

Efficient and fast spline-backfitted kernel smoothing of additive models

JING WANG¹ and LIJIAN YANG^{2*}

¹*Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824 48824, USA. E-mail: wangjin3@msu.edu*

²*Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824 48824, USA. E-mail: *yang@stt.msu.edu*

Short Running Title. Additive Regression Model

A great deal of effort has been devoted to the inference of additive model in the last decade. Among the many existing procedures, the kernel type are too costly to implement for large number of variables or for large sample sizes, while the spline type provide no asymptotic distribution or any measure of uniform accuracy. We propose a synthetic estimator of the component function in an additive regression model, using a one step backfitting, with spline estimators in the first stage and kernel estimators for the second stage. It is established that under very weak conditions, the proposed estimator's pointwise distribution is asymptotically equivalent to an ordinary univariate Nadaraya-Watson estimator, hence the dimension is effectively reduced to one at any point. This dimension reduction holds uniformly over an interval under stronger 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 bands are applied to the Boston housing data for linearity diagnosis.

Keywords: bandwidths; B spline; knots; local linear estimator; Nadaraya-Watson estimator; nonparametric regression

1. Introduction

For the past two decades, non- and semiparametric regression techniques have been widely used in many statistical applications, from biostatistics to econometrics, from engineering to the forecasting of financial time series. Much effort has been devoted to addressing the issue of the “curse of dimensionality”. One popular choice for such purpose is the additive model popularized by the book of Hastie and Tibshirani (1990)

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon, \mathbf{X} = (X_1, \dots, X_d), m(\mathbf{x}) = c + \sum_{\alpha=1}^d m_{\alpha}(x_{\alpha}), \quad (1.1)$$

where the noise satisfies $E(\varepsilon|X) = 0$, $\text{var}(\varepsilon|X) = 1$ and the component functions satisfy the identification conditions $E m_{\alpha}(X_{\alpha}) \equiv 0$, $\alpha = 1, \dots, d$. In addition, we assume that the predictor X_{α} is distributed on a compact interval $[a_{\alpha}, b_{\alpha}]$, $\alpha = 1, \dots, d$. The goal of this paper

is the efficient and fast estimation of the d unknown component functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^d$ based on an i.i.d. sample $\{Y_i, \mathbf{X}_i^T\}_{i=1}^n = \{Y_i, X_{i1}, \dots, X_{id}\}_{i=1}^n$ following model (1.1).

Stone (1985) proposed estimators for functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^d$ and their derivatives, and established optimal rates of convergence. These were later called polynomial spline estimators in the extended context of functional ANOVA model in Stone (1994), Huang (1998). Huang and Yang (2004) further extended these estimators to weakly dependent data and developed consistent BIC model selection procedure based on such estimation. 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 2 covariates. Mammen, Linton and Nielsen (1999) proposed a projection based modification of the backfitting algorithm and established its theoretical properties, which was implemented in Nielsen and Sperlich (2005) and called smooth backfitting estimator. Another viable alternative is the so-called marginal integration method, as first proposed in Tjøstheim and Auestad (1994), Linton and Nielsen (1995), Linton and Härdle (1996), and further developed in various contexts by Fan, Härdle and Mammen (1998), Yang, Härdle and Nielsen (1999), Sperlich, Tjøstheim and Yang (2002), Yang, Sperlich and Härdle (2003), Xue and Yang (2006). Using the wavelet transformation, Härdle, Sperlich and Spokoiny (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”, then one could define a new variable $Y_1 = Y - c - \sum_{\alpha=2}^d m_\alpha(X_\alpha) = m_1(X_1) + \sigma(\mathbf{X})\varepsilon$ which one can use 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 $m_\alpha(X_\alpha)$, $\alpha = 2, \dots, d$ with the marginal integration pilot estimates (kernel-based) and establishing that the error caused by this “cheating” is negligible for estimating function $m_1(x_1)$. The two-step idea for nonparametric regression 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, Linton and Hengartner (1999) provided an computationally efficient two-step estimator, a reduction in computation of 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. The recent works of Huang (2003), Wang and Yang (2006) on spline confidence intervals and confidence bands do not apply to additive model.

In this paper we propose to pre-estimate the functions $\{m_\alpha(x_\alpha)\}_{\alpha=1}^d$ by an under smoothed constant spline procedure. These function estimates are then used as if they were the true functions for constructing the “oracle” estimator. The greatest advantage of our approach over that of Linton (1997) is that ours is much faster, and can be applied to cases of extremely

high dimension data (e.g., the number of predictors, d , can be as large as 50 or 100). We believe that our approach is the first example of marrying the traditionally parallel spline smoothing and kernel smoothing techniques, leading to an estimator with asymptotically normal distribution like a typical kernel estimator, without the formidable computational burden of high dimensional kernel smoothing. Figuratively speaking, spline smoothing can be compared to a sledge-hammer capable of breaking any 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 only one linear least squares problem), but does not guarantee the fine shapes of the broken pieces (i.e., the estimates are not guaranteed to converge at any point or uniformly over an interval, only in the L^2 sense). In contrast, kernel smoothing works like a sharp knife that cuts anything into pieces of precise shapes (i.e., confidence intervals are available at any point based on asymptotic normal distribution, and confidence bands are available 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 proposed new tool can be described as a hammer-knife capable of first slamming any huge clump into many much smaller pieces (i.e., univariate regression problems) in one hit (the spline backfitting step), and then cutting all the smaller pieces into the exact desired shapes (one dimensional kernel smoothing of backfitted pseudo data). In this sense, the method we propose combines the best features of both spline and kernel methods.

Smoothing experts may wonder how one could have all these good features in one method. The success of our method is due to 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, which is the new feature of our proofs. For more details, see Lemmas A.3, A.4 and A.5 in Subsection A.3.

In addition to the above features, uniform confidence bands are provided for all function estimates under mild conditions. Literature on nonparametric confidence bands has been scarce, and as far as we know, is lacking in multivariate regression setting. For kernel smoothing of univariate regression function, Hall and Titterton (1988), Härdle (1989), and Xia (1998) made significant contributions. All of these are based on strong approximation of some empirical processes by the 2-dimensional Brownian bridge, 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, and by extending the confidence band to derivatives of the regression function. Härdle, Huet, Mammen and Sperlich (2004) introduced the bootstrap bands with corrected bias. For spline smoothing of univariate regression function under the most general setting, Wang and Yang (2006) provided simple solutions with asymptotic theory. For additive regression model, however, it seems that this present paper is the one of the few to offer the measure of uniform accuracy with theoretical justifications. The good news is that the confidence band we provide for $m_\alpha(x_\alpha)$ with any $\alpha = 1, \dots, d$, is asymptotically the same confidence band that Härdle (1989) established for univariate regression with kernel smoother, regardless how many regressors there are and what other functions

$m_\alpha(x_\alpha)$, $\alpha = 1, \dots, d$ are. Hence neither the dimension d nor other function components play any role in forming the band for $m_\alpha(x_\alpha)$, at least according to the asymptotic theory. In this sense, our estimator of $m_\alpha(x_\alpha)$ possesses what we would like to call “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. This allows the broadest class of additive model compared to all existing methods, see Remark 3 in Section 2

The rest of the paper is organized as follows. In Section 2 we introduce the spline-backfitted kernel estimator, and state their asymptotic “oracle efficiency” under appropriate assumptions, both pointwise and uniform. In Section 3 we provide some insights into the ideas behind our proofs of the main theoretical results, by decomposing the estimator’s “cheating” error into a bias and a noise part, which will be shown separately to be of negligible order. In Section 4, we present extensive Monte Carlo results to demonstrate that the proposed estimator does indeed possess the claimed asymptotic properties. The simulated examples cover a wide range of sample sizes with correlated structure and some very high dimensions, which would have been either infeasible to handle with kernel smoothing methods, or lacking any measure of confidence, pointwise or global, by spline method. The proposed estimator are applied to the Boston Housing data in Section 5. Section 6 concludes, and all technical proofs are contained in the Appendix.

2. The SBK and SBLLE estimators

In this section, we describe the spline-backfitted kernel estimation procedure. Let $\{Y_i, \mathbf{X}_i^T\}_{i=1}^n = \{Y_i, X_{i1}, \dots, X_{id}\}_{i=1}^n$ be an i.i.d. sample following model (1.1). In what follows, we write all responses as $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, and denote by \mathbf{X} the design matrix $(\mathbf{X}_1, \dots, \mathbf{X}_n)^T$. Without loss of generality, we take all intervals $[a_\alpha, b_\alpha] = [0, 1]$, $\alpha = 1, \dots, d$. We pre select an integer $N_n \sim n^{2/5} \log(n)$, see Assumption (A6) below. Next, we define for any $\alpha = 1, \dots, d$, the indicator function $I_{J,\alpha}(x_\alpha)$ of the $(N + 1)$ equally-spaced subintervals of the finite interval $[0, 1]$, that is

$$I_{J,\alpha}(x_\alpha) = \begin{cases} 1 & JH \leq x_\alpha < (J + 1)H, \\ 0 & \text{otherwise,} \end{cases} \quad H = H_n = (N_n + 1)^{-1}, J = 0, 1, \dots, N. \quad (2.1)$$

Define next the $(1 + dN)$ -dimensional space G of additive spline functions as the linear space spanned by $\{1, I_{J,\alpha}(x_\alpha), \alpha = 1, \dots, d, J = 1, \dots, N\}$, while denote by G_n the subspace of R^n spanned by $\{\{1\}_{i=1}^n, \{I_{J,\alpha}(X_{i\alpha})\}_{i=1}^n, \alpha = 1, \dots, d, J = 1, \dots, N\}$. As $n \rightarrow \infty$, 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 $\{\hat{m}(\mathbf{X}_1), \dots, \hat{m}(\mathbf{X}_n)\}^T$ best approximates the response vector $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. To be precise, we define

$$\hat{m}(\mathbf{x}) = \hat{\lambda}_0 + \sum_{\alpha=1}^d \sum_{J=1}^N \hat{\lambda}_{J,\alpha} I_{J,\alpha}(x_\alpha), \quad (2.2)$$

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

$$\left\{ \hat{\lambda}_0, \hat{\lambda}_{1,1}, \dots, \hat{\lambda}_{N,d} \right\}^T = \operatorname{argmin}_{R^{dN+1}} \sum_{i=1}^n \left\{ Y_i - \lambda_0 - \sum_{\alpha=1}^d \sum_{J=1}^N \lambda_{J,\alpha} I_{J,\alpha}(X_{i\alpha}) \right\}^2. \quad (2.3)$$

The pilot estimators of each component function and the constant are defined as

$$\begin{aligned} \hat{m}_\alpha(x_\alpha) &= \sum_{J=1}^N \hat{\lambda}_{J,\alpha} I_{J,\alpha}(x_\alpha) - n^{-1} \sum_{i=1}^n \sum_{J=1}^N \hat{\lambda}_{J,\alpha} I_{J,\alpha}(X_{i\alpha}), \\ \hat{m}_c &= \hat{\lambda}_0 + n^{-1} \sum_{\alpha=1}^d \sum_{i=1}^n \sum_{J=1}^N \hat{\lambda}_{J,\alpha} I_{J,\alpha}(X_{i\alpha}). \end{aligned} \quad (2.4)$$

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} , to be specific,

$$\hat{Y}_{i1} = Y_i - \hat{c} - \sum_{\alpha=2}^d \hat{m}_\alpha(X_{i\alpha}), Y_{i1} = Y_i - c - \sum_{\alpha=2}^d m_\alpha(X_{i\alpha}), i = 1, \dots, n, \quad (2.5)$$

$\hat{c} = n^{-1} \sum_{i=1}^n Y_i$, where by Central Limit Theorem \hat{c} is a \sqrt{n} -consistent estimator of c . We define next the spline-backfitted kernel (SBK) estimator of $m_1(x_1)$ based on $\left\{ \hat{Y}_{i1}, X_{i1} \right\}_{i=1}^n$ as $\hat{m}_{\text{SBK},1}(x_1)$, which is an attempt to mimic the would-be kernel estimator $\tilde{m}_{\text{K},1}(x_1)$ 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.

$$\hat{m}_{\text{SBK},1}(x_1) = \frac{\sum_{i=1}^n K_h(X_{i1} - x_1) \hat{Y}_{i1}}{\sum_{i=1}^n K_h(X_{i1} - x_1)}, \tilde{m}_{\text{K},1}(x_1) = \frac{\sum_{i=1}^n K_h(X_{i1} - x_1) Y_{i1}}{\sum_{i=1}^n K_h(X_{i1} - x_1)}, \quad (2.6)$$

where \hat{Y}_{i1} and Y_{i1} are defined in (2.5). Similarly, the spline-backfitted local linear (SBLL) estimator $\hat{m}_{\text{SBLL},1}(x_1)$ based on $\left\{ \hat{Y}_{i1}, X_{i1} \right\}_{i=1}^n$ mimics the would-be local linear estimator $\tilde{m}_{\text{LL},1}(x_1)$ based on $\{Y_{i1}, X_{i1}\}_{i=1}^n$

$$\left\{ \hat{m}_{\text{SBLL},1}(x_1), \tilde{m}_{\text{LL},1}(x_1) \right\} = (1, 0) (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W} \left(\hat{\mathbf{Y}}_1, \mathbf{Y}_1 \right), \quad (2.7)$$

in which the oracle and pseudo-response vectors are

$$\mathbf{Y}_1 = (Y_{11}, \dots, Y_{n1})^T, \hat{\mathbf{Y}}_1 = (\hat{Y}_{11}, \dots, \hat{Y}_{n1})^T$$

and the weight and design matrices are

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

Throughout the rest of the paper, the second order smooth function space is defined as $C^{(2)}[a, b] = \{g \mid g'' \in C[a, b]\}$, while the Lipschitz continuous function class is defined as

$$\operatorname{Lip}([a, b], C) = \{m \mid |m(x) - m(x')| \leq C|x - x'|, \forall x, x' \in [a, b]\}.$$

