TEST (2015) 24:632–655
DOI 10.1007/s11749-015-0427-5

ORIGINAL PAPER

A smooth simultaneous confidence band for conditional variance function

Li Cai · Lijian Yang

Received: 21 August 2014 / Accepted: 7 January 2015 / Published online: 22 January 2015

Abstract A smooth simultaneous confidence band (SCB) is obtained for heteroscedastic variance function in nonparametric regression by applying spline regression to the conditional mean function followed by Nadaraya–Waston estimation using the squared residuals. The variance estimator is uniformly oracally efficient, that is, it is as efficient as, up to order less than $n^{-1/2}$, the infeasible kernel estimator when the conditional mean function is known, uniformly over the data range. Simulation experiments provide strong evidence that confirms the asymptotic theory while the computing is extremely fast. The proposed SCB has been applied to test for heteroscedasticity in the well-known motorcycle data and Old Faithful geyser data with different conclusions.

Keywords B spline · Confidence band · Heteroscedasticity · Infeasible estimator · Knots · Nadaraya–Waston estimator · Variance function

Mathematics Subject Classification 62G05 · 62G08 · 62G10 · 62G15 · 62G32

1 Introduction

Conditional variance function is an important ingredient in regression analysis, as many statistical applications require knowledge of the variance function, such as weighted least squares estimation of the mean function and construction of confidence intervals/bands for the mean function. Compared to mean function estimation, the literature on the estimation of variance function is rather sparse. Fan and Yao

L. Cai · L. Yang (✉)
Center for Advanced Statistics and Econometrics Research,
Soochow University, Suzhou 215006, China
e-mail: yanglijian@suda.edu.cn

(1998) proved efficiency of the residual-based kernel variance estimator, Müller and Stadtmüller (1987) and more recently, Levine (2006) and Brown and Levine (2007) proposed difference-based kernel estimator and obtained its asymptotic normality, while Wang et al. (2008) derived the minimax rate of convergence for variance function estimation and constructed minimax rate optimal kernel estimators. For applications of variance estimation in the analysis of assay and microarray data, see Davidian et al. (1988) and Cai and Wang (2008).

Existing literature had mostly overlooked one crucial aspect of the problem, that is, simultaneous confidence band (SCB) for the variance function, which is an extremely powerful tool for inference on the global shape of curves, see for instance, Bickel and Rosenblatt (1973), Hall and Titterington (1988), Härdle (1989), Xia (1998), Claeskens and Van Keilegom (2003), Ma et al. (2012), Wang et al. (2014), Zheng et al. (2014) for theoretical works on SCB. This paper provides a spline–kernel two-step estimator of the variance function that is oracally efficient and comes equipped with a smooth SCB that substantially improves over the spline SCB of Song and Yang (2009), both theoretically and computationally.

To describe the problem, let observations $\{(X_i, Y_i)\}_{i=1}^n$ and unobserved errors $\{\varepsilon_i\}_{i=1}^n$ be i.i.d copies of (X, Y, ε) satisfying the regression model

$$Y = m(X) + \varepsilon, \tag{1}$$

where $\mathsf{E}(\varepsilon \mid X) = 0$, $\mathsf{E}\left(\varepsilon^2 \mid X\right) = \sigma^2(X)$, and the conditional mean function $m(x)$ and variance function $\sigma^2(x)$, defined on a compact interval $[a, b]$, are unknown. Note that with squared errors $Z_i = \varepsilon_i^2, 1 \le i \le n, \mathsf{E}(Z_i \mid X_i) = \sigma^2(X_i)$, hence the variance function $\sigma^2(x)$ is in fact the conditional mean function of Z_i on X_i. If $\sigma^2(\cdot)$ is constant, the model is homoscedastic, otherwise heteroscedastic, see Dette and Munk (1998) for testing of heteroscedasticity, Carroll and Ruppert (1988), Akritas and Van Keilegom (2001), Cai and Wang (2008) for regression methods in the presence of heteroscedastic errors, and Hall and Marron (1990) for rate-optimal estimator of homoscedastic variance.

Suppose for the sake of discussion that the mean function $m(x)$ were known by "oracle", one could obtain a new data set $\{(X_i, Z_i)\}_{i=1}^n$, in which $Z_i = \{Y_i - m(X_i)\}^2, 1 \le i \le n$, and estimate the function $\sigma^2(x)$ by a regressor $\tilde{\sigma}^2(x)$ of the Z_i's on the X_i's, a would-be estimator called "infeasible estimator" as it is based on unavailable knowledge, serves as a useful benchmark against which feasible ones can be compared to. Fan and Yao (1998) had obtained two-step estimator $\hat{\sigma}^2(x)$ of $\sigma^2(x)$ by local linear regression of $\hat{Z}_i = \{Y_i - \hat{m}(X_i)\}^2$ on X_i, in which $\hat{m}(x)$ is a first-step local linear estimator of $m(x)$, and shown that for any fixed $x \in (a, b)$, $\hat{\sigma}^2(x)$ was asymptotically as efficient as the "infeasible local linear estimator" $\tilde{\sigma}^2(x)$. Since this efficiency was merely pointwise, it allowed only the construction of confidence interval for $\sigma^2(x)$ at a single point x, not at every point $x \in [a, b]$ with simultaneous coverage, see also Hall and Carroll (1989) for the negligible effect of mean on the estimation of variance function.

Song and Yang (2009) had formulated a two-step estimator $\hat{\sigma}^2(x)$ of $\sigma^2(x)$ by spline regression of $\hat{Z}_i = \{Y_i - \hat{m}(X_i)\}^2$ on X_i, in which $\hat{m}(x)$ is a first step spline estimator of $m(x)$, and established asymptotic efficiency of $\hat{\sigma}^2(x)$ relative to an "infeasible spline estimator" $\tilde{\sigma}^2(x)$ over the data range $[a, b]$, and as a result an SCB was obtained for the whole variance curve as formulated in Wang and Yang (2009). There are some serious theoretical shortcomings, however with the shortcomings, however, with the spline SCB of Wang and Yang (2009) and hence also of Song and Yang (2009): the constant spline SCB is too wide and inaccurate; the linear spline SCB is narrow but its coverage probability is higher than the nominal level.

We propose a two-step estimator of $\sigma^2(x)$ by spline estimator $\hat{m}(x)$ of $m(x)$ in step one and kernel estimator $\hat{\sigma}^2(x)$ of $\sigma^2(x)$ in step two, which is uniformly as efficient as the infeasible kernel estimator, and hence oracally efficient. It is smooth as it comes from kernel smoothing, and enjoys excellent convergence rate of kernel smoother as well as coverage probability quickly approaching the nominal value. As an illustration, consider the motorcycle data, with Fig. 4 depicting the spline–kernel SCB of its variance function, at confidence levels 99.991 and 98.698 %, overlaid with a constant variance estimate which is either a consistent estimate $n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_{i,p}^2$ or the maximum of lower confidence line, the constant variance hypothesis is rejected in both scenarios, with p value = 0.00009 or 0.01302. While the proposed SCB is superior to the SCB of Song and Yang (2009), the spline–kernel estimator is computationally much faster than the kernel–kernel estimator of Fan and Yao (1998), due to using spline instead of kernel in step one, which cuts computing burden substantially, see Xue and Yang (2006) and Wang and Yang (2007) for speed comparison of spline and kernel smoothing. The new spline–kernel estimator is shown in Theorem 1 to be globally as efficient as the "infeasible kernel estimator" while the kernel–kernel estimator of Fan and Yao (1998) is as efficient as the "infeasible kernel estimator" only at a fixed point, see also Equation (3.2) of Hall and Carroll (1989) for pointwise oracle efficiency. Furthermore, oracle efficiency in Theorem 1 is of order smaller than $n^{-1/2}$, which had not existed in previous works.

The paper is organized as follows. Section 2 presents main theoretical results and Sect. 3 provides insights of proofs, Sect. 4 gives concrete steps to implement the SCB, while Sects. 5 and 6 report simulation results and analysis of the motorcycle data and Old Faithful geyser data. Section 7 concludes, and technical proofs are in the "Appendix".

2 Main result

Without loss of generality, we take $[a, b] = [0, 1]$. An asymptotic $100(1-\alpha)\,\%$ simultaneous confidence band (SCB) for the unknown variance function $\sigma^2(x)$ over a sequence of subintervals $[a_n, b_n] \subseteq [0, 1]$ where $a_n \to 0$, $b_n \to 1$ as $n \to \infty$, consists of an estimator $\hat{\sigma}^2(x)$ of $\sigma^2(x)$, lower and upper confidence limit $\hat{\sigma}^2(x) - l_{n,L}(x)$, $\hat{\sigma}^2(x) + l_{n,U}(x)$ at every $x \in [a_n, b_n]$ such that

$$\lim_{n\to\infty} P\left\{\sigma^2(x) \in \left[\hat{\sigma}^2(x) - l_{n,L}(x), \hat{\sigma}^2(x) + l_{n,U}(x)\right], \quad \forall x \in [a_n, b_n]\right\} = 1-\alpha.$$

Our goal is to construct error bound function $l_{n,L}(x)$, $l_{n,U}(x)$ based on data $\{(X_i, Y_i)\}_{i=1}^{n}$ drawn from model (1). We describe briefly below the ideas of oracally efficient estimation, which will be shown later to yield the SCB.

If the mean function $m(x)$ were known by "oracle ", one could compute the errors $\varepsilon_i = Y_i - m(X_i)$ and the squared errors $Z_i = \varepsilon_i^2$, $1 \le i \le n$, and then smooth the data $\{(X_i, Z_i)\}_{i=1}^n$, taking advantage of the fact that $\mathsf{E}(Z_i \mid X_i) \equiv \sigma^2(X_i)$. Specifically, denote by K a kernel function, $h = h_n$ a sequence of smoothing parameters called bandwidth, and $K_h(u) = K(u/h)/h$, an "infeasible kernel estimator" of the variance function is

$$\tilde{\sigma}_{\mathrm{K}}^2(x) = \frac{\sum_{i=1}^n K_h(X_i - x) Z_i}{\sum_{i=1}^n K_h(X_i - x)}. \tag{2}$$

To mimic this would-be kernel estimator $\tilde{\sigma}_{\mathrm{K}}^2(x)$ of $\sigma^2(x)$, a spline–kernel oracally efficient estimator $\hat{\sigma}_{\mathrm{SK}}^2(x)$ of $\sigma^2(x)$ is

$$\hat{\sigma}_{\mathrm{SK}}^2(x) = \frac{\sum_{i=1}^n K_h(X_i - x)\hat{Z}_i}{\sum_{i=1}^n K_h(X_i - x)}, \tag{3}$$

where $\hat{Z}_i = \hat{\varepsilon}_{i,p}^2$ are the square of residuals $\hat{\varepsilon}_{i,p}$ obtained from spline regression,

$$\hat{\varepsilon}_{i,p} = Y_i - \hat{m}_p(X_i), \quad 1 \le i \le n, \tag{4}$$

the spline estimator $\hat{m}_p(x)$ is defined as follows, for some positive integer p,

$$\hat{m}_p(x) = \underset{g \in G_N^{(p-2)}[0,1]}{\arg\min} \sum\nolimits_{i=1}^n \{Y_i - g(X_i)\}^2, \tag{5}$$

in which $G_N^{(p-2)}$ is the space of functions that are piecewise polynomials of degree $(p-1)$ on interval [0, 1], defined below.

