Electronic Journal of Statistics
Vol. 15 (2021) 1473–1548
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1823

Additive regression for predictors of various natures and possibly incomplete Hilbertian responses*

Jeong Min Jeon, Byeong U. Park and Ingrid Van Keilegom

ORSTAT, KU Leuven
Namsestraat 69, Leuven 3000, Belgium
e-mail: jeongmin.jeon@kuleuven.be; ingrid.vankeilegom@kuleuven.be

Department of Statistics, Seoul National University
Gwanak-ro 1, Seoul 08826, South Korea
e-mail: bupark@stats.snu.ac.kr

Abstract: In this paper we consider a fully nonparametric additive regression model for responses and predictors of various natures. This includes the case of Hilbertian and incomplete (like censored or missing) responses, and continuous, nominal discrete and ordinal discrete predictors. We propose a backfitting technique that estimates this additive model, and establish the existence of the estimator and the convergence of the associated backfitting algorithm under minimal conditions. We also develop a general asymptotic theory for the estimator such as the rates of convergence and asymptotic distribution. We verify the practical performance of the proposed estimator in a simulation study. We also apply the method to various real data sets, including those for a density-valued response regressed on a mixture of continuous and nominal discrete predictors, for a compositional response regressed on a mixture of continuous and ordinal discrete predictors, and for a censored scalar response regressed on a mixture of continuous and nominal discrete predictors.

MSC2020 subject classifications: Primary 62G08; secondary 62G20.
Keywords and phrases: Additive model, smooth backfitting, Hilbertian response, incomplete response, mixed predictor.

Received July 2020.

Contents

*Research of Jeong Min Jeon and Ingrid Van Keilegom was supported by the European Research Council (2016-2021, Horizon 2020/ERC grant agreement No. 694409). Resrach of Byeong U. Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2019R1A2C3007355).

1. Introduction

Data objects that are not Euclidean are now abundant in real world problems so that their analysis becomes one of the important tasks in modern statistics. As part of such task, this paper provides a general structured nonparametric regression technique for Hilbert-space-valued (Hilbertian) responses coupled

with various types of predictors. We consider Hilbert space since it is an important class of data spaces equipped with vector operations and an inner product structure that are vital to most regression tools. It covers a very wide scope of data types such as Euclidean, functional, density-valued and compositional data, etc. Functional data means that each data object itself, corresponding to each subject, is a function. This type of data arises from every corner of our lives [35]. Density-valued data is a kind of functional data, but each data object is a nonnegative function that integrates to one on its domain. Examples include population age distributions in cities [14], the distributions of voxel-to-voxel correlation in fMRI signals [33, 34] and the distributions of metabolite level in the groups of new born babies [37]. Compositional data has Euclidean vectors as data objects whose entries are positive and sum to one. Examples are the proportions of votes earned by candidates in an election, the proportions of races/religions in cities/countries and the proportions of chemical materials constituting bodies/air/sea-water/soil [32, 9]. For formal definitions of such data objects, see Section 2.1.

There have been a few attempts of dealing with Hilbertian responses. A broad review on regression for functional responses may be found in [42]. Other works include a parametric technique for compositional responses [38], the one for density-valued responses [37], and Nadaraya-Watson smoothing [10] and k-nearest neighbor estimation [21] for Hilbertian responses. The latter two works are about nonparametric regression but based on full-dimensional (i.e., unstructured) modeling, so that their approaches suffer from the curse of dimensionality when the number of predictors increases. Recently, nonparametric additive regression has been developed for Hilbertian responses [15]. Additive modeling is known to be an efficient way of avoiding the dimensionality problem. All these works on nonparametric regression do not cover discrete predictors, however. There has been no nonparametric method dealing with density-valued or compositional responses, in particular, together with discrete predictors.

In real world regression problems, discrete predictors are abundant. In many cases it is how to model the effects of discrete predictors, rather than continuous predictors, that determines overall prediction performance. In nonparametric regression, it is the usual practice to assume that the effects of discrete predictors are linear, thus incorporate them into a partially linear [e.g., 36, 45] or a varying coefficient model [e.g., 13, 19, 20]. The usual approach certainly lacks flexibility since it cannot accommodate nonlinear effects, which may result in poor practical performance as illustrated in Section 5. A systematic nonparametric approach to identifying possibly nonlinear effects of discrete predictors on Hilbertian responses, does not exist yet, to the knowledge of the authors. The problem remains unexplored despite of its importance in real world problems.

The aim of the current paper is to develop a nonparametric additive regression approach for general Hilbertian responses that also enhances flexibility in modeling the effects of various types of predictors. Our approach sets both the effects of continuous and those of discrete predictors 'fully nonparametric'. The predictors in our model can be general objects including nominal and ordinal discrete variables as well as continuous ones. Hence, our coverage includes the

case where only continuous predictors are considered [15]. In addition, we allow for incompletely observed responses. Here, by 'incomplete', we mean 'censored' or 'missing'. In particular, our setup covers the case of missing data with general Hilbertian responses. It means that we can deal with missing functional responses, missing density responses, missing compositional responses and so on, and as such it generalizes many current advances in the analysis of missing data.

To introduce our model, let $\mathbb{H}$ denote a separable Hilbert space and $\mathbf{Y}$ be a possibly incompletely observed $\mathbb{H}$-valued random element. Let $\mathbf{W} = (\mathbf{X}, \mathbf{U}, \mathbf{V})$ be mixed predictors taking values in an appropriate space $\mathcal{W}$, where $\mathbf{X} = (X_1, \ldots, X_{d_x})$ is a vector of real-valued continuous predictors, $\mathbf{U} = (U_1, \ldots, U_{d_u})$ is a vector of nominal discrete predictors, and $\mathbf{V} = (V_1, \ldots, V_{d_v})$ is a vector of discrete predictors for which each V_j takes values in a metric space with finite cardinality. Here, some d_x, d_u or d_v are allowed to be zero as long as $d_x + d_u + d_v \geq 2$. Let $\boldsymbol{\epsilon}$ be a $\mathbb{H}$-valued error such that $\mathrm{E}(\boldsymbol{\epsilon}|\mathbf{X}, \mathbf{U}, \mathbf{V}) = \mathbf{0}$, where $\mathbf{0}$ is a zero vector in $\mathbb{H}$, and the conditional expectation is defined in terms of Bochner integral [6]. We consider the following additive model.

$$\mathbf{Y} = \mathbf{m}_0 \oplus \bigoplus_{j=1}^{d_x} \mathbf{m}_{x,j}(X_j) \oplus \bigoplus_{j=1}^{d_u} \mathbf{m}_{u,j}(U_j) \oplus \bigoplus_{j=1}^{d_v} \mathbf{m}_{v,j}(V_j) \oplus \boldsymbol{\epsilon}, \tag{1.1}$$

where $\oplus$ is a vector addition on $\mathbb{H}$, $\mathbf{m}_0$ is an unknown constant in $\mathbb{H}$, and $\mathbf{m}_{x,j}, \mathbf{m}_{u,j}$ and $\mathbf{m}_{v,j}$ are unknown $\mathbb{H}$-valued component maps. The definitions of zero vector $\mathbf{0}$ and vector addition $\oplus$ are different for different Hilbert spaces, see Section 2.1. To the best of our knowledge, this model is the first fully nonparametric version of the standard linear model.

For the estimation of the component maps in (1.1), we develop a new smooth backfitting (SBF) technique. The original idea of SBF was developed for scalar responses and continuous predictors [25]. In comparison with other structured nonparametric methods such as marginal integration [22] and ordinary backfitting [31], the SBF technique was proved to have practical advantages [29] as well as theoretical superiority in various structured nonparametric models [e.g., 26, 23, 46, 18, 12]. All the aforementioned works are for completely observed *real-valued* responses regressed on *continuous* predictors only.

In our theoretical developments, we first prove the existence of the SBF estimator and the convergence of the associated SBF algorithm, in a very general setup where predictors take values in arbitrary σ-finite measure spaces. The general treatment allows not only the three types of predictors in the model (1.1) but also other predictors such as functional and manifold-valued predictors. The SBF algorithm, which evaluates the SBF estimator, requires the estimation of the marginal densities and the marginal regression maps of the predictors. In the existing SBF literature, these are confined to kernel-based estimators. In our theory for the existence of the estimator and the convergence of the algorithm, they are not restricted to this type. Instead, we formulate high-level conditions that are minimally required for the estimators of the marginal densities and marginal regression maps. The conditions allow parametric or nonparametric es-

timators which may not be kernel-based but could be spline- or wavelet-based. Hence, our theory opens the possibility of developing non-kernel-based SBF techniques. Also, the minimal conditions do not ask for the data being identically distributed or independent. Thus, the theory even allows dependent data such as sequential/spatial data. We believe that this new framework makes an important contribution to various future studies, including those on non-kernel-based SBF methods for various structured nonparametric regression models.

Second, we derive the rates of convergence and the asymptotic distribution of the estimator for the model (1.1). The theoretical developments here are based on kernel weighting schemes for the estimation of the marginal densities and the marginal regression maps. They are much more complex than in [15]. The latter work considers only continuous predictors and completely observed responses, and thus all stochastic terms from the asymptotic expansion of their estimator are based on kernels of the same type and bandwidths of the same size. This enables one to apply a unified and relatively simple approach to deriving the theoretical results. In our case, however, there are many more stochastic terms of different natures that arise from different kernel weighting schemes for different types of predictors with smoothing parameters of different rates, and they involve errors due to incompleteness in the response variables. In addition, the component maps for the discrete predictors are not differentiable, so that the techniques dealing with such terms are quite different from those for continuous predictors. Thus, various and sophisticated technical tools are required in our new setting, see the technical details contained in the Appendix A.7.

In Section 2 we describe our methodology in the above general framework. In Section 3 we obtain minimal conditions on the estimators of the marginal densities and regression maps under which we prove the existence of the SBF estimator and the convergence of the SBF algorithm in the general framework. In Section 4 we then specialize these results to the model (1.1), and investigate further the asymptotic properties of the corresponding SBF estimator. We treat in Section A.1 the case where there is no continuous predictor. In Section 5 we demonstrate the numerical superiority of our method and its usefulness in real problems. Auxiliary theoretical results and all proofs are in the Appendix.

2. Methodology for general predictors

2.1. Some examples of Hilbert spaces and vector operations

We give three examples of separable Hilbert spaces. These spaces and Euclidean spaces are the spaces we consider for the response variable $\mathbf{Y}$ in our numerical study.

(i) L^2 *space*. Let S be a Borel subset of $\mathbb{R}^k$. Consider the space of square integrable functions defined on S. For this space the zero vector $\mathbf{0}$ is the identically zero function, and for a scalar $c \in \mathbb{R}$ and for two functions $\mathbf{f} = f(\cdot)$ and $\mathbf{g} = g(\cdot)$, the vector addition $\mathbf{f} \oplus \mathbf{g}$ and scalar multiplication $c \odot \mathbf{f}$ are defined

by $\mathbf{f} \oplus \mathbf{g} = f(\cdot) + g(\cdot)$ and $c \odot \mathbf{f} = c \cdot f(\cdot)$. The inner product and norm are

$$\langle \mathbf{f}, \mathbf{g} \rangle = \int_S f(\mathbf{s}) \cdot g(\mathbf{s})\, d\mathbf{s}, \quad \|\mathbf{f}\| = \left(\int_S f(\mathbf{s})^2\, d\mathbf{s} \right)^{1/2}.$$

(ii) *The space of density functions.* Consider the space of probability density functions f supported on a Borel subset S of $\mathbb{R}^k$ with finite Lebesgue measure such that $\int_S (\log f(\mathbf{s}))^2\, d\mathbf{s} < \infty$. For this space, the zero vector $\mathbf{0}$ is the constant density $f_0(\cdot) \equiv (\mathrm{Leb_k}(S))^{-1}$, where $\mathrm{Leb_k}$ denotes the k-dimensional Lebesgue measure. For a scalar $c \in \mathbb{R}$ and for two densities $\mathbf{f} = f(\cdot)$ and $\mathbf{g} = g(\cdot)$, the vector addition $\mathbf{f} \oplus \mathbf{g}$ and scalar multiplication $c \odot \mathbf{f}$ are defined by

$$\mathbf{f} \oplus \mathbf{g} = \frac{f(\cdot) \cdot g(\cdot)}{\int_S f(\mathbf{s}) \cdot g(\mathbf{s})\, d\mathbf{s}}, \quad c \odot \mathbf{f} = \frac{(f(\cdot))^c}{\int_S (f(\mathbf{s}))^c\, d\mathbf{s}}.$$

The inner product and norm are

$$\langle \mathbf{f}, \mathbf{g} \rangle = \frac{1}{2\mathrm{Leb}_k(S)} \int_{S^2} \log\left(\frac{f(\mathbf{s})}{f(\mathbf{s}')} \right) \log\left(\frac{g(\mathbf{s})}{g(\mathbf{s}')} \right) d\mathbf{s}\, d\mathbf{s}',$$

$$\|\mathbf{f}\| = \left(\frac{1}{2\mathrm{Leb}_k(S)} \int_{S^2} \left[\log\left(\frac{f(\mathbf{s})}{f(\mathbf{s}')} \right) \right]^2 d\mathbf{s}\, d\mathbf{s}' \right)^{1/2}.$$

The space with the inner product forms an infinite-dimensional separable Hilbert space, as proved by [39].

(iii) *The space of compositional vectors.* Consider the space

$$\mathcal{S}^k = \left\{ (a_1, \cdots, a_k) \in (0,1)^k : \sum_{j=1}^k a_j = 1 \right\}.$$

For this space, the zero vector $\mathbf{0}$ is the compositional vector $(1/k, \ldots, 1/k)$ of equalized components. For a scalar $c \in \mathbb{R}$ and two compositional vectors $\mathbf{a}, \mathbf{b} \in \mathcal{S}^k$, the vector addition $\mathbf{a} \oplus \mathbf{b}$ and scalar multiplication $c \odot \mathbf{a}$ are defined by

$$\mathbf{a} \oplus \mathbf{b} = \left(\frac{a_1 \cdot b_1}{a_1 \cdot b_1 + \cdots + a_k \cdot b_k}, \ldots, \frac{a_k \cdot b_k}{a_1 \cdot b_1 + \cdots + a_k \cdot b_k} \right),$$

$$c \odot \mathbf{a} = \left(\frac{a_1^c}{a_1^c + \cdots + a_k^c}, \ldots, \frac{a_k^c}{a_1^c + \cdots + a_k^c} \right).$$

The inner product and norm are

$$\langle \mathbf{a}, \mathbf{b} \rangle = \frac{1}{2k} \sum_{j=1}^k \sum_{l=1}^k \log(a_j/a_l) \log(b_j/b_l),$$

$$\|\mathbf{a}\| = \left(\frac{1}{2k} \sum_{j=1}^k \sum_{l=1}^k \left[\log(a_j/a_l) \right]^2 \right)^{1/2}.$$

It is known that $\mathcal{S}^k$ with the inner product forms a $(k-1)$-dimensional Hilbert space.

2.2. General setting

Here, we consider abstract predictors taking values in general σ-finite measure spaces. We take this route, rather than starting with the specific types of predictors in the model (1.1), for simpler but better exposition of the main idea and also for demonstrating the broad scope of application in terms of the types of predictors we may cover. In technical terms, the general treatment is not direct from the existing literature, however, but actually requires thorough investigation of the way how the SBF idea works.

We let $(\mathcal{Z}_j, \mathscr{A}_j, \nu_j)$, for $1 \le j \le d$, be σ-finite measure spaces. We define $(\mathcal{Z}, \mathscr{A}, \nu) = (\prod_{j=1}^d \mathcal{Z}_j, \bigotimes_{j=1}^d \mathscr{A}_j, \bigotimes_{j=1}^d \nu_j)$, where $\bigotimes_{j=1}^d \mathscr{A}_j$ and $\bigotimes_{j=1}^d \nu_j$ are the product σ-field and product measure, respectively. We let $\mathbf{Z} = (Z_1, \ldots, Z_d)$ be a $\mathcal{Z}$-valued predictor and $\boldsymbol{\epsilon}$ be a $\mathbb{H}$-valued error satisfying $\mathrm{E}(\boldsymbol{\epsilon}|\mathbf{Z}) = \mathbf{0}$ and $\mathrm{E}(\|\boldsymbol{\epsilon}\|^2) < \infty$, where $\|\cdot\|$ denotes a norm of $\mathbb{H}$. We consider the following general additive model:

$$\mathbf{Y} = \mathbf{m}_0 \oplus \bigoplus_{j=1}^d \mathbf{m}_j(Z_j) \oplus \boldsymbol{\epsilon}, \tag{2.1}$$

where $\mathbf{Y}$ is a $\mathbb{H}$-valued response, $\mathbf{m}_0$ is a constant in $\mathbb{H}$ and $\mathbf{m}_j : \mathcal{Z}_j \to \mathbb{H}$ are measurable maps such that

$$\mathrm{E}(\|\mathbf{m}_j(Z_j)\|^2) < \infty, \quad 1 \le j \le d. \tag{2.2}$$

We let $P\mathbf{Z}^{-1}$ denote the distribution of $\mathbf{Z}$ defined by $P\mathbf{Z}^{-1}(A) = P(\mathbf{Z} \in A)$ for $A \in \mathscr{A}$. Likewise, we define $PZ_j^{-1}(A_j) = P(Z_j \in A_j)$ for $A_j \in \mathscr{A}_j$. We assume that $P\mathbf{Z}^{-1}$ is absolutely continuous with respect to ν. We write $dP\mathbf{Z}^{-1}/d\nu = p$, $\int_{\mathcal{Z}_{-jk}} p(\mathbf{z}) d\nu_{-jk}(\mathbf{z}_{-jk}) = p_{jk}(z_j, z_k)$ and $\int_{\mathcal{Z}_{-j}} p(\mathbf{z}) d\nu_{-j}(\mathbf{z}_{-j}) = p_j(z_j)$ for $1 \le j \ne k \le d$, where $(\mathcal{Z}_{-jk}, \mathscr{A}_{-jk}, \nu_{-jk})$ and $(\mathcal{Z}_{-j}, \mathscr{A}_{-j}, \nu_{-j})$ are the respective product measure spaces resulting from omitting the (j,k)th and the jth measure spaces in $(\mathcal{Z}, \mathscr{A}, \nu)$, and $\mathbf{z}_{-jk}$ and $\mathbf{z}_{-j}$ are the respective vectors resulting from omitting (z_j, z_k) and z_j in $\mathbf{z} = (z_1, \ldots, z_d)$.

To take into account various situations where $\mathbf{Y}$ is not completely observed, we consider a 'synthetic' or 'surrogate' response that replaces $\mathbf{Y}$. Such a synthetic response can be obtained by some mean-preserving transformation of 'observed' $\mathbf{Y}$, which may involve the observed predictor $\mathbf{Z}$. For instance, suppose that $\mathbf{Y}$ is subject to missingness. Let $R = 0$ if $\mathbf{Y}$ is missing, and $R = 1$ otherwise. Then, under the Hilbertian MAR (missing at random) condition, $R \perp \mathbf{Y}|\mathbf{Z}$, it holds that

$$\mathrm{E}((1/\pi(\mathbf{Z})) \odot \mathbf{Y}^*|\mathbf{Z}) = \mathrm{E}(\mathbf{Y}|\mathbf{Z}),$$

where $\pi(\mathbf{Z}) = P(R = 1|\mathbf{Z})$, $\odot$ denotes a scalar multiplication on $\mathbb{H}$, and $\mathbf{Y}^* = \mathbf{Y}$ if $R = 1$ and $\mathbf{Y}^* = \mathbf{0}$ otherwise. In this case, we may take $\boldsymbol{\psi}(\mathbf{Z}, \mathbf{Y}^*) := (1/\pi(\mathbf{Z})) \odot \mathbf{Y}^*$ as a surrogate of $\mathbf{Y}$. Another example arises in randomly right-censored regression where $\mathbf{Y}$ is a real-valued survival time subject to censoring, see Example 2 in Section 4.

