Annals of the Institute of Statistical Mathematics (2020) 72:471–509
https://doi.org/10.1007/s10463-018-0697-2

Convergence rates for kernel regression in infinite-dimensional spaces

Joydeep Chowdhury[1] · **Probal Chaudhuri**[1]

Received: 17 September 2016 / Revised: 2 October 2018 / Published online: 17 November 2018

Abstract
We consider a nonparametric regression setup, where the covariate is a random element in a complete separable metric space, and the parameter of interest associated with the conditional distribution of the response lies in a separable Banach space. We derive the optimum convergence rate for the kernel estimate of the parameter in this setup. The small ball probability in the covariate space plays a critical role in determining the asymptotic variance of kernel estimates. Unlike the case of finite-dimensional covariates, we show that the asymptotic orders of the bias and the variance of the estimate achieving the optimum convergence rate may be different for infinite-dimensional covariates. Also, the bandwidth, which balances the bias and the variance, may lead to an estimate with suboptimal mean square error for infinite-dimensional covariates. We describe a data-driven adaptive choice of the bandwidth and derive the asymptotic behavior of the adaptive estimate.

Keywords Adaptive estimate · Bias-variance decomposition · Gaussian process · Maximum likelihood regression · Mean square error · Optimal bandwidth · Small ball probability · t process

1 Introduction

Suppose that we have a nonparametric regression problem, where the covariate $\mathbf{X}$ is a random element in a complete separable metric space, and the response $\mathbf{Y}$ lies in some arbitrary measure space. Our parameter of interest, which is denoted as $\Theta(\mathbf{x})$,

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-018-0697-2) contains supplementary material, which is available to authorized users.

✉ Joydeep Chowdhury
joydeepchowdhury01@gmail.com

Probal Chaudhuri
probal@isical.ac.in

[1] Statistics and Mathematics Unit, Indian Statistical Institute, 203, B.T. Road, Kolkata 700108, India

is a parameter associated with the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. Let $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ be a sample of *i.i.d.* observations from the joint distribution of $(\mathbf{X}, \mathbf{Y})$, and our objective is to estimate $\Theta(\mathbf{x})$ based on this sample. In the particular case, where the response $\mathbf{Y}$ is a real random variable, the covariate space is the q-dimensional Euclidean space $\mathbb{R}^q$ and $\Theta(\mathbf{x}) = \mathbb{E}[\mathbf{Y} \,|\, \mathbf{X} = \mathbf{x}]$, Stone (1980) proved that the optimal convergence rate of a nonparametric estimate $\widehat{\Theta}_n(\mathbf{x})$ of $\Theta(\mathbf{x})$ is $n^{-(\beta/(2\beta+q))}$. Here, β is a positive constant such that $|\Theta(\mathbf{z}) - \Theta(\mathbf{x})| = O(\|\mathbf{z}-\mathbf{x}\|^{\beta})$ as $\mathbf{z} \longrightarrow \mathbf{x}$, with $\|\cdot\|$ being the Euclidean norm in $\mathbb{R}^q$. The optimum achievable convergence rate for nonparametric regression with finite-dimensional covariate was further investigated in Stone (1982), Ibragimov and Haśminskii (1980), Yatracos (1988), Donoho and Liu (1991a, b), etc. However, when the dimension of the covariate space is infinite, the expressions of the optimum rate of convergence derived by these authors are no longer valid.

Recently, nonparametric regression with functional covariates has been studied in Masry (2005), Ferraty et al. (2007), Rachdi and Vieu (2007), etc. These authors investigated nonparametric estimation of the conditional mean when the covariate is functional, and the response is real-valued. They investigated the consistency and the asymptotic normality of kernel estimates as well as data-driven selection of the bandwidth. In Ferraty et al. (2012) and Lian (2012), the problem of nonparametric regression when both the response and the covariate are functions is investigated, where the parameter of interest is the conditional mean of the response given the covariate. In Ferraty et al. (2012), asymptotic normality of the estimate of the conditional mean is derived and a bootstrap implementation is described. In Lian (2012), an upper bound of the convergence rate of the estimate of the conditional mean is established.

The problem of optimum convergence rate of a nonparametric regression estimate was explored in Mas (2012) and Chagny and Roche (2014) when the covariate is infinite-dimensional. In Mas (2012), the usual mean regression problem with a real-valued response was considered, and a lower bound for the rate of convergence of the minimax risk was established (see Theorem 3 in Mas 2012). In Chagny and Roche (2014), the optimum convergence rate was derived for the estimate of the conditional distribution function of a real-valued response given an infinite-dimensional covariate. In both these cases, the methodology developed is restricted to the conditional mean of some real-valued response, and cannot be applied when the response is infinite-dimensional, or the parameter of interest is not the conditional mean.

In most of the existing literature on nonparametric regression with functional data, the authors considered real or multivariate responses and functional covariates. However, regression problems, where the response itself may be infinite-dimensional in nature, are also common. Authors who investigated regression with functional responses and covariates, like Ferraty et al. (2012) and Lian (2012), considered the conditional mean as their parameter of interest. But, one may also be interested in various parameters of the conditional distribution of the response other than the conditional mean, like the conditional variance and covariance, the conditional coefficient of variation, the conditional correlation, etc. In our study, the parameter of interest $\Theta(\mathbf{x})$ lies in some separable Banach space, and it covers a large class of parameters of interest including those stated above. We shall investigate the convergence rate of a large class of kernel estimates in this setup and derive the optimal convergence rate.

In Sect. 2, our regression setup and the kernel estimates are described in detail. In Sect. 3, we discuss an asymptotic bias-variance decomposition of our kernel estimate, and study the asymptotic behavior of the bias and the variance terms. We show that the asymptotic behavior of the variance term critically depends on the small ball probability in the covariate space. The main convergence results for the estimate $\widehat{\Theta}_n(\mathbf{x})$ are presented in Sect. 4. A data-driven method of bandwidth selection along with the asymptotic behavior of the adaptive estimate is presented in Sect. 5. In the same section, we demonstrate the adaptive estimates in simulated datasets from several regression models. Section 6 contains concluding remarks and discussion. The proofs and related mathematical details are provided in Sect. 7 and in the supplement.

2 Kernel estimates

We assume that the covariate $\mathbf{X}$ is a random element in some complete separable metric space $(\mathcal{C}, d)$ with d being the metric, and the response $\mathbf{Y}$ is a random element in some measurable space $\mathcal{R}$ equipped with some appropriate σ-field and probability measure. Denote the conditional probability measure of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ as $\mu(\cdot \,|\, \mathbf{x})$. We want to estimate a parameter $\Theta(\mathbf{x})$ associated with $\mu(\cdot \,|\, \mathbf{x})$ for a fixed $\mathbf{x} \in \mathcal{C}$. We employ the nonparametric kernel regression method developed by Nadaraya (1964) and Watson (1964). Let $K(\cdot)$ be a suitable kernel function with associated bandwidth $h > 0$. To estimate $\Theta(\mathbf{x})$, we first construct the weighted empirical probability measure $\mu_n(\cdot \,|\, \mathbf{x})$ that assigns probability mass

$$W_{i,n} = \frac{K(h^{-1}d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^{n} K(h^{-1}d(\mathbf{x}, \mathbf{X}_i))}$$

to the data point $\mathbf{Y}_i$ for $i = 1, \ldots, n$. The kernel estimate $\widehat{\Theta}_n(\mathbf{x})$ of $\Theta(\mathbf{x})$ is the corresponding parameter associated with $\mu_n(\cdot \,|\, \mathbf{x})$.

We require the concept of a type 2 Banach spaces in our subsequent discussion. A separable Banach space is called type 2 if there is a positive constant c such that for any $n \geq 1$ and independent zero-mean random elements $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ in that Banach space with $\mathbb{E}\|\mathbf{Z}_i\|^2 < \infty$ for $i = 1, \ldots, n$, we have $\mathbb{E}\|\mathbf{Z}_1 + \cdots + \mathbf{Z}_n\|^2 \leq c(\mathbb{E}\|\mathbf{Z}_1\|^2 + \cdots + \mathbb{E}\|\mathbf{Z}_n\|^2)$ (Araujo and Giné 1980, p. 158). Also, a Banach space is said to have a Schauder basis $\{e_n\}$ if for every element v in that space, $v = \sum_{n=1}^{\infty} v_n e_n$ for some sequence of real numbers $\{v_n\}$. Separable Hilbert spaces and $L_p[a, b]$ spaces with $p \geq 2$ and $-\infty \leq a < b \leq \infty$ are well-known examples of type 2 Banach spaces with Schauder bases. Henceforth, $\mathbb{I}(\cdot)$ will denote the usual indicator function.

We now discuss some examples. These examples demonstrate the use of kernel estimates in a diverse class of statistical models. In all our subsequent discussion, the expectation of a random element in a separable Banach space is defined in the sense of Bochner (Araujo and Giné 1980, p. 100).

Example 1 (*Mean regression*): Consider $\Theta(\mathbf{x}) = \mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x}] \in \mathcal{B}$, where $\mathcal{B}$ is a type 2 Banach space, and $\Psi(\cdot)$ is a function from $\mathcal{R}$ to $\mathcal{B}$. The estimate $\widehat{\Theta}_n(\mathbf{x})$ of $\Theta(\mathbf{x})$ is

$$\widehat{\Theta}_n(\mathbf{x}) = \frac{\sum_{i=1}^n \Psi(\mathbf{Y}_i) K(h^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^n K(h^{-1} d(\mathbf{x}, \mathbf{X}_i))}.$$

Some examples of $\Psi(\cdot)$ and the resulting $\Theta(\mathbf{x})$ are the following. Let $\mathbf{Y} \in \mathbb{R}$, and $\Psi(\mathbf{Y}) = \mathbb{I}(\mathbf{Y} \leq y)$, where $y \in \mathbb{R}$. Then, $\Theta(\mathbf{x})$ is the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ at y (see Ferraty et al. (2006), Chagny and Roche (2014), etc.). Alternatively, if $\Psi(\mathbf{Y}) = \mathbf{Y}^r$, $\Theta(\mathbf{x})$ is the conditional rth moment of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. Next, let $\mathbf{Y}$ be a random vector in $\mathbb{R}^q$. For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^q$ with $\mathbf{u} = [u_1, \ldots, u_q]$ and $\mathbf{v} = [v_1, \ldots, v_q]$, let $\mathbf{u} \leq \mathbf{v}$ denote $u_i \leq v_i$ for $i = 1, \ldots, q$. Then, for $\Psi(\mathbf{Y}) = \mathbb{I}(\mathbf{Y} \leq \mathbf{y})$, where $\mathbf{y} \in \mathbb{R}^q$, $\Theta(\mathbf{x})$ becomes the conditional multivariate distribution of $\mathbf{Y}$ at $\mathbf{y}$ given $\mathbf{X} = \mathbf{x}$. When $\mathbf{Y}$ is a univariate or multivariate random variable, the choice $\Psi(\mathbf{Y}) = \mathbf{Y}$ corresponds to the conditional mean of a univariate or multivariate response given $\mathbf{X} = \mathbf{x}$ (see, e.g., Ferraty and Vieu (2006), Ferraty et al. (2007), etc.). Similarly, when $\mathbf{Y} \in \mathcal{B}$ and $\mathcal{B}$ is a separable Hilbert space, the choices $\Psi(\mathbf{Y}) = \mathbf{Y}$ and $\Psi(\mathbf{Y}) = \mathbf{Y} \otimes \mathbf{Y}$ (the outer product of $\mathbf{Y}$ with itself) correspond to the conditional mean and the second conditional moment of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$, respectively (see Ferraty et al. (2012)). Note that when $\mathcal{B} = \mathbb{R}^q$, $\mathbf{Y} \otimes \mathbf{Y}$ becomes the $q \times q$ matrix $\mathbf{Y}\mathbf{Y}^t$.

Example 2 (*Functions of conditional mean*): Let $\mathcal{B}_1, \mathcal{B}_2$ be two separable Banach spaces, $\mathcal{U}$ be an open subset of $\mathcal{B}_1$, and $\Psi(\cdot) : \mathcal{R} \longrightarrow \mathcal{B}_1$ be such that $\mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x}] \in \mathcal{U}$. For $\Gamma(\cdot) : \mathcal{U} \longrightarrow \mathcal{B}_2$, we consider $\Theta(\mathbf{x}) = \Gamma(\mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x}])$. Here, the kernel regression estimate $\widehat{\Theta}_n(\mathbf{x})$ is

$$\widehat{\Theta}_n(\mathbf{x}) = \Gamma\left(\frac{\sum_{i=1}^n \Psi(\mathbf{Y}_i) K(h^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^n K(h^{-1} d(\mathbf{x}, \mathbf{X}_i))}\right).$$

As a special case, let $\mathbf{Y}$ be a real random variable. Let $\Psi(\mathbf{Y}) = (\mathbf{Y}^2, \mathbf{Y})$ and $\mathcal{U} = \{(u, v) \in \mathbb{R}^2 \,|\, u > v^2,\ v > 0\}$. Let $\Gamma(\cdot) : \mathcal{U} \longrightarrow \mathbb{R}$ be defined by $\Gamma(u, v) = v^{-1}\sqrt{u - v^2}$. Then, $\Theta(\mathbf{x}) = \Gamma(\mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x}])$ is the conditional coefficient of variation of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ (see, e.g., Dette and Wieczorek (2009); Dette et al. (2012)).

As another special case, let $\mathbf{Y} = (Y_1, Y_2)$ be a bivariate random variable. Let $\Psi(\mathbf{Y}) = \Psi(Y_1, Y_2) = (Y_1 Y_2, Y_1, Y_2, Y_1^2, Y_2^2)$ and $\mathcal{U} = \{(s, t, u, v, w) \in \mathbb{R}^5 \,|\, v > t^2, w > u^2\}$. Let $\Gamma(\cdot) : \mathcal{U} \longrightarrow \mathbb{R}$ be defined by

$$\Gamma(s, t, u, v, w) = \frac{s - tu}{\sqrt{v - t^2}\sqrt{w - u^2}}.$$

Then, $\Theta(\mathbf{x}) = \Gamma(\mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x}])$ is the conditional correlation coefficient of Y_1 and Y_2 given $\mathbf{X} = \mathbf{x}$ (see, e.g., Klemelä (2014, p. 13)).

As the third special case, let the response space $\mathcal{R}$ be a separable Hilbert space, and $\mathcal{B}_2$ denote the space of Hilbert–Schmidt operators on $\mathcal{R}$. Note that $\mathcal{B}_2$ is a Hilbert space (Bhatia 2009, p. 195). Also, the space of finite rank operators is dense in the space of Hilbert–Schmidt operators (Bhatia 2009, p. 196), and the space of finite rank operators on a separable Hilbert space is itself separable. Consequently, $\mathcal{B}_2$ is a separable Hilbert space. Set $\mathcal{B}_1 = \mathcal{B}_2 \times \mathcal{R}$. Define $\Psi(\mathbf{Y}) = (\mathbf{Y} \otimes \mathbf{Y}, \mathbf{Y})$, $\mathcal{U} = \mathcal{B}_1$

and $\Gamma(\mathbf{u}, \mathbf{v}) = \mathbf{u} - \mathbf{v} \otimes \mathbf{v}$. Then, $\Theta(\mathbf{x}) = \mathbb{COV}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$, which is the conditional covariance operator of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ (see, e.g., Ferraty et al. 2012). Note that when $\mathbf{Y}$ is real random variable, this choice of $\Psi(\mathbf{Y})$ and $\Gamma(\cdot, \cdot)$ corresponds to the conditional variance of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$.

Example 3 (*Maximum likelihood regression*): Nonparametric estimation in a likelihood based regression problem with finite-dimensional covariate was investigated in Staniswalis (1989), Chaudhuri and Dewanji (1995) and Aerts and Claeskens (1997). Let the covariate $\mathbf{X}$ and the response $\mathbf{Y}$ be random elements in the complete separable metric spaces $\mathcal{C}$ and $\mathcal{R}$, respectively. Suppose $\mathbf{Y}$ given $\mathbf{X}$ has a conditional density with respect to some sigma-finite measure in $\mathcal{R}$, and it is given by $f(\cdot \mid \Theta(\mathbf{x}))$ for $\mathbf{X} = \mathbf{x}$, where $\Theta(\cdot) : \mathcal{C} \longrightarrow \mathbb{R}^q$. We assume that the form of the function $f(\cdot \mid \cdot)$ is known, but $\Theta(\cdot)$ is unknown. We are interested in estimating $\Theta(\mathbf{x})$ using maximum weighted likelihood procedure, where $\mathbf{x} \in \mathcal{C}$ is fixed. The kernel estimate $\widehat{\Theta}_n(\mathbf{x})$ of $\Theta(\mathbf{x})$ is given by

$$\widehat{\Theta}_n(\mathbf{x}) = \arg\max_{\mathbf{t} \in \mathbb{R}^q} \prod_{i=1}^{n} [f(\mathbf{Y}_i \mid \mathbf{t})]^{W_{i,n}(\mathbf{x})}, \text{ where } W_{i,n}(\mathbf{x}) = \frac{K(h^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^{n} K(h^{-1} d(\mathbf{x}, \mathbf{X}_i))}. \tag{1}$$

So, when $f(\mathbf{y} \mid \mathbf{t})$ is a differentiable function of $\mathbf{t} \in \mathbb{R}^q$, $\widehat{\Theta}_n(\mathbf{x})$ is the solution (in $\mathbf{t}$) of the likelihood equation

$$\sum_{i=1}^{n} \left[\nabla(\log f(\mathbf{Y}_i \mid \mathbf{t}))\right] W_{i,n}(\mathbf{x}) = \mathbf{0}. \tag{2}$$

Here, ∇ denotes the gradient vector of first partial derivatives with respect to $\mathbf{t}$.

It is well known that when the covariate $\mathbf{X}$ is finite-dimensional, say $\mathbf{X} \in \mathbb{R}^q$, and $\mathbf{X}$ has a continuous positive density at $\mathbf{x}$, one needs to have a sequence of bandwidths $\{h_n\}$ such that $h_n \longrightarrow 0$ and $nh_n^q \longrightarrow \infty$ as $n \longrightarrow \infty$ to ensure the consistency of the kernel regression estimate $\widehat{\Theta}_n(\mathbf{x})$ (see, e.g., chapter 3 in Hardle (1990)). To deal with covariates, which are not necessarily finite-dimensional, define $\phi(\mathbf{z}, h) = \mathbb{P}[d(\mathbf{z}, \mathbf{X}) \leq h]$. The function $\phi(\mathbf{z}, h)$ is known as the small ball probability function, and it plays an important role in the asymptotic properties of nonparametric regression estimates. We make the following assumptions on the kernel and the sequence of bandwidths, which are required to establish the consistency of the estimates and derive their convergence rates.

A(i) The kernel $K(\cdot)$ is supported on $[0, 1]$ with $K(u)$ being bounded and bounded away from 0 for $0 \leq u \leq 1$, i.e., there are constants $0 < l \leq L$ such that $l \leq K(u) \leq L$ for all $0 \leq u \leq 1$.

A(ii) The bandwidth $h_n \longrightarrow 0$, and $n\phi(\mathbf{x}, h_n) \longrightarrow \infty$ as $n \longrightarrow \infty$.

The choice of the kernel $K(\cdot)$ described in Assumption A(i) is equivalent to the type I kernel described in Ferraty and Vieu (2006, p. 42). This is a popular choice of

kernel in the literature on nonparametric regression involving functional covariates (see, e.g., Ferraty et al. (2006), Burba et al. (2009), Chagny and Roche (2014, 2016), etc.). Note that for $\mathbf{X} \in \mathbb{R}^q$ having a continuous positive density at $\mathbf{x}$, the condition $n\phi(\mathbf{x}, h_n) \longrightarrow \infty$ as $n \longrightarrow \infty$ in Assumption A(ii) is equivalent to $nh_n^q \longrightarrow \infty$ as $n \longrightarrow \infty$. Assumption A(ii) is required to ensure the consistency of the kernel estimates involving an infinite-dimensional covariate, and is also used in earlier works (see, e.g., Ferraty et al. 2007).