The interval [0, 1] is divided into $(N+1)$ subintervals $J_j = [t_j, t_{j+1})$, $j = 0, \ldots, N-1$, $J_N = [t_N, 1]$ by a sequence of equally spaced points $\{t_j\}_{j=1}^N$, called interior knots, given as

$$t_0 = 0 < t_1 < \cdots < 1 = t_{N+1}, \quad t_j = jH, \quad j = 0, 1, \ldots, N+1,$$

in which $H = 1/(N+1)$ is the distance between neighboring knots. We denote by $G_N^{(p-2)} = G_N^{(p-2)}[0,1]$ the space of functions that are polynomials of degree $(p-1)$ on each J_j and have continuous $(p-2)$th derivative. In particular, $G_N^{(0)}$ denotes the space of functions that are linear on each J_j and continuous on [0, 1], with linear B-spline basis $\{b_{j,2}(x)\}_{j=-1}^N$ being

$$b_{j,2}(x) = K_0\left(\frac{x - t_{j+1}}{H}\right), \quad j = -1, 0, \ldots, N, \quad \text{for} \quad K_0(u) = (1 - |u|)_+ .$$

Alternatively, one can estimate $\sigma^2(x)$ by spline local linear estimator $\hat{\sigma}_{\mathrm{SLL}}^2(x)$ based on $\{X_i, \hat{Z}_i\}_{i=1}^n$, which mimics the would-be local linear estimator $\tilde{\sigma}_{\mathrm{LL}}^2(x)$ based on $\{X_i, Z_i\}_{i=1}^n$,

$$\left\{\hat{\sigma}^2_{\mathrm{SLL}}(x), \tilde{\sigma}^2_{\mathrm{LL}}(x)\right\} = (1,0)\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\left(\hat{\mathbf{Z}}, \mathbf{Z}\right),$$

in which the oracle and pseudo-response vectors are

$$\mathbf{Z} = (Z_1, \ldots, Z_n), \quad \hat{\mathbf{Z}} = \left(\hat{Z}_1, \ldots, \hat{Z}_n\right)$$

with the same weight and design matrices

$$\mathbf{W} = \mathrm{diag}\left\{K_h(X_i - x)\right\}_{i=1}^n, \quad \mathbf{X}^T = \begin{pmatrix} 1 & ,\ldots, & 1 \\ X_1 - x & ,\ldots, & X_n - x \end{pmatrix}.$$

The idea of synthesizing spline and kernel smoothing in one estimator appeared first in Wang and Yang (2007), Wang and Yang (2009) for additive model and later extended to generalized additive model in Liu et al. (2013).

To formulate the necessary technical assumptions, for sequences of real numbers c_n and d_n, one writes $c_n \ll d_n$ to mean $c_n/d_n \to 0$, as $n \to \infty$.

(A1) The function $m(\cdot) \in C^p[0,1]$, $p > 1$.
(A2) The joint distribution of (X, ε) is bivariate continuous with $\mathsf{E}(\varepsilon \mid X) = 0$, $\mathsf{E}(\varepsilon^2 \mid X) = \sigma^2(X)$, and for some $\eta > 1/2$,$\sup_{x\in[0,1]} \mathsf{E}(|\varepsilon|^{4+2\eta} \mid X = x) = M_\eta < +\infty$.
(A3) The density function $f(x) \in C[0,1]$, the variance function $\sigma^2(x) \in C^2[0,1]$, and $0 < c_f \le f(x) \le C_f < +\infty, 0 < c_\sigma \le \sigma(x) \le C_\sigma < +\infty$ for $x \in [0,1]$.
(A4) The kernel function $K \in C^{(1)}(\mathbb{R})$ is a symmetric probability density function supported on $[-1,1]$.
(A5) The bandwidth h satisfies $n^{2\alpha-1}(\log n)^4 \ll h \ll n^{-1/5}(\log n)^{-1/5}$, for some α such that $\alpha < 2/5$, $\alpha(2+\eta) > 1$, $\alpha(1+\eta) > 2/5$.
(A6) The number of interior knots $N = N_n$ satisfies

$$\max\left\{\left(\frac{n}{h^2}\right)^{1/4p}, \left(\frac{\log n}{h}\right)^{1/2(p-1)}\right\} \ll N \ll \min\left\{n^{1/2}h, \left(\frac{nh}{\log n}\right)^{1/3}, \left(\frac{n}{h}\right)^{1/5}\right\}.$$

Assumptions (A1)–(A3) are adapted from Song and Yang (2009), Assumption (A4) is standard for kernel regression, and Assumptions (A5) and (A6) are general conditions on the choice of number of knots N and bandwidth h to ensure oracle efficient and the extreme distribution result in (6) below. In particular, one may take the mean squared error optimal order $N \sim n^{1/(2p+1)}$ and an undersmoothing $h = n^{-1/5}(\log n)^{-1/5-\delta}$ for any $\delta > 0$, which satisfy all the requirements in Assumptions (A5) and (A6). As an example, data-driven implementation of N and h is given in Sect. 4, aided by explicit formulae (13) for BIC and (15) for rule-of-thumb bandwidth.

It follows from Assumption (A2) that the conditional variance of $Z = \varepsilon^2$ is $v_Z^2(x) \equiv \operatorname{var}(Z \mid X = x) \equiv \mu_4(x) - \sigma^4(x)$ in which $\mu_4(x) \equiv \mathsf{E}(\varepsilon^4 \mid X = x)$. In addition

$$\sup_{x\in[0,1]} \mathsf{E}\left(\left|Z - \sigma^2(X)\right|^{2+\eta} \mid X = x\right) \le \sup_{x\in[0,1]} \mathsf{E}\left(Z^{2+\eta} + \sigma(X)^{4+2\eta} \mid X = x\right) < M_\eta + C_\sigma^{4+2\eta} < +\infty.$$

Consequently, under Assumptions (A2)–(A5), by applying classic SCB theory to the unobservable sample $\{(X_i, Z_i)\}_{i=1}^n$, one has

$$P\left[a_h\left\{\sup_{x\in[h,1-h]}\left|\tilde{\sigma}_{\mathrm{K}}^2(x) - \sigma^2(x)\right|/V_n - b_h\right\} \le t\right] \to e^{-2e^{-t}}, \quad t \in \mathbb{R} \tag{6}$$

where

$$a_h = \sqrt{-2\log h}, \quad b_h = a_h + a_h^{-1}\left\{\log(\sqrt{C(K)}/2\pi)\right\}, \quad C(K) = \sqrt{C_{K'}/C_K},$$
$$C_K = \int K^2(u)\,du, \quad C_{K'} = \int K'^2(u)\,du, \quad V_n = v_Z(x)\{f(x)nh\}^{-1/2}C_K^{1/2}. \tag{7}$$

From (6) one obtains an asymptotic $100(1-\alpha)\,\%$ oracle SCB for $\sigma^2(x)$ over $[h, 1-h]$,

$$\tilde{\sigma}_{\mathrm{K}}^2(x) \pm V_n\left(2\log h^{-1}\right)^{1/2} Q_n(\alpha), \tag{8}$$

where

$$Q_n(\alpha) = 1 + \frac{\log\{C(K)/2\pi\} - \log\left\{-\frac{1}{2}\log(1-\alpha)\right\}}{2\log h^{-1}}. \tag{9}$$

In stating our main theoretical results in the next two Theorems, and throughout this paper, we denote by $\|\cdot\|_\infty$, the supremum norm of a function r on $[0, 1]$, i.e., $\|r\|_\infty = \sup_{x\in[0,1]}|r(x)|$.

Theorem 1 *Under Assumptions (A1)–(A6), as $n \to \infty$, the estimator $\hat{\sigma}_{SK}^2(x)$ is asymptotically as efficient as the "infeasible estimator", $\tilde{\sigma}_K^2(x)$ i.e.,*

$$\left\|\hat{\sigma}_{SK}^2 - \tilde{\sigma}_K^2\right\|_\infty = o_p\left(n^{-1/2}\right).$$

As commented in the introduction, the oracle efficiency stated in Theorem 1 is of the unprecedented small order $o_p\left(n^{-1/2}\right)$, and the next result follows immediately.

Theorem 2 *Under Assumptions (A1)–(A6), an asymptotic* $100\,(1-\alpha)\,\%$ *oracally efficient SCB for* $\sigma^2\,(x)$ *over* $[h, 1-h]$ *is*

$$\hat{\sigma}^2_{SK}\,(x) \pm V_n \left(2\log h^{-1}\right)^{1/2} Q_n\,(\alpha) \tag{10}$$

with V_n *and* $Q_n\,(\alpha)$ *given in* (7) *and* (9) *respectively. In other words,*

$$\lim_{n\to\infty} P\left\{\sigma^2\,(x) \in \hat{\sigma}^2_{SK}\,(x) \pm V_n \left(2\log h^{-1}\right)^{1/2} Q_n\,(\alpha)\,, \forall x \in [h, 1-h]\right\} = 1-\alpha.$$

The proofs of Theorems 1 and 2 depend on Propositions 1, 2 and 3 given in Sect. 3. The proofs of these Propositions are based on Lemmas 1, 4 and 2. All of them are provided in the "Appendix". Both Theorems 1 and 2 remain true with spline–kernel estimator $\hat{\sigma}^2_{\mathrm{SK}}\,(x)$ replaced by spline-local linear estimator $\hat{\sigma}^2_{\mathrm{SLL}}\,(x)$, but detailed proofs for local linear estimator are omitted as in Wang and Yang (2007, 2009).

