The Annals of Statistics
2020, Vol. 48, No. 5, 2671–2697
https://doi.org/10.1214/19-AOS1902

ADDITIVE REGRESSION WITH HILBERTIAN RESPONSES

BY JEONG MIN JEON* AND BYEONG U. PARK**

*Department of Statistics, Seoul National University, *jeongmin.jeon@kuleuven.be; **bupark@stats.snu.ac.kr*

This paper develops a foundation of methodology and theory for the estimation of structured nonparametric regression models with Hilbertian responses. Our method and theory are focused on the additive model, while the main ideas may be adapted to other structured models. For this, the notion of Bochner integration is introduced for Banach-space-valued maps as a generalization of Lebesgue integration. Several statistical properties of Bochner integrals, relevant for our method and theory and also of importance in their own right, are presented for the first time. Our theory is complete. The existence of our estimators and the convergence of a practical algorithm that evaluates the estimators are established. These results are nonasymptotic as well as asymptotic. Furthermore, it is proved that the estimators achieve the univariate rates in pointwise, L^2 and uniform convergence, and that the estimators of the component maps converge jointly in distribution to Gaussian random elements. Our numerical examples include the cases of functional, density-valued and simplex-valued responses, demonstrating the validity of our approach.

1. Introduction. Regression analysis with non-Euclidean data is one of the major challenges in modern statistics. In many cases it is not transparent how one can go beyond traditional Euclidean methods to analyze non-Euclidean objects. The problem we tackle in this paper is particularly the case.

Let $\mathbb{H}$ be a separable Hilbert space with a zero vector $\mathbf{0}$, vector addition $\oplus$ and scalar multiplication $\odot$. For a probability space $(\Omega, \mathscr{F}, P)$, we consider a response $\mathbf{Y} : \Omega \to \mathbb{H}$. Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a predictor taking values in a compact subset of $\mathbb{R}^d$, say $[0, 1]^d$, and $\boldsymbol{\epsilon}$ be a $\mathbb{H}$-valued error satisfying $\mathrm{E}(\boldsymbol{\epsilon}|\mathbf{X}) = \mathbf{0}$. For the definition of the conditional expectation of a $\mathbb{H}$-valued random element, see [4]. In this paper we develop a unified approach to fitting the additive model

$$\mathbf{Y} = \mathbf{m}_0 \oplus \bigoplus_{j=1}^{d} \mathbf{m}_j(X_j) \oplus \boldsymbol{\epsilon}, \tag{1.1}$$

where $\mathbf{m}_0$ is a constant in $\mathbb{H}$ and $\mathbf{m}_1, \ldots, \mathbf{m}_d : [0, 1] \to \mathbb{H}$ are measurable maps.

Additivity is a commonly employed structure with which one is able to avoid the curse of dimensionality in nonparametric regression. A powerful kernel-based method for achieving this is the smooth backfitting (SBF) technique originated by [31]. The idea has been developed for various structured nonparametric models; see [3, 25, 26, 30, 46] and [17], for example. All of them, however, treated the case of Euclidean response.

There have been a few applications of the SBF idea to the case of functional response, in which case $\mathbb{H}$ is a space of real-valued functions defined on a domain $\mathcal{T} \subset \mathbb{R}$. Examples include [47] and [37], but their techniques and theory are essentially the same as in the case of Euclidean response. They applied the SBF technique to a functional response $\mathbf{Y} \equiv Y(\cdot)$

Received December 2018; revised August 2019.
MSC2020 subject classifications. Primary 62G08; secondary 62G20.
Key words and phrases. Additive models, smooth backfitting, Bochner integral, non-Euclidean data, infinite-dimensional spaces, Hilbert spaces, functional responses.

in a *pointwise* manner, that is, to $Y(t)$ for each $t \in \mathcal{T}$, or to a finite number of its singular/principal components that live in a Euclidean space. These methods have certain drawbacks. The pointwise application does not guarantee that the estimate of $\mathbf{m}_j(x_j) \equiv m_j(x_j)(\cdot)$ for each $x_j \in [0, 1]$ belongs to $\mathbb{H}$, particularly in the case where $\mathbb{H}$ is a space of smooth functions, as is typically the case with functional data. Methods based on singular/principal components require choosing the number of included components in a working model, which is very difficult.

Our approach in specialization to functional data does not have these drawbacks. It guarantees that the estimates of $\mathbf{m}_j(x_j)$ belong to the space $\mathbb{H}$ where the targets live. It does not need a dimension reduction procedure to deal with infinite-dimensional responses, while many others do in the functional data literature, in particular, those based on splines and FPCA. Moreover, the computation of our estimators is faster than the pointwise approach as the grid on $\mathcal{T}$ gets denser, since the proposed method estimates $m_j(x_j)(\cdot)$ on the whole $\mathcal{T}$ all at once.

One of the most appealing features of this work is to cover a very wide scope of possible applications that overpass the standard case of curve variables. There are numerous examples of Hilbertian variables. In the next section we introduce three examples, which we also treat in our numerical study in Section 5. These are functional variables, density-valued variables and simplex-valued variables. One may also apply our unified approach to various types of object oriented data, such as images, shapes and manifolds, if the data spaces are equipped with an inner product or after embedding them in a Hilbert space via a transformation. The literature for the analysis of such complex data objects has not been much developed. For recent works on object oriented data analysis, we refer to [34] and the references therein. For a framework of Hilbert space embedding, see [18].

Our work serves as the basic building block of structured nonparametric regression for Hilbertian responses in general. The unified approach involves integral operators acting on Hilbert-space-valued maps such as $\mathbf{m}_j$ in (1.1). The traditional Lebesgue integral theory does not apply here since it is for real-valued functions. Our first task is thus to develop a new theory that generalizes the conventional Lebesgue integration to the case of Hilbert-space-valued maps. For this we take the notion of Bochner integration, which is not very well known in statistics and is for Banach-space-valued maps. We detail the new theory in Section 2. Based on the Bochner integral theory, we present a powerful technique of estimating the model (1.1). The technique consists of solving a system of integral equations, expressed in terms of Bochner integrals, and an associated backfitting algorithm. We establish the existence of the estimator that solves the system of equations and the convergence of the algorithm. The major accomplishment at this stage of the work is to prove that the space of regression maps under the model (1.1) is closed. The results on the existence and convergence include nonasymptotic versions as well as asymptotic ones. The nonasymptotic results do not exist in the literature even, for the case $\mathbb{H} = \mathbb{R}$. Furthermore, we present complete theory for the rates of convergence of the estimators of the component maps $\mathbf{m}_j$ and their asymptotic distributions.

There have been a few attempts of dealing with possibly non-Euclidean responses. Examples include functional Nadaraya–Watson, locally linear and k-nearest neighbor estimation for general Hilbert- or Banach-space-valued responses. These are for full-dimensional estimation without structure on the regression map. For recent trends on this topic, we refer to the survey papers, [29] and [1]. Some others for L^2 or longitudinal responses include [6, 20, 48] and [40]. The first of these considered a representation of a functional response in terms of its finite number of principal components, each of which is expressed as a single index model with a multivariate predictor. The second studied a time-dynamic functional single-index model in a longitudinal data setting. The third discussed function-on-scalar varying coefficient models, and the last was for a spline method. For density-valued data, [38] introduced a transformation approach. [10] considered an additive model that describes a density

response in terms of level sets at different levels. The main focus of the latter paper was to find the optimal levels at which the additive model best depicts the density response, for which they employed a stagewise regression technique considered in [14]. Recently, [41] and [42] treated density- and simplex-valued responses, respectively, but in parametric models.

2. Bochner smooth backfitting. Throughout this paper we use the symbol $\mathbb{B}$ to denote Banach spaces and $\|\cdot\|$ for their norms. We use the symbol $\mathcal{B}(\mathbb{B})$ to denote the Borel σ-field of $\mathbb{B}$. For a set $U \in \mathcal{B}(\mathbb{B})$, we write $U \cap \mathcal{B}(\mathbb{B})$ for the σ-field $\{U \cap B : B \in \mathcal{B}(\mathbb{B})\}$ on U. We let $\mathbb{H}$ denote a separable Hilbert space equipped with an inner product $\langle\cdot,\cdot\rangle$. Let Leb_k denote the Lebesgue measure on $\mathbb{R}^k$.

2.1. *Examples of Hilbertian response.* Here, we introduce three Hilbert spaces. These are the spaces we consider for the response in our numerical study in Section 5.

L^2 and Hilbert–Sobolev spaces. For a set $U \in \mathcal{B}(\mathbb{R}^k)$, consider $L^2(U, U \cap \mathcal{B}(\mathbb{R}^k), \mathrm{Leb}_k)$ and a Hilbert–Sobolev space $W^{l,2}(U)$ for $l \in \mathbb{N}$. For these spaces $\mathbf{0}$ is a zero function, $(f(\cdot) \oplus g(\cdot))(\mathbf{u}) = f(\mathbf{u}) + g(\mathbf{u})$ and $(c \odot f(\cdot))(\mathbf{u}) = c \cdot f(\mathbf{u})$ for $c \in \mathbb{R}$. It is well known that these are separable Hilbert spaces.

Bayes–Hilbert spaces. Consider a space of probability density functions on $U \in \mathcal{B}(\mathbb{R}^k)$ with $\mathrm{Leb}_k(U) < \infty$. Let

$$\mathcal{M} = \big\{\nu : \nu \text{ is a } \sigma\text{-finite measure on } U \cap \mathcal{B}(\mathbb{R}^k) \text{ such that } \nu \ll \mathrm{Leb}_k \text{ and } \mathrm{Leb}_k \ll \nu\big\}.$$

For $\nu \in \mathcal{M}$, let $f_\nu = d\nu/d\mathrm{Leb}_k$. For $\nu, \lambda \in \mathcal{M}$ and $c \in \mathbb{R}$, define $\nu\lambda, \nu^c : U \cap \mathcal{B}(\mathbb{R}^k) \to [0,\infty]$ by $(\nu\lambda)(A) = \int_A f_\nu(\mathbf{u}) f_\lambda(\mathbf{u})\, d\mathbf{u}$ and $(\nu^c)(A) = \int_A (f_\nu(\mathbf{u}))^c\, d\mathbf{u}$, respectively. Then, $\nu\lambda, \nu^c \in \mathcal{M}$. For these measures, $f_{\nu\lambda} = f_\nu \cdot f_\lambda$ a.e. $[\mathrm{Leb}_k]$ and $f_{\nu^c} = (f_\nu)^c$ a.e. $[\mathrm{Leb}_k]$. Define

$$\mathfrak{B}^2\big(U, U \cap \mathcal{B}(\mathbb{R}^k), \mathrm{Leb}_k\big) = \Big\{[f_\nu] : \nu \in \mathcal{M}, \int_U \big(\log f_\nu(\mathbf{u})\big)^2\, d\mathbf{u} < \infty\Big\},$$

where $[f_\nu]$ denotes the class of all measurable functions $g : U \to [0,\infty]$ such that $g = C \cdot f_\nu$ a.e. $[\mathrm{Leb}_k]$ for some constant $C > 0$. Define $\oplus$ and $\odot$ on $\mathfrak{B}^2(U, U \cap \mathcal{B}(\mathbb{R}^k), \mathrm{Leb}_k)$ by $[f_\nu] \oplus [f_\lambda] = [f_{\nu\lambda}] = [f_\nu \cdot f_\lambda]$ and $c \odot [f_\nu] = [f_{\nu^c}] = [(f_\nu)^c]$, respectively. Also, define $\langle\cdot,\cdot\rangle$ by

$$\langle [f_\nu], [f_\lambda]\rangle = \frac{1}{2\mathrm{Leb}_k(U)} \int_{U^2} \log\Big(\frac{f_\nu(\mathbf{u})}{f_\nu(\mathbf{u}')}\Big) \log\Big(\frac{f_\lambda(\mathbf{u})}{f_\lambda(\mathbf{u}')}\Big)\, d\mathbf{u}\, d\mathbf{u}'.$$

Then, $\mathfrak{B}^2(U, U \cap \mathcal{B}(\mathbb{R}^k), \mathrm{Leb}_k)$ is a separable Hilbert space with $\mathbf{0} = [f_{\mathrm{Leb}_k}] = [1]$, as proved by [43].

Simplices. For $s > 0$, consider the space $\mathcal{S}_s^k = \{(v_1, \ldots, v_k) \in (0,s)^k : \sum_{j=1}^k v_j = s\}$. The case $s = 1$ corresponds to the discrete analogue of a Bayes–Hilbert space in the previous example. For $\mathbf{v}, \mathbf{w} \in \mathcal{S}_s^k$ and $c \in \mathbb{R}$, define $\oplus$ and $\odot$, respectively, by $\mathbf{v} \oplus \mathbf{w} = \big(\frac{sv_1w_1}{v_1w_1+\cdots+v_kw_k}, \ldots, \frac{sv_kw_k}{v_1w_1+\cdots+v_kw_k}\big)$ and $c \odot \mathbf{v} = \big(\frac{sv_1^c}{v_1^c+\cdots+v_k^c}, \ldots, \frac{sv_k^c}{v_1^c+\cdots+v_k^c}\big)$. Define $\langle \mathbf{v}, \mathbf{w}\rangle = (2k)^{-1} \sum_{j=1}^k \sum_{l=1}^k \log(v_j/v_l) \cdot \log(w_j/w_l)$. Then, $(\mathcal{S}_s^k, \oplus, \odot, \langle\cdot,\cdot\rangle)$ is a separable Hilbert space with $\mathbf{0} = (s/k, \ldots, s/k)$.

2.2. *Bochner integration.* Our method of estimating the additive model (1.1) is based on the representation of the conditional means of $\mathbf{m}_k(X_k)$ given X_j for $k \neq j$, in terms of the conditional densities of X_k given X_j. This involves integration of $\mathbf{m}_k(x_k)$ over x_k in the support of the corresponding conditional density. Since each component $\mathbf{m}_k$ is a $\mathbb{H}$-valued map, the conventional Lebesgue integration does not apply to the current problem. In this

subsection we study a notion of integration in a more general setting. Specifically, we consider integration of $\mathbb{B}$-valued maps, for which we use a recently introduced notion of Bochner integral. The notion has not been well studied in statistics. Our results below are familiar in Lebesgue integral theory, but their derivation for Bochner integrals requires substantial innovation.

Our definition of Bochner integral is for "strongly measurable" $\mathbb{B}$-valued maps. We briefly introduce it here. For more details, see [7]. Let $(\mathcal{Z}, \mathscr{A}, \mu)$ be a measure space. For a map $\mathbf{f}: \mathcal{Z} \to \mathbb{B}$, we let range($\mathbf{f}$) denote $\{\mathbf{f}(\mathbf{z}) : \mathbf{z} \in \mathcal{Z}\} \subset \mathbb{B}$.

DEFINITION 2.1. A map $\mathbf{f}: (\mathcal{Z}, \mathscr{A}, \mu) \to (\mathbb{B}, \mathcal{B}(\mathbb{B}))$ is called *strongly measurable* if it is $(\mathscr{A}, \mathcal{B}(\mathbb{B}))$-measurable and range($\mathbf{f}$) is separable.

An immediate example of strongly measurable map is μ-simple map. A map $\mathbf{f}: \mathcal{Z} \to \mathbb{B}$ is called μ-simple if $\mathbf{f}(\mathbf{z}) = \bigoplus_{i=1}^{n} 1_{A_i}(\mathbf{z}) \odot \mathbf{b}_i$ for some $\mathbf{b}_i \in \mathbb{B}$ and disjoint $A_i \in \mathscr{A}$ with $\mu(A_i) < \infty$. For such a μ-simple map its Bochner integral is defined by $\int \mathbf{f}(\mathbf{z})\, d\mu(\mathbf{z}) = \bigoplus_{i=1}^{n} \mu(A_i) \odot \mathbf{b}_i$. It can be shown that, if a map $\mathbf{f}$ is strongly measurable and $\|\mathbf{f}\|$ is Lebesgue integrable with respect to μ, then there exist μ-simple maps $\mathbf{f}_n$ such that $\mathbf{f}(\mathbf{z}) = \lim_{n\to\infty} \mathbf{f}_n(\mathbf{z})$ and $\|\mathbf{f}_n(\mathbf{z})\| \le \|\mathbf{f}(\mathbf{z})\|$ for all $\mathbf{z}$ and n.

DEFINITION 2.2. A map $\mathbf{f}: (\mathcal{Z}, \mathscr{A}, \mu) \to (\mathbb{B}, \mathcal{B}(\mathbb{B}))$ is called *Bochner integrable* if it is strongly measurable and $\|\mathbf{f}\|$ is Lebesgue integrable with respect to μ. In this case the Bochner integral of $\mathbf{f}$ is defined by $\int \mathbf{f}\, d\mu = \lim_{n\to\infty} \int \mathbf{f}_n\, d\mu$, where $\mathbf{f}_n$ is a sequence of μ-simple maps such that $\mathbf{f}(\mathbf{z}) = \lim_{n\to\infty} \mathbf{f}_n(\mathbf{z})$ and $\|\mathbf{f}_n(\mathbf{z})\| \le \|\mathbf{f}(\mathbf{z})\|$.