3 Bias-variance decomposition

Let $\mathcal{B}$ be a separable type 2 Banach space with a Schauder basis, and $\Theta(\cdot) : \mathcal{C} \longrightarrow \mathcal{B}$. For $\mathbf{x} \in \mathcal{C}$, we consider the class of kernel regression estimates, which satisfy

$$\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x}) = B_n(\mathbf{x}) + V_n(\mathbf{x}) + R_n(\mathbf{x}), \tag{3}$$

where

$$B_n(\mathbf{x}) = \mathbb{L}_{\mathbf{x}} \left(\frac{\sum_{i=1}^n F(\mathbf{X}_i) K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))} - F(\mathbf{x}) \right), \tag{4}$$

$$V_n(\mathbf{x}) = \mathbb{L}_{\mathbf{x}} \left(\frac{\sum_{i=1}^n [G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i) \,|\, \mathbf{X}_i]] K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))} \right). \tag{5}$$

Here, $F(\cdot) : \mathcal{C} \longrightarrow \mathcal{G}$, $G(\cdot) : \mathcal{R} \longrightarrow \mathcal{G}$, $\mathbb{L}_{\mathbf{x}}(\cdot) : \mathcal{G} \longrightarrow \mathcal{B}$ and $\mathcal{G}$ is a separable Banach space. $\mathbb{L}_{\mathbf{x}}(\cdot)$ is a continuous linear map. The functions $F(\cdot)$, $G(\cdot)$ and the remainder term $R_n(\mathbf{x})$ are assumed to satisfy the following conditions.

- B(i) Let $\beta > 0$ be a constant. Then, $F(\cdot) \in \mathcal{F}(\mathbf{x}, \beta, \mathcal{G})$. Here, $\mathcal{F}(\mathbf{x}, \beta, \mathcal{G})$ is a class of functions $F(\cdot) : \mathcal{C} \longrightarrow \mathcal{G}$ such that for some constant $b_F > 0$, $\|F(\mathbf{z}) - F(\mathbf{x})\| \leq b_F d(\mathbf{x}, \mathbf{z})^\beta$ for all $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$.
- B(ii) $G(\cdot)$ is such that for some $\nu > 2$, $\mathbb{E}[\|G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{z}]\|^\nu \,|\, \mathbf{X} = \mathbf{z}]$ is uniformly bounded for $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$.
- B(iii) $R_n(\mathbf{x}) = o_{\mathbb{P}}(\delta_n)$ as $n \longrightarrow \infty$, where $\delta_n = \max\left\{h_n^\beta, \left[n\phi(\mathbf{x}, h_n)\right]^{-1/2}\right\}$.

Note that $\mathbb{E}[V_n(\mathbf{x}) \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n] = \mathbf{0}$. We can view $B_n(\mathbf{x})$ as the bias term and $V_n(\mathbf{x})$ as the variance term in kernel regression. Note that condition B(i) is related to the smoothness of the regression function. Condition B(ii) imposes a bound on the variability of the residual of the regression. Condition B(iii) essentially states that the remainder term in our bias-variance decomposition is asymptotically negligible. We shall now verify the validity of the above bias-variance decomposition in Examples 1, 2 and 3 in Sect. 2.

Example 1 (continued). Recall Example 1 considered in Sect. 2. We can set

$$B_n(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbb{E}[\Psi(\mathbf{Y}_i) \,|\, \mathbf{X}_i] K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))} - \mathbb{E}[\Psi(\mathbf{Y}) \,|\, \mathbf{X} = \mathbf{x}],$$

$$V_n(\mathbf{x}) = \frac{\sum_{i=1}^n [\Psi(\mathbf{Y}_i) - \mathbb{E}[\Psi(\mathbf{Y}_i) \mid \mathbf{X}_i]] K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))},$$

and $R_n(\mathbf{x}) = \mathbf{0}$.

Then, setting $G(\mathbf{Y}) = \Psi(\mathbf{Y})$, $F(\mathbf{X}) = \mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X}]$ and $\mathbb{L}_{\mathbf{x}}(\cdot)$ to be the identity map on $\mathcal{B}$, (3) holds for any kernel satisfying A(i) and any sequence of bandwidths $\{h_n\}$ satisfying A(ii). Here, condition B(iii) is trivially satisfied. Note that in this case, $F(\mathbf{z}) = \Theta(\mathbf{z})$, and so condition B(i) is satisfied when $\Theta(\mathbf{z})$ is Holder continuous at $\mathbf{x}$ with exponent β, and the class $\mathcal{F}(\mathbf{x}, \beta, \mathcal{B})$ can be taken as the class of all Holder continuous functions at $\mathbf{x}$. Condition B(ii) is satisfied when for some $\nu > 2$, $\mathbb{E}[\|\Psi(\mathbf{Y}) - \mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}]\|^\nu \mid \mathbf{X} = \mathbf{z}]$ is uniformly bounded for $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$. In particular, B(ii) holds for the location-scale type model $\Psi(\mathbf{Y}) = l(\mathbf{X}) + s(\mathbf{X})\mathbf{U}$, where $l(\cdot) : \mathcal{C} \longrightarrow \mathcal{B}$ and $s(\cdot) : \mathcal{C} \longrightarrow (0, \infty)$ are continuous functions, and $\mathbf{U}$ is a zero-mean random element in $\mathcal{B}$, which is independent of $\mathbf{X}$ with $\mathbb{E}[\|\mathbf{U}\|^\nu] < \infty$ for some $\nu > 2$.

Example 2 (continued). Consider again the class of estimators described in Example 2. The following proposition asserts that the bias-variance decomposition (3) holds for those estimators.

Theorem 1 *In Example 2 considered in Sect. 2, let $\mathcal{B}_2$ be a type 2 Banach space. Let the kernel function $K(\cdot)$ satisfy A(i) and the bandwidths $\{h_n\}$ satisfy A(ii). Assume that $\Gamma(\cdot)$ is Fréchet differentiable with derivative $\Gamma'(\cdot)$. Let $\mathbb{L}_{\mathbf{x}}(\cdot) = \Gamma'(\mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}])(\cdot)$, $G(\mathbf{Y}) = \Psi(\mathbf{Y})$, $F(\mathbf{z}) = \mathbb{E}[G(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}]$, and conditions B(i) and B(ii) hold. Then, B(iii) is also satisfied, and consequently the bias-variance decomposition in (3) holds.*

Note that in all the specific cases discussed in Example 2, namely, the coefficient of variation, the correlation coefficient and the covariance operator, the function $\Gamma(\cdot)$ satisfies the differentiability condition stated in Theorem 1, and its derivative can be computed in a straight forward way.

When $\mathbf{Y}$ is a real-valued random variable and $\Theta(\mathbf{x})$ is the conditional coefficient of variation of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ as mentioned in Example 2, Assumption B(i) is satisfied if $\mathbb{E}[\mathbf{Y}^2 \mid \mathbf{X} = \mathbf{z}]$ and $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{z}]$ are both Holder continuous at $\mathbf{x}$ with exponent β. Further, the Assumption B(ii) is satisfied if $\mathbb{E}[\mathbf{Y}^{4+\alpha} \mid \mathbf{X} = \mathbf{z}]$ is uniformly bounded for $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$ for some $\alpha > 0$. Note that conditions $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}] > 0$ and $\mathbb{V}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}] > 0$ ensure that $\mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]$ lies in the domain of $\Gamma(\cdot)$.

When $\mathbf{Y} = (Y_1, Y_2)$ is a bivariate random variable, and $\Theta(\mathbf{x})$ is the conditional correlation between Y_1 and Y_2 given $\mathbf{X} = \mathbf{x}$ as mentioned in Example 2, Assumption B(i) is satisfied if each of $\mathbb{E}[Y_1 Y_2 \mid \mathbf{X} = \mathbf{z}]$, $\mathbb{E}[Y_1 \mid \mathbf{X} = \mathbf{z}]$, $\mathbb{E}[Y_2 \mid \mathbf{X} = \mathbf{z}]$, $\mathbb{E}[Y_1^2 \mid \mathbf{X} = \mathbf{z}]$ and $\mathbb{E}[Y_2^2 \mid \mathbf{X} = \mathbf{z}]$ is Holder continuous at $\mathbf{x}$ with exponent β. Assumption B(ii) is satisfied if $\mathbb{E}[\|\mathbf{Y}\|^{4+\alpha} \mid \mathbf{X} = \mathbf{z}]$ is uniformly bounded for $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$ for some $\alpha > 0$. Further, $\mathbb{V}[Y_1 \mid \mathbf{X} = \mathbf{x}]$, $\mathbb{V}[Y_2 \mid \mathbf{X} = \mathbf{x}] > 0$ ensure that $\mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]$ lies in the domain of $\Gamma(\cdot)$.

One can verify that when $\Theta(\mathbf{x})$ is the conditional covariance of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$, Assumption B(i) is satisfied if $\mathbb{E}[\mathbf{Y} \otimes \mathbf{Y} \mid \mathbf{X} = \mathbf{z}]$ and $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{z}]$ are both Holder continuous at $\mathbf{x}$ with exponent β. Assumption B(ii) holds if $\mathbb{E}[\|\mathbf{Y}\|^{4+\alpha} \mid \mathbf{X} = \mathbf{z}]$ is uniformly bounded for $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$ for some $\alpha > 0$.

Example 3 (continued). In the case of Example 3 considered in Sect. 2, define $g(\mathbf{y} \mid \mathbf{t}) = \log f(\mathbf{y} \mid \mathbf{t})$, where $\mathbf{t} \in \mathbb{R}^q$. Let $\mathcal{T}$ be an open ball in $\mathbb{R}^q$ containing the range of $\Theta(\cdot)$. We now assume some Cramer-type regularity conditions on the log-likelihood $g(\mathbf{y} \mid \mathbf{t})$ that are required for asymptotic analysis of weighted maximum likelihood estimates (see, e.g., Chaudhuri and Dewanji 1995). The support of $f(\mathbf{y} \mid \mathbf{t})$ is assumed to be same for all $\mathbf{t} \in \mathcal{T}$, and $g(\mathbf{y} \mid \mathbf{t})$ is assumed to be thrice continuously differentiable with respect to $\mathbf{t}$ for $\mathbf{t} \in \mathcal{T}$. Denote the Hessian matrix of all second-order partial derivatives of $g(\mathbf{y} \mid \mathbf{t})$ with respect to $\mathbf{t}$ as $\Delta_2(g(\mathbf{y} \mid \mathbf{t}))$, and the array of all third-order partial derivatives of $g(\mathbf{y} \mid \mathbf{t})$ with respect to $\mathbf{t}$ as $\Delta_3(g(\mathbf{y} \mid \mathbf{t}))$. Define $\mathbf{I}(\Theta(\mathbf{z})) = -\mathbb{E}[\Delta_2(g(\mathbf{Y} \mid \Theta(\mathbf{z}))) \mid \mathbf{X} = \mathbf{z}]$, and assume that $\mathbf{I}(\Theta(\mathbf{z}))$ is finite, positive definite and continuous for $\mathbf{z}$ lying in a neighborhood of $\mathbf{x}$. Also, assume that for $\mathbf{t} \in \mathcal{T}$, there exist two nonnegative random variables $\mathbf{D}_1(\mathbf{Y} \mid \mathbf{t})$ and $\mathbf{D}_2(\mathbf{Y} \mid \mathbf{t})$ such that $\mathbb{E}[\mathbf{D}_1(\mathbf{Y} \mid \mathbf{t})]^2 < \infty$, $\mathbb{E}[\mathbf{D}_2(\mathbf{Y} \mid \mathbf{t})] < \infty$, and $\|\Delta_2(g(\mathbf{Y} \mid \mathbf{s}))\| \leq \mathbf{D}_1(\mathbf{Y} \mid \mathbf{t})$, $\|\Delta_3(g(\mathbf{Y} \mid \mathbf{s}))\| \leq \mathbf{D}_2(\mathbf{Y} \mid \mathbf{t})$ for any $\mathbf{s}$ in some neighborhood of $\mathbf{t}$ contained in $\mathcal{T}$.

In the next proposition, we see that the decomposition (3) along with conditions B(i)–B(iii) is satisfied for the weighted maximum likelihood estimate $\widehat{\Theta}_n(\mathbf{x})$ defined in (1).

Theorem 2 *In Example 3 considered in Sect. 2, assume that $\Theta(\cdot) \in \mathcal{F}(\mathbf{x}, \beta, \mathbb{R}^q)$ for some $\beta > 0$, where $\mathcal{F}(\mathbf{x}, \beta, \mathbb{R}^q)$ is as defined in B(i). Let the kernel function $K(\cdot)$ satisfy A(i) and the bandwidths $\{h_n\}$ satisfy A(ii). Then, under the Cramer-type regularity conditions stated above, the decomposition (3) along with conditions B(i)–B(iii) will hold for $\widehat{\Theta}_n(\mathbf{x})$ in (1) if we choose $\mathbb{L}_{\mathbf{x}}(\cdot) = [\mathbf{I}(\Theta(\mathbf{x}))]^{-1}(\cdot)$, $G(\mathbf{Y}) = \nabla g(\mathbf{Y} \mid \Theta(\mathbf{X}))$ and $F(\mathbf{X}) = \mathbf{I}(\Theta(\mathbf{x}))(\Theta(\mathbf{X}))$, where $g(\mathbf{y} \mid \mathbf{t}) = \log f(\mathbf{y} \mid \mathbf{t})$.*

3.1 Asymptotic behavior of the bias and the variance

In this subsection, the orders of convergence of the bias term $B_n(\mathbf{x})$ and the variance term $V_n(\mathbf{x})$ in (3) are investigated. It follows from Assumptions A(ii) and B(i) that $\|B_n(\mathbf{x})\| \leq \|\mathbb{L}_{\mathbf{x}}\| b_F h_n^{\beta}$ for all sufficiently large n. So, for all choices of bandwidths $\{h_n\}$ with $h_n \longrightarrow 0^+$ as $n \longrightarrow \infty$,

$$\mathbb{E}\left[\|B_n(\mathbf{x})\|^2\right] \leq (\|\mathbb{L}_{\mathbf{x}}\| b_F)^2 h_n^{2\beta} \tag{6}$$

for all sufficiently large n. The inequality (6) leads to an upper bound of the rate of convergence of the bias term, and it will be used later to study the asymptotic properties of the estimate $\widehat{\Theta}_n(\mathbf{x})$.

We next discuss the asymptotic behavior of the variance term $V_n(\mathbf{x})$. We derive an upper bound of the convergence rate of $\mathbb{E}[\|V_n(\mathbf{x})\|^2]$ in the theorem below.

Theorem 3 *Under A(i), A(ii) and B(ii), $n\phi(\mathbf{x}, h_n)\mathbb{E}[\|V_n(\mathbf{x})\|^2] = O(1)$ as $n \longrightarrow \infty$.*

The following condition is needed to derive the asymptotic distribution of $V_n(\mathbf{x})$.

B(iv) $\mathcal{B}$ is a separable Hilbert space, and $G(\cdot)$ in (5) is such that the covariance operator $\mathbb{D}(\cdot, \cdot \mid \mathbf{z}) : \mathcal{B} \times \mathcal{B} \longrightarrow \mathbb{R}$ defined by $\mathbb{D}(\mathbf{u}, \mathbf{v} \mid \mathbf{z}) = \mathbb{E}[\langle \mathbf{u}, \mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}])\rangle \langle \mathbf{v}, \mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}])\rangle \mid \mathbf{X} = \mathbf{z}]$, where $\mathbf{u}, \mathbf{v} \in \mathcal{B}$,

converges to $\mathbb{D}(\cdot, \cdot \mid \mathbf{x})$ in the trace norm as $\mathbf{z} \longrightarrow \mathbf{x}$, and $\mathbb{D}(\cdot, \cdot \mid \mathbf{x})$ is a bounded positive definite operator.

Condition B(iv) is related to the smoothness of the conditional distribution of the residual in the regression given the covariate, and it holds in many common models. For example, consider the location-scale type model $\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y})) = l(\mathbf{X}) + s(\mathbf{X})\mathbf{U}$, where $l(\cdot) : \mathcal{C} \longrightarrow \mathcal{B}$ and $s(\cdot) : \mathcal{C} \longrightarrow (0, \infty)$ are continuous functions, and $\mathbf{U}$ is a zero-mean random element in $\mathcal{B}$, which is independent of $\mathbf{X}$, having a bounded positive definite covariance operator.

From Assumption A(i), it follows that $l^j \phi(\mathbf{x}, h) \leq \mathbb{E}\left[K^j(h^{-1}d(\mathbf{x}, \mathbf{X}))\right] \leq L^j \phi(\mathbf{x}, h)$ for any positive integer j and any bandwidth $h > 0$. Define

$$E_n^{(j)}(\mathbf{x}) = [\phi(\mathbf{x}, h_n)]^{-1} \mathbb{E}\left[K^j(h_n^{-1}d(\mathbf{x}, \mathbf{X}))\right] \tag{7}$$

for all positive integer j. Note that

$$0 < L^{-1}l \leq \left[E_n^{(2)}(\mathbf{x})\right]^{-1/2} E_n^{(1)}(\mathbf{x}) \leq l^{-1}L < \infty \tag{8}$$

for all n. In the next theorem, we establish the asymptotic Gaussianity of $V_n(\mathbf{x})$.

Theorem 4 *Let the kernel function $K(\cdot)$ satisfy A(i), and the sequence of bandwidths $\{h_n\}$ satisfy A(ii). Then, under conditions B(ii) and B(iv),*

$$[n\phi(\mathbf{x}, h_n)]^{1/2} \left[E_n^{(2)}(\mathbf{x})\right]^{-1/2} E_n^{(1)}(\mathbf{x}) V_n(\mathbf{x}) \longrightarrow \mathbf{W}$$

in distribution as $n \longrightarrow \infty$, where $\mathbf{W}$ is a zero-mean Gaussian random element in $\mathcal{B}$ with covariance operator $\mathbb{D}(\cdot, \cdot \mid \mathbf{x})$.

Recall that $R_n(\mathbf{x}) = \mathbf{0}$ for the mean-type regression problems described in Example 1. So, for these class of regression problems, from Theorem 4 and (3) we get that

$$[n\phi(\mathbf{x}, h_n)]^{1/2} \left[E_n^{(2)}(\mathbf{x})\right]^{-1/2} E_n^{(1)}(\mathbf{x}) \left[\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x}) - B_n(\mathbf{x})\right] \longrightarrow \mathbf{W}$$

in distribution as $n \longrightarrow \infty$. Define

$$e_n[G(\mathbf{Y}) \mid \mathbf{x}] = \frac{\sum_{i=1}^n G(\mathbf{Y}_i) K(h_n^{-1}d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1}d(\mathbf{x}, \mathbf{X}_i))}.$$

The covariance operator $\mathbb{D}(\cdot, \cdot \mid \mathbf{x})$ of $\mathbf{W}$ may be estimated by

$$\widehat{\mathbb{D}}_n(\mathbf{u}, \mathbf{v} \mid \mathbf{x}) = \frac{\sum_{i=1}^n [\langle \mathbf{u}, \mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}_i) - e_n[G(\mathbf{Y}) \mid \mathbf{x}]) \rangle \langle \mathbf{v}, \mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}_i) - e_n[G(\mathbf{Y}) \mid \mathbf{x}]) \rangle] K(h_n^{-1}d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1}d(\mathbf{x}, \mathbf{X}_i))}.$$

The function $\phi(\mathbf{x}, h)$ plays a central role in determining the convergence rate and the asymptotic distribution of $V_n(\mathbf{x})$, and we discuss it in detail in the next subsection.

3.2 The small ball probability function

When the covariate $\mathbf{X}$ is finite-dimensional, say $\mathbf{X} \in \mathbb{R}^q$, and it has a continuous positive density at $\mathbf{x}$, it follows that $\phi(\mathbf{x}, h) \sim h^q$ as $h \longrightarrow 0^+$. But if $\mathbf{X}$ is a random element in an infinite-dimensional space, getting the asymptotic order of $\phi(\mathbf{x}, h)$ as $h \longrightarrow 0^+$ is much more difficult, and the available results in this area are mostly for the case where $\mathbf{X}$ is a Gaussian process (see, e.g., Lifshits 2013). In the literature, the popular approach has been to first derive the limiting behavior of $\log \phi(\mathbf{0}, h)$ as $h \longrightarrow 0^+$, when $\mathbf{X}$ is a Gaussian random element centered at $\mathbf{0}$. Then, one makes a connection between $\phi(\mathbf{x}, h)$ and $\phi(\mathbf{0}, h)$ for suitable $\mathbf{x}$ and sufficiently small h.