3 Error decomposition

To break the estimation error $\hat{\sigma}^2_{\mathrm{SK}}\,(x) - \tilde{\sigma}^2_{\mathrm{K}}\,(x)$ into simpler parts, we begin by discussing the spline space $G_N^{(p-2)}$ and the representation of the spline estimator of $\hat{m}_p(x)$ in Eq. (5).

Denote by $\|\phi\|_2$ the theoretical L^2 norm of a function ϕ on $[0, 1]$, i.e., $\|\phi\|_2^2 = \mathsf{E}\{\phi^2(X)\} = \int_0^1 \phi^2\,(x)\, f\,(x)\,dx$, the empirical L^2 norm as $\|\phi\|^2_{2,n} = n^{-1}\sum_{i=1}^n \phi^2\,(X_i)$, and then define the rescaled B-spline basis $\{B_{j,p}(x)\}_{j=1-p}^N$ for $G_N^{(p-2)}$, each with theoretical norm equal to 1

$$B_{j,p}\,(x) \equiv b_{j,p}\,(x) \left\|b_{j,p}\,(x)\right\|_2^{-1}, \quad 1-p \le j \le N.$$

The estimator $\hat{m}_p\,(x)$ in Eq. (5) can then be expressed as

$$\hat{m}_p\,(x) = \operatorname*{Proj}_{G_N^{(p-2)}} \mathbf{Y} = \sum_{j=1-p}^{N} \hat{\lambda}_{j,p} B_{j,p}\,(x)\,,$$

where the vector $\{\hat{\lambda}_{1-p,p}, \ldots, \hat{\lambda}_{N,p}\}^T$ solves the following least-squares problem

$$\left\{\hat{\lambda}_{1-p,p}, \ldots, \hat{\lambda}_{N,p}\right\}^T = \operatorname*{arg\,min}_{R^{N+p}} \sum_{i=1}^{n} \left\{Y_i - \sum_{j=1-p}^{N} \hat{\lambda}_{j,p} B_{j,p}\,(X_i)\right\}^2. \tag{11}$$

We write $\mathbf{Y}$ as the sum of a signal vector $\mathbf{m}$ and a noise vector $\mathbf{E}$,

$$\mathbf{Y} = \mathbf{m} + \mathbf{E}, \quad \mathbf{m} = \{m\,(X_1)\,, \ldots, m\,(X_n)\}^T\,, \quad \mathbf{E} = \{\varepsilon_1, \ldots, \varepsilon_n\}^T\,.$$

Projecting this relationship into the space $G_n^{(p-2)}$, one obtains

$$\hat{\mathbf{m}}_p = \left\{\hat{m}_p\left(X_1\right), \ldots, \hat{m}_p\left(X_n\right)\right\}^T = \underset{G_n^{(p-2)}}{\text{Proj}}\, \mathbf{Y} = \underset{G_n^{(p-2)}}{\text{Proj}}\, \mathbf{m} + \underset{G_n^{(p-2)}}{\text{Proj}}\, \mathbf{E}.$$

Correspondingly, in the space $G_N^{(p-2)}$, one has $\hat{m}_p\left(x\right) = \tilde{m}_p\left(x\right) + \tilde{\varepsilon}_p\left(x\right)$, where

$$\tilde{m}_p\left(x\right) = \sum_{J=1-p}^{N} \tilde{\lambda}_{J,p} B_{J,p}\left(x\right), \quad \tilde{\varepsilon}_p\left(x\right) = \sum_{J=1-p}^{N} \tilde{a}_{J,p} B_{J,p}\left(x\right), \tag{12}$$

with the vectors $\left\{\tilde{\lambda}_{1-p,p}, \ldots, \tilde{\lambda}_{N,p}\right\}^T$ and $\left\{\tilde{a}_{1-p,p}, \ldots, \tilde{a}_{N,p}\right\}^T$ being solutions to (11) with Y_i replaced by $m(X_i)$ and ε_i respectively.

Regarding variance estimator in (2) and (3)

$$\hat{\sigma}^2_{\text{SK}}\left(x\right) - \tilde{\sigma}^2_{\text{K}}\left(x\right) = \frac{\sum_{i=1}^{n} K_h\left(X_i - x\right)\left(\text{I}_{i,p} + \text{II}_{i,p} + \text{III}_{i,p}\right)}{\sum_{i=1}^{n} K_h\left(X_i - x\right)}$$
$$= \hat{f}^{-1}\left(x\right)\left\{\text{I} + \text{II} + \text{III}\right\},$$

in which $\hat{f}\left(x\right) = n^{-1}\sum_{i=1}^{n} K_h\left(X_i - x\right)$,

$$\text{I} = \text{I}\left(x\right) = n^{-1}\sum_{i=1}^{n} K_h\left(X_i - x\right)\text{I}_{i,p},$$
$$\text{II} = \text{II}\left(x\right) = n^{-1}\sum_{i=1}^{n} K_h\left(X_i - x\right)\text{II}_{i,p},$$
$$\text{III} = \text{III}\left(x\right) = n^{-1}\sum_{i=1}^{n} K_h\left(X_i - x\right)\text{III}_{i,p},$$
$$\text{I}_{i,p} = \left\{m\left(X_i\right) - \tilde{m}_p\left(X_i\right)\right\}^2 + \tilde{\varepsilon}_p^2\left(X_i\right) + 2\left\{\tilde{m}_p\left(X_i\right) - m\left(X_i\right)\right\}\tilde{\varepsilon}_p\left(X_i\right),$$
$$\text{II}_{i,p} = -2\varepsilon_i\tilde{\varepsilon}_p\left(X_i\right), \text{III}_{i,p} = \left\{m\left(X_i\right) - \tilde{m}_p\left(X_i\right)\right\}\varepsilon_i.$$

By Assumption (A3), $\hat{f}\left(x\right) = f\left(x\right) + u_p\left(1\right) \geq c_f + u_p\left(1\right)$, hence Theorem 1 follows from the next three Propositions on I, II, III.

Proposition 1 *Under Assumptions (A1)–(A6), as $n \to \infty$,*

$$\|\text{I}\|_\infty = \mathcal{O}_p\left\{h^{-1}\left(H^{2p} + (nH)^{-1}\right)\right\} = o_p\left(n^{-1/2}\right).$$

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

$$\|\text{II}\|_\infty = \mathcal{O}_p\left(n^{-1}h^{-1/2}H^{-3/2}\log^{1/2} n + n^{-1}h^{1/2}H^{-5/2}\right) = o_p\left(n^{-1/2}\right).$$

Proposition 3 *Under Assumptions (A1)–(A6), as $n \to \infty$,*

$$\|\mathrm{III}\|_\infty = \mathcal{O}_p\left\{n^{-1/2}h^{-1/2}H^{p-1}\log^{1/2}n + n^{-1/2}h^{1/2}H^{p-2}\right\} = o_p\left(n^{-1/2}\right).$$

4 Implementation

We describe in this section one concrete procedure that implements the oracally efficient SCB in Theorem 2, and is used throughout Sects. 5 and 6 for both simulated and real data examples. Given any sample $\{(X_i, Y_i)\}_{i=1}^n$ from model (1), let $a = \min(X_1, \ldots, X_n), b = \max(X_1, \ldots, X_n)$ and transform the data range from $[a, b]$ into $[0, 1]$ by the linear transformation $x \to (x - a)/(b - a)$. If this linear operation fails to make design variable X conform to Assumption (A3), one applies the quantile transformation $x \to F_n = n^{-1}\sum_{i=1}^n \mathrm{I}(X_i \le x)$.

To select the number of interior knots N, let $\hat{N}^{\mathrm{opt}}$ be the minimizer of BIC defined below, over integers from $[0.5N_r, \min(5N_r, Tb)]$, with $N_r = n^{-1/(2p+1)}$ and $Tb = n/4 - 1$, which ensures that $\hat{N}^{\mathrm{opt}}$ is order of $n^{-1/(2p+1)}$ and the number of parameters in the least-squares estimation is less than $n/4$. The chosen $\hat{N}^{\mathrm{opt}}$ obviously satisfies Assumption (A6), but other choices of N remain open possibility. For any candidate integer $N \in [0.5N_r, \min(5N_r, Tb)]$, denote the predictor for the i-th response Y_i by $\hat{Y}_i = \hat{m}_p(X_i)$, and let $q_n = (1 + N_n)$ be the number of parameters in (11), the BIC value corresponding to N is,

$$\mathrm{BIC} = \log(\mathrm{MSE}) + q_n \log(n)/n,\ \mathrm{MSE} = n^{-1}\sum_{i=1}^n \left\{Y_i - \hat{Y}_i\right\}^2. \tag{13}$$

Algebra shows that the least-squares problem in Eq. (11) can be also solved via the truncated power basis $\left\{1, x, \ldots, x^{p-1}, \left(x - t_j\right)_+^{p-1}, j = 1, 2, \ldots N\right\}$, see de Boor (2001), which is regularly used in implementation. In other words,

$$\hat{m}_p(x) = \sum_{k=0}^{p-1}\hat{r}_k x^k + \sum_{j=1}^N \hat{r}_{j,p}\left(x - t_j\right)_+^k, \tag{14}$$

where the coefficients $\left(\hat{r}_0, \ldots, \hat{r}_{p-1}, \hat{r}_{1,p}, \ldots, \hat{r}_{N,p}\right)^T$ are solutions to the least squares problem

$$\left(\hat{r}_0, \ldots, \hat{r}_{N,p}\right)^T = \underset{(r_0,\ldots,r_{N,p})\in\mathbb{R}^{N+P}}{\operatorname{argmin}} \sum_{i=1}^n \left\{Y_i - \sum_{k=0}^{p-1} r_k X_i^k - \sum_{j=1}^N r_{j,p}\left(X_i - t_j\right)_+^k\right\}^2.$$

To choose an appropriate bandwidth $h = h_n$ for computing $\hat{\sigma}_{\mathrm{SK}}^2(x)$, one adopts the following rule-of-thumb (ROT) bandwidth of Fan and Gijbels (1996), Equation (4.3):

$$h_{\mathrm{rot}} = \left\{ \frac{35 \sum_{i=1}^{n} \left(\hat{Z}_i - \sum_{k=0}^{4} \widehat{a}_k X_i^k \right)^2}{n \sum_{i=1}^{n} \left(2\widehat{a}_2 + 6\widehat{a}_3 X_i + 12\widehat{a}_4 X_i^2 \right)^2} \right\}^{1/5} \tag{15}$$

in which $(\widehat{a}_k)_{k=0}^4 = \operatorname{argmin}_{(a_k)_{k=0}^4 \in \mathbb{R}^5} \sum_{i=1}^n \left(\hat{Z}_i - \sum_{k=0}^4 a_k X_i^k \right)^2$. One then sets $h = h_n = h_{\mathrm{rot}} (\log n)^{-1/2} \sim n^{-1/5} (\log n)^{-1/2}$, which clearly satisfies Assumption (A5), especially the undersmoothing condition $h \ll n^{-1/5} (\log n)^{-1/5}$.

