetric contaminated normal $0.9N(0,1) + 0.1N(0,9^2)$. The MSEs of M-estimator and the LS-estimator are 0.058 and 0.313 respectively. The fitted curves are shown in Figure 2. From Figure 2 we can see that there are two outliers present at the upper right and one present at the lower right. The LS-curve turns in attempting to accommodate them but the M-curve shows little change.

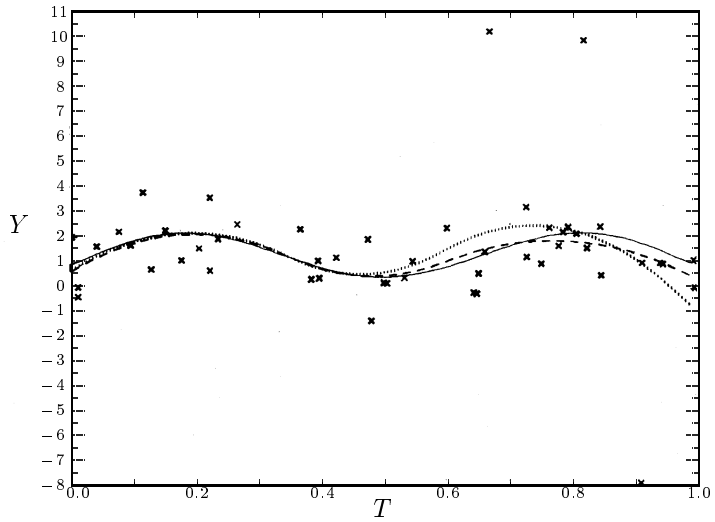

Figure 2. The example of simulated data with sample size $n = 50$. Error distribution is $0.9N(0,1) + 0.1N(0,9^2)$. The solid line is the true curve; the dash line is the regression spline M-fit; the closely spaced dot line is the regression spline LS-fit.

4.2. Simulation study

In our simulated experiment we investigate Huber's M-estimators and the LS estimators which are obtained in (2.1) for $\rho(v) = v^2/2$, where Ψ is Huber ψ-function (Huber (1981, p.177)) with parameter 1.5. When the knot set is determined by the method described in Section 2, Huber's Iteration Procedure (2.3) is employed to produce the final solution. MSE's of the M-estimators and LS estimators are compared by simulation for three kinds of error distributions, the normal, the symmetric contaminated normal, and Cauchy distributions.

Let $g_0(t) = \sin(2\pi t) + \exp\{-3(t-0.5)^2\} + 0.4$. $T_1, \ldots, T_n$ are independently drawn from $U(0,1)$. Thus, $Y_i = g_0(T_i) + u_i$, $1 \le i \le n$, where the u_i are independently taken from either of the following distributions:

1. Normal $N(0,1)$;
2. Symmetric contaminated normal $\mathrm{SCN}(0,9) = 0.85N(0,1) + 0.15N(0,9^2)$;
3. Cauchy $\mathrm{CAU}(0,1)$.

The data in each case of error distributions consist of $K = 250$ replications of samples of sizes $n = 30,\ 50,\ 100$.

Table 1. Means and medians of the 250 MSEs of the M(LS)-estimates

Distri-		M-estimators		LS estimators	
bution	Sample	Mean	Median	Mean	Median
Normal	30	0.282	0.255	0.275	0.245
	50	0.183	0.166	0.179	0.162
	100	0.097	0.086	0.095	0.085
SCN	30	0.966	0.499	2.359	1.492
	50	0.717	0.388	1.714	1.260
	100	0.280	0.264	0.806	0.661
Cauchy	30	1.127	0.715	5234.806	5.695
	50	0.799	0.586	18819.180	5.173
	100	0.576	0.408	14289.802	6.062

The average and median values of the MSEs of the M-type regression spline M-estimators and the regression spline LS estimators are listed in Table 1.