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

- (A1) The component function $m_1 \in C^{(2)} [0, 1]$, while all components $m_\beta \in \text{Lip} ([0, 1], C_\infty)$, $\forall \beta = 1, \dots, d$, $0 < C_\infty < \infty$.
- (A2) The noise ε_i given \mathbf{X}_i are i. i. d. with mean 0 and variance 1, for $i = 1, \dots, n$, 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 $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ given $\tilde{\mathbf{X}} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ 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 density function $K \in \text{Lip}([-1, 1], C_K)$, for some constant $C_K > 0$, and is nonnegative, symmetric, and supported on $[-1, 1]$. We denote in this following $\mu_2(K) = \int u^2 K(u) du$, $\|K\|_2^2 = \int K^2(u) du$.
- (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 $\tilde{m}_{K,1}(x_1)$ and $\tilde{m}_{LL,1}(x_1)$ are well-developed. Under Assumptions (A1)-(A5), according to Theorem 4.2.1 of Härdle (1990), one has for any $x_1 \in [h, 1 - h]$,

$$\begin{aligned} \sqrt{nh} \{ \tilde{m}_{K,1}(x_1) - m_1(x_1) - b_K(x_1) h^2 \} &\xrightarrow{D} N(0, v^2(x_1)), \\ \sqrt{nh} \{ \tilde{m}_{LL,1}(x_1) - m_1(x_1) - b_{LL}(x_1) h^2 \} &\xrightarrow{D} N(0, v^2(x_1)), \end{aligned}$$

where

$$\begin{aligned} b_K(x_1) &= \mu_2(K) \{ m_1''(x_1) f_1(x_1) / 2 + m_1'(x_1) f_1'(x_1) \} f_1^{-1}(x_1), \\ b_{LL}(x_1) &= \mu_2(K) m_1''(x_1) / 2, \\ v^2(x_1) &= \|K\|_2^2 E \{ \sigma^2(x_1, X_2, \dots, X_d) \} f_1^{-1}(x_1). \end{aligned} \tag{2.8}$$

with the equation for $\tilde{m}_{K,1}(x_1)$ 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 $\tilde{m}_{K,1}(x_1)$ is of order $o_p(n^{-2/5})$, which is dominated by the asymptotic size of $\tilde{m}_{K,1}(x_1) - m_1(x_1)$. Hence $\hat{m}_{\text{SBK},1}(x_1)$ will have the same asymptotic distribution as $\tilde{m}_{K,1}(x_1)$.

Theorem 1 Under Assumptions (A1) to (A6), for any $x_1 \in [0, 1]$, estimators $\hat{m}_{\text{SBK},1}(x_1)$ and $\hat{m}_{\text{SBLL},1}(x_1)$ given in (2.6) and (2.7) satisfy

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

Hence with $b_{\text{K}}(x_1)$, $b_{\text{LL}}(x_1)$ and $v^2(x_1)$ defined in (2.8), for any $x_1 \in [h, 1-h]$

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

and with the additional assumption (A7)

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

Theorem 2 Under Assumptions (A1) to (A6) and (A2'), estimators $\hat{m}_{\text{SBK},1}(x_1)$ and $\hat{m}_{\text{SBLL},1}(x_1)$ given in (2.6) and (2.7) satisfy

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

Hence for any z

$$\lim_{n \rightarrow \infty} P \left[\{ \log(h^{-2}) \}^{1/2} \left(\sup_{x_1 \in [0,1]} \frac{\sqrt{nh}}{v(x_1)} |\hat{m}_{\text{SBLL},1}(x_1) - m_1(x_1)| - d_n \right) < z \right] = \exp \{ -2 \exp(-z) \},$$

in which

$$d_n = \{ \log(h^{-2}) \}^{1/2} + \frac{1}{2} \{ \log(h^{-2}) \}^{-1/2} \left[\log \left\{ \int K'^2(u) du \right\} - \log \left\{ 4\pi^2 \int K^2(u) du \right\} \right].$$

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

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

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

$$\hat{m}_{\text{SBLL},1}(x_1) \pm v(x_1) \left(\sqrt{nh} \right)^{-1} \left[d_n - \{ \log(h^{-2}) \}^{-1/2} \log \left\{ -\frac{1}{2} \log(1 - \alpha) \right\} \right]. \quad (2.9)$$

Remark 1. Similar estimators can be constructed for $m_\alpha(x_\alpha)$, $2 \leq \alpha \leq d$ with same oracle properties.

Remark 2. The proofs of Theorems 1 and 2 makes 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 $N'_n \rightarrow \infty$, $n^{-\theta} N'_n \rightarrow 0$ for any $\theta > 0$, while there is no optimal way to choose N'_n . The fact that $N_n^{-1} = o(n^{-2/5})$ 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 (2.3), $1 + dN_n$ has to be small relative to n . Hence we select N_n to be of barely larger order than $n^{2/5}$.

Remark 3. Assumption A1 requires only the Lipschitz continuity for the components except for the component of interest. Obviously all m_α are required to be second order smooth if one needs to estimate all components.

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 $[0, 1]^d$, $\|\phi\|_2^2 = E\{\phi^2(\mathbf{X})\} = \int_{[0,1]^d} \phi^2(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$, 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

$$\begin{aligned} \langle \phi, \varphi \rangle_2 &= \int_{[0,1]^d} \phi(\mathbf{x}) \varphi(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = E\{\phi(\mathbf{X}) \varphi(\mathbf{X})\}, \\ \langle \phi, \varphi \rangle_{2,n} &= n^{-1} \sum_{i=1}^n \phi(\mathbf{X}_i) \varphi(\mathbf{X}_i). \end{aligned} \quad (3.1)$$

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

$$b_{J,\alpha}(x_\alpha) = I_{J+1,\alpha}(x_\alpha) - \|I_{J+1,\alpha}\|_2 \|I_{J,\alpha}\|_2^{-1} I_{J,\alpha}(x_\alpha), \forall \alpha = 1, \dots, d, J = 1, \dots, N, \quad (3.2)$$

where the functions $I_{J,\alpha}(x_\alpha)$'s as defined in (2.1) are indicators on the subintervals $[JH, (J+1)H)$. The standardized one is given for any $\alpha = 1, \dots, d$,

$$B_{J,\alpha}(x_\alpha) = \|b_{J,\alpha}\|_2^{-1} b_{J,\alpha}(x_\alpha), \forall J = 1, \dots, N. \quad (3.3)$$

The additive function space G defined earlier can also be spanned by the linearly independent basis $\{1, B_{J,\alpha}(x_\alpha), J = 1, \dots, N, \alpha = 1, \dots, d\}$, although these new basis involve unknown quantities and therefore can not be computed from the data, they are more convenient for mathematical analysis than the truncated power basis in (2.1). Similarly G_n can be spanned linearly by the basis $\{1, \{B_{J,\alpha}(X_{i\alpha})\}_{i=1}^n, \alpha = 1, \dots, d, J = 1, \dots, N\}$.

For better understanding, we use the projection idea to elaborate the constant spline estimators. The evaluation of constant spline estimator $\hat{m}(\mathbf{x})$ at the n observations results in an n -dimensional vector, $\hat{m}(\tilde{\mathbf{X}}) = \{\hat{m}(\mathbf{X}_1), \dots, \hat{m}(\mathbf{X}_n)\}^T$, which can be considered as the projection of \mathbf{Y} on the space G_n with respect to the empirical inner product $\langle \cdot, \cdot \rangle_{2,n}$ defined in (3.1). In general, for any n -dimensional vector $\mathbf{V} = \{V_1, \dots, V_n\}^T$, 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 $(G_n, \langle \cdot, \cdot \rangle_{2,n})$, 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 square coefficients obtained by

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

which is similar to (2.2) and (2.3) except the basis. Next, the multivariate function $\mathbf{P}_n \mathbf{V}(\mathbf{x})$ is decomposed into direct component $\mathbf{P}_{n,2}^* \mathbf{V}(x_\alpha) = \sum_{J=1}^N \hat{v}_{J,\alpha} B_{J,\alpha}(x_\alpha)$. In order to consider

the identification condition, the pilot estimator in this stage is chosen to be the empirically centered additive components $\mathbf{P}_{n,\alpha} \mathbf{V}(x_\alpha)$, $\alpha = 1, \dots, d$ and the constant component $\mathbf{P}_{n,c} \mathbf{V}$

$$\mathbf{P}_{n,\alpha} \mathbf{V}(x_\alpha) = \sum_{J=1}^N \hat{v}_{J,\alpha} B_{J,\alpha}(x_\alpha) - n^{-1} \sum_{i=1}^n \sum_{J=1}^N \hat{v}_{J,\alpha} B_{J,\alpha}(X_{i\alpha}), \quad (3.4)$$

$$\mathbf{P}_{n,c} \mathbf{V} = \hat{v}_0 + n^{-1} \sum_{\alpha=1}^d \sum_{i=1}^n \sum_{J=1}^N \hat{v}_{J,\alpha} B_{J,\alpha}(x_\alpha), \quad (3.5)$$

in which the centering procedure is the same as (2.4).

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 (2.2) and (2.4) as

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

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

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

and the noise spline components

$$\tilde{\varepsilon}(\mathbf{x}) = \mathbf{P}_n \mathbf{E}(\mathbf{x}), \tilde{\varepsilon}_\alpha(x_\alpha) = \mathbf{P}_{n,\alpha} \mathbf{E}(x_\alpha), \tilde{\varepsilon}_c = \mathbf{P}_{n,c} \mathbf{E}. \quad (3.6)$$

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

$$\hat{m}(\mathbf{x}) = \tilde{m}(\mathbf{x}) + \tilde{\varepsilon}(\mathbf{x}), \hat{m}_\alpha(x_\alpha) = \tilde{m}_\alpha(x_\alpha) + \tilde{\varepsilon}_\alpha(x_\alpha), \hat{m}_c = \tilde{m}_c + \tilde{\varepsilon}_c, \alpha = 1, \dots, d. \quad (3.7)$$

As closer examination is needed later for $\tilde{\varepsilon}(\mathbf{x})$ and $\tilde{\varepsilon}_\alpha(x_\alpha)$, one define that

$$\tilde{\mathbf{a}} = \{\tilde{a}_0, \tilde{a}_{1,1}, \dots, \tilde{a}_{N,d}\}^T = \underset{R^{dN+1}}{\operatorname{argmin}} \sum_{i=1}^n \left\{ \sigma(\mathbf{X}_i) \varepsilon_i - a_0 - \sum_{\alpha=1}^d \sum_{J=1}^N a_{J,\alpha} B_{J,\alpha}(X_{i\alpha}) \right\}^2, \quad (3.8)$$

then $\tilde{\varepsilon}(\mathbf{x})$ in (3.6) can be rewritten as $\tilde{\mathbf{a}}^T \mathbf{B}(\mathbf{x})$, where $\tilde{\mathbf{a}} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{E}$ is the solution of equation (3.8), and matrices $\mathbf{B}(\mathbf{x})$ and \mathbf{B} are defined as

$$\mathbf{B}(\mathbf{x}) = \{1, B_{1,1}(x_1), \dots, B_{N,d}(x_d)\}^T, \mathbf{B} = \{\mathbf{B}(\mathbf{X}_1), \dots, \mathbf{B}(\mathbf{X}_n)\}^T. \quad (3.9)$$

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

$$\tilde{\mathbf{a}} = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \langle B_{J,\alpha}, B_{J',\alpha'} \rangle_{2,n} \end{pmatrix}^{-1} \begin{pmatrix} n^{-1} \sum_{i=1}^n \sigma(\mathbf{X}_i) \varepsilon_i \\ n^{-1} \sum_{i=1}^n B_{J,\alpha}(X_{i\alpha}) \sigma(\mathbf{X}_i) \varepsilon_i \end{pmatrix}_{\substack{1 \leq \alpha, \alpha' \leq d, \\ 1 \leq J, J' \leq N}}. \quad (3.10)$$

Our main objective is to study the difference between smoothed backfitted estimator $\hat{m}_{\text{SBK},1}(x_1)$ and the smoothed ‘‘oracle’’ estimator $\tilde{m}_{\text{K},1}(x_1)$, both given in (??). From now

on, we assume without loss of generality that $d = 2$ for notational brevity. Denote the projection matrix $\mathbf{P}_{0_{N+1}, I_N} = \begin{pmatrix} 0_{N+1} \\ I_N \end{pmatrix}$, we define another auxiliary entity

