Ann Inst Stat Math (2009) 61:663–690
DOI 10.1007/s10463-007-0157-x

Efficient and fast spline-backfitted kernel smoothing of additive models

Jing Wang · Lijian Yang

Received: 2 April 2007 / Revised: 20 June 2007 / Published online: 2 October 2007

Abstract A great deal of effort has been devoted to the inference of additive model in the last decade. Among existing procedures, the kernel type are too costly to implement for high dimensions or large sample sizes, while the spline type provide no asymptotic distribution or uniform convergence. We propose a one step backfitting estimator of the component function in an additive regression model, using spline estimators in the first stage followed by kernel/local linear estimators. Under weak conditions, the proposed estimator's pointwise distribution is asymptotically equivalent to an univariate kernel/local linear estimator, hence the dimension is effectively reduced to one at any point. This dimension reduction holds uniformly over an interval under assumptions of normal errors. Monte Carlo evidence supports the asymptotic results for dimensions ranging from low to very high, and sample sizes ranging from moderate to large. The proposed confidence band is applied to the Boston housing data for linearity diagnosis.

Keywords Bandwidths · B spline · Knots · Local linear estimator · Nadaraya-Watson estimator · Nonparametric regression

Supported in part by NSF awards DMS 0405330, 0706518, BCS 0308420 and SES 0127722.

J. Wang (✉)
Department of Mathematics, Statistics, and Computer Science,
University of Illinois at Chicago, Chicago, IL 60607, USA
e-mail: wangjing@math.uic.edu

L. Yang
Department of Statistics and Probability, Michigan State University,
East Lansing, MI 48824, USA
e-mail: yang@stt.msu.edu

1 Introduction

In the last decade, a great deal of effort has been devoted to the inference of additive model, popularized by the book of Hastie and Tibshirani (1990)

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon, \quad \mathbf{X} = (X_1, \ldots, X_d), \quad m(\mathbf{x}) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha), \qquad (1)$$

where the noise satisfies $E(\varepsilon|X) = 0$, $\text{var}(\varepsilon|X) = 1$ and the component functions satisfy the identification conditions $Em_\alpha(X_\alpha) \equiv 0, \alpha = 1, \ldots, d$. In addition, we assume that the predictor X_α is distributed on a compact interval $[a_\alpha, b_\alpha], \alpha = 1, \ldots, d$. Given an i.i.d. sample $\left\{Y_i, \mathbf{X}_i^T\right\}_{i=1}^n = \{Y_i, X_{i1}, \ldots, X_{id}\}_{i=1}^n$ following model (1), Stone (1985) proposed estimators for $\{m_\alpha(x_\alpha)\}_{\alpha=1}^d$ with optimal rates of convergence. These were later called polynomial spline estimators in Stone (1994), Huang (1998, 2003). Huang and Yang (2004) and Xue and Yang (2006a) further extended these estimators to weakly dependent data. Hastie and Tibshirani (1990) proposed backfitting estimators for functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^d$ without theoretical justifications, while Opsomer and Ruppert (1997) offered partial asymptotic results for the case of $d = 2$ under some strong assumptions. Opsomer (2000) extended the theoretical results to a general case with more than two covariates. Mammen et al. (1999) proposed a modified backfitting algorithm with nice theoretical properties, which was implemented in Nielsen and Sperlich (2005) and called smooth backfitting estimator. Another alternative is the marginal integration method, first proposed in Tjøstheim and Auestad (1994), Linton and Nielsen (1995), Linton and Härdle (1996), and further developed by Fan et al. (1998), Yang et al. (1999, 2003, 2006), Sperlich et al. (2002), and Xue and Yang (2006b). Using the wavelet transformation, Härdle et al. (2001) developed the additivity and the polynomial structural tests. Series estimator in Andrews and Whang (1990) circumvented the curse of dimensionality when interactions are present in the model.

If the last $d - 1$ of the component functions were known by "oracle", one could define a new variable $Y_1 = Y - c - \sum_{\alpha=2}^d m_\alpha(X_\alpha) = m_1(X_1) + \sigma(\mathbf{X})\varepsilon$ and use it to regress on the numerical variable X_1 to estimate the only unknown function $m_1(x_1)$, without the "curse of dimensionality". The basic idea of Linton (1997) was to obtain an approximation to the variable Y_1 by substituting $m_\alpha(X_\alpha), \alpha = 2, \ldots, d$ with the marginal integration pilot estimates (kernel-based) and establishing that the error caused by this "cheating" is negligible. Such two-step estimation idea also later appeared in Fan and Chen (1999) for local quasi-likelihood estimation. It is well known that the kernel estimation in high dimension would be extremely computationally intensive. Kim et al. (1999) provided an computationally efficient two-step estimator, reducing computation by order n compared with marginal integration. The spline method, on the other hand, is very fast, but the rate of convergence is only established in mean squares sense, and there is no pointwise confidence interval or even consistency in additive models.

In this paper we propose to pre-estimate the functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^d$ by an undersmoothed constant spline procedure. These estimates are then used to construct the

"oracle" estimator as if they were the true functions. Our approach is much faster than that of Linton (1997), and can be applied to extremely high dimensional data (e.g., $d = 50, 100$). Our approach marries the traditionally parallel spline and kernel smoothing techniques, keeping the asymptotically normal distribution of kernel estimator, without its computational burden. Figuratively speaking, spline smoothing is a sledge-hammer capable of breaking a huge chunk of material (i.e., a regression problem from data of very high dimension and very large sample size), in one slam (i.e., solving one linear least squares problem), but does not guarantee the fine shapes of the broken pieces (i.e., the estimates may not converge at a point or uniformly over an interval). Kernel smoothing works like a sharp knife that cuts anything into pieces of precise shapes (i.e., normal confidence intervals at any point, and confidence bands over compact intervals), but is too tedious to use for a large chunk of material (i.e., the computation cost is intolerable when dimension is high and/or sample size is large). Our new tool is a hammer-knife that first slams a huge clump into smaller pieces (univariate regression problems) in one hit (the spline step), then cuts each small piece into an exact shape (univariate kernel smoothing). The method we propose therefore combines the best of both spline and kernel methods.

The success of our method lies in the well-known "reducing bias by undersmoothing" and "averaging out the variance" principles, see Propositions 1, 2 and 3. Both goals are accomplished with the joint asymptotics of kernel and spline functions, a new feature of our proofs, see Lemmas 3, 4 and 5 in Sect. A.3. Similar idea has appeared in Horowitz and Mammen (2004) and Horowitz et al. (2006), which essentially have used series estimators in the first step and kernel second step in their theory.

In addition to the above, uniform confidence bands are provided for all component function estimates. Literature on nonparametric confidence bands has been scarce, especially in multivariate setting. For univariate kernel smoothing, Hall and Titterington (1988), Härdle (1989), and Xia (1998) made significant contributions, based on strong approximation results as in Tusnády (1977), which is the same idea used in Bickel and Rosenblatt (1973) for confidence band of probability density function. More recently, Claeskens and Van Keilegom (2003) improved upon Xia (1998) by using smoothed bootstrap, while Härdle et al. (2004) introduced the bootstrap bands with corrected bias. For univariate spline smoothing under general setting, Wang and Yang (2007a) provided simple solutions with asymptotic theory. Bootstrap confidence band has been constructed for additive regression model in Yang (2007). However, it seems that this present paper is the one of the few to offer the measure of uniform accuracy with theoretical justifications. The confidence band we provide for $m_\alpha(x_\alpha), \alpha = 1, \ldots, d$ is asymptotically the same that Härdle (1989) established for univariate kernel regression, regardless what other functions $\{m_\beta(x_\beta)\}_{1\le\beta\le d,\beta\neq\alpha}$ are. Hence neither the dimension d nor other function components play any role in forming the band for $m_\alpha(x_\alpha)$, at least asymptotically. In this sense, our estimator of $m_\alpha(x_\alpha)$ possesses "uniform oracle efficiency", which is much stronger than the "pointwise oracle efficiency" of Linton (1997). Furthermore, components in directions not of interests are only required to be Lipschitz continuous, allowing the broadest class of additive model compared to existing methods, see Remark 3 in Sect. 2.

The paper is organized as follows. In Sect. 2 we introduce the spline-backfitted kernel/local linear estimator, and state their asymptotic "oracle efficiency" under appropriate assumptions. In Sect. 3 we provide insights into the proofs of the main theoretical results. Section 4 presents Monte Carlo results to demonstrate that the proposed spline-backfitted local linear estimator (SBLL) possesses the claimed asymptotic properties. The simulated examples cover a wide range of sample sizes, dependence structure and dimensions. The SBLL estimator is applied to the Boston housing data in Sect. 5. Section 6 concludes, and all technical proofs are in the Appendix.

2 The SBK and SBLL estimators

In this section, we describe the spline-backfitted kernel estimation procedure. Let $\left\{Y_i, \mathbf{X}_i^T\right\}_{i=1}^n = \{Y_i, X_{i1}, \ldots, X_{id}\}_{i=1}^n$ be an i.i.d. sample following model (1). In what follows, we write all responses as $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, and denote by $\mathbf{X}$ the design matrix $(\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$. Without loss of generality, we take all intervals $[a_\alpha, b_\alpha] = [0, 1], \alpha = 1, \ldots, d$. We pre select an integer $N_n \sim n^{2/5} \log(n)$, see Assumption (A6) below. Define for any $\alpha = 1, \ldots, d$, the indicator function $I_{J,\alpha}(x_\alpha)$ of the equally-spaced subintervals of the finite interval [0, 1], i.e. for any $J = 0, 1, \ldots, N$,

$$I_{J,\alpha}(x_\alpha) = \begin{cases} 1 & JH \le x_\alpha < (J+1)H, \quad H = H_n = (N_n+1)^{-1}. \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

Define the $(1 + dN)$-dimensional space G of additive spline functions as the linear space spanned by $\left\{1, I_{J,\alpha}(x_\alpha), \alpha = 1, \ldots, d, J = 1, \ldots, N\right\}$, while denote by G_n the subspace of R^n spanned by $\left\{\{1\}_{i=1}^n, \left\{I_{J,\alpha}(X_{i\alpha})\right\}_{i=1}^n, \alpha = 1, \ldots, d, J = 1, \ldots, N\right\}$. As $n \to \infty$, the dimension of G_n becomes $1 + dN$ with probability approaching one.

The spline estimator of additive function $m(\mathbf{x})$ is the unique element $\hat{m}(\mathbf{x}) = \hat{m}_n(\mathbf{x})$ from the space G so that the vector $\left\{\hat{m}(\mathbf{X}_1), \ldots, \hat{m}(\mathbf{X}_n)\right\}^T$ best approximates the response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. To be precise,

$$\hat{m}(\mathbf{x}) = \hat{\lambda}_0 + \sum_{\alpha=1}^{d} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(x_\alpha), \tag{3}$$

where the coefficients $\hat{\lambda}_0, \hat{\lambda}_{1,1}, \ldots, \hat{\lambda}_{N,d}$ are the solution of the following least squares problem

$$\left\{\hat{\lambda}_0, \hat{\lambda}_{1,1}, \ldots, \hat{\lambda}_{N,d}\right\}^T = \underset{R^{dN+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{Y_i - \lambda_0 - \sum_{\alpha=1}^{d} \sum_{J=1}^{N} \lambda_{J,\alpha} I_{J,\alpha}(X_{i\alpha})\right\}^2. \tag{4}$$

Pilot estimators of component functions $m_\alpha(x_\alpha)$ and the constant c are

$$\hat{m}_\alpha(x_\alpha) = \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(x_\alpha) - n^{-1} \sum_{i=1}^{n} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(X_{i\alpha}),$$
$$\hat{m}_c = \hat{\lambda}_0 + n^{-1} \sum_{\alpha=1}^{d} \sum_{i=1}^{n} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} I_{J,\alpha}(X_{i\alpha}). \tag{5}$$

These pilot estimators are then used to define a set of new pseudo-responses $\hat{Y}_{i1}$ which are estimated versions of the unobservable "oracle" responses Y_{i1},

$$\hat{Y}_{i1} = Y_i - \hat{c} - \sum_{\alpha=2}^{d} \hat{m}_\alpha(X_{i\alpha}), \quad Y_{i1} = Y_i - c - \sum_{\alpha=2}^{d} m_\alpha(X_{i\alpha}), \ i = 1, \ldots, n \tag{6}$$

where $\hat{c} = n^{-1} \sum_{i=1}^{n} Y_i$. By Central Limit Theorem $\hat{c}$ is a $\sqrt{n}$-consistent estimator of c. Define the spline-backfitted kernel (SBK) estimator of $m_1(x_1)$ based on $\left\{\hat{Y}_{i1}, X_{i1}\right\}_{i=1}^{n}$ as $\hat{m}_{\mathrm{SBK},1}(x_1)$, which is an attempt to mimic the would-be kernel estimator $\tilde{m}_{\mathrm{K},1}(x_1)$ of $m_1(x_1)$ based on $\{Y_{i1}, X_{i1}\}_{i=1}^{n}$, had the unobservable "oracle" responses $\{Y_{i1}\}_{i=1}^{n}$ been available, i.e.