REMARK 2.1. There are two other notions of Bochner integrals of which we are aware. One is defined for $(\mathscr{A}, \mathcal{B}(\mathbb{B}))$-measurable maps, and the other is for μ-measurable maps. Both are not relevant for the component maps $\mathbf{m}_j$ in the model (1.1); see the online Supplementary Material S.1 [19] for details.

We present several properties of the Bochner integral that are fundamental in its statistical applications. For $1 \le p < \infty$, define

$$L^p\big((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{B}\big) = \Big\{ \mathbf{f}: \mathcal{Z} \to \mathbb{B} \Big| \mathbf{f} \text{ is strongly measurable and } \Big(\int_{\mathcal{Z}} \|\mathbf{f}(\mathbf{z})\|^p\, d\mu(\mathbf{z}) \Big)^{1/p} < \infty \Big\}.$$

We call it *Lebesgue–Bochner space*. If $\mathbf{f}$ is Bochner integrable with respect to μ, then $\mathbf{f} \in L^1((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{B})$. Note that $L^p((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{R})$ corresponds to the standard L^p space of real-valued functions. It is well known that $L^p((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{R})$ can be made into a Banach space by taking its quotient space $L^p((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{R})/\mathcal{N}$ with respect to the kernel $\mathcal{N}$ of its norm, $\mathcal{N} = \{f : f = 0 \text{ a.e. } [\mu]\}$. This also holds for $L^p((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{B})$. In particular, for $\mathcal{N} = \{\mathbf{f} : \mathbf{f} = \mathbf{0} \text{ a.e. } [\mu]\}$, the quotient space $L^2((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{H})/\mathcal{N}$ is a Hilbert space with an inner product $\langle \cdot, \cdot \rangle_\mu$ defined by $\langle \mathbf{f}, \mathbf{g} \rangle_\mu = \int_{\mathcal{Z}} \langle \mathbf{f}(\mathbf{z}), \mathbf{g}(\mathbf{z}) \rangle\, d\mu(\mathbf{z})$. We adopt the following convention throughout this paper.

CONVENTION 1. With slight abuse of notation, we write $L^p((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{B})$ for $L^p((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{B})/\mathcal{N}$. We also write simply $\mathbf{f} = \mathbf{g}$ for $\mathbf{f}$ and $\mathbf{g}$ with $\mathbf{f} = \mathbf{g}$ a.e. $[\mu]$ unless we need to specify the measure with respect to which the two maps agree almost everywhere. We say simply "measurable" for "strongly measurable" and "integrable" for "Bochner integrable" in the sense of Definition 2.2. We say "μ-integrable" in case we need to specify the underlying measure μ associated with Bochner integration.

For measure spaces $(\mathcal{Z}, \mathscr{A}, \mu)$ and $(\mathcal{W}, \mathscr{B}, \nu)$, let $\mathscr{A} \otimes \mathscr{B}$ denote the product σ-field and $\mu \otimes \nu$ denote a product measure on $\mathscr{A} \otimes \mathscr{B}$. For a $(\mathscr{A}, \mathscr{B})$-measurable mapping $\mathbf{T} : \mathcal{Z} \to \mathcal{W}$, we let $\mu\mathbf{T}^{-1}$ denote a measure on $(\mathcal{W}, \mathscr{B})$ defined by $\mu\mathbf{T}^{-1}(B) = \mu(\mathbf{T}^{-1}(B))$, $B \in \mathscr{B}$. For a probability space $(\Omega, \mathscr{F}, P)$ and a random element $\mathbf{Z} : (\Omega, \mathscr{F}, P) \to (\mathcal{Z}, \mathscr{A}, \mu)$ with σ-finite μ, we write $p_{\mathbf{Z}}$ for its density $dP\mathbf{Z}^{-1}/d\mu$ with respect to μ.

The following two propositions are the basic building blocks of our methodological and theoretical development to be presented later. They are also of interest in their own right. The results are very new in statistics although there are familiar versions in the Lebesgue integral theory. One may find versions of Proposition 2.1 for real-valued functions in standard textbooks; see Theorem 9.4.1 in [24], for example. For a Lebesgue integral version of Proposition 2.2, see Example 4.1.6 in [11]. The proofs of the propositions are in the online Supplementary Material S.4 and S.5.

PROPOSITION 2.1. *Let $(\Omega, \mathscr{F}, P)$ be a probability space and $(\mathcal{Z}, \mathscr{A}, \mu)$ be a σ-finite measure space. Let $\mathbf{Z} : \Omega \to \mathcal{Z}$ be a random element such that $P\mathbf{Z}^{-1} \ll \mu$ and $\mathbf{f} : \mathcal{Z} \to \mathbb{B}$ be a measurable map such that $\mathrm{E}(\|\mathbf{f}(\mathbf{Z})\|) < \infty$. Then, it holds that $\mathrm{E}(\mathbf{f}(\mathbf{Z})) = \int_{\mathcal{Z}} \mathbf{f}(\mathbf{z}) \odot p_{\mathbf{Z}}(\mathbf{z})\, d\mu(\mathbf{z})$.*

PROPOSITION 2.2. *Let $(\Omega, \mathscr{F}, P)$ be a probability space and $(\mathcal{Z}, \mathscr{A}, \mu)$ and $(\mathcal{W}, \mathscr{B}, \nu)$ be σ-finite measure spaces. Let $\mathbf{Z} : \Omega \to \mathcal{Z}$ and $\mathbf{W} : \Omega \to \mathcal{W}$ be random elements such that $P(\mathbf{Z}, \mathbf{W})^{-1} \ll \mu \otimes \nu$. Assume that $p_{\mathbf{W}}(\mathbf{w}) \in (0, \infty)$ for all $\mathbf{w} \in \mathcal{W}$. Let $\mathbf{f} : \mathcal{Z} \to \mathbb{B}$ be a measurable map such that $\mathrm{E}(\|\mathbf{f}(\mathbf{Z})\|) < \infty$. Define $\mathbf{g} : \mathcal{W} \to \mathbb{B}$ by*

$$\mathbf{g}(\mathbf{w}) = \int_{\mathcal{Z}} \mathbf{f}(\mathbf{z}) \odot \frac{p_{\mathbf{Z},\mathbf{W}}(\mathbf{z}, \mathbf{w})}{p_{\mathbf{W}}(\mathbf{w})}\, d\mu(\mathbf{z})$$

if $\mathbf{w} \in D_{\mathcal{W}} := \{\mathbf{w} \in \mathcal{W} : \int_{\mathcal{Z}} \|\mathbf{f}(\mathbf{z})\| p_{\mathbf{Z},\mathbf{W}}(\mathbf{z}, \mathbf{w})\, d\mu(\mathbf{z}) < \infty\}$ and $\mathbf{g}(\mathbf{w}) = \mathbf{g}_0(\mathbf{w})$ if $\mathbf{w} \notin D_{\mathcal{W}}$, where $\mathbf{g}_0 : \mathcal{W} \to \mathbb{B}$ is an arbitrary measurable map. Then, $\mathbf{g}$ is measurable, and $\mathbf{g}(\mathbf{W})$ is a version of $\mathrm{E}(\mathbf{f}(\mathbf{Z})|\mathbf{W})$.

REMARK 2.2. Some earlier uses of Bochner integration may be found in the literature. In [4], for example, Bochner integral is used to express the (conditional) mean of a $\mathbb{B}$-valued random element in probabilistic sense but not in terms of the densities of the involved random elements as we have done in Propositions 2.1 and 2.2. In [35] and [39], as other examples, Bochner integral is also introduced either as kernel mean in the context of reproducing kernel Hilbert space or as the expected value of a Hilbertian random element but is not used to develop a statistical methodology and its theory, as we have done in this paper.

2.3. *Lebesgue–Bochner spaces of additive maps.* We introduce some relevant spaces of $\mathbb{H}$-valued maps for the estimation of the additive model (1.1). For a probability space $(\Omega, \mathscr{F}, P)$ and a separable Hilbert space $\mathbb{H}$, let $\mathbf{Y} : \Omega \to \mathbb{H}$ be a response with $\mathrm{E}(\|\mathbf{Y}\|^2) < \infty$ and $\mathbf{X} : \Omega \to [0, 1]^d$ a d-variate predictor vector. We assume $P\mathbf{X}^{-1} \ll \mathrm{Leb}_d$. For simplicity, we write p, instead of $p_{\mathbf{X}}$, to denote its density $dP\mathbf{X}^{-1}/d\mathrm{Leb}_d$. We also write p_j for $dPX_j^{-1}/d\mathrm{Leb}_1$ and p_{jk} for $dP(X_j, X_k)^{-1}/d\mathrm{Leb}_2$.

The $\mathrm{E}(\mathbf{Y}|X_j)$ and $\mathrm{E}(\mathbf{Y}|\mathbf{X})$, respectively, are $(X_j^{-1}([0, 1] \cap \mathcal{B}(\mathbb{R})), \mathcal{B}(\mathbb{H}))$- and $(\mathbf{X}^{-1}([0, 1]^d \cap \mathcal{B}(\mathbb{R}^d)), \mathcal{B}(\mathbb{H}))$-measurable maps by definition. In general, for a measurable space $(\mathcal{Z}, \mathscr{A})$, a random element $\mathbf{V} : \Omega \to \mathbb{H}$ and a random element $\mathbf{Z} : \Omega \to \mathcal{Z}$, it holds that $\mathbf{V}$ is $(\mathbf{Z}^{-1}(\mathscr{A}), \mathcal{B}(\mathbb{H}))$-measurable if and only if there exists a measurable map $\mathbf{h} : \mathcal{Z} \to \mathbb{H}$ such that $\mathbf{V} = \mathbf{h}(\mathbf{Z})$; see Lemma 1.13 in [21], for example. Thus, there exist measurable maps $\mathbf{h}_j : [0, 1] \to \mathbb{H}$ and $\mathbf{h} : [0, 1]^d \to \mathbb{H}$ such that $\mathrm{E}(\mathbf{Y}|X_j) = \mathbf{h}_j(X_j)$ and $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{h}(\mathbf{X})$. For such measurable maps, we define $\mathrm{E}(\mathbf{Y}|X_j = \cdot) = \mathbf{h}_j$ and $\mathrm{E}(\mathbf{Y}|\mathbf{X} = \cdot) = \mathbf{h}$.

Let $\mathbf{m} : [0,1]^d \to \mathbb{H}$ be defined by $\mathbf{m}(\mathbf{x}) = \mathbf{m}_0 \oplus \bigoplus_{j=1}^d \mathbf{m}_j(x_j)$. We note that $\mathbf{m} = \mathrm{E}(\mathbf{Y}|\mathbf{X} = \cdot)$. As the space where $\mathrm{E}(\mathbf{Y}|\mathbf{X} = \cdot)$ belongs, we consider

$$L_2^{\mathbb{H}}(p) := L^2\big(([0,1]^d, [0,1]^d \cap \mathcal{B}(\mathbb{R}^d), P\mathbf{X}^{-1}), \mathbb{H}\big)$$

and endow $L_2^{\mathbb{H}}(p)$ with the norm $\|\cdot\|_2$ defined by

$$\|\mathbf{f}\|_2^2 = \int_{[0,1]^d} \|\mathbf{f}(\mathbf{x})\|^2 \, dP\mathbf{X}^{-1}(\mathbf{x}) = \int_{[0,1]^d} \|\mathbf{f}(\mathbf{x})\|^2 p(\mathbf{x}) \, d\mathbf{x}.$$

As subspaces of $L_2^{\mathbb{H}}(p)$, define

$$L_2^{\mathbb{H}}(p_j) := \big\{\mathbf{f} \in L_2^{\mathbb{H}}(p) : \exists \text{ a univariate map } \mathbf{f}_j \text{ such that } \mathbf{f}(\mathbf{x}) = \mathbf{f}_j(x_j)\big\}.$$

We note that $L_2^{\mathbb{H}}(p_j)$ depends on p only through its marginalization p_j since, for $\mathbf{f} \in L_2^{\mathbb{H}}(p_j)$, it holds that

$$\int_{[0,1]^d} \|\mathbf{f}(\mathbf{x})\|^2 p(\mathbf{x}) \, d\mathbf{x} = \int_0^1 \|\mathbf{f}_j(x_j)\|^2 p_j(x_j) \, dx_j,$$

where $\mathbf{f}_j$ is a univariate map such that $\mathbf{f}(\mathbf{x}) = \mathbf{f}_j(x_j)$. Let $S^{\mathbb{H}}(p)$ be the sum-space defined by

$$S^{\mathbb{H}}(p) = \left\{ \bigoplus_{j=1}^d \mathbf{f}_j : \mathbf{f}_j \in L_2^{\mathbb{H}}(p_j), 1 \le j \le d \right\} \subset L_2^{\mathbb{H}}(p).$$

To define empirical versions of $L_2^{\mathbb{H}}(p)$, $L_2^{\mathbb{H}}(p_j)$ and $S^{\mathbb{H}}(p)$, we let $K : \mathbb{R} \to [0,\infty)$ be a baseline kernel function. Throughout this paper we assume that K vanishes on $\mathbb{R} \setminus [-1,1]$ and satisfies $\int_{-1}^1 K(u)\,du = 1$. For a bandwidth $h > 0$, we write $K_h(u) = K(u/h)/h$. Define a normalized kernel $K_h(u,v)$ by

$$K_h(u,v) = \frac{K_h(u-v)}{\int_0^1 K_h(t-v)\,dt},$$

whenever $\int_0^1 K_h(t-v)\,dt > 0$, and we set $K_h(u,v) = 0$ otherwise. This type of kernel function has been used in the smooth backfitting literature; see [31, 46] and [26], for example. Note that $\int_0^1 K_h(u,v)\,du = 1$ for all $v \in [0,1]$.

Suppose that we observe $(\mathbf{Y}_i, \mathbf{X}_i)$, $1 \le i \le n$ which follow the model (1.1). We estimate $p_j(x_j)$ and $p_{jk}(x_j, x_k)$ by

$$\hat{p}_j(x_j) = \frac{1}{n}\sum_{i=1}^n K_{h_j}(x_j, X_{ij}), \qquad \hat{p}_{jk}(x_j,x_k) = \frac{1}{n}\sum_{i=1}^n K_{h_j}(x_j, X_{ij}) K_{h_k}(x_k, X_{ik}),$$

respectively, where X_{ij} denotes the jth entry of $\mathbf{X}_i$. Here, we allow the bandwidths h_j to be different for different j. Because of the normalization in defining $K_h(\cdot,\cdot)$, it holds that

$$\int_0^1 \hat{p}_j(x_j)\,dx_j = 1, \qquad \int_0^1 \hat{p}_{jk}(x_j,x_k)\,dx_k = \hat{p}_j(x_j).$$

Let $\hat{p}$ be the multivariate kernel density estimator of p defined by $\hat{p}(\mathbf{x}) = n^{-1} \times \sum_{i=1}^n \prod_{j=1}^d K_{h_j}(x_j, X_{ij})$. The density estimator $\hat{p}$ also has the marginalization properties as p:

$$\int_{[0,1]^{d-1}} \hat{p}(\mathbf{x})\,d\mathbf{x}_{-j} = \hat{p}_j(x_j), \qquad \int_{[0,1]^{d-2}} \hat{p}(\mathbf{x})\,d\mathbf{x}_{-j,k} = \hat{p}_{jk}(x_j,x_k)$$

for $1 \le j \ne k \le d$, where $\mathbf{x}_{-j}$ and $\mathbf{x}_{-j,k}$ denote the respective $(d-1)$- and $(d-2)$-vector resulting from omitting x_j and (x_j,x_k) in $\mathbf{x} = (x_1,\ldots,x_d)$.

Now, define a measure $\hat{P}\mathbf{X}^{-1}$ on $[0,1]^d \cap \mathcal{B}(\mathbb{R}^d)$ by $\hat{P}\mathbf{X}^{-1}(B) = \int_B \hat{p}(\mathbf{x})\,d\mathbf{x}$. With this measure we define $L_2^{\mathbb{H}}(\hat{p})$ and $L_2^{\mathbb{H}}(\hat{p}_j)$ as $L_2^{\mathbb{H}}(p)$ and $L_2^{\mathbb{H}}(p_j)$ with $P\mathbf{X}^{-1}$ in the definition of $L_2^{\mathbb{H}}(p)$ being replaced by $\hat{P}\mathbf{X}^{-1}$. We endow $L_2^{\mathbb{H}}(\hat{p})$ with the norm $\|\cdot\|_{2,n}$ defined by

$$\|\mathbf{f}\|_{2,n}^2 = \int_{[0,1]^d} \|\mathbf{f}(\mathbf{x})\|^2\, d\hat{P}\mathbf{X}^{-1}(\mathbf{x}) = \int_{[0,1]^d} \|\mathbf{f}(\mathbf{x})\|^2 \hat{p}(\mathbf{x})\,d\mathbf{x}.$$

Also, define an analogue of $S^{\mathbb{H}}(p)$ by

$$S^{\mathbb{H}}(\hat{p}) = \left\{ \bigoplus_{j=1}^d \mathbf{f}_j : \mathbf{f}_j \in L_2^{\mathbb{H}}(\hat{p}_j), 1 \le j \le d \right\} \subset L_2^{\mathbb{H}}(\hat{p}).$$

CONVENTION 2. It is often convenient to treat $\mathbf{f}$ in $L_2^{\mathbb{H}}(p_j)$ or in $L_2^{\mathbb{H}}(\hat{p}_j)$ as a univariate map and write $\mathbf{f}(x_j)$ instead of $\mathbf{f}(\mathbf{x})$. This convention is particularly useful in (2.4), for example. Conversely, we may embed a univariate map $\mathbf{f} : [0,1] \to \mathbb{H}$ into $L_2^{\mathbb{H}}(p_j)$ or $L_2^{\mathbb{H}}(\hat{p}_j)$ by considering its version $\mathbf{f}_j^*$ defined by $\mathbf{f}_j^*(\mathbf{x}) = \mathbf{f}(x_j)$ for $\mathbf{x} \in [0,1]^d$. We take the above convention throughout this paper. With this convention we may put $\mathbf{m}_j$ into $L_2^{\mathbb{H}}(p_j)$ if $\mathrm{E}(\|\mathbf{m}_j(X_j)\|^2) < \infty$ and $\mathbf{m}$ into $S^{\mathbb{H}}(p)$ if all $\mathbf{m}_j \in L_2^{\mathbb{H}}(p_j)$.