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

which, in particular, entails that

$$\tilde{\varepsilon}_2^*(X_{i2}) = \tilde{\mathbf{a}}^T \mathbf{P}_{0_{N+1}, I_N} (e_i^T \mathbf{B})^T = \sum_{J=1}^N \tilde{a}_{J,2} B_{J,2}(X_{i2}), \quad (3.11)$$

in which e_i is the n -dimensional unit vector with i -th element 1 and else 0 and hence the i -th row of matrix \mathbf{B} , $e_i^T \mathbf{B} = \mathbf{B}(\mathbf{X}_i)$, is the basis functions corresponding to the i -th observation \mathbf{X}_i . Definitions (3.4) and (3.5) imply that $\tilde{\varepsilon}_2(x_2)$ is simply the empirical centering of $\tilde{\varepsilon}_2^*(x_2)$, i.e. $\tilde{\varepsilon}_2(x_2) \equiv \tilde{\varepsilon}_2^*(x_2) - n^{-1} \sum_{i=1}^n \tilde{\varepsilon}_2^*(X_{i2})$, typically

$$\tilde{\varepsilon}_2(x_2) = \sum_{J=1}^N \tilde{a}_{J,2} B_{J,2}(x_2) - n^{-1} \sum_{i=1}^n \sum_{J=1}^N \tilde{a}_{J,2} B_{J,2}(X_{i2}). \quad (3.12)$$

Making use of the signal noise decomposition (3.7), the difference $\tilde{m}_{K,1}(x_1) - \hat{m}_{\text{SBK},1}(x_1) + \hat{c} - c$ can be treated as the sum of two terms

$$\frac{n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) \{\tilde{m}_2(X_{i2}) - m_2(X_{i2})\}}{n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1)} = \frac{I(x_1) + II(x_1)}{n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1)}, \quad (3.13)$$

where

$$I(x_1) = n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) \cdot \tilde{\varepsilon}_2(X_{i2}), \quad (3.14)$$

$$II(x_1) = n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) \cdot \{\tilde{m}_2(X_{i2}) - m_2(X_{i2})\}. \quad (3.15)$$

The term $I(x_1)$ is related to the noise terms $\tilde{\varepsilon}_2(X_{i2})$, while $II(x_1)$ is induced by the bias terms $\tilde{m}_2(X_{i2}) - m_2(X_{i2})$. Propositions 1 and 2 below show respectively that the term $I(x_1)$ is of order $o_p(n^{-2/5})$, either at a given point or over an interval. This is the most challenging part to be proved, mostly done in Subsection A.1. On the other hand, Proposition 3 below shows that the bias term $II(x_1)$ is uniformly of order $o_p(n^{-2/5})$ for $x_1 \in [0, 1]$, to be proved in Subsection A.2. Standard theory of kernel density estimation ensures that the denominator term in (3.13), $n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1)$, has a positive lower bound for $x_1 \in [0, 1]$. The additional nuisance term $\hat{c} - c$ is of clearly order $O(n^{-1/2})$ and thus $o_p(n^{-2/5})$, which needs no further arguments for the proofs. Hence both Theorems 1 and 2 follow from Propositions 1, 2 and 3. The Appendix, therefore, is devoted exclusively to the proofs of these three propositions, rather than of the main theoretical results, Theorems 1 and 2 themselves.

The next three propositions follow respectively from Lemmas A.10 and A.11, Lemmas A.11 and A.12, Lemmas A.1 and A.2, in the Appendix.

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

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

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 to illustrate the finite-sample behavior of the spline backfitted kernel estimators $\hat{m}_{s,\alpha}(x_\alpha)$ for any $\alpha = 1, \dots, d$.

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

$$m_\alpha(x_\alpha) = \sin(2\pi x_\alpha), \forall \alpha = 1, \dots, d.$$

The predictors X_α are obtained through the transformation $X_\alpha = 2.5 * \{\Phi(Z_\alpha) - 0.5\}$, where Φ is the standard normal distribution function and the variable $Z_\alpha \sim N(0, 1)$, $\alpha = 1, \dots, d$ with the correlation coefficients $\rho_{\alpha\beta} = \rho$, $\alpha \neq \beta$ for any pair of Z 's. Now the correlation between X 's is not ρ any more, it will depend on ρ . In order to validate the assumption that the density is bounded below from 0, we will focus on the estimation inside $[-1, 1]^d$.

Meanwhile, the error term ε follows standard normal distribution and is independent of \mathbf{X} . The conditional standard deviation function is defined by

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

By this choice of $\sigma(\mathbf{x})$, we ensure that our design is heteroscedastic, and the variance is roughly proportional to dimension d . This proportionality is intended to mimic the case when independent copies of the same kind of univariate regression problems are simply added together.

We now describe how the SBLLE estimator are implemented. The first step is to obtain the spline estimator of $\sum_{\alpha=1}^d m_\alpha(X_\alpha)$, using the truncated power B-spline basis as in (2.3). The selection of knots will uniquely define the basis. The knots number N_n will be determined by the sample size and two tuning constants, to be specific

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

in which $[c]$ denotes the integer part of c . In our simulation study, we have used $c_1 = 1 = c_2$. The choice of these constants c_1 and c_2 makes little difference for a large sample. But for small sample size, it does affect the performance to a degree. The additional constraint that

$N \leq (n/4 - 1)d^{-1}$ ensures that the number of terms in the linear least squares problem (2.3), $1 + dN_n$, is no greater than $n/4$, which is necessary when the sample size n is moderate and dimension d is high.

The oracle smoother $\tilde{m}_{K,1}(x_1)$ for comparison is obtained by local linear regression of the unobservable $m_1(X_1) + \sigma(\mathbf{X})\varepsilon$ on X_1 directly, while the oracle SBLl estimators $\hat{m}_{\text{SBLl},1}(x_1)$ are obtained by local linear regression of $\{\hat{Y}_{i1}, X_{i1}\}_{i=1}^n$. To save space, we only implement the local linear version of $\hat{m}_{\text{SBLl},1}(x_1)$, i.e., the SBLl estimator, using the XploRe quantlet “lprextest”. For information on XploRe, see Härdle, Hlávka and Klinke (2000) or visit <http://www.xploRe-stat.de>.

We have run $S = 500$ replications for sample sizes $n = 100, 200, 500$ and 1000 with correlation coefficient $\rho = 0, 0.3$ respectively. The dimensions are taken at $d = 4, 10$. The major objective of this section is to compare the relative efficiency of $\hat{m}_{s,\alpha}$ with respect to $\tilde{m}_{s,\alpha}$

$$\text{eff}_{\alpha,l} = \frac{\frac{1}{n} \sum_{i=1}^n \{\tilde{m}_{s,\alpha}(X_{i\alpha,l}) - m_\alpha(X_{i\alpha,l})\}^2 I_{\{|X_{i\alpha,l}| \leq 1\}}}{\frac{1}{n} \sum_{i=1}^n \{\hat{m}_{s,\alpha}(X_{i\alpha,l}) - m_\alpha(X_{i\alpha,l})\}^2 I_{\{|X_{i\alpha,l}| \leq 1\}}}, \alpha = 1, \dots, d, l = 1, \dots, S$$

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

in which $\{X_{i1,l}, \dots, X_{id,l}\}_{i=1}^n$ is the l -th sample, $l = 1, \dots, S$. Theorems 1 and 2 indicate that the efficiency should be close to 1. In particular, when we have an efficiency value bigger than 1, $\hat{m}_{s,\alpha}(x_\alpha)$ is a better estimator in the sense of mean square distance.

The corresponding mean and the standard error (in the parenthesis) of the relative efficiencies for first and third dimension ($\alpha = 1, 3$) is given in Table 1. For the case of $\rho = 0$, almost of all 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 $\tilde{m}_{s,\alpha}$ to the oracle estimator $\hat{m}_{s,\alpha}$, with faster convergence for a larger sample. At $\rho = 0$ and all the random selected directions, the SBLl performs 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 will treat less noise than the case of direct oracle estimator, the local linear estimator.

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 becomes bigger. Typically, the relative efficiencies are greater than 0.97 for $d = 4$ with sample size 200, and 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 strongly correlated. The standard errors in the parenthesis follow the same trend that less variation is with larger sample size, though it shows slower convergence compared to the case of $\rho = 0$, which is not unexpected.

In addition, several figures display the features of the relative efficiencies in details. In Figures 1 and 2 four types of line characteristics which 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 $\hat{m}_{\text{SBK},1}(x_1)$ and $\tilde{m}_{\text{K},1}(x_1)$. More efficiency values distributed around the vertical line would be confirmative to the conclusions of Theorems 1 and 2.

All the curves in Figures 1 and 2 are the density estimates of relative efficiency distributions for specific sample size n , correlation coefficient ρ and dimension d . With increasing sample sizes, we found that the relative efficiency are becoming closer to the vertical standard line, with narrower spread out. In addition, the curve with $\rho = 0$ shows a faster convergence to the vertical line than those with $\rho = 0.3$ in all cases. 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 above confirms the theorem that SBLL behaves similarly like the oracle local linear estimator.

We have done some more simulation with $d = 50$, and $S = 100$ replications for $\rho = 0, 0.3$, and $n = 500, 1000, 1500, 2000$, the results of which are graphically represented in Figures 3 and 4. The basic graphic pattern is similar to that for the lower dimensions $d = 4, 10$, though with slower convergence rate and relatively lower efficiency. The corresponding statistics are listed in Table 2.

d	n	eff ₁		eff ₃	
		$\rho = 0$	$\rho = 0.3$	$\rho = 0$	$\rho = 0.3$
4	100	1.015 (0.287)	0.958 (0.320)	1.000 (0.268)	0.926 (0.266)
	200	0.992 (0.126)	0.974 (0.164)	1.001 (0.133)	0.973 (0.153)
	500	0.993 (0.060)	0.990 (0.083)	0.995 (0.058)	0.990 (0.083)
	1000	0.998 (0.0416)	1.000 (0.060)	0.998 (0.042)	0.997 (0.057)
10	100	0.899 (0.648)	0.666 (0.597)	0.952 (0.832)	0.641 (0.552)
	200	1.026 (0.434)	0.818 (0.361)	1.045 (0.479)	0.826 (0.395)
	500	1.012 (0.145)	0.977 (0.171)	1.002 (0.138)	0.970 (0.182)
	1000	0.999 (0.078)	0.986 (0.104)	0.989 (0.082)	0.988 (0.105)

Table 1: Relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$ with 500 replications.

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 Rubinfeld (1978) to estimate the housing price index model. Breiman and Friedman (1985)

ρ	n	eff ₁	eff ₁₀	eff ₁₉	eff ₅₀
0	500	1.030 (0.830)	0.995 (0.778)	0.737 (0.567)	0.861 (0.648)
	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)
0.3	500	0.379 (0.297)	0.410 (0.408)	0.352 (0.296)	0.444 (0.721)
	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 2: Relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$ with 100 replications for $d = 50$.

did further analysis to deal with the multi-collinearity among the independent variables. By using a stepwise method, they proposed the alternating conditional expectation method to select a subset of the variables in order to maximize the correlation between the fitted value and the selected covariates. Four variables were selected by penalizing for overfitting. Opsomer and Ruppert (1998) illustrated their automated bandwidth selection for fitting additive models based on the selected four variables. We will use 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 %

One major concern is the big gap in the domain of variables TAX and LSTAT, which will cause severe trouble at the first stage of spline estimation. So logarithmic transformation is done for these two variables before fitting the model. We will fit 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 of each input variable.

In Figure 5, 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 (2.5). 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, $K(u) = \frac{15}{16}(1-u^2)^2$ for $-1 < u < 1$.

Besides the estimation of the component functions, we also use our proposed confidence bands to test the linearity of the components. In Figure 5 the straight solid lines are the regression lines with the least square coefficients. The first figure shows that the linearity null hypothesis $H_0 : m_1(\text{RM}) = a_1 + b_1 \cdot \text{RM}$, 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 $\log(\text{TAX})$ and $\log(\text{LSTAT})$ 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 standard Nadaraya-Watson 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 $\{\hat{Y}_{i1}, X_{i1}\}_{i=1}^n$. Although much mathematics is devoted to proving 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 $n = 500, 100, 1500, 2000$ and dimension as high as $d = 50$ takes about 40 minutes 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 $\tilde{m}_{K,1}(x_1)$ are computed. In addition, the SBK and SBLL estimators inherit the asymptotic confidence bands (2.9) of univariate Nadaraya-Watson and local linear estimators. The combination of speed and global accuracy for very high dimension regression is very appealing.