The asymptotic behavior of $\log \phi(\mathbf{0}, h)$ was investigated in Li (2001) for real-valued centered Gaussian Markov processes on $[0, 1]$ under the L_p-norm, where $1 \leq p \leq \infty$. It was shown there that in such a case, $h^2 \log \phi(\mathbf{0}, h) \longrightarrow -c_1$ as $h \longrightarrow 0^+$, where $c_1 > 0$ is a constant depending on p. For $\mathbf{X}$ being a fractional Brownian motion on $[0, 1]$ with Hurst index $\gamma \in (0, 1)$, it was shown in Theorem 4.6 in Li and Shao (2001) that under the L_∞-norm, $-c_2 h^{-1/\gamma} \leq \log \phi(\mathbf{0}, h) \leq -c_3 h^{-1/\gamma}$ for all $0 < h \leq 1$. Here, c_2 and c_3 are positive constants depending on γ. For $\mathbf{X}$ being an integrated fractional Brownian motion with Hurst index $\gamma \in (0, 1)$, it was established in Theorem 4.10 of Li and Shao (2001) that under the L_∞-norm, $-c_4 h^{-1/(1+\gamma)} \leq \log \phi(\mathbf{0}, h) \leq -c_5 h^{-1/(1+\gamma)}$ for all $0 < h \leq 1$, where c_4 and c_5 are positive constants depending on γ.

For the Lévy fractional Brownian motion on $[0, 1]^q$ with Hurst index $\gamma \in (0, 1)$, it was proved in Theorem 5.1 in Li and Shao (2001) that under the L_∞-norm, $-c_6 h^{-q/\gamma} \leq \log \phi(\mathbf{0}, h) \leq -c_7 h^{-q/\gamma}$ for all $0 < h \leq 1$. Here, c_6 and c_7 are positive constants depending on γ and q. For a Brownian sheet on $[0, 1]^q$, it follows from Theorem 5.3 in Li and Shao (2001) that under the L_2-norm, $-c_8 h^{-2} (\log(1/h))^{(2q-2)} \leq \log \phi(\mathbf{0}, h) \leq -c_9 h^{-2} (\log(1/h))^{(2q-2)}$ as $h \longrightarrow 0^+$, where $c_8, c_9 > 0$ are constants depending on q. It was shown in Theorem 5.4 in Li and Shao (2001) that if $\mathbf{X}$ is a Brownian sheet on $[0, 1]^2$, we have $-c_{10} h^{-2} (\log(1/h))^3 \leq \log \phi(\mathbf{0}, h) \leq -c_{11} h^{-2} (\log(1/h))^3$ under the L_∞-norm, where $c_{10}, c_{11} > 0$ are constants.

3.3 Shifted small ball probability

As we have already mentioned, the asymptotic behavior of $\log \phi(\mathbf{x}, h)$ is derived by establishing some relationship between $\phi(\mathbf{x}, h)$ and $\phi(\mathbf{0}, h)$. As described in subsection 1.2 in Mas (2012), one can establish a relation between $\phi(\mathbf{x}, h)$ and $\phi(\mathbf{0}, h)$ if the probability measure of $\mathbf{X} - \mathbf{x}$ is absolutely continuous with respect to the probability measure of $\mathbf{X}$, and the density of the measure of $\mathbf{X} - \mathbf{x}$ with respect to the measure of $\mathbf{X}$ is suitably smooth. This approach is motivated from the Cameron–Martin theorem describing the Radon–Nikodym derivative of a Weiner measure translated by $\mathbf{x}$ with respect to the centered Weiner measure, where $\mathbf{x}$ is an element of the reproducing kernel Hilbert space associated with the centered Weiner measure (see Cameron and Martin 1944). When $\mathbf{X}$ is a centered Gaussian random element in a separable Banach space, and $\mathbf{x}$ is an element of the associated reproducing kernel Hilbert space, from Theorem 3.1 in Li and Shao (2001) we get that $\exp[-(1/2)\|\mathbf{x}\|_\mu^2] \phi(\mathbf{0}, h) \leq \phi(\mathbf{x}, h) \leq \phi(\mathbf{0}, h)$

for all $h > 0$, where $\|\cdot\|_\mu$ is the norm in the reproducing kernel Hilbert space. But this result is not very useful for our purpose since the probability of the event that an infinite-dimensional Gaussian random element lies in its reproducing kernel Hilbert space is zero (see Corollary 7.1 in Lukić and Beder (2001)). Fortunately, it follows from Remark 2.2 in Dereich and Lifshits (2005) that when $\mathbf{X}$ is a centered Gaussian random element in a separable Banach space, then for almost all $\mathbf{x}$, $(\phi(\mathbf{0}, h/2))^2 \leq \phi(\mathbf{x}, h) \leq \phi(\mathbf{0}, h)$ for all sufficiently small h. How small h needs to be depends on that particular $\mathbf{x}$. On the other hand, it follows from Theorem 2.1 in Hoffmann-Jorgensen et al. (1979) that for $\mathbf{X}$ being a centered Gaussian random element in a separable infinite-dimensional Hilbert space, we have $\exp[-(1/2)\|\mathbf{x}\|^2]\phi(\mathbf{0}, h) \leq \phi(\mathbf{x}, h) \leq \phi(\mathbf{0}, h)$ for all $h > 0$.

Let $\mathbf{X}$ be a centered Gaussian random element in a separable Hilbert space. The Karhunen–Loeve expansion of $\mathbf{X}$ is $\mathbf{X} = \sum_{j=1}^{\infty} \sqrt{\lambda_j} Z_j \boldsymbol{\psi}_j$, where $\{Z_j\}$ is a collection of independent normal random variables with mean 0 and variance 1, $\{\lambda_j\}$ is the sequence of decreasing eigenvalues of the covariance operator of $\mathbf{X}$, and $\{\boldsymbol{\psi}_j\}$ is an orthonormal basis of the Hilbert space. Here, the small ball probability $\phi(\mathbf{x}, h)$ can be related to the rate of decrease of the sequence $\{\lambda_j\}$. As discussed in subsection 4.1 in Chagny and Roche (2014), for certain rates of decrease for $\{\lambda_j\}$, e.g., if for some $\alpha > 1$, $j^\alpha \lambda_j$ is bounded and bounded away from 0 for all j, we may have $c_{12} h^{p_1} \exp(-c_{13} h^{-q_1}) \leq \phi(\mathbf{x}, h) \leq c_{14} h^{p_2} \exp(-c_{15} h^{-q_1})$ for positive constants c_{12}, c_{13}, c_{14}, c_{15}, p_1, p_2 and q_1. Alternatively, for some other rates, e.g., if $j \exp[2j]\lambda_j$ is bounded and bounded away from 0 for all j, we may have $c_{16} h^{p_3} \exp[-c_{17}(\log(1/h))^{q_2}] \leq \phi(\mathbf{x}, h) \leq c_{18} h^{p_4} \exp[-c_{17}(\log(1/h))^{q_2}]$ for positive constants $c_{16}, c_{17}, c_{18}, p_3, p_4$ and $q_2 > 1$ (see subsection 4.1 in Chagny and Roche (2014)). See also Theorem 4.4, Examples 4.5, 4.6 and 4.7 in Hoffmann-Jorgensen et al. (1979) for a discussion on the relation between the small ball probability $\phi(\mathbf{x}, h)$ and the rate of decrease of $\{\lambda_j\}$.

From the discussion on the small ball probability functions above, it is now clear that in a diverse collection of cases, we have

$$C_1 h^{t_1} \exp\left[-C_2 (1/h)^{t_2} (\log(1/h))^{t_3}\right] \leq \phi(\mathbf{x}, h) \leq C_3 h^{t_4} \exp\left[-C_4 (1/h)^{t_2} (\log(1/h))^{t_3}\right] \quad (9)$$

as $h \longrightarrow 0^+$. Here, $C_1, C_2, C_3, C_4 > 0$ and $t_1, t_2, t_3, t_4 \geq 0$ are appropriate constants, all of which, except C_1, are independent of $\mathbf{x}$. C_1 may or may not depend on $\mathbf{x}$, but if it depends on $\mathbf{x}$ then $C_1 = C_1' \exp[-(1/2)\|\mathbf{x}\|^2]$ for some positive constant C_1'. For infinite-dimensional covariates, either $t_2 > 0$, or $t_3 > 1$ is an integer with $C_2 = C_4$. Define

$$m(h) = C_2 (1/h)^{t_2} (\log(1/h))^{t_3} \quad (10)$$

for $0 < h < 1$. We shall derive the optimum convergence rates of the estimates in terms of $m(h)$.

The previous discussion of small ball probabilities are concerned with only Gaussian random elements. We next consider small ball probabilities of some infinite-dimensional non-Gaussian distributions. Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be separable Banach spaces, and $f(\cdot) : \mathcal{B}_2 \longrightarrow \mathcal{B}_1$ be a function such that for any $\mathbf{u} \in \mathcal{B}_2$, there exist constants $r, s > 0$, which may depend on $\mathbf{u}$, such that for any $\mathbf{v} \in \mathcal{B}_2$ sufficiently close to $\mathbf{u}$, we

have $r\|\mathbf{v}-\mathbf{u}\| \leq \|f(\mathbf{v})-f(\mathbf{u})\| \leq s\|\mathbf{v}-\mathbf{u}\|$. Any Frechet differentiable function $f(\cdot)$ with a Frechet differentiable inverse satisfies such a condition. If $\mathbf{T}$ and $\mathbf{G}$ are random elements with $\mathbf{T} = f(\mathbf{G})$, and the small ball probability of $\mathbf{G}$ satisfies the bounds described in (9), then similar bounds also hold for $\mathbf{T}$ (see Proposition 1 in the supplement). An example of such a non-Gaussian process $\mathbf{T}$ is the geometric Brownian motion in an L_2 space, where $f(\cdot)$ is the pointwise exponential map (Øksendal 2003, p. 67).

Next, let $\mathbf{G}$ be a Gaussian process whose small ball probability satisfies the bounds in (9), and $\mathbf{T} = \mathbf{G}/\mathbf{U}$, where $\mathbf{U}$ is a bounded positive random variable independent of $\mathbf{G}$. Then, (9) will also hold for the small ball probabilities of $\mathbf{T}$ (see Proposition 2 in the supplement). Also, bounds similar to (9) can be established for the small ball probabilities of an infinite-dimensional t process, whose corresponding Gaussian process has small ball probabilities satisfying (9) (see Proposition 3 in the supplement).

The bounds in (9) were considered in Ferraty and Vieu (2006, p. 209) with $C_2 = C_4$, $t_1 = t_4 = 0$, and they called it the small ball probability function of an exponential-type process. For $t_2 = 0$, $t_3 = 1$ and appropriate values of the parameters C_1, C_2, C_3 and C_4, (9) yields the case of a finite-dimensional covariate $\mathbf{X}$ with a continuous positive density at $\mathbf{x}$, or a fractal-type process as defined in Ferraty and Vieu (2006, p. 207).

4 Convergence rate

We now derive the optimum achievable convergence rate for kernel estimates satisfying the bias-variance decomposition (3). As we shall see, the function $m(h)$ defined in (10) plays a central role in determining the convergence rate of the estimate $\widehat{\Theta}_n(\mathbf{x})$. We shall consider the covariate space to be infinite-dimensional. The case of finite-dimensional covariates is extensively discussed in the past literature (see, e.g., Stone (1980, 1982), Ibragimov and Haśminskii (1980), Yatracos (1988), Donoho and Liu (1991a,b)). In order to consider only infinite-dimensional covariates, we assume that in (9), either $t_2 > 0$, or $t_3 > 1$ with $C_2 = C_4$ in all subsequent discussions. In that case, $m(h)$ is a strictly decreasing positive function, and $m^{-1}(\cdot)$, which is the inverse function of $m(\cdot)$, is well defined. In the next theorem, we see that $\left(m^{-1}(\log n)\right)^{\beta}$ is an attainable rate of convergence of $\widehat{\Theta}_n(\mathbf{x})$. Also, under certain additional conditions, $\left(m^{-1}(\log n)\right)^{2\beta}$ is an attainable rate of convergence of the mean square error of $\widehat{\Theta}_n(\mathbf{x})$.

Theorem 5 *Suppose that in (9), we have either $t_2 > 0$, or $t_3 > 1$ with $C_2 = C_4$. Then, for any kernel $K(\cdot)$ satisfying A(i) and $\Theta(\mathbf{x})$ satisfying (3) along with conditions B(i)–B(iii), there is a sequence of bandwidths $\{h_n\}$ satisfying A(ii) such that $\left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\| = O_{\mathbb{P}}\left(\left(m^{-1}(\log n)\right)^{\beta}\right)$ as $n \longrightarrow \infty$, where $m(h)$ is as defined in (10). Further, if $\mathbb{E}[\|R_n(\mathbf{x})\|^2] = o\left(\delta_n^2\right)$ as $n \longrightarrow \infty$, where δ_n is as defined in B(iii), $\mathbb{E}\left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\|^2 = O\left(\left(m^{-1}(\log n)\right)^{2\beta}\right)$ as $n \longrightarrow \infty$ for the aforementioned sequence of bandwidths $\{h_n\}$.*

Recall that when the parameter of interest is a conditional mean type function as described in Example 1 in Sect. 2, $R_n(\mathbf{x}) = \mathbf{0}$. So, in that case the condition

$\mathbb{E}[\|R_n(\mathbf{x})\|^2] = o(\delta_n^2)$ assumed in the second part of the above theorem is trivially satisfied.

4.1 Lower bound on the convergence rate

We now proceed to investigate the lower bound of the convergence rate of $\widehat{\Theta}_n(\mathbf{x})$. In the next proposition, we establish an asymptotic lower bound of the sequence of bandwidths $\{h_n\}$ that leads to consistent kernel regression estimates. This result will be needed while deriving the lower bound of the convergence rate of a kernel estimate.

Theorem 6 *Suppose that in the upper and the lower bounds in the shifted small ball probability in (9), we have either $t_2 > 0$, or $t_3 > 1$ with $C_2 = C_4$. Then, for any sequence of bandwidths $\{h_n\}$, which satisfies Assumption A(ii), we have $h_n/m^{-1}(\log n)$ bounded away from 0 as $n \longrightarrow \infty$, where $m(h)$ is as defined in (10).*

Define

$$\tilde{B}_n(\mathbf{x}) = \mathbb{L}_{\mathbf{x}}\left(\mathbb{E}\left[(F(\mathbf{X}) - F(\mathbf{x}))\,\frac{K\left(h_n^{-1}d(\mathbf{x},\mathbf{X})\right)}{E_n^{(1)}(\mathbf{x})\phi(\mathbf{x},h_n)}\right]\right),$$

where $\mathbb{L}_{\mathbf{x}}(\cdot)$ and $F(\cdot)$ are as defined after (5), and $E_n^{(1)}(\mathbf{x})$ is as defined in (7). Also, let $\{\mathbf{e}_1, \mathbf{e}_2, \cdots\}$ be a Schauder basis of $\mathcal{B}$, such that for any $\mathbf{v} \in \mathcal{B}$, $\mathbf{v} = \sum_{n=1}^{\infty} v_n \mathbf{e}_n$ for a sequence of real numbers $\{v_n\}$. Let $\tilde{\phi}_i \in \mathcal{B}^*$ be the projection functional corresponding to $\mathbf{e}_i$, i.e., $\mathbf{v} = \sum_{i=1}^{\infty} \tilde{\phi}_i(\mathbf{v})\mathbf{e}_i$ for all $\mathbf{v} \in \mathcal{B}$. Consider the following assumptions.

C(i) There is $\Theta(\cdot) : \mathcal{C} \longrightarrow \mathcal{B}$ with the corresponding $\mathbb{L}_{\mathbf{x}}(\cdot)$ and $F(\cdot)$ such that for any sequence of bandwidths $\{h_n\}$ satisfying A(ii),

$$h_n^{-\beta}\left\|\tilde{B}_n(\mathbf{x})\right\| > b_1 > 0 \tag{11}$$

for all sufficiently large n.

C(ii) Let $G(\cdot)$ be as defined after (5). For some positive integer i_0, the conditional variance function $\mathbb{V}(\mathbf{z}) : \mathcal{C} \longrightarrow \mathbb{R}$ defined by $\mathbb{V}(\mathbf{z}) = \mathbb{E}\Big[\big(\tilde{\phi}_{i_0}\big(\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}])\big)\big)^2 \,\Big|\, \mathbf{X} = \mathbf{z}\Big]$ converges to $\mathbb{V}(\mathbf{x})$ as $\mathbf{z} \longrightarrow \mathbf{x}$, and $\mathbb{V}(\mathbf{x}) > 0$.

Condition C(ii), like condition B(iv), is related to the smoothness of the conditional distribution of the residual in the regression. In fact, condition C(ii) holds in the same location-scale type models, which we described after condition B(iv). Condition C(i) gives a lower bound on the rate of convergence of the bias part of the estimate. Inequality (6) and condition C(i) together imply that the rate of convergence of the bias part is same as h_n^{β} as $n \longrightarrow \infty$. The following two conditions are sufficient to ensure that C(i) holds.

a. There is a constant $0 < s < 1$ such that $\phi(\mathbf{x}, sh)/\phi(\mathbf{x}, h)$ is bounded away from 1 for all sufficiently small $h > 0$.

b. Let $\mathbb{L}_{\mathbf{x}}(\mathcal{F}(\mathbf{x}, \beta, \mathcal{G}))$ be the class of all functions defined by the composition $\mathbb{L}_{\mathbf{x}} \circ H$, where $H \in \mathcal{F}(\mathbf{x}, \beta, \mathcal{G})$ and $\mathcal{F}(\mathbf{x}, \beta, \mathcal{G})$ is as defined in B(i). Then $\mathbb{L}_{\mathbf{x}}(\mathcal{F}(\mathbf{x}, \beta, \mathcal{G}))$ contains the function $\mathbf{z} \mapsto d(\mathbf{x}, \mathbf{z})^{\beta}\mathbf{v}$, where $\mathbf{v} \in \mathcal{B}$, and $\mathbf{z}$ lies in a neighborhood of $\mathbf{x}$.

Condition (a) is satisfied when in (9), $t_2 > 0$, or $t_1 < t_4$, or $C_2 = C_4$. We observe that at least one of these is true in the examples that we have described in Sect. 3.2.

Now, we derive the lower bound of the order of convergence of the bias term $B_n(\mathbf{x})$ in (3) under B(i) and C(i). Note that $B_n(\mathbf{x}) = \tilde{B}_n(\mathbf{x}) + \tilde{R}_n(\mathbf{x})$, where

$$
\begin{aligned}
&\tilde{R}_n(\mathbf{x}) \\
&= \mathbb{L}_{\mathbf{x}}\left(\frac{\sum_{i=1}^{n}(F(\mathbf{X}_i) - F(\mathbf{x}))\, K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{\sum_{i=1}^{n} K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)} - \tilde{B}_n(\mathbf{x})\right) \\
&= \mathbb{L}_{\mathbf{x}}\left(\frac{\sum_{i=1}^{n}(F(\mathbf{X}_i) - F(\mathbf{x}))\, K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{\sum_{i=1}^{n} K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)} - \frac{1}{n}\sum_{i=1}^{n}(F(\mathbf{X}_i) - F(\mathbf{x}))\, \frac{K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{E_n^{(1)}(\mathbf{x})\,\phi(\mathbf{x}, h_n)}\right) \\
&\quad + \mathbb{L}_{\mathbf{x}}\left(\frac{1}{n}\sum_{i=1}^{n}(F(\mathbf{X}_i) - F(\mathbf{x}))\, \frac{K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{E_n^{(1)}(\mathbf{x})\,\phi(\mathbf{x}, h_n)} - \mathbb{E}\left[(F(\mathbf{X}) - F(\mathbf{x}))\, \frac{K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X})\right)}{E_n^{(1)}(\mathbf{x})\,\phi(\mathbf{x}, h_n)}\right]\right).
\end{aligned}
$$

It follows form condition A(i) and Markov inequality that $(n^{-1}\sum_{i=1}^{n}[E_n^{(1)}(\mathbf{x})\phi(\mathbf{x}, h_n)]^{-1} K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)) - 1) = O_{\mathbb{P}}\left(\left[n\phi(\mathbf{x}, h_n)\right]^{-1/2}\right)$ as $n \longrightarrow \infty$. Hence, from conditions A(i), A(ii) and B(i), we have

$$
\begin{aligned}
&\mathbb{L}_{\mathbf{x}}\left(\frac{\sum_{i=1}^{n}(F(\mathbf{X}_i) - F(\mathbf{x}))\, K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{\sum_{i=1}^{n} K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)} - \frac{1}{n}\sum_{i=1}^{n}(F(\mathbf{X}_i) - F(\mathbf{x}))\, \frac{K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{E_n^{(1)}(\mathbf{x})\,\phi(\mathbf{x}, h_n)}\right) \\
&= o_{\mathbb{P}}\left(h_n^{\beta}\right)
\end{aligned}
$$

as $n \longrightarrow \infty$. Also, from Assumptions A(i), A(ii), B(i) and Markov inequality, it follows that

$$
\begin{aligned}
&\mathbb{L}_{\mathbf{x}}\left(\frac{1}{n}\sum_{i=1}^{n}(F(\mathbf{X}_i) - F(\mathbf{x}))\, \frac{K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{E_n^{(1)}(\mathbf{x})\,\phi(\mathbf{x}, h_n)} - \mathbb{E}\left[(F(\mathbf{X}) - F(\mathbf{x}))\, \frac{K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X})\right)}{E_n^{(1)}(\mathbf{x})\,\phi(\mathbf{x}, h_n)}\right]\right) \\
&= o_{\mathbb{P}}\left(h_n^{\beta}\right)
\end{aligned}
$$

as $n \longrightarrow \infty$. Hence,

$$
\tilde{R}_n(\mathbf{x}) = o_{\mathbb{P}}(h_n^{\beta}) \text{ as } n \longrightarrow \infty. \tag{12}
$$

Note that inequality (11) provides a lower bound of the convergence rate for the bias term $B_n(\mathbf{x})$ in view of (12), and this will be used to determine a lower bound of the rate of convergence of $\widehat{\Theta}_n(\mathbf{x})$. Also note that $\tilde{\phi}_{i_0}\left(\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}))\right)$ in condition C(ii) is a real-valued random variable. So, the convergence condition in C(ii) of the conditional variance may be viewed as a special case of condition B(iv). We now state the theorem on the lower bound of the convergence rate of $\left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\|$.