In case $\mathbf{Y}$ is not completely observed, we assume that there exists a completely observable variable $\mathbf{Y}^* \in \mathbb{H}$ and a $\mathbb{H}$-valued transformation $\boldsymbol{\psi}$ satisfying

$$\mathrm{E}(\boldsymbol{\psi}(\mathbf{Z}, \mathbf{Y}^*)|\mathbf{Z}) = \mathrm{E}(\mathbf{Y}|\mathbf{Z}). \tag{2.3}$$

The equation (2.3) may be used to estimate the additive model based on the observed values of $\boldsymbol{\psi}(\mathbf{Z}, \mathbf{Y}^*)$. However, the transformation $\boldsymbol{\psi}$ usually contains unknown parameters or functions. In the missingness example discussed above, the conditional probability π is unknown. In the censoring example as well, the corresponding $\boldsymbol{\psi}$ involves the distribution function of the censoring variable that is unknown, see Example 2. In such cases, we need to estimate $\boldsymbol{\psi}$. We study the effect of the error in the estimation of $\boldsymbol{\psi}$ on the estimation of the additive model (1.1), see Section 4. Below, we describe our method and theory in terms of $\boldsymbol{\psi}(\mathbf{Z}, \mathbf{Y}^*)$ and its estimator. In the case of completely observed $\mathbf{Y}$, one may simply set $\boldsymbol{\psi}(\mathbf{Z}, \mathbf{Y}^*) = \hat{\boldsymbol{\psi}}(\mathbf{Z}, \mathbf{Y}^*) = \mathbf{Y}$.

2.3. General Bochner SBF estimation

For the estimation of the model (2.1) we first define some relevant spaces of $\mathbb{H}$-valued measurable maps. For any measure space (S, Σ, λ), we define

$$L^2((S, \Sigma, \lambda), \mathbb{H}) = \Big\{\mathbf{f} : S \to \mathbb{H} \,:\, \mathbf{f} \text{ is measurable and } \int_S \|\mathbf{f}(s)\|^2 d\lambda(s) < \infty\Big\}.$$

We note that the measure spaces on which the component maps $\mathbf{m}_j$ and the sum map $\bigoplus_{j=1}^d \mathbf{m}_j$ in (2.1) are defined, correspond to $(S, \Sigma, \lambda) = (\mathcal{Z}_j, \mathscr{A}_j, PZ_j^{-1})$ and $(S, \Sigma, \lambda) = (\mathcal{Z}, \mathscr{A}, P\mathbf{Z}^{-1})$, respectively.

For the identifiability of the component maps $\mathbf{m}_j$ in (2.1), we put the constraints $\mathrm{E}(\mathbf{m}_j(Z_j)) = \mathbf{0}$ for all $1 \le j \le d$, which entails $\mathbf{m}_0 = \mathrm{E}(\mathbf{Y})$. We note that these constraints can be written in *Bochner integrals*. The notion of Bochner integral generalizes that of the conventional Lebesgue integral to maps taking values in Hilbert or more generally in Banach spaces. By Propositions 2.1 and 2.2 in [15], the expected values $\mathrm{E}(\mathbf{m}_j(Z_j))$ and also the conditional expected values $\mathrm{E}(\mathbf{m}_k(Z_k)|Z_j = z_j)$ for $k \neq j$ may be written in Bochner integrals as

$$\begin{aligned} \mathrm{E}(\mathbf{m}_j(Z_j)) &= \int_{\mathcal{Z}_j} \mathbf{m}_j(z_j) \odot p_j(z_j) d\nu_j(z_j), \\ \mathrm{E}(\mathbf{m}_k(Z_k)|Z_j = z_j) &= \int_{\mathcal{Z}_k} \mathbf{m}_k(z_k) \odot \frac{p_{jk}(z_j, z_k)}{p_j(z_j)} d\nu_k(z_k). \end{aligned} \tag{2.4}$$

Here and below, we often write $\mathbf{h} \odot c$ for the scalar multiplication $c \odot \mathbf{h}$ of $\mathbf{h} \in \mathbb{H}$ and $c \in \mathbb{R}$. The representations at (2.4) are valid if (2.2) and the following assumptions on p_j and p_{jk} hold.

Condition (P). *For all* $1 \le j \neq k \le d$ *and* $z_j \in \mathcal{Z}_j, p_j(z_j) > 0$,

$$\int_{\mathcal{Z}_k} \frac{p_{jk}^2(z_j, z_k)}{p_k(z_k)} d\nu_k(z_k) < \infty \quad \textit{and} \quad \int_{\mathcal{Z}_k} \int_{\mathcal{Z}_j} \frac{p_{jk}^2(z_j, z_k)}{p_j(z_j) p_k(z_k)} d\nu_j(z_j) d\nu_k(z_k) < \infty.$$

From the first part at (2.4), the constraints on $\mathbf{m}_j$ are equivalent to

$$\int_{\mathcal{Z}_j} \mathbf{m}_j(z_j) \odot p_j(z_j) d\nu_j(z_j) = \mathbf{0}, \quad 1 \le j \le d. \tag{2.5}$$

We assume $\mathrm{E}(\|\boldsymbol{\psi}(\mathbf{Z}, \mathbf{Y}^*)\|^2) < \infty$ and define

$$\boldsymbol{\mu}(\mathbf{z}) = \mathrm{E}(\boldsymbol{\psi}(\mathbf{Z}, \mathbf{Y}^*)|\mathbf{Z} = \mathbf{z}), \quad \boldsymbol{\mu}_j(z_j) = \mathrm{E}(\boldsymbol{\psi}(\mathbf{Z}, \mathbf{Y}^*)|Z_j = z_j). \tag{2.6}$$

Here, the *marginal regression map* $\boldsymbol{\mu}_j$ is not equal to the component map $\mathbf{m}_j$. The model (2.1) with the constraints (2.5) and the representations of the conditional expectations at (2.4) entail that, under the assumptions (2.2) and (P),

$$\mathbf{m}_j(z_j) = \boldsymbol{\mu}_j(z_j) \ominus \mathbf{m}_0 \ominus \bigoplus_{k \ne j} \int_{\mathcal{Z}_k} \mathbf{m}_k(z_k) \odot \frac{p_{jk}(z_j, z_k)}{p_j(z_j)} d\nu_k(z_k), \tag{2.7}$$

for all $z_j \in \mathcal{Z}_j$ and $1 \le j \le d$, where $\ominus$ is defined by $\mathbf{h}_1 \ominus \mathbf{h}_2 = \mathbf{h}_1 \oplus (-1 \odot \mathbf{h}_2)$ for $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{H}$. Our method of estimating the component maps $\mathbf{m}_j$, which we detail below, is based on the above system of Bochner integral equations. In fact, we estimate the unknown quantities $\boldsymbol{\mu}_j(z_j), \mathbf{m}_0, p_j(z_j)$ and $p_{jk}(z_j, z_k)$ in the system of equations (2.7) to obtain our estimators of the component maps. For this, we consider *general* estimators of $\boldsymbol{\mu}_j$, $\mathbf{m}_0$, p_j and p_{jk} that have no specific forms.

We let $\hat{p}$ be *any* nonnegative estimator of p satisfying

$$\int_{\mathcal{Z}} \hat{p}(\mathbf{z}) d\nu(\mathbf{z}) = 1. \tag{2.8}$$

An example of $\hat{p}$ specialized for the predictors in (1.1) is given in Section 4.2. For $1 \le j \ne k \le d$, we define

$$\hat{p}_{jk}(z_j, z_k) = \int_{\mathcal{Z}_{-jk}} \hat{p}(\mathbf{z}) d\nu_{-jk}(\mathbf{z}_{-jk}), \quad \hat{p}_j(z_j) = \int_{\mathcal{Z}_{-j}} \hat{p}(\mathbf{z}) d\nu_{-j}(\mathbf{z}_{-j}). \tag{2.9}$$

Define a probability measure $\hat{P}\mathbf{Z}^{-1}$ on $\mathscr{A}$ by $\hat{P}\mathbf{Z}^{-1}(A) = \int_A \hat{p}(\mathbf{z}) d\nu(\mathbf{z})$. We also let $\hat{\boldsymbol{\mu}}$ be *any* estimator of $\boldsymbol{\mu}$, as defined at (2.6), satisfying

$$\begin{gathered} \hat{\boldsymbol{\mu}} \in L^1((\mathcal{Z}, \mathscr{A}, \hat{P}\mathbf{Z}^{-1}), \mathbb{H}), \\ \int_{\mathcal{Z}_j} \left\| \int_{\mathcal{Z}_{-j}} \hat{\boldsymbol{\mu}}(\mathbf{z}) \odot \hat{p}(\mathbf{z}) d\nu_{-j}(\mathbf{z}_{-j}) \right\|^2 \hat{p}_j(z_j)^{-1} d\nu_j(z_j) < \infty, \quad 1 \le j \le d. \end{gathered} \tag{2.10}$$

$\hat{\boldsymbol{\mu}}$ is a temporary estimator that induces our regression estimator, and it can be a *full-dimensional* estimator that does not take into account the additive structure of the model (2.1), such as the one considered in Section 4.2.

For the estimation of $\boldsymbol{\mu}_j$, we note that

$$\begin{aligned}
&\int_{\mathcal{Z}_{-j}} \boldsymbol{\mu}(\mathbf{z}) \odot p(\mathbf{z}) d\nu_{-j}(\mathbf{z}_{-j}) \\
&\quad = \int_{\mathcal{Z}_{-j}} \mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)|\mathbf{Z} = \mathbf{z}) \odot p(\mathbf{z}) d\nu_{-j}(\mathbf{z}_{-j}) \\
&\quad = \mathrm{E}(\mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)|\mathbf{Z})|Z_j = z_j) \odot p_j(z_j) \\
&\quad = \mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)|Z_j = z_j) \odot p_j(z_j).
\end{aligned} \tag{2.11}$$

Motivated by (2.11), we estimate $\boldsymbol{\mu}_j = \mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)|Z_j = \cdot)$ by

$$\hat{\boldsymbol{\mu}}_j(z_j) = \int_{\mathcal{Z}_{-j}} \hat{\boldsymbol{\mu}}(\mathbf{z}) \odot (\hat{p}(\mathbf{z})/\hat{p}_j(z_j)) d\nu_{-j}(\mathbf{z}_{-j}) \tag{2.12}$$

whenever the integral exists, and we set $\hat{\boldsymbol{\mu}}_j(z_j) = \mathbf{0}$ otherwise. We note that the integral on the right hand side of (2.12) exists for almost everywhere z_j in the measure ν_j under the first condition at (2.10). Furthermore,

$$\hat{\boldsymbol{\mu}}_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$$

under the second condition at (2.10), where $\hat{P}Z_j^{-1}$ is the probability measure on $\mathscr{A}_j$ defined by $\hat{P}Z_j^{-1}(A_j) = \int_{A_j} \hat{p}_j(z_j) d\nu_j(z_j)$. The square integrability of $\hat{\boldsymbol{\mu}}_j$ is required in our theoretical developments, such as at (3.4) in Section 3, for example. For the unknown Hilbertian constant $\mathbf{m}_0 = \mathrm{E}(\mathbf{Y}) = \mathrm{E}(\boldsymbol{\mu}(\mathbf{Z})) = \int_{\mathcal{Z}} \boldsymbol{\mu}(\mathbf{z}) \odot p(\mathbf{z}) d\nu(\mathbf{z})$, we choose $\hat{\mathbf{m}}_0 = \int_{\mathcal{Z}} \hat{\boldsymbol{\mu}}(\mathbf{z}) \odot \hat{p}(\mathbf{z}) d\nu(\mathbf{z})$. Then, by (2.12) it holds that

$$\int_{\mathcal{Z}_j} \hat{\boldsymbol{\mu}}_j(z_j) \odot \hat{p}_j(x_j) d\nu_j(z_j) = \hat{\mathbf{m}}_0, \quad 1 \le j \le d. \tag{2.13}$$

Now, we define our estimator of $\boldsymbol{\mu} = \mathbf{m}_0 \oplus \bigoplus_{j=1}^d \mathbf{m}_j$ by

$$\hat{\boldsymbol{\mu}}_+ = \hat{\mathbf{m}}_0 \oplus \bigoplus_{j=1}^d \hat{\mathbf{m}}_j, \tag{2.14}$$

where $(\hat{\mathbf{m}}_1, \ldots, \hat{\mathbf{m}}_d) \in \prod_{j=1}^d L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$ is a solution of the system of Bochner integral equations

$$\hat{\mathbf{m}}_j(z_j) = \hat{\boldsymbol{\mu}}_j(z_j) \ominus \hat{\mathbf{m}}_0 \ominus \bigoplus_{k \ne j} \int_{\mathcal{Z}_k} \hat{\mathbf{m}}_k(z_k) \odot \frac{\hat{p}_{jk}(z_j, z_k)}{\hat{p}_j(z_j)} d\nu_k(z_k), \quad 1 \le j \le d. \tag{2.15}$$

We note that (2.15) is an estimating equation obtained by substituting $\hat{\boldsymbol{\mu}}_j$, $\hat{\mathbf{m}}_0$, $\hat{p}_j$ and $\hat{p}_{jk}$ for $\boldsymbol{\mu}_j$, $\mathbf{m}_0$, p_j and p_{jk} in (2.7). In Section 3, we show that the Bochner integrals in (2.15) are well-defined and the system of equations has

a unique solution $\hat{\boldsymbol{\mu}}_+$ under weak conditions. We also demonstrate that each component $\hat{\mathbf{m}}_j$ of $\hat{\boldsymbol{\mu}}_+$ is uniquely determined as an estimator of $\mathbf{m}_j$ under some condition and the constraints

$$\int_{\mathcal{Z}_j} \hat{\mathbf{m}}_j(z_j) \odot \hat{p}_j(z_j) d\nu_j(z_j) = \mathbf{0}, \quad 1 \le j \le d. \tag{2.16}$$

We note that (2.16) is an empirical version of (2.5). We call (2.15) the *general Bochner smooth backfitting (gB-SBF)* system of equations. We also call $\hat{\boldsymbol{\mu}}_+$ and $(\hat{\mathbf{m}}_1, \ldots, \hat{\mathbf{m}}_d)$ the *gB-SBF estimators* of $\boldsymbol{\mu}$ and $(\mathbf{m}_1, \ldots, \mathbf{m}_d)$, respectively.

2.4. General Bochner SBF algorithm

The gB-SBF estimators have no closed form. Hence, to evaluate the estimators, we need an iteration scheme. For an initial estimator in the iteration scheme, we take $(\hat{\mathbf{m}}_1^{[0]}, \ldots, \hat{\mathbf{m}}_d^{[0]})$ satisfying $(\hat{\mathbf{m}}_1^{[0]}, \ldots, \hat{\mathbf{m}}_d^{[0]}) \in \prod_{j=1}^d L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P} Z_j^{-1}), \mathbb{H})$ and the constraints (2.16) for $\hat{\mathbf{m}}_j = \hat{\mathbf{m}}_j^{[0]}$. An immediate choice is $(\hat{\mathbf{m}}_j^{[0]}, \ldots, \hat{\mathbf{m}}_j^{[0]}) \equiv (\mathbf{0}, \ldots, \mathbf{0})$, which obviously satisfies the constraints. Another option is to take $\hat{\mathbf{m}}_j^{[0]} = \hat{\boldsymbol{\mu}}_j - \hat{\mathbf{m}}_0$, which also satisfies the constraints because of (2.13). Put $\hat{\boldsymbol{\mu}}_+^{[0]} = \hat{\mathbf{m}}_0 \oplus \bigoplus_{j=1}^d \hat{\mathbf{m}}_j^{[0]}$. For subsequent updates we apply the gB-SBF system of equations sequentially from $j = 1$ to $j = d$. A step-by-step procedure is described below, which we call the *gB-SBF algorithm*. For $r \ge 0$, $1 \le j \le d$ and a given set of $\hat{\mathbf{m}}_1^{[r]}, \ldots, \hat{\mathbf{m}}_d^{[r]}$, define

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{+j}^{[r]}(z_j) &= \bigoplus_{k \le j-1} \int_{\mathcal{Z}_k} \hat{\mathbf{m}}_k^{[r]}(z_k) \odot \frac{\hat{p}_{jk}(z_j, z_k)}{\hat{p}_j(z_j)} d\nu_k(z_k), \quad 2 \le j \le d, \\ \hat{\boldsymbol{\mu}}_{j+}^{[r]}(z_j) &= \bigoplus_{k \ge j+1} \int_{\mathcal{Z}_k} \hat{\mathbf{m}}_k^{[r]}(z_k) \odot \frac{\hat{p}_{jk}(z_j, z_k)}{\hat{p}_j(z_j)} d\nu_k(z_k), \quad 1 \le j \le d-1, \end{aligned} \tag{2.17}$$

with $\hat{\boldsymbol{\mu}}_{+1}^{[r]} \equiv \mathbf{0} \equiv \hat{\boldsymbol{\mu}}_{d+}^{[r]}$ for all $r \ge 0$.

gB-SBF algorithm.

Initialization: Choose an initial estimate $(\hat{\mathbf{m}}_1^{[0]}, \ldots, \hat{\mathbf{m}}_d^{[0]})$.
Iteration: For $r \ge 1$,

(i) compute $\hat{\mathbf{m}}_j^{[r]}$ for $j = 1, \ldots, d$ according to

$$\hat{\mathbf{m}}_j^{[r]}(z_j) = \hat{\boldsymbol{\mu}}_j(z_j) \ominus \hat{\mathbf{m}}_0 \ominus \hat{\boldsymbol{\mu}}_{+j}^{[r]}(z_j) \ominus \hat{\boldsymbol{\mu}}_{j+}^{[r-1]}(z_j);$$

(ii) compute $\hat{\boldsymbol{\mu}}_+^{[r]} = \hat{\mathbf{m}}_0 \oplus \bigoplus_{j=1}^d \hat{\mathbf{m}}_j^{[r]}$.

Ending: Stop the iteration if $\int_{\mathcal{Z}} \|\hat{\boldsymbol{\mu}}_+^{[r]}(\mathbf{z}) \ominus \hat{\boldsymbol{\mu}}_+^{[r-1]}(\mathbf{z})\|^2 \hat{p}(\mathbf{z}) d\nu(\mathbf{z})$ is sufficiently small. □

The Bochner integrals at (2.17) are well-defined under the condition (S2) to be given in Section 3.1. Indeed, we may prove that $(\hat{\mathbf{m}}_1^{[r]}, \ldots, \hat{\mathbf{m}}_d^{[r]})$ belongs to $\prod_{j=1}^d L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$. In addition, all the subsequent updates $(\hat{\mathbf{m}}_1^{[r]}, \ldots, \hat{\mathbf{m}}_d^{[r]})$ for $r \geq 1$ satisfy the constraints (2.16) for $\hat{\mathbf{m}}_j = \hat{\mathbf{m}}_j^{[r]}$. In Section 3.3, we discuss the convergence of $\hat{\boldsymbol{\mu}}_+^{[r]}$ to $\hat{\boldsymbol{\mu}}_+$ and $(\hat{\mathbf{m}}_1^{[r]}, \ldots, \hat{\mathbf{m}}_d^{[r]})$ to $(\hat{\mathbf{m}}_1, \ldots, \hat{\mathbf{m}}_d)$ as $r \to \infty$.