For constructing the SCB, the unknown functions $v_Z^2(x)$ and $f(x)$ are evaluated and then plugged in, the same approach taken in Hall and Titterington (1988), Härdle (1989), Xia (1998), Wang and Yang (2009), Song and Yang (2009). Let $\tilde{K}(u) = 15 \left(1 - u^2\right)^2 \mathrm{I}\{|u| \leq 1\}/16$ be the quadratic kernel and s_n be the sample standard deviation of $\{X_i\}_{i=1}^n$ and

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} h_{\mathrm{rot},f}^{-1} \tilde{K}\left(\frac{X_i - x}{h_{\mathrm{rot},f}} \right), \quad h_{\mathrm{rot},f} = (4\pi)^{1/10} \left(\frac{140}{3} \right) n^{-1/5} s_n, \tag{16}$$

where $h_{\mathrm{rot},f}$ is the rule-of-thumb bandwidth in Silverman (1986). Define $\nabla^T = \{\nabla_i, 1 \leq i \leq n\}$, $\nabla_i = \{\hat{Z}_i - \hat{\sigma}_{\mathrm{SK}}^2(X_i)\}^2$, and

$$\mathbf{X} = \mathbf{X}(x) = \begin{pmatrix} 1 & ,\ldots, & 1 \\ X_1 - x & ,\ldots, & X_n - x \end{pmatrix}^T, \quad \mathbf{W} = \mathbf{W}(x) = \operatorname{diag}\left\{ \tilde{K}\left(\frac{X_i - x}{h_{\mathrm{rot},\sigma}} \right) \right\}_{i=1}^n$$

where $h_{\mathrm{rot},\sigma}$ is the ROT bandwidth of Fan and Gijbels (1996) Equation (4.3), as h_{rot} in (15), but with the $\hat{Z}_i$'s replaced by ∇_i's, and define the following estimator of $v_Z^2(x)$

$$\hat{v}_Z^2(x) = (1, 0) \left(\mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W} \nabla. \tag{17}$$

The following results follow from Bickel and Rosenblatt (1973) and Fan and Gijbels (1996)

$$\sup_{x \in [0,1]} \left| \hat{v}_Z^2(x) - v_Z^2(x) \right| + \sup_{x \in [0,1]} \left| \hat{f}(x) - f(x) \right| = o_p(1). \tag{18}$$

The function V_n is approximated by the following, with $\hat{f}(x)$ and $\hat{v}_Z^2(x)$ defined in Eqs. (16) and (17)

$$\hat{V}_n = \hat{v}_Z(x) \left\{ \hat{f}(x) nh \right\}^{-1/2} C_K^{1/2}.$$

Then Eq. (18) and Theorem 2 imply that as $n \to \infty$, the SCB below is asymptotically $100(1 - \alpha)\,\%$

$$\hat{\sigma}_{\mathrm{SK}}^2(x) \pm \hat{V}_n \left(2 \log h^{-1} \right)^{1/2} Q_n(\alpha). \tag{19}$$

The construction described above of SCB according to Theorem 2, is over an interior portion of the data range [0, 1], namely $[a_n, b_n] = [h_n, 1 - h_n] \subseteq (0, 1)$, as seen in the SCB plots of Figs. 3, 2, 4 and 5. It should be emphasized, however, that the interval sequence $[h_n, 1 - h_n]$ covers the entire interior (0, 1) as sample size $n \to \infty$ and $h_n \to 0$, which reflects, for instance, the widening range in Fig. 3 of SCB in (c) and (d) over (a) and (b).

Although any spline order $p > 1$ can be employed, we have used only linear splines (with $p = 2$) for simplicity. It is well-known that the choice of kernel function is of less importance, according to Assumptions (A4) and (A5), the kernel function K is chosen to be the quadratic kernel. Simulation comparison will be made in Sect. 5 of the above oracally efficient SCB with the infeasible SCB, which is computed from (8) with $v_Z^2(x)$ and $f(x)$ replaced by $\tilde{v}_Z^2(x)$ and $\hat{f}(x)$ in (16), respectively, where $\tilde{v}_Z^2(x)$ is the right side of (17) with ∇ substituted by $\tilde{\nabla}$, where $\tilde{\nabla}^T = \{\tilde{\nabla}_i, 1 \leq i \leq n\}$, $\tilde{\nabla}_i = \{Z_i - \tilde{\sigma}_{\mathrm{K}}^2(X_i)\}^2$.

5 Simulation

In this section, simulation results are presented to illustrate the finite-sample behavior of the oracally efficient SCB, on data sets generated from model (1), with $X \sim U[-1/2, 1/2]$, and

$$m(x) = \sin(2\pi x), \quad \sigma(x) = 1/2 - cx^2, \quad \varepsilon \mid x \sim N\left\{0, \sigma^2(x)\right\}. \tag{20}$$

We choose $c = 1, c = 0.5$, which have included variance functions $\sigma^2(x)$ that are strongly heteroscedastic ($c = 1$) and nearly homoscedastic ($c = 0.5$), while sample sizes are taken to be $n = 100, 200, 500$ and the confidence levels are $1 - \alpha = 0.99, 0.95$. Table 1 contains the coverage frequency of the true curve $\sigma^2(x)$ at all data

Table 1 Coverage frequency of the oracally efficient SCB in Theorem (2) and the infeasible SCB in (8) from 500 replications

c	n	$1 - \alpha$	Oracally efficient SCB	Infeasible SCB
1.0	100	0.950	0.834	0.852
		0.990	0.932	0.952
	200	0.950	0.912	0.936
		0.990	0.980	0.988
	500	0.950	0.966	0.962
		0.990	0.994	0.994
0.5	100	0.950	0.836	0.870
		0.990	0.934	0.956
	200	0.950	0.922	0.944
		0.990	0.986	0.992
	500	0.950	0.954	0.960
		0.990	0.994	0.996

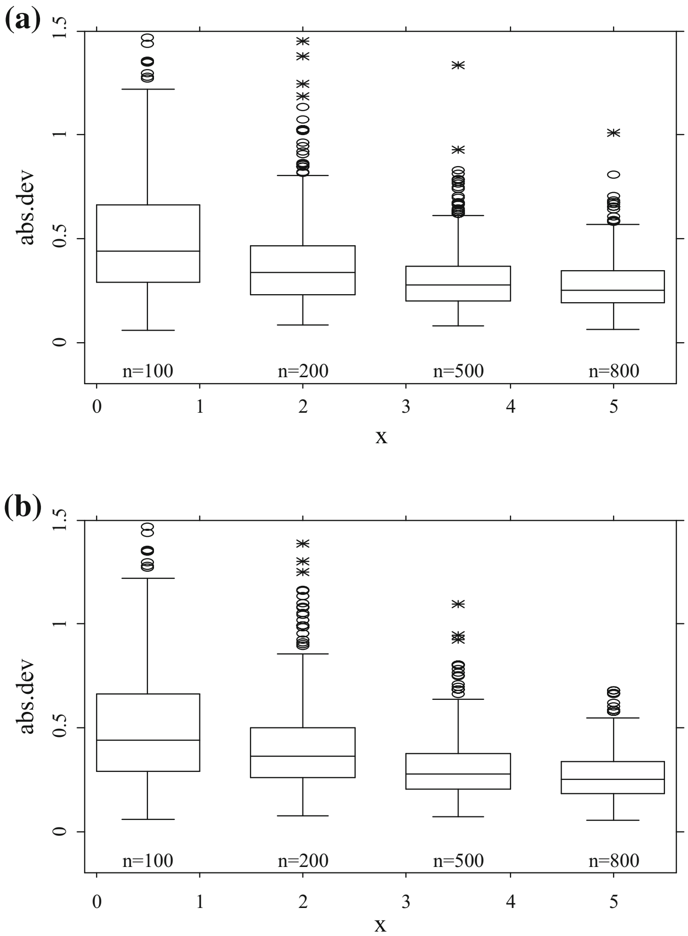

Fig. 1 Boxplots of Δ_n with **a** $c = 1$; **b** $c = 0.5$

points $\{X_i\}_{i=1}^{n}$ by the oracally efficient SCB whose construction details are in Sect. 4 over 500 replications of sample size n. Coverage frequency over the same data sets of the infeasible SCB in (8) is also listed in the table. In all cases, the coverage improves with increasing sample size, which confirms to Theorem 2, and the two SCBs are quite close to each other in terms of coverage frequency, showing positive confirmation of Theorem 1. For both cases $c = 1$ and $c = 0.5$, the oracally efficient SCB has coverage frequency approaching the nominal level for sample size as low as $n = 200$.