Theorem 7 *Suppose that in (9), we have either $t_2 > 0$, or $t_3 > 1$ with $C_2 = C_4$, the kernel $K(\cdot)$ satisfies A(i), the sequence of bandwidths $\{h_n\}$ satisfies A(ii), and the decomposition (3) along with conditions B(i)–B(iii), C(i) and C(ii) hold. Then,*

$$\liminf_{n\longrightarrow\infty} \mathbb{P}\left[\left(m^{-1}(\log n)\right)^{-\beta} \left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\| > c\right] > 0$$

for some constant $c > 0$ depending on $\Theta(\mathbf{x})$, where $m(h)$ is as defined in (10).

Theorem 7 implies that we cannot get a faster rate of convergence than $\left(m^{-1}(\log n)\right)^{\beta}$, since $\left(m^{-1}(\log n)\right)^{-\beta} \left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\|$ does not converge to 0 *in probability* as $n \longrightarrow \infty$. Further, from Theorem 7 it follows that $\left(m^{-1}(\log n)\right)^{2\beta}$ is a lower bound for the rate of convergence of the mean square error $\mathbb{E}\left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\|^2$. Hence, combining Theorems 5 and 7, we get that $\left(m^{-1}(\log n)\right)^{\beta}$ and $\left(m^{-1}(\log n)\right)^{2\beta}$ are the optimum rates of convergence of $\widehat{\Theta}_n(\mathbf{x})$ and its mean square error, respectively, when all the conditions of the two theorems are satisfied. We now deduce simplified expressions of the optimum rates for the specific infinite-dimensional covariate distributions considered in Sect. 3.2.

For $\mathbf{X}$ being a real-valued continuous Gaussian Markov process on [0, 1], under the L_p-norm, we have $\left(m^{-1}(\log n)\right)^{\beta} = O((\log n)^{-\beta/2})$ as $n \longrightarrow \infty$. For fractional Brownian motion on [0, 1] with Hurst index $\gamma \in (0, 1)$, under the L_∞-norm, we have $t_2 = 1/\gamma$, and consequently $\left(m^{-1}(\log n)\right)^{\beta} = O((\log n)^{-\gamma\beta})$ as $n \longrightarrow \infty$. On the other hand, for an integrated fractional Brownian motion with Hurst index γ and under the L_∞-norm, we have $t_2 = 1/(1+\gamma)$ and $\left(m^{-1}(\log n)\right)^{\beta} = O((\log n)^{-(1+\gamma)\beta})$ as $n \longrightarrow \infty$. When $\mathbf{X}$ is a Lévy fractional Brownian motion on $[0, 1]^q$ with Hurst index γ, $t_2 = q/\gamma$ and $\left(m^{-1}(\log n)\right)^{\beta} = O((\log n)^{-\gamma\beta/q})$ as $n \longrightarrow \infty$.

In the class of processes $H_{X,L}$ considered in subsection 4.1 of Chagny and Roche (2014), $t_2 > 0$ and $t_3 = 0$, and we have $\left(m^{-1}(\log n)\right)^{2\beta} = O((\log n)^{-2\beta/t_2})$ as $n \longrightarrow \infty$. On the other hand, for the class of processes $H_{X,M}$ considered by these authors, we have $\left(m^{-1}(\log n)\right)^{2\beta} = O\left(\exp[-2\beta C_2^{-1/t_3}(\log n)^{1/t_3}]\right)$ as $n \longrightarrow \infty$. Note that these rates coincide with the optimal rates of convergence of the mean square error described in Chagny and Roche (2014, Table 1, p. 2363), which were derived when the response is real-valued and the parameter of interest is the conditional distribution function of the response. We have covered this particular case of the parameter of interest in Example 1.

4.2 Asymptotic dominance of bias over variance

Recall that in the case of finite-dimensional covariates, the bias and the variance terms in nonparametric regression have the same rate of convergence (see, e.g., Stone (1980), Ibragimov and Haśminskii (1980)). In fact, Mas (2012) chose the bandwidth of the kernel estimate by balancing the asymptotic orders of the bias and the variance (see Lemma 1 and the preceding discussion in Mas (2012)) even when the covariate is infinite-dimensional. However, as we shall show now, the optimum choice of the

bandwidth in a kernel estimate, as described in the proof of Theorem 5, leads to different asymptotic orders of the bias and the variance when the covariate is infinite-dimensional in nature, i.e., when we have either $t_2 > 0$ or $t_3 > 1$ with $C_2 = C_4$ in (9).

Theorem 8 *Suppose that either $t_2 > 0$ or $t_3 > 1$ with $C_2 = C_4$ in the bounds in (9). Also, let the kernel $K(\cdot)$ satisfy A(i), and the decomposition (3) along with conditions B(i)–B(iii) hold. Then, for any $\Theta(\mathbf{x})$ satisfying C(i), the ratio $\|V_n(\mathbf{x})\|/\|B_n(\mathbf{x})\| = o_{\mathbb{P}}(1)$ as $n \longrightarrow \infty$ for the optimum choice of bandwidth $\{h_n\}$ described in the proof of Theorem 5.*

Theorem 8 illustrates that our optimum bandwidth, which minimizes (25) in the proof of Theorem 5, does not balance the convergence rates of the variance and the bias in kernel regression if the covariate is infinite-dimensional. Instead, the ratio of the variance to the bias for our optimal choice of bandwidth tends to zero as the sample size increases. This phenomenon is due to the exponential decay of the small ball probability function in infinite-dimensional spaces. When the covariate is infinite-dimensional, we may have very small number of observations in a neighborhood in the covariate space. To cope with this problem, one has to use relatively larger bandwidths than what is required for finite-dimensional covariates. This results in an 'over-smoothed' estimate with its bias asymptotically larger than its variance. It will be appropriate to note here that the optimum convergence rate derived in Theorems 5 and 7 is same as the one derived in Mas (2012) for estimation of the conditional mean of a real-valued response, where the chosen bandwidth balances the bias and the variance (Mas 2012, p. 1760). However, our optimum bandwidth, which does not try to balance the bias and the variance in the decomposition (3), will often lead to an estimate with higher statistical precision compared to an estimate based on a bandwidth that balances the bias and the variance. In several cases, the statistical error will be substantially lower when our optimum bandwidth is used as demonstrated in Theorem 9.

Theorem 9 *Suppose Assumptions A(i), A(ii), B(i)–B(iii) and C(ii) are satisfied. Let $\widehat{\Theta}_n^{(b)}(\mathbf{x})$ be an estimate of $\Theta(\mathbf{x})$ constructed using bandwidth $h_n^{(b)}$, which satisfies A(ii) and balances the bias and the variance so that $(h_n^{(b)})^{2\beta} n\phi(\mathbf{x}, h_n^{(b)})$ is bounded and bounded away from 0 as $n \longrightarrow \infty$. Also, let $\widehat{\Theta}_n^{(op)}(\mathbf{x})$ be an estimate of $\Theta(\mathbf{x})$ constructed using our optimum bandwidth minimizing (25) in the proof of Theorem 5. Assume that $t_2 > 0$ in the bounds in (9). Then, for any $\beta_1 > \beta$ and any $\Theta(\cdot)$ for which the corresponding $F(\cdot) \in \mathcal{F}(\mathbf{x}, \beta_1, \mathcal{G}) \subseteq \mathcal{F}(\mathbf{x}, \beta, \mathcal{G})$,*

$$\frac{\left\|\widehat{\Theta}_n^{(op)}(\mathbf{x}) - \Theta(\mathbf{x})\right\|}{\left\|\widehat{\Theta}_n^{(b)}(\mathbf{x}) - \Theta(\mathbf{x})\right\|} = o_{\mathbb{P}}(1) \quad \text{as } n \longrightarrow \infty.$$

Further, if $\mathbb{E}[\|R_n(\mathbf{x})\|^2] = o\left(\delta_n^2\right)$ as $n \longrightarrow \infty$, where δ_n is as defined in B(iii), then

$$\frac{\mathbb{E}\left\|\widehat{\Theta}_n^{(op)}(\mathbf{x}) - \Theta(\mathbf{x})\right\|^2}{\mathbb{E}\left\|\widehat{\Theta}_n^{(b)}(\mathbf{x}) - \Theta(\mathbf{x})\right\|^2} = o(1) \quad \text{as } n \longrightarrow \infty.$$

Recall that for conditional mean type functions described in Example 1 in Sect. 2, $R_n(\mathbf{x}) = \mathbf{0}$, and the condition $\mathbb{E}[\|R_n(\mathbf{x})\|^2] = o(\delta_n^2)$ assumed in the second part of the above theorem is trivially satisfied.

5 Adaptive selection of bandwidths

In practice, one has to choose the bandwidth h by some data-driven adaptive procedure. Such adaptive choice of bandwidth, when the covariate is functional, has been investigated in Chagny and Roche (2014, 2016) for the kernel estimates of the conditional distribution and the conditional mean of a real-valued response. Their data-based bandwidth selection procedure can be suitably extended for more general regression problems considered in this paper.

Let $\mathbb{H}_n$ be a finite collection of bandwidths with cardinality less than or equal to n such that for any $h \in \mathbb{H}_n$, $\phi(\mathbf{x}, h) \leq 2(\log n)^{-1}$ and $\phi(\mathbf{x}, h) \geq n^{-1}(\log n)^2$. Since $\phi(\mathbf{x}, h)$ is a monotone increasing function of h, if a sequence of bandwidths $\{h_n\}$ is such that $h_n \in \mathbb{H}_n$ for all n, then $\{h_n\}$ satisfies condition A(ii). In this section, we shall write $\widehat{\Theta}_n(\mathbf{x}, h)$, $B_n(\mathbf{x}, h)$, $V_n(\mathbf{x}, h)$ and $R_n(\mathbf{x}, h)$ for $\widehat{\Theta}_n(\mathbf{x})$, $B_n(\mathbf{x})$, $V_n(\mathbf{x})$ and $R_n(\mathbf{x})$, respectively, to indicate the dependence of $\widehat{\Theta}_n(\mathbf{x})$, $B_n(\mathbf{x})$, $V_n(\mathbf{x})$ and $R_n(\mathbf{x})$ on the bandwidth h. We assume the following:

D(i) There is a constant $\sigma > 0$ such that for any $\mathbf{z}$ in a certain neighborhood of $\mathbf{x}$ and every integer $k \geq 2$,

$$\mathbb{E}\left[\|\mathbb{L}_{\mathbf{x}}\left(G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}]\right)\|^k \,\middle|\, \mathbf{X} = \mathbf{z}\right] \leq \frac{k!}{2}\sigma^k.$$

D(ii) There are constants $\epsilon_1 > 0$, $\epsilon_2 > 0$ and $M > 0$ such that whenever $h \leq \epsilon_1$ and $\|V_n(\mathbf{x}, h)\| \leq \epsilon_2$, we have $\|R_n(\mathbf{x}, h)\|^2 \leq Mh^{2\beta} + \|V_n(\mathbf{x}, h)\|^2$.

Condition D(i) is similar to assumption (H_ϵ) used in Chagny and Roche (2016, p. 108), which was used to derive the convergence rate of the adaptive estimate of the conditional mean for a real-valued response. Condition D(ii) describes a bound on the remainder of our bias-variance type decomposition in terms of the bound on the bias part and the variance part. Condition D(ii) is trivially satisfied for the conditional mean type estimates described in Example 1 in Sect. 2. It is also satisfied in the class of regression problems described in Example 2 in Sect. 2. Define the empirical shifted small ball probability $\widehat{\phi}(\mathbf{x}, h) = (1/n)\sum_{i=1}^n \mathbb{I}(d(\mathbf{x}, \mathbf{X}_i) \leq h)$. Define

$$D_n(\mathbf{x}, h) = \sigma^2\zeta_n \frac{\log n}{n\hat{\phi}(\mathbf{x}, h)}\mathbb{I}\left(\hat{\phi}(\mathbf{x}, h) > 0\right) + \sigma^2\zeta_n n\mathbb{I}\left(\hat{\phi}(\mathbf{x}, h) = 0\right),$$

where $\{\zeta_n\}$ is a sequence of positive constants independent of h, such that $\zeta_n \longrightarrow \zeta_0 > 0$ as $n \longrightarrow \infty$. The constant ζ_0 is described in the proof of Lemma 11 in the supplement. Also define

$$C_n(\mathbf{x}, h) = \max_{h' \in \mathbb{H}_n}\left(\left\|\widehat{\Theta}_n\left(\mathbf{x}, h'\right) - \widehat{\Theta}_n\left(\mathbf{x}, \max\{h, h'\}\right)\right\|^2 - D_n\left(\mathbf{x}, h'\right)\right)_+.$$

$D_n(\mathbf{x}, h)$ approximates the upper bound of the variance term, and $C_n(\mathbf{x}, h)$ approximates the bias term. The data-driven choice of bandwidth is defined as

$$h_n^* = \arg\min_{h \in \mathbb{H}_n} \left[C_n(\mathbf{x}, h) + D_n(\mathbf{x}, h) \right].$$

The following theorem gives an upper bound on the convergence rate of the adaptive estimate $\widehat{\Theta}_n(\mathbf{x}, h_n^*)$.

Theorem 10 *Define*

$$\lambda_n = \min_{h \in \mathbb{H}_n} \left[h^{2\beta} + \frac{\log n}{n\phi(\mathbf{x}, h)} \right].$$

Let conditions A(i), B(i), D(i) and D(ii) be satisfied. Then,

$$\left\| \widehat{\Theta}_n\left(\mathbf{x}, h_n^*\right) - \Theta(\mathbf{x}) \right\|^2 = O_{\mathbb{P}}(\lambda_n) \quad as\ n \longrightarrow \infty. \tag{13}$$

Further, for the conditional mean type functions described in Example 1 in Sect. 2, we have

$$\mathbb{E} \left\| \widehat{\Theta}_n(\mathbf{x}, h_n^*) - \Theta(\mathbf{x}) \right\|^2 = O(\lambda_n) \quad as\ n \longrightarrow \infty. \tag{14}$$

Equation (13) gives an upper bound for the asymptotic convergence rate of the adaptive estimate. In Chagny and Roche (2016), the adaptive estimate and its convergence rate were derived for the estimation of the conditional mean of a real-valued response in a homoscedastic model. Our setup includes heteroscedastic regression models where the parameter to be estimated is an element of a type 2 Banach space, and it is not necessarily the conditional mean.

5.1 Numerical demonstration

We next demonstrate the adaptive estimate $\widehat{\Theta}_n(\mathbf{x}, h_n^*)$ in several regression models. In all the examples, we consider the covariate $\mathbf{X}$ to be a random element in $L_2[0, 1]$. The usual norm in $L_2[0, 1]$ is denoted as $\|\cdot\|_2$. We denote the adaptive choice of the bandwidth as h_n^*. We take $\zeta_n = \min\{\sqrt{n}, 1500\}$ in our computation, the validity of which is ensured from (73) in the supplement. We substitute $\phi(\mathbf{x}, h)$ by $\hat{\phi}(\mathbf{x}, h)$ in the construction of the collection of bandwidths $\mathbb{H}_n$, as done in Chagny and Roche (2016). The parameter σ^2 used to define $D_n(\mathbf{x}, h)$ and mentioned in D(i) also needs to be estimated. This is done based on the regression model. Note that for σ^2, which satisfies

$$\sigma^2 \geq \|\mathbb{L}_{\mathbf{x}}\|^2 \sup_{\mathbf{z}} \mathbb{E}\left[\|G(\mathbf{Y}) - \mathbb{E}[G(\mathbf{Y}) \mid \mathbf{X} = \mathbf{z}]\|^2 \,\middle|\, \mathbf{X} = \mathbf{z} \right] \tag{15}$$

for $\mathbf{z}$ lying in some neighborhood of $\mathbf{x}$, condition D(i) will hold. Since by construction $\max \mathbb{H}_n \longrightarrow 0$ as $n \longrightarrow \infty$, and the kernel $K(u) = I(0 \leq u \leq 1)$ satisfies

condition A(i), it is enough to consider the supremum in (15) over the $\mathbf{X}_i$s such that $d(\mathbf{x}, \mathbf{X}_i) \leq \max \mathbb{H}_n$ for estimating σ^2. So, if $\hat{\sigma}_1^2$ is an estimated upper bound of $\|\mathbb{L}_{\mathbf{x}}\|^2$, and $\hat{\sigma}_2^2(\mathbf{X}_i)$ is an estimated upper bound of $\mathbb{E}\left[\|G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i) \mid \mathbf{X}_i]\|^2 \,\middle|\, \mathbf{X}_i\right]$, then we can take

$$\hat{\sigma}^2 = \hat{\sigma}_1^2 \max\left\{\hat{\sigma}_2^2(\mathbf{X}_i) \mid d(\mathbf{x}, \mathbf{X}_i) \leq \max \mathbb{H}_n\right\}$$

as an estimate of σ^2. Let $h_{n,1} = \min \mathbb{H}_n$ and $h_{n,2} = \max \mathbb{H}_n$. Denote

$$W_{i,n}^{(1)}(\mathbf{z}) = \frac{K\left({h_{n,1}}^{-1} d\left(\mathbf{z}, \mathbf{X}_i\right)\right)}{\sum_{i=1}^n K\left({h_{n,1}}^{-1} d\left(\mathbf{z}, \mathbf{X}_i\right)\right)}.$$

In the case of the mean regression model as described in Example 1, an estimate of σ^2 is

$$\hat{\sigma}^2 = \max\left\{\sum_{j=1}^n \left\|\Psi(\mathbf{Y}_j) - \left(\sum_{k=1}^n \Psi(\mathbf{Y}_k) W_{k,n}^{(1)}(\mathbf{X}_j)\right)\right\|^2 W_{j,n}^{(1)}(\mathbf{X}_i) \,\middle|\, d(\mathbf{x}, \mathbf{X}_i) \leq h_{n,2}\right\}.$$

The function $\Psi(\cdot)$ is as described in Example 1. The rationale for using the weights $W_{i,n}^{(1)}(\mathbf{z})$ is the same as that described in subsection 4.1.2 in Chagny and Roche (2016). In case the parameter to be estimated is a function of the conditional mean as discussed in Example 2, or in the case of a maximum likelihood regression model as described in Example 3, we need to additionally estimate an upper bound of the term $\|\mathbb{L}_{\mathbf{x}}\|^2$ in (15). For a function of conditional mean type estimate, we have seen in Theorem 1 that $\mathbb{L}_{\mathbf{x}}(\cdot) = \Gamma'\left(\mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]\right)(\cdot)$, where $\Gamma(\cdot)$ and $\Psi(\cdot)$ are as described in Example 2. Hence, we can take $\hat{\sigma}_1^2 = \left\|\Gamma'\left(\sum_{i=1}^n \Psi(\mathbf{Y}_i) W_{i,n}^{(1)}(\mathbf{x})\right)\right\|^2$, and

$$\hat{\sigma}^2 = \hat{\sigma}_1^2 \max\left\{\sum_{j=1}^n \left\|\Psi(\mathbf{Y}_j) - \left(\sum_{k=1}^n \Psi(\mathbf{Y}_k) W_{k,n}^{(1)}(\mathbf{X}_j)\right)\right\|^2 W_{j,n}^{(1)}(\mathbf{X}_i) \,\middle|\, d(\mathbf{x}, \mathbf{X}_i) \leq h_{n,2}\right\}.$$