$$\hat{m}_{\mathrm{SBK},1}(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i1} - x_1)\, \hat{Y}_{i1}}{\sum_{i=1}^{n} K_h(X_{i1} - x_1)}, \quad \tilde{m}_{\mathrm{K},1}(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i1} - x_1)\, Y_{i1}}{\sum_{i=1}^{n} K_h(X_{i1} - x_1)}, \tag{7}$$

where $\hat{Y}_{i1}$ and Y_{i1} are defined in (6). Similarly, the spline-backfitted local linear (SBLL) estimator $\hat{m}_{\mathrm{SBLL},1}(x_1)$ based on $\left\{\hat{Y}_{i1}, X_{i1}\right\}_{i=1}^{n}$ mimics the would-be local linear estimator $\tilde{m}_{\mathrm{LL},1}(x_1)$ based on $\{Y_{i1}, X_{i1}\}_{i=1}^{n}$

$$\left\{\hat{m}_{\mathrm{SBLL},1}(x_1), \tilde{m}_{\mathrm{LL},1}(x_1)\right\} = (1, 0)\left(\mathbf{Z}^T \mathbf{W} \mathbf{Z}\right)^{-1} \mathbf{Z}^T \mathbf{W} \left(\hat{\mathbf{Y}}_1, \mathbf{Y}_1\right), \tag{8}$$

in which the oracle and pseudo-response vectors are

$$\mathbf{Y}_1 = (Y_{11}, \ldots, Y_{n1})^T, \quad \hat{\mathbf{Y}}_1 = \left(\hat{Y}_{11}, \ldots, \hat{Y}_{n1}\right)^T$$

and the weight and design matrices are

$$\mathbf{W} = \mathrm{diag}\left\{K_h(X_{i1} - x_1)\right\}_{i=1}^{n}, \quad \mathbf{Z}^T = \begin{pmatrix} 1 & \ldots, & 1 \\ X_{11} - x_1 & \ldots, & X_{n1} - x_1 \end{pmatrix}.$$

Throughout this paper, for a function $g \in L_2[0, 1]$, denote $c^2(g) = \int g^2(u)\, du$. Second order smooth function space is defined by $C^{(2)}[0, 1] = \left\{g \mid g'' \in C[0, 1]\right\}$,

while the Lipschitz continuous function class is defined by

$$\text{Lip}\left([0,1],C\right)=\left\{m||m(x)-m(x')|\leq C|x-x'|,\forall x,x'\in[0,1]\right\}.$$

Before presenting the main theoretical results, we state the assumptions.

(A1) The component function $m_1 \in C^{(2)}[0,1]$, while all components $m_\beta \in \text{Lip}\left([0,1],C_\infty\right)$, $\forall \beta = 1,\ldots,d$, $0 < C_\infty < \infty$.

(A2) The noise ε_i given $\mathbf{X}_i$ are i. i. d. with mean 0 and variance 1, for $i = 1,\ldots,n$, while the conditional standard deviation function $\sigma(\mathbf{x})$ is continuous on $[0,1]^d$. Denote $C_\sigma = \max_{\mathbf{x}\in[0,1]^d}\sigma(\mathbf{x})$.

(A2′) The conditional distribution of noise $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_n)$ given $\mathbf{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_n)^T$ is n-dimensional standard normal.

(A3) The density function $f(\mathbf{x})$ of $\mathbf{X}$ is continuous and bounded away from zero and infinity on $[0,1]^d$, i.e.

$$0 < c_f \leq \inf_{\mathbf{x}\in[0,1]^d}\{f(\mathbf{x})\} \leq \sup_{\mathbf{x}\in[0,1]^d}\{f(\mathbf{x})\} \leq C_f < \infty.$$

(A4) The kernel function $K \in \text{Lip}\left([-1,1],C_K\right)$, for some constant $C_K > 0$, is a symmetric probability density function supported on $[-1,1]$ with the second moment $\mu_2(K) = \int u^2 K(u)\,\mathrm{d}u$.

(A5) The bandwidth h of the kernel K is assumed to be of order $n^{-1/5}$, i.e., $c_h n^{-1/5} \leq h \leq C_h n^{-1/5}$ for some positive constants c_h, C_h.

(A6) The number of interior knots $N_n \sim n^{2/5}\log(n)$, i.e., $c_N n^{2/5}\log(n) \leq N_n \leq C_N n^{2/5}\log(n)$ for some positive constants c_N, C_N, and the interval width $H = (N_n+1)^{-1}$.

(A7) The marginal density $f_1(x_1)$ of X_1 has continuous derivative on $[0,1]$.

Asymptotic properties of smoothers $\tilde{m}_{\text{K},1}(x_1)$ and $\tilde{m}_{\text{LL},1}(x_1)$ are well-developed. Under Assumptions (A1)–(A5), according to Theorem 4.2.1 of Härdle (1990), one has for any $x_1 \in [h, 1-h]$,

$$\sqrt{nh}\left\{\tilde{m}_{\text{K},1}(x_1) - m_1(x_1) - b_{\text{K}}(x_1)h^2\right\} \xrightarrow{D} N\left(0, v^2(x_1)\right),$$
$$\sqrt{nh}\left\{\tilde{m}_{\text{LL},1}(x_1) - m_1(x_1) - b_{\text{LL}}(x_1)h^2\right\} \xrightarrow{D} N\left(0, v^2(x_1)\right),$$

where

$$\begin{aligned} b_{\text{K}}(x_1) &= \mu_2(K)\left\{m_1''(x_1)f_1(x_1)/2 + m_1'(x_1)f_1'(x_1)\right\}f_1^{-1}(x_1), \\ b_{\text{LL}}(x_1) &= \mu_2(K)\,m_1''(x_1)/2, \\ v^2(x_1) &= \|K\|_2^2\, E\left\{\sigma^2(x_1,X_2,\ldots,X_d)\right\}f_1^{-1}(x_1) \end{aligned} \tag{9}$$

with the equation for $\tilde{m}_{\text{K},1}(x_1)$ requiring the additional assumption (A7). The next two theorems state that the asymptotic magnitude of difference between $\hat{m}_{\text{SBK},1}(x_1)$ and $\tilde{m}_{\text{K},1}(x_1)$ is of order $o_p\left(n^{-2/5}\right)$, both pointwise and uniformly, which is dominated by the asymptotic size of $\tilde{m}_{\text{K},1}(x_1) - m_1(x_1)$. Hence $\hat{m}_{\text{SBK},1}(x_1)$ will have the same asymptotic distribution as $\tilde{m}_{\text{K},1}(x_1)$. The same is true for $\hat{m}_{\text{SBLL},1}(x_1)$ and $\tilde{m}_{\text{LL},1}(x_1)$.

Theorem 1 *Under Assumptions* (A1) *to* (A6)*, for any* $x_1 \in [0,1]$*, the estimators* $\hat{m}_{\mathrm{SBK},1}(x_1)$ *and* $\hat{m}_{\mathrm{SBLL},1}(x_1)$ *given in* (7) *and* (8) *satisfy*

$$\left|\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{\mathrm{K},1}(x_1)\right| + \left|\hat{m}_{\mathrm{SBLL},1}(x_1) - \tilde{m}_{\mathrm{LL},1}(x_1)\right| = o_p\left(n^{-2/5}\right).$$

Hence with $b_{\mathrm{K}}(x_1)$, $b_{\mathrm{LL}}(x_1)$ *and* $v^2(x_1)$ *defined in* (9)*, for any* $x_1 \in [h, 1-h]$

$$\sqrt{nh}\left\{\hat{m}_{\mathrm{SBLL},1}(x_1) - m_1(x_1) - b_{\mathrm{LL}}(x_1)h^2\right\} \xrightarrow{D} N\left(0, v^2(x_1)\right),$$

and with the additional assumption (A7) *we have*

$$\sqrt{nh}\left\{\hat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1) - b_{\mathrm{K}}(x_1)h^2\right\} \xrightarrow{D} N\left(0, v^2(x_1)\right).$$

Theorem 2 *Under Assumptions* (A1) *to* (A6) *and* (A2′)*, estimator* $\hat{m}_{\mathrm{SBK},1}(x_1)$ *given in* (7) *and* $\hat{m}_{\mathrm{SBLL},1}(x_1)$ *in* (8) *satisfy*

$$\sup_{x_1\in[0,1]}\left|\hat{m}_{\mathrm{SBK},1}(x_1) - \tilde{m}_{\mathrm{K},1}(x_1)\right| + \left|\hat{m}_{\mathrm{SBLL},1}(x_1) - \tilde{m}_{\mathrm{LL},1}(x_1)\right| = o_p\left(n^{-2/5}\right).$$

Hence for any z

$$\lim_{n\to\infty} P\left[\left\{\log\left(h^{-2}\right)\right\}^{1/2}\left(\sup_{x_1\in[h,1-h]}\frac{\sqrt{nh}}{v(x_1)}\left|\hat{m}_{\mathrm{SBLL},1}(x_1) - m_1(x_1)\right| - d_n\right) < z\right] = \exp\left\{-2\exp(-z)\right\},$$

in which $d_n = \left\{\log\left(h^{-2}\right)\right\}^{1/2} + \left\{\log\left(h^{-2}\right)\right\}^{-1/2}\log\left\{c\left(K'\right)(2\pi)^{-1}c^{-1}(K)\right\}$.

With the additional assumption (A7), it is also true that

$$\lim_{n\to\infty} P\left[\left\{\log\left(h^{-2}\right)\right\}^{1/2}\left(\sup_{x_1\in[h,1-h]}\frac{\sqrt{nh}}{v(x_1)}\left|\hat{m}_{\mathrm{SBK},1}(x_1) - m_1(x_1)\right| - d_n\right) < z\right] = \exp\left\{-2\exp(-z)\right\}.$$

For any $\alpha \in (0,1)$*, an asymptotic* $100(1-\alpha)\%$ *confidence band for* $m_1(x_1)$ *over interval* $[h, 1-h]$ *is*

$$\hat{m}_{\mathrm{SBLL},1}(x_1) \pm v(x_1)(nh)^{-1/2}\left[d_n - \log^{-1/2}\left(h^{-2}\right)\log\left\{-\frac{\log(1-\alpha)}{2}\right\}\right]. \quad (10)$$

Remark 1 Similar estimators $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ and $\hat{m}_{\mathrm{SBLL},\alpha}(x_\alpha)$ can be constructed for $m_\alpha(x_\alpha)$, $2 \le \alpha \le d$ with same oracle properties.

Remark 2 Proofs of Theorems 1 and 2 make it clear that the number of knots can be of the more general form $N_n \sim n^{2/5} N_n'$, where the sequence N_n' satisfies $N_n' \to \infty$, $n^{-\theta} N_n' \to 0$ for all $\theta > 0$, while there is no optimal way to choose N_n'. The fact that $N_n^{-1} = o\left(n^{-2/5}\right)$ ensures that the bias in the spline pilot estimators is negligible compared to the bias of h^2 in the kernel/local linear smoothing stage. On the other hand, one does not allow N_n to be too large for practical reasons: the number of terms in (4), $1 + dN_n$ has to be small relative to n. Hence we select N_n to be of order barely larger than $n^{2/5}$.

Remark 3 Assumption (A1) requires only the Lipschitz continuity for the components except for the component of interest.