2.5. gB-SBF for Euclidean and functional responses

The Euclidean and L^2 spaces are two common types of Hilbert spaces. The cases with compositional and density responses introduced in Section 2.1 may be treated within these spaces by executing some transformations, see the Appendix A.3. For $\mathbb{H} = \mathbb{R}^D$ with $D \geq 1$, the component and marginal regression maps $\hat{\mathbf{m}}_j$ and $\hat{\boldsymbol{\mu}}_j$ are $\mathbb{R}^D$-valued functions. For instance, $\hat{\mathbf{m}}_j(\cdot) = (\hat{m}_{j1}(\cdot), \ldots, \hat{m}_{jD}(\cdot))^\top$, where $\hat{m}_{jl}$ for $1 \leq l \leq D$ are real-valued functions. Then, writing $\hat{\mathbf{m}}_0 = (\hat{m}_{01}, \ldots, \hat{m}_{0D})^\top \in \mathbb{R}^D$, the gB-SBF equations at (2.15) reduce to

$$\begin{pmatrix} \hat{m}_{j1}(z_j) \\ \vdots \\ \hat{m}_{jD}(z_j) \end{pmatrix} = \begin{pmatrix} \hat{\mu}_{j1}(z_j) \\ \vdots \\ \hat{\mu}_{jD}(z_j) \end{pmatrix} - \begin{pmatrix} \hat{m}_{01} \\ \vdots \\ \hat{m}_{0D} \end{pmatrix} - \sum_{k \neq j} \int_{\mathcal{Z}_k} \begin{pmatrix} \hat{m}_{k1}(z_k) \\ \vdots \\ \hat{m}_{kD}(z_k) \end{pmatrix} \times \frac{\hat{p}_{jk}(z_j, z_k)}{\hat{p}_j(z_j)} d\nu_k(z_k). \tag{2.18}$$

Now, in case $\mathbb{H}$ is $L^2(S)$, the space of square integrable real-valued functions defined on $S \subset \mathbb{R}^D$ for some $D \geq 1$, we may write $\hat{\mathbf{m}}_j(z_j) = \hat{m}_j(z_j, \cdot)$, where $\hat{m}_j : \mathcal{Z}_j \times S \to \mathbb{R}$. Likewise, $\hat{\boldsymbol{\mu}}_j(z_j) = \hat{\mu}_j(z_j, \cdot)$ with $\hat{\mu}_j : \mathcal{Z}_j \times S \to \mathbb{R}$. Write $\hat{\mathbf{m}}_0 = \hat{m}_0(\cdot) : S \to \mathbb{R}$. Then, we may write the gB-SBF equations at (2.15) as

$$\hat{m}_j(z_j, \mathbf{s}) = \hat{\mu}_j(z_j, \mathbf{s}) - \hat{m}_0(\mathbf{s}) - \sum_{k \neq j} \int_{\mathcal{Z}_k} \hat{m}_k(z_k, \mathbf{s}) \times \frac{\hat{p}_{jk}(z_j, z_k)}{\hat{p}_j(z_j)} d\nu_k(z_k), \quad \mathbf{s} \in S. \tag{2.19}$$

The gB-SBF systems of equations at (2.18) and (2.19) can be implemented by the gB-SBF algorithm described in Section 2.4 with the specializations of the operations $\oplus$ and $\odot$.

3. Theory for general predictors

In this section, we prove the existence of the gB-SBF estimators and the convergence of the gB-SBF algorithm in various modes. These are basic properties we need to establish before we study other statistical properties of the gB-SBF method. For other backfitting-based methods [e.g., 3, 31, 30, 43], fairly strong conditions are imposed to guarantee the corresponding existence and

convergence results. Even the initial SBF work for scalar responses [25] assumes somewhat strong conditions to get such properties. We demonstrate the basic properties under much weaker conditions, by deep investigation of the SBF technique. Except for Theorem 3 and Corollary 1 in Section 3.3, the conditions are imposed on datasets rather than on the data generating model, and thus the corresponding theoretical results are non-asymptotic and valid for datasets satisfying the conditions. The data-specific results are more useful to practitioners since it is direct and feasible to check the data-specific conditions with a dataset at hand.

3.1. Minimal conditions

The system of Bochner integral equations at (2.15) involves the estimators $\hat{\boldsymbol{\mu}}_j$, $\hat{\mathbf{m}}_0$, $\hat{p}_j$ and $\hat{p}_{jk}$, all of which are determined by the estimators of the joint density p and regression map $\boldsymbol{\mu}$. We state a set of weak conditions for general estimators $\hat{p}$ and $\hat{\boldsymbol{\mu}}$, under which we prove the existence of the gB-SBF estimators and the convergence of the gB-SBF algorithm. The formulation of the conditions in this general framework, which boils down to those as given below, is non-trivial since it requires careful investigation of all the steps of the way how the SBF technique works theoretically.

Condition (S1). *The estimators $\hat{p}$ and $\hat{\boldsymbol{\mu}}$ satisfy (2.8) and (2.10).*

Condition (S2). *For all $1 \le j \ne k \le d$ and $z_j \in \mathcal{Z}_j, \hat{p}_j(z_j) > 0$,*

$$\int_{\mathcal{Z}_k} \frac{\hat{p}_{jk}^2(z_j, z_k)}{\hat{p}_k(z_k)} d\nu_k(z_k) < \infty \quad \textit{and} \quad \int_{\mathcal{Z}_k} \int_{\mathcal{Z}_j} \frac{\hat{p}_{jk}^2(z_j, z_k)}{\hat{p}_j(z_j)\hat{p}_k(z_k)} d\nu_j(z_j) d\nu_k(z_k) < \infty.$$

Note that the condition (S2) is an empirical version of the condition (P). In Section 4.2, we specialize the conditions for the specific estimators we consider in the case of mixed predictors.

3.2. Existence of gB-SBF estimators

In this subsection, we prove the existence and the uniqueness of the gB-SBF estimators. Let $\mathcal{S}^{\mathbb{H}}(\hat{p})$ be the 'sum-space' such that

$$\begin{aligned} \mathcal{S}^{\mathbb{H}}(\hat{p}) := & \left\{ \bigoplus_{j=1}^{d} \mathbf{f}_j : \mathbf{f}_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H}),\ 1 \le j \le d \right\} \\ & \subset L^2((\mathcal{Z}, \mathscr{A}, \hat{P}\mathbf{Z}^{-1}), \mathbb{H}), \end{aligned} \tag{3.1}$$

which is the space in which we seek the solution $\hat{\boldsymbol{\mu}}_+ = \hat{\mathbf{m}}_0 \oplus \bigoplus_{j=1}^{d} \hat{\mathbf{m}}_j$ of (2.15). To give a key idea for the proof of the existence of the solution, consider the

functional $F : L^2((\mathcal{Z}, \mathscr{A}, P\mathbf{Z}^{-1}), \mathbb{H}) \to \mathbb{R}$ defined by

$$\begin{aligned} F(\mathbf{f}) &= \mathrm{E}(\|\boldsymbol{\psi}(\mathbf{Z}, \mathbf{Y}^*) \ominus \mathbf{f}(\mathbf{Z})\|^2) \\ &= \int_{\mathcal{Z}} \|\mathbf{f}(\mathbf{z})\|^2 p(\mathbf{z}) d\nu(\mathbf{z}) - 2 \int_{\mathcal{Z}} \langle \mathbf{f}(\mathbf{z}), \boldsymbol{\mu}(\mathbf{z}) \rangle p(\mathbf{z}) d\nu(\mathbf{z}) \\ &\quad + \int_{\mathcal{Z}} \mathrm{E}(\|\boldsymbol{\psi}(\mathbf{Z}, \mathbf{Y}^*)\|^2 | \mathbf{Z} = \mathbf{z}) p(\mathbf{z}) d\nu(\mathbf{z}). \end{aligned} \tag{3.2}$$

The true regression map $\boldsymbol{\mu} = \mathbf{m}_0 \oplus \bigoplus_{j=1}^d \mathbf{m}_j$ is the minimizer of (3.2). By Theorem 5.3.19 in [4], $\boldsymbol{\mu}$ satisfies $DF(\boldsymbol{\mu})(\mathbf{g}) = 0$ for all $\mathbf{g} \in L^2((\mathcal{Z}, \mathscr{A}, P\mathbf{Z}^{-1}), \mathbb{H})$, provided that the Gâteaux derivative $DF(\boldsymbol{\mu}) : L^2((\mathcal{Z}, \mathscr{A}, P\mathbf{Z}^{-1}), \mathbb{H}) \to \mathbb{R}$ of F at $\boldsymbol{\mu}$ exists. In this case, one may verify that $DF(\boldsymbol{\mu})(\cdot) \equiv 0$ induces (2.7), which is a population version of (2.15).

Based on the above observation, we formulate the existence of the gB-SBF estimators satisfying (2.15) as the existence of a minimizer of the objective functional $\hat{F} : \mathcal{S}^{\mathbb{H}}(\hat{p}) \to \mathbb{R}$ defined by

$$\hat{F}(\mathbf{f}) = \|\mathbf{f}\|_{2,n}^2 - 2 \int_{\mathcal{Z}} \langle \mathbf{f}(\mathbf{z}), \hat{\boldsymbol{\mu}}(\mathbf{z}) \rangle \, \hat{p}(\mathbf{z}) d\nu(\mathbf{z}). \tag{3.3}$$

We note that $\hat{F}$ is an empirical version of F with the last integral at (3.2), which is irrelevant in the minimization, being omitted. The functional $\hat{F}$ is well-defined on $\mathcal{S}^{\mathbb{H}}(\hat{p})$ since, for all $\mathbf{f} = \bigoplus_{j=1}^d \mathbf{f}_j \in \mathcal{S}^{\mathbb{H}}(\hat{p})$, it holds that

$$|\hat{F}(\mathbf{f})| \leq \|\mathbf{f}\|_{2,n}^2 + 2 \sum_{j=1}^d \|\mathbf{f}_j\|_{2,n} \left(\int_{\mathcal{Z}_j} \|\hat{\boldsymbol{\mu}}_j(z_j)\|^2 \hat{p}_j(z_j) d\nu_j(z_j) \right)^{1/2} < \infty \tag{3.4}$$

by (S1), where $\hat{\boldsymbol{\mu}}_j$ is defined at (2.12). Formally, we prove the following theorem.

Theorem 1. *Assume the conditions (S1) and (S2). Then, there exists a solution* $\hat{\boldsymbol{\mu}}_+ = \hat{\mathbf{m}}_0 \oplus \bigoplus_{j=1}^d \hat{\mathbf{m}}_j$ *with* $\hat{\mathbf{m}}_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$ *satisfying the system of equations at (2.15), and the solution is unique up to measure zero with respect to* $\hat{P}\mathbf{Z}^{-1}$. *Furthermore, each component* $\hat{\mathbf{m}}_j$ *is uniquely determined up to measure zero with respect to* ν_j *under the constraints (2.16), provided that* $\hat{p} > 0$ *on* $\mathcal{Z}$.

3.3. Convergence of gB-SBF algorithm

Here, we present the convergence of $\hat{\boldsymbol{\mu}}_+^{[r]} = \hat{\mathbf{m}}_0 \oplus \bigoplus_{j=1}^d \hat{\mathbf{m}}_j^{[r]}$ and $(\hat{\mathbf{m}}_1^{[r]}, \ldots, \hat{\mathbf{m}}_d^{[r]})$ in various modes. We have the following non-asymptotic result for the convergence of $\hat{\boldsymbol{\mu}}_+^{[r]}$ to $\hat{\boldsymbol{\mu}}_+$.

Theorem 2. *Assume that the conditions (S1) and (S2) hold. Then,*

$$\int_{\mathcal{Z}} \|\hat{\boldsymbol{\mu}}_+(\mathbf{z}) \ominus \hat{\boldsymbol{\mu}}_+^{[r]}(\mathbf{z})\|^2 \hat{p}(\mathbf{z}) d\nu(\mathbf{z}) \leq \hat{c}^* \cdot \hat{\gamma}^r \text{ for all } r \geq 0,$$

where $\hat{c}^* > 0$ *is a constant that does not depend on* r *but only on* $\hat{p}$, $\hat{\boldsymbol{\mu}}$ *and the initial estimator* $(\hat{\mathbf{m}}_j^{[0]} : 1 \le j \le d)$, *and* $\hat{\gamma} \in (0,1)$ *is a constant that does not depend on* r *but only on* $\hat{p}$.

The above theorem establishes that $\hat{\boldsymbol{\mu}}_+^{[r]}$ converges to $\hat{\boldsymbol{\mu}}_+$ at a geometric speed. The next theorem is an asymptotic version of Theorem 2, which deals with the convergence of the individual components $\hat{\mathbf{m}}_j^{[r]}$ to their respective targets $\hat{\mathbf{m}}_j$. For this, we introduce the following high-level conditions.

Condition (A). *The condition (2.8) and the first one at (2.10) hold with probability tending to one. Also, there exists a constant* $C > 0$ *such that*

$$\begin{aligned}
&\lim_{n\to\infty} P\Big(\max_{1\le j\le d} \sup_{z_j\in\mathcal{Z}_j} \hat{p}_j(z_j)^{-1} < C\Big) = 1,\\
&\lim_{n\to\infty} P\Big(\max_{1\le j\ne k\le d} \sup_{z_j\in\mathcal{Z}_j, z_k\in\mathcal{Z}_k} \hat{p}_{jk}(z_j,z_k) < C\Big) = 1,\\
&\lim_{n\to\infty} P\Big(\max_{1\le j\le d} \sup_{z_j\in\mathcal{Z}_j} \|\hat{\boldsymbol{\mu}}_j(z_j)\| < C\Big) = 1,\\
&\lim_{n\to\infty} P\Big(\max_{1\le j\le d} \int_{\mathcal{Z}_j} \|\hat{\mathbf{m}}_j^{[0]}(z_j)\|^2 d\nu_j(z_j) < C\Big) = 1,
\end{aligned} \tag{3.5}$$

and the one- and two-dimensional density estimators satisfy

$$\begin{aligned}
\max_{1\le j\le d} \int_{\mathcal{Z}_j} (\hat{p}_j(z_j) - p_j(z_j))^2 \, d\nu_j(z_j) &= o_p(1),\\
\max_{1\le j\ne k\le d} \int_{\mathcal{Z}_j\times\mathcal{Z}_k} (\hat{p}_{jk}(z_j,z_k) - p_{jk}(z_j,z_k))^2 \, d\nu_j\otimes\nu_k(z_j,z_k) &= o_p(1).
\end{aligned} \tag{3.6}$$

We note that the last condition at (3.5) for initial component estimators $\hat{\mathbf{m}}_j^{[0]}$ is not restrictive. It is satisfied by the choice $(\hat{\mathbf{m}}_1^{[0]},\dots,\hat{\mathbf{m}}_d^{[0]}) \equiv (\mathbf{0},\dots,\mathbf{0})$, for example. Others at (3.5) are mild conditions on the estimators of the marginal densities and regression maps. The conditions at (3.6) are some L^2-consistency conditions on the marginal density estimators.

Theorem 3. *Assume that* p *is bounded away from zero and infinity on* $\mathcal{Z}$ *and that* ν_j *are finite measures for all* $1 \le j \le d$. *Then, under the condition (A), there exist constants* $c^{**} > 0$ *and* $\gamma \in (0,1)$ *such that*

$$\lim_{n\to\infty} P\Big(\max_{1\le j\le d} \int_{\mathcal{Z}_j} \|\hat{\mathbf{m}}_j(z_j) \ominus \hat{\mathbf{m}}_j^{[r]}(z_j)\|^2 p_j(z_j) d\nu_j(z_j) \le c^{**}\cdot\gamma^r \text{ for all } r \ge 0\Big) = 1.$$

Theorem 3 is about the L^2-convergence of $\hat{\mathbf{m}}_j^{[r]}$ with a geometric rate. From the theorem we may deduce an almost everywhere convergence of $\hat{\mathbf{m}}_j^{[r]}$, which is also of interest. Indeed, Theorem 3 implies that

$$\sum_{r=1}^{\infty} \int_{\mathcal{Z}_j} \|\hat{\mathbf{m}}_j(z_j) \ominus \hat{\mathbf{m}}_j^{[r]}(z_j)\|^2 p_j(z_j) d\nu_j(z_j) < \infty$$

with probability tending to one. This entails that, with probability tending to one, $\sum_{r=1}^{\infty} \|\hat{\mathbf{m}}_j(z_j) \ominus \hat{\mathbf{m}}_j^{[r]}(z_j)\|^2 p_j(z_j) < \infty$ a.e. with respect to ν_j, which gives the following corollary.

Corollary 1. *Assume that the conditions in Theorem 3. Let E_j denote the set of z_j in $\mathcal{Z}_j$ such that $\hat{\mathbf{m}}_j^{[r]}(z_j) \to \hat{\mathbf{m}}_j(z_j)$ as $r \to \infty$. Then, for all $1 \le j \le d$, it holds that $\lim_{n\to\infty} P\left(\nu_j(\mathcal{Z}_j) = \nu_j(E_j)\right) = 1$.*

4. Theory for mixed predictors

This section deals with the specification of the general results in Section 3 to the model (1.1) with $d_x, d_u, d_v \ge 1$. The cases of $d_u = 0$ or $d_v = 0$ follow immediately with trivial modification. The case of $d_x = 0$ is treated in Section A.1. We consider X_j taking values in $[0,1]$ for $1 \le j \le d_x$, U_j with values in a finite set $\mathcal{U}_j$ for $1 \le j \le d_u$, and V_j with values in a metric space $\mathcal{V}_j$ with finite cardinality for $1 \le j \le d_v$. The latter general setting allows V_j to be an ordinal discrete predictor or a continuous predictor on a fixed design. We let $\mathbf{W} = (\mathbf{X}, \mathbf{U}, \mathbf{V})$ so that $\mathbf{W}$ takes the role of $\mathbf{Z}$ in the model (2.1). The product space $\mathcal{Z}$ for $\mathbf{Z}$ in (2.1) corresponds to $\mathcal{W} = [0,1]^{d_x} \times \prod_{j=1}^{d_u} \mathcal{U}_j \times \prod_{j=1}^{d_v} \mathcal{V}_j$ and the product measure ν in (2.1) is specialized to $\bigotimes_{j=1}^{d_x} \mathrm{Leb} \otimes \bigotimes_{j=1}^{d_u} \mathrm{C}_{u,j} \otimes \bigotimes_{j=1}^{d_v} \mathrm{C}_{v,j}$, where Leb is the Lebesgue measure on $\mathbb{R}$, $\mathrm{C}_{u,j}$ is the counting measure on $\mathcal{U}_j$ and $\mathrm{C}_{v,j}$ is the counting measure on $\mathcal{V}_j$. We continue to use p to denote the joint density of $\mathbf{W}$ with respect to $\bigotimes_{j=1}^{d_x} \mathrm{Leb} \otimes \bigotimes_{j=1}^{d_u} \mathrm{C}_{u,j} \otimes \bigotimes_{j=1}^{d_v} \mathrm{C}_{v,j}$, and use $\boldsymbol{\mu}$ to denote the regression map $\mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)|\mathbf{W} = \cdot)$ in the same spirit as in (2.6), where $\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)$ is a synthetic response that satisfies (2.3) for $\mathbf{Z} = \mathbf{W}$ and $\mathrm{E}(\|\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)\|^2) < \infty$. In Section 4.2, we exemplify $\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)$.