From the results presented in Table 1, we observe that when the random errors are normally distributed the M-estimators are as good as LS estimators; however, when the random errors are drawn from a symmetrically contaminated normal distribution the M-estimators are superior to LS estimators; and when the random errors are distributed as Cauchy distribution the M-estimators seem

acceptable, but the LS estimators behave poorly.

Our computations were done on COMPAG 386. It took at most 40 seconds of CPU time for each data set.

5. Proof of Theorem 1

The remainder of the paper is devoted to proving Theorem 1 based mainly on the monotonicity of $\Psi(\cdot)$. Note, from the quasi-uniform partition assumption, that $D_0^{-1}k_n^{-1} \le t_i - t_{i-1} \le D_0 k_n^{-1}$. For the sake of simplicity and convenience, we assume uniform partitions $t_i = i/k_n$, $i = 1, \dots, k_n$. Only nonessential modifications are needed to deal with the general quasi-uniform partitions described in Section 2.

First, some notation is needed. Let $V_n' = (B(T_1), \dots, B(T_n))_{N\times n}$, $H_n^2 = V_n'V_n$, $\widehat{\theta}_n = H_n\widehat{\theta}$, $\theta_0 = H_n\theta^*$, $z_i = H_n^+ B(T_i)$, $R_{ni} = R_{nT_i}$, and H_n^+ stand for the Moore inverse of H_n.

Outline of a proof of Theorem 1. From Condition 2 and Corollary 6.21 in Schumaker (1981), p.227 (cf. also Theorem XII.4 in de Boor (1978), p.178), we conclude that there exists a constant M_1 depending only on m and M_0 such that

$$\begin{cases} \sup\limits_{t\in[0,1]} |R_{nt}| \le M_1 k_n^{-(m+\gamma)}, \\ g_0(t) = B(t)'\theta^* - R_{nt}, \end{cases} \tag{5.1}$$

where θ^* is a vector depending on g_0. From (5.1), the triangular inequality, $k_n \sim n^{1/[2(m+\gamma)+1]}$ and Lemmas 8 and 9 of Stone (1985), to prove Theorem 1, we need only verify

$$\sum_{i=1}^{n} \left(B(T_i)'(\widehat{\theta} - \theta^*)\right)^2 = O_P(k_n). \tag{5.2}$$

Lemma 5.1 below and (2.2) imply

$$\sum_{i=1}^{n} \Psi(u_i - z_i'(\widehat{\theta}_n - \theta_0) - R_{ni}) z_i = 0 \quad \text{a.s.} \tag{5.3}$$

Write $U(\theta, L) = \sum_{i=1}^n \Psi(u_i - L\, z_i'\theta - R_{ni}) z_i'\theta$ for $L \in R^1$ and $\theta \in R^N$. From (5.3) and the monotonicity of $\Psi(\cdot)$,

$$0 = U\Big(\frac{\widehat{\theta}_n - \theta_0}{|\widehat{\theta}_n - \theta_0|}, |\widehat{\theta}_n - \theta_0|\Big) \le \sup_{|\theta|=1} U(\theta, Lk_n^{1/2}) \quad \text{a.s.}$$

Thus,

$$\mathcal{P}\left\{|\widehat{\theta}_n - \theta_0| \ge Lk_n^{1/2} \,|\, W^*\right\} \le \mathcal{P}\left\{\sup_{|\theta|=1} k_n^{-1/2} U(\theta, Lk_n^{1/2}) \ge 0 \,|\, W^*\right\},$$

for almost all $W^* = (T_1, T_2, \ldots)$. From this fact and (5.2), we need only check

$$\lim_{L\to\infty} \limsup_{n\to\infty} \mathcal{P}\left\{ \sup_{|\theta|=1} k_n^{-1/2} U(\theta, Lk_n^{1/2}) \geq 0 \,|\, W^* \right\} = 0 \tag{5.4}$$

for almost all W^*, which will be done by decomposing $U(\theta, Lk_n^{1/2})$ into three parts based mainly on Conditions 3 and 4. These statements will be made precise latter.