Here, the function $\Psi(\cdot)$ is as described in Example 2. In a maximum likelihood regression model (Example 3), we have seen from Theorem 2 that $\mathbb{L}_{\mathbf{x}}(\cdot) = [\mathbf{I}(\Theta(\mathbf{x}))]^{-1}(\cdot)$. So, we need to estimate $\mathbf{I}(\Theta(\mathbf{x})) = -\mathbb{E}[\Delta_2(g(\mathbf{Y} \mid \Theta(\mathbf{x}))) \mid \mathbf{X} = \mathbf{x}]$, which we estimate by $\hat{\mathbf{I}} = -\sum_{i=1}^n \Delta_2(g(\mathbf{Y}_i \mid \widehat{\Theta}_n^{(1)}(\mathbf{x}))) W_{i,n}^{(1)}(\mathbf{x})$, where $\widehat{\Theta}_n^{(1)}(\mathbf{x})$ is defined as the solution of the likelihood equation (2) with the bandwidth being $h_{n,1}$. So, we can take $\hat{\sigma}_1^2 = \left\|\hat{\mathbf{I}}\right\|^{-2}$. In this case, $G(\mathbf{Y}) = \nabla g(\mathbf{Y} \mid \Theta(\mathbf{X}))$, so that $\mathbb{E}[G(\mathbf{Y}_i) \mid \mathbf{X}_i] = \mathbf{0}$ for all i. Since $\Theta(\mathbf{X}_i)$ is unknown, we estimate $G(\mathbf{Y}_i)$ by $\nabla g(\mathbf{Y}_i \mid \widehat{\Theta}_n^{(1)}(\mathbf{X}_i))$, where $\widehat{\Theta}_n^{(1)}(\mathbf{X}_i)$ is the solution of the likelihood equation (2) with $\mathbf{x}$ replaced by $\mathbf{X}_i$ and the bandwidth being $h_{n,1}$. Consequently, here we have

$$\hat{\sigma}^2 = \hat{\sigma}_1^2 \max\left\{\sum_{j=1}^n \left\|\nabla g(\mathbf{Y}_j \mid \widehat{\Theta}_n^{(1)}(\mathbf{X}_j))\right\|^2 W_{j,n}^{(1)}(\mathbf{X}_i) \,\middle|\, d(\mathbf{x}, \mathbf{X}_i) \leq h_{n,2}\right\}.$$

As our first example, we consider $\mathbf{Y}$ following a normal distribution with mean zero and variance $\|\mathbf{X}\|_2^2$. We consider two distributions for $\mathbf{X}$, namely the standard Brownian motion and the fractional Brownian motion with Hurst index 0.8. We want to estimate the conditional variance $\mathbb{V}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$, which we do in two ways. In the first case, we estimate $\mathbb{V}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ by

$$\widehat{\mathbb{V}}_n^{(1)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}] = \sum_{i=1}^{n} \left[\mathbf{Y}_i - \left(\sum_{i=1}^{n} \mathbf{Y}_i W_{i,n}(\mathbf{x}) \right) \right]^2 W_{i,n}(\mathbf{x}),$$

where

$$W_{i,n}(\mathbf{x}) = \frac{K\left((h_n^*)^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{\sum_{i=1}^{n} K\left((h_n^*)^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}.$$

So, this estimate belongs to the class of estimates described in Example 2. In the second case, we estimate $\mathbb{V}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ using the weighted maximum likelihood procedure described in Example 3, with the conditional density of $\mathbf{Y}$ given $\mathbf{X}$ being the density of the normal random variable with mean zero and variance $\|\mathbf{X}\|_2^2$. In this case, our estimate turns out to be

$$\widehat{\mathbb{V}}_n^{(2)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}] = \sum_{i=1}^{n} \mathbf{Y}_i^2 W_{i,n}(\mathbf{x}).$$

We randomly generate 100 values of $\mathbf{x}$ from the distribution of $\mathbf{X}$ and compute the estimates $\widehat{\mathbb{V}}_n^{(1)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ and $\widehat{\mathbb{V}}_n^{(2)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ for each of them. We plot the estimates $\widehat{\mathbb{V}}_n^{(1)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ and $\widehat{\mathbb{V}}_n^{(2)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ along with the actual $\mathbb{V}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ values for different sample sizes against the values of $\|\mathbf{x}\|$ in Fig. 1, where $\mathbf{X}$ is a standard Brownian motion. When $\mathbf{X}$ is a fractional Brownian motion with Hurst index 0.8, we plot the estimates along with the actual values in Fig. 2. from the plots, we observe that the two estimates $\widehat{\mathbb{V}}_n^{(1)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ and $\widehat{\mathbb{V}}_n^{(2)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ have no noticeable differences. Also, there appears to be some underestimation when the value of $\mathbb{V}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ is high. This is due to the fact that the $\mathbb{V}[\mathbf{Y} \mid \mathbf{X} = \mathbf{X}_i]$ values for $\mathbf{X}_i$ lying in a neighborhood of $\mathbf{x}$ tend to be smaller than $\mathbb{V}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$, and the kernel estimate is based on those $\mathbf{X}_i$ and their corresponding $\mathbf{Y}_i$ values. We also observe that the deviations of the estimated values from the actual values are marginally less when the covariate is a fractional Brownian motion with Hurst index 0.8, compared to the case where the covariate is a standard Brownian motion. This may be due to the fact that the distribution of $\|\mathbf{X}\|_2$ is more concentrated at lower values when $\mathbf{X}$ is a fractional Brownian motion with Hurst index 0.8 compared to the distribution of the same when $\mathbf{X}$ is a standard Brownian motion.

In the second example, we take $\mathbf{Y}$ to be a Bernoulli random variable, with $\mathbb{P}[\mathbf{Y} = 1 \mid \mathbf{X} = \mathbf{x}] = 1 - \mathbb{P}[\mathbf{Y} = 0 \mid \mathbf{X} = \mathbf{x}] = 1/(1 + \|\mathbf{X}\|_2)$. Here, our parameter of interest is $\mathbb{P}[\mathbf{Y} = 1 \mid \mathbf{X} = \mathbf{x}]$, and we estimate it using the weighted maximum likelihood procedure described in Example 3, while employing the adaptive choice

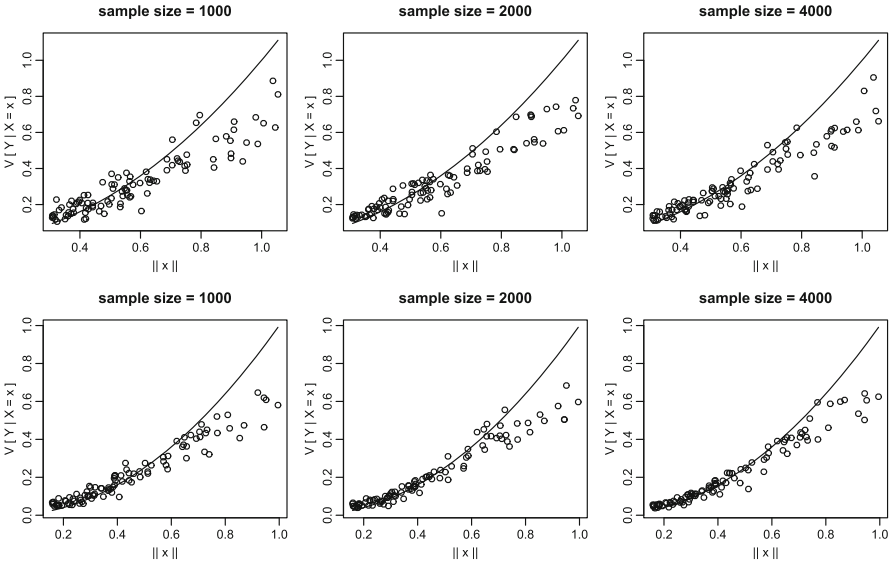

Fig. 1 Plots of actual conditional variance $\mathbb{V}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ (line) and its estimates $\widehat{\mathbb{V}}_n^{(1)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ (points, first row) and $\widehat{\mathbb{V}}_n^{(2)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ (points, second row) for different sample sizes. The covariate is a standard Brownian motion

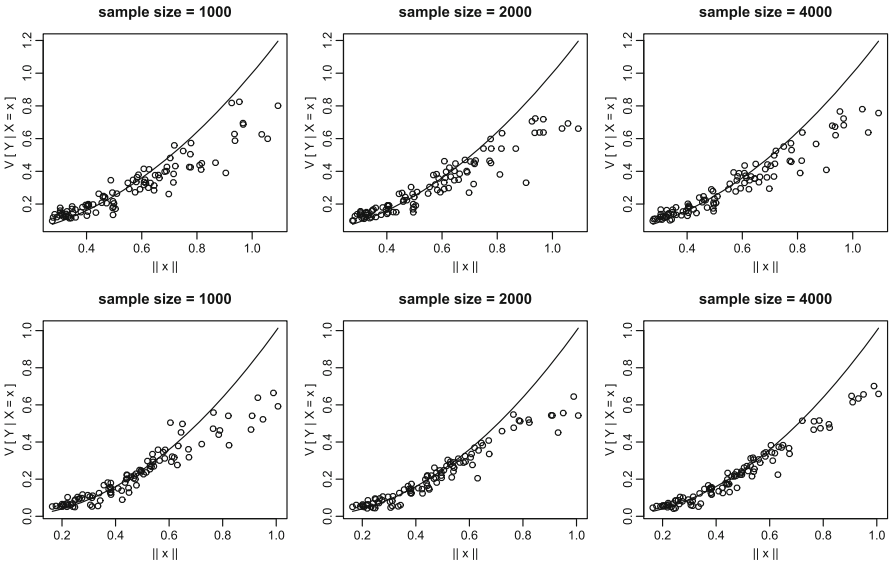

Fig. 2 Plots of actual conditional variance $\mathbb{V}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ (line) and its estimates $\widehat{\mathbb{V}}_n^{(1)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ (points, first row) and $\widehat{\mathbb{V}}_n^{(2)}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ (points, second row) for different sample sizes. The covariate is a fractional Brownian motion with Hurst index 0.8

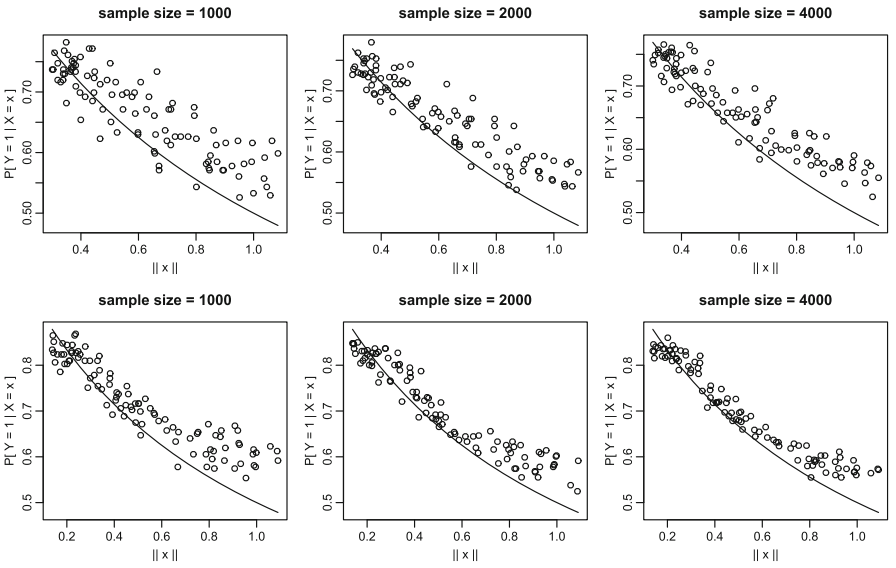

Fig. 3 Plots of estimated probability $\widehat{\mathbb{P}}_n[\mathbf{Y} = 1 \,|\, \mathbf{X} = \mathbf{x}]$ (points) and actual probability $\mathbb{P}[\mathbf{Y} = 1 \,|\, \mathbf{X} = \mathbf{x}]$ (line). The covariate is the standard Brownian motion in the first row, and the fractional Brownian motion with Hurst index 0.8 in the second row

of the bandwidth. We again consider two distributions of $\mathbf{X}$, namely the standard Brownian motion and the fractional Brownian motion with Hurst index 0.8, randomly generate 100 values of $\mathbf{x}$ from the distribution of $\mathbf{X}$ and plot the estimated values and the actual probabilities against the values of $\|\mathbf{x}\|$ in Fig. 3 for three different sample sizes. The improvement in accuracy of the estimate over the sample sizes is noticeable. We also observe that there appears to be some overestimation for small values of $\mathbb{P}[\mathbf{Y} = 1 \,|\, \mathbf{X} = \mathbf{x}]$, which is due to the fact that values of $\mathbb{P}[\mathbf{Y} = 1 \,|\, \mathbf{X} = \mathbf{X}_i]$ for $\mathbf{X}_i$ lying in a neighborhood of $\mathbf{x}$ tend to be larger than $\mathbb{P}[\mathbf{Y} = 1 \,|\, \mathbf{X} = \mathbf{x}]$ in such a case. Further, like in the first example, we observe that the deviations of the estimated values from the actual values are less when the covariate is a fractional Brownian motion with Hurst index 0.8, compared to the case where the covariate is a standard Brownian motion.

In the third example, we consider a functional response $\mathbf{Y}$, defined by $\mathbf{Y}(t) = \int_0^t \mathbf{X}(t)\mathrm{d}t + \mathbf{E}(t)$, where $\mathbf{E}(\cdot)$ is a Brownian motion independent of $\mathbf{X}(\cdot)$ with the covariance operator $\mathbb{COV}(\mathbf{E}(s), \mathbf{E}(t)) = 0.25 \min\{s, t\}$. We want to estimate the conditional mean curve $\mathbb{E}[\mathbf{Y} \,|\, \mathbf{X} = \mathbf{x}]$, for some fixed value $\mathbf{x}$ of the covariate. We again consider two distributions of $\mathbf{X}$, namely the standard Brownian motion and the fractional Brownian motion with Hurst index 0.8. In each case, we generate 3 random curves as values of $\mathbf{x}$ and plot the adaptive estimates of the corresponding conditional means for different sample sizes in Fig. 4. In the first column, we have plotted the curves chosen as values of $\mathbf{x}$. The first three rows in Fig. 4 present the estimated conditional mean curves and the actual conditional mean curves for different sample sizes corresponding to the respective values of $\mathbf{x}$ in the particular rows when the covariate is a standard Brownian motion. The last three rows in Fig. 4 present the

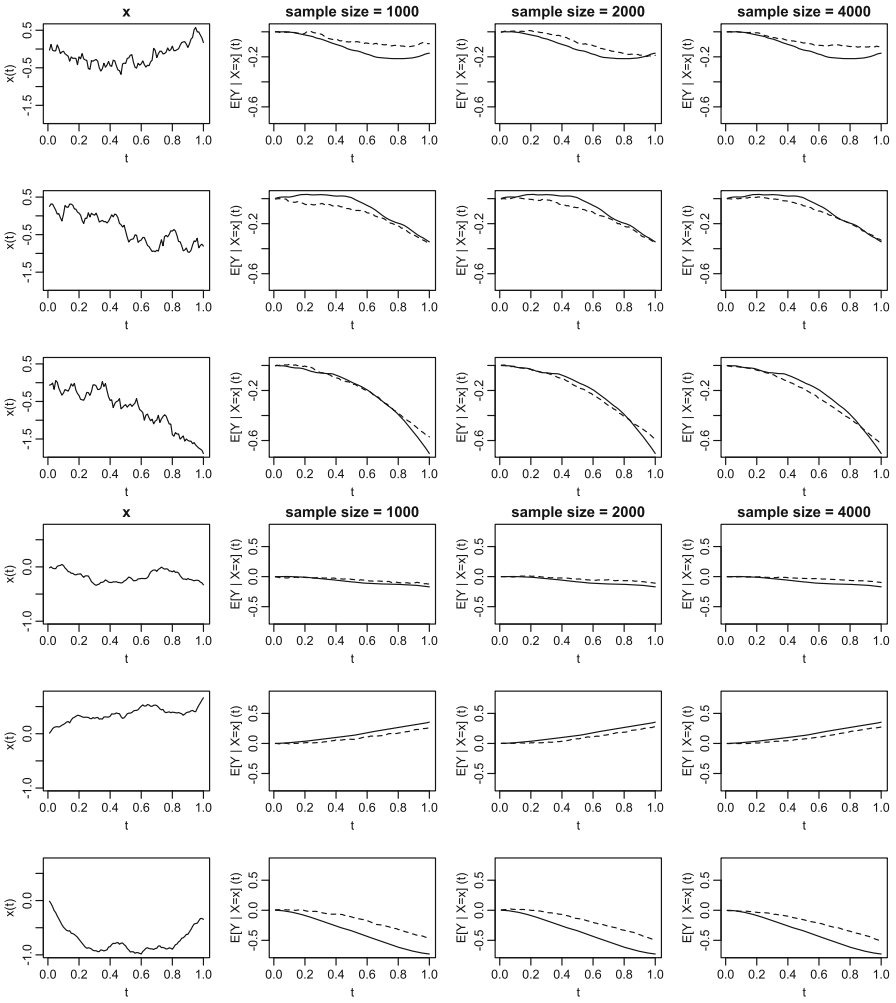

Fig. 4 Plots of estimated conditional mean curves $\widehat{\mathbb{E}}_n[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ (dashed line) and actual conditional mean curves $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ (solid line). The covariate is a standard Brownian motion in the first three rows, and a fractional Brownian motion with Hurst index 0.8 in the last three rows

estimated conditional mean curves and the actual conditional mean curves for different sample sizes when the covariate is a fractional Brownian motion with Hurst index 0.8. We observe that in all the cases, the estimates follow the actual curves closely.