The marginalization of p along the coordinates that are of interest in $\mathcal{W}$ defines the densities of X_j, U_j, V_j, (X_j, X_k), (U_j, U_k), (V_j, V_k), (X_j, U_k), (X_j, V_k) and (U_j, V_k). We denote them, respectively, by $p_{x,j}$, $p_{u,j}$, $p_{v,j}$, $p_{xx,jk}$, $p_{uu,jk}$, $p_{vv,jk}$, $p_{xu,jk}$, $p_{xv,jk}$ and $p_{uv,jk}$. We denote the marginal regression maps, which correspond to $\boldsymbol{\mu}_j$ in (2.6), by $\boldsymbol{\mu}_{x,j}(x_j) = \mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)|X_j = x_j)$, $\boldsymbol{\mu}_{u,j}(u_j) = \mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)|U_j = u_j)$ and $\boldsymbol{\mu}_{v,j}(v_j) = \mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)|V_j = v_j)$. Below we discuss the estimation of these marginal densities and regression maps.

4.1. Estimation of marginal densities and regression maps

To estimate the joint density p and regression map $\boldsymbol{\mu}$, we use kernel-based estimators. For this, we introduce a kernel weighting scheme for each of the three types of predictors. First, for smoothing across on $[0,1]$ where X_j takes values, let $K_h(t) = K(t/h)/h$, where $h > 0$ is a bandwidth and $K : \mathbb{R} \to [0, \infty)$ is a baseline kernel function. Throughout this paper, we assume that K vanishes on $\mathbb{R} \setminus [-1, 1]$ and satisfies $\int_{-1}^{1} K(t)dt = 1$. Define a normalized kernel $K_h(x, x')$ by

$$K_h(x, x') = \frac{K_h(x - x')}{\int_0^1 K_h(t - x')dt} \tag{4.1}$$

whenever $\int_0^1 K_h(t-x')dt > 0$, and we set $K_h(x,x') = 0$ otherwise. This kernel has been used in the SBF literature [e.g., 25]. We note that it has the normalization property

$$\int_0^1 K_h(x,x')dx = 1 \quad \text{for all } x' \in [0,1]. \tag{4.2}$$

Now, for smoothing across $\mathcal{U}_j$ we take a discrete kernel $L_{\lambda_j} : \mathcal{U}_j \times \mathcal{U}_j \to [0,1]$ defined by

$$L_{\lambda_j}(u_j, u_j') = (1-\lambda_j)I(u_j = u_j') + (\lambda_j/(c_j-1))I(u_j \neq u_j'),$$

where $\lambda_j \in [0,1]$ is a smoothing parameter and c_j is the cardinality of $\mathcal{U}_j$. This kernel was introduced by [1]. We note that L_{λ_j} has the normalization property

$$\sum_{u_j \in \mathcal{U}_j} L_{\lambda_j}(u_j, u_j') = 1 \quad \text{for all } u_j' \in \mathcal{U}_j. \tag{4.3}$$

Next, for smoothing across $\mathcal{V}_j$ with a metric δ_j we define a new metric-based discrete kernel $W_{s_j} : \mathcal{V}_j \times \mathcal{V}_j \to [0,1]$ by

$$W_{s_j}(v_j, v_j') = \left(1 - \sum_{v_j'' \in \mathcal{V}_j : v_j'' \neq v_j'} s_j^{\delta_j(v_j'', v_j')}\right) I(v_j = v_j') + s_j^{\delta_j(v_j, v_j')} I(v_j \neq v_j'),$$

where $0 \leq s_j < 1$ is a smoothing parameter that is sufficiently small so that $0 \leq W_{s_j}(v_j, v_j') \leq 1$ for all $v_j, v_j' \in \mathcal{V}_j$. Basically, this kernel gives more weights when v_j' gets closer to v_j in the metric δ_j. It also has the normalization property

$$\sum_{v_j \in \mathcal{V}_j} W_{s_j}(v_j, v_j') = 1 \quad \text{for all } v_j' \in \mathcal{V}_j. \tag{4.4}$$

Now, suppose that we have n observations $\{(\mathbf{W}_i, \mathbf{Y}_i^*) : 1 \leq i \leq n\}$ which are not necessarily i.i.d. Writing $\mathbf{w} = (\mathbf{x}, \mathbf{u}, \mathbf{v})$ for vectors $\mathbf{x} \in [0,1]^{d_x}$, $\mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j$ and $\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j$, we let $\kappa_i(\mathbf{w}) = \prod_{j=1}^{d_x} K_{h_j}(x_j, X_{ij}) \cdot \prod_{j=1}^{d_u} L_{\lambda_j}(u_j, U_{ij}) \cdot \prod_{j=1}^{d_v} W_{s_j}(v_j, V_{ij})$. We estimate p by $\hat{p}(\mathbf{w}) = n^{-1} \sum_{i=1}^n \kappa_i(\mathbf{w})$ and $\boldsymbol{\mu}$ by

$$\hat{\boldsymbol{\mu}}(\mathbf{w}) = \begin{cases} (n \cdot \hat{p}(\mathbf{w}))^{-1} \odot \bigoplus_{i=1}^n (\kappa_i(\mathbf{w}) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*)), & \text{if } \hat{p}(\mathbf{w}) > 0 \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{4.5}$$

for an appropriate estimator $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$. We show that these $\hat{p}$ and $\hat{\boldsymbol{\mu}}$ satisfy the non-asymptotic condition (S1). Because of the normalization properties (4.2), (4.3) and (4.4), $\hat{p}$ clearly satisfies (2.8). Also, the full-dimensional estimator $\hat{\boldsymbol{\mu}}$ satisfies (2.10). To see this, for a vector $\mathbf{x} \in [0,1]^{d_x}$ let $\mathbf{x}_{-j}$ denote the $(d_x - 1)$-vector resulting from omitting the jth entry of $\mathbf{x}$ and likewise define $\mathbf{u}_{-j}$ and

$\mathbf{v}_{-j}$ for $\mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j$ and $\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j$, respectively. The first condition at (2.10) is satisfied since

$$\int_{[0,1]^{d_x}} \sum_{\mathbf{u} \in \prod_{j=1}^{d} \mathcal{U}_j} \sum_{\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j} \|\hat{\boldsymbol{\mu}}(\mathbf{w})\| \hat{p}(\mathbf{w}) d\mathbf{x} \leq n^{-1} \sum_{i=1}^{n} \|\hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*)\| < \infty.$$

For the second condition, we note that

$$\begin{aligned}
&\int_0^1 \Bigg\| \bigoplus_{\mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j} \bigoplus_{\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j} \int_{[0,1]^{d_x - 1}} \hat{\boldsymbol{\mu}}(\mathbf{w}) \odot \hat{p}(\mathbf{w}) d\mathbf{x}_{-j} \Bigg\|^2 \hat{p}_{x,j}(x_j)^{-1} dx_j \\
&\quad = \int_0^1 \Bigg\| n^{-1} \bigoplus_{i=1}^{n} K_{h_j}(x_j, X_{ij}) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*) \Bigg\|^2 \hat{p}_{x,j}(x_j)^{-1} dx_j \\
&\quad \leq \max_{1 \leq i \leq n} \|\hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*)\|^2 < \infty
\end{aligned}$$

and the same bound applies to the other integrals involved in the second condition.

Moreover, we may get estimators of the marginal densities and regression maps by integrating $\hat{p}$ and $\hat{\boldsymbol{\mu}}$ over appropriate domains as in (2.9) and (2.12). In particular, from the normalization properties (4.2), (4.3) and (4.4) we get

$$\hat{p}_{x,j}(x_j) = n^{-1} \sum_{i=1}^{n} K_{h_j}(x_j, X_{ij}) \tag{4.6}$$

and similarly $\hat{p}_{u,j}(u_j)$ and $\hat{p}_{v,j}(v_j)$ simply by substituting the discrete kernel weights $L_{\lambda_j}(u_j, U_{ij})$ and $W_{s_j}(v_j, V_{ij})$, respectively, for $K_{h_j}(x_j, X_{ij})$. We also obtain the two dimensional density estimator

$$\hat{p}_{xx,jk}(x_j, x_k) = n^{-1} \sum_{i=1}^{n} K_{h_j}(x_j, X_{ij}) K_{h_k}(x_k, X_{ik}) \tag{4.7}$$

when $d_x \geq 2$ by integrating $\hat{p}$ over $(x_l : l \neq j, k) \in [0,1]^{d_x - 2}$, and likewise $\hat{p}_{uu,jk}$, $\hat{p}_{vv,jk}$, $\hat{p}_{xu,jk}$, $\hat{p}_{xv,jk}$ and $\hat{p}_{uv,jk}$. Similarly, using the normalization properties (4.2), (4.3) and (4.4) again, we get the following estimators of the marginal regression maps.

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_{x,j}(x_j) &= (n \cdot \hat{p}_{x,j}(x_j))^{-1} \odot \bigoplus_{i=1}^{n} K_{h_j}(x_j, X_{ij}) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*), \\
\hat{\boldsymbol{\mu}}_{u,j}(u_j) &= (n \cdot \hat{p}_{u,j}(u_j))^{-1} \odot \bigoplus_{i=1}^{n} L_{\lambda_j}(u_j, U_{ij}) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*), \\
\hat{\boldsymbol{\mu}}_{v,j}(v_j) &= (n \cdot \hat{p}_{v,j}(v_j))^{-1} \odot \bigoplus_{i=1}^{n} W_{s_j}(v_j, V_{ij}) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*).
\end{aligned}$$

4.2. Existence and algorithm convergence

Here, we apply the general results in Theorems 1–3 and Corollary 1 in Section 3 to the model (1.1). In the gB-SBF system of equations (2.15), we consider the cases $d = d_x + d_u + d_v$ and

$$\hat{\boldsymbol{\mu}}_j = \hat{\boldsymbol{\mu}}_{x,j},\ 1 \le j \le d_x; \quad \hat{\boldsymbol{\mu}}_{d_x+j} = \hat{\boldsymbol{\mu}}_{u,j},\ 1 \le j \le d_u;$$
$$\hat{\boldsymbol{\mu}}_{d_x+d_u+j} = \hat{\boldsymbol{\mu}}_{v,j},\ 1 \le j \le d_v.$$

Likewise, we enumerate the collection of $\hat{p}_{x,j}$, $\hat{p}_{u,j}$ and $\hat{p}_{v,j}$ into $\hat{p}_1, \dots, \hat{p}_{d_x+d_u+d_v}$ and that of the two-dimensional density estimators $\hat{p}_{xx,jk}, \dots, \hat{p}_{uv,jk}$ as well in an obvious manner. We let $(\hat{\mathbf{m}}_{x,1}, \dots, \hat{\mathbf{m}}_{x,d_x}; \hat{\mathbf{m}}_{u,1}, \dots, \hat{\mathbf{m}}_{u,d_u}; \hat{\mathbf{m}}_{v,1}, \dots, \hat{\mathbf{m}}_{v,d_v})$ be the solution of the resulting gB-SBF system of equations with $\hat{\mathbf{m}}_0 = n^{-1} \odot \bigoplus_{i=1}^n \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*)$. We let $\hat{\mathbf{m}}_{x,j}^{[r]}$, $\hat{\mathbf{m}}_{u,j}^{[r]}$ and $\hat{\mathbf{m}}_{v,j}^{[r]}$ be the rth updates in the resulting gB-SBF algorithm corresponding to $\hat{\mathbf{m}}_{x,j}$, $\hat{\mathbf{m}}_{u,j}$ and $\hat{\mathbf{m}}_{v,j}$, respectively. For more concrete description of the resulting gB-SBF system of equations and algorithm, we refer to the Appendix A.2. We also write

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_+(\mathbf{w}) &= \hat{\mathbf{m}}_0 \oplus \bigoplus_{j=1}^{d_x} \hat{\mathbf{m}}_{x,j}(x_j) \oplus \bigoplus_{j=1}^{d_u} \hat{\mathbf{m}}_{u,j}(u_j) \oplus \bigoplus_{j=1}^{d_v} \hat{\mathbf{m}}_{v,j}(v_j), \\
\hat{\boldsymbol{\mu}}_+^{[r]}(\mathbf{w}) &= \hat{\mathbf{m}}_0 \oplus \bigoplus_{j=1}^{d_x} \hat{\mathbf{m}}_{x,j}^{[r]}(x_j) \oplus \bigoplus_{j=1}^{d_u} \hat{\mathbf{m}}_{u,j}^{[r]}(u_j) \oplus \bigoplus_{j=1}^{d_v} \hat{\mathbf{m}}_{v,j}^{[r]}(v_j)
\end{aligned} \tag{4.8}$$

as in Section 2. As we verified in the previous subsection, the full-dimensional kernel estimators $\hat{p}$ and $\hat{\boldsymbol{\mu}}$ satisfy the condition (S1). Below, we give a set of sufficient conditions on the smoothing parameters, the baseline kernel K and a dataset under which the non-asymptotic condition (S2), tailored for the case of mixed predictors, are valid. The sufficient conditions are actually minimal in the sense that they are required even for one-dimensional regression smoothing across $[0,1]$, $\mathcal{U}_j$ and $\mathcal{V}_j$ to be well-posed.

Condition (S*).

(S1*) $a := \max_{1 \le j \le d_x} \max\left\{X_{(1),j}, 1 - X_{(n),j}, \max_{1 \le i \le n-1}(X_{(i+1),j} - X_{(i),j})/2\right\} / h_j < 1$, *where* $(X_{(i),j} : 1 \le i \le n)$ *is the order statistics of* $(X_{ij} : 1 \le i \le n)$.

(S2*) K *is bounded and* $\inf_{t \in [-a,a]} K(t) > 0$, *where* a *is the constant in (S1*).*

(S3*) *For each* $1 \le j \le d_u$ *and* $u_j \in \mathcal{U}_j$, *there exists an observation* U_{ij} *such that* $U_{ij} = u_j$ *and* $\lambda_j < 1$.

(S4*) *For each* $1 \le j \le d_v$ *and* $v_j \in \mathcal{V}_j$, *there exists an observation* V_{ij} *such that* $V_{ij} = v_j$ *and* $\sum_{v_j' \in \mathcal{V}_j, v_j' \ne v_j} s_j^{\delta_j(v_j', v_j)} < 1$.

We note that the conditions (S1*) and (S2*) imply that $\inf_{x_j \in [0,1]} \hat{p}_{x,j}(x_j) > 0$ for all $1 \le j \le d_x$. The conditions (S3*) and (S4*) imply, respectively, that $\hat{p}_{u,j}(u_j) > 0$ for all $u_j \in \mathcal{U}_j$ and $1 \le j \le d_u$ and that $\hat{p}_{v,j}(v_j) > 0$ for all $v_j \in \mathcal{V}_j$

and $1 \leq j \leq d_v$. With these observations, one can show that the condition (S*) implies the condition (S2). Then, the following non-asymptotic result is immediate.

Corollary 2. *Assume the condition (S*). Then, the corresponding versions of Theorems 1 and 2 hold for* $\hat{\mathbf{m}}_{x,j}$, $\hat{\mathbf{m}}_{u,j}$, $\hat{\mathbf{m}}_{v,j}$, $\hat{\boldsymbol{\mu}}_+$ *and* $\hat{\boldsymbol{\mu}}_+^{[r]}$.

Now, we present the corresponding versions of Theorem 3 and Corollary 1. Recall that we do not impose the assumption of i.i.d. data for Theorem 3 and Corollary 1, but state the results with the higher-level condition (A). Here, we focus on the case where we have n i.i.d. observations $\{(\mathbf{W}_i, \mathbf{Y}_i^*) : 1 \leq i \leq n\}$, to present a set of sufficient conditions that imply the condition (A). Finding sufficient conditions for non-i.i.d. data is more challenging, particulary in Hilbert spaces, but may be solved using the techniques in [6], for example. The sufficient conditions under the i.i.d. assumption are given below.

Condition (B).

(B1) $\mathrm{E}(\|\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)\|^\alpha) < \infty$ *for some* $\alpha > 2$, *and* $\mathrm{E}(\|\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)\|^2 | X_j = \cdot)$ *are bounded on* $[0,1]$ *for all* $1 \leq j \leq d_x$.

(B2) *The joint density* p *is bounded away from zero and infinity on* $\mathcal{W}$. *For all* j, k, u_k *and* v_k, $p_{xu,jk}(\cdot, u_k)$ *and* $p_{xv,jk}(\cdot, v_k)$ *are continuous on* $[0,1]$. *When* $d_x \geq 2$, $p_{xx,jk}$ *are continuous on* $[0,1]^2$.

(B3) K *is Lipschitz continuous and* $\int_{-1}^0 K(t)dt \wedge \int_0^1 K(t)dt > 0$.

(B4) *For all* j, *it holds that* $h_j, \lambda_j, s_j = o(1)$ *and* $\inf_n n^{c_j} h_j > 0$ *for some* $c_j < (\alpha - 2)/\alpha$, *where* α *is the constant in (B1). Also,* $\log n/(nh_1) = o(1)$ *when* $d_x = 1$, *and* $\log n/(nh_j h_k) = o(1)$ *when* $d_x \geq 2$.

(B5) $P\left(\max_{1 \leq i \leq n} \|\hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*) \ominus \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)\| < M\right) \to 1$ *for some constant* $M > 0$.

We note that the conditions (B2)-(B4) are standard in the kernel smoothing theory. The condition on h_j in (B4) allows the optimal bandwidth rate $h_j \asymp n^{-1/5}$ if α in (B1) is larger than 5/2. When $\mathbf{Y}$ is completely observed, we take $\hat{\boldsymbol{\psi}}(\mathbf{W}, \mathbf{Y}^*) = \boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*) = \mathbf{Y}$, in which case (B1) reduces to a standard condition in the kernel smoothing theory, and (B5) is automatically satisfied. Below after the statement of a corollary, we give two other examples where the conditions (B1) and (B5) are satisfied.

Corollary 3. *Assume the condition (B) and the version of the last condition at (3.5) corresponding to the mixed predictor case. Then, the corresponding versions of Theorem 3 and Corollary 1 hold for* $\hat{\mathbf{m}}_{x,j}$, $\hat{\mathbf{m}}_{u,j}$, $\hat{\mathbf{m}}_{v,j}$ *and* $\hat{\mathbf{m}}_{x,j}^{[r]}$, $\hat{\mathbf{m}}_{u,j}^{[r]}$, $\hat{\mathbf{m}}_{v,j}^{[r]}$.

