Statistica Sinica **10**(2000), 93-108

ON DERIVATIVE ESTIMATION IN SPLINE REGRESSION

Shanggang Zhou and Douglas A. Wolfe

Kansas State University and The Ohio State University

Abstract: We consider the problem of estimating the derivatives of a regression function by the corresponding derivatives of regression splines. Unlike kernel smoothers, these spline derivative estimators do not have boundary problems. In addition, they have simple expressions and are easy to compute. In this paper, we study the local asymptotic properties of these derivative estimators. Under regularity conditions, the asymptotic bias and variance of these estimators are derived, and asymptotic normality is established. Furthermore, we extend the results to random designs and heteroscedastic errors.

Key words and phrases: Asymptotic bias, asymptotic normality, asymptotic variance, Bernoulli polynomial, B-splines, derivative estimators, heteroscedastic errors, random designs, regression splines.

1. Introduction

In nonparametric regression, it is often of interest to estimate some functionals of a regression function, such as its derivatives. For example, in the study of growth curves, the first (speed) and second (spurt) derivatives of the height as a function of age are important parameters for study (Müller (1988)). Other needs for derivative estimation often arise in nonparametric regressions themselves. For example, in constructing interval estimates for a regression function (Eubank and Speckman (1993)) and kernel bandwidth selection (Ruppert, Sheather and Wand (1995)), estimators of higher order derivatives are employed in estimating the leading bias terms. In this work, we study the estimation of derivatives of regression functions using regression spline estimators.

Suppose we observe

$$y_j = g(x_j) + \epsilon_j,\ j = 1, \ldots, n, \tag{1}$$

where the ϵ_j's are uncorrelated with $E\epsilon_j = 0$ and $E\epsilon_j^2 = \sigma^2 > 0$. Here the design points $\{x_j\}_{j=1}^n$ are either deterministic or random, and we assume that each $x_j \in [0, 1]$. Our goal is to estimate the derivatives $g^{(i)}$ $(i \leq m-2)$ provided $g^{(m)}$ exists.

There is a long history, and extensive studies, on the use of spline functions in nonparametric regression. As a result, many of their properties are already

well understood. For example, Agarwal and Studden (1980) have shown that the regression spline can achieve the optimal rate of convergence in the univariate case with fixed design points. Stone (1985, 1994) has extended this result to additive and multivariate models, among other things. Friedman and Silverman (1989) and Friedman (1991) have studied the implementation of regression splines. The application of regression splines to model real data has also been studied by many authors; see, for example, Stone and Koo (1986) and Friedman and Silverman (1989).

In the context of derivative estimation, Stone (1985) has shown that spline derivative estimators can achieve the optimal L_2 rate of convergence. However, it appears that their asymptotic bias and variance properties have not been studied. In this paper, we derive the asymptotic form of the leading terms in the bias and variance of regression spline derivative estimators, and establish asymptotic normality for the proposed derivative estimators. Furthermore, we extend our results to random designs, heteroscedastic errors, and weighted least squares regressions. The present work is a generalization of the work by Zhou, Shen and Wolfe (1998), where the asymptotic bias and variance for regression spline estimators are derived.

In Section 2, we describe the spline regression and provide expressions for the spline derivative estimator. In Section 3, we derive forms for the asymptotic bias and variance for the proposed estimator, and establish the asymptotic normality. In Section 4, we generalize our results to heteroscedastic errors and random designs. Technical details are given in Section 5.

2. Spline Derivative Estimators

Spline regression is a natural generalization of polynomial regression. In polynomial regression, a single polynomial function is used to fit the data. The main drawback of this method is that the estimated curve may be unstable and oscillative in some regions due to the nature of polynomials, especially when higher order polynomials are fitted. To overcome this problem, spline regression fits the model by piecewise polynomials with some smooth constraints at the joints. The use of spline functions as an approximation tool has been extensively studied and its properties are well understood. However, by comparison, its theoretical properties in statistics are much less understood. Schumaker (1981) contains a good overview on this topic.

In order to describe regression splines in detail, we need some notation. For any $l \leq m$, let $\underline{t} = (t_0(=0), t_1, \ldots, t_{k+1}(=1))$ be a partition of $[0,1]$ with $t_l < t_j$ if $l < j$, and let $\mathbf{N}_l(x) = (N_{1,l}(t), \ldots, N_{k+l,l}(t))'$ be a vector of normalized lth order B-splines associated with an extended partition of $[0,1]$ generated by $\{t_j\}_{j=0}^{k+1}$ (Schmaker (1981), p.224).

Let $S(l, \underline{t})$ be the space spanned by $\{N_{j,l}\}_{j=1}^{k+l}$. The regression spline of order m is defined as $\hat{g}(x) = \underline{\hat{a}}\mathbf{N}_m$ with $\underline{\hat{a}}$ minimizing

$$\sum_{i=1}^{n}(y_i - \underline{\hat{a}}\mathbf{N}_m)^2.$$

After some simple algebra, we have the equivalent expression

$$\hat{g}(x) = \mathbf{N}_m^{'}(x)G^{-1}\mathbf{XY}, \tag{2}$$

provided G^{-1} exists, where $\mathbf{Y} = (y_1, \ldots, y_n)^{'}$, $\mathbf{X} = n^{-1}(\mathbf{N}_m(x_1), \ldots, \mathbf{N}_m(x_n))$, and $G = n\mathbf{X}\mathbf{X}^{'}$.