2.4. *Bochner integral equations and backfitting algorithm.* In this section we describe the estimation of the component maps $\mathbf{m}_j$ in the model (1.1) using Bochner integrals. Throughout this paper we assume that $\mathbf{m}_j \in L_2^{\mathbb{H}}(p_j)$ for all $1 \le j \le d$. Furthermore, we make the following assumptions:

CONDITION (A). For all $1 \le j \ne k \le d$ and $x_j \in [0,1]$, $p_j(x_j) > 0$,

$$\int_0^1 \frac{p_{jk}^2(x_j, x_k)}{p_k(x_k)}\,dx_k < \infty \quad \text{and} \quad \int_{[0,1]^2} \frac{p_{jk}^2(x_j, x_k)}{p_j(x_j)p_k(x_k)}\,dx_j\,dx_k < \infty.$$

We also use the following analogue of the condition (A) for $\hat{p}_j$ and $\hat{p}_{jk}$:

CONDITION (S). For all $1 \le j \ne k \le d$ and $x_j \in [0,1]$, $\hat{p}_j(x_j) > 0$,

$$\int_0^1 \frac{\hat{p}_{jk}^2(x_j, x_k)}{\hat{p}_k(x_k)}\,dx_k < \infty \quad \text{and} \quad \int_{[0,1]^2} \frac{\hat{p}_{jk}^2(x_j, x_k)}{\hat{p}_j(x_j)\hat{p}_k(x_k)}\,dx_j\,dx_k < \infty.$$

We note that the condition (S) always holds if

$$c := \max_{1\le j\le d} h_j^{-1} \max\left\{ X_j^{(1)}, 1 - X_j^{(n)}, \max_{1\le i\le n-1} \{(X_j^{(i+1)} - X_j^{(i)})/2\} \right\} < 1,$$

K is bounded and $\inf_{u\in[-c,c]} K(u) > 0$, where $X_j^{(1)} < \cdots < X_j^{(n)}$ are the order statistics of $(X_{ij} : 1 \le i \le n)$.

Under the condition (A) we also get that

$$\int_0^1 \|\mathbf{m}_k(x_k)\|\, p_{jk}(x_j, x_k)\,dx_k < \infty \tag{2.1}$$

for all $x_j \in [0,1]$ and $1 \le j \ne k \le d$. The property (2.1) is a simple consequence of an application of Hölder's inequality. Then, by Proposition 2.2, $\mathrm{E}(\mathbf{m}_k(X_k)|X_j) = \int_0^1 \mathbf{m}_k(x_k) \odot$

$[p_{jk}(X_j, x_k)/p_j(X_j)]\,dx_k$. Thus, by the definition of $\mathrm{E}(\mathbf{Y}|X_j = \cdot)$ and from the model (1.1), we get

$$\mathrm{E}(\mathbf{Y}|X_j = x_j) = \mathbf{m}_0 \oplus \mathbf{m}_j(x_j) \oplus \bigoplus_{k\neq j} \int_0^1 \mathbf{m}_k(x_k) \odot \frac{p_{jk}(x_j, x_k)}{p_j(x_j)}\,dx_k, \tag{2.2}$$
$$1 \le j \le d.$$

For the identifiability of $\mathbf{m}_j$ in the model, we put the constraints $\mathrm{E}(\mathbf{m}_j(X_j)) = \mathbf{0}$, $1 \le j \le d$. By Proposition 2.1 the constraints are equivalent to

$$\int_0^1 \mathbf{m}_j(x_j) \odot p_j(x_j)\,dx_j = \mathbf{0}, \quad 1 \le j \le d. \tag{2.3}$$

The constraints entail $\mathbf{m}_0 = \mathrm{E}(\mathbf{Y})$.

Now, we describe the estimation of $\mathbf{m}_j$ based on the Bochner integral equations at (2.2). We estimate $\mathrm{E}(\mathbf{Y}|X_j = x_j)$ by the Nadaraya–Watson-type estimator

$$\tilde{\mathbf{m}}_j(x_j) = \big[\hat{p}_j(x_j)^{-1} n^{-1}\big] \odot \bigoplus_{i=1}^n K_{h_j}(x_j, X_{ij}) \odot \mathbf{Y}_i$$

and $\mathrm{E}(\mathbf{Y})$ by the sample mean $\bar{\mathbf{Y}} = n^{-1} \odot \bigoplus_{i=1}^n \mathbf{Y}_i$. Let $\ominus$ be defined by $\mathbf{b}_1 \ominus \mathbf{b}_2 = \mathbf{b}_1 \oplus (-1 \odot \mathbf{b}_2)$. We solve the estimated system of Bochner integral equations

$$\hat{\mathbf{m}}_j(x_j) = \tilde{\mathbf{m}}_j(x_j) \ominus \bar{\mathbf{Y}} \ominus \bigoplus_{k\neq j} \int_0^1 \hat{\mathbf{m}}_k(x_k) \odot \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\,dx_k, \tag{2.4}$$
$$1 \le j \le d$$

for $(\hat{\mathbf{m}}_1, \dots, \hat{\mathbf{m}}_d)$ in the space of d-tuple maps $\{(\mathbf{f}_1, \dots, \mathbf{f}_d) : \mathbf{f}_j \in L_2^{\mathbb{H}}(\hat{p}_j), 1 \le j \le d\}$, subject to the constraints

$$\int_0^1 \hat{\mathbf{m}}_j(x_j) \odot \hat{p}_j(x_j)\,dx_j = \mathbf{0}, \quad 1 \le j \le d. \tag{2.5}$$

We note that the Bochner integrals at (2.4) are well defined for $\hat{\mathbf{m}}_j \in L_2^{\mathbb{H}}(\hat{p}_j)$ under the condition (S).

In the next section we will show that there exists a solution $(\hat{\mathbf{m}}_j : 1 \le j \le d)$ of (2.4) with $\hat{\mathbf{m}}_j \in L_2^{\mathbb{H}}(\hat{p}_j)$ and that their sum $\bigoplus_{j=1}^d \hat{\mathbf{m}}_j$ is unique, only under the condition (S). The estimator of the regression map $\mathbf{m} = \mathrm{E}(\mathbf{Y}|\mathbf{X} = \cdot) : [0,1]^d \to \mathbb{H}$ is defined by $\hat{\mathbf{m}}$, where $\hat{\mathbf{m}}(\mathbf{x}) = \bar{\mathbf{Y}} \oplus \bigoplus_{j=1}^d \hat{\mathbf{m}}_j(x_j)$. For the estimator $\hat{\mathbf{m}}$ we will also prove that the constraints (2.5) uniquely determine the component tuple $(\hat{\mathbf{m}}_j : 1 \le j \le d)$ under some additional assumption. Our estimator of $(\mathbf{m}_1, \dots, \mathbf{m}_d)$ is then the solution $(\hat{\mathbf{m}}_1, \dots, \hat{\mathbf{m}}_d)$. We call $\hat{\mathbf{m}}$ and $\hat{\mathbf{m}}_j$ *Bochner smooth backfitting estimators*, or *B-SBF estimators* in short, and the system of equations (2.4) *Bochner smooth backfitting equation*, or *B-SBF equation* in short. Our approach guarantees that $\hat{\mathbf{m}}_j(x_j)$ and $\hat{\mathbf{m}}(\mathbf{x})$ belong to $\mathbb{H}$, the space of the true values of the maps $\mathbf{m}_j$ and $\mathbf{m}$ as well as the values of $\mathbf{Y}$.

To solve (2.4), we take an initial estimator $(\hat{\mathbf{m}}_1^{[0]}, \dots, \hat{\mathbf{m}}_d^{[0]}) \in \prod_{j=1}^d L_2^{\mathbb{H}}(\hat{p}_j)$ that satisfies the constraints (2.5). We update the estimator $(\hat{\mathbf{m}}_1^{[r]}, \dots, \hat{\mathbf{m}}_d^{[r]})$ for $r \ge 1$ by

$$\begin{aligned}\hat{\mathbf{m}}_j^{[r]}(x_j) = \tilde{\mathbf{m}}_j(x_j) \ominus \bar{\mathbf{Y}} \ominus \bigoplus_{k<j} \int_0^1 \hat{\mathbf{m}}_k^{[r]}(x_k) \odot \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\,dx_k \\ \ominus \bigoplus_{k>j} \int_0^1 \hat{\mathbf{m}}_k^{[r-1]}(x_k) \odot \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\,dx_k, \quad 1 \le j \le d.\end{aligned} \tag{2.6}$$

Then, all subsequent updates $(\hat{\mathbf{m}}_1^{[r]}, \ldots, \hat{\mathbf{m}}_d^{[r]})$ for $r \geq 1$ are in $\prod_{j=1}^d L_2^{\mathbb{H}}(\hat{p}_j)$ under the condition (S) and satisfy (2.5) due to the normalization property $\int_0^1 K_{h_j}(u, \cdot)\, du \equiv 1$ on $[0, 1]$. We let $\hat{\mathbf{m}}^{[r]}(\mathbf{x}) = \bar{\mathbf{Y}} \oplus \bigoplus_{j=1}^d \hat{\mathbf{m}}_j^{[r]}(x_j)$. We call (2.6) *Bochner smooth backfitting algorithm*, or *B-SBF algorithm* in short. In the next section we will show that the B-SBF algorithm converges always in $\|\cdot\|_{2,n}$ norm under the condition (S). We will also show that it converges in $\|\cdot\|_2$ norm with probability tending to one under weak conditions on p, K and h_j.

2.5. *Practical implementation.* Bochner integrals are defined in an abstract way. In this subsection we present an innovative way of implementing the B-SBF algorithm using the usual numerical integration techniques. The key idea is to use the fact that, for any measure space $(\mathcal{Z}, \mathscr{A}, \mu)$,

$$\text{(Bochner)} \int_{\mathcal{Z}} f(\mathbf{z}) \odot \mathbf{b}\, d\mu(\mathbf{z}) = \text{(Lebesgue)} \int_{\mathcal{Z}} f(\mathbf{z})\, d\mu(\mathbf{z}) \odot \mathbf{b}, \tag{2.7}$$

where f is a real-valued integrable function on $\mathcal{Z}$ and $\mathbf{b}$ is a constant in a Banach space. Suppose that we choose

$$\hat{\mathbf{m}}_j^{[0]}(x_j) = n^{-1} \odot \bigoplus_{i=1}^n w_{ij}^{[0]}(x_j) \odot \mathbf{Y}_i$$

with the weights $w_{ij}^{[0]}(x_j) \in \mathbb{R}$ satisfying $\int_0^1 w_{ij}^{[0]}(x_j) \hat{p}_j(x_j)\, dx_j = 0$. This is not a restriction since we can take $w_{ij}^{[0]} \equiv 0$ for all $1 \leq j \leq d$ and $1 \leq i \leq n$. Define

$$\begin{aligned} w_{ij}^{[r]}(x_j) = {} & \frac{K_{h_j}(x_j, X_{ij})}{\hat{p}_j(x_j)} - 1 - \sum_{k<j} \int_0^1 w_{ik}^{[r]}(x_k) \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\, dx_k \\ & - \sum_{k>j} \int_0^1 w_{ik}^{[r-1]}(x_k) \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\, dx_k, \quad r \geq 1. \end{aligned}$$

Then, by using (2.7) we may express (2.6) as follows.

$$\hat{\mathbf{m}}_j^{[r]}(x_j) = n^{-1} \odot \bigoplus_{i=1}^n w_{ij}^{[r]}(x_j) \odot \mathbf{Y}_i, \quad 1 \leq j \leq d. \tag{2.8}$$

Thus, it turns out that the algorithm (2.6) reduces to a simple iteration scheme that updates the weight functions $w_{ij}^{[r]}$ based on Lebesgue integrals.

The equation (2.8) reveals that $\hat{\mathbf{m}}_j^{[r]}$ for $r \geq 1$ are linear smoothers if the initial $\hat{\mathbf{m}}_j^{[0]}$ are. It also demonstrates explicitly that the values of $\hat{\mathbf{m}}_j^{[r]}(x_j)$ for each x_j belong to the space of the values of $\mathbf{Y}_i$ and $\mathbf{m}_j(x_j)$. The idea of using (2.7) in the evaluation of Bochner integrals appears to be important in the analysis of more general object-oriented data belonging to a Banach space. One may develop a similar idea for nonparametric structural regression dealing with various types of random objects.

3. Existence and algorithm convergence.

3.1. *Projection operators.* Our theory for the existence of the B-SBF estimators and the convergence of the B-SBF algorithm rely heavily on the theory of projection operators that map $L_2^{\mathbb{H}}(p)$ to $L_2^{\mathbb{H}}(p_j)$, or $L_2^{\mathbb{H}}(\hat{p})$ to $L_2^{\mathbb{H}}(\hat{p}_j)$. Let $L_2^{\mathbb{B}}(p)$, $L_2^{\mathbb{B}}(\hat{p})$, $L_2^{\mathbb{B}}(p_j)$ and $L_2^{\mathbb{B}}(\hat{p}_j)$ be defined as $L_2^{\mathbb{H}}(p)$, $L_2^{\mathbb{H}}(\hat{p})$, $L_2^{\mathbb{H}}(p_j)$ and $L_2^{\mathbb{H}}(\hat{p}_j)$ but with $\mathbb{H}$ being replaced by a Banach space

$\mathbb{B}$. We start with a proposition that characterizes $L_2^{\mathbb{B}}(p_j)$ and $L_2^{\mathbb{B}}(\hat{p}_j)$, respectively, as closed subspaces of $L_2^{\mathbb{B}}(p)$ and $L_2^{\mathbb{B}}(\hat{p})$. These topological properties in the case where $\mathbb{B} = \mathbb{H}$ are essential to defining relevant projection operators. We write $\mathbb{B}_1 \le \mathbb{B}_2$ if $\mathbb{B}_1$ is a *closed* subspace of a Banach space $\mathbb{B}_2$. Also, define a σ-field $\mathscr{B}_j = \{[0,1]^{j-1} \times B_j \times [0,1]^{d-j} : B_j \in [0,1] \cap \mathcal{B}(\mathbb{R})\}$ on $[0,1]^d$. We let $\mathscr{B}_j^*$ denote the smallest σ-field such that $\mathscr{B}_j \subset \mathscr{B}_j^*$ and $\{B \in [0,1]^d \cap \mathcal{B}(\mathbb{R}^d) : P\mathbf{X}^{-1}(B) = 1\} \subset \mathscr{B}_j^*$. Lemma S.6 in the Supplementary Material asserts that $L_2^{\mathbb{B}}(p_j) = L^2(([0,1]^d, \mathscr{B}_j^*, P\mathbf{X}^{-1}), \mathbb{B})$ and $L_2^{\mathbb{B}}(\hat{p}_j) = L^2(([0,1]^d, \mathscr{B}_j^*, \hat{P}\mathbf{X}^{-1}), \mathbb{B})$. The following proposition is immediate from this and the fact that a complete subspace of a metric space is closed.

PROPOSITION 3.1. *It holds that* $L_2^{\mathbb{B}}(p_j) \le L_2^{\mathbb{B}}(p)$ *and* $L_2^{\mathbb{B}}(\hat{p}_j) \le L_2^{\mathbb{B}}(\hat{p})$.

We define the operators $\pi_j : L_2^{\mathbb{H}}(p) \to L_2^{\mathbb{H}}(p_j)$ by

$$\pi_j(\mathbf{f})(x_j) = \int_{[0,1]^{d-1}} \mathbf{f}(\mathbf{x}) \odot \frac{p(\mathbf{x})}{p_j(x_j)}\, d\mathbf{x}_{-j}$$

for $x_j \in D_j(\mathbf{f}) := \{x_j \in [0,1] : \int_{[0,1]^{d-1}} \|\mathbf{f}(\mathbf{x})\| p(\mathbf{x})\, d\mathbf{x}_{-j} < \infty\}$ and simply let $\pi_j(\mathbf{f})(x_j) = \mathbf{0}$ for $x_j \notin D_j(\mathbf{f})$. Likewise, we define the operators $\hat{\pi}_j : L_2^{\mathbb{H}}(\hat{p}) \to L_2^{\mathbb{H}}(\hat{p}_j)$ with p and p_j being replaced by $\hat{p}$ and $\hat{p}_j$, respectively. The following proposition demonstrates that both π_j and $\hat{\pi}_j$ are projection operators on the respective spaces.