Appendix

A.1 Variance reduction

In this subsection we prove Propositions 1 and 2. The magnitude of the variance term $I(x_1)$ in (3.14) can be measured by its conditional second moment given $\mathbf{X}_1, \dots, \mathbf{X}_n$. Based on (3.12) and (3.14), the conditional second moment $E \left\{ I(x_1) | \tilde{\mathbf{X}} \right\}^2$ of $I(x_1)$ given $\tilde{\mathbf{X}} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$

is

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

It is clear that

$$E \left\{ I(x_1) | \tilde{\mathbf{X}} \right\}^2 = E \left\{ I_1^2(x_1) | \tilde{\mathbf{X}} \right\} - E \left\{ I_2^2(x_1) | \tilde{\mathbf{X}} \right\},$$

where for brevity, we write

$$I_1(x_1) = n^{-1} \sum_{l=1}^n K_h(X_{l1} - x_1) \tilde{\varepsilon}_2^*(X_{l2}), I_2(x_1) = n^{-1} \sum_{l=1}^n K_h(X_{l1} - x_1) \cdot n^{-1} \sum_{i=1}^n \tilde{\varepsilon}_2^*(X_{i2}). \quad (\text{A.1})$$

If further one denotes

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

then

$$I_1(x_1) = n^{-1} \sum_{l=1}^n K_h(X_{l1} - x_1) \sum_{J=1}^N \tilde{a}_{J,2} B_{J,2}(X_{l2}) = n^{-1} \sum_{l=1}^n \sum_{J=1}^N \tilde{a}_{J,2} \xi_J(\mathbf{X}_l, x_1). \quad (\text{A.3})$$

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 $\sum_{J=1}^N |\tilde{a}_{J,2}|$, in Lemma A.3, A.4 and A.7. Consequently, Lemma A.10 shows that $\sup_{x_1 \in [0,1]} E \left\{ I_1^2(x_1) | \tilde{\mathbf{X}} \right\} = O_p(n^{-1})$. In Lemma A.11 we have $\sup_{x_1 \in [0,1]} |I_2(x_1)| = O_p(Nn^{-1}\sqrt{\log n})$. Based on the selection of $N \sim n^{2/5} \log n$, Proposition 1 is thus proved.

There is one more Assumption (A2') in addition to Assumptions (A1) to (A6) in Lemma A.12. The order of $I_1(x_1)$ under the new restrictions is obtained uniformly over $[0, 1]$ inflated only by a factor of $\{\log(n)\}^{1/2}$ compared with the pointwise case, one has $\sup_{x_1 \in [0,1]} |I_1(x_1)| = O_p(\sqrt{\log(n)/n})$. Now again, due to the selection of the interval width $H \sim (n^{2/5} \log n)^{-1}$, the order $O_p(Nn^{-1}\sqrt{\log n})$ of $\sup_{x_1 \in [0,1]} |I_2(x_1)|$ in Lemma A.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 (3.15). We first cite one important result from page 149 of de Boor (2001).

Theorem A.1 *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 $\forall \alpha = 1, \dots, d$*

$$\|g_\alpha - m_\alpha\|_\infty \leq C_\infty H. \quad (\text{A.4})$$

Lemma A.1 Under Assumptions (A1), (A3) and (A6), for the spline function g_2 satisfying (A.4), one has

$$\sup_{x_1 \in [0,1]} \left| \frac{\sum_{i=1}^n K_h(X_{i1} - x_1) \{g_2(X_{i2}) - m_2(X_{i2})\}}{\sum_{i=1}^n K_h(X_{i1} - x_1)} \right| \leq C_\infty H, \quad (\text{A.5})$$

and for $\alpha = 1, 2$

$$|E_n g_\alpha(X_\alpha)| = \left| n^{-1} \sum_{i=1}^n g_\alpha(X_{i\alpha}) \right| = O_p(n^{-1/2} + H). \quad (\text{A.6})$$

PROOF. The first inequality (A.5) follows trivially from (A.4). 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 (A.4) leads to $|E_n g_\alpha(X_\alpha) - E_n m_\alpha(X_\alpha)| \leq C_\infty H$, while $E m_\alpha(X_\alpha) = 0$ leads to $E_n m_\alpha(X_\alpha) = O_p(n^{-1/2})$. Putting these together, one has $|E_n g_\alpha(X_\alpha)| \leq C_\infty H + O_p(n^{-1/2})$, which establishes (A.6). \square

In order to show that the bias term $II(x_1)$ defined in (3.15) is uniformly $o_p(n^{-2/5})$, the following lemma suffices.

Lemma A.2 Under Assumptions (A1) to (A6), as $n \rightarrow \infty$

$$\sup_{x_1 \in [0,1]} \left| \frac{\sum_{i=1}^n K_h(X_{i1} - x_1) \{\tilde{m}_2(X_{i2}) - g_2(X_{i2}) + E_n g_2(X_2)\}}{\sum_{i=1}^n K_h(X_{i1} - x_1)} \right| = O_p(n^{-1/2} + H). \quad (\text{A.7})$$

PROOF. Lemma A.1 and Lemma A.8 would entail that

$$\|\tilde{m} - g + E_n g_1(X_1) + E_n g_2(X_2)\|_2 = O_p(n^{-1/2} + H). \quad (\text{A.8})$$

To complete the proof of the lemma, we write

$$(\tilde{m} - g)(\mathbf{x}) + E_n g_1(X_1) + E_n g_2(X_2) = a + \sum_{\alpha=1}^2 \sum_{J=1}^N a_{J,\alpha} B_{J,\alpha}^*(x_\alpha),$$

where the empirically centered spline basis are $B_{J,\alpha}^*(x_\alpha) = B_{J,\alpha}(x_\alpha) - E_n B_{J,\alpha}(X_\alpha)$, $1 \leq J \leq N$, $1 \leq \alpha \leq 2$. Then for $\alpha = 1, 2$,

$$\tilde{m}_\alpha(x_\alpha) - g_\alpha(x_\alpha) + E_n g_\alpha(X_\alpha) = \sum_{J=1}^N a_{J,\alpha} B_{J,\alpha}^*(x_\alpha),$$

and according to (A.17) one has

$$\begin{aligned} & \|\tilde{m} - g + E_n g_1(X_1) + E_n g_2(X_2)\|_2^2 \\ & \geq c_0 \left[\left\{ a + \sum_{\alpha=1}^2 \sum_{J=1}^N a_{J,\alpha} E_n B_{J,\alpha}(X_\alpha) \right\}^2 + \sum_{\alpha=1}^2 \sum_{J=1}^N a_{J,\alpha}^2 \right]. \end{aligned} \quad (\text{A.9})$$

Now $n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) \{\tilde{m}_2(X_{i2}) - g_2(X_{i2}) + E_n g_2(X_2)\}$ is bounded by

$$\sum_{J=1}^N |a_{J,2}| \sup_{1 \leq J \leq N} \left| n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) B_{J,2}^*(X_{i2}) \right| \leq$$

$$\sum_{J=1}^N |a_{J,2}| \left\{ \sup_{1 \leq J \leq N} \left| n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) B_{J,2}(X_{i2}) \right| + \left| n^{-1} \sum_{i=1}^n K_h(X_{i1} - x_1) \right| \sup_{1 \leq J \leq N} |E_n B_{J,2}(X_2)| \right\}$$

which can be rewritten as the following according to the definitions of $\xi_J(\mathbf{X}_l, x_1)$ in (A.2) and of $A_{n,1}^*$ in (A.26)

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

Minkowski inequality, Lemma A.5, (A.27) and standard properties of kernel density estimator now imply that

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

$$\leq \sqrt{N \sum_{J=1}^N a_{J,2}^2} \left\{ O_p(\sqrt{H}) + O_p\left(\sqrt{\frac{\log n}{n}}\right) \right\} = O_p\left(\sqrt{\sum_{J=1}^N a_{J,2}^2}\right)$$

$$= O_p\left(\left[\left\{ \hat{a} + \sum_{\alpha=1}^2 \sum_{J=1}^N \hat{a}_{J,\alpha} E_n B_{J,\alpha}(X_\alpha) \right\}^2 + \sum_{\alpha=1}^2 \sum_{J=1}^N a_{J,\alpha}^2 \right]^{1/2}\right),$$

which according to (A.8) and (A.9) is

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

thus proving (A.7). □

Now combining Lemmas A.1 and A.2, one immediately gets

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

which establishes Proposition 3.

A.3 Technical lemmas

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

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

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

PROOF. Define for $\alpha = 1, 2, J = 1, \dots, N + 1$, $c_{J,\alpha} = \|I_{J,\alpha}\|_2^2 = \int I_{J,\alpha}^2(x_\alpha) f_\alpha(x_\alpha) dx_\alpha$, then $b_{J,\alpha}(x_\alpha)$ in (3.2) can be written as $b_{J,\alpha}(x_\alpha) = I_{J+1,\alpha}(x_\alpha) - c_{J+1,\alpha} I_{J,\alpha}(x_\alpha) / c_{J,\alpha}$ and

$$\|b_{J,\alpha}\|_2^2 = c_{J+1,\alpha} (1 + c_{J+1,\alpha}/c_{J,\alpha}), \forall \alpha = 1, 2, J = 1, \dots, N.$$

In Assumption (A3) the two positive constants c_f, C_f are the upper and lower bounds of all the marginal densities $f_\alpha(x_\alpha)$, then for all $J = 1, \dots, N + 1, \alpha = 1, 2$

$$c_f H \leq c_{J,\alpha} \leq C_f H. \quad (\text{A.10})$$

Then for all $\alpha = 1, 2, J = 1, \dots, N$, $\|b_{J,\alpha}\|_2^2 \sim H$, or specifically

$$c_f (1 + c_f/C_f) H \leq \|b_{J,\alpha}\|_2^2 \leq C_f (1 + C_f/c_f) H. \quad (\text{A.11})$$

The absolute expected value of $\xi_J(\mathbf{X}_l, x_1)$ is

$$|E\xi_J(\mathbf{X}_l, x_1)| \leq \int \int K_h(u_1 - x_1) |B_{J,2}(u_2)| f(u_1, u_2) du_1 du_2$$

The boundedness of the joint density f and the Lipschitz continuity of the kernel K will then imply that

$$\sup_{x_1 \in [0,1]} \sup_{1 \leq J \leq N} \int \int K(v_1) I_{J,2}(u_2) f(hv_1 + x_1, u_2) dv_1 du_2 \leq C_K C_f H,$$

the proof of the lemma is then completed. \square

Lemma A.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 $c_D n^6 \leq M_n \leq C_D n^6$, then under Assumptions (A3) to (A6)

$$\sup_{x_1 \in D_n} \sup_{1 \leq J \leq N} \left| n^{-1} \sum_{l=1}^n \{\xi_J(\mathbf{X}_l, x_1) - E\xi_J(\mathbf{X}_l, x_1)\} \right| = O_p\left((nh)^{-1/2} \log^{1/2} n\right). \quad (\text{A.12})$$

PROOF. For simplicity, denote $\xi_J^*(\mathbf{X}_l, x_1) = \xi_J(\mathbf{X}_l, x_1) - E\xi_J(\mathbf{X}_l, x_1)$. First we will compute the moments of the theoretical centered random variable $\xi_J^*(\mathbf{X}_l, x_1)$ for later use in Bernstein's inequality, $E\{\xi_J^*(\mathbf{X}_l, x_1)\}^2 = E\xi_J^2(\mathbf{X}_l, x_1) - \{E\xi_J(\mathbf{X}_l, x_1)\}^2$. Based on the bounded assumption of the joint density f and the Lipschitz continuity of the kernel K , there exist constants $c', C' > 0$, such that $c'h^{-1} \leq E\xi_J^2(\mathbf{X}_l, x_1) \leq C'h^{-1}$. 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 $\lim_{n \rightarrow \infty} b_n/a_n = 0$. Hence $E\{\xi_J^*(\mathbf{X}_l, x_1)\}^2 = E\xi_J^2(\mathbf{X}_l, x_1) - \{E\xi_J(\mathbf{X}_l, x_1)\}^2 \geq c^*h^{-1}$, for positive constant $c^* < c'$.

When $k \geq 3$, the k -th moment $E|\xi_J(\mathbf{X}_l, x_1)|^k$ can be bounded as follows

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

Lemma A.3 implies $|E\xi_J(\mathbf{X}_l, x_1)|^k \leq CH^{k/2}$, then $E|\xi_J(\mathbf{X}_l, x_1)|^k \gg |E\xi_J(\mathbf{X}_l, x_1)|^k$.

$$\begin{aligned} E|\xi_J^*(\mathbf{X}_l, x_1)|^k &\leq 2^{k-1} \left(E|\xi_J(\mathbf{X}_l, x_1)|^k + |E\xi_J(\mathbf{X}_l, x_1)|^k \right) \\ &\leq \{C_1 h^{-1} H^{-1/2}\}^{(k-2)} k! E|\xi_J^*(\mathbf{X}_l, x_1)|^2 \end{aligned}$$

then there exists such a constant $c = C_1 h^{-1} H^{-1/2}$ such that

$$E|\xi_J^*(\mathbf{X}_l, x_1)|^k \leq c^{k-2} k! E|\xi_J^*(\mathbf{X}_l, x_1)|^2,$$

that means the sequence of random variables $\{\xi_J^*(\mathbf{X}_l, x_1)\}_{l=1}^n$ satisfies the Cramér's condition, hence by the Bernstein's inequality we have

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

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

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

Borel-Cantelli Lemma implies (A.12). □

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

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

PROOF. Denote for $x \in [0, 1]$, $\Lambda(x) = \sup_{1 \leq J \leq N} |n^{-1} \sum_{l=1}^n \xi_J(\mathbf{X}_l, x)|$. If we choose the subset D_n as in Lemma A.4 to consist of equally spaced endpoints in $[0, 1]$, specifically

$$D_n = \{x_{1,k}, 0 \leq k \leq M_n; 0 = x_{1,0} < x_{1,1} < \dots < x_{1,M_n} = 1\},$$

then the consecutive endpoints make a total of M_n subintervals with length M_n^{-1} . Employing the discretization method, we have

$$\sup_{x_1 \in [0,1]} |\Lambda(x_1)| = \sup_{0 \leq k \leq M_n} |\Lambda(x_{1,k})| + \sup_{1 \leq k \leq M_n} \sup_{x_1 \in [x_{1,k-1}, x_{1,k}]} |\Lambda(x_1) - \Lambda(x_{1,k})|. \quad (\text{A.13})$$