Since $\hat{g}(x)$ is an estimator of $g(x)$, it is natural to consider $\hat{g}^{(i)}(x)$ as an estimator of $g^{(i)}(x)$ for any $i = 1, \ldots, m-2$. Meanwhile, because $\hat{g}(x)$ is a spline function, its derivatives also have simple expressions and are, therefore, relatively easy to analyze. From de Boor (1972), we know that the derivatives of spline functions can be simply expressed in terms of lower order spline functions. More precisely, let $s(t) = \sum_{j=1}^{k+m} a_j N_{j,m}(t) \in S(m, \underline{t})$ be any spline function. Then we have

$$s^{(l)}(x) = \sum_{j=1}^{k+m-l} a_j^{(l)} N_{j,m-l}(x), \tag{3}$$

where $a_j^{(0)} = a_j$, $1 \le j \le k+m$, and

$$a_j^{(l)} = (m-l)(a_{j+1}^{(l-1)} - a_j^{(l-1)})/(t_j - t_{j-m+l}), \quad 1 \le j \le k+m-l. \tag{4}$$

We easily obtain the following expression for $\hat{g}^{(i)}(x)$ from (3):

$$\hat{g}^{(i)}(x) = \mathbf{N}_{m-i}^{'}(x)D^{(i)}G^{-1}\mathbf{XY}, \tag{5}$$

where $D^{(i)} = M_i^{'}M_{i-1}^{'}\cdots M_1^{'}$, with

$$M_l = (m-l)\begin{pmatrix} \frac{-1}{t_1 - t_{1-m+l}} & 0 & 0 & \cdots 0 & 0 \\ \frac{1}{t_1 - t_{1-m+l}} & \frac{-1}{t_2 - t_{2-m+l}} & 0 & \cdots 0 & 0 \\ 0 & \frac{1}{t_2 - t_{2-m+l}} & \frac{-1}{t_3 - t_{3-m+l}} & \cdots 0 & 0 \\ \vdots & \vdots & \vdots & \ddots \vdots & \vdots \\ 0 & 0 & 0 & \cdots 0 & \frac{1}{t_{k+m-l} - t_k} \end{pmatrix}, \tag{6}$$

for $1 \le l \le i \le m-2$.

From (6), it is easy to see that the additional expense of computing derivative estimators is minimal. After the regression coefficient vector $\underline{\hat{a}}$ is obtained by the standard linear least squares estimation procedure, $\underline{\hat{a}}^{(i)}$ can be computed inductively by (4) in less than $4i(k+m)$ extra steps.

3. Local Asymptotic Results

In this section, we study the asymptotic properties of the bias and variance of $\hat{g}^{(i)}(x)$. In order to analyze $\hat{g}^{(i)}(x)$ in detail, it is necessary to characterize the elements of G and G^{-1}. Note that the elements of G depend on the distributions of the knots t_i and the design points. Therefore we need some restrictions on both the knots and the design points. For simplicity, we first discuss the fixed design and then generalize the result to random designs in Section 3. For a fixed design, the following two assumptions are sufficient.

A1 (**Condition on the Knots**)

$$\max_{1\le j\le k} |h_{j+1}/h_j - 1| = o(1) \text{ and } h/\min_{1\le j\le k+1} h_j \le c_1, \tag{7}$$

where $h_j = t_j - t_{j-1}$, $h = \max_{1\le j\le k+1} h_j$, $o(1) \to 0$ as $k \to 0$, and $c_1 > 0$ is a pre-determined constant.

The upper bound on the global mesh ratio implies that $c_1^{-1} < hk < c_1$ and therefore $O(h^l) = O(k^{-l})$ for any $l \in (-\infty, \infty)$. This bound is not essential and can be replaced by a much weaker assumption. In fact, under the assumption that $n \min_{1\le j\le k+1} h_j \to \infty$ as $n \to \infty$, it can be shown that the (l, j)th element of G^{-1} is bounded by $c_m \gamma_m^{|l-j|}/\sqrt{(t_l - t_{l-m})(t_j - t_{j-m})}$ for some constant $c_m > 0$ and $\gamma_m \in (0, 1)$. As a result, results similar to those in Zhou, Shen and Wolfe (1998) can be derived under the additional assumption that there is a constant $\gamma \in (\gamma_m, 1)$ such that $\max_{1\le l,j\le k+1}\{(t_l - t_{l-m})(t_j - t_{j-m})^{-1}\gamma^{|l-j|}\} < \infty$. However, our analysis is more elegant under this restriction, as we shall see in the proofs.

A2 (**Condition on Fixed Designs**)

For fixed designs, we assume that

$$\sup_{x\in[0,1]} |Q_n(x) - Q(x)| = o(k^{-1}), \tag{8}$$

where $Q_n(x)$ is the empirical distribution function for $\{x_j\}_{j=1}^n$, and $Q(x)$ is a distribution with a positive, continuous density $Q(x)$.

Assumption (A2) is a mild restriction on the design points. For instance, if the design points are generated according to a positive continuous density $q(x)$ (called *regular sequences* by Sacks and Ylvisaker (1970)) such that

$$\int_{x_j}^{x_{j+1}} q(x)dx = 1/n,$$

for all $j = 1, \ldots, n-1$, then it is easy to verify that (8) holds, provided $k/n \to 0$.

In the context of L_2 approximation, it is known that the approximation error for a spline function and its derivatives behave like scaled Bernoulli polynomials in L_2 norm (Barrow and Smith (1979)). Under Assumptions (A1) and (A2),

this property is preserved for least squares approximation and, in fact, the bias behavior of $\hat{g}^{(i)}(x)$ is like a scaled Bernoulli polynomial under the L_∞ norm, as stated in the next theorem.

Theorem 3.1. *Under Assumptions* (A1) *and* (A2), *if* $g \in C^m[0,1]$ *with* $m > 2$ *and if* $k/n \to 0$, *then for any* $i = 1, \ldots, m-2$,

$$E(\hat{g}^{(i)}(x)) - g^{(i)}(x) = b_i(x) + o(h^{m-i}),$$

where

$$b_i(x) = \frac{g^{(m)}(x) h_{j+1}^{m-i}}{(m-i)!} B_{m-i}\Big(\frac{x - t_j}{h_{j+1}}\Big), \quad \text{if } t_j < x < t_{j+1},\ j = 0, \ldots, k,$$

and $B_{m-i}(\cdot)$ *is a Bernoulli polynomial* (*Ghizzetti and Ossicini* (1970)).

From (5), the following expression for $\mathrm{Var}\,(\hat{g}^{(i)}(x))$ can be obtained:

$$\mathrm{Var}\,(\hat{g}^{(i)}(x)) = \frac{\sigma^2}{n} \mathbf{N}_{m-i}'(x) D^{(i)} G^{-1} (D^{(i)})' \mathbf{N}_{m-i}(x). \tag{9}$$

The asymptotic form of $\mathrm{Var}\,(\hat{g}^{(i)}(x))$ is provided by the next theorem.