PROPOSITION 3.2. *If* $p_j(x_j) > 0$ *for all* $x_j \in [0,1]$, *then,* $\pi_j(\mathbf{f}) \in L_2^{\mathbb{H}}(p_j)$ *and* $\mathbf{f} - \pi_j(\mathbf{f}) \perp L_2^{\mathbb{H}}(p_j)$ *for all* $\mathbf{f} \in L_2^{\mathbb{H}}(p)$. *Also, if* $\hat{p}_j(x_j) > 0$ *for all* $x_j \in [0,1]$, *then,* $\hat{\pi}_j(\mathbf{f}) \in L_2^{\mathbb{H}}(\hat{p}_j)$ *and* $\mathbf{f} - \hat{\pi}_j(\mathbf{f}) \perp L_2^{\mathbb{H}}(\hat{p}_j)$ *for all* $\mathbf{f} \in L_2^{\mathbb{H}}(\hat{p})$.

For Banach spaces $\mathbb{B}_1$ and $\mathbb{B}_2$, let $\mathcal{L}(\mathbb{B}_1, \mathbb{B}_2)$ denote the space of all bounded linear operators that map $\mathbb{B}_1$ to $\mathbb{B}_2$. We write simply $\mathcal{L}(\mathbb{B})$ for $\mathcal{L}(\mathbb{B}, \mathbb{B})$. Let $\pi_j|L_2^{\mathbb{H}}(p_k) : L_2^{\mathbb{H}}(p_k) \to L_2^{\mathbb{H}}(p_j)$ denote the operator π_j restricted to $L_2^{\mathbb{H}}(p_k)$ for $k \ne j$. Under the condition (A), $\pi_j|L_2^{\mathbb{H}}(p_k)$ are integral operators with the kernel $\mathbf{k}_{jk} : [0,1]^d \times [0,1]^d \to \mathcal{L}(\mathbb{H})$ defined by

$$\mathbf{k}_{jk}(\mathbf{u}, \mathbf{v})(\mathbf{h}) = \mathbf{h} \odot \frac{p_{jk}(u_j, v_k)}{p_j(u_j) p_k(v_k)}.$$

To see this, we note that the condition (A) implies $\int_{[0,1]^{d-1}} \|\mathbf{f}_k(\mathbf{x})\| p(\mathbf{x})\, d\mathbf{x}_{-j} < \infty$ for all $x_j \in [0,1]$ if $\mathbf{f}_k \in L_2^{\mathbb{H}}(p_k)$, so that $D_j(\mathbf{f}_k) = [0,1]$ for all $\mathbf{f}_k \in L_2^{\mathbb{H}}(p_k)$. Thus, it holds that

$$\begin{aligned}\pi_j(\mathbf{f}_k)(u_j) &= \int_{[0,1]^d} \mathbf{f}_k(\mathbf{x}) \odot \frac{p_{jk}(u_j, x_k)}{p_j(u_j) p_k(x_k)}\, dP\mathbf{X}^{-1}(\mathbf{x}) \\ &= \int_{[0,1]^d} \mathbf{k}_{jk}(\mathbf{u}, \mathbf{x})\big(\mathbf{f}_k(\mathbf{x})\big)\, dP\mathbf{X}^{-1}(\mathbf{x}).\end{aligned}$$

Similarly, under the condition (S), $\hat{\pi}_j|L_2^{\mathbb{H}}(\hat{p}_k)$ are integral operators with the kernel $\hat{\mathbf{k}}_{jk} : [0,1]^d \times [0,1]^d \to \mathcal{L}(\mathbb{H})$ defined by $\hat{\mathbf{k}}_{jk}(\mathbf{u}, \mathbf{v})(\mathbf{h}) = \mathbf{h} \odot \frac{\hat{p}_{jk}(u_j, v_k)}{\hat{p}_j(u_j)\hat{p}_k(v_k)}$.

3.2. *Compactness of projection operators.* In the case where $\mathbb{H} = \mathbb{R}$, a common approach to establishing the existence of the SBF estimators and the convergence of the SBF algorithm is to prove that $\pi_j|L_2^{\mathbb{H}}(p_k)$ or $\hat{\pi}_j|L_2^{\mathbb{H}}(\hat{p}_k)$ for all $1 \le j \ne k \le d$ are *compact operators*; see [31] or a more recent [33], for example. Indeed, it follows from Proposition A.4.2 in [2] that if $\pi_j|L_2^{\mathbb{H}}(p_k)$ for all $1 \le j \ne k \le d$ are compact, then

$$S^{\mathbb{H}}(p) \le L_2^{\mathbb{H}}(p). \tag{3.1}$$

Moreover, according to Corollary 4.3 in [45], (3.1) implies

$$\|T\|_{\mathcal{L}(S^{\mathbb{H}}(p))} < 1, \tag{3.2}$$

where T is an operator in $\mathcal{L}(S^{\mathbb{H}}(p))$ defined by $T = (I - \pi_d) \circ \cdots \circ (I - \pi_1)$ and I is the identity operator. The same properties hold for $S^{\mathbb{H}}(\hat{p})$ and for $\hat{T}$, defined in the same way as T with π_j being replaced by $\hat{\pi}_j$, if $\hat{\pi}_j|L_2^{\mathbb{H}}(\hat{p}_k)$ are compact. The two properties at (3.1) and (3.2) and their empirical versions are essential to the existence of the B-SBF estimators and the convergence of the B-SBF algorithm.

The compactness of $\pi_j|L_2^{\mathbb{H}}(p_k)$ or $\hat{\pi}_j|L_2^{\mathbb{H}}(\hat{p}_k)$ has been unknown when $\mathbb{H} \neq \mathbb{R}$. Some sufficient conditions for the compactness of integral operators, defined on Lebesgue–Bochner spaces of "μ-measurable maps" were studied by [5] and [44], among others. But the case for "strongly measurable maps" which are relevant in statistical applications and on which our theoretical development is based, has never been studied. Below we present two general theorems in the latter case. The first one gives a sufficient condition for compactness, and the second is about noncompactness for certain integral operators. The two theorems have important implications in our theoretical development, while they are also of interest in their own right. Their proofs are in the online Supplementary Material S.7 and S.8.

In the following two theorems, $(\mathcal{Z}, \mathscr{A}, \mu)$ and $(\mathcal{W}, \mathscr{B}, \nu)$ are measure spaces, and $\mathbb{B}_1$ and $\mathbb{B}_2$ are Banach spaces. We denote by $\|\cdot\|_{\mathcal{L}(\mathbb{B}_1,\mathbb{B}_2)}$ the operator norm of $\mathcal{L}(\mathbb{B}_1, \mathbb{B}_2)$. Let $1 < p, q < \infty$ satisfy $p^{-1} + q^{-1} = 1$. Let $\mathbf{k} : \mathcal{Z} \times \mathcal{W} \to \mathcal{L}(\mathbb{B}_1, \mathbb{B}_2)$ be a *measurable* map such that $\int_{\mathcal{Z}\times\mathcal{W}} \|\mathbf{k}(\mathbf{z}, \mathbf{w})\|^q_{\mathcal{L}(\mathbb{B}_1,\mathbb{B}_2)}\, d\mu \otimes \nu(\mathbf{z}, \mathbf{w}) < \infty$. For $\mathbf{f} \in L^p((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{B}_1)$, define $L(\mathbf{f}) : \mathcal{W} \to \mathbb{B}_2$ by

$$L(\mathbf{f})(\mathbf{w}) = \begin{cases} \displaystyle\int_{\mathcal{Z}} \mathbf{k}(\mathbf{z}, \mathbf{w})\big(\mathbf{f}(\mathbf{z})\big)\, d\mu(\mathbf{z}) & \text{if } \mathbf{w} \in D_{\mathcal{W}}, \\ L_0(\mathbf{f})(\mathbf{w}) & \text{otherwise,} \end{cases} \tag{3.3}$$

where $D_{\mathcal{W}} = \{\mathbf{w} \in \mathcal{W} : \int_{\mathcal{Z}} \|\mathbf{k}(\mathbf{z}, \mathbf{w})\|^q_{\mathcal{L}(\mathbb{B}_1,\mathbb{B}_2)}\, d\mu(\mathbf{z}) < \infty\}$ and L_0 is any linear map from $L^p((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{B}_1)$ to $\{\mathbf{g} : \mathcal{W} \to \mathbb{B}_2 | \mathbf{g} \text{ is measurable}\}$. Finally, we let $\mathcal{C}(\mathbb{B}_1, \mathbb{B}_2)$ denote the space of all compact operators from $\mathbb{B}_1$ to $\mathbb{B}_2$.

THEOREM 3.1. *The mapping* $\mathbf{f} \mapsto L(\mathbf{f})$ *with* $L(\mathbf{f})$ *at* (3.3) *defines a bounded linear operator* $L : L^p((\mathcal{Z}, \mathscr{A}, \mu), \mathbb{B}_1) \to L^q((\mathcal{W}, \mathscr{B}, \nu), \mathbb{B}_2)$. *Furthermore, if* $\operatorname{range}(\mathbf{k}) \subset \mathcal{C}(\mathbb{B}_1, \mathbb{B}_2)$, *then* L *is compact.*

One may compare the above theorem with those in Lebesgue integral theory; see Proposition 4.7 in [8], for example. In the application of Theorem 3.1 to $L = \pi_j|L_2^{\mathbb{H}}(p_k)$ or to $L = \hat{\pi}_j|L_2^{\mathbb{H}}(\hat{p}_k)$ with $\mathbf{k} = \mathbf{k}_{jk}$ or $\mathbf{k} = \hat{\mathbf{k}}_{jk}$, respectively, we may prove that $\mathbf{k}_{jk}(\mathbf{u}, \mathbf{v})$ and $\hat{\mathbf{k}}_{jk}(\mathbf{u}, \mathbf{v})$ belong to $\mathcal{C}(\mathbb{H}, \mathbb{H})$ for all $\mathbf{u}, \mathbf{v} \in [0, 1]^d$ under the conditions (A) and (S), respectively, if $\mathbb{H}$ is finite dimensional. Furthermore, $\mathbf{k}_{jk}$ and $\hat{\mathbf{k}}_{jk}$ are measurable since $\mathcal{C}(\mathbb{H}, \mathbb{H})$ is separable, according to a lemma in [15].

COROLLARY 3.1. *Suppose that* $\mathbb{H}$ *is finite dimensional. Then,* $\pi_j|L_2^{\mathbb{H}}(p_k)$ *and* $\hat{\pi}_j|L_2^{\mathbb{H}}(\hat{p}_k)$ *for all* $1 \le j \neq k \le d$ *are compact under the conditions* (A) *and* (S), *respectively.*

At the beginning we thought that $\pi_j|L_2^{\mathbb{H}}(p_k)$ and $\hat{\pi}_j|L_2^{\mathbb{H}}(\hat{p}_k)$ might be also compact when $\mathbb{H}$ is infinite dimensional. However, we find that the conclusion of Corollary 3.1 is not valid for infinite dimensional $\mathbb{H}$ which follows from an application of the following theorem:

THEOREM 3.2. *Suppose that $\mu(\mathcal{Z}) < \infty$. Let $\kappa : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}$ be a measurable function such that $\int_{\mathcal{Z}\times\mathcal{W}} |\kappa(\mathbf{z}, \mathbf{w})|^q \, d\mu \otimes \nu(\mathbf{z}, \mathbf{w}) < \infty$ and $0 < \int_{\mathcal{W}} |\int_{\mathcal{Z}} \kappa(\mathbf{z}, \mathbf{w}) \, d\mu(\mathbf{z})|^q \, d\nu(\mathbf{w}) < \infty$. Let $C \in \mathcal{L}(\mathbb{B}_1, \mathbb{B}_2)$ be a noncompact operator. Then, L at (3.3) with $\mathbf{k}(\mathbf{z}, \mathbf{w})(\mathbf{h}) = \kappa(\mathbf{z}, \mathbf{w}) \odot C(\mathbf{h})$ defines a bounded linear noncompact operator.*

For the application of Theorem 3.2 to $\pi_j|L_2^{\mathbb{H}}(p_k)$ and $\hat{\pi}_j|L_2^{\mathbb{H}}(\hat{p}_k)$, we take $\kappa_{jk} : [0,1]^d \times [0,1]^d \to \mathbb{R}$ such that $\kappa_{jk}(\mathbf{u}, \mathbf{v}) = p_{jk}(u_j, v_k)/(p_j(u_j)p_k(v_k))$ for κ in the theorem, and the identity operator $I_{\mathbb{H}} : \mathbb{H} \to \mathbb{H}$ for C. Note that $I_{\mathbb{H}}$ is noncompact since the unit closed balls in infinite-dimensional Hilbert spaces are not compact. Also, it holds that

$$0 < \int_{[0,1]^d} \left| \int_{[0,1]^d} \kappa_{jk}(\mathbf{u}, \mathbf{v}) \, dP\mathbf{X}^{-1}(\mathbf{u}) \right|^2 dP\mathbf{X}^{-1}(\mathbf{v}) = 1 < \infty$$

under the condition (A). The same holds for $\hat{\kappa}_{jk}$ defined by $\hat{\kappa}_{jk}(\mathbf{u}, \mathbf{v}) = \hat{p}_{jk}(u_j, v_k)/(\hat{p}_j(u_j)\hat{p}_k(v_k))$ under the condition (S). Therefore, surprisingly we have the following corollary of Theorem 3.2:

COROLLARY 3.2. *Suppose that $\mathbb{H}$ is infinite dimensional. Then, for all $1 \le j \ne k \le d$, $\pi_j|L_2^{\mathbb{H}}(p_k)$ and $\hat{\pi}_j|L_2^{\mathbb{H}}(\hat{p}_k)$ are noncompact under the conditions (A) and (S), respectively.*

3.3. *Existence of B-SBF estimators.* Noncompactness of $\pi_j|L_2^{\mathbb{H}}(p_k)$ and $\hat{\pi}_j|L_2^{\mathbb{H}}(\hat{p}_k)$ raises a major difficulty in proving (3.1) and (3.2) and their empirical versions since the earlier proofs of them for the case $\mathbb{H} = \mathbb{R}$ use the compactness of the respective projection operators. To tackle the difficulty, we rely on the following equivalence result, which is a direct consequence of applying Lemma S.7 in the Supplementary Material and Proposition 3.2. We state the result only for the empirical versions $S^{\mathbb{H}}(\hat{p})$ and $\hat{T}$, but an obvious analogue holds for $S^{\mathbb{H}}(p)$ and T as well. Let $\overline{S^{\mathbb{H}}(\hat{p})}$ denote the closure of $S^{\mathbb{H}}(\hat{p})$.

PROPOSITION 3.3. *Assume that $\hat{p}_j(x_j) > 0$ for all $x_j \in [0,1]$ and $1 \le j \le d$. Then, the followings are equivalent: (a) $S^{\mathbb{H}}(\hat{p}) \le L_2^{\mathbb{H}}(\hat{p})$; (b) $\|\hat{T}\|_{\mathcal{L}(\overline{S^{\mathbb{H}}(\hat{p})})} < 1$; (c) there exists $\hat{c} > 0$ such that, for all $\mathbf{f} \in S^{\mathbb{H}}(\hat{p})$, there exist $\mathbf{f}_1 \in L_2^{\mathbb{H}}(\hat{p}_1), \ldots, \mathbf{f}_d \in L_2^{\mathbb{H}}(\hat{p}_d)$ satisfying $\bigoplus_{j=1}^d \mathbf{f}_j = \mathbf{f}$ and $\sum_{j=1}^d \|\mathbf{f}_j\|_{2,n}^2 \le \hat{c}\|\mathbf{f}\|_{2,n}^2$.*

The above proposition does not say that one of (a)–(c) is true which has never been known. With an innovative use of Corollary 3.1, we are able to show that the "compatibility" condition (c) for sum-maps holds.

THEOREM 3.3. *Assume that the condition (S) holds. Then, the statements in Proposition 3.3 are true.*

We are now ready to discuss the existence of the B-SBF estimators. For this we consider an objective functional $\hat{F} : S^{\mathbb{H}}(\hat{p}) \to \mathbb{R}$ defined by

$$\hat{F}(\mathbf{f}) = \int_{[0,1]^d} n^{-1} \sum_{i=1}^n \|\mathbf{Y}_i \ominus \mathbf{f}(\mathbf{x})\|^2 \cdot \prod_{j=1}^d K_{h_j}(x_j, X_{ij}) \, d\mathbf{x}.$$

The map $\hat{F}$ is well-defined since $\hat{F}(\mathbf{f}) \le 2(\max_{1\le i\le n} \|\mathbf{Y}_i\|^2 + \|\mathbf{f}\|_{2,n}^2) < \infty$. Now, the Gâteaux differential at $\mathbf{f} \in S^{\mathbb{H}}(\hat{p})$ is given by

$$D\hat{F}(\mathbf{f})(\mathbf{g}) := -2 \int_{[0,1]^d} n^{-1} \sum_{i=1}^n \langle \mathbf{Y}_i \ominus \mathbf{f}(\mathbf{x}), \mathbf{g}(\mathbf{x}) \rangle \prod_{j=1}^d K_{h_j}(x_j, X_{ij}) \, d\mathbf{x}. \tag{3.4}$$

Clearly, $D\hat{F}(\mathbf{f}) : S^{\mathbb{H}}(\hat{p}) \to \mathbb{R}$ is a linear operator. It is also bounded, which we may verify by using Hölder's inequality. Hence, $\hat{F}$ is Gâteaux differentiable.