3 Decomposition

In this section, we introduce some additional notations in order to shed some light on the ideas behind the proofs of Theorems 1 and 2. Denote by $\|\phi\|_2$ the theoretical L_2 norm of a function ϕ on $[0,1]^d$, $\|\phi\|_2^2 = E\left\{\phi^2(\mathbf{X})\right\} = \int_{[0,1]^d} \phi^2(\mathbf{x}) f(\mathbf{x}) \, d\mathbf{x}$, and the empirical L_2 norm as $\|\phi\|_{2,n}^2 = n^{-1} \sum_{i=1}^n \phi^2(\mathbf{X}_i)$. For any L_2-integrable functions ϕ, φ on $[0,1]^d$, the corresponding inner products are defined by

$$\begin{aligned} \langle \phi, \varphi \rangle_2 &= \int_{[0,1]^d} \phi(\mathbf{x}) \varphi(\mathbf{x}) f(\mathbf{x}) \, d\mathbf{x} = E\{\phi(\mathbf{X}) \varphi(\mathbf{X})\}, \\ \langle \phi, \varphi \rangle_{2,n} &= n^{-1} \sum_{i=1}^n \phi(\mathbf{X}_i) \varphi(\mathbf{X}_i). \end{aligned} \tag{11}$$

A function ϕ on $[0,1]^d$ is called theoretically centered (empirically centered) if $\langle 1, \varphi \rangle_2 = 0$ ($\langle 1, \varphi \rangle_{2,n} = 0$). Define the theoretically centered spline basis

$$b_{J,\alpha}(x_\alpha) = I_{J+1,\alpha}(x_\alpha) - \frac{\left\| I_{J+1,\alpha} \right\|_2}{\left\| I_{J,\alpha} \right\|_2} I_{J,\alpha}(x_\alpha), \quad \forall \alpha = 1, \ldots, d, \ J = 1, \ldots, N, \tag{12}$$

where the functions $I_{J,\alpha}(x_\alpha)$'s defined in (2) are indicators on the subintervals $[JH, (J+1)H)$. The standardized one is given for any $\alpha = 1, \ldots, d$,

$$B_{J,\alpha}(x_\alpha) = \left\| b_{J,\alpha} \right\|_2^{-1} b_{J,\alpha}(x_\alpha), \quad \forall J = 1, \ldots, N. \tag{13}$$

The additive function space G is also spanned by the linearly independent basis $\left\{1, B_{J,\alpha}(x_\alpha), J = 1, \ldots, N, \alpha = 1, \ldots, d\right\}$. These new basis involve unknown quantities and are not computable from the data, but are more convenient for mathematical analysis than the truncated power basis in (2). Similarly G_n is spanned linearly by the basis $\left\{1, \left\{B_{J,\alpha}(X_{i\alpha})\right\}_{i=1}^n, \alpha = 1, \ldots, d, J = 1, \ldots, N\right\}$.

The n-dimensional vector, $\hat{m}(\mathbf{X}) = \left\{\hat{m}(\mathbf{X}_1), \ldots, \hat{m}(\mathbf{X}_n)\right\}^T$, is the projection of $\mathbf{Y}$ on the space G_n relative to the empirical inner product $\langle \cdot, \cdot \rangle_{2,n}$. In general,

for any n-dimensional vector $\mathbf{V} = \{V_1, \ldots, V_n\}^T$, we define $\mathbf{P}_n\mathbf{V}(\mathbf{x})$ as the spline function constructed from the projection of $\mathbf{V}$ on the inner product space $\left(G_n, \langle\cdot,\cdot\rangle_{2,n}\right)$, i.e., $\mathbf{P}_n\mathbf{V}(\mathbf{x}) = \hat{v}_0 + \sum_{\alpha=1}^{d}\sum_{J=1}^{N}\hat{v}_{J,\alpha}B_{J,\alpha}(x_\alpha)$, with the least squares coefficients obtained by

$$\left\{\hat{v}_0, \hat{v}_{1,1}, \ldots, \hat{v}_{N,d}\right\}^T = \underset{R^{dN+1}}{\operatorname{argmin}} \sum_{i=1}^{n}\left\{V_i - v_0 - \sum_{\alpha=1}^{d}\sum_{J=1}^{N} v_{J,\alpha}B_{J,\alpha}(X_{i\alpha})\right\}^2,$$

which is similar to (3) and (4) except the basis. Next, the multivariate function $\mathbf{P}_n\mathbf{V}(\mathbf{x})$ is decomposed into direct component $\mathbf{P}^*_{n,\alpha}\mathbf{V}(x_\alpha) = \sum_{J=1}^{N}\hat{v}_{J,\alpha}B_{J,\alpha}(x_\alpha)$. In order to consider the identification condition, the pilot estimator in this stage is chosen to be the empirically centered additive components $\mathbf{P}_{n,\alpha}\mathbf{V}(x_\alpha), \alpha = 1, \ldots, d$ and the constant component $\mathbf{P}_{n,c}\mathbf{V}$

$$\mathbf{P}_{n,\alpha}\mathbf{V}(x_\alpha) = \sum_{J=1}^{N}\hat{v}_{J,\alpha}B_{J,\alpha}(x_\alpha) - n^{-1}\sum_{i=1}^{n}\sum_{J=1}^{N}\hat{v}_{J,\alpha}B_{J,\alpha}(X_{i\alpha}), \tag{14}$$

$$\mathbf{P}_{n,c}\mathbf{V} = \hat{v}_0 + n^{-1}\sum_{\alpha=1}^{d}\sum_{i=1}^{n}\sum_{J=1}^{N}\hat{v}_{J,\alpha}B_{J,\alpha}(x_\alpha), \tag{15}$$

in which the centering procedure is the same as (5).

With these new notations, we can rewrite the constant spline estimators $\hat{m}(\mathbf{x}), \hat{m}_\alpha(x_\alpha), \hat{m}_c$ defined in (3) and (5) as

$$\hat{m}(\mathbf{x}) = \mathbf{P}_n\mathbf{Y}(\mathbf{x}), \quad \hat{m}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha}\mathbf{Y}(x_\alpha), \quad \hat{m}_c = \mathbf{P}_{n,c}\mathbf{Y}.$$

Based on the relation $\mathbf{Y} = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon = m(\mathbf{X}) + \mathbf{E}$, with noise vector $\mathbf{E} = \{\sigma(\mathbf{X}_i)\varepsilon_i\}_{i=1}^n$, one defines similarly noiseless spline smoothers

$$\tilde{m}(\mathbf{x}) = \mathbf{P}_n\{m(\mathbf{X})\}(\mathbf{x}), \tilde{m}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha}\{m(\mathbf{X})\}(x_\alpha), \tilde{m}_c = \mathbf{P}_{n,c}\{m(\mathbf{X})\},$$

and spline components of the noise

$$\tilde{\varepsilon}(\mathbf{x}) = \mathbf{P}_n\mathbf{E}(\mathbf{x}), \quad \tilde{\varepsilon}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha}\mathbf{E}(x_\alpha), \quad \tilde{\varepsilon}_c = \mathbf{P}_{n,c}\mathbf{E}. \tag{16}$$

Due to the linearity of operators $\mathbf{P}_n, \mathbf{P}_{n,\alpha}, \mathbf{P}_{n,c}, \alpha = 1, \ldots, d$, one has the following decomposition, which is crucial to prove Theorems 1 and 2

$$\hat{m}(\mathbf{x}) = \tilde{m}(\mathbf{x}) + \tilde{\varepsilon}(\mathbf{x}), \quad \hat{m}_\alpha(x_\alpha) = \tilde{m}_\alpha(x_\alpha) + \tilde{\varepsilon}_\alpha(x_\alpha), \quad \hat{m}_c = \tilde{m}_c + \tilde{\varepsilon}_c, \tag{17}$$

for $\alpha = 1, \ldots, d$.

As closer examination is needed later for $\tilde{\varepsilon}(\mathbf{x})$ and $\tilde{\varepsilon}_\alpha(x_\alpha)$, one defines vector $\tilde{\mathbf{a}} = \left\{\tilde{a}_0, \tilde{a}_{1,1}, \ldots, \tilde{a}_{N,d}\right\}^T$ such that

$$\tilde{\mathbf{a}} = \underset{R^{dN+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ \sigma\left(\mathbf{X}_i\right) \varepsilon_i - a_0 - \sum_{\alpha=1}^{d} \sum_{J=1}^{N} a_{J,\alpha} B_{J,\alpha}\left(X_{i\alpha}\right) \right\}^2, \tag{18}$$

then $\tilde{\varepsilon}\left(\mathbf{x}\right)$ in (16) can be rewritten as $\tilde{\mathbf{a}}^T \mathbf{B}\left(\mathbf{x}\right)$, where $\tilde{\mathbf{a}}$ is the solution of equation (18), and matrices $\mathbf{B}\left(\mathbf{x}\right)$ and $\mathbf{B}$ are defined as

$$\mathbf{B}\left(\mathbf{x}\right) = \left\{1, B_{1,1}\left(x_1\right), \ldots, B_{N,d}\left(x_d\right)\right\}^T, \quad \mathbf{B} = \left\{\mathbf{B}\left(\mathbf{X}_1\right), \ldots, \mathbf{B}\left(\mathbf{X}_n\right)\right\}^T. \tag{19}$$

To be specific, the least squares solution of the noise is

$$\begin{aligned} \tilde{\mathbf{a}} &= \left(\mathbf{B}^T \mathbf{B}\right)^{-1} \mathbf{B}^T \mathbf{E} \\ &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \left\langle B_{J,\alpha}, B_{J',\alpha'} \right\rangle_{2,n} \end{pmatrix}^{-1}_{\substack{1 \le \alpha, \alpha' \le d \\ 1 \le J, J' \le N}} \begin{pmatrix} n^{-1} \sum_{i=1}^{n} \sigma\left(\mathbf{X}_i\right) \varepsilon_i \\ n^{-1} \sum_{i=1}^{n} B_{J,\alpha}\left(X_{i\alpha}\right) \sigma\left(\mathbf{X}_i\right) \varepsilon_i \end{pmatrix}_{\substack{1 \le J \le N \\ 1 \le \alpha \le d}} \end{aligned} \tag{20}$$

Our objective is to study the difference between the estimator $\hat{m}_{\mathrm{SBK},1}\left(x_1\right)$ and the "oracle" smoother $\tilde{m}_{K,1}\left(x_1\right)$, and between $\hat{m}_{\mathrm{SBLL},1}\left(x_1\right)$ and $\tilde{m}_{LL,1}\left(x_1\right)$. For notational brevity, we assume without loss of generality that $d = 2$ and we focus on the proof for SBK estimator. Denote the projection matrix $\mathbf{P}_{0_{N+1}, I_N} = \left({}^{0_{N+1}}{}_{I_N}\right)$, and we define another auxiliary entity

$$\tilde{\varepsilon}_2^*\left(x_2\right) = \mathbf{P}_{n,2}^* \mathbf{E}\left(x_2\right) = \tilde{\mathbf{a}}^T \mathbf{P}_{0_{N+1}, I_N} \left(\mathbf{B}\left(\mathbf{x}\right)\right)^T = \sum_{J=1}^{N} \tilde{a}_{J,2} B_{J,2}\left(x_2\right),$$

which, in particular, implies that

$$\tilde{\varepsilon}_2^*\left(X_{i2}\right) = \tilde{\mathbf{a}}^T \mathbf{P}_{0_{N+1}, I_N} \left(e_i^T \mathbf{B}\right)^T = \sum_{J=1}^{N} \tilde{a}_{J,2} B_{J,2}\left(X_{i2}\right), \tag{21}$$

in which e_i is the n-dimensional unit vector with ith element 1 and else 0. Hence the ith row of matrix $\mathbf{B}$, $e_i^T \mathbf{B} = \mathbf{B}\left(\mathbf{X}_i\right)$, is the basis functions corresponding to the ith observation $\mathbf{X}_i$. Definitions (14) and (15) imply that $\tilde{\varepsilon}_2\left(x_2\right)$ is simply the empirical centering of $\tilde{\varepsilon}_2^*\left(x_2\right)$, i.e., $\tilde{\varepsilon}_2\left(x_2\right) \equiv \tilde{\varepsilon}_2^*\left(x_2\right) - n^{-1} \sum_{i=1}^{n} \tilde{\varepsilon}_2^*\left(X_{i2}\right)$, typically

$$\tilde{\varepsilon}_2\left(X_{i2}\right) = \sum_{J=1}^{N} \tilde{a}_{J,2} B_{J,2}\left(X_{i2}\right) - n^{-1} \sum_{i=1}^{n} \sum_{J=1}^{N} \tilde{a}_{J,2} B_{J,2}\left(X_{i2}\right). \tag{22}$$

Making use of the signal noise decomposition (17), the difference $\tilde{m}_{\mathrm{K},1}\left(x_1\right) - \hat{m}_{\mathrm{SBK},1}\left(x_1\right) + \hat{c} - c$ can be treated as the sum of two terms

$$\frac{n^{-1}\sum_{i=1}^{n} K_h(X_{i1}-x_1)\{\hat{m}_2(X_{i2})-m_2(X_{i2})\}}{n^{-1}\sum_{i=1}^{n} K_h(X_{i1}-x_1)} = \frac{I(x_1)+II(x_1)}{n^{-1}\sum_{i=1}^{n} K_h(X_{i1}-x_1)}, \tag{23}$$

where

$$I(x_1) = n^{-1}\sum_{i=1}^{n} K_h(X_{i1}-x_1)\cdot\tilde{\varepsilon}_2(X_{i2}), \tag{24}$$

$$II(x_1) = n^{-1}\sum_{i=1}^{n} K_h(X_{i1}-x_1)\cdot\{\tilde{m}_2(X_{i2})-m_2(X_{i2})\}. \tag{25}$$

The term $I(x_1)$ relates to the noise terms $\tilde{\varepsilon}_2(X_{i2})$, while $II(x_1)$ the bias terms $\tilde{m}_2(X_{i2})-m_2(X_{i2})$. Propositions 1 and 2 below, both proved in Sect. A.1, bound $I(x_1)$, while Proposition 3, proved in Sect. A.2, bounds $II(x_1)$. Standard theory of kernel smoothing ensures that the denominator term in (23), $n^{-1}\sum_{i=1}^{n} K_h(X_{i1}-x_1)$, has a positive lower bound for $x_1 \in [0,1]$. The additional term $\hat{c}-c$ is of clearly order $O\left(n^{-1/2}\right)$ and thus $o_p\left(n^{-2/5}\right)$. Hence both Theorems 1 and 2 follow from Propositions 1, 2 and 3. The Appendix is devoted to the proofs of these propositions, rather than Theorems 1 and 2. If one were to prove the corresponding results for SBLL estimator, one would need to extend Propositions 1 and 2 to include also the term $n^{-1}\sum_{i=1}^{n} K_h(X_{i1}-x_1)\left(\frac{X_{i1}-x_1}{h}\right)\cdot\tilde{\varepsilon}_2(X_{i2})$, and Proposition 3 to include also the term $n^{-1}\sum_{i=1}^{n} K_h(X_{i1}-x_1)\left(\frac{X_{i1}-x_1}{h}\right)\cdot\{\tilde{m}_2(X_{i2})-m_2(X_{i2})\}$. These do not add a great deal of difficulty.