We only need to bound the second term, as Lemmas A.3 and A.4, and the fact $H^{1/2} \gg \sqrt{\log n / (nh)}$ yield

$$\sup_{0 \leq k \leq M_n} |\Lambda(x_{1,k})| = \sup_{x_1 \in D_n} \sup_{1 \leq J \leq N} \left| n^{-1} \sum_{l=1}^n \xi_J(\mathbf{X}_l, x_1) \right| = O_p(H^{1/2}). \quad (\text{A.14})$$

Employing Lipschitz continuity of kernel K , one has

$$\sup_{1 \leq k \leq M_n} \sup_{x_1 \in [x_{1,k-1}, x_{1,k}]} |K_h(X_{l1} - x_1) - K_h(X_{l1} - x_{1,k})| \leq C_K M_n^{-1} h^{-2} \quad (\text{A.15})$$

Hence we have

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

$$\sup_{1 \leq k \leq M_n} \sup_{x_1 \in [x_{1,k-1}, x_{1,k}]} |\Lambda(x_1) - \Lambda(x_{1,k})| = O(M_n^{-1} h^{-2} H^{-1/2}) = o(n^{-1}) \quad (\text{A.16})$$

since $c_D n^6 \leq M_n \leq C_D n^6$ in Lemma A.4. The lemma follows instantly from (A.13), (A.14) and (A.16). \square

Lemma A.6 *Under Assumptions (A3) and (A6), there exist constants $C_0 > c_0 > 0$ such that*

$$c_0 \left(a_0^2 + \sum_{J,\alpha} a_{J,\alpha}^2 \right) \leq \left\| a_0 + \sum_{J,\alpha} a_{J,\alpha} B_{J,\alpha} \right\|_2^2 \leq C_0 \left(a_0^2 + \sum_{J,\alpha} a_{J,\alpha}^2 \right), \quad (\text{A.17})$$

for any $\mathbf{a} = (a_0, a_{1,1}, \dots, a_{N,1}, a_{1,2}, \dots, a_{N,2})^T \in R^{2N+1}$.

PROOF. According to Lemma 1 in Stone (1985), there exists a constant $c_0 > 0$ such that

$$\left\| a_0 + \sum_{J,\alpha} a_{J,\alpha} B_{J,\alpha} \right\|_2^2 \geq c_0 \left(a_0^2 + \left\| \sum_{J=1}^N a_{J,1} B_{J,1} \right\|_2^2 + \left\| \sum_{J=1}^N a_{J,2} B_{J,2} \right\|_2^2 \right),$$

If it can be proved that there exist constants $C'_0 > c'_0 > 0$ such that for $\alpha = 1, 2$

$$c'_0 \sum_{J=1}^N a_{J,\alpha}^2 \leq \left\| \sum_{J=1}^N a_{J,\alpha} B_{J,\alpha} \right\|_2^2 \leq C'_0 \sum_{J=1}^N a_{J,\alpha}^2, \quad (\text{A.18})$$

then (A.17) follows. To prove (A.18), the original B-Spline basis is employed. Without loss of generality we only provide the proof for $\alpha = 1$. We pick the constant basis $\{I_{J,1}(x_1)\}_{J=1}^{N+1}$ and represent the term $\sum_{J=1}^N a_{J,1} B_{J,1}(x_1)$ as follows

$$\sum_{J=1}^N a_{J,1} B_{J,1}(x_1) = \sum_{J=1}^{N+1} d_{J,1} I_{J,1}(x_1). \quad (\text{A.19})$$

Theorem 5.4.2 in Devore & Lorentz (1993) says that there is an equivalent relationship between the L_p ($p > 0$) norm of a B-spline function and the sequence of B-spline coefficients. To be specific, in our case, $\left\| \sum_{J=1}^{N+1} d_{J,1} I_{J,1} \right\|_{L_2}^2 = \int \left\{ \sum_{J=1}^{N+1} d_{J,1} I_{J,1}(x_1) \right\}^2 dx_1 = \sum_{J=1}^{N+1} d_{J,1}^2 H$. As in Assumption (A3) the joint density bounded between c_f and C_f , we have

$c_f \left\| \sum_{J=1}^{N+1} d_{J,1} I_{J,1} \right\|_{L_2}^2 \leq \left\| \sum_{J=1}^{N+1} d_{J,1} I_{J,1} \right\|_2^2 \leq C_f \left\| \sum_{J=1}^{N+1} d_{J,1} I_{J,1} \right\|_{L_2}^2$. The equality (A.19) and (A.11) leads to

$$c_d \sum_{J=1}^N a_{J,1}^2 H^{-1} \leq \sum_{J=1}^{N+1} d_{J,1}^2 \leq C_d \sum_{J=1}^N a_{J,1}^2 H^{-1},$$

for positive constants c_d and C_d . Therefore,

$$c'_0 \sum_{J=1}^N a_{J,1}^2 \leq \left\| \sum_{J=1}^N a_{J,1} B_{J,1} \right\|_2^2 = \left\| \sum_{J=1}^{N+1} d_{J,1} I_{J,1} \right\|_2^2 \leq C'_0 \sum_{J=1}^N a_{J,1}^2,$$

i.e. (A.18) holds given $c'_0 = c_f c_d$, $C'_0 = C_f C_d$. \square

Lemma A.7 *Under Assumptions (A1) to (A6), the least square solution $\tilde{\mathbf{a}}$ defined in (3.8) satisfies*

$$\tilde{\mathbf{a}}^T \tilde{\mathbf{a}} = \tilde{a}_0^2 + \sum_{J=1}^N \sum_{\alpha=1}^2 \tilde{a}_{J,\alpha}^2 = O_p \left(\frac{N}{n} \right). \quad (\text{A.20})$$

PROOF. According to (3.8), $\tilde{\mathbf{a}} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{E}$, then $\tilde{\mathbf{a}}^T \mathbf{B}^T \mathbf{B} \tilde{\mathbf{a}} = \tilde{\mathbf{a}}^T (\mathbf{B}^T \mathbf{E})$. Replacing $\mathbf{B}^T \mathbf{B}$ with matrix of the inner products $\langle B_{J,\alpha}, B_{J',\alpha'} \rangle_{2,n}$, as the matrix \mathbf{B} is given in (3.9), one has

$$\|\mathbf{B} \tilde{\mathbf{a}}\|_{2,n}^2 = \tilde{\mathbf{a}}^T \begin{pmatrix} \mathbf{1} \\ \langle B_{J,\alpha}, B_{J',\alpha'} \rangle_{2,n} \end{pmatrix} \tilde{\mathbf{a}} = \tilde{\mathbf{a}}^T (n^{-1} \mathbf{B}^T \mathbf{E}). \quad (\text{A.21})$$

Based on (A.17), one has

$$(1 - A_n) \|\mathbf{B} \tilde{\mathbf{a}}\|_{2,n}^2 = (1 - A_n) \left\| \tilde{a}_0 + \sum_{J,\alpha} \tilde{a}_{J,\alpha} B_{J,\alpha} \right\|_2^2 \geq c_0 (1 - A_n) \left(\tilde{a}_0^2 + \sum_{J,\alpha} \tilde{a}_{J,\alpha}^2 \right), \quad (\text{A.22})$$

where A_n is of order $o_p(1)$ in Lemma A.8. Meanwhile by the Cauchy-Schwartz inequality, the right hand side of (A.21) is bounded from above by

$$|\tilde{\mathbf{a}}^T (n^{-1} \mathbf{B}^T \mathbf{E})| \leq \left(\tilde{a}_0^2 + \sum_{J,\alpha} \tilde{a}_{J,\alpha}^2 \right)^{1/2} \{n^{-2} \mathbf{E}^T \mathbf{B} \mathbf{B}^T \mathbf{E}\}^{1/2}. \quad (\text{A.23})$$

Now (A.21), (A.22) and (A.23) will lead to

$$\tilde{a}_0^2 + \sum_{J,\alpha} \tilde{a}_{J,\alpha}^2 \leq c_0^{-2} (1 - A_n)^{-2} \{n^{-2} \mathbf{E}^T \mathbf{B} \mathbf{B}^T \mathbf{E}\}.$$

Note next that it is trivial to verify that $E \{n^{-2} \mathbf{E}^T \mathbf{B} \mathbf{B}^T \mathbf{E}\} = O(n^{-1} N)$. Therefore (A.20) holds. \square

Lemma A.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

$$A_n = \sup_{g_1, g_2 \in G^{(-1)}} \frac{|\langle g_1, g_2 \rangle_{2,n} - \langle g_1, g_2 \rangle_2|}{\|g_1\|_2 \|g_2\|_2} = O_p \left(\sqrt{\frac{\log n}{nH}} \right) = o_p(1). \quad (\text{A.24})$$

PROOF. Let

$$\begin{aligned} g_1(X_1, X_2) &= a_0 + \sum_{J=1}^N \sum_{\alpha=1}^2 a_{J,\alpha} B_{J,\alpha}(X_\alpha), \\ g_2(X_1, X_2) &= a'_0 + \sum_{J'=1}^N \sum_{\alpha'=1}^2 a'_{J',\alpha'} B_{J',\alpha'}(X_{\alpha'}), \end{aligned}$$

in which for any $J, J' = 1, \dots, N, \alpha, \alpha' = 1, 2$, $a_{J,\alpha}$ and $a'_{J',\alpha'}$ are real constants.

The difference between the empirical and theoretical inner products of g_1 and g_2 is $|\langle g_1, g_2 \rangle_{2,n} - \langle g_1, g_2 \rangle_2|$, which is bounded by

$$\begin{aligned} & \left| \sum_{J,\alpha} \langle a'_0, a_{J,\alpha} B_{J,\alpha} \rangle_{2,n} \right| + \left| \sum_{J',\alpha'} \langle a_0, a'_{J',\alpha'} B_{J',\alpha'} \rangle_{2,n} \right| \\ & + \sum_{J,J',\alpha,\alpha'} |a_{J,\alpha}| |a'_{J',\alpha'}| \left| \langle B_{J,\alpha}, B_{J',\alpha'} \rangle_{2,n} - \langle B_{J,\alpha}, B_{J',\alpha'} \rangle_2 \right|. \end{aligned} \quad (\text{A.25})$$

The equivalence of norms given in equation (A.17) leads to

$$\begin{aligned} & \left| \sum_{J,\alpha} \langle a'_0, a_{J,\alpha} B_{J,\alpha} \rangle_{2,n} \right| \leq A_{n,1}^* \cdot |a'_0| \cdot \left\| \sum_{J,\alpha} a_{J,\alpha} B_{J,\alpha} \right\|_2 \\ & \leq C_0 A_{n,1}^* |a'_0|^{1/2} \left(\sum_{J,\alpha} \tilde{a}_{J,\alpha}^2 \right)^{1/2} \leq C_{A,1} A_{n,1}^* \|g_1\|_2 \|g_2\|_2, \end{aligned}$$

where

$$A_{n,1}^* = \sup_{J,\alpha} \left| \langle 1, B_{J,\alpha} \rangle_{2,n} - \langle 1, B_{J,\alpha} \rangle_2 \right| = \sup_{J,\alpha} \left| \langle 1, B_{J,\alpha} \rangle_{2,n} \right|. \quad (\text{A.26})$$

Similarly it holds for the second term in (A.25) that

$$\left| \sum_{J,\alpha} \langle a_0, a'_{J',\alpha'} B_{J',\alpha'} \rangle_{2,n} \right| \leq C'_{A,1} A_{n,1}^* \|g_1\|_2 \|g_2\|_2.$$

It is easy to show by Bernstein's inequality that

$$A_{n,1}^* = \sup_{J,\alpha} \left| n^{-1} \sum_{i=1}^n B_{J,\alpha}(X_{i\alpha}) \right| = O_p \left(\sqrt{\log n/n} \right). \quad (\text{A.27})$$

The third term in (A.25) will be in probability less than

$$\begin{aligned} & \sum_{J,J',\alpha,\alpha'} |a_{J,\alpha}| |a'_{J',\alpha'}| \left| \langle B_{J,\alpha}, B_{J',\alpha'} \rangle_{2,n} - \langle B_{J,\alpha}, B_{J',\alpha'} \rangle_2 \right| \\ & \leq \sum_{J,J',\alpha,\alpha'} |a_{J,\alpha}| |a'_{J',\alpha'}| A_{n,2}^* \leq C_{A,2} A_{n,2}^* \left\{ \sum_{J,\alpha} a_{J,\alpha}^2 \right\}^{1/2} \left\{ \sum_{J',\alpha'} a'^2_{J',\alpha'} \right\}^{1/2} \\ & \leq C_{A,2} A_{n,2}^* \|g_1\|_2 \|g_2\|_2, \end{aligned}$$

where

$$A_{n,2}^* = \sup_{J,J',\alpha,\alpha'} \left| \langle B_{J,\alpha}, B_{J',\alpha'} \rangle_{2,n} - \langle B_{J,\alpha}, B_{J',\alpha'} \rangle_2 \right|.$$

Now since

$$\left| \langle g_1, g_2 \rangle_{2,n} - \langle g_1, g_2 \rangle_2 \right| \leq \{ (C_{A,1} + C'_{A,1}) A_{n,1}^* + C_{A,2} A_{n,2}^* \} \|g_1\|_2 \|g_2\|_2,$$

if we can show that

$$A_{n,2}^* = O_p \left(\sqrt{\log n / (nH)} \right), \quad (\text{A.28})$$

plus the fact that $\sqrt{\log n / (nH)} \gg \sqrt{\log n / n}$, based on the selection of $H^{-1} \sim n^{2/5} \log n$, then there exists a constant $C_A > 0$

$$\left| \langle g_1, g_2 \rangle_{2,n} - \langle g_1, g_2 \rangle_2 \right| \|g_1\|_2^{-1} \|g_2\|_2^{-1} \leq (C_{A,1} + C'_{A,1}) A_{n,1}^* + C_{A,2} A_{n,2}^* \leq C_A A_{n,2}^*,$$

the order $O_p \left(\sqrt{\log n / (nH)} \right)$ of A_n will be established as in the statement (A.24).