THEOREM 3.4. *Assume that the condition (S) holds. Then, there exists a unique solution* $\hat{\mathbf{m}} \in S^{\mathbb{H}}(\hat{p})$ *of (2.4). Furthermore, if* $\hat{p}(\mathbf{x}) > 0$ *for all* $\mathbf{x} \in [0,1]^d$, *then there exists a unique decomposition* $\hat{\mathbf{m}} = \bar{\mathbf{Y}} \oplus \hat{\mathbf{m}}_1 \oplus \cdots \oplus \hat{\mathbf{m}}_d$ *with* $\hat{\mathbf{m}}_j \in L_2^{\mathbb{H}}(\hat{p}_j)$ *that satisfy (2.5).*

REMARK 3.1. In the proof of the second part of Theorem 3.4 in the online Supplementary Material, it is worthwhile to note that $\bigoplus_{j=1}^d \hat{\mathbf{g}}_j = \mathbf{0}$ a.e. $[\hat{P}\mathbf{X}^{-1}]$ does not always imply $\bigoplus_{j=1}^d \hat{\mathbf{g}}_j = \mathbf{0}$ a.e. $[\hat{P}X_j^{-1} \otimes \hat{P}\mathbf{X}_{-j}^{-1}]$ under the condition (S) only. Thus, one can not argue directly from the condition (S) and (2.5) that $\mathbf{0} = \int_{[0,1]^{d-1}} \bigoplus_{k=1}^d \hat{\mathbf{g}}_k(x_k) \odot \hat{p}_{\mathbf{X}_{-j}}(\mathbf{x}_{-j})\, d\mathbf{x}_{-j} = \hat{\mathbf{g}}_j(x_j)$ a.e. $[\hat{P}X_j^{-1}]$.

3.4. *Convergence of B-SBF algorithm.* In this subsection we establish the convergence of the B-SBF algorithm (2.6). We first consider convergence in the empirical norm, $\|\cdot\|_{2,n}$, for fixed n and given observations $(\mathbf{X}_i, \mathbf{Y}_i)$, $1 \le i \le n$. Then, we study convergence in $\|\cdot\|_2$ norm, where we let n diverge to infinity. We note that all works in the smooth backfitting literature treated only the latter asymptotic version for $\mathbb{H} = \mathbb{R}$. The following theorem is a nonasymptotic version of the convergence of the B-SBF algorithm.

THEOREM 3.5. *Assume that the condition (S) holds. Then,* $\|\hat{T}\|_{\mathcal{L}(S^{\mathbb{H}}(\hat{p}))} < 1$, *and there exists* $\hat{c}' > 0$ *such that*

$$\int_{[0,1]^d} \|\hat{\mathbf{m}}(\mathbf{x}) \ominus \hat{\mathbf{m}}^{[r]}(\mathbf{x})\|^2 \hat{p}(\mathbf{x})\, d\mathbf{x} \le \hat{c}' \|\hat{T}\|^{2r}_{\mathcal{L}(S^{\mathbb{H}}(\hat{p}))} \quad \text{for all } r \ge 0.$$

We now turn to the asymptotic version of the convergence of the B-SBF algorithm in $\|\cdot\|_2$ norm. For this we need the following additional conditions:

CONDITION (B).

(B1) $\mathrm{E}(\|\mathbf{Y}\|^\alpha) < \infty$ for some $\alpha > 2$ and $\mathrm{E}(\|\mathbf{Y}\|^2 | X_j = \cdot)$ is bounded on $[0,1]$ for $1 \le j \le d$.

(B2) p is bounded away from zero and infinity on $[0,1]^d$, and p_{jk} are continuous on $[0,1]^2$ for $1 \le j \ne k \le d$.

(B3) K is Lipschitz continuous and $\int_{-1}^0 K(u)\, du \wedge \int_0^1 K(u)\, du > 0$.

(B4) $h_j = o(1)$ as $n \to \infty$, and there exist $0 < c_j < (\alpha - 2)/\alpha$ with $c_j + c_k < 1$ for all $1 \le j \ne k \le d$ such that $n^{c_j} h_j$ are bounded away from zero for all $1 \le j \le d$.

(B5) $\max_{1 \le j \le d} \|\hat{\mathbf{m}}_j^{[0]}\|_{2,n}^2 < C$ for an absolute constant $0 < C < \infty$.

THEOREM 3.6. *Assume that the condition (B) holds. Then, there exist constants* $c > 0$ *and* $\gamma \in (0,1)$ *such that*

$$\lim_{n \to \infty} P\left(\max_{1 \le j \le d} \int_0^1 \|\hat{\mathbf{m}}_j(x_j) \ominus \hat{\mathbf{m}}_j^{[r]}(x_j)\|^2 p_j(x_j)\, dx_j \le c\gamma^r \text{ for all } r \ge 0 \right) = 1.$$

Theorem 3.6 is about the L^2-convergence of the B-SBF algorithm, like all other results in the literature on smooth backfitting for $\mathbb{H} = \mathbb{R}$. Here, we add an almost everywhere convergence result which is also of interest. We note that the theorem implies $\sum_{r=1}^\infty \int_0^1 \|\hat{\mathbf{m}}_j(x_j) \ominus \hat{\mathbf{m}}_j^{[r]}(x_j)\|^2 p_j(x_j)\, dx_j < \infty$ with probability tending to one. This entails that, with probability tending to one, $\sum_{r=1}^\infty \|\hat{\mathbf{m}}_j(x_j) \ominus \hat{\mathbf{m}}_j^{[r]}(x_j)\|^2 p_j(x_j) < \infty$ a.e. $x_j \in [0,1]$ with respect to Leb_1, which gives the following corollary:

COROLLARY 3.3. *Under the condition of Theorem 3.6,*

$$\lim_{n\to\infty} P\big(\hat{\mathbf{m}}_j^{[r]}(x_j) \to \hat{\mathbf{m}}_j(x_j) \text{ as } r \to \infty \text{ a.e. } x_j \in [0,1] \text{ w.r.t. } \mathrm{Leb}_1\big) = 1$$

for $1 \le j \le d$.

4. Asymptotic properties.

4.1. *Rates of convergence.* Below we collect the assumptions for our asymptotic theory:

CONDITION (C).

(C1) $\mathrm{E}(\|\boldsymbol{\epsilon}\|^\alpha) < \infty$ for some $\alpha > 5/2$ and $\mathrm{E}(\|\boldsymbol{\epsilon}\|^2 | X_j = \cdot)$ is bounded on $[0,1]$ for $1 \le j \le d$.

(C2) The true maps $\mathbf{m}_j$ for $1 \le j \le d$ are twice continuously Fréchet differentiable on $[0,1]$.

(C3) The condition (B2) holds. Also, p_{jk} are C^1 on $[0,1]^2$ for $1 \le j \ne k \le d$.

(C4) The condition (B3) holds. Also, $\int_{-1}^1 u K(u)\, du = 0$.

(C5) $n^{1/5} h_j \to \alpha_j$ for some constant $\alpha_j > 0$, $1 \le j \le d$.

The conditions on $\boldsymbol{\epsilon}$ and the Fréchet differentiability of the maps $\mathbf{m}_j : [0,1] \to \mathbb{H}$, respectively, are natural generalizations of the usual conditions on Euclidean errors and the smoothness assumptions on real-valued functions. In the theory we need functional calculus for Fréchet derivatives and Bochner integrals. Other assumptions on the baseline kernel K and the density p are typical in the kernel smoothing theory.

Let $I_j = [2h_j, 1 - 2h_j]$ and I_j^c denote its complement in $[0,1]$. The following theorem demonstrates that our estimators achieve the univariate error rates:

THEOREM 4.1. *Assume that the condition (C) holds. Then, for* $1 \le j \le d$, (i) $\|\hat{\mathbf{m}}_j(x_j) \ominus \mathbf{m}_j(x_j)\| = O_p(n^{-2/5})$ *for* $x_j \in I_j$ *and* $\|\hat{\mathbf{m}}_j(x_j) \ominus \mathbf{m}_j(x_j)\| = O_p(n^{-1/5})$ *for* $x_j \in I_j^c$; (ii) $\int_{I_j} \|\hat{\mathbf{m}}_j(x_j) \ominus \mathbf{m}_j(x_j)\|^2 p_j(x_j)\, dx_j = O_p(n^{-4/5})$ *and* $\int_0^1 \|\hat{\mathbf{m}}_j(x_j) \ominus \mathbf{m}_j(x_j)\|^2 \times p_j(x_j)\, dx_j = O_p(n^{-3/5})$; (iii) $\sup_{x_j \in I_j} \|\hat{\mathbf{m}}_j(x_j) \ominus \mathbf{m}_j(x_j)\| = O_p(n^{-2/5}\sqrt{\log n})$ *and* $\sup_{x_j \in [0,1]} \|\hat{\mathbf{m}}_j(x_j) \ominus \mathbf{m}_j(x_j)\| = O_p(n^{-1/5})$.

4.2. *Asymptotic distribution and asymptotic independence.* Recall that, for a mean-zero random element $\mathbf{Z} : \Omega \to \mathbb{H}$, its covariance operator $C : \mathbb{H} \to \mathbb{H}$ is characterized by

$$\langle C(\mathbf{h}), \mathbf{g} \rangle = \mathrm{E}\big(\langle \mathbf{Z}, \mathbf{h} \rangle \cdot \langle \mathbf{Z}, \mathbf{g} \rangle\big), \quad \mathbf{h}, \mathbf{g} \in \mathbb{H}.$$

Also, recall that a $\mathbb{H}$-valued random element $\mathbf{Z}$ is called Gaussian if $\langle \mathbf{Z}, \mathbf{h} \rangle$ is normally distributed for any $\mathbf{h} \in \mathbb{H}$. We denote a Gaussian random element with mean zero and covariance operator C, by $\mathbf{G}(\mathbf{0}, C)$.

Let $\{\mathbf{e}_l\}_{l=1}^L$ be an orthonormal basis of $\mathbb{H}$, where we allow $L = \infty$ for infinite-dimensional $\mathbb{H}$. Define

$$a_{j,kl}(x_j) = \alpha_j^{-1} p_j(x_j)^{-1} \int_{-1}^1 K^2(u)\, du \cdot \mathrm{E}\big(\langle \boldsymbol{\epsilon}, \mathbf{e}_k \rangle \cdot \langle \boldsymbol{\epsilon}, \mathbf{e}_l \rangle | X_j = x_j\big)$$

for α_j defined at (C5). Let $C_{j,x_j} : \mathbb{H} \to \mathbb{H}$ be a covariance operator characterized by

$$\langle C_{j,x_j}(\mathbf{h}), \mathbf{e}_k \rangle = \sum_{l=1}^L \langle \mathbf{h}, \mathbf{e}_l \rangle \cdot a_{j,kl}(x_j).$$

Define $\tilde{\mathbf{m}}_j^A(x_j) = [\hat{p}_j(x_j)^{-1} n^{-1}] \odot \bigoplus_{i=1}^n K_{h_j}(x_j, X_{ij}) \odot \boldsymbol{\epsilon}_i$. The following theorem plays an important role in determining the distributions of $\hat{\mathbf{m}}_j(x_j)$:

THEOREM 4.2. *Fix* $\mathbf{x} = (x_1, \ldots, x_d) \in (0,1)^d$. *Assume that the condition* (C5) *holds, that* K *is bounded, and that, for all* $1 \le j \ne k \le d$, (i) $\mathrm{E}(\|\boldsymbol{\epsilon}\|^\alpha) < \infty$ *for some* $\alpha > 2$, $\mathrm{E}(\|\boldsymbol{\epsilon}\|^\alpha \mid X_j = \cdot), \mathrm{E}(\langle \boldsymbol{\epsilon}, \mathbf{e}_l \rangle \cdot \langle \boldsymbol{\epsilon}, \mathbf{e}_{l'} \rangle \mid X_j = \cdot, X_k = \cdot)$ *and* p_{jk} *are bounded on a neighborhood of* x_j, *of* (x_j, x_k) *and of* (x_j, x_k), *respectively, and* $\mathrm{E}(\langle \boldsymbol{\epsilon}, \mathbf{e}_l \rangle \cdot \langle \boldsymbol{\epsilon}, \mathbf{e}_{l'} \rangle \mid X_j = \cdot)$, *for all* l *and* l', *are continuous on a common neighborhood of* x_j; (ii) p_j *is continuous on a neighborhood of* x_j *and* $p_j(x_j) > 0$. *Then,* $(n^{2/5} \odot \tilde{\mathbf{m}}_j^A(x_j) : 1 \le j \le d)$ *converges in distribution to* $(\mathbf{G}(\mathbf{0}, C_{j,x_j}) : 1 \le j \le d)$, *where* $\mathbf{G}(\mathbf{0}, C_{j,x_j})$ *are independent.*

Now, we are ready to present a theorem that demonstrates the asymptotic distribution and independence of our estimators of the component maps $\mathbf{m}_j$. In addition to (C), we need the following condition:

CONDITION (D). For all l, l' and $1 \le j \ne k \le d$, the followings hold:

(D1) $\mathrm{E}(\|\boldsymbol{\epsilon}\|^\alpha \mid X_j = \cdot)$ and $\mathrm{E}(\langle \boldsymbol{\epsilon}, \mathbf{e}_l \rangle \cdot \langle \boldsymbol{\epsilon}, \mathbf{e}_{l'} \rangle \mid X_j = \cdot, X_k = \cdot)$ are bounded on $[0,1]$ and $[0,1]^2$, respectively, for the constant α in (C1), and $\mathrm{E}(\langle \boldsymbol{\epsilon}, \mathbf{e}_l \rangle \cdot \langle \boldsymbol{\epsilon}, \mathbf{e}_{l'} \rangle \mid X_j = \cdot)$ are continuous on $[0,1]$.

(D2) $\partial p(\mathbf{x})/\partial x_j$ exist and are bounded on $[0,1]^d$.

To state the theorem, we need to introduce more terminologies. For a twice Fréchet differentiable $\mathbf{f} : [0,1] \to \mathbb{H}$, we let $D\mathbf{f} : [0,1] \to \mathcal{L}(\mathbb{R}, \mathbb{H})$ denote its first Fréchet derivative and $D^2\mathbf{f} : [0,1] \to \mathcal{L}(\mathbb{R}, \mathcal{L}(\mathbb{R}, \mathbb{H}))$ its second Fréchet derivative. Let p_j' denote the first derivative of p_j and define

$$\boldsymbol{\delta}_j(x_j) = \left[\frac{p_j'(x_j)}{p_j(x_j)} \cdot \int_{-1}^1 u^2 K(u)\, du \right] \odot D\mathbf{m}_j(x_j)(1),$$

$$\boldsymbol{\delta}_{jk}(x_j, x_k) = \left[\frac{\partial p_{jk}(x_j, x_k)/\partial x_k}{p_{jk}(x_j, x_k)} \cdot \int_{-1}^1 u^2 K(u)\, dt \right] \odot D\mathbf{m}_k(x_k)(1),$$

$$\tilde{\boldsymbol{\Delta}}_j(x_j) = \alpha_j^2 \odot \boldsymbol{\delta}_j(x_j) \oplus \bigoplus_{k \ne j} \int_0^1 \boldsymbol{\delta}_{jk}(x_j, x_k) \odot \left[\alpha_k^2 \frac{p_{jk}(x_j, x_k)}{p_j(x_j)} \right] dx_k.$$

Let $(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_d) \in \prod_{j=1}^d L_2^{\mathbb{H}}(p_j)$ be a solution of the system of equations

$$\boldsymbol{\Delta}_j(x_j) = \tilde{\boldsymbol{\Delta}}_j(x_j) \ominus \bigoplus_{k \ne j} \int_0^1 \boldsymbol{\Delta}_k(x_k) \odot \frac{p_{jk}(x_j, x_k)}{p_j(x_j)}\, dx_k, \quad 1 \le j \le d \tag{4.1}$$

satisfying the constraints

$$\int_0^1 \boldsymbol{\Delta}_j(x_j) \odot p_j(x_j)\, dx_j = \alpha_j^2 \odot \int_0^1 \boldsymbol{\delta}_j(x_j) \odot p_j(x_j)\, dx_j, \quad 1 \le j \le d. \tag{4.2}$$

Below in Theorem 4.3, we prove that the equation (4.1) subject to (4.2) has a unique solution. Define $\mathbf{c}_j(x_j) = \frac{1}{2} \int_{-1}^1 u^2 K(u)\, du \odot D^2\mathbf{m}_j(x_j)(1)(1)$ and $\boldsymbol{\Theta}_j(x_j) = \alpha_j^2 \odot \mathbf{c}_j(x_j) \oplus \boldsymbol{\Delta}_j(x_j)$.