Proposition 1 *Under Assumptions* (A1) *to* (A6)*, for any* $x_1 \in [0,1]$

$$|I(x_1)| = O_p\left(n^{-1/2}\right) = o_p\left(n^{-2/5}\right).$$

Proposition 2 *Under Assumptions* (A1) *to* (A6) *and* (A2′)

$$\sup_{x_1\in[0,1]} |I(x_1)| = O_p\left(n^{-1/2}\{\log n\}^{1/2}\right) = o_p\left(n^{-2/5}\right).$$

Proposition 3 *Under Assumptions* (A1)*, and* (A3) *to* (A6)

$$\sup_{x_1\in[0,1]} |II(x_1)| = O_p\left(n^{-1/2}+H\right) = o_p\left(n^{-2/5}\right).$$

4 Simulation results

In this section, we present simulated results on the finite-sample behavior of the SBLL estimators $\hat{m}_{\mathrm{SBLL},\alpha}(x_\alpha)$, $\alpha = 1,\ldots d$. The SBK estimator is not implemented as it is inferior to the SBLL estimator, see Fan and Gijbels (1996).

The data set is generated from the regression model $Y = \sum_{\alpha=1}^{d} m_\alpha(X_\alpha)+\sigma(\mathbf{X})\cdot\varepsilon$. The additive elements are assumed to be $m_\alpha(x_\alpha) = \sin(2\pi x_\alpha)$, $\forall\alpha = 1,\ldots,d$. The

predictors X_α are $X_\alpha = 2.5 * \{\Phi(Z_\alpha) - 0.5\}$, where Φ is the distribution function of the variable $Z_\alpha \sim N(0,1)$, $\alpha = 1, \ldots, d$ with the correlation $\rho_{\alpha\beta} = \rho$, $\alpha \neq \beta$ for any pair of Z's. Note that the dependence among the X's increases with ρ. To ensure that the density is bounded below from 0, estimation is carried out only at data points $\mathbf{X}_i$, $i = 1, \ldots, n$, which satisfy $\max_{1 \le \alpha \le d} |X_{i\alpha}| \le 1$, following Nielsen and Sperlich (2005).

Meanwhile, the error term $\varepsilon \sim N(0,1)$ and is independent of $\mathbf{X}$. The conditional standard deviation function is defined by

$$\sigma(\mathbf{x}) = \frac{\sqrt{d}}{2} \cdot \frac{100 - \exp\left\{\sum_{\alpha=1}^{d} |x_\alpha| / d\right\}}{100 + \exp\left\{\sum_{\alpha=1}^{d} |x_\alpha| / d\right\}}.$$

This choice of $\sigma(\mathbf{x})$ ensures design heteroscedasticity, with variance roughly proportional to dimension d. This proportionality mimics the case when independent copies of the same univariate regression problems are added together.

To implement the SBLL estimator, one first obtain the spline estimator of $\sum_{\alpha=1}^{d} m_\alpha(X_\alpha)$, using the truncated power B-spline basis as in (4). The number of knots N_n is

$$N_n = \min\left(\left[c_1 n^{2/5} \log n\right] + c_2, \left[(n/4 - 1) d^{-1}\right]\right), \tag{26}$$

in which $[\cdot]$ denotes the integer part, and c_1, c_2 are tuning constants. The choice of these constants c_1 and c_2 makes little difference for a large sample, while for small sample size, it does affect the performance to a degree. In our simulation study, we have used $c_1 = 1 = c_2$. The additional constraint, $N \le (n/4 - 1) d^{-1}$, ensures that the number of terms in the linear least squares problem (4), $1 + dN_n$, is no greater than $n/4$, which is necessary when the sample size n is moderate and dimension d is high.

The oracle smoother $\tilde{m}_{\mathrm{LL},\alpha}(x_\alpha)$ for comparison is obtained by local linear regression of the unobservable $m_\alpha(X_\alpha) + \sigma(\mathbf{X})\varepsilon$ on X_α directly, while the oracle SBLL estimators $\hat{m}_{\mathrm{SBLL},\alpha}(x_\alpha)$ are obtained by local linear regression of $\left\{\hat{Y}_{i\alpha}, X_{i\alpha}\right\}_{i=1}^{n}$, using the XploRe quantlet "lpregxest" with the rule-of-thumb (ROT) bandwidth of Fan and Gijbels (1996). For information on XploRe, see Härdle et al. (2000) or visit http://www.xplore-stat.de.

We have run $S = 500$ replications for sample sizes $n = 100, 200, 500$ and $1,000$ with $\rho = 0, 0.3, 0.9$ respectively. The dimensions are taken at $d = 4, 10$. Denote by $\left\{Y_i, X_{i1,l}, \ldots, X_{id,l}\right\}_{i=1}^{n}$ the lth sample, $1 \le l \le S$. The main objective is to compare the relative efficiency of $\tilde{m}_{\mathrm{LL},\alpha}$ with respect to $\hat{m}_{\mathrm{SBLL},\alpha}$,

$$\mathrm{eff}_{\alpha,l} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{m}_{\mathrm{LL},\alpha}\left(X_{i\alpha,l}\right) - m_\alpha\left(X_{i\alpha,l}\right)\right\}^2 I_{\{|X_{i\alpha,l}| \le 1\}}}{\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}_{\mathrm{SBLL},\alpha}\left(X_{i\alpha,l}\right) - m_\alpha\left(X_{i\alpha,l}\right)\right\}^2 I_{\{|X_{i\alpha,l}| \le 1\}}}$$

$$\mathrm{eff}_\alpha = \frac{1}{S}\sum_{l=1}^{S} \mathrm{eff}_{\alpha,l},\ 1 \le \alpha \le d,$$

Table 1 Relative efficiency of $\tilde{m}_{\mathrm{LL},1}$ against $\hat{m}_{\mathrm{SBLL},1}$ with 500 replications

		eff_1		
d	n	$\rho = 0$	$\rho = 0.3$	$\rho = 0.9$
	100	1.015 (0.287)	0.958 (0.320)	0.731 (0.458)
4	200	0.992 (0.126)	0.974 (0.164)	0.940 (0.487)
	500	0.993 (0.060)	0.990 (0.083)	1.037 (0.378)
	1000	0.998 (0.0416)	1.000 (0.060)	1.024 (0.241)
	100	0.899 (0.648)	0.666 (0.597)	0.027 (0.012)
10	200	1.026 (0.434)	0.818 (0.361)	0.160 (0.070)
	500	1.012 (0.145)	0.977 (0.171)	0.946 (0.583)
	1000	0.999 (0.078)	0.986 (0.104)	1.027 (0.457)

Theorems 1 and 2 indicate that the efficiency should be close to 1.

The corresponding mean and the standard error (in the parenthesis) of the relative efficiencies for the first dimension ($\alpha = 1$) is given in Table 1. For the cases of $\rho = 0$, almost all of the mean values are around 1 without noticeable influence from the sample size and the correlation. The trend of standard errors confirm the comparability of SBLL estimator $\hat{m}_{\mathrm{SBLL},\alpha}$ to the oracle smoother $\tilde{m}_{\mathrm{LL},\alpha}$, with faster convergence for larger samples.

In the cases of $\rho = 0.3$, the trend to relative efficiency 1 is very clear regardless of the dimension d. All the means are becoming larger accordingly and approaching to 1 steadily when the sample size increases. Typically, the relative efficiencies are 0.974 for $d = 4$ with sample size 200, and 0.977 for $d = 10$ with sample size 500 respectively. We believe that in high dimensional cases the convergence rate is slower than in lower dimensional cases when the predictors are highly dependent. The standard errors in the parenthesis follow the same trend that less variation is with larger sample size, though it has slower convergence compared to the cases of $\rho = 0$, which is not unexpected.

For the highly dependent cases of $\rho = 0.9$, the efficiency follows the similar trend as the cases of $\rho = 0, 0.3$, but with less efficiency for sample sizes below 200, and much better performance for sample sizes higher than 500. In particular, when $n = 1{,}000$, the relative efficiency reaches the surprisingly high 1.024 and 1.027 when $d = 4, 10$ respectively. This offers some assurance that the SBLL can work well even in the presence of strong dependence among predictors, as long as the sample size is large.

Several figures display the features of the relative efficiencies in details. In Fig. 1 four types of line characteristics correspond to the four sample sizes, the solid line (100), the dotted line (200), the thin line (500) and the thick line (1000). The vertical line at efficiency 1 is the standard line for the comparison of $\hat{m}_{\mathrm{SBLL},1}(x_1)$ and $\tilde{m}_{\mathrm{LL},1}(x_1)$. More efficiency values distributed around the vertical line would be confirmative to the conclusions of Theorems 1 and 2.

All the curves in Fig. 1 are the density estimates of relative efficiency distributions for $n = 100, 200, 500, 1{,}000$, $\rho = 0, 0.3$ and $d = 4, 10$. With increasing sample sizes, the relative efficiency distributions are becoming closer to the vertical standard

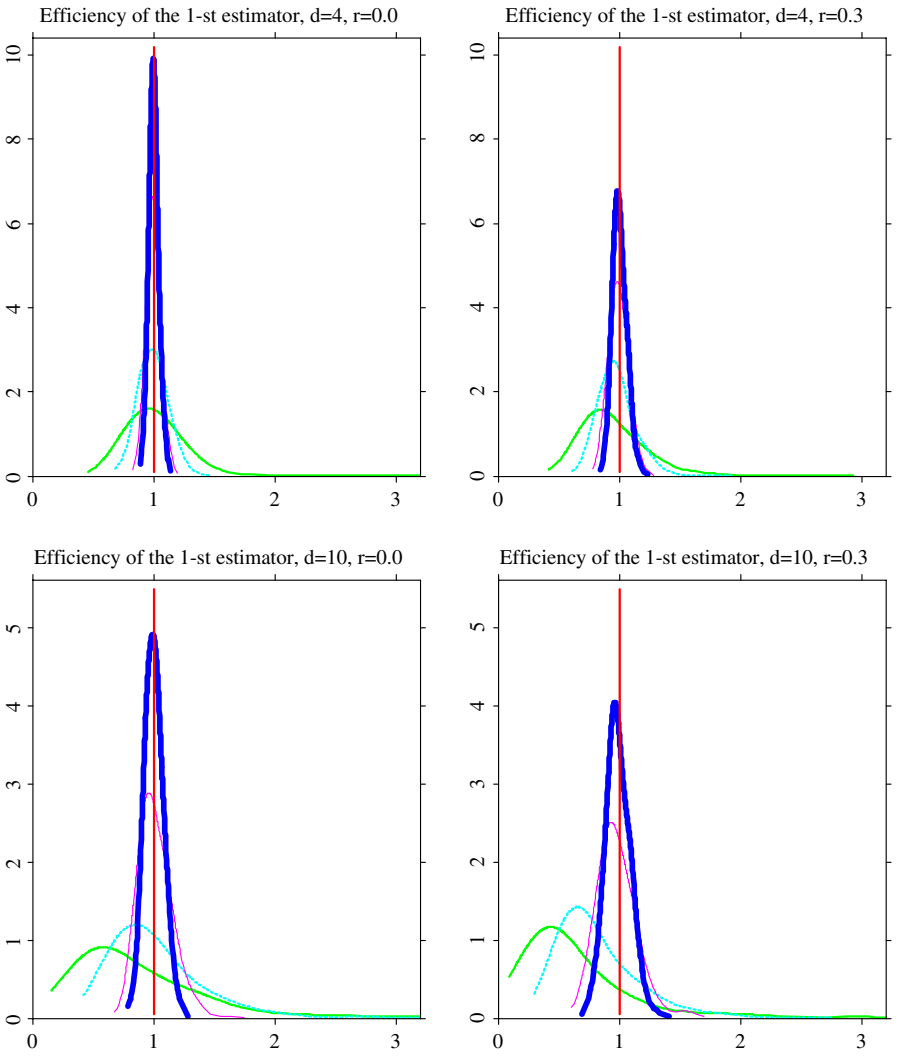

Fig. 1 Empirical distribution of relative efficiency of $\tilde{m}_{\mathrm{LL},\alpha}$ against $\hat{m}_{\mathrm{SBLL},\alpha}$, $d = 4, 10$

line, with narrower spread. In addition, curves with $\rho = 0$ shows a faster convergence to the vertical line than those with $\rho = 0.3$. An interesting point is that almost of all the peak points of the thick line (with the largest sample size) fall very close to the vertical lines. All of these confirm the theorem that SBLL behaves similarly to the oracle local linear estimator.