Theorem 3.2. *Under Assumptions* (A1) *and* (A2), *if* $k/n \to 0$, *then*

$$\mathrm{Var}\,(\hat{g}^{(i)}(x)) = \frac{\sigma^2}{n} \mathbf{N}_{m-i}'(x) D^{(i)} G^{-1}(q) (D^{(i)})' \mathbf{N}_{m-i}(x) + o\left(h^{-2i-1} n^{-1}\right),$$

where

$$G(q) = \int_0^1 \mathbf{N}_m(x) \mathbf{N}_m'(x) q(x) dx.$$

Using Theorems 3.1 and 3.2, we establish the asymptotic normality of $\hat{g}^{(i)}(x)$ in the next theorem.

Theorem 3.3. *In addition to Assumptions* (A1) *and* (A2), *suppose that the* ϵ_j *are i.i.d. from some distribution with mean* 0 *and variance* σ^2. *If* $g \in C^m[0,1]$ *and* $k \geq c_2 n^{1/(2m+1)}$ *for some positive constant* c_2, *then*

$$\frac{\hat{g}^{(i)}(x) - g^{(i)}(x) - b_i(x)}{\sqrt{\mathrm{Var}\,(\hat{g}^{(i)}(x))}} \xrightarrow{d} N(0,1).$$

Remark 1. From Theorems 3.1 and 3.2, the bias of $\hat{g}^{(i)}(x)$ is $O(h^{m-i})$ and the variance is $O(h^{-2i-1} n^{-1})$ (see also Lemma 5.4). As a result, if the number of knots k is of order $O(n^{1/(2m+1)})$, the mean square error is

$$MSE(\hat{g}^{(i)}(x)) = E(\hat{g}^{(i)}(x) - g^{(i)}(x))^2 = O\left(n^{2(m-i)/(2m+1)}\right), \tag{10}$$

for any $x \in [0,1]$. This agrees with the result on local convergence rate for a nonparametric regression estimator in Stone (1982). An interesting fact about the optimal number of knots is that its order in magnitude does not depend on i. This may provide a clue on how to choose the optimal number of knots for $\hat{g}^{(i)}(x)$. We will discuss this topic in a forthcoming paper.

Remark 2. Unlike kernel smoothers, $\hat{g}^{(i)}(x)$ achieves the optimal rate noted in (10) for any $x \in [0,1]$. Meanwhile, we should note that the variance of $\hat{g}^{(i)}(x)$ near the boundary of $[0,1]$ is significantly larger than in the interior. This can be seen clearly in Figure 1 which plots the ratio of boundary $\sqrt{\mathrm{Var}\,(\hat{g}^{(i)}(x))}$ to the average interior $\sqrt{\mathrm{Var}\,(\hat{g}^{(i)})}$. This phenomenon can easily be explained by the fact that there are fewer observations near the boundary contributing to the regression.

Remark 3. The local asymptotic results in Zhou, Shen and Wolfe (1998) can be considered a special case of the results in this work. In fact, if we define $D^{(0)} = I_{k+m}$, where I_{k+m} is the $(k+m)$ by $(k+m)$ identity matrix, then Theorems 3.1 and 3.2 also provide the local asymptotic expressions for $\hat{g}(x)$, corresponding to the case $i = 0$.

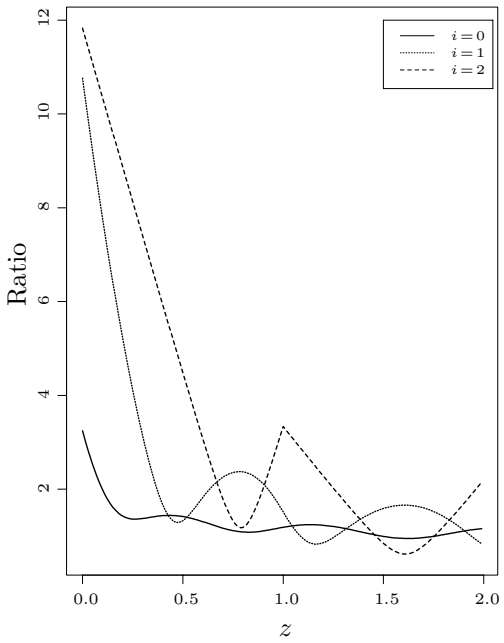

Figure 1. Ratio of boundary $\sqrt{\mathrm{Var}\,(\hat{g}^{(i)}(z))}$ to average $\sqrt{\mathrm{Var}\,(\hat{g}^{(i)})}$ in the interior, where $\hat{g}$ is a cubic spline with $k = 9$ equally spaced knots, where $z = x/h$.

4. Extensions

Random Design. In addition to the fixed design where the experimenters can determine or control the design points, another possible sampling scheme corresponds to selecting the design points by a random process. This is often the case in observational studies. Under certain conditions, an appropriate model to describe this situation corresponds to $x_1, \ldots, x_n$ being a random sample from a distribution $Q(x)$, as stated in the following assumption.

(A2′) (**Condition on the random Design**)

$x_1, \ldots, x_n$ are i.i.d. with cdf $Q(x)$, where $Q(x)$ is a continuous distribution function with a positive, continuous density $q(x)$ on $[0, 1]$.

For random designs, results similar to those in Section 3 can be established for $\hat{g}^{(i)}(x)$, as stated in the following theorems.