Example 1. (Missing data). *Suppose that* $\mathbf{Y}$ *is subject to missing. Let* R *be the indicator defined by* $R = I(\mathbf{Y}$ *is not missing). In this case, we observe* $\mathbf{Y}^* = \mathbf{Y}$ *if* $R = 1$ *and* $\mathbf{Y}^* = \mathbf{0}$ *otherwise. Suppose that the Hilbertian MAR condition* $R \perp \mathbf{Y} | \mathbf{W}$ *holds. For the unbiased transformation* $\boldsymbol{\psi}$, *we take the inverse probability weighting map defined by* $\boldsymbol{\psi}(\mathbf{w}, \mathbf{h}) = (1/\pi(\mathbf{w})) \odot \mathbf{h}$, *where* $\pi(\mathbf{w}) = P(R =$

$1|\mathbf{W}=\mathbf{w})$. *Then,* $\boldsymbol{\psi}$ *satisfies (B1), provided that* $\mathrm{E}(\|\mathbf{Y}\|^{\alpha})<\infty$ *for some* $\alpha>2$, $\mathrm{E}(\|\mathbf{Y}\|^{2}|X_j=\cdot)$ *are bounded on* $[0,1]$ *for all* $1\le j\le d_x$, *and* $\inf_{\mathbf{w}}\pi(\mathbf{w})>0$. *For the validity of (B5) we note that* $\max_{1\le i\le n}\|\hat{\boldsymbol{\psi}}(\mathbf{W}_i,\mathbf{Y}_i^*)\ominus\boldsymbol{\psi}(\mathbf{W}_i,\mathbf{Y}_i^*)\|\le \max_{1\le i\le n}(|1/\hat{\pi}(\mathbf{W}_i)-1/\pi(\mathbf{W}_i)|\cdot\|\mathbf{Y}_i\|)$, *where* $\hat{\boldsymbol{\psi}}(\mathbf{w},\mathbf{h})=(1/\hat{\pi}(\mathbf{w}))\odot\mathbf{h}$ *and* $\hat{\pi}$ *is an estimator of* π. *For the estimation of* π, *suppose that the predictors in* $\mathbf{V}$ *are real-valued and one applies logistic linear regression. Let* β_j *denote the regression coefficients in the logistic linear regression and* $\hat{\beta}_j$ *be their estimators. Then, under either of the assumptions: (i)* $\mathbf{Y}$ *is a bounded random element and* $|\hat{\beta}_j-\beta_j|=o_p(1)$ *for all* j; *(ii)* $\mathrm{E}(\|\mathbf{Y}\|^{\alpha})<\infty$ *for some* $\alpha>2$ *and* $|\hat{\beta}_j-\beta_j|=O_p(n^{-1/2})$ *for all* j, *we get* $\max_{1\le i\le n}(|1/\hat{\pi}(\mathbf{W}_i)-1/\pi(\mathbf{W}_i)|\cdot\|\mathbf{Y}_i\|)=o_p(1)$, *which gives (B5). We note that both assumptions on* $\hat{\beta}_j$ *in (i) and (ii) are standard results in logistic linear regression.*

Example 2. (Censored data). *Let* $\mathbf{Y}\equiv Y\in(0,\tau)$ *for* $\tau<\infty$ *be a survival time subject to random censoring. Let* $C>0$ *be the random censoring time with distribution function* G *satisfying* $G(\tau)<1$, $Y\perp C$ *and* $P(Y\le C|\mathbf{W},Y)=P(Y\le C|Y)$. *These conditions on* C *are commonly adopted in the literature on censored regression. Let* $T=Y\wedge C$ *denote the observed time and* $\Delta=I(Y\le C)$ *denote the censoring indicator. In this case, we observe the random vector* $\mathbf{Y}^*\equiv Y^*=(T,\Delta)$ *instead of* Y *and* C. *Let* $\boldsymbol{\psi}\equiv\psi$ *be the unbiased transformation* $\psi(\mathbf{W},Y^*)=\Delta\cdot T/(1-G(T))$ *proposed by [16]. In this case,* ψ *does not depend on* $\mathbf{W}$ *and* ψ *clearly satisfies (B1). Also, (B5) holds for* $\hat{\psi}(\mathbf{W},Y^*)=\Delta\cdot T/(1-\hat{G}(T))$, *where* $\hat{G}$ *is the Kaplan-Meier estimator of* G. *To see the latter, we note that*

$$\max_{1\le i\le n}|\hat{\psi}(\mathbf{W}_i,Y_i^*)-\psi(\mathbf{W}_i,Y_i^*)|\le\frac{\tau}{(1-G(\tau))(1-\hat{G}(\tau))}\cdot\sup_{t\le\tau}|\hat{G}(t)-G(t)|.$$

The standard theory in survival analysis [41] gives $\sup_{t\le\tau}|\hat{G}(t)-G(t)|=o_p(1)$, *from which (B5) follows.*

4.3. Rates of convergence

In this subsection, we demonstrate that the gB-SBF estimator does not have the dimensionality problem by showing that it achieves the optimal univariate error rate. Let $\boldsymbol{\epsilon}_+=\boldsymbol{\psi}(\mathbf{W},\mathbf{Y}^*)\ominus\mathbf{Y}\oplus\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is the error term at (1.1). Here and in Section 4.4, we assume that $\{(\mathbf{W}_i,\mathbf{Y}_i^*):1\le i\le n\}$ are i.i.d. To obtain the rates of convergence, we make use of the following assumptions.

Condition (C).

(C1) (i) $\mathrm{E}(\|\boldsymbol{\epsilon}_+\|^{\alpha})<\infty$ *for some* $\alpha>5/2$ *and* (ii) $\mathrm{E}(\|\boldsymbol{\epsilon}_+\|^{2}|X_j=\cdot)$ *are bounded on* $[0,1]$ *for all* $1\le j\le d_x$.

(C2) *The component maps* $\mathbf{m}_{x,j}$ *for* $1\le j\le d_x$ *are twice continuously Fréchet differentiable on* $[0,1]$.

(C3) *The condition on* p *in (B2) holds. In addition, for all* j,k,u_k *and* v_k, $p_{xu,jk}(\cdot,u_k)$ *and* $p_{xv,jk}(\cdot,v_k)$ *are* C^1 *on* $[0,1]$. *When* $d_x\ge 2$, $p_{xx,jk}$ *are* C^1 *on* $[0,1]^2$.

(C4) *The condition (B3) holds. In addition,* $\int_{-1}^{1} tK(t)dt = 0$.
(C5) *For all* j*, it holds that* $h_j \asymp n^{-1/5}$ *and* $\lambda_j, s_j = o(1)$.

When $\mathbf{Y}$ is completely observed so that $\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*) = \mathbf{Y}$, the condition (C1) reduces to the one with $\boldsymbol{\epsilon}_+$ being replaced by $\boldsymbol{\epsilon}$. The latter is a standard condition in the kernel smoothing theory. Even when $\mathbf{Y}$ is incompletely observed as in Examples 1 and 2, (C1) is easily satisfied in the examples. (C2) is a natural extension of the usual condition for real-valued component maps ($\mathbb{H} = \mathbb{R}$) to the current Hilbertian case. (C3)–(C5) are standard conditions in the kernel smoothing theory. Define $s_* = \max\{s_j^{\delta_j^*} : 1 \le j \le d_v\}$ where $\delta_j^* = \min\{\delta_j(v_j, v_j') : v_j, v_j' \in \mathcal{V}_j, v_j \neq v_j'\}$, the minimum nonzero distance in $\mathcal{V}_j$, and $\lambda_* = \max\{\lambda_j : 1 \le j \le d_u\}$. Let $\max_{1\le i\le n} \|\hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*) \ominus \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)\| = O_p(a_n)$ hold for some sequence a_n. Let $I_j = [2h_j, 1 - 2h_j]$. We note that $K_{h_j}(x, x') = K_{h_j}(x - x')$ for $(x, x') \in I_j \times [0, 1]$ and that $\int_0^1 K_{h_j}(x, x')dx' = \int_{-1}^{1} K(t)dt = 1$ for $x \in I_j$.

Theorem 4. *Assume the condition (C). Then, the followings hold for all* j.

(i) (Pointwise convergence)

$$\|\hat{\mathbf{m}}_{x,j}(x_j) \ominus \mathbf{m}_{x,j}(x_j)\| = O_p(n^{-2/5} + \lambda_* + s_* + a_n) \quad \text{for } x_j \in I_j,$$
$$\|\hat{\mathbf{m}}_{x,j}(x_j) \ominus \mathbf{m}_{x,j}(x_j)\| = O_p(n^{-1/5} + \lambda_* + s_* + a_n) \quad \text{for } x_j \in [0,1] \setminus I_j.$$

(ii) (L_2 convergence)

$$\int_{I_j} \|\hat{\mathbf{m}}_{x,j}(x_j) \ominus \mathbf{m}_{x,j}(x_j)\|^2 p_{x,j}(x_j)dx_j = O_p(n^{-4/5} + \lambda_*^2 + s_*^2 + a_n^2),$$
$$\int_{0}^{1} \|\hat{\mathbf{m}}_{x,j}(x_j) \ominus \mathbf{m}_{x,j}(x_j)\|^2 p_{x,j}(x_j)dx_j = O_p(n^{-3/5} + \lambda_*^2 + s_*^2 + a_n^2).$$

(iii) (Uniform convergence)

$$\sup_{x_j \in I_j} \|\hat{\mathbf{m}}_{x,j}(x_j) \ominus \mathbf{m}_{x,j}(x_j)\| = O_p(n^{-2/5}\sqrt{\log n} + \lambda_* + s_* + a_n),$$
$$\sup_{x_j \in [0,1]} \|\hat{\mathbf{m}}_{x,j}(x_j) \ominus \mathbf{m}_{x,j}(x_j)\| = O_p(n^{-1/5} + \lambda_* + s_* + a_n),$$
$$\max_{u_j} \|\hat{\mathbf{m}}_{u,j}(u_j) \ominus \mathbf{m}_{u,j}(u_j)\| = O_p(n^{-2/5} + \lambda_* + s_* + a_n),$$
$$\max_{v_j} \|\hat{\mathbf{m}}_{v,j}(v_j) \ominus \mathbf{m}_{v,j}(v_j)\| = O_p(n^{-2/5} + \lambda_* + s_* + a_n).$$

Remark 1. *We give some remarks on the magnitude of* λ_*, s_* *and* a_n. *One can show that the Nadaraya-Watson-type full-dimensional estimator based on the observations of* $\mathbf{U}_i, \mathbf{V}_i$ *and completely observed* $\mathbf{Y}_i$, *defined by*

$$\tilde{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}(\mathbf{u}, \mathbf{v}) = \left(\sum_{i=1}^{n} \prod_{j=1}^{d_u} L_{\lambda_j}(u_j, U_{ij}) \cdot \prod_{k=1}^{d_v} W_{s_k}(v_k, V_{ik})\right)^{-1} \odot \bigoplus_{i=1}^{n} \left(\prod_{j=1}^{d_u} L_{\lambda_j}(u_j, U_{ij}) \cdot \prod_{k=1}^{d_v} W_{s_k}(v_k, V_{ik})\right) \odot \mathbf{Y}_i, \tag{4.9}$$

achieves the optimal error rate when $\lambda_* = O(n^{-1/2})$ *and* $s_* = O(n^{-1/2})$. *Hence, it makes sense to assume that* $\lambda_* = O(n^{-C_\lambda})$ *for some* $C_\lambda \geq 2/5$ *and* $s_* = O(n^{-C_s})$ *for some* $C_s \geq 2/5$. *When* $\mathbf{Y}$ *is completely observed, the term* a_n *does not appear in the rates in Theroem 4. In the missing data setting (Example 1), we get* $a_n = n^{-c}$ *for some* $c > 2/5$ *when* $\mathrm{E}(\|\mathbf{Y}\|^\alpha) < \infty$ *for some* $\alpha > 10$ *and* $|\hat{\beta}_j - \beta_j| = O_p(n^{-1/2})$ *for all* j. *In the censored data setting (Example 2), we have* $a_n = n^{-1/2}\sqrt{\log n}$ *when* G *is continuous, since the Kaplan-Meier estimator* $\hat{G}$ *satisfies* $\sup_{t \leq \tau} |\hat{G}(t) - G(t)| = O_p(n^{-1/2}\sqrt{\log n})$, *as proved by [24]. Therefore,* $a_n = o(n^{-2/5})$ *under these mild conditions.*

Theorem 4 together with Remark 1 demonstrates that the gB-SBF estimator may achieve the optimal univariate rates of convergence even though there are multiple predictors. This is an important property in nonparametric inference, which is not shared by estimators based on full-dimensional approaches. The result is particularly notable since it shows the dimension-free convergence rates in the general data setting.

4.4. Asymptotic distribution

In this subsection, we present the asymptotic joint distribution of the gB-SBF estimator. For this, we make further assumptions. Let $\{\mathbf{e}_l : 1 \leq l \leq L\}$ denote an orthonormal basis of $\mathbb{H}$. Our theory covers both $L < \infty$ and $L = \infty$.

Condition (D).

(D1) *For the constant* α *in (C1)-(i) and for all* l, l', u_k, v_k *and* $1 \leq j \leq d_x$, *the functions* $\mathrm{E}(\|\boldsymbol{\epsilon}_+\|^\alpha | X_j = \cdot)$, $\mathrm{E}(\langle \boldsymbol{\epsilon}_+, \mathbf{e}_l \rangle \langle \boldsymbol{\epsilon}_+, \mathbf{e}_{l'} \rangle | X_j = \cdot, U_k = u_k)$, $\mathrm{E}(\langle \boldsymbol{\epsilon}_+, \mathbf{e}_l \rangle \langle \boldsymbol{\epsilon}_+, \mathbf{e}_{l'} \rangle | X_j = \cdot, V_k = v_k)$ *and* $\mathrm{E}(\langle \boldsymbol{\epsilon}_+, \mathbf{e}_l \rangle \langle \boldsymbol{\epsilon}_+, \mathbf{e}_{l'} \rangle | X_j = \cdot, X_k = \cdot)$ *are bounded on their respective domains, and* $\mathrm{E}(\langle \boldsymbol{\epsilon}_+, \mathbf{e}_l \rangle \langle \boldsymbol{\epsilon}_+, \mathbf{e}_{l'} \rangle | X_j = \cdot)$ *are continuous on* $[0, 1]$.

(D2) *For all* $1 \leq j \leq d_x$, $\partial p(\mathbf{w})/\partial x_j$ *exist and are bounded on* $\mathcal{W}$.

(D3) *For all* j, $n^{1/5} h_j \to \alpha_j$, $n^{2/5} \lambda_j \to \beta_j$ *and* $n^{2/5} s_j^{\delta_j^*} \to \gamma_j$ *for some constants* $\alpha_j > 0$, $\beta_j \geq 0$ *and* $\gamma_j \geq 0$. *Also,* $\max_{1 \leq i \leq n} \|\hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*) \ominus \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)\| = o_p(n^{-2/5})$.

The conditions on λ_j and s_j in (D3) are satisfied with the sizes of the smoothing parameters discussed in Remark 1. The condition on $\hat{\boldsymbol{\psi}}$ is also valid under the mild conditions given there. The remaining ones in (D) are weak regularity conditions. 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, where $\mathcal{L}(\mathbb{B}_1, \mathbb{B}_2)$ for two Banach spaces $\mathbb{B}_1$ and $\mathbb{B}_2$ denotes the space of bounded linear operators that map $\mathbb{B}_1$ to $\mathbb{B}_2$. The first derivative $D\mathbf{f}(x) : \mathbb{R} \to \mathbb{H}$ at $x \in [0, 1]$ in our setting is defined by $D\mathbf{f}(x)(s) = s \odot D\mathbf{f}(x)(1)$, where $D\mathbf{f}(x)(1)$ satisfies

$$\lim_{\varepsilon \to 0} |\varepsilon|^{-1} \cdot \big\| \mathbf{f}(x + \varepsilon) \ominus \mathbf{f}(x) \ominus \big(\varepsilon \odot D\mathbf{f}(x)(1) \big) \big\| = 0.$$

Let $D^2\mathbf{f} : [0,1] \to \mathcal{L}\big(\mathbb{R}, \mathcal{L}(\mathbb{R}, \mathbb{H})\big)$ denote the second Fréchet derivative of $\mathbf{f}$. The second derivative $D^2\mathbf{f}(x)$ at $x \in [0,1]$, now as a map from $\mathbb{R}$ to $\mathcal{L}(\mathbb{R}, \mathbb{H})$ or as a map from $\mathbb{R}^2$ to $\mathbb{H}$, is defined by $D^2\mathbf{f}(x)(s,t) = t \odot D^2\mathbf{f}(x)(s,1)$, where $D^2\mathbf{f}(x)(s,1)$ satisfies

$$\lim_{\varepsilon \to 0} |\varepsilon|^{-1} \cdot \big\| D\mathbf{f}(x+\varepsilon)(s) \ominus D\mathbf{f}(x)(s) \ominus \big(\varepsilon \odot D^2\mathbf{f}(x)(s,1)\big) \big\| = 0.$$

Define $\mathbf{c}_j(x_j) = \frac{1}{2}\int_{-1}^{1} t^2 K(t) dt \odot D^2\mathbf{m}_{x,j}(x_j)(1,1)$ and $\mathbf{\Theta}_j(x_j) = \alpha_j^2 \odot \mathbf{c}_j(x_j) \oplus \mathbf{\Delta}_{x,j}(x_j)$, where α_j are the constants in (D3) and $\mathbf{\Delta}_{x,j}$ together with $\mathbf{\Delta}_{u,j}$ and $\mathbf{\Delta}_{v,j}$ are defined in the Appendix A.7.1. They constitute the asymptotic bias of the joint distribution of the estimated component maps, as is demonstrated in Theorem 5 below. In fact, re-enumerating

$$(\mathbf{\Delta}_{x,1}, \ldots, \mathbf{\Delta}_{x,d_x}; \mathbf{\Delta}_{u,1}, \ldots, \mathbf{\Delta}_{u,d_u}; \mathbf{\Delta}_{v,1}, \ldots, \mathbf{\Delta}_{v,d_v})$$

as $(\mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_{d_x+d_u+d_v})$, the $(d_x+d_u+d_v)$-tuple is nothing else than the solution of a system of equations

$$\mathbf{\Delta}_j(z_j) = \tilde{\mathbf{\Delta}}_j(z_j) \ominus \bigoplus_{k \neq j}^{d_x+d_u+d_v} \int_{\mathcal{Z}_k} \mathbf{\Delta}_k(z_k) \odot \frac{p_{jk}(z_j, z_k)}{p_j(z_j)} d\nu_k(z_k), \quad 1 \leq j \leq d_x + d_u + d_v.$$

Note that the similarity between the above system of equations and the one at (2.7). Here,

$$(\tilde{\mathbf{\Delta}}_{x,1}, \ldots, \tilde{\mathbf{\Delta}}_{x,d_x}; \tilde{\mathbf{\Delta}}_{u,1}, \ldots, \tilde{\mathbf{\Delta}}_{u,d_u}; \tilde{\mathbf{\Delta}}_{v,1}, \ldots, \tilde{\mathbf{\Delta}}_{v,d_v})$$

with $\tilde{\mathbf{\Delta}}_{x,j}$, $\tilde{\mathbf{\Delta}}_{u,j}$ and $\tilde{\mathbf{\Delta}}_{v,j}$ being defined in the Appendix A.7.1 is re-enumerated as $(\tilde{\mathbf{\Delta}}_1, \ldots, \tilde{\mathbf{\Delta}}_{d_x+d_u+d_v})$. Also, $\mathcal{Z}_j$ are $[0,1]$, $\mathcal{U}_j$ or $\mathcal{V}_j$ depending on the position of j in the re-enumeration, p_j and p_{jk} are the marginal and 2-dimensional joint densities of the corresponding predictors in the re-enumeration and ν_j are the associated Lebesgue or counting measures. The terms $\mathbf{\Theta}_j$ arise from an expansion of the kernel weighted averages of $\mathbf{m}_{x,j}(X_{ij}) \ominus \mathbf{m}_{x,j}(x_j)$, $\mathbf{m}_{u,j}(U_{ij}) \ominus \mathbf{m}_{u,j}(u_j)$ and $\mathbf{m}_{v,j}(V_{ij}) \ominus \mathbf{m}_{v,j}(v_j)$, see the Appendix A.7.4 for details.