We have experimented with high dimensionality $d = 50$, $S = 100$ replications, $\rho = 0, 0.3$, and $n = 500, 1,000, 1,500, 2,000$, the results of which are graphically represented in Fig. 2. The basic graphic pattern is similar to that for the lower dimensions $d = 4, 10$, though with slower convergence rate and relatively poorer efficiency. The corresponding statistics are listed in Table 2. We agree with the referee's com-

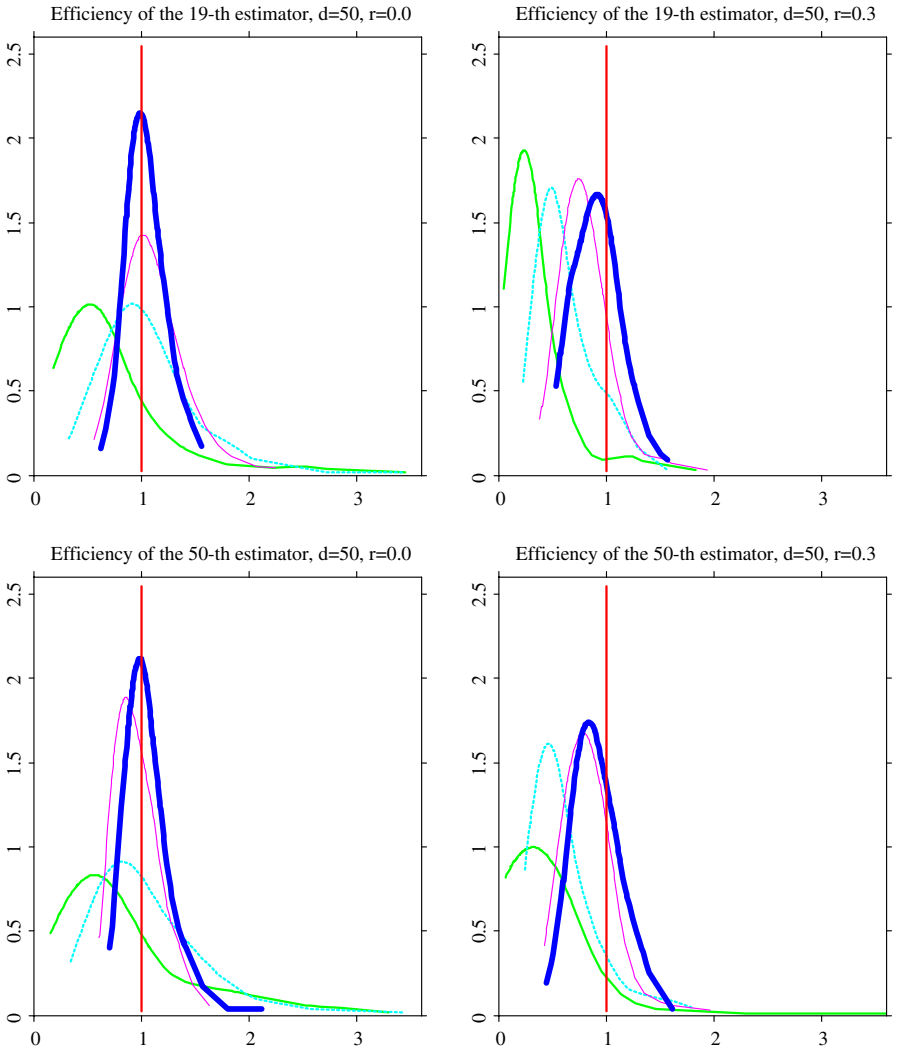

Fig. 2 Empirical distribution of relative efficiency of $\tilde{m}_{\mathrm{LL},\alpha}$ against $\hat{m}_{\mathrm{SBLL},\alpha}$, $d = 50$, $\alpha = 19, 50$

ment that the combination of very high d and moderate n makes the number of knots N_n defined in (26) too small for the first step spline smoothing, thus undersmoothing impossible, resulting in the significant loss of efficiency, at least when $\rho = 0.3$. The good news is that at $\rho = 0$, the SBLL performs on average better than the oracle local linear estimator in most cases because the independent components can be well-estimated at the first stage, then univariate local linear smoothing at the second stage faces smaller noise than the oracle local linear estimator.

To understand further what causes low efficiency when $\rho > 0$, the referee suggested that decomposing the mean squared error into the bias and variance could provide additional insights. We have therefore run $S = 500$ replications with $n = 500$, $d = 10$,

Table 2 Relative efficiency of $\tilde{m}_{\mathrm{LL},\alpha}$ against $\hat{m}_{\mathrm{SBLL},\alpha}$, $\alpha = 1, 10, 19, 50$, with 100 replications for $d = 50$

ρ	n	eff_1	eff_{10}	eff_{19}	eff_{50}
	500	1.030 (0.830)	0.995 (0.778)	0.737 (0.567)	0.861 (0.648)
0	1000	1.130 (0.756)	1.015 (0.523)	1.055 (0.467)	1.056 (0.509)
	1500	1.022 (0.318)	1.029 (0.248)	1.107 (0.302)	0.957 (0.205)
	2000	1.029 (0.197)	1.016 (0.194)	1.045 (0.188)	1.061 (0.223)
	500	0.379 (0.297)	0.410 (0.408)	0.352 (0.296)	0.444 (0.721)
0.3	1000	0.618 (0.269)	0.604 (0.290)	0.623 (0.268)	0.607 (0.311)
	1500	0.864 (0.345)	0.843 (0.280)	0.806 (0.254)	0.831 (0.250)
	2000	0.915 (0.247)	0.872 (0.194)	0.917 (0.221)	0.907 (0.221)

Table 3 Averaged squared bias and averaged variance of $\hat{m}_{\mathrm{SBLL},3}$, ratio = AVAR / ASB. Dimensionality $d = 10$, sample size $n = 500$, number of replications $S = 500$

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
ASB	0.023	0.023	0.02	0.023	0.024	0.025	0.028	0.032
AVAR	0.581	0.593	0.560	0.604	0.621	0.632	0.627	0.654
Ratio	25.15	25.58	25.8	25.8	25.6	24.9	22.6	20.70
ρ	0.8	0.85	0.9	0.92	0.94	0.96	0.98	
ASB	0.037	0.044	0.057	0.069	0.084	0.117	0.192	
AVAR	0.665	0.693	0.740	0.762	0.816	0.906	1.044	
Ratio	17.8	15.7	13.1	11.1	9.7	7.7	5.4	

and estimated the third component by SBLL procedure. The statistics of interest are averaged squared bias ASB and the averaged variance AVAR, where the averaging is taken over all replications and data points. Empirical values of ASB and AVAR associated with various ρ have been listed in Table 3. Also listed are the ratios AVAR / ASB, which indicates the relative magnitude of the squared bias and variance. For all values of ρ from 0 to 0.98, the variance dominates squared bias in magnitude, ranging from 25.15 for $\rho = 0$ to 17.8 for $\rho = 0.8$, and to 5.4 for $\rho = 0.98$. We do not have a theoretical interpretation at this point as to why the variance dominates, but the fact that it does leads us to believe that the variance is most to blame for the loss of efficiency when there exists high dependence among predictors. Similar phenomenon has been noticed in Nielsen and Sperlich (2005).

5 Application to Boston housing data

In this section we apply our method to the Boston housing data. The data files bostonh.dat is available in the software of XploRe. The data set contains 506 different houses from a variety of locations in Boston Standard Metropolitan Statistical Area in 1970. The median value and 13 sociodemographic statistics values of the Boston houses were first studied by Harrison and Rubinfel (1978) to estimate the housing

price index model. Breiman and Friedman (1985) did further analysis to deal with the multi-collinearity among the independent variables. Four variables were selected after penalizing for overfitting by using a stepwise method. We used the same four covariates for our model fitting and current analysis. The response and explanatory variables of interest are:

MEDV: Median value of owner-occupied homes in \$1000's
RM: average number of rooms per dwelling
TAX: full-value property-tax rate per \$10, 000
PTRATIO: pupil-teacher ratio by town school district
LSTAT: proportion of population that is of "lower status" in %.

In order to ease off the trouble caused by big gaps in the domain of variables TAX and LSTAT, logarithmic transformation is done for both variables before fitting the model. We fitted an additive model as follows:

$$\text{MEDV} = \mu + m_1(\text{RM}) + m_2(\log(\text{TAX})) + m_3(\text{PTRATIO}) + m_4(\log(\text{LSTAT})) + \varepsilon.$$

Although the transformation has shrunk the gap in the domain, some compromise will be necessary to estimate the components since we select the same knots number for each direction. In this case we choose a large number of knots, $N = 5$. In the smoothing step, we use the SBLL estimator to get the final function estimate for each input variable.

In Fig. 3, the univariate function estimates and corresponding confidence bands are displayed together with the "pseudo data points" with pseudo response as the backfitted response after subtracting the sum function of the remaining three covariates as in (6). All the function estimates are represented by the dotted lines, "data points" by circles, and confidence bands by upper and lower thin lines. The kernel used in SBLL estimator is quartic kernel, $K(u) = \frac{15}{16}\left(1-u^2\right)^2$ for $-1 < u < 1$.

The proposed confidence bands are used to test the linearity of the components. In Fig. 3 the straight solid lines are the least squares regression lines. The first figure shows that the null hypothesis H_0: $m_1(\text{RM}) = a_1 + b_1\text{RM}$, will be rejected since the confidence bands with 0.99 confidence couldn't totally cover the straight regression line, i.e, the p-value is less than 0.01. Similarly the linearity of the component functions for log (TAX) and log (LSTAT) are not accepted at the significance level 0.01. While the least square straight line of variable PTRATIO in the upper right figure totally falls between the upper and lower 95% confidence bands, thus the linearity null hypothesis $H_0 : m_3(\text{PTRATIO}) = a_3 + b_3 \cdot \text{PTRATIO}$ is accepted at the significance level 0.05.

In addition we add up all the SBLL estimates of component functions and the mean response as a estimate for the response (MEDV). The correlation between the estimate and the raw value of MEDV is as high as 0.80112, implying rather satisfactory fit.