Theorem 4.1. *Under Assumptions* (A1) *and* (A2′), *if* $g \in C^m[0,1]$ *with* $m > 2$ *and if* $k/n \to 0$, *then for any* $i = 1, \ldots, m-2$ *and* $t_j < x < t_{j+1}$, *we have*

$$E(\hat{g}^{(i)}(x)|x_1, \ldots, x_n) - g^{(i)}(x) = b_i(x) + o_P(h^{m-i}).$$

Theorem 4.2. *Under Assumptions* (A1) *and* (A2′), *if* $k/n \to 0$, *then*

$$\mathrm{Var}\,(\hat{g}^{(i)}(x)|x_1, \ldots, x_n) = \frac{\sigma^2}{n}\mathbf{N}_{m-i}'(x)D^{(i)}G^{-1}(q)(D^{(i)})\mathbf{N}_{m-i}(x) + o_P(h^{-2i-1}n^{-1}).$$

Theorem 4.3. *In addition to Assumptions* (A1) *and* (A2′), *suppose that the* ϵ_j *are i.i.d. from some distribution with mean* 0 *and variance* σ^2. *If* $g \in C^m[0,1]$ *and* $k \geq c_2 n^{1/(2m+1)}$ *for some positive constant* c_2, *then we have*

$$\frac{\hat{g}^{(i)}(x) - g^{(i)}(x) - b_i(x)}{\sqrt{\mathrm{Var}\,(\hat{g}^{(i)}(x)|x_1, \ldots, x_n)}} \xrightarrow{d} N(0,1).$$

Unconditional bias and variance. For a random design, there always exists a probability that the Gram matrix G is singular and therefore that $\hat{g}^{(i)}(x)$ is not defined. However, if we slightly modify the definition of $\hat{g}^{(i)}(x)$ by using the idea of ridge regression to guard against the singularity of $\mathbf{G}$, $\hat{g}^{(i)}(x)$ can still have the desired unconditional bias and variance. For example, if we define

$$\hat{g}_a^{(i)}(x) = \mathbf{N}_{m-i}'(x)D^{(i)}(\mathbf{G} + n^{-2}\mathbf{I})^{-1}\mathbf{X}\mathbf{Y}, \tag{11}$$

where $\mathbf{I}$ is the $(k+m) \times (k+m)$ identity matrix, then we have the following theorem on the unconditional asymptotic bias and variance of $\hat{g}_a^{(i)}(x)$.

Theorem 4.4. *Under Assumptions* (A1) *and* (A2′), *if* $g \in C^m[0,1]$ *and* $k \to \infty$ *but* $k/n^\delta \to 0$ *for some constant* $\delta \in (0, 1/2)$ *as* $n \to \infty$, *then*

$$E(\hat{g}_a^{(i)}(x)) - g^{(i)}(x) = b_i(x) + o(h^{m-i}). \tag{12}$$

$$\operatorname{Var}(\hat{g}_a^{(i)}(x)) = \frac{\sigma^2}{n}\mathbf{N}'_{m-i}(x)D^{(i)}G^{-1}(q)(D^{(i)})'\mathbf{N}_{m-i}(x) + o(h^{2(m-i)} + h^{-2i-1}n^{-1}). \tag{13}$$

Heteroscedastic errors. In practice, there are situations where the observations exhibit heteroscedastic errors. Hence the regression model becomes

$$y_j = g(x_j) + v^{1/2}(x_j)\epsilon_j,$$

where $v(x)$ is a positive function on $[0,1]$. For such settings, it may be more appropriate to use a weighted least squares criterion to define $\hat{g}(x)$. For example, $\hat{g}_v(x) = \sum_{j=1}^{k+m} a_j N_{j,m}(x)$ can be considered, where $\underline{a}$ is the minimizer of

$$\sum_{j=1}^{n} \frac{1}{v(x_j)}(y_j - \underline{a}\mathbf{N}_m(x_j))^2.$$

It is known (see Zhou, Shen and Wolfe (1998)) that, if $v(x)$ is continuous,

$$E(\hat{g}_v(x)) - g(x) = b_0(x) + o(h^m),$$

$$\operatorname{Var}(\hat{g}_v(x)) = \frac{\sigma^2}{n}\mathbf{N}'_m(x)G_v^{-1}(q)\mathbf{N}_m(x) + o((nh)^{-1}),$$

where $G_v(q) = \int_0^1 v(x)\mathbf{N}_m(x)\mathbf{N}'_m(x)q(x)dx$.

The results in Section 3 easily extend to $\hat{g}_v(x)$, as stated in the following theorems.

Theorem 4.5. *Under Assumptions* (A1) *and* (A2), *if* $g \in C^m[0,1]$ *with* $m > 2$ *and if* $k/n \to 0$, *then for any* $i = 1, \ldots, m-2$,

$$E(\hat{g}_v^{(i)}(x)) - g^{(i)}(x) = b_i(x) + o(h^{m-i}),$$

if $t_j < x < t_{j+1}$.

Theorem 4.6. *Under Assumptions* (A1) *and* (A2), *if* $k/n \to 0$,

$$\operatorname{Var}(\hat{g}_v^{(i)}(x)) = \frac{\sigma^2}{n}\mathbf{N}'_{m-i}(x)D^{(i)}G_v^{-1}(q)(D^{(i)})'\mathbf{N}_{m-i}(x) + o\left(h^{-2i-1}n^{-1}\right).$$

Theorem 4.7. *In addition to Assumptions* (A1) *and* (A2), *suppose that the* ϵ_j *are i.i.d. from some distribution with mean* 0 *and variance* σ^2. *If* $g \in C^m[0,1]$ *and* $k \geq c_2 n^{1/(2m+1)}$ *for some positive constant* c_2, *then*

$$\frac{\hat{g}_v^{(i)}(x) - g^{(i)}(x) - b_i(x)}{\sqrt{\operatorname{Var}(\hat{g}_v^{(i)}(x))}} \xrightarrow{d} N(0,1).$$

Proof of the above three theorems are similar to those in Section 3 and are, therefore, omitted.

5. Proofs

We need some preliminary lemmas for the proof of Theorem 3.1.

Lemma 5.1. *For any $g \in C^m[0,1]$, there exists an $s_g(x) \in S(m, \underline{t})$ such that*

$$\left\| ((k+1)^m (g - s_g)(x) - b_0(x))^{(i)} \right\|_{L_\infty[0,1]} = o(1),$$

for any $i = 0, \ldots, m-2$.