THEOREM 4.3. *Assume that the conditions* (C) *and* (D) *hold. Then, the solution of* (4.1) *subject to* (4.2) *is unique. Furthermore, for a.e. fixed* $\mathbf{x} \in (0,1)^d$, $(n^{2/5} \odot (\hat{\mathbf{m}}_j(x_j) \ominus \mathbf{m}_j(x_j)) : 1 \le j \le d)$ *converges in distribution to* $(\boldsymbol{\Theta}_j(x_j) \oplus \mathbf{G}(\mathbf{0}, C_{j,x_j}) : 1 \le j \le d)$, *where* $\boldsymbol{\Theta}_j(x_j) \oplus \mathbf{G}(\mathbf{0}, C_{j,x_j})$ *are independent. Moreover,* $n^{2/5} \odot (\hat{\mathbf{m}}(\mathbf{x}) \ominus \mathbf{m}(\mathbf{x}))$ *converges in distribution to* $\bigoplus_{j=1}^d \boldsymbol{\Theta}_j(x_j) \oplus \mathbf{G}(\mathbf{0}, \sum_{j=1}^d C_{j,x_j})$.

Let $\hat{\mathbf{m}}_j^{\text{ora}}$ be the oracle estimator of $\mathbf{m}_j$ under the knowledge of all other component maps $\mathbf{m}_k, k \neq j$. Using Theorem 4.2, we may prove

$$n^{2/5} \odot (\hat{\mathbf{m}}_j^{\text{ora}}(x_j) \ominus \mathbf{m}_j(x_j)) \xrightarrow{d} \alpha_j^2 \odot [\boldsymbol{\delta}_j(x_j) \oplus \mathbf{c}_j(x_j)] \oplus \mathbf{G}(\mathbf{0}, C_{j,x_j}).$$

Therefore, $\hat{\mathbf{m}}_j$ and $\hat{\mathbf{m}}_j^{\text{ora}}$ have the same asymptotic covariance operator but differ in their asymptotic biases. The difference of asymptotic biases is $[\alpha_j^2 \odot \boldsymbol{\delta}_j(x_j)] \ominus \boldsymbol{\Delta}_j(x_j) =: \boldsymbol{\beta}_j(x_j)$, and it holds that $\int_0^1 \boldsymbol{\beta}_j(x_j) \odot p_j(x_j)\,dx_j = \mathbf{0}$ by (4.2).

5. Numerical study. In the simulation and real data examples presented here, we took Epanechnikov kernel $K(u) = (3/4)(1-u^2)I(|u|<1)$. We chose the initial estimators

$$\hat{\mathbf{m}}_j^{[0]}(x_j) = n^{-1} \odot \bigoplus_{i=1}^{n} \left(\frac{K_{h_j}(x_j, X_{ij})}{\hat{p}_j(x_j)} - 1 \right) \odot \mathbf{Y}_i =: n^{-1} \odot \bigoplus_{i=1}^{n} w_{ij}^{[0]}(x_j) \odot \mathbf{Y}_i,$$

so that they satisfy $\int_0^1 w_{ij}^{[0]}(x_j)\hat{p}_j(x_j)\,dx_j = 0$. However, any other choices with $\int_0^1 \hat{\mathbf{m}}_j^{[0]}(x_j) \odot \hat{p}_j(x_j)\,dx_j = \mathbf{0}$ would work. For example, one may choose $\hat{\mathbf{m}}_j^{[0]} \equiv \mathbf{0}$ for all $1 \le j \le d$. For the convergence criterion of the B-SBF algorithm, we set

$$\max_{1 \le j \le d} \int_0^1 \|\hat{\mathbf{m}}_j^{[r]}(x_j) \ominus \hat{\mathbf{m}}_j^{[r-1]}(x_j)\|^2 \hat{p}_j(x_j)\,dx_j < 10^{-4}.$$

With the above criterion we found that the B-SBF algorithm converged within eight iterations in all the cases of the numerical studies to be presented in Sections 5.2–5.4. The R codes used in this numerical study are available in a Github repository at https://github.com/jeong-min-jeon/Add-Reg-Hilbert-Res.

5.1. *Bandwidth selection.* Searching for the bandwidths h_j on a full-dimensional grid is not feasible when d is large. One way often adopted in multivariate smoothing is to set $h_1 = \cdots = h_d$ and perform a one-dimensional grid search. Obviously, this is not desirable since it ignores different degrees of smoothness for different target functions. Recently, [16] and [17] used a method called "bandwidth shrinkage". The method first selects $\hat{h}_j$ for each j that is good for estimating marginal regression function of X_j and then tunes $c > 0$ for $(c\hat{h}_1, \ldots, c\hat{h}_d)$. The latter method also searches bandwidths on a restricted class of options.

Here, we suggest a new scheme called "CBS" (coordinate-wise bandwidth selection) based on cross-validation. We used the CBS method, as described below, in our numerical study. Let $\text{CV}(h_1, \ldots, h_d)$ denote a cross-validatory criterion for bandwidths $h_1, \ldots, h_d$.

CBS ALGORITHM. Take a grid $\mathcal{G} = \prod_{j=1}^d \{g_{j1}, \ldots, g_{jL_j}\}$. Choose an initial bandwidth $h_j^{(0)}$ from $\{g_{j1}, \ldots, g_{jL_j}\}$ for $1 \le j \le d$. For $t = 1, 2, \ldots$, find

$$h_j^{(t)} = \underset{g_j \in \{g_{j1}, \ldots, g_{jL_j}\}}{\arg\min} \text{CV}(h_1^{(t)}, \ldots, h_{j-1}^{(t)}, g_j, h_{j+1}^{(t-1)}, \ldots, h_d^{(t-1)}), \quad 1 \le j \le d.$$

Repeat the procedure until $(h_1^{(t)}, \ldots, h_d^{(t)}) = (h_1^{(t-1)}, \ldots, h_d^{(t-1)})$.

We chose $\mathcal{G} = \prod_{j=1}^d \{a_j + 0.01 \times k : k = 0, \ldots, 20\}$ in our simulation and $\mathcal{G} = \prod_{j=1}^d \{a_j + 0.005 \times k : k = 0, \ldots, 100\}$ in the real data examples, for some small values a_j that satisfy $c < 1$ for c defined in Section 2.4 and used a 10-fold cross-validation. Let

$$T = \min\{t \ge 1 : (h_1^{(t)}, \ldots, h_d^{(t)}) = (h_1^{(t-1)}, \ldots, h_d^{(t-1)})\}.$$

We note that T is finite since the grid size is finite. In our numerical work the algorithm converged very fast. In all cases $T \le 4$. We also note that $(h_1^{(T)}, \dots, h_d^{(T)})$ is a coordinate-wise minimum that satisfies

$$\mathrm{CV}(h_1^{(T)}, \dots, h_d^{(T)}) = \min_j \min_{g_j} \mathrm{CV}(h_1^{(T)}, \dots, h_{j-1}^{(T)}, g_j, h_{j+1}^{(T)}, \dots, h_d^{(T)}).$$

A bandwidth selected by the CBS algorithm does not always match with a bandwidth selected by the full-dimensional grid search. However, we found that they coincided in most cases in our numerical study; see Table 1. Based on this observation, we suggest to consider the CBS algorithm in high dimension as a promising solution to the infeasibility of the full-dimensional search.

5.2. *Simulation study with density response.* We considered the case where $Y(\cdot)$ is a probability density on a domain $U \in \mathcal{B}(\mathbb{R})$ such that $\mathbf{Y} := [Y(\cdot)] \in \mathfrak{B}^2(U, U \cap \mathcal{B}(\mathbb{R}), \mathrm{Leb}_1)$. In this case, simply writing $w_{i,j,r}(x_j) = n^{-1} w_{ij}^{[r]}(x_j)$ for brevity, we get

$$\hat{\mathbf{m}}_j^{[r]}(x_j) = \left[\left(\int_U \prod_{i=1}^n Y_i(u)^{w_{i,j,r}(x_j)}\, du \right)^{-1} \prod_{i=1}^n Y_i(\cdot)^{w_{i,j,r}(x_j)} \right],$$

$$\bar{\mathbf{Y}} \oplus \bigoplus_{j=1}^d \hat{\mathbf{m}}_j^{[r]}(x_j) = \left[\left(\int_U \prod_{i=1}^n Y_i(u)^{n^{-1}+\sum_{j=1}^d w_{i,j,r}(x_j)}\, du \right)^{-1} \times \prod_{i=1}^n Y_i(\cdot)^{n^{-1}+\sum_{j=1}^d w_{i,j,r}(x_j)} \right], \tag{5.1}$$

whenever the denominators are nonzero and finite. We predicted $[Y(\cdot)]$ at $\mathbf{X} = \mathbf{x}$ for an out-of-sample $(\mathbf{X}, Y(\cdot))$ by $\bar{\mathbf{Y}} \oplus \bigoplus_{j=1}^d \hat{\mathbf{m}}_j^{[r]}(x_j)$ given in the above formula (5.1). We note that

TABLE 1

Average computing times in minutes with a personal PC, Inter(R) Xeon(R) CPU 23-1245 v3@3.50 GHz, and percentages of the cases where the bandwidth from the CBS algorithm (CBS) coincided with that from the full-dimensional search (Full), based on $M = 100$ pseudo samples. MSPE ratio was (MSPE with 'Full' bandwidth)/(MSPE with 'CBS' bandwidth). In the computation of MSPE ratio according to (5.5), the cases where CBS = Full were deleted

Scenario	d	n	Computing time			CBS = Full	MSPE Ratio
			CBS	Full	Ratio		
Additive	2	100	0.02	0.07	3.50	97%	1.00
(5.2)		400	0.38	1.29	3.39	99%	0.99
	3	100	0.06	2.71	45.17	94%	1.02
		400	1.41	65.12	46.18	97%	1.00
Non-Add. I	2	100	0.02	0.07	3.50	100%	–
(5.3)		400	0.39	1.29	3.31	100%	–
	3	100	0.06	2.72	45.33	94%	0.98
		400	1.38	66.16	47.94	93%	0.99
Non-Add. II	2	100	0.02	0.06	3.00	91%	1.00
(5.4)		400	0.34	1.17	3.44	98%	1.01
	3	100	0.05	2.34	46.80	83%	1.00
		400	1.27	58.62	46.16	99%	1.00

the denominators are nonzero and finite for all $w_{ij}^{[r]}(x_j) \in \mathbb{R}$ if $Y_i(\cdot)$'s are essentially bounded away from zero and infinity on U with $\mathrm{Leb}_1(U) < \infty$. In this simulation study our focus is to demonstrate that: (i) the CBS algorithm for bandwidth selection works well, and (ii) the prediction based on the proposed estimators $\hat{\mathbf{m}}_j$ and $\hat{\mathbf{m}}$ is valid for small sample sizes, avoiding the curse of dimensionality.

We generated $Y(\cdot)$ on $U = [-1/2, 1/2]$, according to the following formula:

$$Y(\cdot) = \left(\int_U \prod_{j=1}^{d} f_j(X_j)(u)\epsilon(u)\,du \right)^{-1} \cdot \prod_{j=1}^{d} f_j(X_j)(\cdot)\epsilon(\cdot), \tag{5.2}$$

where $f_j(x_j)(\cdot) : U \to \mathbb{R}$ are some measurable functions, ϵ is an error process and X_j are uniform $[0, 1]$ random variables. Specifically, we considered $d = 2$ and $d = 3$. We took $f_j(x_j)(u) = -\exp(-jx_j^j u^j)$ for $1 \le j \le 3$ and $\epsilon(u) = \exp(-Zu^4)$ with Z being a uniform $[-1, 1]$ random variable. By considering the operations $\oplus$ and $\odot$ for the quotient space $\mathbb{H} = \mathfrak{B}^2(U, U \cap \mathcal{B}(\mathbb{R}), \mathrm{Leb}_1)$ and the equivalence class $[Y(\cdot)]$ as introduced in Section 2, we clearly see that (5.2) falls into the additive model (1.1).

For a sensitivity analysis we also considered two nonadditive models for each d. For the first nonadditive scenario we took

$$\begin{aligned} Y(u) &= \frac{\prod_{j=1}^{2} f_j(X_j)(u) \cdot f_{12}(X_1, X_2)(u) \cdot \epsilon(u)}{\int_{-1/2}^{1/2} \prod_{j=1}^{2} f_j(X_j)(u) \cdot f_{12}(X_1, X_2)(u) \cdot \epsilon(u)\,du} \quad (d = 2), \\ Y(u) &= \frac{\prod_{j=1}^{3} f_j(X_j)(u) \cdot f_{123}(X_1, X_2, X_3)(u) \cdot \epsilon(u)}{\int_{-1/2}^{1/2} \prod_{j=1}^{3} f_j(X_j)(u) \cdot f_{123}(X_1, X_2, X_3)(u) \cdot \epsilon(u)\,du} \quad (d = 3), \end{aligned} \tag{5.3}$$

where f_j and ϵ are as defined in the additive scenario, $f_{12}(X_1, X_2)(u) = \exp(-X_1X_2u^2)$ and $f_{123}(X_1, X_2, X_3)(u) = \exp[-(X_1X_2 + X_1X_3 + X_2X_3)u^2]$.

For the second nonadditive scenario we considered

$$\begin{aligned} Y(u) &= \frac{\log((X_1/2 + X_2/2)u + 2)\epsilon(u)}{\int_{-1/2}^{1/2} \log((X_1/2 + X_2/2)u + 2)\epsilon(u)\,du} \quad (d = 2), \\ Y(u) &= \frac{\log((X_1/2 + X_2/2 + X_3/2)u + 2)\epsilon(u)}{\int_{-1/2}^{1/2} \log((X_1/2 + X_2/2 + X_3/2)u + 2)\epsilon(u)\,du} \quad (d = 3). \end{aligned} \tag{5.4}$$

The latter two models correspond to single-index models of the form $\mathbf{Y} = \mathbf{m}(\sum_{j=1}^{d} \theta_j X_j) \oplus \boldsymbol{\epsilon}$.

We repeatedly generated a training sample of size n and a test sample of size $N = 100$ for $M = 100$ times. As a measure of performance, we computed the mean squared prediction error (MSPE) defined by

$$\mathrm{MSPE} = M^{-1} \sum_{m=1}^{M} N^{-1} \sum_{i=1}^{N} \| [Y_i^{\mathrm{test}(m)}(\cdot)] \ominus [\hat{Y}_i^{\mathrm{test}(m)}(\cdot)] \|^2, \tag{5.5}$$

where $Y_i^{\mathrm{test}(m)}(\cdot)$ is the ith response in the mth test sample and $\hat{Y}_i^{\mathrm{test}(m)}(\cdot)$ is the prediction of $Y_i^{\mathrm{test}(m)}(\cdot)$ based on the mth training sample. We note that

$$\begin{aligned} &\| [Y_i^{\mathrm{test}(m)}(\cdot)] \ominus [\hat{Y}_i^{\mathrm{test}(m)}(\cdot)] \|^2 \\ &\quad = \frac{1}{2} \int_{[-1/2,1/2]^2} \left[\log\left(\frac{Y_i^{\mathrm{test}(m)}(u)}{Y_i^{\mathrm{test}(m)}(u')} \right) - \log\left(\frac{\hat{Y}_i^{\mathrm{test}(m)}(u)}{\hat{Y}_i^{\mathrm{test}(m)}(u')} \right) \right]^2 du\,du'. \end{aligned}$$

Table 1 suggests that the CBS algorithm is much faster than the full-dimensional grid search, especially for higher dimension. The former would also win the latter by wider margin as the grid $\mathcal{G}$ gets denser. The table also reveals that the bandwidths obtained by the CBS algorithm and by the full-dimensional search matched in most cases. This might be due to the fact that $\mathrm{CV}(h_1, \ldots, h_d)$ is coordinate-wise convex as is often the case in practice. Even in the case where the two were different, the CBS bandwidths gave comparable prediction results to the full-dimensional grid search, as the ratios in the last column of the table show.

In the simulation we also compared the prediction based on our approach with those based on full-dimensional estimators. We considered the functional Nadaraya–Watson (NW) estimator proposed by [9, 12] and [13] and the kernel-based functional k-nearest neighbor (k-NN) estimator proposed by [27] and [28]. For the NW estimator we used Epanechnikov kernel and tuned bandwidth on $\{b + 0.001 \times l : 1 \leq l \leq 200\}$ for some small b. For the k-nearest neighbor estimator we selected k from $\{1, 2, \ldots, 30\}$. We chose both the bandwidth and k by 10-fold cross-validation. Table 2 demonstrates that the proposed method won these methods, except in the lower dimensional ($d = 2$) nonadditive scenario (5.3). As the dimension gets higher in the nonadditive models, the SBF estimator deteriorated much less than the full-dimensional competitors, so that in the case $d = 3$ it won the latter two by large margins.

We conducted additional simulation with $d = 4$ and compared the performance of our proposal with those of oracle estimators that were based on the knowledge of some of the true component maps. The performance of the B-SBF was comparable with those of the oracle estimators; see the online Supplementary Material S.18 for more details.