6 Conclusions

In this paper we have proposed SBK and SBLL estimators for the component functions in an additive regression model. These estimators behave asymptotically like the stan-

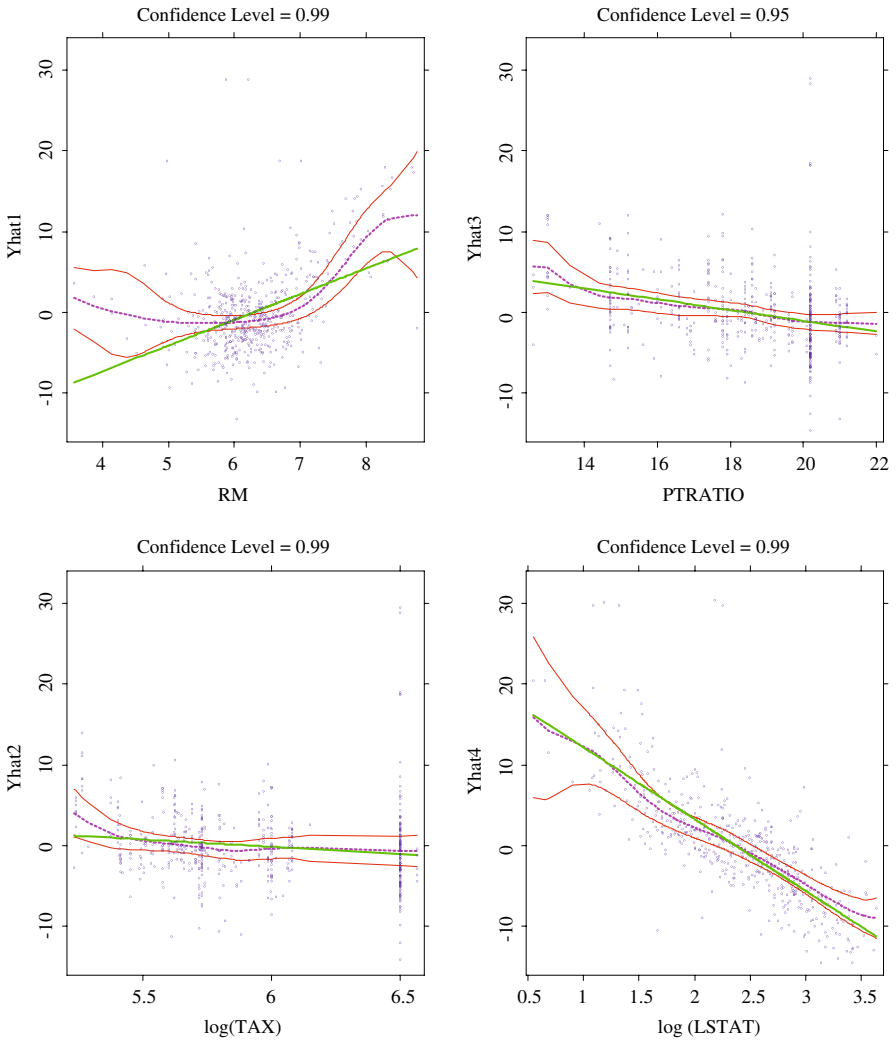

Fig. 3 Linearity test for the Boston housing data. Plots of null hypothesis curves of $H_0 : m_\alpha(x_\alpha) = a_\alpha + b_\alpha \cdot x_\alpha, \alpha = 1, 2, 3, 4$ (*solid line*), linear confidence bands (*upper* and *lower thin lines*), the linear spline estimator (*dotted line*) and the data (*circle*)

dard kernel and local linear estimators in one dimension, thus breaking the problem of d-dimensional additive regression to d univariate regression problems. This is achieved by approximating the unobservable sample $\{Y_{i1}, X_{i1}\}_{i=1}^n$ with the spline estimated sample $\left\{\hat{Y}_{i1}, X_{i1}\right\}_{i=1}^n$. Although much mathematics is devoted to the proof that this approximation works, the implementation is very easy. To give some idea of how fast the procedure is, to run 100 replications for sample sizes $n = 500, 100, 1, 500, 2, 000$ and dimension as high as $d = 50$ takes about 40 min on a Dell notebook. In other words, within this time span, a total of $100 \times 4 = 400$ SBLL estimators $\hat{m}_{s,\alpha}(x_\alpha)$

and the same number of oracle smoothers $\tilde{m}_{s,1}(x_1)$ are computed. In addition, the SBK and SBLL estimators inherit the asymptotic confidence bands (10) of univariate kernel and local linear estimators. The combination of speed and global accuracy for very high dimension regression is very appealing.

Acknowledgments This research is part of the first author's dissertation under the supervision of the second author. The helpful comments from an anonymous referee and an Associate Editor are gratefully acknowledged.

Appendix

A.1 Variance reduction

In this subsection, we prove Propositions 1 and 2. Based on (22) and (24), the conditional second moment $E\{I(x_1)|\mathbf{X}\}^2$ of $I(x_1)$ is

$$E\left\{\left[n^{-1}\sum_{l=1}^{n}K_h(X_{l1}-x_1)\left\{\tilde{\varepsilon}_2^*(X_{l2})-n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_2^*(X_{i2})\right\}\right]^2\middle|\mathbf{X}\right\}.$$

It is clear that $E\{I(x_1)|\mathbf{X}\}^2 = E\left\{I_1^2(x_1)\middle|\mathbf{X}\right\} - E\left\{I_2^2(x_1)\middle|\mathbf{X}\right\}$. Denote

$$I_1(x_1) = n^{-1}\sum_{l=1}^{n}K_h(X_{l1}-x_1)\,\tilde{\varepsilon}_2^*(X_{l2}),$$

$$I_2(x_1) = n^{-1}\sum_{l=1}^{n}K_h(X_{l1}-x_1)\cdot n^{-1}\sum_{i=1}^{n}\tilde{\varepsilon}_2^*(X_{i2}), \tag{27}$$

$$\xi_J(\mathbf{X}_l, x_1) = K_h(X_{l1}-x_1)\,B_{J,2}(X_{l2}). \tag{28}$$

In (22), one has $\tilde{\varepsilon}_2^*(X_{i2}) = \sum_{J=1}^{N}\tilde{a}_{J,2}B_{J,2}(X_{i2})$. Then

$$I_1(x_1) = n^{-1}\sum_{l=1}^{n}\sum_{J=1}^{N}\tilde{a}_{J,2}\xi_J(\mathbf{X}_l, x_1). \tag{29}$$

In order to obtain the order of the conditional second moment of $I_1(x_1)$, we first find the supremum magnitudes of $E\xi_J(\mathbf{X}_l, x_1)$, $\xi_J(\mathbf{X}_l, x_1) - E\xi_J(\mathbf{X}_l, x_1)$ and the size of $\sum_{J=1}^{N}\left|\tilde{a}_{J,2}\right|$, in Lemmas 3, 4 and 7. Consequently, Lemma 10 shows that $\sup_{x_1\in[0,1]}E\left\{I_1^2(x_1)\middle|\mathbf{X}\right\} = O_p\left(n^{-1}\right)$. In Lemma 11, we have $\sup_{x_1\in[0,1]}|I_2(x_1)| = O_p\left(Nn^{-1}\sqrt{\log n}\right)$. Based on the selection of $N \sim n^{2/5}\log n$, Proposition 1 is thus proved.

Under the additional assumption (A2′), the order of $I_1(x_1)$ is obtained uniformly over [0, 1] inflated only by a factor of $\{\log(n)\}^{1/2}$ compared with the pointwise case, one has $\sup_{x_1\in[0,1]}|I_1(x_1)| = O_p\left(\sqrt{\log(n)/n}\right)$. Now again, due to the selection of the

interval width $H \sim \left(n^{2/5} \log n\right)^{-1}$, the order $O_p\left(Nn^{-1}\sqrt{\log n}\right)$ of $\sup_{x_1\in[0,1]} |I_2(x_1)|$ in Lemma 11 is negligible compared with order of $\sup_{x_1\in[0,1]} |I_1(x_1)|$. So under the Assumptions (A1) to (A6) and (A2′), we have established the uniform bound over [0, 1] of Proposition 2.

A.2 Bias reduction

Now we prove Proposition 3 by bounding the bias term $II(x_1)$ in (25). We first cite one important result from page 149 of de Boor (2001).

Theorem 3 *Under Assumption* (A1) $m_\alpha \in \text{Lip}([0,1], C_\infty)$, *then there exists a function* $g_\alpha \in G[0,1]$ *such that* $\forall \alpha = 1, \ldots, d$

$$\|g_\alpha - m_\alpha\|_\infty \leq C_\infty H. \tag{30}$$

Lemma 1 *Under Assumptions* (A1), (A3) *and* (A6), *for the spline function* g_2 *satisfying* (30), *one has*

$$\sup_{x_1\in[0,1]} \left| \frac{\sum_{i=1}^n K_h(X_{i1} - x_1)\{g_2(X_{i2}) - m_2(X_{i2})\}}{\sum_{i=1}^n K_h(X_{i1} - x_1)} \right| \leq C_\infty H, \tag{31}$$

and for $\alpha = 1, 2$

$$|E_n g_\alpha(X_\alpha)| = \left| n^{-1} \sum_{i=1}^n g_\alpha(X_{i\alpha}) \right| = O_p\left(n^{-1/2} + H\right). \tag{32}$$

Proof The first inequality (31) follows trivially from (30). To prove the second inequality, define a function $g(\mathbf{x}) = c + \sum_{\alpha=1}^2 g_\alpha(x_\alpha)$, then $\|g - m\|_\infty \leq 2C_\infty H$ and hence $\|g - m\|_{2,n} \leq 2C_\infty H$. The definition of projection in Hilbert space then implies that $\|\tilde{m} - m\|_{2,n} \leq \|g - m\|_{2,n} \leq 2C_\infty H$ where $\tilde{m}$ is the projection of m to the space G with respect to $\langle \cdot, \cdot \rangle_{2n}$, the triangular inequality implies that $\|\tilde{m} - g\|_{2,n} \leq 4C_\infty H$.

Now (30) leads to $|E_n g_\alpha(X_\alpha) - E_n m_\alpha(X_\alpha)| \leq C_\infty H$, while $Em_\alpha(X_\alpha) = 0$ leads to $E_n m_\alpha(X_\alpha) = O_p\left(n^{-1/2}\right)$. Then one has $|E_n g_\alpha(X_\alpha)| \leq C_\infty H + O_p\left(n^{-1/2}\right)$, which establishes (32). □

In order to show that the bias term $II(x_1)$ defined in (25) is uniformly $o_p\left(n^{-2/5}\right)$, the following lemma suffices.

Lemma 2 *Under Assumptions* (A1) *to* (A6), *as* $n \to \infty$

$$\sup_{x_1\in[0,1]} \left| \frac{\sum_{i=1}^n K_h(X_{i1} - x_1)\{\tilde{m}_2(X_{i2}) - g_2(X_{i2}) + E_n g_2(X_2)\}}{\sum_{i=1}^n K_h(X_{i1} - x_1)} \right| = O_p\left(n^{-1/2} + H\right).$$

Proof Lemmas 1 and 8 would entail that

$$\|\tilde{m} - g + E_n g_1(X_1) + E_n g_2(X_2)\|_2 = O_p\left(n^{-1/2} + H\right). \tag{33}$$

Write next $(\tilde{m}-g)(\mathbf{x})+E_n g_1(X_1)+E_n g_2(X_2)=a+\sum_{\alpha=1}^{2}\sum_{J=1}^{N}a_{J,\alpha}B^*_{J,\alpha}(x_\alpha)$, where $B^*_{J,\alpha}(x_\alpha)=B_{J,\alpha}(x_\alpha)-E_nB_{J,\alpha}(X_\alpha)$, $1\le J\le N, 1\le\alpha\le 2$, which is the empirically centered spline basis. Then for $\alpha=1,2$, $\tilde{m}_\alpha(x_\alpha)-g_\alpha(x_\alpha)+E_ng_\alpha(X_\alpha)=\sum_{J=1}^{N}a_{J,\alpha}B^*_{J,\alpha}(x_\alpha)$, and according to (39) one has

$$\begin{aligned}&\|\tilde{m}-g+E_ng_1(X_1)+E_ng_2(X_2)\|_2^2\\&\quad\ge c_0\left[\left\{a+\sum_{\alpha=1}^{2}\sum_{J=1}^{N}a_{J,\alpha}E_nB_{J,\alpha}(X_\alpha)\right\}^2+\sum_{\alpha=1}^{2}\sum_{J=1}^{N}a_{J,\alpha}^2\right].\end{aligned}\tag{34}$$

We bound $n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)\{\tilde{m}_2(X_{i2})-g_2(X_{i2})+E_ng_2(X_2)\}$ by

$$\begin{aligned}&\sum_{J=1}^{N}|a_{J,2}|\sup_{1\le J\le N}\left|n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)B^*_{J,2}(X_{i2})\right|\\&\quad\le\sum_{J=1}^{N}|a_{J,2}|\left\{\sup_{1\le J\le N}\left|n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)B_{J,2}(X_{i2})\right|\right.\\&\qquad\left.+\left|n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)\right|\sup_{1\le J\le N}\left|E_nB_{J,2}(X_2)\right|\right\}\end{aligned}$$

which can be rewritten as the following according to the definitions of $\xi_J(\mathbf{X}_l,x_1)$ in (28) and of A_n in Lemma 8

$$\sum_{J=1}^{N}|a_{J,2}|\left\{\sup_{1\le J\le N}\left|n^{-1}\sum_{l=1}^{n}\xi_J(\mathbf{X}_l,x_1)\right|+A_n\left|n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)\right|\right\}.$$

Minkowski inequality, Lemmas 5, 8 and standard properties of kernel density estimator now imply that

$$\begin{aligned}&\sup_{x_1\in[0,1]}\left|n^{-1}\sum_{i=1}^{n}K_h(X_{i1}-x_1)\{\tilde{m}_2(X_{i2})-g_2(X_{i2})+E_ng_2(X_2)\}\right|\\&\le\sqrt{N\sum_{J=1}^{N}a_{J,2}^2}\left\{O_p\left(\sqrt{H}\right)+O_p\left(\sqrt{\frac{\log n}{n}}\right)\right\}=O_p\left(\sqrt{\sum_{J=1}^{N}a_{J,2}^2}\right)\\&=O_p\left(\left[\left\{\hat{a}+\sum_{\alpha=1}^{2}\sum_{J=1}^{N}\hat{a}_{J,\alpha}E_nB_{J,\alpha}(X_\alpha)\right\}^2+\sum_{\alpha=1}^{2}\sum_{J=1}^{N}a_{J,\alpha}^2\right]^{1/2}\right),\end{aligned}$$

which according to (33), (34) is $O_p\left(\|\tilde{m}-g+E_ng_1(X_1)+E_ng_2(X_2)\|_2\right)=O_p\left(n^{-1/2}+H\right)$, thus proving the lemma. □

Lemmas 1, 2 lead to the following and then Proposition 3

$$\sup_{x_1\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^{n}K_h(X_{i1}-x_1)\{\tilde{m}_2(X_{i2})-m_2(X_{i2})\}\right|=O_p\left(n^{-1/2}+H\right)=o_p\left(n^{-2/5}\right).$$

A.3 Technical lemmas

In this subsection we have collected all the auxiliary results used in Sects. A.1 and A.2.