Figure 1 depicts the boxplots over 500 replications of $\Delta_n = \sqrt{n}\max\left|\tilde{\sigma}^2_{\mathrm{K}}\left(x_j\right) - \hat{\sigma}^2_{\mathrm{SK}}\left(x_j\right)\right|$, where $\{x_j, j = 1, 2, \ldots n_{\mathrm{grid}}\}$ points on $[-0.5 + h, 0.5 - h]$ with $n_{\mathrm{grid}} = 401$, h being the chosen bandwidth of estimator (3). it can be seen that the boxplot of Δ_n becomes narrower as n increases, implying that difference between the spline–kernel variance estimator and the infeasible estimator with known mean function is asymptotically of smaller order than $n^{-1/2}$, which confirms Theorem 1. For visual impression of the SCB, Figs. 3 and 2 are created based on sample sizes $n = 100$, 500,

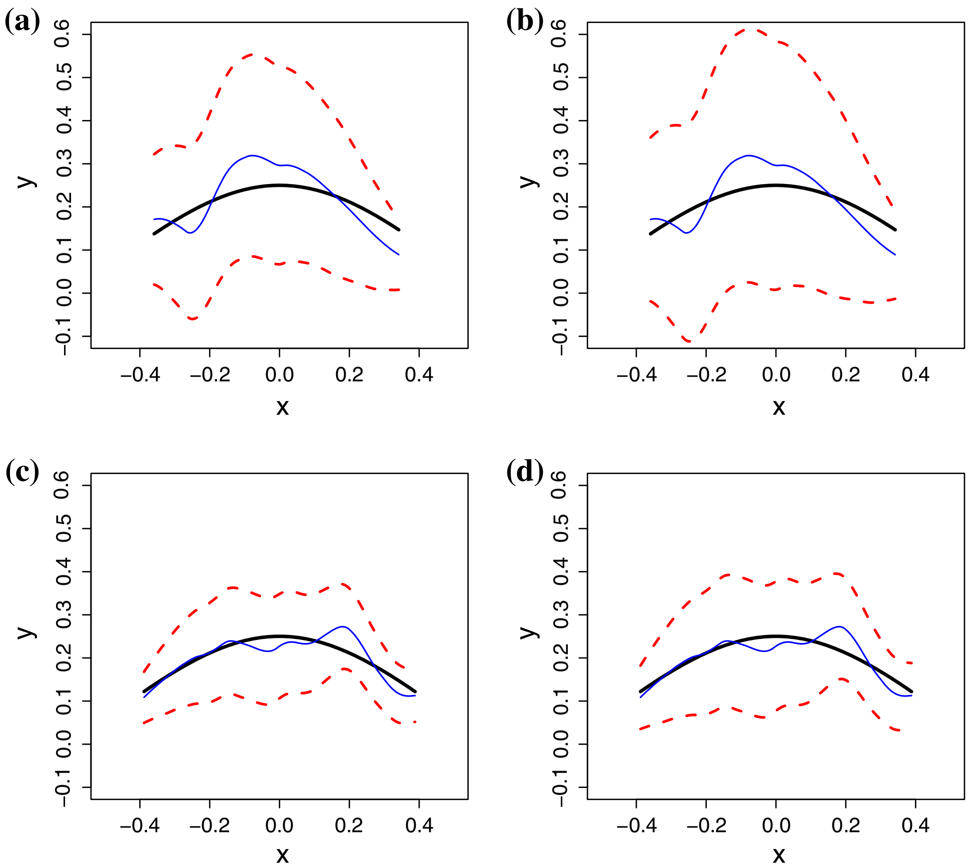

Fig. 2 Plots of SCB for variance function (*dashed*) which is computed according to (19), the estimator $\hat{\sigma}^2_{\text{SK}}(x)$ (*solid*), the true function $\sigma^2(x)$ with $c = 0.5$ (*thick*). **a** $n = 100$, 95 % SCB; **b** $n = 100$, 99 % SCB; **c** $n = 500$, 95 % SCB; **d** $n = 500$, 99 % SCB

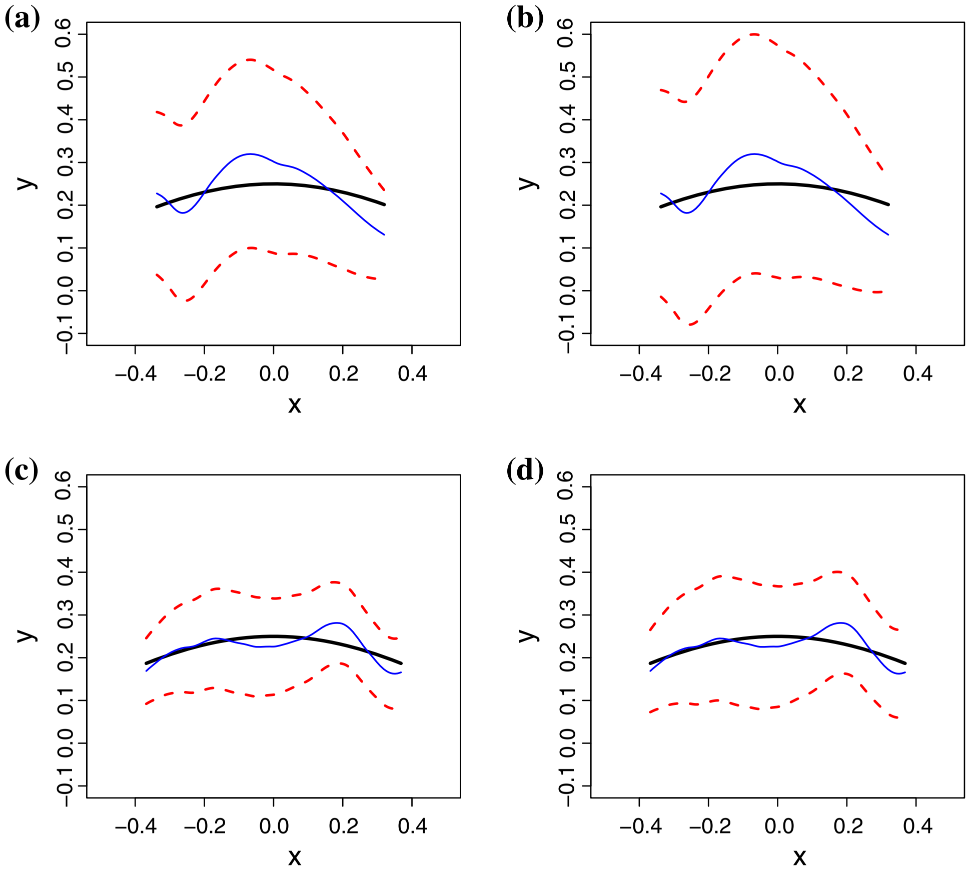

Fig. 3 Plots of SCB for variance function (*dashed*) which is computed according to (19), the estimator $\hat{\sigma}^2_{\mathrm{SK}}(x)$ (*solid*), the true function $\sigma^2(x)$ with $c = 1$ (*thick*). **a** $n = 100$, 95% SCB; **b** $n = 100$, 99% SCB; **c** $n = 500$, 95 % SCB; **d** $n = 500$, 99 % SCB

and $c = 1, 0.5$, respectively, each with symbols: center thick line (true curve), center solid line (the estimated curve), upper- and lower-dashed line (SCB). In all figures, the SCB becomes narrower and fit better for $n = 500$ than for $n = 100$.

6 Empirical examples

In this section, we test the null hypothesis of homoscedasticity $H_0 : \sigma^2(x) = \sigma_0^2 > 0$ for two well-known data sets. The first is the motorcycle data with $n = 133$ observations, with X = time (in milliseconds) after a simulated impact on motorcycles, Y = the head acceleration of a PTMO (post mortem human test object). The data can be called in R by the command "data(motorcycledata)", see http://www.inside-r.org/node/52453. In Fig. 4, the center thick lines are the spline–kernel estimator $\hat{\sigma}^2_{\mathrm{SK}}(x)$ for $\sigma^2(x)$, the upper/lower solid lines represent the SCB for the variance function. Since the $100(1 - 0.00009)\%$ SCB in (a) does not contain the consistent estimate of σ_0^2 under the null hypothesis, which equals $n^{-1} \sum_{i=1}^{n} \hat{\varepsilon}^2_{i,p}$, one rejects the null hypothesis of homoscedasticity with p value < 0.00009.

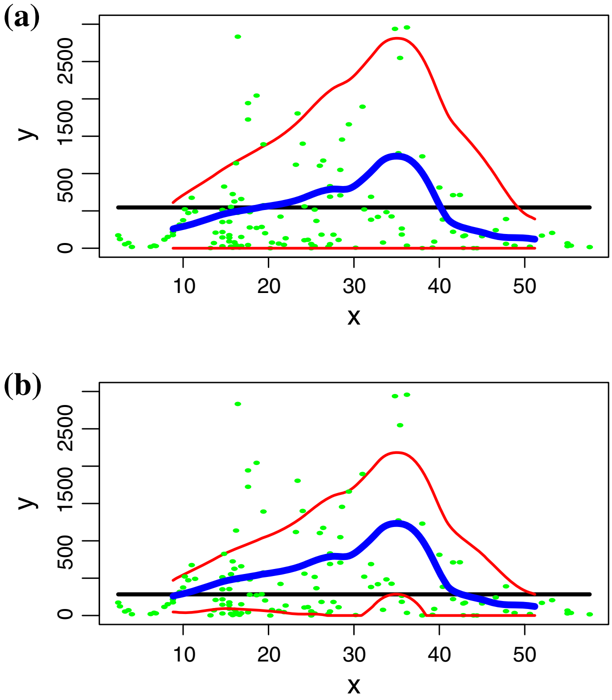

Fig. 4 For the motorcycle data, plots of SCB (*solid*) computed according to the (19), the spline–kernel estimator $\hat{\sigma}^2_{\mathrm{SK}}(x)$ (*thick*), the scatterplot of $\hat{Z}_i = \hat{\varepsilon}^2_{i,p}$ **a** 99.991% SCB, a constant variance fit which equals $n^{-1} \sum_{i=1}^{n} \hat{\varepsilon}^2_{i,p}$, $\alpha = 0.00009$; **b** 98.698 % SCB, a constant variance fit which equals the maximum of upper SCB, $\alpha = 0.01302$

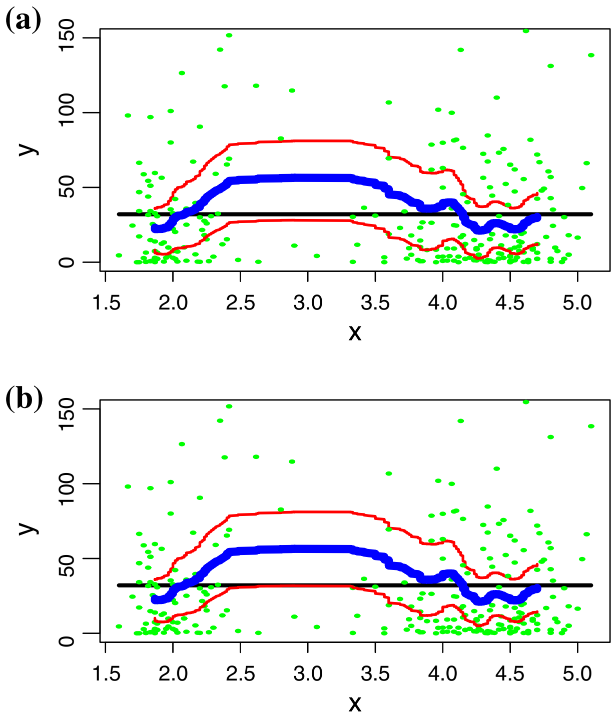

Fig. 5 For the Old Faithful geyser data, plots of SCB (*solid*) computed according to the (19), the spline–kernel estimator $\hat{\sigma}^2_{\mathrm{SK}}(x)$ (*thick*), a constant variance fit which equals $n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}^2_{i,p}$, the scatterplot of $\hat{Z}_i = \hat{\varepsilon}^2_{i,p}$ **a** 95 % SCB, $\alpha = 0.05$, **b** 88 % SCB, $\alpha = 0.12$