5.3. *Real data analysis with functional response.* Predicting electricity consumption pattern is, nationally, an important issue because it may identify peak demand that can cause blackout. Recently, the Korea meteorological administration released a report that meteorological information can be useful in the prediction of national electricity consumption. Motivated by this report, we considered function-on-scalar regression taking electricity consumption trajectory as the response and some meteorological variables as predictors. Specifically, we took the monthly average of daylong home electricity consumption trajectories as the response, and the monthly average temperatures and amounts of clouds

TABLE 2
MSPE, multiplied by 10^3*, of the proposed method, the functional Nadaraya–Watson and the kernel-based functional k-NN methods*

Scenario	d	n	Proposed with CBS	Functional Nadaraya–Watson	Kernel-based functional k-NN
Additive (5.2)	2	100	0.1422	0.2573	0.2660
		400	0.1020	0.1310	0.1377
	3	100	0.1698	0.6862	0.7017
		400	0.1060	0.2834	0.2963
Non-Add. I (5.3)	2	100	0.1879	0.2541	0.2658
		400	0.1409	0.1345	0.1458
	3	100	0.3308	0.8287	0.8435
		400	0.2360	0.3258	0.3406
Non-Add. II (5.4)	2	100	0.1143	0.1367	0.1440
		400	0.1010	0.1067	0.1096
	3	100	0.1580	0.2856	0.2970
		400	0.1299	0.1558	0.1628

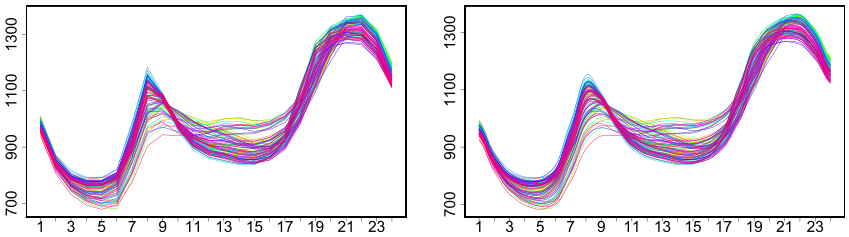

FIG. 1. *Raw electricity consumption curves in the electricity data (left) and their pre-smoothed curves (right). Each raw curve is presmoothed by kernel smoothing with the Gaussian kernel ($sd = 0.5$) and leave-one-out cross-validatory bandwidth.*

as two predictors. The data on these variables were available for the period January 2008–December 2016, in KOSIS (Korean Statisticl Information Service) http://kosis.kr/statHtml/statHtml.do?orgId=310&tblId=DT_3664N_2008 and the Korea meteorological administration https://data.kma.go.kr/cmmn/main.do. The observed responses are normalized hourly electricity loads, $Y(t_j) = 1000 \times Z(t_j) \times (\sum_{l=1}^{24} Z(t_l)/24)^{-1}$, where $Z(t)$ denotes the electricity load at time t. Thus, we focus on predicting relative consumption pattern during a day, not absolute electricity consumption itself. We presmoothed the hourly observed $Y_i(\cdot)$ and applied our method to the presmoothed versions. Figure 1 depicts the observed trajectories of $Y(\cdot)$ and their presmoothed curves. We computed the leave-one-curve-out average squared prediction error (ASPE) defined by

$$\text{ASPE} = n^{-1} \sum_{i=1}^{n} \|Y_i(\cdot) \ominus \hat{Y}_i^{(-i)}(\cdot)\|^2 = n^{-1} \sum_{i=1}^{n} \int_0^1 \big(Y_i(s) - \hat{Y}_i^{(-i)}(s)\big)^2 \, ds$$

with $n = 108$ (9 years $\times$ 12 months), where $\hat{Y}_i^{(-i)}(\cdot)$ is the prediction of $Y_i(\cdot)$ based on the sample without the ith observation.

For this example we compared our method with those of [6] and [40] and with the functional NW and the kernel-based functional k-NN estimators. [6] considered the model $\mathrm{E}(Y(t)|\mathbf{X}) = \mathrm{E}(Y(t)) + \sum_{k=1}^{L} g_k(\boldsymbol{\beta}_k^\top \mathbf{X})\rho_k(t)$, where $L \geq 1$, $\boldsymbol{\beta}_k \in \mathbb{R}^d$ and real-valued functions g_k are unknown and ρ_k are the (unknown) eigenfunctions of the autocovariance operator of the response process $Y(\cdot)$. [40] studied the model $\mathrm{E}(Y(t)|\mathbf{X}) = g_0(t) + g_1(X_1, t) + g_2(X_2, t)$ based on spline expansions of g_j. In the comparison we included, as well the function-on-scalar linear model, $\mathrm{E}(Y(t)|\mathbf{X}) = g_0(t) + X_1 g_1(t) + X_2 g_2(t)$, which was also discussed in [40]. To implement the method of [6], we used "FQR" function in the matlab package "PACE" (version 2.17) with bandwidth for mean curve being selected by leave-one-curve-out cross-validation and bandwidth for covariance surface being selected by GCV. For the method of [40], we used "pffr" function in the R package "refund" (version 0.1-16) with 100 cubic B-spline basis functions and smoothing parameter selected by REML. The penalty was "first-order-difference", which is the default option of "pffr" in "refund". We chose the bandwidth for the NW method and the smoothing parameter k for the k-NN estimator using 10-fold cross-validation.

Table 3 contains the results, which suggest that our method outperforms all competitors. While the performances of the functional NW and k-NN estimators were inferior to those of the approaches based on structured models, such as the B-SBF, [6] and [40], they were computationally cheaper. We found that methods of [6] and [40] were computationally very expensive while the B-SBF was moderate but slower than the NW and k-NN.

Figure 2 depicts the fitted component maps based on our method. The first component map demonstrates that, when the weather is hot or cold, people use more electricity in the afternoon than in normal temperature. The second component map illustrates that, when it is

TABLE 3
Comparison of ASPE for electricity data

Method	ASPE
B-SBF with CBS	315.54
Function-on-scalar additive model (Scheipl et al. (2015))	355.46
Function-on-scalar multiple-index model (Chiou et al. (2003))	355.64
Kernel-based functional k-NN	361.99
Functional Nadaraya–Watson	425.19
Function-on-scalar linear model (Scheipl et al. (2015))	704.90

cloudy, people use more electricity in the afternoon than in less cloudy condition. Combining these two component maps, we may identify peak demand hours depending on the weather conditions. We found that the peak hour predicted by the B-SBF was not much affected by cloudiness, while the spline additive regression approach studied by [40] produced a different result showing that the peak hour was influenced by both temperature and cloudiness. For the plots of the peak hours against the mean temperature and cloudiness, see the online Supplementary Material S.20.

5.4. *Real data analysis with simplex-valued response*. It is a general belief that age, educational level and richness are main factors in determining people's political orientation. There have not been many statistical analysis checking the belief using an advanced method. Here, we analyzed the 2017 Korea presidential election data collected from https://github.com/OhmyNews/2017-Election. The dataset contains the voting results and some demographic information for each of the 250 electoral districts in Korea. The voting results are the proportions of votes earned by the top five candidates, and the demographic information consists of average age(X_1), average years of education(X_2), average housing price per square meter(X_3) and average paid national health insurance premium(X_4). The last one is considered as an indicator of richness because those who get high salary pay more national health insurance premium. Since the election was mainly focused on the candidates from three major parties representing progressivism, conservatism and centrism, we took the

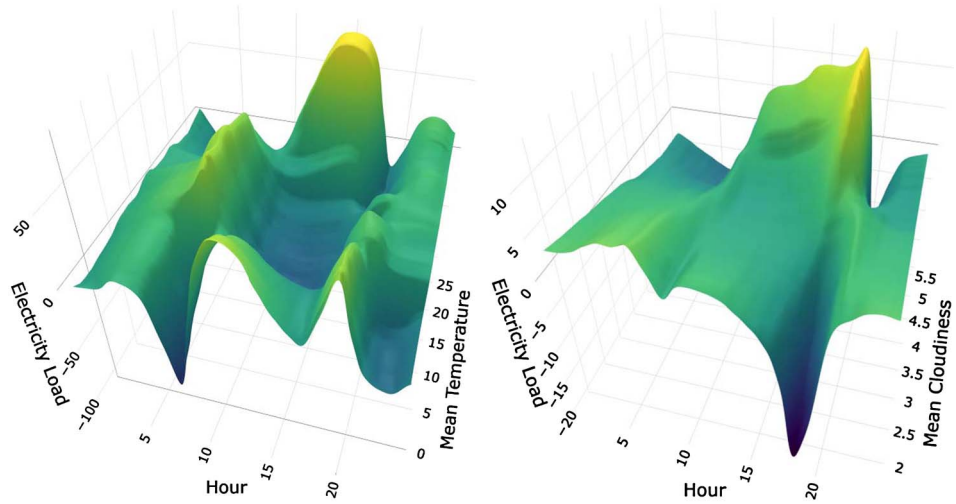

FIG. 2. *The fitted component maps for electricity data based on the B-SBF method for the monthly averaged temperature* (*left*) *and for the monthly averaged cloudiness* (*right*).

three-dimensional compositional vector $\mathbf{Y} = (Y_1, Y_2, Y_3) \in \mathcal{S}_1^3 = \mathbb{H}$ as the response, where Y_j is the proportion of votes earned by the jth candidate among the three. We divided the 250 observations into 10 partitions S_k, $1 \le k \le 10$, with each partition having 25 observations. We then computed the 10-fold average squared prediction error (ASPE) defined by

$$\text{ASPE} = 10^{-1} \sum_{k=1}^{10} |S_k|^{-1} \sum_{i \in S_k} \left\| \mathbf{Y}_i \ominus \hat{\mathbf{Y}}_i^{(-S_k)} \right\|^2,$$

where $|S_k|$ is the number of observations in S_k and $\hat{\mathbf{Y}}_i^{(-S_k)}$ is the prediction of $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})$ based on the sample with the observations in S_k being removed.

To find a model that best describes the relationship between $\mathbf{Y}$ and the four-dimensional predictor, we compared our method with four others. These are in the R package "Compositional." The first one is Dirichlet regression named as "diri.reg". It assumes that the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is given by Dirichlet$(\phi\alpha_1(\mathbf{x}), \phi\alpha_2(\mathbf{x}), \phi\alpha_3(\mathbf{x}))$, where $\phi > 0$, $\alpha_1(\mathbf{x}) = 1/[1 + \sum_{l=2}^3 \exp(\beta_{l0} + \boldsymbol{\beta}_l^\top \mathbf{x})]$ and $\alpha_j(\mathbf{x}) = \exp(\beta_{j0} + \boldsymbol{\beta}_j^\top \mathbf{x})/[1 + \sum_{l=2}^3 \exp(\beta_{l0} + \boldsymbol{\beta}_l^\top \mathbf{x})]$ for $j \ge 2$ and estimates $\phi, \beta_{j0} \in \mathbb{R}$ and $\boldsymbol{\beta}_j \in \mathbb{R}^4$ by maximum likelihood. The second one named as "ols.compreg" is to assume the multinomial logistic model, $\mathrm{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = (\alpha_1(\mathbf{x}), \alpha_2(\mathbf{x}), \alpha_3(\mathbf{x}))$, and estimates β_{j0} and $\boldsymbol{\beta}_j$ by least squares. The third one named as "kl.compreg" also assumes the multinomial logistic model but estimates β_{j0} and $\boldsymbol{\beta}_j$ by minimizing the Kullback–Leibler divergence $\sum_{i=1}^n \sum_{j=1}^3 Y_{ij} \log(Y_{ij}/\alpha_j(\mathbf{X}_i))$. The last one is the method of [42] named as "alfa.reg" and for this we tuned "alpha" on $\{-1 + 0.2 \times k : 0 \le k \le 10\}$ by 10-fold cross-validation. As for the computing time, the four competitors were faster than the B-SBF.

The results are presented in Table 4. It demonstrates that our method is most predictive. Figure 3 depicts the component maps fitted by the proposed method. The first, third and fourth fitted component maps suggest that districts where older or richer people live are politically more conservative. The second map tells an interesting story. As people are more educated, from low to medium level, their political orientation moves to the direction of conservatism, while it is reversed for people at medium to medium-high education level. For people at medium-high to high education level, it shows a zigzagging pattern. The applications of other methods offered somewhat different conclusions. Basically, all other methods are based on a model that is monotone in each of the predictors, so that we may not expect the zigzagging pattern that we observed in the application of our method. For example, the coefficients $\boldsymbol{\beta}_j$ estimated by the Kullback–Leibler-divergence-based regression "kl.compreg" tell that, as people get more educated, they tend to support both conservatism and centrism. For the estimated coefficients of other methods, we refer to the tables in the online Supplementary Material S.21.

TABLE 4
Comparison of ASPE for election data

Method	ASPE
B-SBF with CBS	0.82
Alpha transformation method (Tsagris (2015))	1.07
Kullback-Leibler-divergence-based regression	1.07
Dirichlet regression	1.07
Multinomial logistic regression	1.08

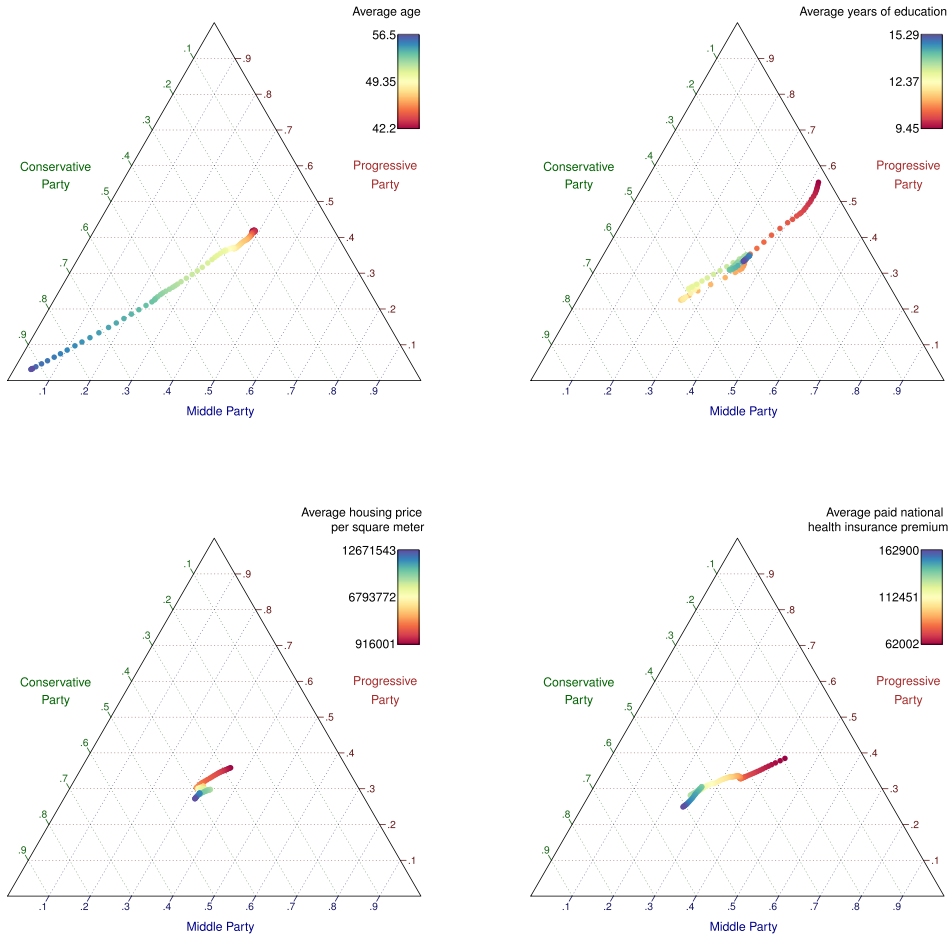

FIG. 3. *The values of the fitted component maps for election data based on the B-SBF method, depicted on the simplex* $\mathcal{S}_1^3$, *for the average age* (*top left*), *average years of education* (*top right*), *average housing price per square meter* (*bottom left*) *and average paid national health insurance premium* (*bottom right*).

6. Discussion. We have focused on the case where the predictors X_j are real. One challenging extension is to consider infinite-dimensional predictors. Consider the model (1.1), now for infinite-dimensional $\mathbf{X}_j$ taking values in Banach spaces $\mathbb{B}_j$. For this we assume that there exist measures μ_j on $(\mathbb{B}_j, \mathcal{B}(\mathbb{B}_j))$ such that

$$P\mathbf{X}^{-1} \ll \mu_1 \otimes \cdots \otimes \mu_d, \tag{6.1}$$

where $\mathbf{X} \equiv (\mathbf{X}_1, \ldots, \mathbf{X}_d)$. Define the density of $\mathbf{X}$ by $p = dP\mathbf{X}^{-1}/d\bigotimes_{j=1}^d \mu_j$. Also, define p_j and p_{jk} to be the integrations of p over $\mathbb{B}_{-j} := \prod_{l\neq j} \mathbb{B}_l$ and $\mathbb{B}_{-jk} := \prod_{l\neq j,k} \mathbb{B}_l$, respectively, with the measures $\mu_{-j} := \bigotimes_{l\neq j} \mu_l$ and $\mu_{-jk} := \bigotimes_{l\neq j,k} \mu_l$. With these new definitions, the estimating equation (2.2) with obvious changes holds. We may estimate these densities and $\mathrm{E}(\mathbf{Y}|\mathbf{X}_j = \mathbf{x}_j)$ in the same way as in Section 2 but with the modified kernel scheme

$$K_{h_j}(\mathbf{x}_j, \mathbf{X}_{ij}) = \frac{K(h_j^{-1}\|\mathbf{x}_j \ominus \mathbf{X}_{ij}\|)}{\int_{\mathbb{B}_j} K(h_j^{-1}\|\mathbf{x}_j \ominus \mathbf{X}_{ij}\|)\, d\mu_j(\mathbf{x}_j)} \tag{6.2}$$

for a baseline kernel function K, now defined on $[0, \infty)$. Let $\hat{p}$, $\hat{p}_j$, $\hat{p}_{jk}$ and $\tilde{\mathbf{m}}_j$ continue to denote the estimators. Then, we would get analogues of (2.4) and (2.6) for the B-SBF equation and algorithm. Also, by going through the development in Section 3 we may establish

versions of Theorems 3.4 and 3.5 under a modified version of the condition (S). We refer to the online Supplementary Material S.2 for further details.