Lemma 3 *Under Assumptions* (A3) *to* (A6)*, one has*

$$\sup_{x_1\in[0,1]}\sup_{1\le J\le N}|E\xi_J(\mathbf{X}_l,x_1)|=O\left(H^{1/2}\right).$$

Proof See Wang and Yang (2007b).

Lemma 4 *Denote by* D_n *a set of endpoints in* $[0,1]$*, with cardinality* $M_n=|D_n|$ *of order* n^6*, i.e. there exist constants* $0<c_D<C_D$ *such that* $c_Dn^6\le M_n\le C_Dn^6$*, then under Assumptions* (A3) *to* (A6)

$$\sup_{x_1\in D_n}\sup_{1\le J\le N}\left|n^{-1}\sum_{l=1}^{n}\{\xi_J(\mathbf{X}_l,x_1)-E\xi_J(\mathbf{X}_l,x_1)\}\right|=O_p\left((nh)^{-1/2}\log^{1/2}n\right).$$

Proof Based on Assumptions (A3) and (A4), there exist constants $c',C'>0$, such that $c'h^{-1}\le E\xi_J^2(\mathbf{X}_l,x_1)\le C'h^{-1}$. Then $E\xi_J^2(\mathbf{X}_l,x_1)\gg\{E\xi_J(\mathbf{X}_l,x_1)\}^2$ where $a_n\gg b_n$ means $\lim_{n\to\infty}b_n/a_n=0$. For simplicity, denote $\xi_J^*(\mathbf{X}_l,x_1)=\xi_J(\mathbf{X}_l,x_1)-E\xi_J(\mathbf{X}_l,x_1)$. There exists a positive constant $c^*<c'$ such that

$$E\left\{\xi_J^*(\mathbf{X}_l,x_1)\right\}^2=E\xi_J^2(\mathbf{X}_l,x_1)-\{E\xi_J(\mathbf{X}_l,x_1)\}^2\ge c^*h^{-1}.$$

When $k\ge3$, the kth moment $E|\xi_J(\mathbf{X}_l,x_1)|^k$ can be bounded as follows

$$c_k'h^{(1-k)}H^{(1-k/2)}\left(1+\frac{c_f^k}{C_f^k}\right)\le E|\xi_J(\mathbf{X}_l,x_1)|^k\le C_k'h^{(1-k)}H^{(1-k/2)}\left(1+\frac{C_f^k}{c_f^k}\right).$$

Lemma 3 implies that $E|\xi_J(\mathbf{X}_l,x_1)|^k\gg|E\xi_J(\mathbf{X}_l,x_1)|^k$. There exists a positive constant c such that

$$\begin{aligned}E\left|\xi_J^*(\mathbf{X}_l,x_1)\right|^k&\le2^{k-1}\left(E|\xi_J(\mathbf{X}_l,x_1)|^k+|E\xi_J(\mathbf{X}_l,x_1)|^k\right)\\&\le c^{k-2}k!E\left|\xi_J^*(\mathbf{X}_l,x_1)\right|^2,\end{aligned}$$

the sequence of $\left\{\xi_J^*\left(\mathbf{X}_l, x_1\right)\right\}_{l=1}^n$ satisfies the Cramér's condition. By the Bernstein's inequality we have

$$P\left\{\left|n^{-1}\sum_{l=1}^n \xi_J^*\left(\mathbf{X}_l, x_1\right)\right| \geq \delta\sqrt{\frac{\log n}{(nh)}}\right\} \leq 2\exp\left\{\frac{-\delta^2 \log n}{c^* + 2C_2\delta H^{-1/2}\sqrt{\log n/(nh)}}\right\},$$

for $\delta > 0$ large enough such that $\delta^2\left\{c^* + 2C_2\delta H^{-1/2}\sqrt{\log n/(nh)}\right\}^{-1} > 10$, then

$$\sum_{n=1}^{\infty} P\left\{\sup_{x_1\in D_n}\sup_{1\leq J\leq N}\left|n^{-1}\sum_{l=1}^n \xi_J^*\left(\mathbf{X}_l, x_1\right)\right| \geq \delta\sqrt{\frac{\log n}{(nh)}}\right\} \leq 2C_D\sum_{n=1}^{\infty} n^{-3} < \infty.$$

Borel-Cantelli Lemma implies the desirable result in the lemma.

Lemma 5 *Under Assumptions* (A3) *to* (A6)

$$\sup_{x_1\in[0,1]}\sup_{1\leq J\leq N}\left|n^{-1}\sum_{l=1}^n \xi_J\left(\mathbf{X}_l, x_1\right)\right| = O_p\left(H^{1/2}\right).$$

Proof Denote for $x \in [0, 1]$, $\Lambda(x) = \sup_{1\leq J\leq N}\left|n^{-1}\sum_{l=1}^n \xi_J\left(\mathbf{X}_l, x\right)\right|$. Let the subset D_n in Lemma 4 be equally spaced in $[0, 1]$, specifically $D_n = \left\{x_{1,k}, 0 \leq k \leq M_n; 0 = x_{1,0} < x_{1,1} < \ldots < x_{1,M_n} = 1\right\}$, then the consecutive endpoints make a total of M_n subintervals with length M_n^{-1}. Employing the discretization method, we have

$$\sup_{x_1\in[0,1]}|\Lambda(x_1)| = \sup_{0\leq k\leq M_n}\left|\Lambda\left(x_{1,k}\right)\right| + \sup_{1\leq k\leq M_n}\sup_{x_1\in\Delta_k}\left|\Lambda(x_1) - \Lambda\left(x_{1,k}\right)\right|, \tag{35}$$

where $\Delta_k = \left[x_{1,k-1}, x_{1,k}\right]$. We only need to bound the second term, as Lemmas 3 and 4, and the fact $H^{1/2} \gg \sqrt{\log n/(nh)}$ yield

$$\sup_{0\leq k\leq M_n}\left|\Lambda\left(x_{1,k}\right)\right| = \sup_{x_1\in D_n}\sup_{1\leq J\leq N}\left|n^{-1}\sum_{l=1}^n \xi_J\left(\mathbf{X}_l, x_1\right)\right| = O_p\left(H^{1/2}\right). \tag{36}$$

Employing Lipschitz continuity of kernel K, one has

$$\sup_{1\leq k\leq M_n}\sup_{x_1\in\Delta_k}\left|K_h\left(X_{l1} - x_1\right) - K_h\left(X_{l1} - x_{1,k}\right)\right| \leq C_K M_n^{-1} h^{-2} \tag{37}$$

Hence we have

$$\left|\Lambda(x_1) - \Lambda\left(x_{1,k}\right)\right| \leq \left|K_h\left(X_{l1} - x_1\right) - K_h\left(X_{l1} - x_{1,k}\right)\right| \sup_{1\leq J\leq N}\sum_{l=1}^n \frac{\left|B_{J,2}\left(X_{l2}\right)\right|}{n},$$

$$\sup_{1\le k\le M_n}\sup_{x_1\in\Delta_k}\left|\Lambda\left(x_1\right)-\Lambda\left(x_{1,k}\right)\right|=O\left(M_n^{-1}h^{-2}H^{-1/2}\right)=o\left(n^{-1}\right) \tag{38}$$

since $c_D n^6 \le M_n \le C_D n^6$ in Lemma 4. The lemma follows instantly from (35), (36) and (38). □

Lemma 6 *Under Assumptions* (A3) *and* (A6)*, there exist constants* $C_0 > c_0 > 0$ *such that for any* $\mathbf{a} = \left(a_0, a_{1,1}, \ldots, a_{N,1}, a_{1,2}, \ldots, a_{N,2}\right)^T \in R^{2N+1}$

$$c_0\left(a_0^2+\sum_{J,\alpha}a_{J,\alpha}^2\right)\le\left\|a_0+\sum_{J,\alpha}a_{J,\alpha}B_{J,\alpha}\right\|_2^2\le C_0\left(a_0^2+\sum_{J,\alpha}a_{J,\alpha}^2\right). \tag{39}$$

Proof See Wang and Yang (2007b).

Lemma 7 *Under Assumptions* (A1) *to* (A6)*, the least square solution* $\tilde{\mathbf{a}}$ *defined in* (18) *satisfies*

$$\tilde{\mathbf{a}}^T\tilde{\mathbf{a}}=\tilde{a}_0^2+\sum_{J=1}^{N}\sum_{\alpha=1}^{2}\tilde{a}_{J,\alpha}^2=O_p\left(\frac{N}{n}\right). \tag{40}$$

Proof According to (18), $\tilde{\mathbf{a}}=\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}$, then $\tilde{\mathbf{a}}^T\mathbf{B}^T\mathbf{B}\tilde{\mathbf{a}}=\tilde{\mathbf{a}}^T\left(\mathbf{B}^T\mathbf{E}\right)$. Replacing $\mathbf{B}^T\mathbf{B}$ with matrix of the inner products $\left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_{2,n}$, as the matrix $\mathbf{B}$ is given in (19), one has

$$\|\mathbf{B}\tilde{\mathbf{a}}\|_{2,n}^2=\tilde{\mathbf{a}}^T\begin{pmatrix}\mathbf{1}\\ \left\langle B_{J,\alpha}, B_{J',\alpha'}\right\rangle_{2,n}\end{pmatrix}\tilde{\mathbf{a}}=\tilde{\mathbf{a}}^T\left(n^{-1}\mathbf{B}^T\mathbf{E}\right). \tag{41}$$

Based on (39), one has

$$(1-A_n)\|\mathbf{B}\tilde{\mathbf{a}}\|_2^2=(1-A_n)\left\|\tilde{a}_0+\sum_{J,\alpha}\tilde{a}_{J,\alpha}B_{J,\alpha}\right\|_2^2\ge c_0(1-A_n)\left(\tilde{a}_0^2+\sum_{J,\alpha}\tilde{a}_{J,\alpha}^2\right), \tag{42}$$

where A_n is of order $o_p(1)$ in Lemma 8. Meanwhile by the Cauchy–Schwartz inequality, the right-hand side of (41) is bounded from above by

$$\left|\tilde{\mathbf{a}}^T\left(n^{-1}\mathbf{B}^T\mathbf{E}\right)\right|\le\left(\tilde{a}_0^2+\sum_{J,\alpha}\tilde{a}_{J,\alpha}^2\right)^{1/2}\left\{n^{-2}\mathbf{E}^T\mathbf{B}\mathbf{B}^T\mathbf{E}\right\}^{1/2}. \tag{43}$$

Now (41), (42), (43) imply $\tilde{a}_0^2+\sum_{J,\alpha}\tilde{a}_{J,\alpha}^2\le c_0^{-2}(1-A_n)^{-2}n^{-2}\mathbf{E}^T\mathbf{B}\mathbf{B}^T\mathbf{E}$. It is trivial to verify that $E\left\{n^{-2}\mathbf{E}^T\mathbf{B}\mathbf{B}^T\mathbf{E}\right\}=O\left(n^{-1}N\right)$, so (40) holds. □

Lemma 8 *Under Assumptions* (A3) *and* (A4)*, the uniform supremum of the rescaled difference between* $\langle g_1, g_2\rangle_{2,n}$ *and* $\langle g_1, g_2\rangle_2$ *is*

$$A_n = \sup_{g_1, g_2 \in G^{(-1)}} \frac{\left|\langle g_1, g_2\rangle_{2,n} - \langle g_1, g_2\rangle_2\right|}{\|g_1\|_2 \|g_2\|_2} = O_p\left(\sqrt{\frac{\log n}{nH}}\right) = o_p(1). \tag{44}$$

Proof See Wang and Yang (2007b).

The positive definiteness of matrix $\left(n^{-1}\mathbf{B}^T\mathbf{B}\right)^{-1}$ is a sufficient step to achieve Lemma 10. For simplicity it is denoted by $S = \left(s_{jj'}\right)_{j,j'=1}^{dN+1}$.