Lemma 5.1 was established by Barrow and Smith (1979) (see Lemma 1 of that paper) under the condition that the knot sequence $\underline{t}$ is generated according to a continuous positive density. However, it can be seen that their proof needs only a few changes to establish Lemma 5.1 under Assumption (A1). We omit the details. From Lemma 5.1,

$$(g - s_g)^{(i)}(x) = b_i(x) + o(h^{m-i}). \tag{14}$$

Lemma 5.2. *For any $1 \le i \le m-2$,*

$$\| D^{(i)} \|_\infty = O(h^{-i}).$$

Proof of Lemma 5.2. By the definition of $D^{(i)}$,

$$\| D^{(i)} \|_\infty = \| M_i \cdots M_1 \|_\infty \le \| M_i \|_\infty \cdots \| M_1 \|_\infty .$$

From (6), we have $\| M_i \|_\infty = O(h^{-1})$. Hence it follows from the above inequality that

$$\| D^{(i)} \|_\infty = O(h^{-i}).$$

For convenience, we next introduce several results from Zhou, Shen and Wolfe (1998). For proofs, see Lemmas 5.3, 5.4 and 5.5 of that paper.

1. If A and B are $l \times l$ nonnegative matrices, then

$$\lambda^A_{min} Tr(B) \le Tr(AB) \le \lambda^A_{max} Tr(B), \tag{15}$$

where λ^A_{min} and λ^A_{max} are the minimum and maximum eigenvalues of A, respectively.

2.
$$\| G^{-1} \|_\infty = O(h^{-1}). \tag{16}$$

3.
$$\max_{1 \le l,j \le k+m} |\alpha_{lj} - \alpha_{lj}(q)| = o(h^{-1}), \tag{17}$$

where α_{lj} and $\alpha_{lj}(q)$ are the (l,j)th element of G^{-1} and $G^{-1}(q)$, respectively.

Proof of Theorem 3.1. Let

$$R_m^{(i)}(x) = \frac{g^{(m)}(t_j)h_{j+1}^{m-i}}{(m-i)!}B_{m-i}\Big(\frac{x-t_j}{h_{j+1}}\Big),\ t_j < x < t_{j+1},\ j = 0, \ldots, k.$$

Note that since $g \in C^m[0,1]$, we have $g^{(m)}(x) = g^{(m)}(t_j) + o(1)$ and

$$R_m^{(i)}(x) - b_i(x) = o(h^{m-i}).$$

Hence it suffices to show that

$$g^{(i)}(x) - E(\hat{g}^{(i)}(x)) = R_m^{(i)}(x) + o(h^{m-i}).$$

From (14), we know that

$$(g - s_g)^{(i)}(x) = R_m^{(i)}(x) + o(h^{m-i}).$$

Hence Theorem 3.1 is established if we show that

$$D^{(i)}(E\hat{g}(x) - s_g(x)) = o(h^{m-i}). \tag{18}$$

In the proof of Theorem 2.1 of Zhou, Shen and Wolfe (1998) (see (24) of that paper), it has been shown that

$$E(\hat{g}(x)) - s_g(x) = o(h^m).$$

It follows from Lemma 5.2 that

$$\| D^{(i)}(E\hat{g}(x) - s_g(x)) \|_{L_\infty[0,1]} \leq \| D^{(i)} \|_\infty \| E\hat{g}(x) - s_g(x) \|_{L_\infty[0,1]} = o(h^{m-i}),$$

and the proof of Theorem 3.1 is complete.

Proof of Theorem 3.2. Let j_x be the integer such that $x \in [t_{j_x-1}, t_{j_x}]$. By the definition of B-spline functions (see, e.g., de Boor (1972), p.52), we have

$$N_{j,m-i}(x) = 0 \qquad \text{if } j < j_x \text{ or } j > j_x + m - i - 1. \tag{19}$$

Let $A_i(x) = \mathbf{N}'_{m-i}(x)D^{(i)} = (a_1(x), \ldots, a_k(x))'$. By the definition of M_i, it is easy to verify that

$$a_j(x) = 0 \qquad \text{if } j < j_x \text{ or } j > j_x + m - 1. \tag{20}$$

Hence from (9), we have

$$\text{Var}\,(\hat{g}^{(i)}(x)) = \frac{\sigma^2}{n}\sum_{j=j_x}^{j_x+m-1}\sum_{l=j_x}^{j_x+m-1}\alpha_{jl}a_j(x)a_l(x), \tag{21}$$

where α_{jl} is the (j,l)th element of G^{-1}. Let $\alpha_{jl}(q)$ be the (j,l)th element of $G^{-1}(q)$. Using similar arguments, we have

$$\frac{\sigma^2}{n}\mathbf{N}'_{m-i}(x)D^{(i)}G^{-1}(q)(D^{(i)})'\mathbf{N}_{m-i}(x) = \frac{\sigma^2}{n}\sum_{j=j_x}^{j_x+m-1}\sum_{l=j_x}^{j_x+m-1}\alpha_{jl}(q)a_j(x)a_l(x). \tag{22}$$

Let

$$V(x) = \mathrm{Var}\,(\hat{g}^{(i)}(x)) - \frac{\sigma^2}{n}\mathbf{N}'_{m-i}(x)D^{(i)}G^{-1}(q)(D^{(i)})'\mathbf{N}_{m-i}(x).$$

It follows from (21) and (22) that

$$V(x) = \frac{\sigma^2}{n}\sum_{j=j_x}^{j_x+m-1}\sum_{l=j_x}^{j_x+m-1}(\alpha_{jl}-\alpha_{jl}(q))a_j(x)a_l(x).$$

Hence,

$$\| V(x) \|_{L_\infty[0,1]} \leq \frac{\sigma^2}{n}m^2 \max_{1\leq j,l\leq k+m}|\alpha_{jl}-\alpha_{jl}(q)| \,\| \mathbf{N}'_{m-i}D^{(i)} \|^2_\infty . \tag{23}$$

By (23), (17) and Lemma 5.2, we have

$$\| V(x) \|_{L_\infty[0,1]} = o(h^{-1})O(h^{-2i})n^{-1} = o(h^{-2i-1}n^{-1}),$$

and the proof of Theorem 3.2 is complete.