Remark 5.1. From (5.1) and (5.2), we see that the approximating error (modeling error) has the order of magnitude $k_n^{-(m+\gamma)}$ and the estimating error is of the order of magnitude $(k_n/n)^{1/2}$. Thus, the estimator of g_0 achieves the best rates of convergence only when $k_n^{-(m+\gamma)} \sim (k_n/n)^{1/2}$, which results in the choice of k_n having the order of magnitude $n^{1/[2(m+\gamma)+1]}$.

Our proofs need a result that we state here for convenience, the proof of which will be given later. First, denote the smallest eigenvalue of $(N/n)H_n^2$ by Λ_n and let $\Lambda = b\left((m+1)^2(5(m+2))^{2(m+1)}\right)^{-1}$. From Lemma 5.1 (i) below, we can find a set S_0 such that $P(S_0) = 1$ and for any $W^* = (T_1, T_2, \ldots) \in S_0$, there is an $n_1(W^*) > 0$ for which $V_n'V_n$, H_n and H_n^+ are all positive definite when $n \geq n_1(W^*)$.

Lemma 5.1. *If Condition* 1 *holds and* $\lim_{n\to\infty} n^{\delta_0 - 1} k_n^2 = 0$ *for some positive constant* δ_0, *then as* $n \to \infty$

(i) $\Lambda_n \geq \Lambda$ a.s.;

(ii) $\max_{1\leq i\leq n} |z_i|^2 \leq 2(m+3)N/(n\Lambda)$ a.s.

Proof of (5.4). Observe, from the definition of $U(\cdot,\cdot)$, that

$$\begin{aligned} k_n^{-1/2} U(\theta, k_n^{1/2} L) = & \sum_{i=1}^n \Psi(u_i) z_i'\theta k_n^{-1/2} - L \sum_{i=1}^n D(u_i, -Lk_n^{1/2} z_i'\theta - R_{ni})(z_i'\theta)^2 \\ & - \sum_{i=1}^n D(u_i, -Lk_n^{1/2} z_i'\theta - R_{ni}) z_i'\theta R_{ni} k_n^{-1/2} \\ \widehat{=} & \; J_2(\theta) - LJ_4(\theta) - J_3(\theta). \end{aligned} \tag{5.5}$$

By (5.1) and Lemma 5.1, for any $W^* \in S_0$ and $L > 0$, there exists an $n_2(W^*, L) \geq n_1(W^*)$ such that for $n \geq n_2(W^*, L)$,

$$k_n^{1/2} L|z_i| + |R_{ni}| \leq \min(c_2, c_3). \tag{5.6}$$

Thus,

$$E\left\{\sum_{i=1}^n (z_i'\theta)^2 I(|u_i| \leq c_1) \,|\, W^*\right\} = \theta' \sum_{i=1}^n z_i z_i' \theta \mathcal{P}\{|u_i| \leq c_1\} = q$$

and

$$\sum_{i=1}^{n} D(u_i, -Lk_n^{1/2} z_i'\theta - R_{ni})(z_i'\theta)^2 \geq d_1 \sum_{i=1}^{n} (z_i'\theta)^2 I\{|u_i| \leq c_1\}$$

$$= d_1 q - d_1 \left(\mathcal{P}\{\sum_{i=1}^{n} (z_i'\theta)^2 I\{|u_i| \leq c_1\} \mid W^*\} - \sum_{i=1}^{n} (z_i'\theta)^2 I\{|u_i| \leq c_1\} \right)$$

$$\widehat{=} d_1 q - J_1(\theta) \tag{5.7}$$

for $n \geq n_2(W^*)$. Equations (5.5) and (5.7) imply

$$\sup_{|\theta|=1} k_n^{-1/2} U(\theta, Lk_n^{1/2}) + Ld_1 q \;\leq\; L \sup_{|\theta|=1} |J_1(\theta)| + \sup_{|\theta|=1} |J_2(\theta)| + \sup_{|\theta|=1} |J_3(\theta)|$$