Lemma 9 *Under Assumptions* (A3) *and* (A4)*, for the matrix S defined above, there exist constants* $C_S > c_S > 0$ *such that with probability approaching to* 1*, one has*

$$c_S I_{2N+1} \le S^{-1} \le C_S I_{2N+1}. \tag{45}$$

Proof See Wang and Yang (2007b).

Lemma 10 *Under Assumptions* (A1) *to* (A6)*, for any* $x_1 \in [0,1]$ *and* $I_1(x_1)$ *defined in* (27)*, one has*

$$\sup_{x_1 \in [0,1]} E\left\{I_1^2(x_1)\middle|\mathbf{X}\right\} = O_p\left(n^{-1}\right). \tag{46}$$

Proof The conditional mean square of $\tilde{\varepsilon}_2^*(X_{l2})$ given $\mathbf{X}$ is

$$E\left[\left\{\tilde{\varepsilon}_2^*(X_{l2})\right\}^2\middle|\mathbf{X}\right] = E\left(\left\{\tilde{\mathbf{a}}^T\mathbf{P}_{0_{N+1},I_N}\left(e_l^T\mathbf{B}\right)^T\right\}^T\left\{\tilde{\mathbf{a}}^T\mathbf{P}_{0_{N+1},I_N}\left(e_{l'}^T\mathbf{B}\right)^T\right\}\middle|\mathbf{X}\right).$$

Based on Assumption (A2), we have $E\left\{\left(\mathbf{E}\cdot\mathbf{E}^T\right)\middle|\mathbf{X}_1,\ldots,\mathbf{X}_n\right\} \le C_\sigma^2 I_n$ in the matrix sense. Knowing that $\tilde{\mathbf{a}} = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}$, and apply the matrices to a quadratic form with vector $\mathbf{E}^T\left\{\mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{P}_{0_{N+1},I_N}\mathbf{B}^T e_{l'}\right\}$, one has

$$E\left[\left\{\tilde{\varepsilon}_2^*(X_{l2})\right\}^2\middle|\mathbf{X}\right] \le n^{-1}C_\sigma^2 \sum_{1\le J,J'\le N} B_{J,2}(X_{l2})\, s_{J+N+1,J'+N+1} B_{J',2}(X_{l'2}),$$

where the $s_{J+N+1,J'+N+1}$'s are elements of S in Lemma 9. Plugging in the above term, and employing (29), $E\left\{I_1^2(x_1)\middle|\mathbf{X}\right\}$ is less than

$$\begin{aligned}&\frac{C_\sigma^2}{n}\sum_{1\le J,J'\le N} s_{J+N+1,J'+N+1}\sum_{1\le J\le N}\left\{n^{-1}\sum_{l=1}^n K_h(X_{l1}-x_1)B_{J,2}(X_{l2})\right\}^2\\ &\quad\le \frac{C_\sigma^2}{n}C_S\sum_{1\le J\le N}\left\{n^{-1}\sum_{l=1}^n K_h(X_{l1}-x_1)B_{J,2}(X_{l2})\right\}^2,\end{aligned}$$

where C_S is as in (45). Using Lemma 5, $\sup_{x_1\in[0,1]} E\left\{I_1^2(x_1)\middle|\mathbf{X}\right\} \le C_\sigma^2 n^{-1} C_S \sum_{1\le J\le N} H = Cn^{-1}$ in probability, which implies (46). □

Lemma 11 *Under Assumptions* (A1) *to* (A6)*, for* $I_2(x_1)$ *as defined in* (27)*, one has*

$$\sup_{x_1\in[0,1]} |I_2(x_1)| = \sup_{x_1\in[0,1]} \left|\frac{1}{n}\sum_{l=1}^{n} K_h(X_{l1}-x_1)\frac{1}{n}\sum_{i=1}^{n}\tilde{\varepsilon}_2^*(X_{i2})\right| = O_p\left(\frac{N}{n}\sqrt{\log n}\right).$$

Proof See Wang and Yang (2007b).

Lemma 12 *Under Assumptions* (A1) *to* (A6) *and* (A2′)*, and with* $I_1(x_1)$ *defined in* (27)*, one has*

$$\sup_{x_1\in[0,1]} |I_1(x_1)| = \sup_{x_1\in[0,1]} \left|\frac{1}{n}\sum_{l=1}^{n} K_h(X_{l1}-x_1)\tilde{\varepsilon}_2^*(X_{l2})\right| = O_p\left(\sqrt{\frac{\log n}{n}}\right).$$

Proof Using the same discretization as in Lemma 5, we start with

$$\sup_{x_1\in[0,1]} |I_1(x_1)| = \sup_{0\le k\le M_n} \left|I_1\left(x_{1,k}\right)\right| + \sup_{1\le k\le M_n}\sup_{x_1\in\Delta_k} \left|I_1(x_1) - I_1\left(x_{1,k}\right)\right|. \quad (47)$$

Note that for any $x_1 \in [0, 1]$, (21) and (27) imply that

$$I_1(x_1) = n^{-1}\sum_{l=1}^{n} K_h(X_{l1}-x_1)\left(e_l^T\mathbf{B}\right)\mathbf{P}_{0_{N+1},I_N}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{E}.$$

Under Assumption (A2′), $\mathcal{L}\left\{R\left(\mathbf{X},x_{1,k}\right)|\mathbf{X}\right\}$ is standard normal. There exists some $c > 0$, such that $1-\Phi(x) \le c\phi(x)$ for large x, where $\Phi(x)$, $\phi(x)$ are the standard normal distribution and density functions. Take $t_n = \sqrt{16\log n}$, then there is a constant c such that for n large enough

$$\sum_{n=1}^{n} P\left\{\sup_{0\le k\le M_n}\left|R\left(\mathbf{X},x_{1,k}\right)\right| \ge t_n\middle|\mathbf{X}\right\} = \sum_{n=1}^{n} P\left\{\sup_{0\le k\le M_n}|Z| \ge t_n\right\}$$
$$\le \sum_{n=1}^{n} M_n\cdot P\{|Z|\ge t_n\} \le c\sum_{n=1}^{n} M_n\cdot\exp\left\{-t_n^2/2\right\} < \infty,$$

where $Z \sim N(0,1)$. Consequently for a large value $\delta > 0$, we have

$$\sum_{n=1}^{n} P\left\{\sup_{0\le k\le M_n}\left|R\left(\mathbf{X},x_{1,k}\right)\right| \ge \delta\sqrt{\log n}\right\} < \infty,$$

Borel-Cantelli Lemma implies that $\sup_{0\le k\le M_n}\left|R\left(\mathbf{X},x_{1,k}\right)\right| = O_p\left(\sqrt{\log n}\right)$. The conditional variance of $I_1\left(x_{1,k}\right)$ given $\mathbf{X}$ is naturally defined as follows:

$$\operatorname{var}\left\{I_1\left(x_{1,k}\right)\middle|\mathbf{X}\right\} = E\left[\left\{I_1\left(x_{1,k}\right) - EI_1\left(x_{1,k}\right)\right\}^2\middle|\mathbf{X}\right] = E\left\{I_1^2\left(x_{1,k}\right)\middle|\mathbf{X}\right\}.$$

Now Lemma 10 implies that $\sup_{0\le k\le M_n} \operatorname{var}\left\{I_1\left(x_{1,k}\right)\middle|\mathbf{X}\right\} = O_p\left(n^{-1}\right)$ and

$$\begin{aligned}\sup_{0\le k\le M_n}\left|I_1\left(x_{1,k}\right)\right| &\le \sup_{0\le k\le M_n}\left|R\left(\mathbf{X},x_{1,k}\right)\right| \sup_{0\le k\le M_n}\sqrt{\operatorname{var}\left\{I_1\left(x_{1,k}\right)\middle|\mathbf{X}\right\}} \\ &= O_p\left(\sqrt{n^{-1}\log n}\right). \end{aligned} \tag{48}$$

Next, with (21), (37), and (40), it leads to

$$\sup_{1\le k\le M_n}\sup_{x_1\in\Delta_k}\left|I_1\left(x_1\right) - I_1\left(x_{1,k}\right)\right| = O_p\left(M_n^{-1}h^{-2}N^{3/2}n^{-1/2}\right) = o_p\left(n^{-1}\right) \tag{49}$$

due to the choice of $c_D n^6 \le M_n \le C_D n^6$ in Lemma 4. Now (47), (48) and (49) establish the lemma. □

References

Andrews, D., Whang, Y. (1990). Additive interactive regression models: circumvention of the curse of the dimensionality. *Econometric Theory*, *6*, 466–479.

Breiman, L., Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, *80*, 580–619.

Bickel, P. J., Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics*, *1*, 1071–1095.

Claeskens, G., Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, *31*, 1852–1884.

de Boor, C. (2001). *A practical guide to splines*. New York: Springer.

Fan, J., Chen, J. (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society: Series B*, *61*, 927–934.

Fan, J., Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.

Fan, J., Härdle, W., Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *Annals of Statistics*, *26*, 943–971.

Hall, P., Titterington, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *Journal of Multivariate Analysis*, *27*, 228–254.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis*, *29*, 163–179.

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge: Cambridge University Press

Härdle, W., Hlávka, Z., Klinke, S. (2000). *XploRe application guide*. Berlin: Springer.

Härdle, W., Huet, S., Mammen, E., Sperlich, S. (2004). Bootstrap inference in semiparametric generalized additive models. *Econometric Theory*, *20*, 265–300.

Härdle, W., Sperlich, S., Spokoiny, V. (2001). Structural tests in additive regression. *Journal of the American Statistical Association*, *96*, 1333–1347.

Harrison, D., Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for cleaning air. *Journal of Economics and Management*, *5*, 81–102.

Hastie, T. J., Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman and Hall

Horowitz, J. L., Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *Annals of Statistics*, *32*, 2412–2443.

Horowitz, J. L., Klemelä, J., Mammen, E. (2006). Optimal estimation in additive regression models. *Bernoulli*, *12*, 271–298.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Annals of Statistics*, *26*, 242–272.
Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics*, *31*, 1600–1635.
Huang, J. Z., Yang, L. (2004). Identification of nonlinear additive autoregression models. *Journal of the Royal Statistical Society: Series B*, *66*, 463–477.
Kim, W., Linton, O. B., Hengartner, N. (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, *8*, 278–297.
Linton, O. B., Nielsen, J. P. (1995). Estimating structured nonparametric regression models by the kernel method. *Biometrika*, *82*, 93–101.
Linton, O. B., Härdle, W. (1996). Estimating additive regression models with known links. *Biometrika*, *83*, 529–540.
Linton, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika*, *84*, 469–473.
Mammen, E., Linton, O., Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, *27*, 1443–1490.
Nielsen, J. P., Sperlich, S. (2005). Smooth backfitting in practice. *Journal of the Royal Statistical Society: Series B*, *67*, 43–61.
Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, *73*, 166–179.
Opsomer, J. D., Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, *25*, 186–211.
Sperlich, S., Tjøstheim, D., Yang, L. (2002). Nonparametric estimation and testing of interaction in additive models. *Econometric Theory*, *18*, 197–251.
Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, *13*, 689–705.
Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics*, *22*, 118–184.
Tjøstheim, D., Auestad, B. (1994). Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association*, *89*, 1398–1409.
Tusnády, G. (1977). A remark on the approximation of the sample df in the multidimensional case. *Periodica Mathematica Hungarica*, *8*, 53–55.
Wang, J., Yang, L. (2007a). Polynomial spline confidence bands for regression curves. Manuscript.
Wang, J., Yang, L. (2007b). Efficient and fast spline-backfitted kernel smoothing of additive models. http://www.stt.msu.edu/~yangli/SBKAISMfull.pdf.
Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society: Series B*, *60*, 797–811.
Xue, L., Yang, L. (2006a). Additive coefficient modeling via polynomial spline. *Statistica Sinica*, *16*, 1423–1446.
Xue, L., Yang, L. (2006b). Estimation of semiparametric additive coefficient model. *Journal of Statistical Planning and Inference*, *136*, 2506–2534.
Yang, L. (2007). Confidence band for additive regression model. *Journal of Data Science*, forthcoming.
Yang, L., Härdle, W., Nielsen, J. P. (1999). Nonparametric autoregression with multiplicative volatility and additive mean. *Journal of Time Series Analysis*, *20*, 579–604.
Yang, L., Sperlich, S., Härdle, W. (2003). Derivative estimation and testing in generalized additive models. *Journal of Statistical Planning and Inference*, *115*, 521–542.
Yang, L., Park, B. U., Xue, L., Härdle, W. (2006). Estimation and testing of varying coefficients in additive models with marginal integration. *Journal of the American Statistical Association*, *101*, 1212–1227.