Let $M_{m-i,j}$ be the $(m-i+1)\times(m-i)$ sub-matrix of M_i formed by the elements of M_i at row u, $j{+}1 \leq u \leq j{+}m{-}i$ and column l, $j{+}1 \leq l \leq j{+}m{-}i{+}1$, that is,

$$\frac{M_{m-i,j}}{m-i} = \begin{pmatrix} \frac{-1}{t_j-t_{j-m+i}} & 0 & \cdots 0 & 0 \\ \frac{1}{t_j-t_{j-m+i}} & \frac{-1}{t_{j+1}-t_{j+1-m+i}} & \cdots 0 & 0 \\ 0 & \frac{1}{t_{j+1}-t_{j+1-m+i}} & \cdots 0 & 0 \\ \vdots & \vdots & \ddots \vdots & \vdots \\ 0 & 0 & \cdots 0 & \frac{1}{t_{j+m-i-1}-t_{j-1}} \end{pmatrix}. \tag{24}$$

We need the following lemmas in the proof of Theorem 3.3.

Lemma 5.3. *Let $\{\lambda_l^i\}_{l=1}^{m-i}$ be the set of eigenvalues of $M'_{m-i,j}M_{m-i,j}$. There exist $\lambda_{max} > \lambda_{min} > 0$, depending only on m, such that,*

$$\lambda_{min}h^{-2} \leq \lambda_l^i \leq \lambda_{max}h^{-2}, \;\; 0 \leq l \leq m-i,$$

for all $1 \leq i \leq m-2$ and $1 \leq j \leq k+m-i$.

Proof of Lemma 5.3. Define

$$U_{m-i,j} = (m-i)\Big(\frac{1}{t_{j+1} - t_{j+1-m+i}}, \ldots, \frac{1}{t_{j+m-i} - t_j}\Big)',$$

and

$$V_{m-i} = \begin{pmatrix} -1 & 0 & 0 \ldots 0\, 0 \\ 1 & -1 & 0 \ldots 0\, 0 \\ 0 & 1 & -1 \ldots 0\, 0 \\ \vdots & \vdots & \vdots \ddots \vdots\, \vdots \\ 0 & 0 & 0 \ldots 0\, 1 \end{pmatrix}.$$

It is easy to verify that $M_{m-i,j} = V_{m-i}U_{m-i,j}$. Let $B_{m-i} = (b_1, \ldots, b_{m-i})'$ be any real unit vector. Then we have

$$B'_{m-i}V'_{m-i}V_{m-i}B_{m-i} = 2\Big(\sum_{l=1}^{m-i} b_l^2 - \sum_{l=1}^{m-i-1} b_l b_{l+1}\Big) \geq 2/m.$$

Hence,

$$\begin{aligned} B'_{m-i}M'_{m-i,j}M_{m-i,j}B_{m-i} &\geq \frac{2}{m}B'_{m-i}U'_{m-i,j}U_{m-i,j}B_{m-i} \\ &= \frac{2}{m}\sum_{l=1}^{m-i}\Big(\frac{(m-i)b_l}{t_{l+j} - t_{l+j-m+i}}\Big)^2 \geq \frac{2h^{-2}}{m}. \end{aligned}$$

It follows that $\lambda_l^i \geq (2/m)h^{-2}$ for any i and l. To show the right side inequality in Lemma 5.3, note that

$$V'_{m-i}V_{m-i} = \begin{pmatrix} 2 & -1 & 0 \ldots & 0\, 0 \\ -1 & 2 & -1 \ldots & 0\, 0 \\ \vdots & \vdots & \vdots \ddots & \vdots\, \vdots \\ 0 & 0 & 0 \ldots & -1\, 2 \end{pmatrix}.$$

By Gerschgorin's Theorem (see, e.g., Ortega (1987), p.227), the maximum eigenvalue of $V'_{m-i}V_{m-i}$ is less than 4. Thus we have

$$B'_{m-i}M'_{m-i,j}M_{m-i,j}B_{m-i} \leq 4\sum_{l=1}^{m-i}\Big(\frac{(m-i)b_l}{t_{l+j} - t_{l+j-m+i}}\Big)^2 \leq 4\Big(\frac{m}{\min_{1\leq l\leq k+1} h_l}\Big)^2.$$

It follows from (7) that $\lambda_l^i \leq 4(c_1 m)^2 h^{-2}$ for any i and l, which completes the proof of Lemma 5.3.

Lemma 5.4. *For any $1 \leq i \leq m-2$, there exist two constants $L_{max} > L_{min} > 0$ such that*

$$\frac{L_{min}\sigma^2}{nh^{2i+1}} \leq \mathrm{Var}\,(\hat{g}^{(i)}(x)) \leq \frac{L_{max}\sigma^2}{nh^{2i+1}}. \tag{25}$$

Proof of Lemma 5.4. Let $\mathbf{N}_{m-i}^{j_x}(x) = (N_{j_x,m-i}(x), \ldots, N_{j_x+m-i-1,m-i}(x))'$ and $P_{i,x} = M_{m-1,j_x} \cdots M_{m-i,j_x}$. By the definition of M_{m-i,j_x} in (24), it follows from (19) that, for any $l = 1, \ldots, m-2$,

$$Tr\left(D^{(i)} N_{m-i}(x) N'_{m-i}(x) (D^{(i)})'\right) = Tr\left(P_{i,x} \mathbf{N}_{m-i}^{j_x} (\mathbf{N}_{m-i}^{j_x}(x))' P'_{i,x}\right). \tag{26}$$

In addition, using (15) and Lemma 5.3, we have

$$\begin{aligned} Tr\Big(P_{i,x} \mathbf{N}_{m-i}^{j_x}(x) (\mathbf{N}_{m-i}^{j_x}(x))' P'_{i,x}\Big) &\geq (\lambda_{min})^i h^{-2i} \Big\{ \sum_{l=j_x}^{j_x+m-i} N_{l,m-i}(x) \Big\}^2 \\ &\geq (\lambda_{min})^i h^{-2i}/m \end{aligned} \tag{27}$$

and

$$\begin{aligned} Tr\Big(P_{i,x} \mathbf{N}_{m-i}^{j_x}(x) (\mathbf{N}_{m-i}^{j_x}(x))' P'_{i,x}\Big) &\leq (\lambda_{max})^i h^{-2i} \Big\{ \sum_{l=j_x}^{j_x+m-i} N_{l,m-i}(x) \Big\}^2 \\ &\leq (\lambda_{max})^i h^{-2i}. \end{aligned} \tag{28}$$