for $n \geq n_2(W^*, L)$. Therefore, to prove (5.4), it suffices to verify

$$\limsup_{n\to\infty} \mathcal{P}\{\sup_{|\theta|=1} |J_1(\theta)| \geq \frac{d_1 q}{3} \mid W^*\} = 0, \tag{5.8}$$

$$\lim_{L\to\infty} \limsup_{n\to\infty} \mathcal{P}\{\sup_{|\theta|=1} |J_k(\theta)| \geq \frac{Ld_1 q}{3} \mid W^*\} = 0 \quad \text{for } k = 2, 3. \tag{5.9}$$

First, by a simple calculation we conclude, from Lemma 5.1 and $k_n \sim n^{1/[2(m+\gamma)+1]}$, that

$$E\{\sup_{|\theta|=1} |J_1(\theta)|^2 \mid W^*\} \leq 2d_1^2 q(1-q)(m+3)^2 (k_n + m)^2/(n\Lambda) \longrightarrow 0$$

for any $W^* \in S_0$ as $n \to \infty$. This fact and Tchebychev's Inequality imply (5.8).

(5.6) and Condition 4 yield

$$|D(u_i, -Lk_n^{1/2} z_i'\theta - R_{ni})| \leq d_2$$

for all $|\theta| \leq 1$ when $n \geq n_2(W^*, L)$. Hence, by (5.1) and Lemma 5.1

$$\sup_{|\theta|=1} |J_3(\theta)| \;\leq\; \frac{d_2}{2} \sup_{|\theta|=1} \left(\sum_{i=1}^{n} (z_i'\theta)^2 + \sum_{i=1}^{n} R_{ni}^2 k_n^{-1} \right) \;\leq\; \frac{d_2}{2} \left(1 + n M_1^2 k_n^{-[2(m+\gamma)+1]} \right)$$

for any $W^* \in S_0$ and $L > 0$ when $n \geq n_2(W^*, L)$. Consequently, (5.9) follows for $k = 3$ from Tchebychev's Inequality and $k_n \sim n^{1/[2(m+\gamma)+1]}$.

Finally, Condition 3 and Lemma 5.1 give

$$E\left(\sup_{|\theta|=1} \left| \sum_{i=1}^{n} \Psi(u_i) z_i'\theta k_n^{-1/2} \right|^2 \,\middle|\, W^*\right) \leq v_0 k_n^{-1} \mathrm{trace}\left(\sum_{i=1}^{n} z_i z_i' \right) = v_0 (k_n + m)/k_n$$

for any $W^* \in S_0$ when $n \geq n_1(W^*)$. Thus, (5.9) follows for $k = 2$ from this inequality and Tchebychev's Inequality.

Proof of Lemma 5.1. (ii) It is easily seen, from the definition of $B_i(\cdot)$, that $\sup_{t\in[0,1]}|B_i(t)|\le 1$ and $B_i(t)=B_i(t)I(s_i\le t\le s_{i+m+1})$ for $i=1,\dots,N$. Hence

$$\begin{aligned}\sup_{t\in[0,1]}B(t)'B(t)&=\max_i\sup_{t\in[t_i^*,t_{i+1}^*]}\sum_{j=1}^N B_j(t)^2\\&=\max_i\sup_{t\in[t_i^*,t_{i+1}^*]}\sum_{\substack{\{j:[s_j,s_{j+m+1}]\cap[t_i^*,t_{i+1}^*)\neq\emptyset\\ j=1,\dots,n\}}}B_j(t)^2\le m+3.\end{aligned}$$

From this inequality and (i), we obtain

$$|z_i|^2=B(T_i)'\left(V_n'V_n\right)^{-1}B(T_i)\le 2(m+3)N/(n\Lambda)$$

for any $W^*\in S_0$ when $n\ge n_1(W^*)$. Therefore, assertion (ii) of Lemma 5.1 follows.