5.2 Demonstration in real data

We now demonstrate the adaptive estimates of several regression parameters in the Tecator data. The Tecator data is a popular dataset available in the R package 'caret.' This dataset contains the percentage values of moisture, fat and protein contents of 215 meat samples along with their absorbance spectra in the wavelength range 850–

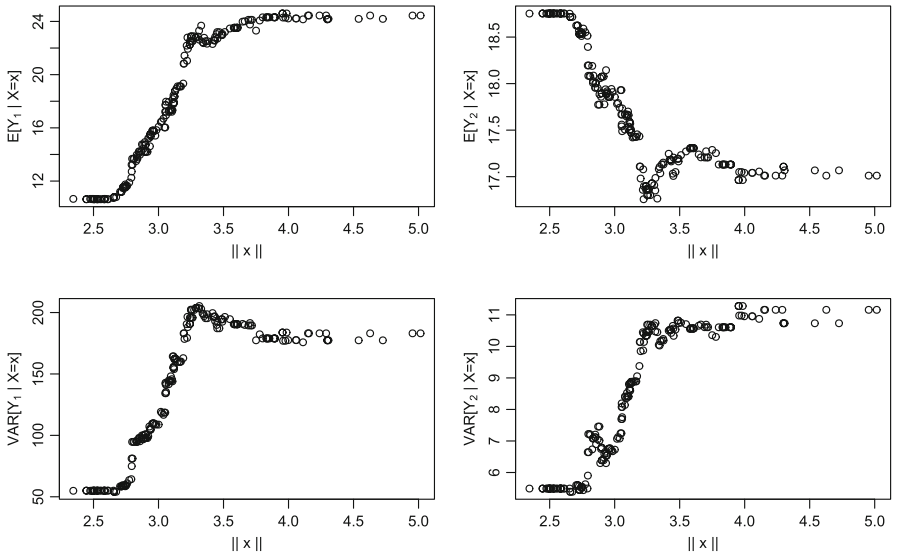

Fig. 5 Plots of adaptive estimates of $\mathbb{E}[\mathbf{Y}_1 \mid \mathbf{X} = \mathbf{x}]$, $\mathbb{E}[\mathbf{Y}_2 \mid \mathbf{X} = \mathbf{x}]$, $\mathbb{VAR}[\mathbf{Y}_1 \mid \mathbf{X} = \mathbf{x}]$ and $\mathbb{VAR}[\mathbf{Y}_2 \mid \mathbf{X} = \mathbf{x}]$ against the L_2-norm of $\mathbf{x}$ in the Tecator data

1050 nm measured by a Tecator spectroscope. The chemical contents of the meat samples are measured by analytical chemistry, which is expensive. The spectra of the samples are measured using a Tecator spectroscope, which is relatively inexpensive compared to the analytical chemistry method. So, it is economically beneficial to be able to predict the chemical contents of a sample from its spectra. Hence, we consider the fat and the protein content values as the response and the curve of the absorbance spectra as the covariate. We denote the percentage values of the fat and the protein contents as $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively, and curve of the absorbance spectra as $\mathbf{X}$. So, the covariate $\mathbf{X}$ is a random function here, which we consider as a random element in the L_2 space. We consider 5 regression parameters of interest, namely $\mathbb{E}[\mathbf{Y}_1 \mid \mathbf{X} = \mathbf{x}]$, $\mathbb{E}[\mathbf{Y}_2 \mid \mathbf{X} = \mathbf{x}]$, $\mathbb{VAR}[\mathbf{Y}_1 \mid \mathbf{X} = \mathbf{x}]$, $\mathbb{VAR}[\mathbf{Y}_2 \mid \mathbf{X} = \mathbf{x}]$ and $\mathbb{COR}[\mathbf{Y}_1, \mathbf{Y}_2 \mid \mathbf{X} = \mathbf{x}]$. We compute the adaptive estimates of all this parameters, where $\mathbf{x}$ varies over all the sample curves of the absorbance spectra. We plot the adaptive estimates of $\mathbb{E}[\mathbf{Y}_1 \mid \mathbf{X} = \mathbf{x}]$, $\mathbb{E}[\mathbf{Y}_2 \mid \mathbf{X} = \mathbf{x}]$, $\mathbb{VAR}[\mathbf{Y}_1 \mid \mathbf{X} = \mathbf{x}]$ and $\mathbb{VAR}[\mathbf{Y}_2 \mid \mathbf{X} = \mathbf{x}]$ against the L_2 norm of $\mathbf{x}$ in Fig. 5, and the adaptive estimate of $\mathbb{COR}[\mathbf{Y}_1, \mathbf{Y}_2 \mid \mathbf{X} = \mathbf{x}]$ against the L_2 norm of $\mathbf{x}$ in Fig. 6. The clear patterns of variation of the regression parameters over the covariate values are noticeable in each of the plots.

Next, we demonstrate the adaptive estimates of the conditional mean in another dataset, where both the response and the covariate are random functions. The dataset we consider is the Cigar data, which is available in the 'Ecdat' package in R. This dataset contains information about cigarette sales in packs per capita, per capita net disposable income (NDI) and other economic parameters in 46 states in the USA over a 30 years period from 1963 to 1992. We consider the curve of NDI over 30 years as the covariate $\mathbf{X}$, and the curve of cigarette sales over 30 years as the response $\mathbf{Y}$. So, both the response and the covariate in this setup are random functions, and our

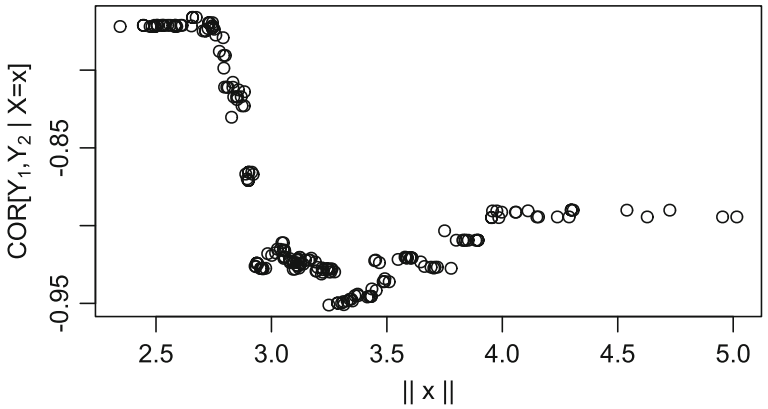

Fig. 6 Plot of adaptive estimates of $\mathbb{COR}[\mathbf{Y}_1, \mathbf{Y}_2 \mid \mathbf{X} = \mathbf{x}]$ against the L_2-norm of x in the Tecator data

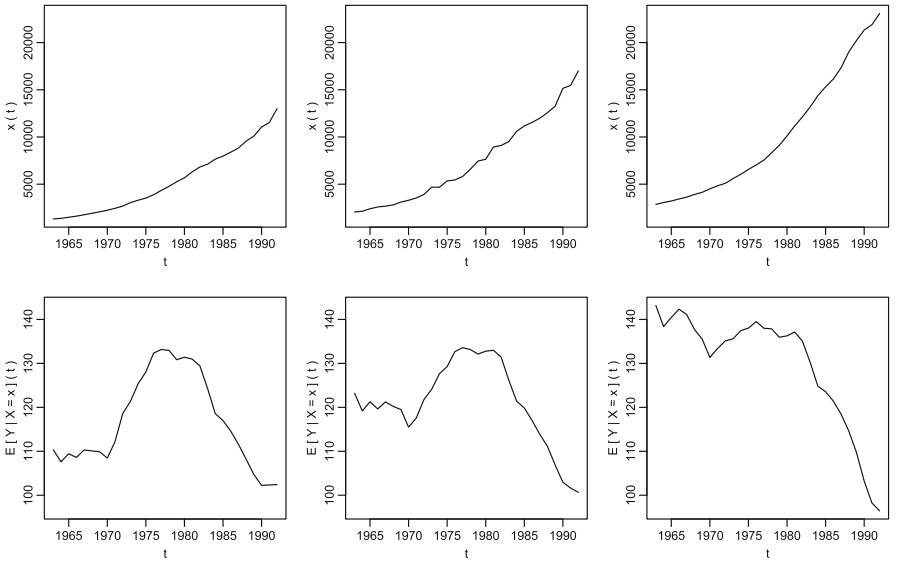

Fig. 7 Plots of adaptive estimates of $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ for 3 values of $\mathbf{x}$ in the Cigar data

sample size is 46. We choose 3 sample covariate curves as values of $\mathbf{x}$, and compute the adaptive estimates of $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ for these 3 values of $\mathbf{x}$. We plot the estimated curves along with the respective covariate curves in Fig. 7, where the first row contains the plots of the 3 curves chosen as values of $\mathbf{x}$, and the second row contains the plots of the corresponding adaptive estimates of $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$. The 3 estimated curves reflect the variation of $\mathbb{E}[\mathbf{Y} \mid \mathbf{X} = \mathbf{x}]$ over $\mathbf{x}$.

6 Concluding remarks

In this paper, we have derived the optimum convergence rate for a wide class of kernel regression estimates when the covariate as well as the response may be infinite-dimensional. It is shown that the convergence rates of such estimates do not depend on the dimension of the response, but they depend critically on the dimension of the covariate. We have seen that, for a wide class of covariates having infinite-dimensional Gaussian distributions, the convergence rate is much slower than the optimum achievable rate for finite-dimensional covariates. For instance, if the covariate is a real-valued continuous Gaussian Markov process in $L_p[0, 1]$, the convergence rate is $O((\log n)^{-\delta})$ for some $\delta > 0$. Theorem 4 implies that if $h_n^{2\beta} n\phi(\mathbf{x}, h_n) \longrightarrow 0$ as $n \longrightarrow \infty$, $[n\phi(\mathbf{x}, h_n)]^{1/2} c_n[\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})]$ converges in distribution to a Gaussian random element with zero mean as $n \longrightarrow \infty$, where $c_n = [E_n^{(2)}(\mathbf{x})]^{-1/2} E_n^{(1)}(\mathbf{x})$ is a sequence of positive numbers bounded and bounded away from 0. Note that this corresponds to an under-smoothed kernel estimate of $\Theta(\mathbf{x})$. On the other hand, if $h_n^{2\beta} n\phi(\mathbf{x}, h_n) \longrightarrow \infty$ as $n \longrightarrow \infty$, which includes the case of our optimum bandwidth obtained in Theorem 5, we have $h_n^{-\beta}[\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})] - h_n^{-\beta} \tilde{B}_n(\mathbf{x}) \longrightarrow 0$ *in probability* as $n \longrightarrow \infty$. Here, $\tilde{B}_n(\mathbf{x})$ is a non-random deterministic object described at the beginning of Sect. 4.1.

In Ferraty and Vieu (2006), Ferraty et al. (2006, 2010) and Chaouch and Laïb (2013, 2015), asymptotic properties of nonparametric regression estimates of different parameters other than the mean of the conditional distribution of the response were investigated. However, they only considered finite-dimensional responses, and they did not investigate the problem of optimum convergence rates of nonparametric regression estimates.

The problem of slow convergence rate of the regression estimates with infinite-dimensional covariates that has been derived in this paper may be coped with using an appropriate dimension reduction procedure on the covariate. Some procedures for such dimension reduction for infinite-dimensional covariates available in the literature are the uses of functional sliced inverse regression (Ferré and Yao 2003, 2005), functional average derivative regression (Ferraty et al. 2011) and distance correlation maximization (Vepakomma et al. 2016). If the covariate with the reduced dimension is adequate for regression analysis, the new small ball probability function in the reduced covariate space will lead to better convergence rates.

7 Proofs and mathematical details

Proof of Theorem 1 From the definitions of $\mathbb{L}_{\mathbf{x}}(\cdot)$, $G(\mathbf{Y})$ and $F(\mathbf{z})$ in the statement of Theorem 1, and from (4) and (5), we have

$$B_n(\mathbf{x}) = \Gamma'(\mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]) \left(\frac{\sum_{i=1}^n \mathbb{E}[\Psi(\mathbf{Y}_i) \mid \mathbf{X}_i] K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{\sum_{i=1}^n K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))} - \mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}] \right),$$

$$V_n(\mathbf{x}) = \Gamma'\left(\mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]\right) \left(\frac{\sum_{i=1}^n [\Psi(\mathbf{Y}_i) - \mathbb{E}[\Psi(\mathbf{Y}_i) \mid \mathbf{X}_i]] K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^n K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)} \right).$$

Set $R_n(\mathbf{x}) = [\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})] - B_n(\mathbf{x}) - V_n(\mathbf{x})$. From A(ii) and (6), we have $\|B_n(\mathbf{x})\| \longrightarrow 0$ as $n \longrightarrow \infty$. From A(ii) and Theorem 3, we have $\mathbb{E}[\|V_n(\mathbf{x})\|^2] \longrightarrow 0$ as $n \longrightarrow \infty$, and consequently $\|V_n(\mathbf{x})\| \longrightarrow 0$ *in probability* as $n \longrightarrow \infty$. So, $\|B_n(\mathbf{x}) + V_n(\mathbf{x})\| \longrightarrow 0$ *in probability* as $n \longrightarrow \infty$. Therefore,

$$\begin{aligned}
&\|R_n(\mathbf{x})\| \\
&= \left\| [\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})] - B_n(\mathbf{x}) - V_n(\mathbf{x}) \right\| \\
&= \left\| \Gamma\left(\frac{\sum_{i=1}^n \Psi(\mathbf{Y}_i) K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{\sum_{i=1}^n K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))} \right) - \Gamma(\mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]) \right. \\
&\quad \left. - \Gamma'\left(\mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}]\right) \left(\frac{\sum_{i=1}^n \Psi(\mathbf{Y}_i) K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\sum_{i=1}^n K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))} - \mathbb{E}[\Psi(\mathbf{Y}) \mid \mathbf{X} = \mathbf{x}] \right) \right\| \\
&= o\left(\|B_n(\mathbf{x}) + V_n(\mathbf{x})\|\right) \quad \text{whenever } \|B_n(\mathbf{x}) + V_n(\mathbf{x})\| \longrightarrow 0, \\
&= o\left(\|B_n(\mathbf{x})\| + \|V_n(\mathbf{x})\|\right) \quad \text{whenever } \|B_n(\mathbf{x})\| + \|V_n(\mathbf{x})\| \longrightarrow 0, \\
&= o_{\mathbb{P}}\left(\max\left\{h_n^\beta, [n\phi(\mathbf{x}, h_n)]^{-1/2}\right\}\right) \quad \text{as } n \longrightarrow 0, \qquad (16)
\end{aligned}$$

since from (6), we have $\|B_n(\mathbf{x})\| = O(h_n^\beta)$ as $n \longrightarrow \infty$, and from Theorem 3, we have $\|V_n(\mathbf{x})\| = O_{\mathbb{P}}([n\phi(\mathbf{x}, h_n)]^{-1/2})$ as $n \longrightarrow \infty$. So, B(iii) is satisfied. □

Proof of Theorem 2 Under the assumptions stated in Example 3, it follows that condition B(i) holds from the Holder continuity of $\Theta(\mathbf{z})$. The continuity of the linear operator $\mathbb{L}_{\mathbf{x}}(\cdot)$ follows from the invertibility of $\mathbf{I}(\Theta(\mathbf{x}))$, and B(ii) follows from the assumptions stated in Example 3 using arguments similar to those used in the proof of Theorem 3.2 in Chaudhuri and Dewanji (1995). We now proceed to verify condition B(iii).

Using arguments similar to those in the proof of Theorem 3.1 in Chaudhuri and Dewanji (1995), we get $\widehat{\Theta}_n(\mathbf{x}) \longrightarrow \Theta(\mathbf{x})$ *in probability* as $n \longrightarrow \infty$. Using this fact, (2) and a Taylor expansion of $\nabla g(\mathbf{Y}_i \mid \mathbf{t})$ at $\mathbf{t} = \widehat{\Theta}_n(\mathbf{x})$, we get

$$\begin{aligned}
&\sum_{i=1}^n \nabla g(\mathbf{Y}_i \mid \Theta(\mathbf{X}_i)) W_{i,n}(\mathbf{x}) = \sum_{i=1}^n \Delta_2(g(\mathbf{Y}_i \mid \eta_i(\mathbf{x}))) \left(\Theta(\mathbf{X}_i) - \widehat{\Theta}_n(\mathbf{x})\right) W_{i,n}(\mathbf{x}) \\
&\implies \widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x}) \\
&= \left[\sum_{i=1}^n \Delta_2(g(\mathbf{Y}_i \mid \eta_i(\mathbf{x}))) W_{i,n}(\mathbf{x}) \right]^{-1} \left(\sum_{i=1}^n \Delta_2(g(\mathbf{Y}_i \mid \eta_i(\mathbf{x}))) (\Theta(\mathbf{X}_i) - \Theta(\mathbf{x})) W_{i,n}(\mathbf{x}) \right) \\
&\quad - \left[\sum_{i=1}^n \Delta_2(g(\mathbf{Y}_i \mid \eta_i(\mathbf{x}))) W_{i,n}(\mathbf{x}) \right]^{-1} \left(\sum_{i=1}^n \nabla g(\mathbf{Y}_i \mid \Theta(\mathbf{X}_i)) W_{i,n}(\mathbf{x}) \right),
\end{aligned}$$

where $\eta_i(\mathbf{x})$ lies between $\Theta(\mathbf{X}_i)$ and $\widehat{\Theta}_n(\mathbf{x})$. Also, under the assumptions in Example 3, using arguments similar to those used in the proofs of Theorems 3.1 and 3.2 in Chaudhuri and Dewanji (1995), we get that $\| \sum_{i=1}^n \Delta_2(g(\mathbf{Y}_i \mid \eta_i(\mathbf{x})))W_{i,n}(\mathbf{x}) + \mathbf{I}(\Theta(\mathbf{x}))\| \longrightarrow 0$ *in probability* as $n \longrightarrow \infty$. Also, since $\Theta(\mathbf{z}) \in \mathcal{F}(\mathbf{x}, \beta, \mathbb{R}^q)$, we have $\max\{\|\Theta(\mathbf{X}_i) - \Theta(\mathbf{x})\| W_{i,n}(\mathbf{x}) \mid i = 1, \ldots, n\} \leq c h_n^\beta$ for all n, where $c > 0$ is a constant. Consequently, it follows that

$$
\begin{aligned}
&\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x}) \\
&\quad = [\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^n \mathbf{I}(\Theta(\mathbf{x}))(\Theta(\mathbf{X}_i) - \Theta(\mathbf{x}))W_{i,n}(\mathbf{x}) \right) \\
&\qquad + [\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^n \nabla g(\mathbf{Y}_i \mid \Theta(\mathbf{X}_i))W_{i,n}(\mathbf{x}) \right) \\
&\qquad + o_P\left(h_n^\beta\right) + o_P\left(\left\| [\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^n \nabla g(\mathbf{Y}_i \mid \Theta(\mathbf{X}_i))W_{i,n}(\mathbf{x}) \right) \right\| \right).
\end{aligned}
$$

Taking

$$
\begin{aligned}
V_n(\mathbf{x}) &= [\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^n \nabla g(\mathbf{Y}_i \mid \Theta(\mathbf{X}_i))W_{i,n}(\mathbf{x}) \right) \\
\text{and } B_n(\mathbf{x}) &= [\mathbf{I}(\Theta(\mathbf{x}))]^{-1} \left(\sum_{i=1}^n \mathbf{I}(\Theta(\mathbf{x}))(\Theta(\mathbf{X}_i) - \Theta(\mathbf{x}))W_{i,n}(\mathbf{x}) \right),
\end{aligned}
$$

we have $R_n(\mathbf{x}) = o_{\mathbb{P}}\left(h_n^\beta + \|V_n(\mathbf{x})\|\right)$ as $n \longrightarrow \infty$, and the proof is complete using the convergence rate of $\mathbb{E}[\|V_n(\mathbf{x})\|^2]$ as described in the proof on Theorem 1. □

Proof of Theorem 3 The arguments used in this proof are closely related to the arguments in the proof of Proposition 1 in Chagny and Roche (2016). Define $W_n(\mathbf{x}) = n^{-1} \sum_{i=1}^n [E_n^{(1)}(\mathbf{x})\phi(\mathbf{x}, h_n)]^{-1} K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))$, where $E_n^{(1)}(\mathbf{x})$ is as defined in (7). It follows from Bernstein's inequality (Serfling 2009, p. 95) and condition A(i) that

$$
\mathbb{P}[|W_n(\mathbf{x}) - 1| > (1/2)] \leq 2\exp(-c_1 n\phi(\mathbf{x}, h_n)), \tag{17}
$$

where c_1 is a positive constant. Note that

$$
\mathbb{E}[\|V_n(\mathbf{x})\|^2] = \mathbb{E}[\|V_n(\mathbf{x})\|^2 \mathbb{I}(W_n(\mathbf{x}) < (1/2))] + \mathbb{E}[\|V_n(\mathbf{x})\|^2 \mathbb{I}(W_n(\mathbf{x}) \geq (1/2))]. \tag{18}
$$

For the first term on the RHS in (18), using the fact that $\mathcal{B}$ is a type 2 Banach space and conditions A(i), A(ii) and B(ii), we have from (17),

$$
\begin{aligned}
&\mathbb{E}[\|V_n(\mathbf{x})\|^2 \mathbb{I}(W_n(\mathbf{x}) < (1/2))] = \mathbb{E}[\mathbb{E}[\|V_n(\mathbf{x})\|^2 \mathbb{I}(W_n(\mathbf{x}) < (1/2)) \,|\, \mathbf{X}_1, \ldots, \mathbf{X}_n]] \\
&\quad \le c_2 \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{E}[\|G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i) \,|\, \mathbf{X}_i]\|^2 \,|\, \mathbf{X}_i] K^2\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{\left(\sum_{i=1}^n K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)\right)^2} \mathbb{I}\left(W_n(\mathbf{x}) < \frac{1}{2}\right) \right] \\
&\quad \le c_3 \mathbb{E}\left[\frac{\sum_{i=1}^n K^2(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{\left(\sum_{i=1}^n K\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)\right)^2} \mathbb{I}\left(W_n(\mathbf{x}) < \frac{1}{2}\right) \right] \\
&\quad \le c_3 \mathbb{P}[|W_n(\mathbf{x}) - 1| > (1/2)] \le 2c_3 \exp(-c_1 n \phi(\mathbf{x}, h_n)) \qquad (19)
\end{aligned}
$$

for all sufficiently large n, where c_2 and c_3 are positive constants. Since $ue^{-u} \le e^{-1}$ for $u > 0$, from (19), we get that for all sufficiently large n,

$$
n\phi(\mathbf{x}, h_n) \mathbb{E}\left[\|V_n(\mathbf{x})\|^2 \mathbb{I}(W_n(\mathbf{x}) < (1/2)) \right] \le \frac{2c_3}{c_1 e}. \qquad (20)
$$