From (9),

$$\mathrm{Var}\,(\hat{g}^{(i)}(x)) = \frac{\sigma^2}{n} Tr\left(D^{(i)} N_{m-i}(x) N'_{m-i}(x) (D^{(i)})' G^{-1}\right).$$

It follows from (16) and (15) that

$$\begin{aligned} \frac{c_4 \sigma^2}{nh} Tr\Big(D^{(i)} N_{m-i}(x) N'_{m-i}(x) (D^{(i)})'\Big) &\leq \mathrm{Var}\,(\hat{g}^{(i)}(x)) \\ &\leq \frac{c_5 \sigma^2}{nh} Tr\Big(D^{(i)} N_{m-i}(x) N'_{m-i}(x) (D^{(i)})'\Big), \end{aligned}$$

for constants $c_5 > c_4 > 0$, and Lemma 5.4 follows from (26), (27), (28) and the above inequality.

Proof of Theorem 3.3. If $k \geq c_2 n^{1/(2m+1)}$, then by Theorems 3.1 and 3.2,

$$\| E\hat{g}^{(i)}(x) - g^{(i)}(x) - b_i(x) \|_{L_\infty[0,1]} = o(n^{\frac{-i}{2m+1}}), \text{ and } \sqrt{\mathrm{Var}\,(\hat{g}^{(i)}(x))} = O(n^{\frac{-i}{2m+1}}).$$

Therefore it is enough to show that

$$\frac{\hat{g}^{(i)}(x) - E\hat{g}^{(i)}(x) - b_i(x)}{\sqrt{\mathrm{Var}\,(\hat{g}^{(i)}(x))}} \xrightarrow{d} N(0,1).$$

By (5), we have

$$\hat{g}^{(i)}(x) - E\hat{g}^{(i)}(x) = \mathbf{N}'_{m-i}(x) D^{(i)} G^{-1} \mathbf{N}_{\underline{x}} \epsilon = \sum_{j=1}^{n} w_j \epsilon_j,$$

where $w_j(x) = \mathbf{N}'_{m-i}(x)D^{(i)}G^{-1}\mathbf{N}_m(x_j)/n$. To verify that the Lindeberg-Feller condition holds, it suffices to show that

$$\max_{1\le j\le n}\left(w_j^2\right) = o\Big(\sum_{j=1}^{n} w_j^2\Big) = o\left(\mathrm{Var}\,(\hat{f}^{(i)}(x))\right). \tag{29}$$

By (15), (16) and Lemma 5.2, we have

$$\begin{aligned} w_j^2 n^2 &= \mathbf{N}'_{m-i}(x)D^{(i)}G^{-1}\mathbf{N}_m(x_j)\mathbf{N}'_m(x_j)G^{-1}(D^{(i)})'\mathbf{N}_{m-i}(x) \\ &= Tr(\mathbf{N}_{m-i}(x)\mathbf{N}'_{m-i}(x)D^{(i)})G^{-1}\mathbf{N}_m(x_j)\mathbf{N}'_m(x_j)G^{-1}(D^{(i)})') \\ &\le e_{m-i}\,Tr((D^{(i)})'D^{(i)}G^{-1}\mathbf{N}_m(x_j)\mathbf{N}'_m(x_j)G^{-1}) \\ &\le e_{m-i}\;\|\;(D^{(i)})'\;\|_\infty\|\;D^{(i)}\;\|_\infty\,Tr(G^{-2}\mathbf{N}_m(x_j)\mathbf{N}'_m(x_j)) \\ &\le e_{m-i}\,O(h^{-2i-2})\sum_{l=1}^{k+m} N_{l,m}^2(x_j), \end{aligned}$$

where

$$e_{m-i} = \max_{x\in[0,1]}\{\lambda(x):\ \ \lambda(x)\ \text{ is the maximum eigenvalue of }\ \mathbf{N}_{m-i}(x)\mathbf{N}'_{m-i}(x)\}.$$

By definition, for any $x \in [0,1]$ and $1 \le i \le m-1$, $0 \le N_{l,m-i}(x) \le 1$ and $\sum_{l=1}^{k+m} N_{l,m}(x) = 1$, $1 \le l \le k$. This implies that $\sum_{l=1}^{k+m} N_{l,m}^2(x_j) \le 1$ and $e_{m-i} \le 1$. Therefore $w_j^2 n^2 = O(h^{-2i-2})$. Hence (29) follows from Lemma 5.4 and the assumption that $k/n \to 0$ ($hn \to \infty$), and the proof of Theorem 3.3 is complete.

Proof of Theorems 4.1, 4.2, 4.3. By the Glivenko-Cantelli Theorem (see, e.g., Gaenssler and Wellner (1981)), we have $\max_{0\le x\le 1}|Q_n(x) - Q(x)| = O_p(n^{-1/2})$. Using arguments similar to those in the proofs of Theorems 3.1, 3.2 and 3.3, the desired results follow.

Proof of (12) of Theorem 4.4. Set

$$\Omega_\delta = \{(x_1,\dots,x_n):\ \ \max_x |Q_n(x) - Q(x)| \le k^{-1}n^{-(1-2\delta)/4}\},$$

and let Ω_δ^c be the complement of Ω_δ. For any $\underline{z} \in \Omega_\delta$, using arguments similar to those in the proof of Theorems 3.1 and 3.2, we have

$$E(\hat{g}_a^{(i)}(x)|\underline{x}=\underline{z}) - g^{(i)}(x) = b_i(x) + o(h^m), \tag{30}$$

$$\mathrm{Var}\,(\hat{g}_a^{(i)}(x)|\underline{x}=\underline{z}) = \frac{\sigma^2}{n}\mathbf{N}'_{m-i}(x)D^{(i)}G^{-1}(q)(D^{(i)})'\mathbf{N}_{m-i}(x) + o\Big(h^{-2i-1}n^{-1}\Big). \tag{31}$$