The proof of (A.28) could be proved case by case with various α, α', J and J' , via Bernstein's inequality. For details, please see www.stt.msu.edu/~yangli/SBK.pdf. \square

The next lemma on the positive definiteness of matrix $(n^{-1} \mathbf{B}^T \mathbf{B})^{-1}$ is a sufficient step to achieve Lemma A.10.

Lemma A.9 *Under Assumptions (A3) and (A4), for the matrix $S = (s_{jj'})_{j,j'=1}^{dN+1} = (n^{-1} \mathbf{B}^T \mathbf{B})^{-1}$, there exist constants $C_S > c_S > 0$ such that with probability approaching to 1, one has*

$$c_S I_{2N+1} \leq S^{-1} \leq C_S I_{2N+1}. \quad (\text{A.29})$$

PROOF. Take a real vector $\boldsymbol{\varsigma} = (u_0, u_{1,1}, \dots, u_{N,1}, u_{1,2}, \dots, u_{N,2})^T \in R^{2N+1}$, one has

$$\|\boldsymbol{\varsigma}^T B_*\|_{2,n}^2 = \boldsymbol{\varsigma}^T \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \langle B_{J,\alpha}, B_{J',\alpha'} \rangle_{2,n} \end{pmatrix} \boldsymbol{\varsigma} = \boldsymbol{\varsigma}^T S^{-1} \boldsymbol{\varsigma}, \quad (\text{A.30})$$

where we denote $B_* = \{1, B_{1,1}(X_1), \dots, B_{N,2}(X_2)\}^T$. Meanwhile, the definition of A_n in (A.24) entails in particular that

$$\|\boldsymbol{\varsigma}^T B_*\|_2^2 (1 + A_n) \geq \|\boldsymbol{\varsigma}^T B_*\|_{2,n}^2 \geq \|\boldsymbol{\varsigma}^T B_*\|_2^2 (1 - A_n),$$

while (A.17) means that there exist constants $C_S > c_S > 0$ such that

$$C_S \cdot \boldsymbol{\varsigma}^T \boldsymbol{\varsigma} \geq \|\boldsymbol{\varsigma}^T B_*\|_2^2 \geq c_S \cdot \boldsymbol{\varsigma}^T \boldsymbol{\varsigma},$$

hence

$$C_S \cdot \boldsymbol{\varsigma}^T \boldsymbol{\varsigma} (1 + A_n) \geq \|\boldsymbol{\varsigma}^T B_*\|_{2,n}^2 \geq c_S \cdot \boldsymbol{\varsigma}^T \boldsymbol{\varsigma} (1 - A_n). \quad (\text{A.31})$$

Putting together (A.30), (A.31), one concludes that with probability approaching 1

$$C_S \cdot \boldsymbol{\varsigma}^T \boldsymbol{\varsigma} \geq \boldsymbol{\varsigma}^T S^{-1} \boldsymbol{\varsigma} \geq c_S \cdot \boldsymbol{\varsigma}^T \boldsymbol{\varsigma},$$

which gives (A.29). \square

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

$$\sup_{x_1 \in [0,1]} E \left\{ I_1^2(x_1) \mid \tilde{\mathbf{X}} \right\} = O_p(n^{-1}). \quad (\text{A.32})$$

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

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

Based on Assumption (A2), we have $E \{ (\mathbf{E} \cdot \mathbf{E}^T) \mid \mathbf{X}_1, \dots, \mathbf{X}_n \} \leq C_\sigma^2 I_n$ in the matrix sense. It is known that $\tilde{\mathbf{a}} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{E}$, then applying these two matrices to a quadratic form with vector $\left\{ \mathbf{B} (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{P}_{0_{N+1}, I_N} \mathbf{B}^T e_{l'} \right\}$, one has

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

where the $s_{J+N+1, J'+N+1}$'s are elements of S in Lemma A.9. Plugging in the above term, and employing (A.3), $E \left\{ I_1^2(x_1) \mid \tilde{\mathbf{X}} \right\}$ is less than

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

where C_S is the same as in (A.29). Now using Lemma A.5, one has with probability approaching to 1

$$\sup_{x_1 \in [0,1]} E \left\{ I_1^2(x_1) \mid \tilde{\mathbf{X}} \right\} \leq C_\sigma^2 n^{-1} C_S \sum_{1 \leq J \leq N} H = C n^{-1},$$

which implies (A.32). \square

Lemma A.11 Under Assumptions (A1) to (A6), for $I_2(x_1)$ as defined in (A.1), one has

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

PROOF. Based on (3.11), $n^{-1} \sum_{i=1}^n \tilde{\varepsilon}_2^*(X_{i2}) = \sum_{J=1}^N \tilde{a}_{J,2} \{ n^{-1} \sum_{i=1}^n B_{J,2}(X_{i2}) \}$. Lemma A.7 helps to get

$$\left| \sum_{J=1}^N \tilde{a}_{J,2} \right| \leq \left\{ N \cdot \sum_{J=1}^N \tilde{a}_{J,2}^2 \right\}^{1/2} \leq \{ N \cdot \tilde{\mathbf{a}}^T \tilde{\mathbf{a}} \}^{1/2} = O_p(N n^{-1/2}).$$

Now it is clear from (A.26) and (A.27) that

$$\sup_{1 \leq J \leq N} \left| n^{-1} \sum_{i=1}^n B_{J,2}(X_{i2}) \right| \leq A_{n,1}^* = O_p \left(\sqrt{n^{-1} \log n} \right),$$

hence

$$n^{-1} \sum_{i=1}^n \tilde{\varepsilon}_2^*(X_{i2}) \leq \left| \sum_{J=1}^N \tilde{a}_{J,2} \right| \sup_{1 \leq J \leq N} \left| n^{-1} \sum_{i=1}^n B_{J,2}(X_{i2}) \right| = O_p \left(\frac{N}{n} \sqrt{\log n} \right). \quad (\text{A.33})$$

By Assumption (A4) on the kernel function K , standard theory on kernel density estimation entails that $\sup_{x_1 \in [0,1]} |n^{-1} \sum_{l=1}^n K_h(X_{l1} - x_1)| = O_p(1)$. Thus with (A.33) the lemma follows immediately. \square

Lemma A.12 *Under Assumptions (A1) to (A6) and (A2'), and with $I_1(x_1)$ defined in (A.1), one has*

$$\sup_{x_1 \in [0,1]} |I_1(x_1)| = \sup_{x_1 \in [0,1]} \left| n^{-1} \sum_{l=1}^n K_h(X_{l1} - x_1) \cdot \tilde{\varepsilon}_2^*(X_{l2}) \right| = O_p \left(\sqrt{\log n/n} \right). \quad (\text{A.34})$$

PROOF. The discretization idea will be employed again in this lemma, by dividing the interval $[0, 1]$ into M_n equally spaced intervals with disjoint endpoints $0 = x_{1,0} < x_{1,1} < \dots < x_{1,M_n} = 1$. As in (A.13), we start with

$$\sup_{x_1 \in [0,1]} |I_1(x_1)| = \sup_{0 \leq k \leq M_n} |I_1(x_{1,k})| + \sup_{1 \leq k \leq M_n} \sup_{x_1 \in [x_{1,k-1}, x_{1,k}]} |I_1(x_1) - I_1(x_{1,k})|. \quad (\text{A.35})$$

Note that for any $x_1 \in [0, 1]$, (3.11) and (A.1) imply that

$$\begin{aligned} \tilde{\varepsilon}_2^*(X_{l2}) &= (e_l^T \mathbf{B}) \mathbf{P}_{0_{N+1}, I_N} (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{E} \\ I_1(x_1) &= n^{-1} \sum_{l=1}^n K_h(X_{l1} - x_1) (e_l^T \mathbf{B}) \mathbf{P}_{0_{N+1}, I_N} (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{E}. \end{aligned}$$

Since $I_1(x_1)$ is a linear combination of the noise terms in \mathbf{E} , its conditional distribution given $\tilde{\mathbf{X}}$ is normal with mean 0, under Assumption (A2'). Let

$$R(\tilde{\mathbf{X}}, x_{1,k}) = \left(\text{var} \left\{ I_1(x_{1,k}) \mid \tilde{\mathbf{X}} \right\} \right)^{-1/2} I_1(x_{1,k}),$$

then the conditional distribution of $R(\tilde{\mathbf{X}}, x_{1,k})$ given $\tilde{\mathbf{X}}$ is standard normal. In what follows, we use the well-known tail property of the normal distribution, i.e. $1 - \Phi(x) \leq \phi(x)/x$, for $x \geq 0$, hence there exists some $c > 0$, such that $1 - \Phi(x) \leq c\phi(x)$ for large x , where $\Phi(x)$ and $\phi(x)$ are the cumulative distribution function and the density function of the standard normal. Take $t_n = \sqrt{16 \log n}$, then there exists a constant c such that for large enough n

$$\begin{aligned} &\sum_{n=1}^n P \left\{ \sup_{0 \leq k \leq M_n} \left| R(\tilde{\mathbf{X}}, x_{1,k}) \right| \geq t_n \mid \tilde{\mathbf{X}} \right\} = \sum_{n=1}^n P \left\{ \sup_{0 \leq k \leq M_n} |Z| \geq t_n \right\} \\ &\leq \sum_{n=1}^n M_n \cdot P \left\{ |Z| \geq t_n \right\} \leq c \sum_{n=1}^n M_n \cdot \exp \left\{ -t_n^2/2 \right\} < \infty, \end{aligned}$$

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

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

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

$$\text{var} \left\{ I_1(x_{1,k}) \mid \tilde{\mathbf{X}} \right\} = E \left[\left\{ I_1(x_{1,k}) - E I_1(x_{1,k}) \right\}^2 \mid \tilde{\mathbf{X}} \right] = E \left\{ I_1^2(x_{1,k}) \mid \tilde{\mathbf{X}} \right\}.$$

Now Lemma A.10 implies that $\sup_{0 \leq k \leq M_n} \text{var} \left\{ I_1(x_{1,k}) \mid \tilde{\mathbf{X}} \right\} = O_p(n^{-1})$. Therefore

$$\sup_{0 \leq k \leq M_n} |I_1(x_{1,k})| \leq \sup_{0 \leq k \leq M_n} \left| R \left(\tilde{\mathbf{X}}, x_{1,k} \right) \right| \sup_{0 \leq k \leq M_n} \sqrt{\text{var} \left\{ I_1(x_{1,k}) \mid \tilde{\mathbf{X}} \right\}} = O_p \left(\sqrt{n^{-1} \log n} \right). \quad (\text{A.36})$$

Next, with (3.11), (A.15), and (A.20), it leads to

$$\sup_{1 \leq k \leq M_n} \sup_{x_1 \in [x_{1,k-1}, x_{1,k}]} |I_1(x_1) - I_1(x_{1,k})| = O_p \left(M_n^{-1} h^{-2} N \cdot N^{1/2} n^{-1/2} \right) = o_p(n^{-1}). \quad (\text{A.37})$$

due to the choice of $c_D n^6 \leq M_n \leq C_D n^6$ in Lemma A.4.

Now (A.35), (A.36) and (A.37) establish the lemma. \square

Acknowledgements

This research is part of the first author's dissertation under the supervision of the second author and has been supported in part by NSF grants DMS 0405330, BCS 0308420 and SES 0127722. The authors gratefully acknowledge the financial and staff support provided by Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, People's Republic of China, during the second author's visit.

References

- Andrews, D. and Whang, Y. (1990) Additive interactive regression models: circumvention of the curse of the dimensionality. *Econometric Theory*, **6**, 466–479.
- Breiman, L. and Friedman, J.H. (1985) Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.*, **80**, 580–619.
- Bickel, P.J. and Rosenblatt, M. (1973) On some global measures of the deviations of density function estimates. *Ann. Statist.*, **1**, 1071–1095.
- Claeskens, G. and Van Keilegom, I. (2003) Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.*, **31**, 1852–1884.