The implementation of the B-SBF algorithm involves integrating real-valued functions over $\mathbb{B}_k$. In case $\mathbb{B}_k = \mathbb{R}$ and μ_k is the Lebesgue measure, it has no difficulty as we have seen in previous sections. However, in the case of infinite-dimensional $\mathbb{B}_k$ there exists no reference measure such as Lebesgue measure. It has not been well studied how to implement integrals with respect to general measures on Banach spaces. In case there exist random elements $\mathbf{W}_k$ whose distributions equal μ_k and one can generate i.i.d. $\mathbf{W}_{ik}$ for $1 \le i \le M$ from μ_k, then one may approximate $\int_{\mathbb{B}_k} g_k(\mathbf{x}_k)\mu_k(\mathbf{x}_k)$ for real-valued functions g_k by $M^{-1}\sum_{l=1}^{M} g_k(\mathbf{W}_{ik})$. The statistical properties of this Monte Carlo approximation and their implications to the properties of the estimators $\hat{\mathbf{m}}_j$ are yet to be developed.

In the case where $\mathbf{X}_j$ are functional predictors, there is a general way of dimension reduction based on functional principal components (FPC), as originated from [36]. In fact, [16] applied this approach to the case of a scalar response and multiple functional predictors. We may repeat it in our case with Hilbertian responses. Let ξ_{jk} denote the kth FPC of $\mathbf{X}_j$. Then, it is straightforward to extend our methodology and theory to estimating Hilbertian additive models of the form $\mathbf{Y} = \mathbf{m}_0 \oplus \bigoplus_{j=1}^{d} \bigoplus_{k=1}^{L_j} \mathbf{m}_{jk}(\xi_{jk}) \oplus \boldsymbol{\epsilon}$, under some additional conditions in [16]. Another way of dimension reduction is through functional single-index modeling. This approach basically takes $L_j \equiv 1$ and replaces $\xi_j \equiv \xi_{j1}$ by $\langle \theta_j, \mathbf{X}_j \rangle$ for some unknown real-valued functions θ_j.

The SBF technique is largely based on the method of alternating orthogonal projections acting on the spaces $L_2^{\mathbb{H}}(p)$ and $L_2^{\mathbb{H}}(\hat{p})$. One might be interested in extending our method to $\mathbb{B}$-valued responses. The main difficulty with the extension is that the spaces $L_2^{\mathbb{B}}(p)$ and $L_2^{\mathbb{B}}(\hat{p})$, defined as $L_2^{\mathbb{H}}(p)$ and $L_2^{\mathbb{H}}(\hat{p})$ but with $\mathbb{H}$ being replaced by $\mathbb{B}$, are non-Hilbertian. There are a number of hurdles to overcome to extend our results to $L_2^{\mathbb{B}}(p)$ and $L_2^{\mathbb{B}}(\hat{p})$. To list some of them, one needs analogues of Proposition A.4.2 of [2] and Lemma S.7 for non-Hilbertian spaces of maps. For asymptotic distribution one also need to develop a Lindeberg-type CLT for sequences of $\mathbb{B}$-valued random elements. [23] provided one in their "Extension of Theorem 1.1," but checking the third condition of the theorem is infeasible since there is no distributive law to be applied to the norm of a Banach space. For these reasons a unified treatment of $\mathbb{B}$-valued responses seems too early considering the current state of related theory in mathematics. One may handle each Banach-space-valued response case by case. A useful approach to this option might be to think of a transformation that maps the underlying Banach space, where the response takes values, to a Hilbert space and then apply the methods and tools we explored in this paper to the transformed response variable.

We may strengthen Theorem 4.1 in two directions. An easy one is to relax the bandwidth condition (C5). Suppose that $n^{\beta}h_j \to \alpha_j \in (0, \infty)$ for some $0 < \beta < \min\{1/2, (\alpha - 2)/\alpha\}$, where α is the constant in (C1). Let $h \asymp h_j$, $1 \le j \le d$. Then, we may show that the pointwise rates $n^{-2/5}$ in the interior I_j and $n^{-1/5}$ at the boundaries I_j^c in Theorem 4.1 are now modified to $h^2 + n^{-1/2}h^{-1/2}$ and $h + n^{-1/2}h^{-1/2}$, respectively. We may also prove that the squared L_2 rate on the interior and the one on the whole interval $[0, 1]$, respectively, are $h^4 + n^{-1}h^{-1}$ and $h^3 + n^{-1}h^{-1}$, and that the respective sup-rates are $h^2 + n^{-1/2}h^{-1/2}(\log n)^{1/2}$ and $h + n^{-1/2}h^{-1/2}(\log n)^{1/2}$. The more challenging extension is to get uniform rates over $h_j \in [a_n, b_n]$ for some sequences $a_n < b_n$. This is important as it gives error rates for the $\hat{\mathbf{m}}_j$ with data-driven bandwidth choices such as those discussed in Section 5.1 in case they are chosen to minimize the CV criterion in the range $[a_n, b_n]$. We believe it is possible to get some results that are similar in flavor to [32] and [22], for example. This could be the subject of a new paper.

The present paper considered only the estimation problem. One challenging topic for future study is to develop a procedure for testing if some of the components in the model (1.1)

are zero. Another direction for future study is to treat responses or predictors taking values in a locally compact space, which encompasses the case of Riemannian manifolds.

Acknowledgements. We thank an Associate Editor and three referees for their helpful comments on the earlier versions of the paper.

This work was supported by Samsung Science and Technology Foundation under Project Number SSTF-BA1802-01.

SUPPLEMENTARY MATERIAL

Supplement to "Additive regression with Hilbertian responses" (DOI: 10.1214/19-AOS1902SUPP; .pdf). The Supplementary Material [19] contains additional lemmas and propositions with their proofs and the proofs of Propositions 2.1, 2.2, 3.2 and Theorems 3.1–3.6, 4.1–4.3. It also introduces two other notions of Bochner integrals, gives the extensions of Theorems 3.4 and 3.5 to the case of infinite-dimensional predictors, and presents additional numerical results.

REFERENCES

[1] ANEIROS, G., CAO, R., FRAIMAN, R., GENEST, C. and VIEU, P. (2019). Recent advances in functional data analysis and high-dimensional statistics. *J. Multivariate Anal.* **170** 3–9. MR3913024 https://doi.org/10.1016/j.jmva.2018.11.007

[2] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

[3] BISSANTZ, N., DETTE, H., HILDEBRANDT, T. and BISSANTZ, K. (2016). Smooth backfitting in additive inverse regression. *Ann. Inst. Statist. Math.* **68** 827–853. MR3520045 https://doi.org/10.1007/s10463-015-0517-x

[4] BOSQ, D. (2000). *Linear Processes in Function Spaces*: *Theory and Applications*. *Lecture Notes in Statistics* **149**. Springer, New York. MR1783138 https://doi.org/10.1007/978-1-4612-1154-9

[5] BUSBY, R. C., SCHOCHETMAN, I. and SMITH, H. A. (1972). Integral operators and the compactness of induced representations. *Trans. Amer. Math. Soc.* **164** 461–477. MR0295099 https://doi.org/10.2307/1995990

[6] CHIOU, J.-M., MÜLLER, H.-G. and WANG, J.-L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 405–423. MR1983755 https://doi.org/10.1111/1467-9868.00393

[7] COHN, D. L. (2013). *Measure Theory*, 2nd ed. *Birkhäuser Advanced Texts*: *Basler Lehrbücher.* [*Birkhäuser Advanced Texts*: *Basel Textbooks*]. Birkhäuser/Springer, New York. MR3098996 https://doi.org/10.1007/978-1-4614-6956-8

[8] CONWAY, J. B. (1985). *A Course in Functional Analysis*. *Graduate Texts in Mathematics* **96**. Springer, New York. MR0768926 https://doi.org/10.1007/978-1-4757-3828-5

[9] DABO-NIANG, S. and RHOMARI, N. (2009). Kernel regression estimation in a Banach space. *J. Statist. Plann. Inference* **139** 1421–1434. MR2485136 https://doi.org/10.1016/j.jspi.2008.06.015

[10] DELICADO, P. and VIEU, P. (2017). Choosing the most relevant level sets for depicting a sample of densities. *Comput. Statist.* **32** 1083–1113. MR3703579 https://doi.org/10.1007/s00180-017-0746-y

[11] DURRETT, R. (2019). *Probability—Theory and Examples*. *Cambridge Series in Statistical and Probabilistic Mathematics* **49**. Cambridge Univ. Press, Cambridge. Fifth edition of [MR1068527]. MR3930614 https://doi.org/10.1017/9781108591034

[12] FERRATY, F., LAKSACI, A., TADJ, A. and VIEU, P. (2011). Kernel regression with functional response. *Electron. J. Stat.* **5** 159–171. MR2786486 https://doi.org/10.1214/11-EJS600

[13] FERRATY, F., VAN KEILEGOM, I. and VIEU, P. (2012). Regression when both response and predictor are functions. *J. Multivariate Anal.* **109** 10–28. MR2922850 https://doi.org/10.1016/j.jmva.2012.02.008

[14] FERRATY, F. and VIEU, P. (2009). Additive prediction and boosting for functional data. *Comput. Statist. Data Anal.* **53** 1400–1413. MR2657100 https://doi.org/10.1016/j.csda.2008.11.023

[15] GOLDBERG, S. (1959). Some properties of the space of compact operators on a Hilbert space. *Math. Ann.* **138** 329–331. MR0121673 https://doi.org/10.1007/BF01344152

[16] HAN, K., MÜLLER, H.-G. and PARK, B. U. (2018). Smooth backfitting for additive modeling with small errors-in-variables, with an application to additive functional regression for multiple predictor functions. *Bernoulli* **24** 1233–1265. MR3706793 https://doi.org/10.3150/16-BEJ898

[17] HAN, K. and PARK, B. U. (2018). Smooth backfitting for errors-in-variables additive models. *Ann. Statist.* **46** 2216–2250. MR3845016 https://doi.org/10.1214/17-AOS1617

[18] JAYASUMANA, S., SALZMANN, M., LI, H. and HARANDI, M. (2013). A framework for shape analysis via Hilbert space embedding. In *Proceedings of the IEEE International Conference on Computer Vision* 1249–1256.

[19] JEON, J. M. and PARK, B. U. (2020). Supplement to "Additive regression with Hilbertian responses." https://doi.org/10.1214/19-AOS1902SUPP.

[20] JIANG, C.-R. and WANG, J.-L. (2011). Functional single index models for longitudinal data. *Ann. Statist.* **39** 362–388. MR2797850 https://doi.org/10.1214/10-AOS845

[21] KALLENBERG, O. (1997). *Foundations of Modern Probability*. *Probability and Its Applications* (*New York*). Springer, New York. MR1464694

[22] KARA-ZAITRI, L., LAKSACI, A., RACHDI, M. and VIEU, P. (2017). Uniform in bandwidth consistency for various kernel estimators involving functional data. *J. Nonparametr. Stat.* **29** 85–107. MR3597219 https://doi.org/10.1080/10485252.2016.1254780

[23] KUNDU, S., MAJUMDAR, S. and MUKHERJEE, K. (2000). Central limit theorems revisited. *Statist. Probab. Lett.* **47** 265–275. MR1747487 https://doi.org/10.1016/S0167-7152(99)00164-9

[24] LEADBETTER, R., CAMBANIS, S. and PIPIRAS, V. (2014). *A Basic Course in Measure and Probability*: *Theory for Applications*. Cambridge Univ. Press, Cambridge. MR3445362

[25] LEE, Y. K., MAMMEN, E. and PARK, B. U. (2010). Backfitting and smooth backfitting for additive quantile models. *Ann. Statist.* **38** 2857–2883. MR2722458 https://doi.org/10.1214/10-AOS808

[26] LEE, Y. K., MAMMEN, E. and PARK, B. U. (2012). Flexible generalized varying coefficient regression models. *Ann. Statist.* **40** 1906–1933. MR3015048 https://doi.org/10.1214/12-AOS1026

[27] LIAN, H. (2011). Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electron. J. Stat.* **5** 31–40. MR2773606 https://doi.org/10.1214/11-EJS595

[28] LIAN, H. (2012). Convergence of nonparametric functional regression estimates with functional responses. *Electron. J. Stat.* **6** 1373–1391. MR2988451 https://doi.org/10.1214/12-EJS716

[29] LING, N. and VIEU, P. (2018). Nonparametric modelling for functional data: Selected survey and tracks for future. *Statistics* **52** 934–949. MR3827427 https://doi.org/10.1080/02331888.2018.1487120

[30] LINTON, O., SPERLICH, S. and VAN KEILEGOM, I. (2008). Estimation of a semiparametric transformation model. *Ann. Statist.* **36** 686–718. MR2396812 https://doi.org/10.1214/009053607000000848

[31] MAMMEN, E., LINTON, O. and NIELSEN, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490. MR1742496 https://doi.org/10.1214/aos/1017939137

[32] MAMMEN, E. and PARK, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *Ann. Statist.* **33** 1260–1294. MR2195635 https://doi.org/10.1214/009053605000000101

[33] MAMMEN, E., PARK, B. U. and SCHIENLE, M. (2014). Additive models: Extensions and related models. In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics* 176–211. Oxford Univ. Press, Oxford. MR3306926

[34] MARRON, J. S. and ALONSO, A. M. (2014). Overview of object oriented data analysis. *Biom. J.* **56** 732–753. MR3258083 https://doi.org/10.1002/bimj.201300072

[35] MUANDET, K., SRIPERUMBUDUR, B., FUKUMIZU, K., GRETTON, A. and SCHÖLKOPF, B. (2016). Kernel mean shrinkage estimators. *J. Mach. Learn. Res.* **17** Paper No. 48, 41. MR3504608

[36] MÜLLER, H.-G. and YAO, F. (2008). Functional additive models. *J. Amer. Statist. Assoc.* **103** 1534–1544. MR2504202 https://doi.org/10.1198/016214508000000751

[37] PARK, B. U., CHEN, C.-J., TAO, W. and MÜLLER, H.-G. (2018). Singular additive models for function to function regression. *Statist. Sinica* **28** 2497–2520. MR3839871

[38] PETERSEN, A. and MÜLLER, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Ann. Statist.* **44** 183–218. MR3449766 https://doi.org/10.1214/15-AOS1363

[39] PINI, A., STAMM, A. and VANTINI, S. (2017). Hotelling in Wonderland. In *Functional Statistics and Related Fields*. *Contrib. Stat.* 211–216. Springer, Cham. MR3823575

[40] SCHEIPL, F., STAICU, A.-M. and GREVEN, S. (2015). Functional additive mixed models. *J. Comput. Graph. Statist.* **24** 477–501. MR3357391 https://doi.org/10.1080/10618600.2014.901914

[41] TALSKÁ, R., MENAFOGLIO, A., MACHALOVÁ, J., HRON, K. and FIŠEROVÁ, E. (2018). Compositional regression with functional response. *Comput. Statist. Data Anal.* **123** 66–85. MR3777086 https://doi.org/10.1016/j.csda.2018.01.018

[42] TSAGRIS, M. (2015). Regression analysis with compositional data containing zero values. *Chil. J. Stat.* **6** 47–57. MR3407274

[43] VAN DEN BOOGAART, K. G., EGOZCUE, J. J. and PAWLOWSKY-GLAHN, V. (2014). Bayes Hilbert spaces. *Aust. N. Z. J. Stat.* **56** 171–194. MR3226435 https://doi.org/10.1111/anzs.12074

[44] VÄTH, M. (2000). *Volterra and Integral Equations of Vector Functions*. *Monographs and Textbooks in Pure and Applied Mathematics* **224**. Dekker, New York. MR1738341

[45] XU, J. and ZIKATANOV, L. (2002). The method of alternating projections and the method of subspace corrections in Hilbert space. *J. Amer. Math. Soc.* **15** 573–597. MR1896233 https://doi.org/10.1090/S0894-0347-02-00398-3

[46] YU, K., PARK, B. U. and MAMMEN, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36** 228–260. MR2387970 https://doi.org/10.1214/009053607000000596

[47] ZHANG, X., PARK, B. U. and WANG, J.-L. (2013). Time-varying additive models for longitudinal data. *J. Amer. Statist. Assoc.* **108** 983–998. MR3174678 https://doi.org/10.1080/01621459.2013.778776

[48] ZHU, H., LI, R. and KONG, L. (2012). Multivariate varying coefficient model for functional responses. *Ann. Statist.* **40** 2634–2666. MR3097615 https://doi.org/10.1214/12-AOS1045