For a $k\times k$ matrix A, let $\|A\|_2=(\sup_{|\xi|=1}\xi'A'A\xi)^{1/2}$. Let P denote the probability measure corresponding to the distribution of T and P_n denote the empirical probability measure associated with $T_1,\dots,T_n$. The notation Ph stands for the expectation of $h(T)$, i.e. $Ph\mathrel{\widehat{=}}\int h dP$. Set $\mathcal{H}_n=\{N\,B_i(t)B_j(t):\ i,j=1,\dots,N\}$. From Corollary 2.3.2 in Golub and Loan (1989), p.58,

$$\begin{aligned}&\|(P_n-P)B(t)B(t)'N\|_2^2\\&\le N^2\max_{1\le l\le N}\sum_{i=1}^N|(P_n-P)B_i(t)B_l(t)|\max_{1\le i\le N}\sum_{l=1}^N|(P_n-P)B_i(t)B_l(t)|.\end{aligned}$$

Thus, noting that each $B_i(\cdot)$ vanishes outside of $[s_i,s_{i+m+1}]$, we conclude

$$\begin{aligned}\|(P_n-P)B(t)B(t)'N\|_2^2&=N^2\left[\max_{1\le l\le N}\sum_{i=1\vee(l-m-2)}^{(l+m+2)\wedge N}|(P_n-P)B_i(t)B_l(t)|\right]^2\\&\le 4(m+3)^2\left[\sup_{h\in\mathcal{H}_n}|(P_n-P)h|\right]^2.\end{aligned}$$

This inequality and Lemmas 5.2 and 5.3 below yield assertion (i) of Lemma 5.1.

Lemma 5.2. *If* $\lim_{n\to\infty}n^{\delta_0-1}k_n^2=0$ *for some positive constant* δ_0, *then*

$$\sup_{h\in\mathcal{H}_n}|P_nh-Ph|\longrightarrow 0\qquad\text{a.s.}$$

Proof. By using the symmetrization method (Pollard (1984, p.15)), from the Hoeffding inequality (cf. Pollard (1984, p.191)) and Borel-Contelli Lemma, Lemma 5.2 can easily be established. We omit the details here.

Lemma 5.3. *If Condition* 1 *is satisfied, then*

$$\liminf_{n\to\infty} \Lambda_n^* \geq \Lambda,$$

where Λ_n^* *denotes the smallest eigenvalue of* $P(B(t)B(t)'\,N)$.

Proof. By Condition 1, for all $k_n > 1$

$$\inf_{|\theta|=1} \theta' P(B(t)B(t)'\,N)\theta \geq b\,N \inf_{|\theta|=1} \int_0^1 \left(\sum_{i=1}^N \theta_i B_i(t)\right)^2 d\,t.$$

It is easily seen that

$$\sum_{j=1}^N \int_{s_j}^{s_{j+m+1}} \left(\sum_{i=1}^N \theta_i B_i(t)\right)^2 d\,t \leq (m+1) \int_0^1 \left(\sum_{i=1}^N \theta_i B_i(t)\right)^2 d\,t$$

for all $\theta = (\theta_1, \ldots, \theta_N)' \in R^N$ and $k_n > 1$. Therefore, to prove Lemma 5.3, we need only check

$$\sum_{i=1}^N \theta_i^2 \leq (m+1)(5(m+2))^{2(m+1)} k_n \sum_{j=1}^N \int_{s_j}^{s_{j+m+1}} \left(\sum_{i=1}^N \theta_i B_i(t)\right)^2 d\,t \qquad (5.10)$$

for all $k_n > 1$ and $\theta \in R^N$.