Now, for the second term on the RHS in (18), again using the fact that $\mathcal{B}$ is a type 2 Banach space, conditions A(i), A(ii), B(ii) and inequality (8), we get that for all sufficiently large n,

$$
\begin{aligned}
&\mathbb{E}[\|V_n(\mathbf{x})\|^2 \mathbb{I}(W_n(\mathbf{x}) \ge (1/2))] \\
&\quad \le \|\mathbb{L}_{\mathbf{x}}\|^2 \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n [G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i) \,|\, \mathbf{X}_i]] \frac{K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{E_n^{(1)}(\mathbf{x}) \phi(\mathbf{x}, h_n)} \right\|^2 \frac{\mathbb{I}(W_n(\mathbf{x}) \ge (1/2))}{(W_n(\mathbf{x}))^2} \right] \\
&\quad = \|\mathbb{L}_{\mathbf{x}}\|^2 \mathbb{E}\left[\mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n [G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i) \,|\, \mathbf{X}_i]] \frac{K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i))}{E_n^{(1)}(\mathbf{x}) \phi(\mathbf{x}, h_n)} \right\|^2 \Big| \mathbf{X}_1, \ldots, \mathbf{X}_n \right] \frac{\mathbb{I}(W_n(\mathbf{x}) \ge (1/2))}{(W_n(\mathbf{x}))^2} \right] \\
&\quad \le \|\mathbb{L}_{\mathbf{x}}\|^2 c_4 \mathbb{E}\left[\sum_{i=1}^n \mathbb{E}\left[\|G(\mathbf{Y}_i) - \mathbb{E}[G(\mathbf{Y}_i) \,|\, \mathbf{X}_i]\|^2 \Big| \mathbf{X}_i \right] \frac{K^2(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)) \mathbb{I}(W_n(\mathbf{x}) \ge (1/2))}{(W_n(\mathbf{x}))^2 (E_n^{(1)}(\mathbf{x}))^2 n^2 (\phi(\mathbf{x}, h_n))^2} \right] \\
&\quad \le c_5 \mathbb{E}\left[\sum_{i=1}^n \frac{K^2(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)) \mathbb{I}(W_n(\mathbf{x}) \ge (1/2))}{(W_n(\mathbf{x}))^2 (E_n^{(1)}(\mathbf{x}))^2 n^2 (\phi(\mathbf{x}, h_n))^2} \right] \le 4c_5 \mathbb{E}\left[\sum_{i=1}^n \frac{K^2\left(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)\right)}{\left(E_n^{(1)}(\mathbf{x})\right)^2 n^2 (\phi(\mathbf{x}, h_n))^2} \right] \\
&\quad = \frac{4c_5 E_n^{(2)}(\mathbf{x})}{\left(E_n^{(1)}(\mathbf{x})\right)^2} \frac{1}{n\phi(\mathbf{x}, h_n)} \le \frac{4c_5 L^2}{l^2} \frac{1}{n\phi(\mathbf{x}, h_n)} \\
&\implies \phi(\mathbf{x}, h_n) \mathbb{E}\left[\|V_n(\mathbf{x})\|^2 \mathbb{I}(W_n(\mathbf{x}) \ge (1/2)) \right] \le \frac{4c_5 L^2}{l^2}, \qquad (21)
\end{aligned}
$$

where c_4 and c_5 are positive constants. From (18), (20) and (21), we get $n\phi(\mathbf{x}, h_n)$ $\mathbb{E}[\|V_n(\mathbf{x})\|^2] = O(1)$ as $n \longrightarrow \infty$. □

Proof of Theorem 4 Note that

$$[n\phi(\mathbf{x},h_n)]^{1/2}\,[E_n^{(2)}(\mathbf{x})]^{-1/2}E_n^{(1)}(\mathbf{x})V_n(\mathbf{x}) = \frac{\sum_{i=1}^n \frac{K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{[E_n^{(2)}(\mathbf{x})]^{1/2}[n\phi(\mathbf{x},h_n)]^{1/2}}\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}_i)-\mathbb{E}[G(\mathbf{Y}_i)\,|\,\mathbf{X}_i])}{n^{-1}\sum_{i=1}^n\left[E_n^{(1)}(\mathbf{x})\phi(\mathbf{x},h_n)\right]^{-1}K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}. \tag{22}$$

Define

$$V_n^*(\mathbf{x}) = \sum_{i=1}^n \frac{K(h_n^{-1}d(\mathbf{x},\mathbf{X}_i))}{[E_n^{(2)}(\mathbf{x})]^{1/2}\,[n\phi(\mathbf{x},h_n)]^{1/2}}\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y}_i)-\mathbb{E}[G(\mathbf{Y}_i)\,|\,\mathbf{X}_i]).$$

The covariance operator of $V_n^*(\mathbf{x})$, denoted as $\mathbb{D}_n(\cdot,\cdot\,|\,\mathbf{x})$, is given by

$$\mathbb{D}_n(\mathbf{u},\mathbf{v}\,|\,\mathbf{x}) = \mathbb{E}\left[\langle\mathbf{u},\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y})-\mathbb{E}[G(\mathbf{Y})\,|\,\mathbf{X}])\rangle\,\langle\mathbf{v},\mathbb{L}_{\mathbf{x}}(G(\mathbf{Y})-\mathbb{E}[G(\mathbf{Y})\,|\,\mathbf{X}])\rangle\,\frac{K^2(h_n^{-1}d(\mathbf{x},\mathbf{X}))}{E_n^{(2)}(\mathbf{x})\phi(\mathbf{x},h_n)}\right]$$

for $\mathbf{u},\mathbf{v}\in\mathcal{B}$. Under conditions A(i), A(ii) and B(iv), $\mathbb{D}_n(\cdot,\cdot\,|\,\mathbf{x})$ converges to $\mathbb{D}(\cdot,\cdot\,|\,\mathbf{x})$ in the trace norm as $n\longrightarrow\infty$. Consequently, conditions (i) and (ii) in Theorem 1.1 in Kundu et al. (2000) are satisfied. Define

$$\mathbf{U}_{n,i}(\mathbf{x}) = \frac{K\left(h_n^{-1}d\left(\mathbf{x},\mathbf{X}_i\right)\right)}{[E_n^{(2)}(\mathbf{x})]^{1/2}}\mathbb{L}_{\mathbf{x}}\left(G(\mathbf{Y}_i)-\mathbb{E}\left[G(\mathbf{Y}_i)\,|\,\mathbf{X}_i\right]\right).$$

Given $\epsilon>0$ and $\mathbf{b}\in\mathcal{B}$, define

$$L_n(\epsilon,\mathbf{b}) = \sum_{i=1}^n \mathbb{E}\left[\left\langle\frac{\mathbf{U}_{n,i}(\mathbf{x})}{[n\phi(\mathbf{x},h_n)]^{1/2}},\mathbf{b}\right\rangle^2\mathbb{I}\left[\left|\left\langle\frac{\mathbf{U}_{n,i}(\mathbf{x})}{[n\phi(\mathbf{x},h_n)]^{1/2}},\mathbf{b}\right\rangle\right|>\epsilon\right]\right].$$

From A(i), A(ii) and B(ii), we have for any $\mathbf{b}$ with $\|\mathbf{b}\|=1$,

$$\begin{aligned}
L_n(\epsilon,\mathbf{b}) &= \mathbb{E}\left[\frac{\langle\mathbf{U}_{n,1}(\mathbf{x}),\mathbf{b}\rangle^2}{\phi(\mathbf{x},h_n)}\mathbb{I}\left[|\langle\mathbf{U}_{n,1}(\mathbf{x}),\mathbf{b}\rangle|>\epsilon\,[n\phi(\mathbf{x},h_n)]^{1/2}\right]\right]\\
&\le \mathbb{E}\left[\frac{\langle\mathbf{U}_{n,i}(\mathbf{x}),\mathbf{b}\rangle^2}{\phi(\mathbf{x},h_n)}\left[\frac{|\langle\mathbf{U}_{n,1}(\mathbf{x}),\mathbf{b}\rangle|}{\epsilon\,[n\phi(\mathbf{x},h_n)]^{1/2}}\right]^{\nu-2}\right]\\
&\le \left[\frac{1}{\epsilon\,[n\phi(\mathbf{x},h_n)]^{1/2}}\right]^{\nu-2}\mathbb{E}\left[\frac{\|\mathbb{L}_{\mathbf{x}}\|^{\nu}\,\|G(\mathbf{Y})-\mathbb{E}[G(\mathbf{Y})\,|\,\mathbf{X}]\|^{\nu}\,K^{\nu}(h_n^{-1}d(\mathbf{x},\mathbf{X}))}{[E_n^{(2)}(\mathbf{x})]^{\nu/2}\phi(\mathbf{x},h_n)}\right]\\
&\le c\,[n\phi(\mathbf{x},h_n)]^{-\frac{\nu-2}{2}}\longrightarrow 0
\end{aligned}$$

as $n \longrightarrow \infty$, where $\nu > 2$ is the constant mentioned in B(ii). Hence, condition (iii) in Theorem 1.1 in Kundu et al. (2000) is satisfied. Consequently,

$$V_n^*(\mathbf{x}) \longrightarrow \mathbf{W} \tag{23}$$

in distribution as $n \longrightarrow \infty$. Now, under conditions A(i), A(ii) and an application of the Markov inequality, we get

$$n^{-1} \sum_{i=1}^{n} \left[E_n^{(1)}(\mathbf{x}) \phi(\mathbf{x}, h_n) \right]^{-1} K(h_n^{-1} d(\mathbf{x}, \mathbf{X}_i)) \longrightarrow 1 \tag{24}$$

in probability as $n \longrightarrow \infty$. The proof is completed from (22), (23), (24) and an application of Slutsky's Theorem. □

Proof of Theorem 5 From the upper bounds of $\mathbb{E}\|B_n(\mathbf{x})\|^2$ and $\mathbb{E}\|V_n(\mathbf{x})\|^2$ in (6) and Theorem 3, respectively, and the lower bound of $\phi(\mathbf{x}, h_n)$ in (9), we have

$$\mathbb{E}\|B_n(\mathbf{x}) + V_n(\mathbf{x})\|^2 \leq 2\left[\mathbb{E}\|B_n(\mathbf{x})\|^2 + \mathbb{E}\|V_n(\mathbf{x})\|^2\right] \leq f_1(h_n)$$

for all sufficiently large n, where

$$f_1(h_n) = a h_n^{2\beta} + \frac{b}{nC_1}(1/h_n)^{t_1} \exp\left[m(h_n)\right], \tag{25}$$

and $a, b > 0$ are some constants. We establish below that the choice of bandwidths $\{h_n\}$ described in the statement of Theorem 5 is one which minimizes (25). Note that $m(h)$, which is defined in (10), is a differentiable function of h, and

$$m'(h) = -m(h)(1/h)\left(t_2 + \frac{t_3}{\log(1/h)}\right). \tag{26}$$

Consequently, $f_1(h)$ is differentiable for all n, and

$$\begin{aligned} f_1'(h) &= 2\beta a h^{2\beta-1} - \frac{bt_1}{nC_1}(1/h)^{t_1+1} \exp[m(h)] \\ &\quad - \frac{b}{nC_1}(1/h)^{t_1+1} \exp[m(h)] m(h) \left(t_2 + \frac{t_3}{\log(1/h)}\right) \qquad (27) \\ &= \exp[m(h)] \left[\frac{2\beta a h^{2\beta-1}}{\exp[m(h)]} - \frac{bt_1}{nC_1}(1/h)^{t_1+1} \right. \\ &\quad \left. - \frac{b}{nC_1}(1/h)^{t_1+1} m(h) \left(t_2 + \frac{t_3}{\log(1/h)}\right) \right]. \qquad (28) \end{aligned}$$

From (28), we get that for every fixed n, $f_1'(h) \longrightarrow -\infty$ as $h \longrightarrow 0^+$, and for any $0 < s < 1$, $f_1'(s) > 0$ for all sufficiently large n. Since $f_1'(h)$ is continuous in h for $0 < h < 1$, given any $0 < s < 1$, $f_1'(h)$ must have a root in $(0, s)$ for all sufficiently

large n. For any fixed n, consider $h_0 = \inf\{h \mid f_1'(h) = 0\}$. Again, since $f_1'(h)$ is continuous in h, we have $f_1'(h_0) = 0$. Further, since $f_1'(h) \longrightarrow -\infty$ as $h \longrightarrow 0^+$, from the continuity of $f_1'(h)$ we have $f_1'(h) < 0$ for $h < h_0$, which implies that $f_1(h)$ is a decreasing function for $h < h_0$. Also, for any $0 < s < s' < 1$, we have for all sufficiently large n, $f_1'(h) > 0$ for all $s \le h \le s'$, which implies $f_1(h)$ is increasing in $s \le h \le s'$. Therefore, $f_1(h)$ must have a minima for all sufficiently large n, whose corresponding h will satisfy $f_1'(h) = 0$. Now, from (27), $f_1'(h_n) = 0$ implies that

$$
\begin{aligned}
2\beta a h_n^{2\beta-1} &= n^{-1}(1/h_n)^{t_1+1}\exp[m(h_n)] \\
&\quad \times \left[\frac{bt_1}{C_1} + \frac{b}{C_1} m(h_n)\left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right] \\
\iff h_n^{2\beta} &= n^{-1}(1/h_n)^{t_1}\exp[m(h_n)] \\
&\quad \times \left[\frac{bt_1}{2\beta a C_1} + \frac{b}{2\beta a C_1} m(h_n)\left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right] \qquad (29) \\
\iff n &= (1/h_n)^{2\beta+t_1}\exp[m(h_n)] \\
&\quad \times \left[\frac{bt_1}{2\beta a C_1} + \frac{b}{2\beta a C_1} m(h_n)\left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right] \qquad (30) \\
\iff \frac{\log n}{m(h_n)} &= 1 + (2\beta + t_1)\frac{\log(1/h_n)}{m(h_n)} + \frac{1}{m(h_n)} \\
&\quad \times \log\left(\left[\frac{bt_1}{2\beta a C_1} + \frac{b}{2\beta a C_1} m(h_n)\left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right]\right). \qquad (31)
\end{aligned}
$$

Let $\{h_n\}$ be such that $f_1'(h_n) = 0$ for all n. If either $t_2 > 0$ or $t_3 > 1$, then from (30), we get that $h_n \longrightarrow 0^+$ as $n \longrightarrow \infty$. Consequently, from (29), we have

$$
nC_1 h_n^{t_1}\exp[-m(h_n)] = h_n^{-2\beta}\left[\frac{bt_1}{2\beta a} + \frac{b}{2\beta a} m(h_n)\left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right] \longrightarrow \infty
$$

as $n \longrightarrow \infty$, which implies $n\phi(\mathbf{x}, h_n) \longrightarrow \infty$ as $n \longrightarrow \infty$ from the lower bound of $\phi(\mathbf{x}, h_n)$ in (9). Therefore, $\{h_n\}$ satisfies A(ii). Also, $h_n \longrightarrow 0^+$ as $n \longrightarrow \infty$ implies that

$$
\begin{aligned}
&\frac{\log(1/h_n)}{m(h_n)} \longrightarrow 0 \text{ as } n \longrightarrow \infty, \qquad (32) \\
&\frac{1}{m(h_n)}\log\left(\left[\frac{bt_1}{2\beta a C_1} + \frac{b}{2\beta a C_1} m(h_n)\left(t_2 + \frac{t_3}{\log(1/h_n)}\right)\right]\right) \longrightarrow 0 \qquad (33)
\end{aligned}
$$

as $n \longrightarrow \infty$. Combining (31), (32) and (33), we have

$$
\frac{\log n}{m(h_n)} \longrightarrow 1 \text{ as } n \longrightarrow \infty. \qquad (34)
$$

Consequently, when either $t_2 > 0$ or $t_3 > 1$, we have for all sufficiently large n,

$$ah_n^{2\beta} < aC_2'(m^{-1}(\log n))^{2\beta}, \tag{35}$$

where C_2' is a positive constant depending on C_2 and β. From (25) and (29), we get that $ah_n^{2\beta} < f_1(h_n) < 2ah_n^{2\beta}$ for all sufficiently large n, and consequently $f_1(h_n) < 2aC_2'(m^{-1}(\log n))^{2\beta}$ for all sufficiently large n. Hence, for the bandwidth sequence $\{h_n\}$ minimizing $f_1(h)$ for every fixed n, we have

$$\mathbb{E}\|B_n(\mathbf{x}) + V_n(\mathbf{x})\|^2 < 2aC_2'\left(m^{-1}(\log n)\right)^{2\beta} \tag{36}$$

for all sufficiently large n, which implies $\|B_n(\mathbf{x}) + V_n(\mathbf{x})\| = O_{\mathbb{P}}\left(\left(m^{-1}(\log n)\right)^{\beta}\right)$ as $n \longrightarrow \infty$. Also, when either $t_2 > 0$ or $t_3 > 1$, from (29) and the lower bound of $\phi(\mathbf{x}, h)$ in (9), we get

$$h_n^{2\beta} / \left[[n\phi(\mathbf{x}, h_n)]^{-1}\right] \longrightarrow \infty \tag{37}$$

as $n \longrightarrow \infty$. Hence, from (35) and (37), we get that $\left(m^{-1}(\log n)\right)^{-\beta} \|R_n(\mathbf{x})\| = o_{\mathbb{P}}(1)$ as $n \longrightarrow \infty$. Therefore, $\left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\| = O_{\mathbb{P}}\left(\left(m^{-1}(\log n)\right)^{\beta}\right)$ as $n \longrightarrow \infty$.