Song and Yang (2009) had obtained a p value of 0.008 with spline SCB, as minimum of the upper confidence line equals the maximum of the lower confidence line for the spline SCB of confidence level 99.2 % $=$ 1 $-$ 0.008. The 99.2 % spline SCB therefore contains completely a horizontal line, even though its height is not equal to $n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}^2_{i,p}$. For comparison, we have computed the confidence level at which the upper and lower lines of the spline–kernel SCB coincide, which turns out to be 98.698 %, thus one rejects the null hypothesis of homoscedasticity with p value $\leq$ 0.01302. Figure 4b depicts the 98.698 % spline–kernel SCB and the horizontal line that completely fits inside the SCB. We have also constructed ad hoc local linear SCBs by substituting $\hat{\sigma}^2_{\mathrm{SK}}(x)$ in (19) with the two-step local linear estimator of $\sigma^2(x)$ in Fan and Yao (1998), and with minimum of upper line and maximum of lower line equal, the confidence level is 99.999865 %, thus the p value is 0.00000135 for rejecting the null hypothesis of homoscedasticity. To sum up, for the motorcycle data, homoscedasticity is rejected by all four approaches, with p values ranging from 0.00000135 to 0.01302.

The second data set is the Old Faithful geyser data, which can be downloaded from http://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat. Geysers are a special kind of hot springs that erupt a mixture of hot water, steam and other gases, and by studying geysers scientists obtain useful information about the structure and the dynamics of earth's crust. The data consists of $n = 272$ observations for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA: $X =$ eruption time in mins, $Y =$ the waiting time to next eruption. Figure 5 shows that for the geyser data one can not reject the null hypothesis of homoscedasticity with a p value 0.12.

7 Conclusions

A spline–kernel estimator is proposed for the conditional variance function in nonparametric regression model, which is shown to be oracally efficient, that is, it uniformly approximates an infeasible kernel variance estimator at the rate of $o_p\left(n^{-1/2}\right)$. A powerful technical Lemma 4 is used in the proofs of Propositions 2 and 3, both indispensable in establishing oracle efficiency. A data-driven procedure implements the kernel SCB centered around the oracally efficient two-step estimator, with limiting coverage probability equal to that of the infeasible kernel SCB. As illustrated by both the motorcycle and the Old Faithful geyser data, the theoretically justified kernel SCB is also a useful tool for testing hypotheses on the conditional variance function, and is expected to find wide applications in many scientific disciplines.

Acknowledgments This work has been supported by NSF award DMS 1007594, Jiangsu Specially-Appointed Professor Program SR10700111, Jiangsu Key-Discipline Program ZY107992, National Natural Science Foundation of China award 11371272, and Research Fund for the Doctoral Program of Higher Education of China award 20133201110002. The authors thank the Editor and two Reviewers for helpful comments.

Appendix A

Throughout this Appendix, we denote by $\|\xi\|$ the Euclidean norm and $|\xi|$ means the largest absolute value of the elements of any vector ξ. We use c, C to denote any positive constants in the generic sense. We denote for any given constant $C > 0$, a class of Lipschitz continuous functions by $\text{Lip}\left([0,1], C\right) = \left\{\varphi \left| \left|\varphi(x) - \varphi\left(x'\right)\right| \le C\left|x - x'\right|, \forall x, x' \in [0,1]\right.\right\}$.

A.1 Preliminaries

The Lemmas of this Subsection are needed for the proof of Propositions 1, 2 and 3. These Propositions clearly establish Theorems 1 and 2.

Lemma 1 *Under Assumptions (A1)–(A5), there exists a constant* $C_p > 0$, $p > 1$, *such that for any* $m \in C^p[0,1]$ *there is a spline function* $g_p \in G_N^{(p-2)}$ *satisfying* $\left\|m - g_p\right\|_\infty \le CH^p$ *and* $m - g_p \in Lip\left([0,1], CH^{p-1}\right)$. *The function* $\tilde{m}_p(x)$ *given in Equation (12)*

$$\left\|\tilde{m}_p(x) - m(x)\right\|_\infty \le C_p \inf_{g \in G_N^{(p-2)}} \|g - m\|_\infty = \mathcal{O}_p\left(H^p\right).$$

Moreover, for the function $\tilde{\varepsilon}_p(x)$ given in Equation (12)

$$\left\|\tilde{\varepsilon}_p(x)\right\|_{2,n} = \mathcal{O}_p\left(n^{-1/2}N^{1/2}\right), \left\|\tilde{\varepsilon}_p(x)\right\|_{\infty} = \mathcal{O}_p\left(n^{-1/2}N^{1/2}(\log n)^{1/2}\right).$$

See Lemma A.1 of Song and Yang (2009), and also Wang and Yang (2009) for detailed proof.

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

$$\sup_{x\in[0,1]}\left\{\left|\left\{B_{j,p}(x)\right\}_{j=1-p}^{N}\right| + \left\|\left\{B_{j,p}(x)\right\}_{j=1-p}^{N}\right\|\right\} = \mathcal{O}\left(H^{-1/2}\right) \tag{21}$$

See Lemma A.4 of Song and Yang (2009) for detailed proof.

The strong approximation result of Tusnády (1977) is also needed.

Lemma 3 *Let $U_1, \ldots, U_n$ be i.i.d. r.v.'s on the 2 -dimensional unit square with $P(U_i < \mathbf{t}) = \lambda(\mathbf{t}), \mathbf{0} \le \mathbf{t} \le \mathbf{1}$, where $\mathbf{t} = (t_{1,}t_2)$ and $\mathbf{1} = (1, 1)$ are 2-dimensional vectors, $\lambda(\mathbf{t}) = t_1t_2$. The empirical distribution function $F_n^u(\mathbf{t})$ based on sample $(U_1, \ldots, U_n)$ is defined as $F_n^u(\mathbf{t}) = n^{-1}\sum_{i=1}^n \mathrm{I}_{\{U_i<\mathbf{t}\}}$ for $\mathbf{0} \le \mathbf{t} \le \mathbf{1}$. The 2-dimensional Brownian bridge $B(\mathbf{t})$ is defined by $B(\mathbf{t}) = W(\mathbf{t}) - \lambda(\mathbf{t})W(\mathbf{1})$ for $\mathbf{0} \le \mathbf{t} \le \mathbf{1}$, where $W(\mathbf{t})$ is a 2-dimensional Wiener process. Then there is a version $B_n(\mathbf{t})$ of $B(\mathbf{t})$ such that*

$$P\left[\sup_{\mathbf{0}\le\mathbf{t}\le\mathbf{1}}\left|n^{1/2}\left\{F_n^u(\mathbf{t}) - \lambda(\mathbf{t})\right\} - B_n(\mathbf{t})\right| > (C\log n + x)\frac{\log n}{n^{1/2}}\right] < Ke^{-\lambda x}$$

holds for all x, where C, K, λ are positive constants.

Denote the well-known Rosenblatt transformation for bivariate continuous (X, ε) as

$$\left(X', \varepsilon'\right) = M(X, \varepsilon) = \left\{F_X(x), F_{\varepsilon|X}(\varepsilon|x)\right\}, \tag{22}$$

so that $\left(X', \varepsilon'\right)$ has uniform distribution on $[0, 1]^2$, therefore

$$Z_n\left\{M^{-1}\left(x', \varepsilon'\right)\right\} = Z_n(x, \varepsilon) = \sqrt{n}\left\{F_n(x, \varepsilon) - F(x, \varepsilon)\right\}, \tag{23}$$

with $F_n(x, \varepsilon)$ denoting the empirical distribution of (X, ε). Lemma 3 implies that there exists a version B_n of 2-dimensional Brownian bridge such that

$$\sup_{x,\varepsilon}\left|Z_n(x, \varepsilon) - B_n\{M(x, \varepsilon)\}\right| = \mathcal{O}_{a.s.}\left(n^{-1/2}\log^2 n\right). \tag{24}$$

Lemma 4 *Under Assumptions (A2)-(A5), as $n \to \infty$, for any sequence of functions $r_n \in Lip([0, 1], l_n), l_n > 0$ with $\|r_n\|_\infty = \rho_n \ge 0$*

$$n^{-1}\sum_{i=1}^{n}K_h\left(X_i-x\right)r_n\left(X_i\right)\varepsilon_i=\mathcal{U}_p\left(n^{-1/2}h^{-1/2}\rho_n\log^{1/2}n+n^{-1/2}h^{1/2}l_n\right) \tag{25}$$

Proof Step 1. We first discretize the problem by letting $0=x_0<x_1<\cdots<x_{M_n}=1$, $M_n=n^4$ by equally spaced points, the smoothness of kernel K in Assumption (A4) imples that

$$\begin{aligned}\sup_{x\in[0,1]}\left|n^{-1}\sum_{i=1}^{n}K_h\left(X_i-x\right)r_n\left(X_i\right)\varepsilon_i\right| &\le \max_{0\le j\le M_n}\left|n^{-1}\sum_{i=1}^{n}K_h\left(X_i-x_j\right)r_n\left(X_i\right)\varepsilon_i\right|\\ &+\max_{0\le j<M_n}\sup_{x\in\left[x_j,x_{j+1}\right]}\left|n^{-1}\sum_{i=1}^{n}K_h\left(X_i-x_j\right)r_n\left(X_i\right)\varepsilon_i-n^{-1}\sum_{i=1}^{n}K_h\left(X_i-x\right)r_n\left(X_i\right)\varepsilon_i\right|\\ &\le\max_{0\le j<M_n}\left|n^{-1}\sum_{i=1}^{n}K_h\left(X_i-x_j\right)r_n\left(X_i\right)\varepsilon_i\right|+CM_n^{-1}h^{-2}n^{-1}\sum_{i=1}^{n}\left|r_n\left(X_i\right)\varepsilon_i\right|\end{aligned}$$

and the moment conditions on error ε in Assumption (A2), the rate of h in Assumption (A5) imply next that

$$\begin{aligned}&\sup_{x\in[0,1]}\left|n^{-1}\sum_{i=1}^{n}K_h\left(X_i-x\right)r_n\left(X_i\right)\varepsilon_i\right|\\ &\quad\le\max_{0\le j<M_n}\left|n^{-1}\sum_{i=1}^{n}K_h\left(X_i-x_j\right)r_n\left(X_i\right)\varepsilon_i\right|+\mathcal{O}_p\left\{\rho_n n^{-2}\right\}.\end{aligned} \tag{26}$$