To prove (5.10), we follow essentially the notation of Chapter 4 of Schumaker (1981). Let $\mathcal{L}_2[0,1]$ denote a class of functions such that for each $h \in \mathcal{L}_\in[\prime, \infty]$, it satisfies $\int_0^1 h^2(t)\,d\,t < \infty$. Let $t_i^* = \cos((m+1-i)\pi/(m+1))$, $i = 0, 1, \ldots, m+1$, and $B_0(t) = 2^{m-1}((t-t_0^*)_+^m - 2(t-t_1^*)_+^m + \cdots + 2(-1)^m(t-t_m^*)_+^m + (-1)^{m+1}(t - t_{m+1}^*)_+^m)$, where t_+^k is the truncated power function. Set

$$h_0(t) = \begin{cases} 0, & \text{for } t < -1, \\ \int_{-1}^t B_0(s)ds, & \text{for } |t| \leq 1, \\ 1, & \text{for } t \geq 1, \end{cases}$$

$$H_j(t) = h_0\left(\frac{2t - s_j - s_{j+m+1}}{s_{j+m+1} - s_j}\right), \quad h_j(t) = H_j(t)\xi_j(t),\ \xi_j(t) = \prod_{l=1}^m (t - s_{j+l})/m!,$$

$$\kappa_j(h) = \int_{s_j}^{s_{j+m+1}} h(t) \frac{d_+^{m+1}}{ds^{m+1}} h_j(s)|_{s=t} d\,t, \quad \text{for } h \in \mathcal{L}_\in[\prime, \infty],\ | = \infty, \ldots, \mathcal{N}, \qquad (5.11)$$

where $\frac{d_+^k}{ds^k}$ is the kth right derivative operator. Note that $\frac{d_+^{m+1}}{ds^{m+1}} h_j(s) \in \mathcal{L}_\in[\prime, \infty]$. Hence, $\kappa_j(\cdot)$ is well defined. From the proof of Theorem 4.41 in Schumaker (1981) (see also Shi (1991)), we conclude that for i, $j = 1, \ldots, N$

$$\kappa_j(B_i) = m!(s_{i+m+1} - s_i)[s_i, \ldots, s_{i+m+1}]h_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

which together with the definition of $\kappa_j(\cdot)$ implies

$$\kappa_j(\sum_{i=1}^{N}\theta_i B_i)=\sum_{i=1}^{N}\theta_i\kappa_j(B_i)=\theta_j,\quad j=1,\ldots,N. \tag{5.12}$$

It is easily seen that

$$\sup_{t\in[s_j,s_{j+m+1}]}|\xi_j^{(k)}|\le\frac{((m+1)\delta_n)^{m-k}}{(m-k)!}\ \text{ and }\ \sup_{t\in[0,1]}|\frac{d_+^{m+1}}{dt^{m+1}}H_j(t)|\le(4/\delta_n)^{m+1}\,m!/4, \tag{5.13}$$

where $\delta_n=1/k_n$. From de Boor (1976), we get $\sup_{[-1,1]}|\frac{d_+^l}{dt^l}B_0(t)|\le 2^{l+1}m!/(m-l-1)!$ for $l=0,1,\ldots,m-1$. This inequality gives

$$\sup_{[s_j,s_{j+m+1}]}|\frac{d_+^{m-k+1}}{dt^{m-k+1}}H_j(t)|\le(4/\delta_n)^{m-k+1}\,\frac{m!}{(k-1)!} \tag{5.14}$$

for $1\le k\le m+1$. Note that $\frac{d_+^{m+1}}{d\,t^{m+1}}\xi_j(t)\equiv 0$. From (5.13) and (5.14), we have

$$\begin{aligned}&\delta_n\sup_{t\in[s_j,s_{j+m+1}]}|\frac{d_+^{m+1}}{d\,t^{m+1}}h_j(t)|\\ &\le\delta_n\sum_{k=0}^{m+1}\binom{m+1}{k}\sup_{t\in[s_j,s_{j+m+1}]}|\frac{d_+^k}{d\,t^k}\xi_j(t)|\sup_t|\frac{d_+^{m+1-k}}{d\,t^{m+1-k}}H_j(t)|.\\ &\le(5(m+2))^{m+1}.\end{aligned}$$

This inequality and the Schwarz inequality yield

$$\kappa_j(h)^2\le(m+1)\delta_n^{-1}(5(m+2))^{2(m+1)}\int_{s_j}^{s_{j+m+1}}h^2(t)dt.$$

Consequently, (5.10) follows from the last inequality and (5.12).