Next, if $\mathbb{E}[\|R_n(\mathbf{x})\|^2] = o\left(\delta_n^2\right)$ as $n \longrightarrow \infty$, where $\delta_n = \max\left\{h_n^{\beta}, \left[n\phi(\mathbf{x}, h_n)\right]^{-1/2}\right\}$, we get from (35) and (37) that $\left(m^{-1}(\log n)\right)^{-2\beta} \mathbb{E}\|R_n(\mathbf{x})\|^2 \longrightarrow 0$ as $n \longrightarrow \infty$. From (3), we have $\mathbb{E}\left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\|^2 \leq 2\mathbb{E}\left\|B_n(\mathbf{x}) + V_n(\mathbf{x})\right\|^2 + 2\mathbb{E}\|R_n(\mathbf{x})\|^2$. Therefore, from (36), we get $\mathbb{E}\left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\|^2 = O\left(\left(m^{-1}(\log n)\right)^{2\beta}\right)$ as $n \longrightarrow \infty$. □

Proof of Theorem 6 For a sequence of bandwidths $\{h_n\}$ satisfying A(ii), it follows from the upper bound of $\phi(\mathbf{x}, h)$ in (9) and the definition of $m(h)$ in (10) that

$$\begin{aligned} & nC_3h_n^{t_4}\exp\left[-(C_4/C_2)m(h_n)\right] \geq n\phi(\mathbf{x}, h_n) \longrightarrow \infty \text{ as } n \longrightarrow \infty \\ & \Longrightarrow \log n - t_4\log(1/h_n) - (C_4/C_2)m(h_n) \longrightarrow \infty \\ & \Longleftrightarrow m(h_n)\left[\frac{\log n}{m(h_n)} - t_4\frac{\log(1/h_n)}{m(h_n)} - \frac{C_4}{C_2}\right] \longrightarrow \infty \end{aligned} \tag{38}$$

as $n \longrightarrow \infty$. Now, since either $t_2 > 0$ or $t_3 > 1$, and $h_n \longrightarrow 0$ as $n \longrightarrow \infty$ under assumption A(ii), we have

$$m(h_n) \longrightarrow \infty \text{ as } n \longrightarrow \infty \quad \text{and} \quad \frac{\log(1/h_n)}{m(h_n)} \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{39}$$

Hence, for (38) to be satisfied, in view of (39), we must have, for all sufficiently large n,

$$\frac{\log n}{m(h_n)} - \frac{C_4}{C_2} > 0 \iff m^{-1}\left(\frac{\log n}{(C_4/C_2)}\right) < h_n \implies \frac{h_n}{m^{-1}(\log n)} > c_1 > 0, \tag{40}$$

where c_1 is a constant depending on C_2, C_4. Clearly, when $C_2 = C_4$, $c_1 = 1$. □

Proof of Theorem 7 Suppose, if possible,

$$\liminf_{n\longrightarrow\infty} \mathbb{P}\left[\left(m^{-1}(\log n)\right)^{-\beta} \left\|\widehat{\Theta}_n(\mathbf{x}) - \Theta(\mathbf{x})\right\| > c\right] = 0 \tag{41}$$

for every $c > 0$. Then, given any $c > 0$, there is a subsequence $\{n'\}$ such that

$$\lim_{n'\longrightarrow\infty} \mathbb{P}\left[\left(m^{-1}(\log n')\right)^{-\beta} \left\|\widehat{\Theta}'_n(\mathbf{x}) - \Theta(\mathbf{x})\right\| > c\right] = 0. \tag{42}$$

Consider the bandwidth sequence $\{h_{n'}\}$. If $\liminf_{n'\longrightarrow\infty} h_{n'}^{2\beta} n\phi(\mathbf{x}, h_{n'}) = 0$, then there exists a further subsequence $\{n''\}$ such that $h_{n''}^{2\beta} n\phi(\mathbf{x}, h_{n''}) \longrightarrow 0$ as $n'' \longrightarrow \infty$. But in this case, we get a contradiction of (42) from Lemma 2 in the supplement. On the other hand, if $\limsup_{n'\longrightarrow\infty} h_{n'}^{2\beta} n\phi(\mathbf{x}, h_{n'}) = \infty$, then there exists a further subsequence $\{n''\}$ such that $h_{n''}^{2\beta} n\phi(\mathbf{x}, h_{n''}) \longrightarrow \infty$ as $n'' \longrightarrow \infty$. But again, we get a contradiction of (42) from Lemma 3 in the supplement. We consider the only remaining case, which is $0 < \liminf_{n'\longrightarrow\infty} h_{n'}^{2\beta} n\phi(\mathbf{x}, h_{n'}) \leq \limsup_{n'\longrightarrow\infty} h_{n'}^{2\beta} n\phi(\mathbf{x}, h_{n'}) < \infty$. Then, there exist $\epsilon_1 > 0$, $\epsilon_2 > 0$ and a further subsequence $\{n''\}$ such that $0 < \epsilon_1 < h_{n''}^{2\beta} n\phi(\mathbf{x}, h_{n''}) < \epsilon_2$ for all sufficiently large n''. But in this case also, we get a contradiction of (42) from Lemma 4 in the supplement. Therefore, the assertion (41) is not possible, and this completes the proof. □

Proof of Theorem 8 From (29) in the proof of Theorem 5 and the lower bound of $\phi(\mathbf{x}, h)$ in (9), it follows that

$$h_n^{2\beta} n\phi(\mathbf{x}, h_n) \longrightarrow \infty \tag{43}$$

as $n \longrightarrow \infty$. Now, choose $\Theta(\cdot)$ as in Theorem 7 such that $h_n^{-\beta}\|\tilde{B}_n(\mathbf{x})\| \geq b_1 > 0$ for a constant b_1 and all sufficiently large n. So, $\mathbb{P}[h_n^{-\beta}\|B_n(\mathbf{x})\| > b_1/2] \longrightarrow 1$ as $n \longrightarrow \infty$. Hence, for this choice of $\Theta(\cdot)$ and using Theorem 4 and (43), we have

$$\frac{\|V_n(\mathbf{x})\|}{\|B_n(\mathbf{x})\|} = \frac{1}{[n\phi(\mathbf{x}, h_n)]^{1/2} h_n^{\beta}} \frac{[n\phi(\mathbf{x}, h_n)]^{1/2}\|V_n(\mathbf{x})\|}{h_n^{-\beta}\|B_n(\mathbf{x})\|} = o_{\mathbb{P}}(1) \text{ as } n \longrightarrow \infty.$$

□

Proof of Theorem 9 Let $\mathbf{U}$ and $\mathbf{V}$ be two nonnegative random variables. Then, given any $\epsilon > 0$ and $\delta > 0$, we have

$$\mathbb{P}\left[\frac{\mathbf{U}}{\mathbf{V}} < \epsilon\right] \geq \mathbb{P}\left[\mathbf{U} < \epsilon\delta,\ \mathbf{V} > \delta\right] \geq \mathbb{P}\left[\mathbf{U} < \epsilon\delta\right] + \mathbb{P}\left[\mathbf{V} > \delta\right] - 1. \tag{44}$$

We denote our optimum bandwidth minimizing (25) in the proof of Theorem 5 as $h_n^{(op)}$. Given any $\epsilon > 0$, from Lemma 7 in the supplement, we get that there is $\delta > 0$ such that

$$\mathbb{P}\left[(h_n^{(b)})^{-\beta}\left\|\widehat{\Theta}_n^{(b)}(\mathbf{x}) - \Theta(\mathbf{x})\right\| > \delta\right] > 1 - \epsilon \tag{45}$$

for all sufficiently large n. Further, from Lemma 6 in the supplement, we get that for this constant δ,

$$\mathbb{P}\left[(h_n^{(op)})^{-\beta}\left\|\widehat{\Theta}_n^{(op)}(\mathbf{x}) - \Theta(\mathbf{x})\right\| < \epsilon\delta\right] > 1 - \epsilon \tag{46}$$

for all sufficiently large n. Therefore, from (44), (45) and (46), we get that

$$\frac{(h_n^{(op)})^{-\beta}\left\|\widehat{\Theta}_n^{(op)}(\mathbf{x}) - \Theta(\mathbf{x})\right\|}{(h_n^{(b)})^{-\beta}\left\|\widehat{\Theta}_n^{(b)}(\mathbf{x}) - \Theta(\mathbf{x})\right\|} = o_{\mathbb{P}}(1) \quad \text{as } n \longrightarrow \infty. \tag{47}$$

Hence, from (47) and Lemma 5 in the supplement, we have

$$\frac{\left\|\widehat{\Theta}_n^{(op)}(\mathbf{x}) - \Theta(\mathbf{x})\right\|}{\left\|\widehat{\Theta}_n^{(b)}(\mathbf{x}) - \Theta(\mathbf{x})\right\|} = o_{\mathbb{P}}(1) \quad \text{as } n \longrightarrow \infty.$$

On the other hand, from Lemmas 5, 6 and 7 in the supplement, we have

$$\frac{\mathbb{E}\left\|\widehat{\Theta}_n^{(op)}(\mathbf{x}) - \Theta(\mathbf{x})\right\|^2}{\mathbb{E}\left\|\widehat{\Theta}_n^{(b)}(\mathbf{x}) - \Theta(\mathbf{x})\right\|^2} = o(1) \quad \text{as } n \longrightarrow \infty.$$

□

Proof of Theorem 10 This proof is partly based on arguments used in Chagny and Roche (2014, 2016). For every $h \in \mathbb{H}_n$, we have

$$\begin{aligned}\left\|\widehat{\Theta}_n\left(\mathbf{x}, h_n^*\right) - \Theta\left(\mathbf{x}\right)\right\|^2 &\leq 3\left[\left\|\widehat{\Theta}_n\left(\mathbf{x}, h_n^*\right) - \widehat{\Theta}_n\left(\mathbf{x}, \max\{h_n^*, h\}\right)\right\|^2\right. \\ &\quad \left. + \left\|\widehat{\Theta}_n\left(\mathbf{x}, h\right) - \widehat{\Theta}_n\left(\mathbf{x}, \max\{h_n^*, h\}\right)\right\|^2\right] + 3\left\|\widehat{\Theta}_n(\mathbf{x}, h) - \Theta(\mathbf{x})\right\|^2 \\ &\leq 3\left[\left(C_n(\mathbf{x}, h) + D_n\left(\mathbf{x}, h_n^*\right)\right) + \left(C_n\left(\mathbf{x}, h_n^*\right) + D_n(\mathbf{x}, h)\right)\right]\end{aligned}$$

$$+ 3 \left\| \widehat{\Theta}_n(\mathbf{x}, h) - \Theta(\mathbf{x}) \right\|^2$$
$$\leq 6 \left[C_n(\mathbf{x}, h) + D_n(\mathbf{x}, h) \right] + 3 \left\| \widehat{\Theta}_n(\mathbf{x}, h) - \Theta(\mathbf{x}) \right\|^2. \tag{48}$$

From Lemmas 8, 9 and 11 in the supplement, we get

$$C_n(\mathbf{x}, h) \leq C_n^{(1)}(\mathbf{x}, h) + C_n^{(2)}(\mathbf{x}, h). \tag{49}$$

Here, for all sufficiently large n,

$$\mathbb{E}\left[C_n^{(1)}(\mathbf{x}, h) \right] \leq c_1 h^{2\beta} + \frac{1}{n \log n} \tag{50}$$

for all $h \in \mathbb{H}_n$ and some constant $c_1 > 0$ independent of h. Also,

$$\mathbb{P}\left[C_n^{(2)}(\mathbf{x}, h) > n^{-2} \right] = O\left(n^{-2} \right) \quad \text{as } n \longrightarrow \infty. \tag{51}$$

Further, $C_n^{(2)}(\mathbf{x}, h) = 0$ for all h if $R_n(\mathbf{x}, h) = 0$ for all h. From Lemma 9 in the supplement, we get that for all sufficiently large n,

$$\mathbb{E}\left[D_n(\mathbf{x}, h) \right] \leq c_2 \frac{\log n}{n \phi(\mathbf{x}, h)} \tag{52}$$

for all $h \in \mathbb{H}_n$, where $c_2 > 0$ is a constant independent of h. On the other hand, from decomposition (3), we get

$$\left\| \widehat{\Theta}_n(\mathbf{x}, h) - \Theta(\mathbf{x}) \right\|^2 \leq 3 \left[\left(\| B_n(\mathbf{x}, h) \|^2 + M h^{2\beta} \right) + 2 \| V_n(\mathbf{x}, h) \|^2 \right]$$
$$+ 3 \left(\| R_n(\mathbf{x}, h) \|^2 - \left(M h^{2\beta} + \| V_n(\mathbf{x}, h) \|^2 \right) \right)_+, \tag{53}$$

where M is the constant described in condition D(ii). From inequality (6) and Theorem 3, we have

$$\mathbb{E}\left[\left(\| B_n(\mathbf{x}, h) \|^2 + M h^{2\beta} \right) + 2 \| V_n(\mathbf{x}, h) \|^2 \right] \leq c_3 h^{2\beta} + \frac{c_4}{n \phi(\mathbf{x}, h)} \tag{54}$$

for all sufficiently large n and some constants $c_3 > 0$ and $c_4 > 0$ independent of h. Also, from Lemma 11 in the supplement, we have

$$\max_{h \in \mathbb{H}_n} \left(\| R_n(\mathbf{x}, h) \|^2 - \left(M h^{2\beta} + \| V_n(\mathbf{x}, h) \|^2 \right) \right)_+ = o_{\mathbb{P}}\left(n^{-2} \right) \tag{55}$$

as $n \longrightarrow \infty$. Therefore, from (48)–(55), we get that

$$\begin{aligned}
&\left\|\widehat{\Theta}_n\left(\mathbf{x}, h_n^*\right) - \Theta(\mathbf{x})\right\|^2 \\
&\quad \leq \left[6C_n^{(1)}(\mathbf{x}, h) + 6D_n(\mathbf{x}, h) + 9\left(\left(\|B_n(\mathbf{x}, h)\|^2 + Mh^{2\beta}\right) + 2\|V_n(\mathbf{x}, h)\|^2\right)\right] \\
&\quad + \left[6C_n^{(2)}(\mathbf{x}, h) + 9\left(\|R_n(\mathbf{x}, h)\|^2 - \left(Mh^{2\beta} + \|V_n(\mathbf{x}, h)\|^2\right)\right)_+\right],
\end{aligned}$$

where

$$\begin{aligned}
&\mathbb{E}\left[6C_n^{(1)}(\mathbf{x}, h) + 6D_n(\mathbf{x}, h) + 9\left(\left(\|B_n(\mathbf{x}, h)\|^2 + Mh^{2\beta}\right) + 2\|V_n(\mathbf{x}, h)\|^2\right)\right] \\
&\quad = O\left(h^{2\beta} + \frac{\log n}{n\phi(\mathbf{x}, h)}\right) \qquad (56)
\end{aligned}$$

and

$$\begin{aligned}
&\max_{h \in \mathbb{H}_n}\left[6C_n^{(2)}(\mathbf{x}, h) + 9\left(\|R_n(\mathbf{x}, h)\|^2 - \left(Mh^{2\beta} + \|V_n(\mathbf{x}, h)\|^2\right)\right)_+\right] \\
&\quad = o_{\mathbb{P}}\left(n^{-2}\right) \quad \text{as } n \longrightarrow \infty. \qquad (57)
\end{aligned}$$

Further, if $R_n(\mathbf{x}, h) = 0$ for all h, then

$$\max_{h \in \mathbb{H}_n}\left[6C_n^{(2)}(\mathbf{x}, h) + 9\left(\|R_n(\mathbf{x}, h)\|^2 - \left(Mh^{2\beta} + \|V_n(\mathbf{x}, h)\|^2\right)\right)_+\right] = 0 \quad \text{for all } h.$$

From (56) and (57), we get

$$\left\|\widehat{\Theta}_n\left(\mathbf{x}, h_n^*\right) - \Theta(\mathbf{x})\right\|^2 = O_{\mathbb{P}}(\lambda_n) \quad \text{as } n \longrightarrow \infty,$$

and if $R_n(\mathbf{x}, h) = 0$ for all h, then

$$\mathbb{E}\left\|\widehat{\Theta}_n\left(\mathbf{x}, h_n^*\right) - \Theta(\mathbf{x})\right\|^2 = O(\lambda_n) \quad \text{as } n \longrightarrow \infty.$$

□

Supplement

The supplement contains some results and mathematical details required to prove the theorems in the paper. It has four sections. The first section contains a few results on small ball probabilities of some non-Gaussian processes. The second, the third and the fourth sections contain some technical details required to prove Theorems 7, 9 and 10, respectively.

Acknowledgements We thank the Editor, the Associate Editor and three reviewers for their extremely careful reading and valuable comments and suggestions that led to a substantially revised and significantly improved version of the paper.

References

Aerts, M., Claeskens, G. (1997). Local polynomial estimation in multiparameter likelihood models. *Journal of the American Statistical Association*, *92*(440), 1536–1545.

Araujo, A., Giné, E. (1980). *The central limit theorem for real and Banach valued random variables*. New York: Wiley.

Bhatia, R. (2009). *Notes on functional analysis*. New Delhi: Hindustan Book Agency.

Burba, F., Ferraty, F., Vieu, P. (2009). k-Nearest neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics*, *21*(4), 453–469.

Cameron, R. H., Martin, W. T. (1944). Transformations of Weiner integrals under translations. *Annals of Mathematics*, *45*(2), 386–396.

Chagny, G., Roche, A. (2014). Adaptive and minimax estimation of the cumulative distribution function given a functional covariate. *Electronic Journal of Statistics*, *8*(2), 2352–2404.

Chagny, G., Roche, A. (2016). Adaptive estimation in the functional nonparametric regression model. *Journal of Multivariate Analysis*, *146*, 105–118.

Chaouch, M., Laïb, N. (2013). Nonparametric multivariate L_1-median regression estimation with functional covariates. *Electronic Journal of Statistics*, *7*, 1553–1586.

Chaouch, M., Laïb, N. (2015). Vector-on-function quantile regression for stationary ergodic processes. *Journal of the Korean Statistical Society*, *44*(2), 161–178.

Chaudhuri, P., Dewanji, A. (1995). On a likelihood-based approach in nonparametric smoothing and cross-validation. *Statistics & Probability Letters*, *22*(1), 7–15.

Dereich, S., Lifshits, M. (2005). Probabilities of randomly centered small balls and quantization in Banach spaces. *The Annals of Probability*, *33*(4), 1397–1421.

Dette, H., Wieczorek, G. (2009). Testing for a constant coefficient of variation in nonparametric regression. *Journal of Statistical Theory and Practice*, *3*(3), 587–612.

Dette, H., Marchlewski, M., Wagener, J. (2012). Testing for a constant coefficient of variation in nonparametric regression by empirical processes. *Annals of the Institute of Statistical Mathematics*, *64*(5), 1045–1070.

Donoho, D. L., Liu, R. C. (1991a). Geometrizing rates of convergence, II. *The Annals of Statistics*, *19*(2), 633–667.

Donoho, D. L., Liu, R. C. (1991b). Geometrizing rates of convergence, III. *The Annals of Statistics*, *19*(2), 668–701.

Ferraty, F., Vieu, P. (2006). *Nonparametric functional data analysis*: *Theory and practice*. New York: Springer.

Ferraty, F., Laksaci, A., Vieu, P. (2006). Estimating some characteristics of the conditional distribution in nonparametric functional models. *Statistical Inference for Stochastic Processes*, *9*(1), 47–76.

Ferraty, F., Mas, A., Vieu, P. (2007). Nonparametric regression on functional data: Inference and practical aspects. *Australian & New Zealand Journal of Statistics*, *49*(3), 267–286.

Ferraty, F., Laksaci, A., Tadj, A., Vieu, P. (2010). Rate of uniform consistency for nonparametric estimates with functional variables. *Journal of Statistical Planning and Inference*, *140*(2), 335–352.

Ferraty, F., Park, J., Vieu, P. (2011). Estimation of a functional single index model. In F. Ferraty (Ed.), *Recent advances in functional data analysis and related topics, chapter 17*, pp. 111–116. New York: Springer.

Ferraty, F., Van Keilegom, I., Vieu, P. (2012). Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, *109*, 10–28.

Ferré, L., Yao, A. (2005). Smoothed functional inverse regression. *Statistica Sinica*, *15*(3), 665.

Ferré, L., Yao, A.-F. (2003). Functional sliced inverse regression analysis. *Statistics*, *37*(6), 475–488.

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

Hoffmann-Jorgensen, J., Shepp, L. A., Dudley, R. M. (1979). On the lower tail of Gaussian seminorms. *The Annals of Probability*, *7*(2), 319–342.

Ibragimov, I. A., Haśminskii, R. Z. (1980). On nonparametric estimation of regression. *Soviet Mathematics Doklady*, *21*, 810–814.
Klemelä, J. S. (2014). *Multivariate nonparametric regression and visualization*: *With R and applications to finance*. Hoboken: Wiley.
Kundu, S., Majumdar, S., Mukherjee, K. (2000). Central limit theorems revisited. *Statistics and Probability Letters*, *47*(3), 265–275.
Li, W. V. (2001). Small ball probabilities for Gaussian Markov processes under the L_p-norm. *Stochastic Processes and Their Applications*, *92*(1), 87–102.
Li, W. V., Shao, Q.-M. (2001). Gaussian processes: Inequalities, small ball probabilities and applications. *Stochastic Processes*: *Theory and Methods*, *19*, 533–597.
Lian, H. (2012). Convergence of nonparametric functional regression estimates with functional responses. *Electronic Journal of Statistics*, *6*, 1373–1391.
Lifshits, M. A. (2013). *Gaussian random functions*, Vol. 322. Dordrecht: Springer.
Lukić, M., Beder, J. (2001). Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, *353*(10), 3945–3969.
Mas, A. (2012). Lower bound in regression for functional data by representation of small ball probabilities. *Electronic Journal of Statistics*, *6*, 1745–1778.
Masry, E. (2005). Nonparametric regression estimation for dependent functional data: Asymptotic normality. *Stochastic Processes and Their Applications*, *115*(1), 155–177.
Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, *9*(1), 141–142.
Øksendal, B. (2003). *Stochastic differential equations*: *An introduction with applications*. New York: Springer.
Rachdi, M., Vieu, P. (2007). Nonparametric regression for functional data: Automatic smoothing parameter selection. *Journal of Statistical Planning and Inference*, *137*(9), 2784–2801.
Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, Vol. 162. Hoboken: Wiley.
Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, *84*(405), 276–283.
Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, *8*(6), 1348–1360.
Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, *10*(4), 1040–1053.
Vepakomma, P., Tonde, C., Elgammal, A. (2016). Supervised dimensionality reduction via distance correlation maximization. arXiv preprint arXiv:1601.00236.
Watson, G. S. (1964). Smooth regression analysis. *Sankhyā*: *The Indian Journal of Statistics, Series A*, *26*(4), 359–372.
Yatracos, Y. G. (1988). A lower bound on the error in nonparametric regression type problems. *The Annals of Statistics*, *16*(3), 1180–1187.