The asymptotic variance comes from the stochastic part of the marginal regression estimators $\hat{\boldsymbol{\mu}}_{x,j}$. For this, let $\boldsymbol{\epsilon}_+ \otimes \boldsymbol{\epsilon}_+ : \mathbb{H} \to \mathbb{H}$ be the operator defined by $(\boldsymbol{\epsilon}_+ \otimes \boldsymbol{\epsilon}_+)(\mathbf{h}) = \langle \boldsymbol{\epsilon}_+, \mathbf{h} \rangle \odot \boldsymbol{\epsilon}_+$ and let $C_{j,x_j} : \mathbb{H} \to \mathbb{H}$ be the covariance operator defined by

$$C_{j,x_j}(\mathbf{h}) = \alpha_j^{-1} p_{x,j}(x_j)^{-1} \int_{-1}^{1} K^2(t) dt \cdot \mathrm{E}((\boldsymbol{\epsilon}_+ \otimes \boldsymbol{\epsilon}_+)(\mathbf{h}) | X_j = x_j). \tag{4.10}$$

Let $\mathbf{G}(\mathbf{0}, C_{j,x_j})$ denote a $\mathbb{H}$-valued Gaussian random element with mean $\mathbf{0}$ and covariance operator C_{j,x_j}. It is a random element such that $\langle \mathbf{G}(\mathbf{0}, C_{j,x_j}), \mathbf{h} \rangle$ is normally distributed with mean 0 and variance $\langle C_{j,x_j}(\mathbf{h}), \mathbf{h} \rangle$ for all $\mathbf{h} \in \mathbb{H}$. When $\mathbb{H} = \mathbb{R}$, it reduces to a normal random variable.

We let $\hat{\mathbb{L}}_{\mathbf{x},\mathbf{u},\mathbf{v}}$ denote the joint distribution of $(n^{2/5} \odot (\hat{\mathbf{m}}_{x,j}(x_j) \ominus \mathbf{m}_{x,j}(x_j)) : 1 \le j \le d_x)$, $(n^{2/5} \odot (\hat{\mathbf{m}}_{u,j}(u_j) \ominus \mathbf{m}_{u,j}(u_j)) : 1 \le j \le d_u)$ and $(n^{2/5} \odot (\hat{\mathbf{m}}_{v,j}(v_j) \ominus \mathbf{m}_{v,j}(v_j)) : 1 \le j \le d_v)$. Similarly, we write $\mathbb{L}_{\mathbf{x},\mathbf{u},\mathbf{v}}$ for the joint distribution of $(\boldsymbol{\Theta}_j(x_j) \oplus \mathbf{G}(\mathbf{0}, C_{j,x_j}) : 1 \le j \le d_x)$, $(\boldsymbol{\Delta}_{u,j}(u_j) : 1 \le j \le d_u)$ and $(\boldsymbol{\Delta}_{v,j}(v_j) : 1 \le j \le d_v)$.

Theorem 5. *Assume the conditions (C1)-(i),(C2)-(C4) and (D). Then, the following results hold: (i) The joint distribution* $\hat{\mathbb{L}}_{\mathbf{x},\mathbf{u},\mathbf{v}}$ *converges weakly to* $\mathbb{L}_{\mathbf{x},\mathbf{u},\mathbf{v}}$ *for a.e. fixed* $\mathbf{x} \in (0,1)^{d_x}$ *with respect to* $\bigotimes_{j=1}^{d_x} \mathrm{Leb}$ *and for all* $\mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j$ *and* $\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j$*; (ii) For* $\hat{\boldsymbol{\mu}}_+(\mathbf{w})$ *defined at (4.8),*

$$n^{2/5} \odot (\hat{\boldsymbol{\mu}}_+(\mathbf{w}) \ominus \boldsymbol{\mu}(\mathbf{w})) \xrightarrow{d} \bigoplus_{j=1}^{d_x} \boldsymbol{\Theta}_j(x_j) \oplus \bigoplus_{j=1}^{d_u} \boldsymbol{\Delta}_{u,j}(u_j) \oplus \bigoplus_{j=1}^{d_v} \boldsymbol{\Delta}_{v,j}(v_j) \oplus \mathbf{G}\Big(\mathbf{0}, \sum_{j=1}^{d_x} C_{j,x_j}\Big).$$

Let $\hat{\mathbf{m}}_{x,j}^{\mathrm{ora}}$ be the oracle estimator of $\mathbf{m}_{x,j}$ obtained by using the knowledge of all other component maps. Then, the asymptotic distribution for $\hat{\mathbf{m}}_{x,j}^{\mathrm{ora}}$ is given by

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

where $\boldsymbol{\delta}_{x,j} = \big(dp_{x,j}(x_j)/dx_j \cdot p_{x,j}(x_j)^{-1} \cdot \int_{-1}^{1} t^2 K(t) dt\big) \odot D\mathbf{m}_{x,j}(x_j)(1)$. This means that $\hat{\mathbf{m}}_{x,j}$ and $\hat{\mathbf{m}}_{x,j}^{\mathrm{ora}}$ have the same asymptotic covariance operator, but differ in their asymptotic biases, so that the gB-SBF estimator achieves a 'semi-oracle property'. The difference of the asymptotic biases is $(\alpha_j^2 \odot \boldsymbol{\delta}_{x,j}(x_j)) \ominus \boldsymbol{\Delta}_{x,j}(x_j) =: \boldsymbol{\beta}_j(x_j)$ and it holds that $\mathrm{E}(\boldsymbol{\beta}_j(X_j)) = \int_0^1 \boldsymbol{\beta}_j(x_j) \odot p_{x,j}(x_j) dx_j = \mathbf{0}$ by (A.33) in the Appendix.

5. Numerical properties

In this section we report the results of simulation studies and real data applications. Details on the practical implementation of the gB-SBF algorithm including smoothing parameter selection can be found in the Appendix A.3.

5.1. Simulation study

In the first simulation study, we compared our gB-SBF method with the SBF method based on partially linear additive models [SBF-PLAM, 45]. Since the latter model can only deal with completely observed scalar responses, we considered the case $\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*) = Y$, where Y is a scalar response. Partially linear (additive) models are widely used when responses are real-valued and both continuous-type and discrete-type predictors are present. Hence, the comparison we made here is a meaningful check of how our new class of methods works in

comparison with the existing class of methods. In the simulation, we estimated the following additive model:

$$Y = m_{x,1}(X_1) + m_{x,2}(X_2) + m_{u,1}(U_1) + m_{u,2}(U_2) + m_{v,1}(V_1) + m_{v,2}(V_2) + \epsilon,$$

where X_1 and X_2 are independent uniform $[0,1]$ random variables, U_1 and U_2 are *nominal* discrete random variables taking values in $\{1,2\}$ and $\{1,2,3\}$, respectively, V_1 and V_2 are *ordinal* discrete random variables taking values in $\{1,2,3,4\}$ and $\{0, 0.25, 0.75, 1.5, 2.5\}$, respectively, and ϵ is a $N(0, 0.5^2)$ random variable. We took $m_{x,1}(X_1) = \sin(2\pi X_1)$, $m_{x,2}(X_2) = \cos(2\pi X_2)$, $m_{u,1}(U_1) = -3 \cdot I(U_1 = 1) + 3 \cdot I(U_1 = 2)$, $m_{u,2}(U_2) = -5 \cdot I(U_2 = 1) + 0 \cdot I(U_2 = 2) + 5 \cdot I(U_2 = 3)$. For $m_{v,1}$ and $m_{v,2}$, we considered the two cases:

$$m_{v,1}(V_1) = 2V_1, \quad m_{v,2}(V_2) = -2V_2, \qquad \text{(Linear)}$$

$$m_{v,1}(V_1) = 2(V_1 - 2.5)^2, \quad m_{v,2}(V_2) = -\exp(V_2)/2. \qquad \text{(Nonlinear)}$$

We note that the SBF-PLAM technique is designed for the case where $m_{v,1}$ and $m_{v,2}$ are linear, while the gB-SBF method is for general component maps.

For the generation of U_1, U_2, V_1 and V_2, we considered two scenarios. To describe them, let $\mathcal{M}_k(q_1, \ldots, q_k)$ denote a k-variate multinomial distribution with sampling probabilities $q_j \geq 0$ such that $q_1 + \cdots + q_k = 1$. In both of the following scenarios, U_1, U_2, V_1 and V_2 are mutually independent.

(a) $(\mathbf{U}, \mathbf{V})$ depending on $\mathbf{X}$:

$$\begin{aligned}
&U_1|\mathbf{X} \sim \mathcal{M}_2(1 - (X_1^2 + X_2^2)/2,\ (X_1^2 + X_2^2)/2);\\
&U_2|\mathbf{X} \sim \mathcal{M}_3(\sin(X_1\pi/2)/2,\ \cos(X_2\pi/2)/2,\\
&\qquad\qquad 1 - \sin(X_1\pi/2)/2 - \cos(X_2\pi/2)/2);\\
&V_1|\mathbf{X} \sim \mathcal{M}_4(X_1/2,\ 1/2 - X_1/2,\ X_2/2,\ 1/2 - X_2/2);\\
&V_2|\mathbf{X} \sim \mathcal{M}_5(2X_1/5,\ 2/5 - 2X_1/5,\ 2X_2/5,\ 2/5 - 2X_2/5,\ 1/5).
\end{aligned}$$

(b) $(\mathbf{U}, \mathbf{V})$ independent of $\mathbf{X}$: U_1, U_2, V_1 and V_2 are from $\mathcal{M}_k(q_1, \ldots, q_k)$ for $k = 2$, 3, 4 and 5, respectively with $q_1 = \cdots = q_k = 1/k$.

We note that the SBF-PLAM technique gains some efficiency, in comparison with the partially linear approach (without additive modeling for the effect of $\mathbf{X}$), only when $\mathrm{E}(U_j|\mathbf{X}) \neq \mathrm{E}(U_j)$ or $\mathrm{E}(V_j|\mathbf{X}) \neq \mathrm{E}(V_j)$ for some j, which is violated in the scenario (b).

We generated a training sample of size n and a test sample of size $N = 100$ for $M = 500$ times. We computed the gB-SBF estimator based on the gB-SBF algorithm. We compared the gB-SBF and the SBF-PLAM via the mean squared prediction error defined by

$$\mathrm{MSPE} = M^{-1} \sum_{m=1}^{M} N^{-1} \sum_{i=1}^{N} (Y_i^{\mathrm{test}(m)} - \hat{Y}_i^{\mathrm{test}(m)})^2, \qquad (5.1)$$

TABLE 1
The values of the ratio MSPE(SBF-PLAM)/MSPE(gB-SBF).

	Scenario			
	Linear		Nonlinear	
n	(a)	(b)	(a)	(b)
100	1.58	1.65	14.78	16.98
200	1.19	1.21	15.24	17.46
400	1.04	1.05	15.39	17.86

where $Y_i^{\text{test}(m)}$ is the ith response in the mth test sample and $\hat{Y}_i^{\text{test}(m)}$ is the prediction of $Y_i^{\text{test}(m)}$ based on the mth training sample. Table 1 gives the MSPE ratios of the gB-SBF relative to those of the SBF-PLAM.

The table indicates that the prediction based on the gB-SBF estimator performs better than the one based on the SBF-PLAM estimator, even when $m_{v,1}$ and $m_{v,2}$ are linear. Our interpretation for this is that the SBF-PLAM procedure, after estimating the parametric and nonparametric parts based on a profiling method, does not update the estimators further, which might have degraded its performance. We think that the inferior performance might be also the case with other methods based on one-step update. On the contrary, our gB-SBF method operates an iterative algorithm, as described in (A.3) in the Supplement, until convergence. In the nonlinear case, there is a large gap in MSPE between the SBF-PLAM and the gB-SBF methods and the gap grows further as n increases. The results suggest that the gB-SBF procedure is a powerful option.

In the second simulation study, we considered a functional response that is observed at discrete time points with noise. We generated discrete points T_{ik} uniformly on $[0,1]$ for $1 \le k \le N_i$ and $1 \le i \le n$, where N_i are uniform random integers between 25 and 60. We took $n = 100, 200$ and 400. We considered the case where $d_x = d_u = d_v = 1$ and generated X, U and V uniformly from $[0,1]$, $\{1,2\}$ and $\{-1/2, 0, 1\}$, respectively. We set $\mathbf{m}_s(s) \equiv m_s(s)(\cdot)$ for $s = x, u, v$ by

$$\begin{aligned} m_x(x)(t) &= \log(x + t + 1) - f_1(t), \\ m_u(u)(t) &= \sin(2\pi t) \cdot I(u = 1) + \cos(2\pi t) \cdot I(u = 2) - f_2(t), \\ m_v(v)(t) &= \exp(vt) - f_3(t) \end{aligned}$$

for $t \in [0,1]$, where f_1, f_2 and f_3 are some functions that make the component maps satisfy the constraints (2.5). We then generated $Y_i(T_{ik})$ according to the additive model,

$$Y_i(T_{ik}) = m_x(X_{i1})(T_{ik}) + m_u(U_{i1})(T_{ik}) + m_v(V_{i1})(T_{ik}) + (T_{ik} - 0.5)\epsilon_i + \delta_i,$$

where ϵ_i are i.i.d. standard normal random variables and δ_i are i.i.d. random noises from $N(0, 0.1^2)$. We obtained $\mathbf{Y}_i = Y_i(\cdot)$ by smoothing the observations $\{Y_i(T_{ik}) : 1 \le k \le N_i\}$ for each $1 \le i \le n$. At this pre-smoothing stage, we used the kernel-weighting approach with the standard Gaussian kernel and bandwidths chosen by the leave-one-out cross-validation. Based on this generation

TABLE 2
The values of IMSE, ISB and IV, multiplied by 10^2.

	$m_x(\cdot)$			$m_u(\cdot)$			$m_v(\cdot)$		
n	IMSE	ISB	IV	IMSE	ISB	IV	IMSE	ISB	IV
100	0.37	0.04	0.33	0.41	0.07	0.34	0.39	0.02	0.37
200	0.19	0.03	0.16	0.22	0.05	0.17	0.20	0.01	0.19
400	0.10	0.01	0.09	0.12	0.05	0.07	0.10	0.00	0.10

process, we obtained $R = 100$ pseudo samples $\{(Y_i^{(r)}(\cdot), X_i^{(r)}, U_i^{(r)}, V_i^{(r)}) : 1 \leq i \leq n\}$, $1 \leq r \leq 100$.

The measures of performance we chose are integrated mean squared error (IMSE), integrated squared bias (ISB) and integrated variance (IV), defined by

$$\begin{aligned}
\text{IMSE} &= R^{-1}\sum_{r=1}^{R}\int\int_0^1 \left(m_s(s)(t) - \hat{m}_s^{(r)}(s)(t)\right)^2 dt\, p_s(s)d\nu_s(s) = \text{ISB} + \text{IV},\\
\text{ISB} &= \int\int_0^1 \left(m_s(s)(t) - R^{-1}\sum_{r=1}^{R}\hat{m}_s^{(r)}(s)(t)\right)^2 dt\, p_s(s)d\nu_s(s),\\
\text{IV} &= R^{-1}\sum_{r=1}^{R}\int\int_0^1 \left(R^{-1}\sum_{r=1}^{R}\hat{m}_s^{(r)}(s)(t) - \hat{m}_s^{(r)}(s)(t)\right)^2 dt\, p_s(s)d\nu_s(s)
\end{aligned}$$

for $s = x, u, v$, where $\hat{m}_s^{(r)}$ is the estimate of m_s based on the rth pseudo sample and ν_s is either the Lebesgue measure Leb or the counting measure depending on s. Table 2 shows the result of the estimation performance. The table demonstrates that the values of IMSE, ISB and IV are decreasing as the sample size increases. This indicates that our method after a pre-smoothing procedure works quite well even in the case of discretely and nosily observed functional responses.

5.2. Real data analysis

In this subsection we analyse four datasets. They are the cases of density-valued response, compositional response, missing scalar response and randomly right-censored scalar response.

5.2.1. Density-valued response

Dose-response data contains several dose of drugs and their effects on a response variable. Dose-response data analysis is important in finding an appropriate dose. In this analysis, we analyzed cytotoxicity experiments data on the pediatric cancer Ewing sarcoma obtained from the R package 'braidrm' by combining 'es1data', 'es8data' and 'ew8data' there. The data contains several toxicity levels of several drug types. For each drug type (U_1) and log(toxicity level) (X_1), multiple (ranging from 32 to 112) log-transformed CellTiter-Glo intensity of

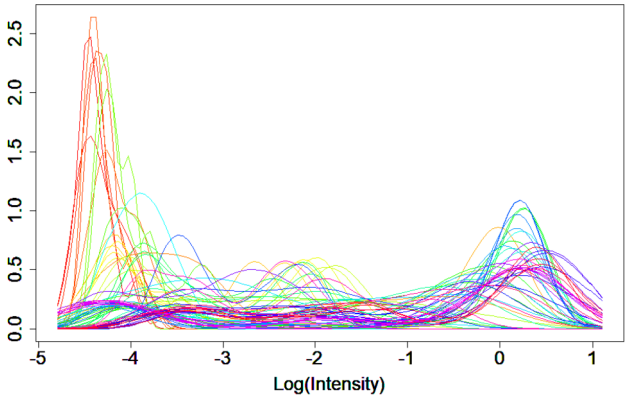

FIG 1. *Plot of densities* $Y_i(\cdot)$ *for* $1 \le i \le 72$.

the tumor are obtained from Ewing sarcoma tumor cells. In the usual experimental data analysis, such multiple observations at each treatment (X_1, U_1) are aggregated into its sample mean or other statistic. However, such aggregation causes huge loss of information. The multiple log-transformed CellTiter-Glo intensity at each (X_1, U_1) can be understood as a random sample from its conditional distribution given (X_1, U_1). Hence, based on the random sample, we estimated the conditional density of the log(intensity) by a kernel density estimator, and treated it as a density response $\mathbf{Y} \equiv Y(\cdot)$. This procedure corresponds to the pre-smoothing in the usual functional data analysis, and similar procedures are adopted in the literature on density-valued data. With the procedure, we obtained the dataset $\{(Y_i(\cdot), X_{1i}, U_{1i}) : 1 \le i \le n\}$, where $n = 72$ is the number of combinations of (X_1, U_1). Figure 1 shows the plot of the densities $Y_i(\cdot)$.

Estimating $\mathrm{E}(Y(\cdot)|X_1, U_1)$ is very important since we can estimate the conditional distributions of outcomes given new values of (X_1, U_1) without conducting new time-consuming and expensive experiments. Clearly, estimating the conditional distributions of outcomes gives much more information than estimating the conditional means of outcomes. We believe that this approach will provide a useful tool in experimental data analysis. The same idea can be also applied to various data analysis such as predicting the distribution of survival times at each treatment, distribution of sales at each sale condition, distribution of income/housing price at each national tax condition and distribution of outputs/defect rates of an item at each process condition in a factory.

We note that our method is the unique nonparametric method for density-valued responses and mixed predictors. To see how the discrete predictor U_1 helps in predicting $Y(\cdot)$, we compared the prediction performance of our estimator with those of the Nadaraya-Watson [10] and the k-nearest neighbor [21] estimators for Hilbertian responses. For the latter two nonparametric estimators, we used only the continuous predictor X_1. The measure of performance was the

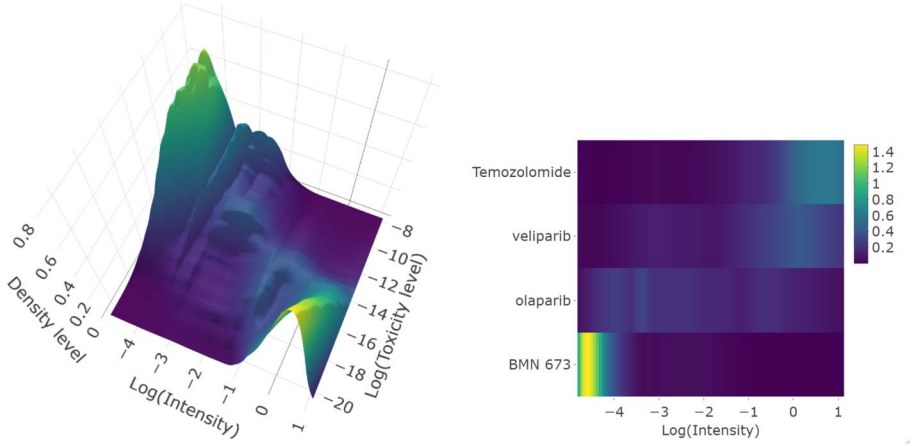

FIG 2. *Estimated component maps for 'toxicity level' (left) and 'drug types' (right).*

leave-one-observation-out average squared prediction error (ASPE) defined by

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

where $\hat{Y}_i^{(-i)}(\cdot)$ is the prediction of $Y_i(\cdot)$ based on the sample without the ith observation. Here, for two density functions $f(\cdot)$ and $g(\cdot)$ supported on a Borel set $S \subset \mathbb{R}$,

$$\|f(\cdot) \ominus g(\cdot)\|^2 = \frac{1}{2\text{Leb}(S)} \int_{S^2} \left(\log\left(\frac{f(\mathbf{s})}{f(\mathbf{s}')}\right) - \log\left(\frac{g(\mathbf{s})}{g(\mathbf{s}')}\right) \right)^2 d\mathbf{s}\, d\mathbf{s}'.$$

We found that the value of ASPE was 23.29 for our estimator, 41.67 for the Nadaraya-Watson estimator and 36.61 for the k-nearest neighbor estimator. This reveals that the discrete predictor and the additive structure may improve substantially the prediction accuracy.