It follows from (30) that

$$\begin{aligned}
&|E(\hat{g}_a^{(i)}(x)) - g^{(i)}(x) - b_i(x)| \\
&\le |E\{[E(\hat{g}_a^{(i)}(x)|\underline{x}) - g^{(i)}(x) - b_i(x)]1_{\Omega_\delta}(\underline{x})\}| \\
&\quad +(1-p_\delta)\,|g^{(i)}(x) - b_i(x)| + |E\{E(\hat{g}_a^{(i)}(x)|\underline{x})1_{\Omega_\delta^c}(\underline{x})\}| \\
&\le o(h^{m-i})p_\delta + (1-p_\delta)|g^{(i)}(x) - b_i(x))| + (1-p_\delta)\max_{\underline{x}}|E(\hat{g}_a^{(i)}(x)|\underline{x})|,
\end{aligned}$$

where $p_\delta = P(\underline{x} \in \Omega_\delta)$. From Gaenssler and Wellner (1981), we know that $1 - p_\delta \le c_6 e^{-c_7 n^{(1-2\delta)/2}}$ for some positive constants c_6 and c_7. Hence, it suffices to show that

$$\max_{\underline{x}}|E(\hat{g}_a^{(i)}(x)|\underline{x})| = O(n^{i+3}). \tag{32}$$

From (11),

$$\begin{aligned}
\max_{\underline{x}}|E(\hat{g}_a^{(i)}(x)|\underline{x})| &= \max_{\underline{x}}|\mathbf{N}_{m-i}'(x)D^{(i)}(\mathbf{G} + n^{-2}\mathbf{I})^{-1}\mathbf{X}g(\underline{x})| \\
&\le \max_{\underline{x}}\{\| \,(\mathbf{G} + n^{-2}\mathbf{I})^{-1}\,\|_\infty \| \,D^{(i)}\,\|_\infty \| \,\mathbf{X}\,\|_\infty\} \,\| \,g\,\|_{L_\infty[0,1]} \\
&\le \| \,g\,\|_{L_\infty[0,1]}\,(k+m)n^2h^{-i}\max_{\underline{x}}\Big\{\max_{1\le l\le k+m}\sum_{j=1}^{n}N_{l,m}(x_j)/n\Big\} \\
&\le \| \,g\,\|_{L_\infty[0,1]}\,n^{i+3},
\end{aligned}$$

where $g(\underline{x}) = (g(x_1), \ldots, g(x_n))'$. Hence (32) holds and the proof is complete.

Proof of (13) of Theorem 4.4. Noting that

$$\mathrm{Var}\,(\hat{g}_a^{(i)}(x)) = E\{\mathrm{Var}\,(\hat{g}_a^{(i)}(x)|\underline{x})\} + E\{E(\hat{g}_a^{(i)}(x)|\underline{x}) - E\hat{g}_a^{(i)}(x)\}^2,$$

it suffices to show that

$$E\{\mathrm{Var}\,(\hat{g}_a^{(i)}(x)|\underline{x})\} = \frac{\sigma^2}{n}\mathbf{N}_{m-i}'(x)D^{(i)}G^{-1}(q)(D^{(i)})'\mathbf{N}_{m-i}(x) + o(h^{-2i-1}n^{-1}) \tag{33}$$

$$E\{E(\hat{g}_a^{(i)}(x)|\underline{x}) - E\hat{g}_a^{(i)}(x)\}^2 = o(h^{2(m-i)} + h^{-2i-1}n^{-1}). \tag{34}$$

Equality (33) follows from (30) and arguments similar to those in the proof of (12). Now let us turn to (34). By (12), (30) and (32),

$$\begin{aligned}
E\{E(\hat{g}_a^{(i)}(x)|\underline{x}) - E\hat{g}_a^{(i)}(x)\}^2 &= E\{[E(\hat{g}_a^{(i)}(x)|\underline{x}) - E\hat{g}_a^{(i)}(x)]1_{\Omega_\delta}(\underline{x})\}^2 \\
&\quad +E\{[E(\hat{g}_a^{(i)}(x)|\underline{x}) - E\hat{g}_a^{(i)}(x)]1_{\Omega_\delta^c}(\underline{x})\}^2 \\
&\le o(h^{2(m-i)})p_\delta + (1-p_\delta)\{\max_{\underline{x}}|2\,E(\hat{g}_a^{(i)}(x)|\underline{x})|\}^2 \\
&\le o(h^{2(m-i)})p_\delta + (1-p_\delta)O(n^{2i+6}),
\end{aligned}$$

and (34) follows.

Acknowledgements

We would like to thank the referees for their valuable comments and suggestions. This research was supported in part by NSF Grant DMS-9802358.

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.

Barrow, D. L. and Smith, P.W. (1979). Efficient L_2 approximation by splines. *Numer. Math.* **33**, 101-114.

Chu, C. K. and Marron, J. S. (1991). Choosing a kernel regression estimator (with discussion). *Statist. Sci.* **74**, 829-836.

de Boor, C. (1972). On calculating with B-splines. *J. Approx. Theor.* **6**, 50-62.

Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* **88**, 1287-1301.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1-141.

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

Gaenssler, P. and Wellner, J. A. (1981). Glivenko-Cantelli Theorems. *Encyclopedia of Statistical Science* **3**, 442-445.

Gasser, Th. and Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11**, 171-185.

Ghizzetti, A. and Ossicini, A. (1970). *Quadrature Formulae*. Academic Press, New York.

Müller, H. G. (1988). *Nonparametric Analysis of Longitudinal Data*. Springer, Berlin.

Ortega, J. M. (1987). *Matrix Theory*. Plenum Press, New York.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90**, 1257-1270.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346-1370.

Sacks, J. and Ylvisaker, D. (1970). Designs for regression problems with correlated errors III. *Ann. Math. Statist.* **41**, 2057-2074.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

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

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

Stone, C. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22**, 118-184.

Stone, C. and Koo, C. Y. (1986). Additive splines in statistics. In *Proceedings of the Statistical Computing Section*, 45-48. Amer. Statist. Assoc., Washington, DC.

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

Department of Statistics, 101 Dickens Hall, Kansas State University, Manhattan, KS 66506, U.S.A.

E-mail: zhou@stat.ksu.edu

Department of Statistics, The Ohio State University.

E-mail: daw@stat.osu.edu

(Received April 1997; accepted February 1999)