Step 2. To truncate the error, we denote $D_n=n^{\alpha}$ with α as in Assumption (A5). Assumption (A5) implies that $D_n n^{-1/2}h^{-1/2}\log^2 n\to 0$, $n^{1/2}h^{1/2}D_n^{-(1+\eta)}\to 0$, $\sum_{n=1}^{\infty}D_n^{-(2+\eta)}<\infty$. Write $\varepsilon_i=\varepsilon_{i,1}^{D_n}+\varepsilon_{i,2}^{D_n}$, where $\varepsilon_{i,1}^{D_n}=\varepsilon_i\mathrm{I}\left\{\left|\varepsilon_i\right|>D_n\right\}$, $\varepsilon_{i,2}^{D_n}=\varepsilon_i\mathrm{I}\left\{\left|\varepsilon_i\right|\le D_n\right\}$, and denote $\mu^{D_n}(x)=\mathsf{E}\left\{\varepsilon_i\mathrm{I}\left\{\left|\varepsilon_i\right|\le D_n\right\}\mid X_i=x\right\}$. One immediately obtains that

$$\begin{aligned}\sup_{x\in[0,1]}\left|\mu^{D_n}(x)\right|&\le\mathsf{E}\left\{|\varepsilon|^{2+\eta}\mid X=x\right\}D_n^{-(1+\eta)}=o\left(n^{-1/2}h^{-1/2}\right),\\ \sup_{x\in[0,1]}\left|\mathsf{E}n^{-1}\sum_{i=1}^{n}K_h\left(X_i-x\right)r_n\left(X_i\right)\mu^{D_n}\left(X_i\right)\right|&=o\left(n^{-1/2}h^{-1/2}\rho_n\right).\end{aligned} \tag{27}$$

Next, since $P\left(\left|\varepsilon_i\right|>D_n\right)\le\mathsf{E}|\varepsilon|^{2+\eta}D_n^{-(2+\eta)}$, $\sum_{n=1}^{\infty}P\left(\left|\varepsilon_n\right|>D_n\right)\le\mathsf{E}|\varepsilon|^{2+\eta}\sum_{n=1}^{\infty}D_n^{-(2+\eta)}<+\infty$, Borel–Cantelli Lemma then implies that

$$P\left\{\omega \mid \exists N(\omega), \varepsilon_{i,1}^{D_n}(\omega) = 0, i = 1, 2, \ldots n, \text{for } n > N(\omega)\right\} = 1.$$

So one has for any $k > 0$

$$\sup_{x\in[0,1]} n^{-1}\left|\sum\nolimits_{i=1}^{n} K_h(X_i - x) r_n(X_i)\varepsilon_{i,1}^{D_n}\right| = \mathcal{O}_{a.s.}\left(n^{-k}\rho_n\right). \tag{28}$$

Step 3. The truncated sum $n^{-1}\sum_{i=1}^{n} K_h(X_i - x) r_n(X_i)\varepsilon_{i,2}^{D_n}$ equals $\int_{|\varepsilon|\le D_n} K_h(u-x)$ $r_n(u)\varepsilon dF_n(u,\varepsilon)$, while

$$\int_{|\varepsilon|\le D_n} K_h(u-x) r_n(u)\varepsilon dF(u,\varepsilon) = \mathsf{E}n^{-1}\sum_{i=1}^{n} K_h(X_i - x) r_n(X_i)\varepsilon_{i,2}^{D_n}$$
$$= \mathsf{E}n^{-1}\sum_{i=1}^{n} K_h(X_i - x) r_n(X_i)\mu^{D_n}(X_i) = u\left(n^{-1/2}h^{-1/2}\rho_n\right),$$

according to (27). The above two Equations imply that

$$n^{-1}\sum_{i=1}^{n} K_h(X_i - x) r_n(X_i)\varepsilon_{i,2}^{D_n} = n^{-1/2}\int_{|\varepsilon|\le D_n} K_h(u-x) r_n(u)\varepsilon dZ_n(u,\varepsilon)$$
$$+u\left(n^{-1/2}h^{-1/2}\rho_n\right). \tag{29}$$

Step 4. The term $n^{-1/2}\int_{|\varepsilon|\le D_n} K_h(u-x) r_n(u)\varepsilon dZ_n(u,\varepsilon)$ equals

$$-n^{-1/2}\int_{|\varepsilon|\le D_n} d\left\{K_h(u-x) r_n(u)\varepsilon\right\} Z_n(u,\varepsilon)$$
$$= n^{-1/2}\int_{|\varepsilon|\le D_n} d\left\{K_h(u-x) r_n(u)\varepsilon\right\}\left[B_n\left\{M(u,\varepsilon)\right\} - Z_n(u,\varepsilon)\right]$$
$$-n^{-1/2}\int_{|\varepsilon|\le D_n} d\left\{K_h(u-x) r_n(u)\varepsilon\right\} B_n\left\{M(u,\varepsilon)\right\}.$$

Note that

$$n^{-1/2}\int_{|\varepsilon|\le D_n} d\left\{K_h(u-x) r_n(u)\varepsilon\right\}\left[B_n\left\{M(u,\varepsilon)\right\} - Z_n(u,\varepsilon)\right]$$
$$= n^{-1/2}\int_{|\varepsilon|\le D_n}\left\{dK_h(u-x) r_n(u) + K_h(u-x) dr_n(u)\right\} d\varepsilon\left[B_n\{M(u,\varepsilon)\} - Z_n(u,\varepsilon)\right]$$
$$= U_{a.s.}\left\{n^{-1/2}\log^2 n \times D_n\left(h^{-1}\rho_n + l_n\right) n^{-1/2}\right\}$$
$$= U_{a.s.}\left\{D_n n^{-1}h^{-1}\log^2 n\left(\rho_n + hl_n\right)\right\}$$
$$= u_{a.s.}\left\{n^{-1/2}h^{-1/2}\left(\rho_n + hl_n\right)\right\} \tag{30}$$

by the growth constraint on $D_n = n^{\alpha}$. Note also that

$$\begin{aligned}
&-n^{-1/2} \int_{|\varepsilon| \le D_n} d\left\{K_h(u-x) r_n(u) \varepsilon\right\} B_n\left\{M(u,\varepsilon)\right\} \\
&= n^{-1/2} \int_{|\varepsilon| \le D_n} K_h(u-x) r_n(u) \varepsilon d B_n\left\{M(u,\varepsilon)\right\} \\
&= n^{-1/2} \int_{|\varepsilon| \le D_n} K_h(u-x) r_n(u) \varepsilon d W_n\left\{M(u,\varepsilon)\right\} \\
&\quad -n^{-1/2} \int_{|\varepsilon| \le D_n} K_h(u-x) r_n(u) \varepsilon d F(\varepsilon\,|u) f(u) du W_n(1,1)
\end{aligned}$$

in which

$$\begin{aligned}
&\left| n^{-1/2} \int_{|\varepsilon| \le D_n} K_h(u-x) r_n(u) \varepsilon d F(\varepsilon\,|u) f(u) du W_n(1,1) \right| \\
&\le n^{-1/2} \int_{|\varepsilon| \le D_n} |\varepsilon| \, dF(\varepsilon\,|u) \int K_h(u-x) \left|r_n(u)\right| f(u) du \left|W_n(1,1)\right| \\
&\le n^{-1/2} \mathsf{E} |\varepsilon|^{2+\eta} D_n^{-(1+\eta)} \rho_n C_f \left|W_n(1,1)\right| = U_p\left(n^{-1/2} D_n^{-(1+\eta)} \rho_n\right). \qquad (31)
\end{aligned}$$

Meanwhile

$$\begin{aligned}
&\mathsf{E}\left[n^{-1/2} \int_{|\varepsilon| \le D_n} K_h(u-x) r_n(u) \varepsilon d W_n\left\{M(u,\varepsilon)\right\}\right]^2 \\
&= n^{-1} \int_{|\varepsilon| \le D_n} K_h(u-x)^2 r_n^2(u) \varepsilon^2 d F(\varepsilon\,|u) f(u) du \\
&= n^{-1} \int \left\{ \int_{|\varepsilon| \le D_n} \varepsilon^2 d F(\varepsilon\,|u) \right\} K_h(u-x)^2 r_n^2(u) f(u) du \\
&\le n^{-1} \int K_h(u-x)^2 r_n^2(u) \sigma^2(u) f(u) du \le n^{-1} h^{-1} \rho_n^2 C_\sigma^2 C_f,
\end{aligned}$$

so the M_n Gaussian variables $n^{-1/2} \int_{|\varepsilon| \le D_n} K_h\left(u - x_j\right) r_n(u) \varepsilon d W_n\left\{M(u,\varepsilon)\right\}, 0 \le j < M_n$ each has variance less than $n^{-1} h^{-1} \rho_n^2 C_\sigma^2 C_f$, hence

$$\begin{aligned}
&\max_{0 \le j < M_n} \left| n^{-1/2} \int_{|\varepsilon| \le D_n} K_h\left(u-x_j\right) r_n(u) \varepsilon d W_n\left\{M(u,\varepsilon)\right\} \right| \\
&= \mathcal{O}_p\left(n^{-1/2} h^{-1/2} \rho_n \log^{1/2} n\right). \qquad (32)
\end{aligned}$$

Finally, putting together Eqs. (26), (28), (29), (30), (31) and (32) proves the Lemma.