Acknowledgments

This work was supported by the National Natural Science Foundation of China.

References

Agarwal, G. G. and Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. Ann. Statist. **8**, 1307-1325.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory, pp. 267-281.

Atilgan, T. (1988). Basis selection for density estimation and regression. AT&T Bell Laboratories technical memorandum.

Bai, Z. D., Wu, Y. H., Chen, X. R. and Miao, B. Q. (1990). On solvability of an equation arising in the theory of M-estimates. Comm. Statist. Theory Methods **19**, 363-380.

de Boor, C. (1976). On local linear functions which vanish at all B-splines but one. In Theory of Approximation with Applications (Edited by A. G. Law and B. N. Sahney), Academic Press, New York.

de Boor, C. (1978). A Practical Guide to Splines. Springer-Verlag, New York.

Cox, D. D. (1983). Asymptotics for M-type smoothing splines. Ann. Statist. **11**, 530-551.

Cunningham, J. K., Eubank, R. L. and Hsing, T. (1991). M-type smoothing splines with auxiliary scale estimation. Comput. Statist. Data Anal. **11**, 43-51.

Eubank, R. L. (1988). Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York and Basel.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). Ann. Statist. **19**, 1-67.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). Technometrics **31**, 3-21.

Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression functions. In Smoothing Techniques for Curve Estimation (Edited by T. Gasser and M. Rosenblatt), Lecture Notes in Math. **757**, 23-68, Springer-Verlag.

Golub, G. H. and Loan, C. F. V. (1989). Matrix Computations, 2nd edition. The John Hopkins University Press.

Härdle, W. (1984). Robust regression function estimation. J. Multivariate Anal. **14**, 169-180.

Härdle, W. (1990). Applied Nonparametric Regression. Cambridge University Press, New York.

Härdle, W. and Gasser, T. (1984). Robust nonparametric function fitting. J. Roy. Statist. Soc. Ser.B **46**, 42-51.

Härdle, W. and Luckhaus, S. (1984). Uniform consistency of a class of regression function estimators. Ann. Statist. **12**, 612-623.

Härdle, W. and Tsybakov, A. B. (1988). Robust nonparametric regression with simultaneous scale curve estimation. Ann. Statist. **16**, 120-135.

Huber, P. J. (1979). Robust smoothing. In Robustness in Statistics (Edited by Launer and Wilkinson), 33-47. Academic Press, New York.

Huber, P. J. (1981). Robust Statistics. John Wiley, New York.

Lenth, R. V. (1977). Robust splines. Comm. Statist. **A6**, 847-854.

Pollard, D. (1984). Convergence of Stochastic Processes. Springer-Verlag, New York.

Rockafellar, R. T. (1970). Convex Analysis. Princeton University Press.

Schumaker, L. L. (1981). Spline Functions. John Wiley, New York.

Shi, P. D. (1991). Asymptotic behavior of piecewise polynomial and B-spline M-estimates in regression. Ph.D. Thesis, Institute of Systems Science, Academia Sinica, Beijing.

Shi, P. D. (1992). Personal communication with professor Chih-Ling Tsai.

Shi, P. D. (1993). Automatic selection of parameters in spline regression via Kullback-Leibler information. To appear in Sys. Sci. & Math. Scis.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. Ann. Statist. **8**, 1348-1360.

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

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

Stone, C. J. and Koo, C. Y. (1985). Additive splines in statistics. In Proceedings, Annual Meeting of Amer. Statist. Assoc., Statist. Comp. Section, August, 45-48.
Su, B. and Liu, D. (1989). Computational Geometry: Curve and Surface Modeling. Academic Press, New York.

Institute of Systems Science, Academia Sinica, Beijing 100080.

(Received November 1992; accepted May 1994)