Figure 2 shows the estimated component maps $\hat{\mathbf{m}}_{x,1}$ and $\hat{\mathbf{m}}_{u,1}$. The definitions of $\mathbf{0}, \oplus$ and $\odot$ used to obtain $\hat{\mathbf{m}}_{x,1}$ and $\hat{\mathbf{m}}_{u,1}$ for this case can be found in Section 2.1. We note that, in Figure 2, each line or bar along the log(intensity) at each log(toxicity level) or drug type represents a density. The first plot indicates that, as the toxicity level decreases, the density of the log(intensity) is gradually skewed to the left, while the density tends to be right-skewed as the toxicity level increases. This shows that strong toxicity level tends to kill more tumor cells. It also demonstrates that the toxicity level around $x_1 = -12$ has a similar effect on the intensity of the tumor to those at higher levels ($x_1 > -12$), and thus indicates that the level $x_1 = -12$ is a right dosage for such effect. The second plot says that, as the drug type moves from 'Temozolomide' to 'BMN 673', the density is gradually skewed to the right. This reveals that 'BMN 673' is the most effective drug for the tumor.

5.2.2. Compositional response

It is a general belief that a political election is determined by population characteristics and underlying political orientation. Recently, [15] analyzed the 2017 Korea presidential election data to see the effects of these factors. In that study, however, the effect of underlying political orientation, which is believed to be one of the most important factors, could not be analyzed, because the earlier method can only deal with continuous predictors. In fact, there have been no nonparametric regression method for compositional responses and mixed predictors. This motivated us to analyze the data with the gB-SBF method.

The original dataset analyzed in [15] contains the election result and population characteristics for 250 electoral districts in Korea. For each electoral district as the subject unit we have the proportion of votes earned by five candidates, people's average age (X_1), people's average years of education (X_2), average housing price (X_3) and people's average paid national health insurance premium (X_4). The variables X_3 and X_4 are measures of richness. Since the election was mainly focused on who would be elected among the candidates from the three major parties representing progressive, conservative and middle party, we considered the three-dimensional compositional vector $\mathbf{Y} = (Y_1, Y_2, Y_3)$ as a response with $Y_1 + Y_2 + Y_3 \equiv 1$, where Y_1, Y_2 and Y_3 are the proportions of votes earned by the progressive, conservative and middle party, respectively, divided by the sum of the three proportions.

To incorporate the effect of the underlying political orientation, we added three discrete predictors V_1, V_2 and V_3 representing the number of congress members from the progressive, conservative and middle party, respectively, elected from the 2016 Korea parliamentary election. We excluded two electoral districts since there was a mismatch between the 2017 presidential and the 2016 parliament elections. We also removed two other cases, one with $V_1 = 4$ and the other with $V_2 = 3$ since those values are not well supported by the data. This resulted in a total of $n = 246$ observations with all V_j in the range $\{0, 1, 2\}$, which we actually used in our study.

To assess the prediction performance, we divided the 246 observations into 10 partitions $S_k, 1 \le k \le 10$, with each partition having 24 or 25 observations, and then computed the 10-fold average squared prediction error (ASPE) defined by

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

where $|S_k|$ is the number of observations in S_k and $\hat{\mathbf{Y}}_i^{(-S_k)}$ is the prediction of $\mathbf{Y}_i$ based on the sample without the observations in S_k. Here, for two compositional vectors $\mathbf{a} = (a_1, a_2, a_3)$ and $\mathbf{b} = (b_1, b_2, b_3)$, $\|\mathbf{a} \ominus \mathbf{b}\|^2 = (2 \times 3)^{-1} \sum_{j=1}^{3} \sum_{k=1}^{3} (\log(a_j/a_k) - \log(b_j/b_k))^2$.

To see how the discrete predictors (V_1–V_3) help in predicting $\mathbf{Y}$, we compared the ASPE of our method that was based on the seven predictors of mixed types (X_1–X_4 and V_1–V_3), with the B-SBF estimator [15] that was based on the four

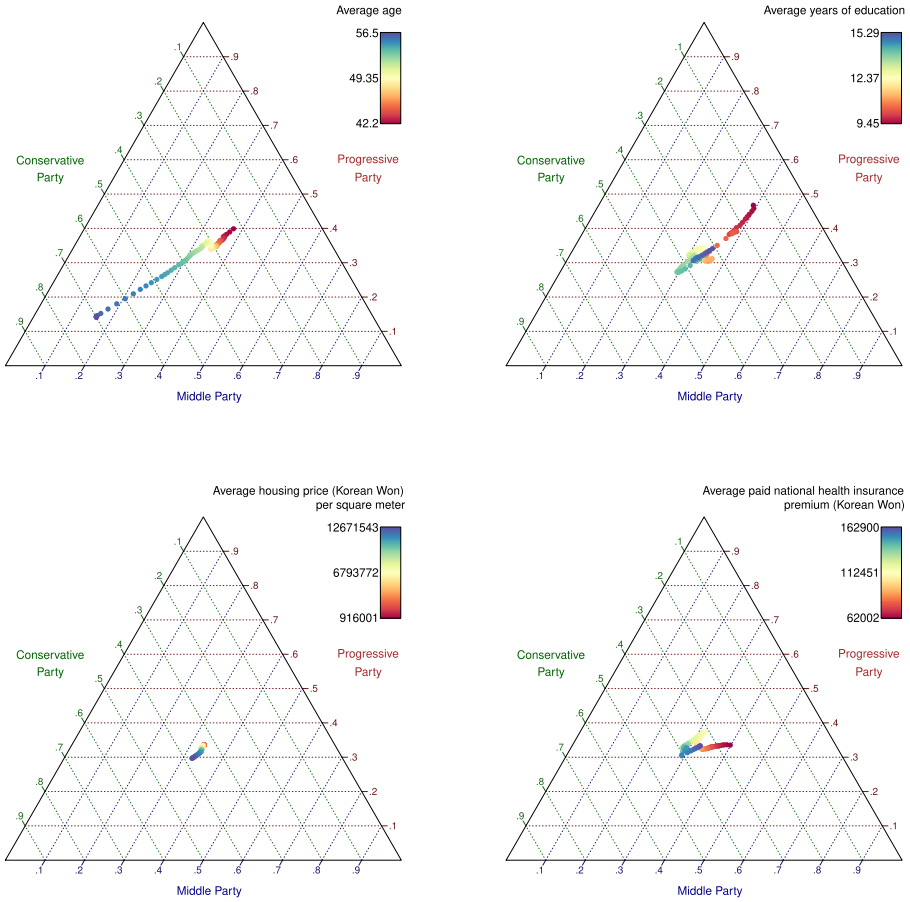

FIG 3. *Component maps estimated by the gB-SBF method: continuous predictors.*

continuous predictors. To see the effect of the additive structure, we also included in the comparison the full-dimensional Nadaraya-Watson-type estimator [27], which was also based on the four continuous predictors. We note that there exists no non-parametric or semi-parametric competitor designed for compositional responses with mixed predictors. We found that the values of ASPE was 0.31 for the gB-SBF, 0.82 for the B-SBF estimator and 0.99 for the full-dimensional Nadaraya-Watson-type estimator. This reveals that the discrete predictors and the additive structure are helpful in predicting $\mathbf{Y}$, confirming the general belief that political orientation is an important factor in election results.

Figures 3 and 4 depict the component maps estimated by the proposed method. The top-left component map in Figure 3 demonstrates clearly that older people are politically more conservative. Overall, richness drives people to conservatism, as indicated in the two component maps at the bottom, al-

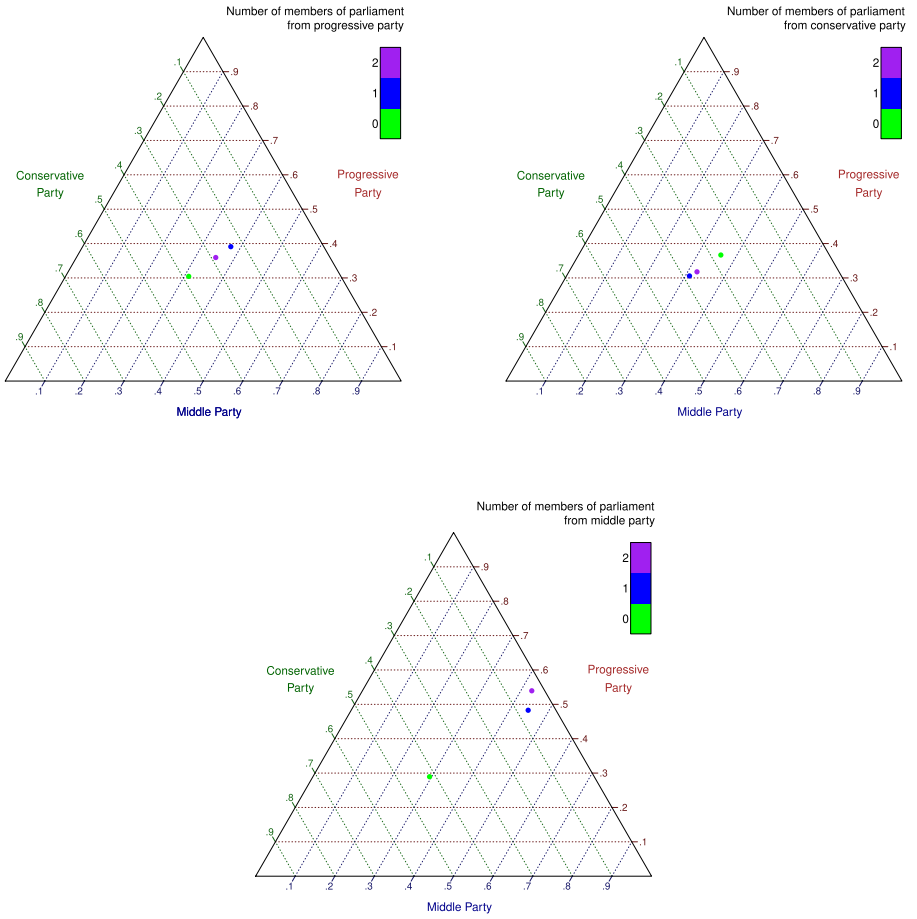

FIG 4. *Component maps estimated by the gB-SBF method: discrete predictors.*

though the strength is much lower than age. As for education level, whose effect is shown in the top-right map, people get more conservative as they are more educated until reaching a level of medium-high and then it is reversed from that level to the highest. The esimated component maps in Figure 4 suggest that larger number in the parliament does not always lead to higher proportion of votes. This indicates that there is some non-monotone relationship between $\mathbf{V} = (V_1, V_2, V_3)$ and $\mathbf{Y}$, contrary to general expectation, which one would not detect with the other methods.

5.2.3. Missing scalar response

We analyzed the 'ACTG175' data in the R package 'speff2trial' (Version 1.0.4). This dataset contains the treatment history of 2,139 patients infected by HIV

type I. The prognosis of the treatment depends on factors, known at the beginning of the treatment, such as the CD4 T cell count at baseline (X_1), hemophilia (yes or no, U_1), history of intravenous drug use (yes or no, U_2), race (white or non-white, U_3), type of treatment (zidovudine only or others, U_4), Karnofsky score ($\{70, 80, 90, 100\}$, V_1) and antiretroviral history stratification ($\{1, 2, 3\}$, V_2). Here, higher V_1 means healthier. For the values of antiretroviral history stratification, $V_2 = 1$ means no prior antiretroviral therapy, $V_2 = 2$ a prior antiretroviral therapy but for a period less than or equal to 52 weeks, and $V_2 = 3$ means a prior antiretroviral therapy for a period more than 52 weeks. We took, as a response variable Y, the CD4 T cell count observed at 96 ± 5 weeks after the observation of X_1. The response variable has 797 missing observations caused by the health condition of patients and/or the cease of the treatment.

We applied the gB-SBF method with the correction for missingness detailed in Example 1 (gB-SBF-correct), and also to the dataset consisting of non-missing observations only (gB-SBF-non-missing). There were 9 observations with V_1 taking the value 70 in the dataset, among which 7 had missing responses. We deleted them from the dataset, so that we could apply a 10-fold cross-validation to the remaining dataset of size 2,130 to compute the prediction error

$$\text{ASPE} = 10^{-1} \sum_{k=1}^{10} n_k^{-1} \sum_{i \in S_k,\ Y_i \text{ is observed}} (Y_i - \hat{Y}_i^{(-S_k)})^2,$$

where $S_k, 1 \le k \le 10$, are partitions each of which is of size 213, n_k is the number of non-missing observations in S_k and $\hat{Y}_i^{(-S_k)}$ is the prediction of Y_i based on the 'gB-SBF-correct' or the 'gB-SBF-non-missing' applied to the observations not in S_k. The ASPE was 14,683 for the 'gB-SBF-correct' while it was 16,383 for the 'gB-SBF-non-missing'. This suggests that our correction for missing observations improves the prediction performance.

Figure 5 depicts the estimated component function of the predictor X_1 based on the 'gB-SBF-correct'. It demonstrates that those having more CD4 T cells at baseline do not necessarily have more CD4 T cell count at 96 ± 5 weeks. For the estimated component functions of the discrete predictors (U_1–U_4 and V_1–V_2), we got $\hat{m}_{U_1}(\text{no}) = 1.02$, $\hat{m}_{U_1}(\text{yes}) = -0.09$, $\hat{m}_{U_2}(\text{no}) = -4.79$, $\hat{m}_{U_2}(\text{yes}) = 0.72$, $\hat{m}_{U_3}(\text{white}) = 3.54$, $\hat{m}_{U_3}(\text{non-white}) = -6.40$, $\hat{m}_{U_4}(\text{zidovudine only}) = -36.32$, $\hat{m}_{U_4}(\text{others}) = 13.30$, $\hat{m}_{V_1}(80) = -47.37$, $\hat{m}_{V_1}(90) = -6.16$, $\hat{m}_{V_1}(100) = 9.27$, $\hat{m}_{V_2}(1) = 24.65$, $\hat{m}_{V_2}(2) = -6.31$ and $\hat{m}_{V_2}(3) = -20.81$. The results suggest that no hemophilia, using an intravenous drug, being white, getting a treatment other than zidovudine, having higher Karnofsky score and/or getting shorter prior antiretroviral therapy, improve the prognosis.

5.2.4. Randomly right-censored scalar response

We next considered the "BMT" data in the R package "KMsurv". In this dataset, there are 137 patients who received a bone marrow transplant as a treatment for their acute leukemia. The prognosis of the transplant depends

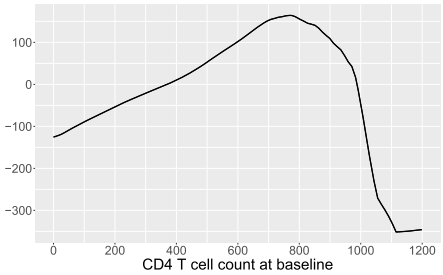

FIG 5. *Estimated component function for 'CD4 T cell count at baseline' based on the gB-SBF-correct method.*

on some risk factors known at the time of transplantation, such as patient and donor's age and gender. To focus on the effects of age and gender, we considered four predictors: patient's age (X_1), donor's age (X_2), patient's gender (U_1) and donor's gender (U_2), to predict survival time. The survival times of 56 patients were censored, so the censoring proportion is about 40%.

We applied to the dataset the gB-SBF method and the SBF method based on varying-coefficient models [44]. The latter models take the form of linear models but the coefficients are functions of continuous predictors. Also, different coefficients should be functions of different continuous predictors. Hence, when there are not enough continuous predictors compared to the number of discrete predictors, it is not possible to apply the latter approach. For this data, there are only two models that the latter approach can deal with:

$$Y = m_1(X_1)U_1 + m_2(X_2)U_2 + \epsilon, \qquad \text{(VCM 1)}$$

$$Y = m_1(X_1)U_2 + m_2(X_2)U_1 + \epsilon, \qquad \text{(VCM 2)}$$

where Y is the survival time as described in Example 2 in Section 4.2. We compared the prediction performance based on the average squared prediction error (ASPE) defined by

$$\text{ASPE} = n_1^{-1} \sum_{i:Y_i \text{ is uncensored}} (Y_i - \hat{Y}_i^{(-i)})^2,$$

where $n_1 = 81$ is the number of uncensored observations and $\hat{Y}_i^{(-i)}$ is the prediction of Y_i based on the sample without the ith uncensored observation. We used the unbiased transformation given in Example 2 to obtain $\hat{Y}_i^{(-i)}$. We also considered regression without the unbiased transformation, which ignores the censoring information.

The results are presented in Table 3. We find that the methods with the unbiased transformation are more predictive than those without it. The results also indicate that the proposed new class of methods, based on the fully nonparametric modeling for both continuous and discrete predictors, works better than the existing class of methods based on the varying-coefficient models.

TABLE 3
Prediction results for the "BMT" data.

Methods	ASPE with transformation	ASPE without transformation
gB-SBF	143552	396573
VCM 1	195266	377007
VCM 2	194146	424525

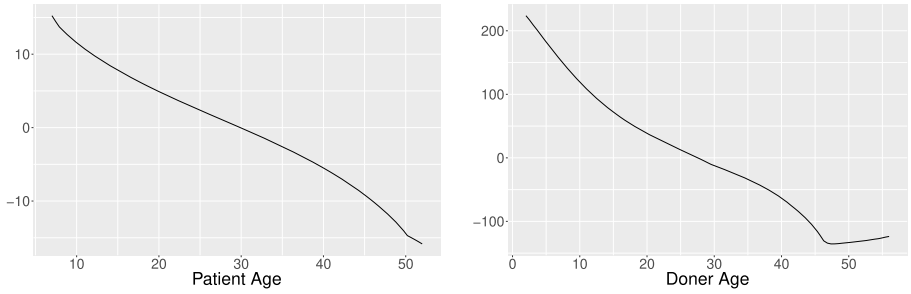

FIG 6. *Estimated component functions for 'patient's age' (left) and 'donor's age' (right) based on the gB-SBF method.*

Figure 6 depicts the estimated component functions for the two continuous predictors based on the proposed method. As for the discrete predictors, we found $\hat{m}_{u,1}(\text{Female}) = 18.35$, $\hat{m}_{u,1}(\text{Male}) = -17.39$, $\hat{m}_{u,2}(\text{Female}) = 73.07$ and $\hat{m}_{u,2}(\text{Male}) = -58.13$. This suggests that, if the patient or the donor is younger or female, then the patient survives longer. In particular, donor's age and gender are more significant factors than patient's age and gender. We remark that, as demonstrated in this application, the gB-SBF method gives easier interpretation than the methods based on varying-coefficient models, since we may assess the effects of the predictors separately.