- de Boor, C. (2001) *A Practical Guide to Splines*. New York: Springer-Verlag.
- DeVore, R.A. and Lorentz, G.G. (1993) *Constructive Approximation*. Berlin: Springer-Verlag.
- Fan, J. and Chen, J. (1999) One-step local quasi-likelihood estimation. *J. Roy. Statist. Soc. Ser. B*, **61**, 927–934.
- Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.
- Fan, J., Härdle, W. and Mammen, E. (1998) Direct estimation of low-dimensional components in additive models. *Ann. Statist.*, **26**, 943–971.
- Hall, P. and Titterton, D. M. (1988) On confidence bands in nonparametric density estimation and regression. *J. Multivariate Anal.*, **27**, 228–254.
- Härdle, W. (1989) Asymptotic maximal deviation of M-smoothers. *J. Multivariate Anal.*, **29**, 163–179.
- Härdle, W. (1990) *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
- Härdle, W., Hlávka, Z. and Klinke, S. (2000) *XploRe Application Guide*. Berlin: Springer-Verlag.
- Härdle, W., Huet, S., Mammen, E., and Sperlich, S. (2004) Bootstrap inference in semi-parametric generalized additive models. *Econometric Theory*, **20**, 265–300.
- Härdle, W., Sperlich, S. and Spokoiny, V. (2001) Structural tests in additive regression. *J. Amer. Statist. Assoc.*, **96**, 1333–1347.
- Harrison, D. and Rubinfeld, D.L. (1978) Hedonic housing prices and the demand for cleaning air. *Journal of Economics and Management*, **5**, 81–102.
- Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*. London: Chapman and Hall.
- Huang, J.Z. (1998) Projection estimation in multiple regression with application to functional ANOVA models. *Ann. Statist.*, **26**, 242–272.
- Huang, J.Z. (2003) Local asymptotics for polynomial spline regression. *Ann. Statist.*, **31**, 1600–1635.
- Huang, J.Z. and Yang, L. (2004) Identification of nonlinear additive autoregression models. *J. Roy. Statist. Soc. Ser. B*, **66**, 463–477.

- Kim, W., Linton, O.B., and Hengartner, N. (1999) A Computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *J. Comput. Graph. Statist.*, **8**, 278–297.
- Linton, O.B. and Nielsen, J.P. (1995) Estimating structured nonparametric regression models by the kernel method. *Biometrika*, **82**, 93–101.
- Linton, O.B. and 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. and Nielsen, J. (1999) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, **27**, 1443–1490.
- Nielsen, J.P. and Sperlich, S. (2005) Smooth backfitting in practice, *J. Roy. Statist. Soc. Ser. B*, **67**, 43–61.
- Opsomer, J.D. (2000) Asymptotic properties of backfitting estimators. *J. Multivariate Anal.*, **73**, 166–179.
- Opsomer, J.D. and Ruppert, D. (1997) Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.*, **25**, 186–211.
- Opsomer, J.D. and Ruppert, D. (1998) A Fully automated bandwidth selection method for fitting additive models. *J. Amer. Statist. Assoc.*, **93**, 605–619.
- Sperlich, S., Tjøstheim, D. and 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. *Ann. Statist.*, **13**, 689–705.
- Stone, C.J. (1994) The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, **22**, 118–184.
- Tjøstheim, D. and Auestad, B. (1994) Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.*, **89**, 1398–1409.
- Tusnády, G. (1977) A remark on the approximation of the sample df in the multidimensional case. *Period. Math. Hungarian*, **8**, 53–55.
- Wang, J. and Yang, L. (2006) Polynomial spline confidence bands for regression curves. *Ann. Statist.*, **tentatively accepted**.

- Xia, Y. (1998) Bias-corrected confidence bands in nonparametric regression. *J. Roy. Statist. Soc. Ser. B*, **60**, 797–811.
- Xue, L. and Yang, L. (2006) Estimation of semiparametric additive coefficient model. *J. Statist. Plan. Infer.*, **136**, 2506–2534.
- Yang, L., Härdle, W. and Nielsen, J.P. (1999) Nonparametric autoregression with multiplicative volatility and additive mean. *J. Time Ser. Anal.*, **20**, 579–604.
- Yang, L., Sperlich, S. and Härdle, W. (2003) Derivative estimation and testing in generalized additive models. *J. Statist. Plan. Infer.*, **115**, 521–542.

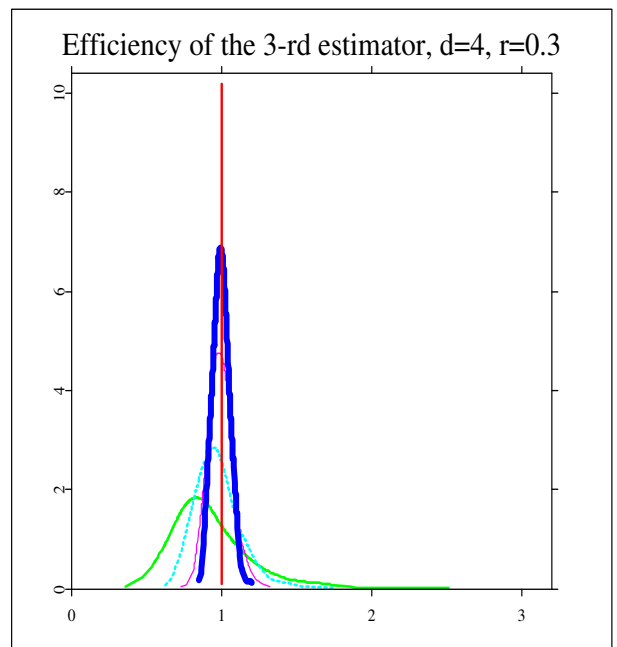
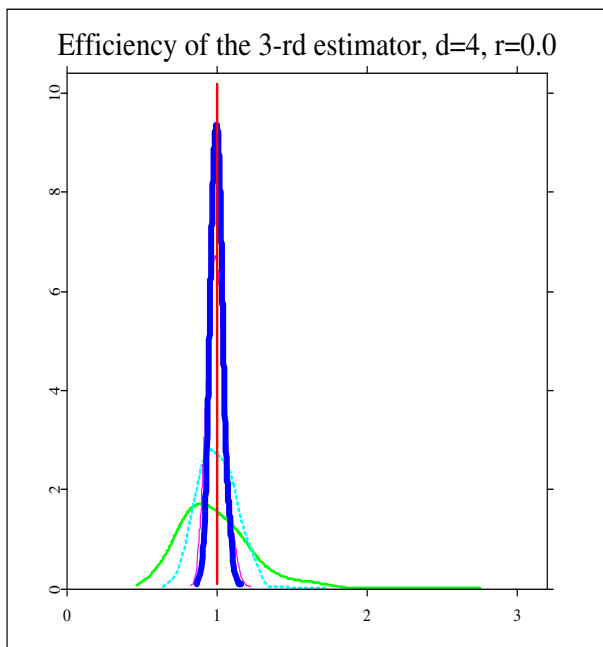
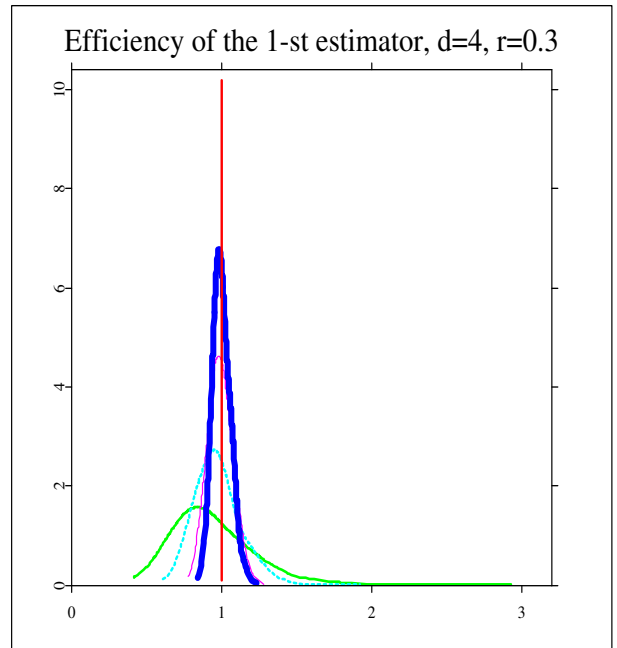
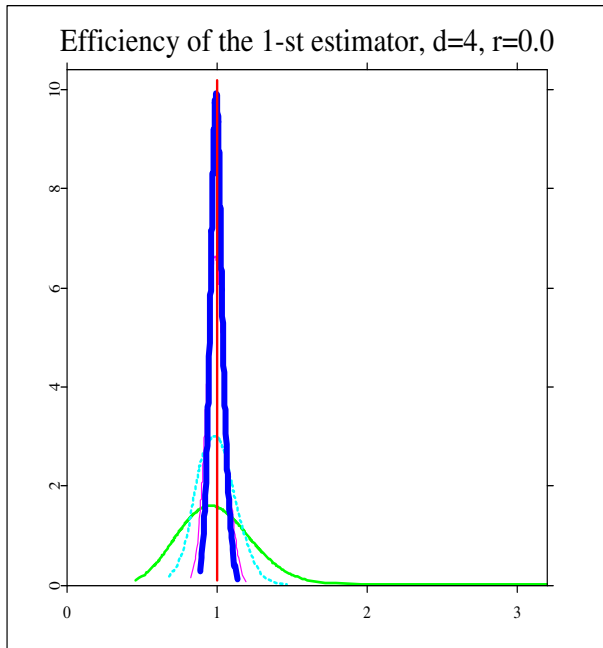


Figure 1: Empirical distribution of relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$, $d = 4$.

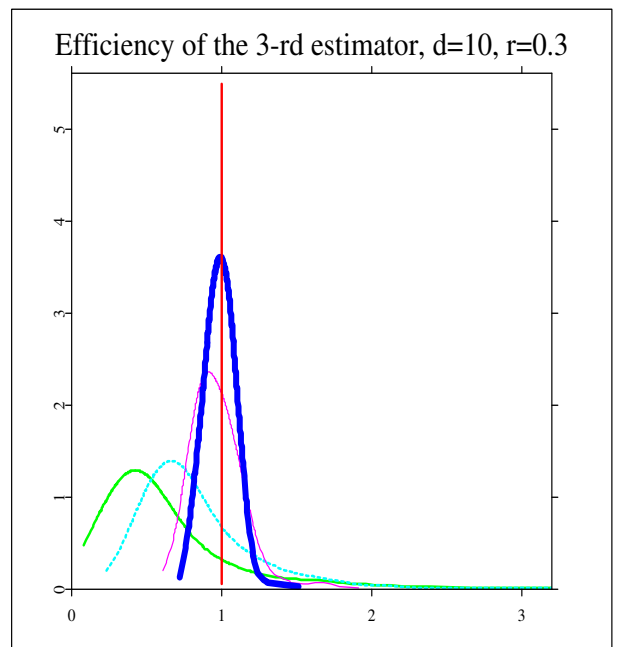
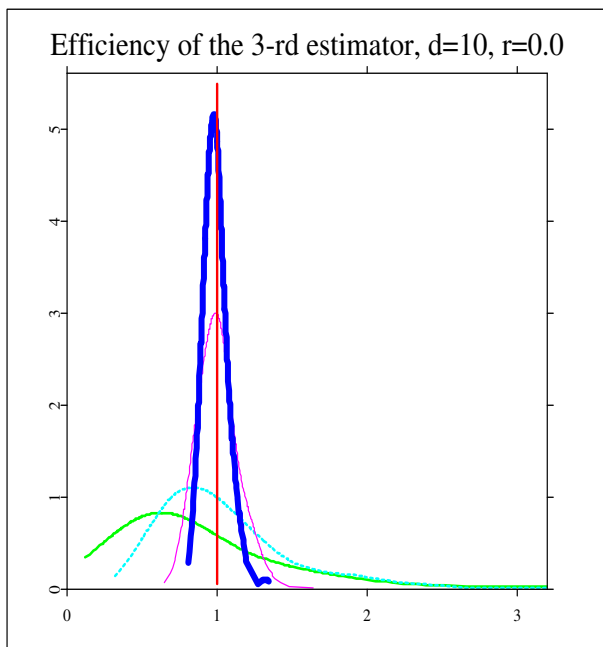
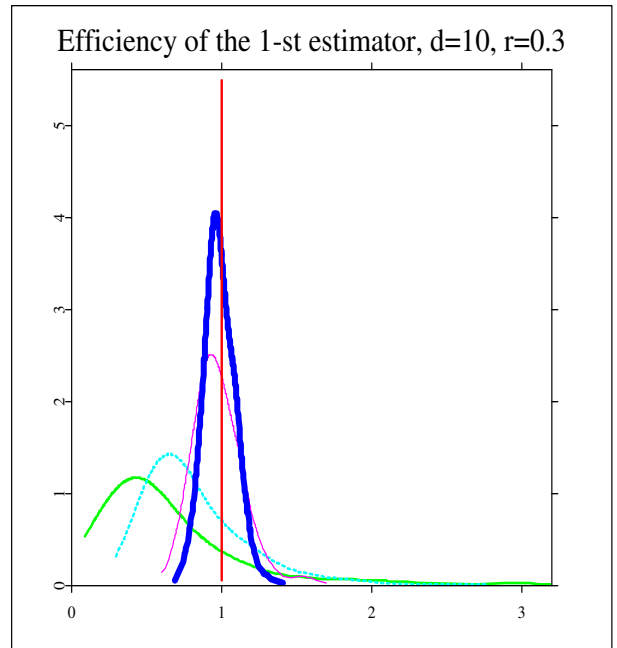
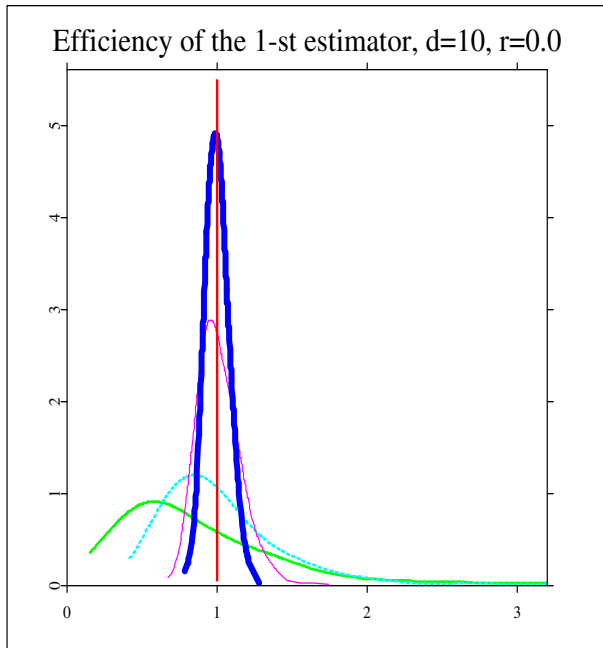


Figure 2: Empirical distribution of relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$, $d = 10$.