A.2 Proof of Propositions

Proof of Proposition 1 It is obvious that $\left|\mathrm{I}_{i,p}\right| \leq 2\left\{\tilde{m}_p(X_i) - m(X_i)\right\}^2 + 2\tilde{\varepsilon}_p^2(X_i)$. Meanwhile applied Lemma 1, $\left\|m - \tilde{m}_p\right\|_{2,n}^2 \leq \left\|m - \tilde{m}_p\right\|_{\infty}^2 = \mathcal{O}_p\left(H^{2p}\right)$, |I| is bounded by

$$\begin{aligned} & 2n^{-1}h^{-1}c^{-1}\sum\nolimits_{i=1}^{n}\left|K\left\{(X_i - x)/h\right\}\right|\left\{\left(m(X_i) - \tilde{m}_p(X_i)\right)^2 + \tilde{\varepsilon}_p^2(X_i)\right\} \\ & \leq 2h^{-1}c^{-1}\left\|K\right\|_{\infty}\left\{\left\|m - \tilde{m}_p\right\|_{2,n}^2 + \left\|\tilde{\varepsilon}_p^2\right\|_{2,n}^2\right\} \\ & \leq 2c^{-1}\left\|K\right\|_{\infty}\left\{h^{-1}\left(H^{2p} + n^{-1}H^{-1}\right)\right\}. \end{aligned}$$

Proof of Proposition 2 By (12), $\tilde{\varepsilon}_p(X_i) = \sum_{J=1-p}^{N}\tilde{a}_{J,p}B_{J,p}(X_i)$, Lemma 1 and Wang and Yang (2009) entail that $(\sum_{J=1-p}^{N}\tilde{a}_{J,p}^2)^{1/2} = \mathcal{O}_p(\|\tilde{\varepsilon}_p(x)\|_{2,n}) = \mathcal{O}_p(n^{-1/2}N^{1/2})$. Set $r_n(x) = B_{J,p}(x)$, then Lemma 2 entails that $\rho_n = \mathcal{O}\left(H^{-1/2}\right)$ and it is easy to verify that $l_n = \mathcal{O}\left(H^{-3/2}\right)$. Applying Lemma 4, one obtains that

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}K_h(X_i - x)\,\varepsilon_i B_{J,p}(X_i) \\ & = \mathcal{U}_p\left(n^{-1/2}h^{-1/2}H^{-1/2}\log^{1/2}n + n^{-1/2}h^{1/2}H^{-3/2}\right) \end{aligned} \tag{33}$$

and hence

$$\begin{aligned} |\mathrm{II}(x)| &= \left|n^{-1}\sum_{i=1}^{n}K_h(X_i - x)\,2\varepsilon_i\tilde{\varepsilon}_p(X_i)\right| \\ &= \left|2n^{-1}\sum_{i=1}^{n}K_h(X_i - x)\,\varepsilon_i\sum_{J=1-p}^{N}\tilde{a}_{J,p}B_{J,p}(X_i)\right| \\ &\leq c^{-1}\left\{\sum_{J=1-p}^{N}\tilde{a}_{J,p}^2\sum_{J=1-p}^{N}\left\{2n^{-1}\sum_{i=1}^{n}K_h(X_i - x)\,\varepsilon_i B_{J,p}(X_i)\right\}^2\right\}^{1/2} \\ &= \mathcal{O}_p\left(n^{-1/2}N^{1/2}\right) \times (N + p)^{1/2} \\ &\quad \times\mathcal{U}_p\left(n^{-1/2}h^{-1/2}H^{-1/2}\log^{1/2}n + n^{-1/2}h^{1/2}H^{-3/2}\right) \\ &= \mathcal{U}_p\left(n^{-1}h^{-1/2}H^{-3/2}\log^{1/2}n + n^{-1}h^{1/2}H^{-5/2}\right), \end{aligned}$$

the lemma is proved.

Proof of Proposition 3

$$\begin{aligned}\mathrm{III}(x) &= 2n^{-1}\sum_{i=1}^{n} K_h(X_i - x)\left\{\left(m(X_i) - g_p(X_i)\right)\varepsilon_i\right\} \\ &\quad + 2n^{-1}\sum_{i=1}^{n} K_h(X_i - x)\left\{\left(g_p(X_i) - \tilde{m}_p(X_i)\right)\varepsilon_i\right\}\end{aligned}$$

in which the spline function $g_p \in G_N^{(p-2)}$ satisfies $\left\|m - g_p\right\|_\infty \le CH^p, m - g_p \in \mathrm{Lip}\left([0,1], CH^{p-1}\right)$ as in Lemma 1. Set $r_n(x) = m(x) - g_p(x)$, then $\rho_n = \mathcal{O}(H^p), l_n = \mathcal{O}\left(H^{p-1}\right)$, so applying Lemma 4 yields

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n} K_h(X_i - x)\left\{\left(m(X_i) - g_p(X_i)\right)\varepsilon_i\right\} \\ &= \mathcal{U}_p\left\{(nh/\log n)^{-1/2}H^p + n^{-1/2}h^{1/2}H^{p-1}\right\}. \end{aligned} \tag{34}$$

Denoting $g_p(x) - \tilde{m}_p(x) = \sum_{J=1-p}^{N}\gamma_{J,p}B_{J,p}(x)$ and applying Lemma 1, one has

$$\left(\sum\nolimits_{J=1-p}^{N}\gamma_{J,p}^2\right)^{1/2} \le C\left\|g_p(x) - \tilde{m}_p(x)\right\|_2 = \mathcal{O}\left(H^p\right),$$

which, together with (33) imply that

$$\begin{aligned} & \left|2n^{-1}\sum_{i=1}^{n} K_h(X_i - x)\left\{g_p(X_i) - \tilde{m}_p(X_i)\right\}\varepsilon_i\right| \\ &= \left|2n^{-1}\sum_{J=1-p}^{N}\gamma_{J,p}\sum\nolimits_{i=1}^{n} K_h(X_i - x)B_{J,p}(X_i)\varepsilon_i\right| \\ &\le \left(\sum_{J=1-p}^{N}\gamma_{J,p}^2\right)^{1/2}\left[\sum_{J=1-p}^{N}\left\{2n^{-1}\sum\nolimits_{i=1}^{n} K_h(X_i - x)B_{J,p}(X_i)\varepsilon_i\right\}^2\right]^{1/2} \\ &= \mathcal{O}\left(H^p\right) \times \mathcal{O}\left(N^{1/2}\right) \times \mathcal{U}_p\left(n^{-1/2}h^{-1/2}H^{-1/2}\log^{1/2}n + n^{-1/2}h^{1/2}H^{-3/2}\right) \\ &= \mathcal{U}_p\left(n^{-1/2}h^{-1/2}H^{p-1}\log^{1/2}n + n^{-1/2}h^{1/2}H^{p-2}\right), \end{aligned}$$

which, together with (34), prove the lemma.

References

Akritas MG, Van Keilegom I (2001) ANCOVA methods for heteroscedastic nonparametric regression models. J Am Stat Assoc 96:220–232

Bickel PJ, Rosenblatt M (1973) On some global measures of deviations of density function estimates. Ann Stat 31:1852–1884

Brown DL, Levine M (2007) variance estimation in nonparametric regression via the difference sequence method. Ann Stat 35:2219–2232

Cai T, Wang L (2008) Adaptive variance function estimation in heteroscedastic nonparametric regression. Ann Stat 36:2025–2054

Carroll RJ, Wang Y (2008) Nonparametric variance estimation in the analysis of microarray data: a measurement error approach. Biometrika 95:437–449

Carroll RJ, Ruppert D (1988) Transformations and weighting in regression. Champman and Hall, London

Claeskens G, Van Keilegom I (2003) Bootstrap confidence bands for regression curves and their derivatives. Ann Stat 31:1852–1884

Davidian M, Carroll RJ, Smith W (1988) Variance functions and the minimum detectable concentration in assays. Biometrika 75:549–556

De Boor C (2001) A practical guide to splines. Springer, New York

Dette H, Munk A (1998) Testing heteroscedasticity in nonparametric regression. J R Stat Soc Ser B 60:693–708

Fan J, Gijbels T (1996) Local polynomial modelling and its applications. Champman and Hall, London

Fan J, Yao Q (1998) Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85:645–660

Hall P, Titterington MD (1988) On confidence bands in nonparametric density estimation and regression. J Multivar Anal 27:228–254

Hall P, Carroll RJ (1989) Variance function estimation in regression: the effect of estimating the mean. J R Stat Soc Ser B 51:3–14

Hall P, Marron JS (1990) On variance estimation in nonparametric regression. Biometrika 77:415–419

Härdle W (1989) Asmptotic maximal deviation of M-smoothers. J Multivar Anal 29:163–179

Härdle W (1992) Applied nonparametric regression. Cambridge University Press, Cambridge

Levine M (2006) Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: a possible approach. Comput Stat Data Anal 50:3405–3431

Liu R, Yang L, Härdle W (2013) Oracally efficient two-step estimation of generalized additive model. J Am Stat Assoc 108:619–631

Ma S, Yang L, Carroll RJ (2012) A simultaneous confidence band for sparse longitudinal regression. Stat Sin 22:95–122

Müller HG, Stadtmüller U (1987) Estimation of heteroscedasticity in regression analysis. Ann Stat 15:610–625

Silverman WB (1986) Density estimation for statistics and data analysis. Chapman and Hall, London

Song Q, Yang L (2009) Spline confidence bands for variance functions. J Nonparametr Stat 5:589–609

Tusnády G (1977) A remark on the approximation of the sample df in the multidimensional case. Periodica Mathematica Hungarica 8:53–55

Wang L, Brown LD, Cai T, Levine M (2008) Effect of mean on variance function estimation in nonparametric regression. Ann Stat 36:646–664

Wang J, Liu R, Cheng F, Yang L (2014) Oracally efficient estimation of autoregressive error distribution with simultaneous confidence band. Ann Stat 42:654–668

Wang L, Yang L (2007) Spline-backfitted kernel smoothing of nonlinear additive autoregression model. Ann Stat 35:2474–2503

Wang J, Yang L (2009a) Polynomial spline confidence bands for regression curves. Stat Sin 19:325–342

Wang L, Yang L (2009b) Spline estimation of single-index models. Stat Sin 19:765–783

Wang J, Yang L (2009c) Efficient and fast spline-backfitted kernel smoothing of additive models. Ann Inst Stat Math 61:663–690

Xia Y (1998) Bias-corrected confidence bands in nonparametric regression. J R Stat Soc Ser B 60:797–811

Xue L, Yang L (2006) Additive coefficient modeling via polynomial spline. Stat Sin 16:1423–1446

Zheng S, Yang L, Härdle W (2014) A smooth simultaneous confidence corridor for the mean of sparse functional data. J Am Stat Assoc 109:661–673