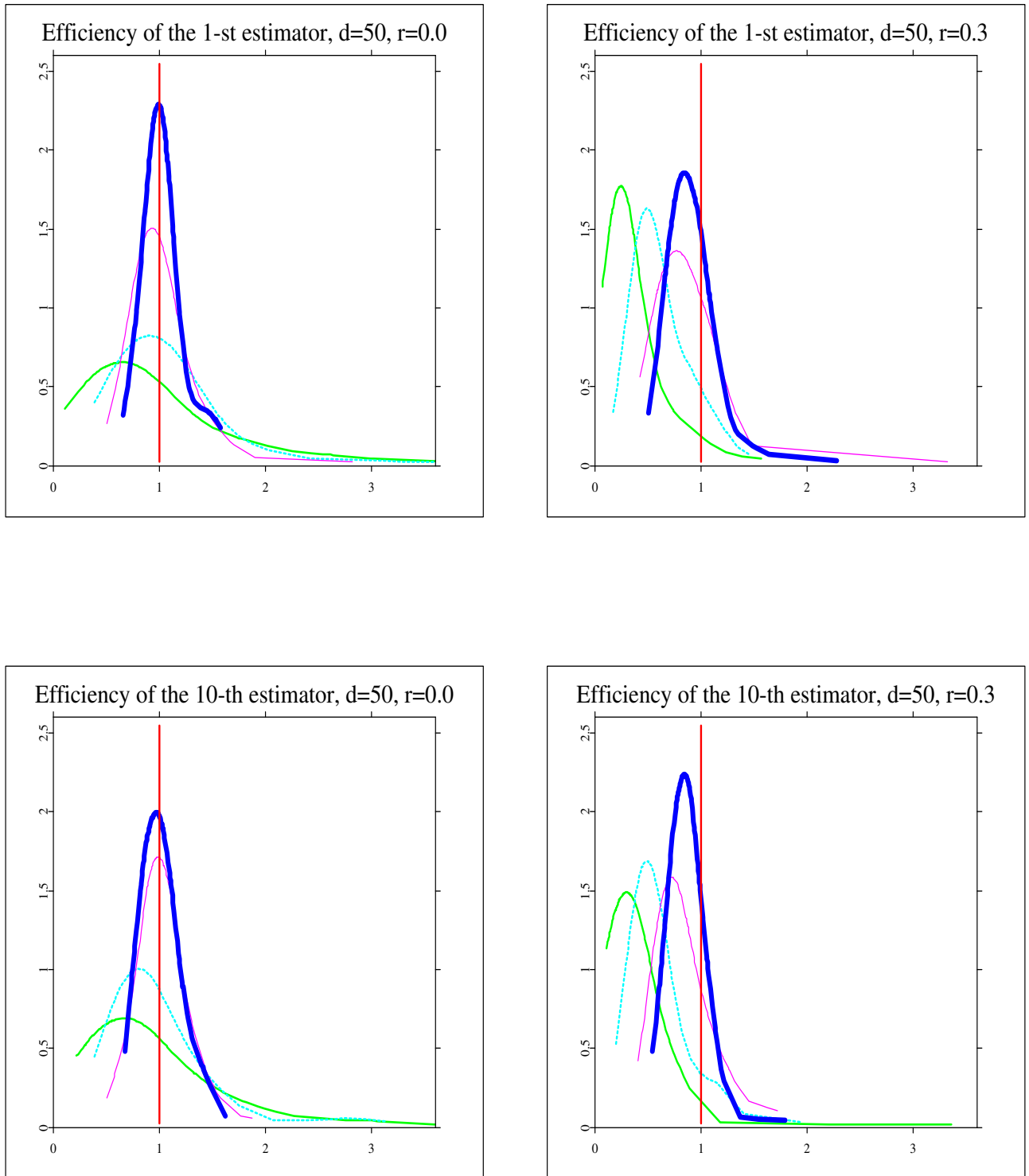


Figure 3: Empirical distribution of relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$, $d = 50, \alpha = 1, 10$.

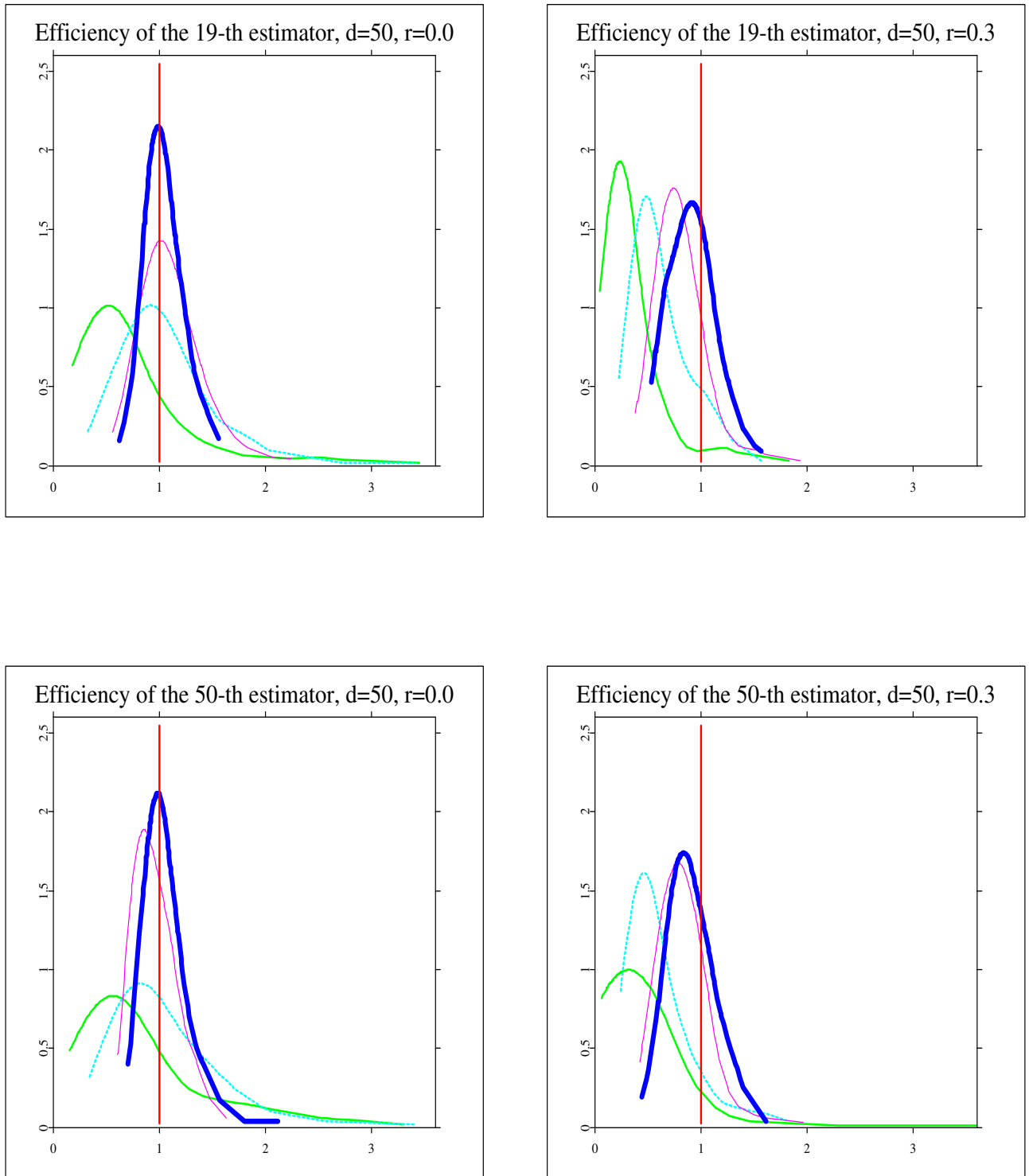


Figure 4: Empirical distribution of relative efficiency of $\tilde{m}_{s,\alpha}$ against $\hat{m}_{s,\alpha}$, $d = 50$, $\alpha = 19, 50$.

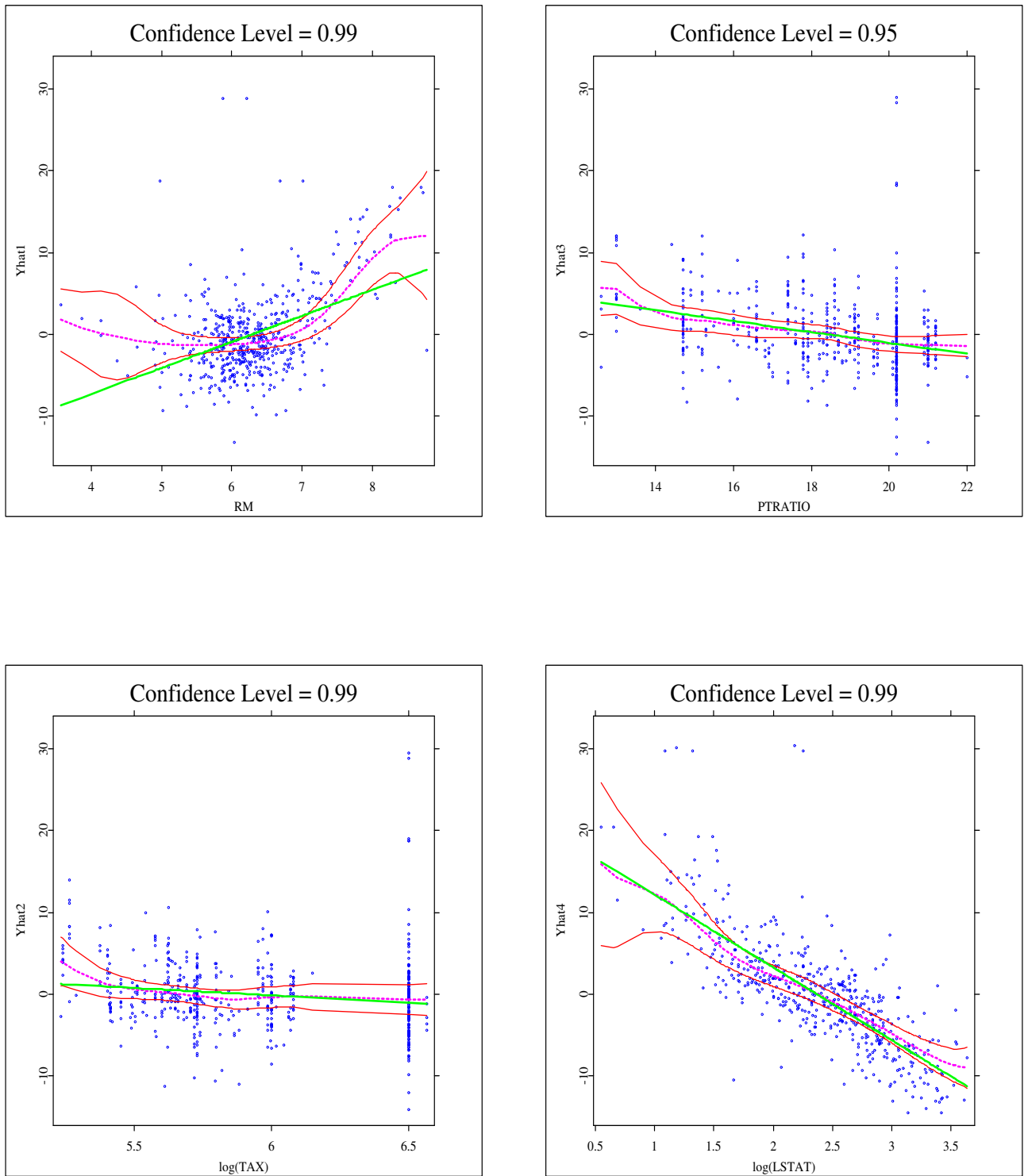


Figure 5: Linearity test for the Boston housing data. Plots of null hypothesis curves of $H_0 : m_\alpha(x_\alpha) = a_\alpha + b_\alpha \cdot x_\alpha, \alpha = 1, 2, 3, 4$ (solid line), linear confidence bands (upper and lower thin lines), the linear spline estimator (dotted line) and the data (circle).