6. Conclusion and discussion

In this paper, we propose and study a new class of methods for mixed predictors, which is the first fully nonparametric version of the standard linear model. Our framework covers the cases of (in)completely observed Hilbertian responses, such as Euclidean, functional, density-valued and compositional responses, with various types of predictors. Our unified approach based on the novel idea of backfitting the discrete predictors as well as the continuous predictors possesses many advantages over the existing approaches. Also, it is a unique nonparametric method for certain types of responses such as density-valued and compositional responses. The proposed method is supported by a complete theory, and its superiority and a wide spectrum of applications are illustrated via extensive numerical experiments.

In particular, for the existence of the proposed estimators and the convergence of the algorithm materializing them, our theory is developed in full generality under minimal conditions. The conditions are data-specific and are valid

in most cases, which is of great significance and importance to practitioners. The general theory includes non-asymptotic results, works for predictors taking values in general σ-finite measure spaces and for general estimators of the main ingredients of the SBF methodology, such as the densities of the predictors and the marginal regression maps, and even allows dependent data. We think that the general theory and method we develop here casts a long shadow to future study on the subject areas.

Our coverage of compositional responses does not include those with zero entries since spaces that contain such compositional vectors do not form a Hilbert space. For such compositional data with zero entries, one may apply the parametric approach in [38], for example. For density-valued responses with continuous predictors, [11] considered an additive model but on the transformed conditional Fréchet mean via log-quantile and log-hazard transformations studied in [33]. It is different from our additive model that assumes additivity directly on the conditional mean based on the 'Aitchison' geometry given in Section 2.1. The target in [11] is to estimate the conditional Fréchet mean minimizing the expected Wasserstein distance from $\mathbf{Y}$, while our target is to estimate the conditional mean minimizing $\mathrm{E}(\|\mathbf{Y} \ominus \cdot\|^2)$ with the Aitchison norm $\|\cdot\|$.

Appendix

Here, we first present the results for the case of no continuous predictor. Then, we provide some additional details in the methodology and its implementation, followed by additional theoretical results and all technical proofs.

A.1. Case of no continuous predictor

Recall that the results in Section 4 are valid as long as $d_x \geq 1$ and $d_x + d_u + d_v \geq 2$, even in the cases where d_u or d_v equals zero, with trivial modification. Here we complement Section 4 by adding some results for the case where there is no continuous predictor. In the latter case the rates of convergence are different from the case $d_x \geq 1$, so that it is of theoretical interest. It is also important in practice since we encounter many occasions where we only have discrete predictors. For brevity we state the results here for the case $d_u, d_v \geq 1$. However, the results hold as long as $d_u + d_v \geq 2$, even if d_u or d_v is zero, which is clearly seen with trivial modification.

In the case where $d_x = 0$, the corresponding additive model is

$$\mathbf{Y} = \mathbf{m}_0 \oplus \bigoplus_{j=1}^{d_u} \mathbf{m}_{u,j}(U_j) \oplus \bigoplus_{j=1}^{d_v} \mathbf{m}_{v,j}(V_j) \oplus \boldsymbol{\epsilon}, \tag{A.1}$$

where $\mathrm{E}(\mathbf{m}_{u,j}(U_j)) = \mathrm{E}(\mathbf{m}_{v,j}(V_j)) = \mathbf{0}$ for all j and $\mathrm{E}(\boldsymbol{\epsilon}|\mathbf{U}, \mathbf{V}) = \mathbf{0}$. The new definitions of $\hat{\mathbf{m}}_0, \hat{\mathbf{m}}_{u,j}, \hat{\mathbf{m}}_{v,j}, \hat{\mathbf{m}}_{u,j}^{[r]}$ and $\hat{\mathbf{m}}_{v,j}^{[r]}$ are immediate by omitting

$\prod_{j=1}^{d_x} K_{h_j}(x_j, X_{ij})$ in the definition of $\kappa_i(\mathbf{w})$ in Section 4. In this case, we put $\hat{\boldsymbol{\mu}}_+ = \hat{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}$ and $\hat{\boldsymbol{\mu}}_+^{[r]} = \hat{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}^{[r]}$, where

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}(\mathbf{u},\mathbf{v}) &= \hat{\mathbf{m}}_0 \oplus \bigoplus_{j=1}^{d_u} \hat{\mathbf{m}}_{u,j}(u_j) \oplus \bigoplus_{j=1}^{d_v} \hat{\mathbf{m}}_{v,j}(v_j),\\ \hat{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}^{[r]}(\mathbf{u},\mathbf{v}) &= \hat{\mathbf{m}}_0 \oplus \bigoplus_{j=1}^{d_u} \hat{\mathbf{m}}_{u,j}^{[r]}(u_j) \oplus \bigoplus_{j=1}^{d_v} \hat{\mathbf{m}}_{v,j}^{[r]}(v_j).\end{aligned} \tag{A.2}$$

We first note that the corresponding versions of Theorems 1 and 2 hold under the conditions (S3*) and (S4*). For these, we do not need i.i.d. data. However, for the corresponding versions of Theorem 3 and Corollary 1 and for Theorem 6 below, we assume that $\{(\mathbf{W}_i, \mathbf{Y}_i^*) : 1 \le i \le n\}$ are i.i.d. for brevity.

Let p_{uv} be the joint density of $(\mathbf{U}, \mathbf{V})$ with respect to $\bigotimes_{j=1}^{d_u} \mathrm{C}_{u,j} \otimes \bigotimes_{j=1}^{d_v} \mathrm{C}_{v,j}$, where $\mathrm{C}_{u,j}$ and $\mathrm{C}_{v,j}$ are the counting measures on $\mathcal{U}_j$ and $\mathcal{V}_j$, respectively.

Condition (B*).

(B1*) *The joint density p_{uv} is strictly positive on $\prod_{j=1}^{d_u} \mathcal{U}_j \times \prod_{j=1}^{d_v} \mathcal{V}_j$.*
(B2*) *The smoothing parameters satisfy $\lambda_j, s_j = o(1)$ for all j.*
(B3*) $P\left(n^{-1}\sum_{i=1}^n \|\hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*) \ominus \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)\| < M\right) \to 1$ *for some constant* $M > 0$.

We note that (B3*) is weaker than the condition (B5), and under the condition (B*) the corresponding versions of Theorem 3 and Corollary 1 hold. The next theorem shows that the discrete gB-SBF estimator for the model (A.1) may also achieve the univariate error rate.

Theorem 6. *Let $n^{-1}\sum_{i=1}^n \|\hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*) \ominus \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)\| = O_p(b_n)$ hold for some sequence b_n. Assume the conditions (B1*) and (B2*). Then, for all j it holds that*

$$\begin{aligned}\max_{u_j} \|\hat{\mathbf{m}}_{u,j}(u_j) \ominus \mathbf{m}_{u,j}(u_j)\| &= O_p(n^{-1/2} + \lambda_* + s_* + b_n),\\ \max_{v_j} \|\hat{\mathbf{m}}_{v,j}(v_j) \ominus \mathbf{m}_{v,j}(v_j)\| &= O_p(n^{-1/2} + \lambda_* + s_* + b_n).\end{aligned}$$

In a simulation study we present below, we compared the discrete gB-SBF estimator $\hat{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}$ defined at (A.2) with the full-dimensional discrete kernel estimator $\tilde{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}$ defined as in (4.9). Since $\tilde{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}$ is model-free, our focus in this comparison was then to see how $\hat{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}$ perform in non-additive scenarios or in higher-dimension as well as in additive and lower-dimensional settings.

We considered two scenarios, one for a 4-dimensional and the other for a 6-dimensional predictor:

$$Y = m_{u,1}(U_1) + m_{u,2}(U_2) + m_{v,1}(V_1) + m_{v,2}(V_2) + \rho \cdot \frac{m_{u,2}(U_2) \cdot m_{v,2}(V_2)}{m_{u,1}(U_1) \cdot m_{v,2}(V_2)} + \epsilon;$$

TABLE 4
The values of the ratio MSPE($\tilde{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}$)/MSPE($\hat{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}$).

		Additive	Non-Additive			
n	Dimension $(d_u + d_v)$	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$	$\rho = 1$
100	4	4.01	3.10	1.87	1.17	0.81
	6	26.30	5.02	2.09	1.43	1.16
200	4	2.10	1.60	0.97	0.60	0.40
	6	15.02	2.94	1.27	0.93	0.80
400	4	1.37	1.05	0.62	0.36	0.24
	6	8.57	1.73	0.82	0.65	0.57

$$Y = m_{u,1}(U_1) + m_{u,2}(U_2) + m_{v,1}(V_1) + m_{v,2}(V_2) + m_{v,3}(V_3) + m_{v,4}(V_4) + \rho \cdot \frac{m_{u,2}(U_2) \cdot m_{v,2}(V_2) \cdot m_{v,4}(V_4)}{m_{u,1}(U_1) \cdot m_{v,1}(V_1) \cdot m_{v,3}(V_3)} + \epsilon.$$

Here, $m_{u,1}$ and $m_{u,2}$ are the same as those in the first simulation, and $m_{v,1}$ and $m_{v,2}$ are the functions in the linear case. The predictors U_1, U_2, V_1 and V_2 are the same as those in (b) and ϵ is again a $N(0, 0.5^2)$ random variable. The additional ordinal discrete predictors V_3 and V_4 have the same distributions as V_1 and V_2, and $m_{v,3}$ and $m_{v,4}$ are the same as $m_{v,1}$ and $m_{v,2}$ in the nonlinear case in the first simulation. We note that ρ controls the departure from additivity. We took $\rho = 0, 0.25, 0.5, 0.75$ and 1 in both scenarios. We considered (5.1) as a measure of performance with the same n, N and M as in the first simulation. Table 4 shows the results. It shows that the performance of $\tilde{\boldsymbol{\mu}}_{\mathbf{U},\mathbf{V}}$ gets worse as $d_u + d_v$ increases. This indicates that, when the true model departs moderately from additive or the number of discrete predictors is large, the discrete gB-SBF estimator can be a better option than the full-dimensional estimator.

A.2. gB-SBF equation and algorithm for mixed predictors

Here, we articulate the gB-SBF system of equations at (2.15) and its algorithm in Section 2.4 for the model (1.1). For succinct presentation, we first introduce some terminologies. Let $\hat{\mathbf{m}}^{\text{tup}}_{x,+j}$ denote the $(j-1)$-tuple of component maps obtained by taking the first $(j-1)$ components up to $\hat{\mathbf{m}}_{x,j-1}$ from

$$\hat{\mathbf{m}}^{\text{tup}} \equiv (\hat{\mathbf{m}}_{x,1}, \ldots, \hat{\mathbf{m}}_{x,d_x}; \hat{\mathbf{m}}_{u,1}, \ldots, \hat{\mathbf{m}}_{u,d_u}; \hat{\mathbf{m}}_{v,1}, \ldots, \hat{\mathbf{m}}_{v,d_v}),$$

and $\hat{\mathbf{m}}^{\text{tup}}_{x,j+}$ the tuple consisting of those from $\hat{\mathbf{m}}_{x,j+1}$ to $\hat{\mathbf{m}}_{v,d_v}$. Similarly, let $\hat{\mathbf{m}}^{\text{tup}}_{u,+j} = (\hat{\mathbf{m}}_{x,1}, \ldots, \hat{\mathbf{m}}_{u,j-1})$ denote the $(d_x + j - 1)$-tuple and let $\hat{\mathbf{m}}^{\text{tup}}_{u,j+} = (\hat{\mathbf{m}}_{u,j+1}, \ldots, \hat{\mathbf{m}}_{v,d_v})$. Also, let $\hat{\mathbf{m}}^{\text{tup}}_{v,+j} = (\hat{\mathbf{m}}_{x,1}, \ldots, \hat{\mathbf{m}}_{v,j-1})$ be the $(d_x + d_u + j - 1)$-tuple and $\hat{\mathbf{m}}^{\text{tup}}_{v,j+} = (\hat{\mathbf{m}}_{v,j+1}, \ldots, \hat{\mathbf{m}}_{v,d_v})$. For $1 \le j \le d_x$, define $\hat{\boldsymbol{\mu}}_{x,+j}(\cdot;\cdot)$ and $\hat{\boldsymbol{\mu}}_{x,j+}(\cdot;\cdot)$ by

$$\hat{\boldsymbol{\mu}}_{x,+j}(x_j; \hat{\mathbf{m}}^{\text{tup}}_{x,+j}) = \bigoplus_{k \le j-1} \int_0^1 \hat{\mathbf{m}}_{x,k}(x_k) \odot \frac{\hat{p}_{xx,jk}(x_j, x_k)}{\hat{p}_{x,j}(x_j)} dx_k,$$

$$\hat{\boldsymbol{\mu}}_{x,j+}(x_j;\hat{\mathbf{m}}^{\mathrm{tup}}_{x,j+}) = \Bigg(\bigoplus_{k\geq j+1}\int_0^1 \hat{\mathbf{m}}_{x,k}(x_k)\odot\frac{\hat{p}_{xx,jk}(x_j,x_k)}{\hat{p}_{x,j}(x_j)}dx_k\Bigg) \oplus\Bigg(\bigoplus_{k=1}^{d_u}\bigoplus_{u_k\in\mathcal{U}_k}\hat{\mathbf{m}}_{u,k}(u_k)\odot\frac{\hat{p}_{xu,jk}(x_j,u_k)}{\hat{p}_{x,j}(x_j)}\Bigg) \oplus\Bigg(\bigoplus_{k=1}^{d_v}\bigoplus_{v_k\in\mathcal{V}_k}\hat{\mathbf{m}}_{v,k}(v_k)\odot\frac{\hat{p}_{xv,jk}(x_j,v_k)}{\hat{p}_{x,j}(x_j)}\Bigg).$$

Likewise, define $\hat{\boldsymbol{\mu}}_{t,+j}(\cdot;\cdot)$ and $\hat{\boldsymbol{\mu}}_{t,j+}(\cdot;\cdot)$ for $t = u$ and v. For example,

$$\hat{\boldsymbol{\mu}}_{u,+j}(u_j;\hat{\mathbf{m}}^{\mathrm{tup}}_{u,+j}) = \Bigg(\bigoplus_{k=1}^{d_x}\int_0^1 \hat{\mathbf{m}}_{x,k}(x_k)\odot\frac{\hat{p}_{ux,jk}(u_j,x_k)}{\hat{p}_{u,j}(u_j)}dx_k\Bigg) \oplus\Bigg(\bigoplus_{k\leq j-1}\bigoplus_{u_k\in\mathcal{U}_k}\hat{\mathbf{m}}_{u,k}(u_k)\odot\frac{\hat{p}_{uu,jk}(u_j,u_k)}{\hat{p}_{u,j}(u_j)}\Bigg),$$

$$\hat{\boldsymbol{\mu}}_{u,j+}(u_j;\hat{\mathbf{m}}^{\mathrm{tup}}_{u,j+}) = \Bigg(\bigoplus_{k\geq j+1}\bigoplus_{u_k\in\mathcal{U}_k}\hat{\mathbf{m}}_{u,k}(u_k)\odot\frac{\hat{p}_{uu,jk}(u_j,u_k)}{\hat{p}_{u,j}(u_j)}\Bigg) \oplus\Bigg(\bigoplus_{k=1}^{d_v}\bigoplus_{v_k\in\mathcal{V}_k}\hat{\mathbf{m}}_{v,k}(v_k)\odot\frac{\hat{p}_{uv,jk}(u_j,v_k)}{\hat{p}_{u,j}(u_j)}\Bigg).$$

Then, the system of equations defining $\hat{\mathbf{m}}^{\mathrm{tup}}$ is given by

$$\begin{aligned}
\hat{\mathbf{m}}_{x,j}(x_j) &= \hat{\boldsymbol{\mu}}_{x,j}(x_j)\ominus\hat{\mathbf{m}}_0\ominus\hat{\boldsymbol{\mu}}_{x,+j}(x_j;\hat{\mathbf{m}}^{\mathrm{tup}}_{x,+j})\ominus\hat{\boldsymbol{\mu}}_{x,j+}(x_j;\hat{\mathbf{m}}^{\mathrm{tup}}_{x,j+}),\ 1\leq j\leq d_x,\\
\hat{\mathbf{m}}_{u,j}(u_j) &= \hat{\boldsymbol{\mu}}_{u,j}(u_j)\ominus\hat{\mathbf{m}}_0\ominus\hat{\boldsymbol{\mu}}_{u,+j}(u_j;\hat{\mathbf{m}}^{\mathrm{tup}}_{u,+j})\ominus\hat{\boldsymbol{\mu}}_{u,j+}(u_j;\hat{\mathbf{m}}^{\mathrm{tup}}_{u,j+}),\ 1\leq j\leq d_u,\\
\hat{\mathbf{m}}_{v,j}(v_j) &= \hat{\boldsymbol{\mu}}_{v,j}(v_j)\ominus\hat{\mathbf{m}}_0\ominus\hat{\boldsymbol{\mu}}_{v,+j}(v_j;\hat{\mathbf{m}}^{\mathrm{tup}}_{v,+j})\ominus\hat{\boldsymbol{\mu}}_{v,j+}(v_j;\hat{\mathbf{m}}^{\mathrm{tup}}_{v,j+}),\ 1\leq j\leq d_v.
\end{aligned}$$

The constraints corresponding to (2.16) are

$$\begin{aligned}
&\int_0^1 \hat{\mathbf{m}}_{x,j}(x_j)\odot\hat{p}_{x,j}(x_j)dx_j = \mathbf{0},\quad 1\leq j\leq d_x,\\
&\bigoplus_{u_j\in\mathcal{U}_j}\hat{\mathbf{m}}_{u,j}(u_j)\odot\hat{p}_{u,j}(u_j) = \mathbf{0},\quad 1\leq j\leq d_u,\\
&\bigoplus_{v_j\in\mathcal{V}_j}\hat{\mathbf{m}}_{v,j}(v_j)\odot\hat{p}_{v,j}(v_j) = \mathbf{0},\quad 1\leq j\leq d_v.
\end{aligned}$$

Next, to express the gB-SBF algorithm, we let $\hat{\mathbf{m}}^{\mathrm{tup},[r]}_{t,+j}$ and $\hat{\mathbf{m}}^{\mathrm{tup},[r]}_{t,j+}$ for $t = x$, u and v denote the versions of $\hat{\mathbf{m}}^{\mathrm{tup}}_{t,+j}$ and $\hat{\mathbf{m}}^{\mathrm{tup}}_{t,j+}$, respectively, taken from the tuple of the rth updates $(\hat{\mathbf{m}}^{[r]}_{x,1},\ldots,\hat{\mathbf{m}}^{[r]}_{x,d_x};\hat{\mathbf{m}}^{[r]}_{u,1},\ldots,\hat{\mathbf{m}}^{[r]}_{u,d_u};\hat{\mathbf{m}}^{[r]}_{v,1},\ldots,\hat{\mathbf{m}}^{[r]}_{v,d_v})$. Then

the gB-SBF algorithm for the case of mixed predictors is then given by

$$\begin{aligned}
\hat{\mathbf{m}}_{x,j}^{[r]}(x_j) &= \hat{\boldsymbol{\mu}}_{x,j}(x_j) \ominus \hat{\mathbf{m}}_0 \ominus \hat{\boldsymbol{\mu}}_{x,+j}(x_j; \hat{\mathbf{m}}_{x,+j}^{\mathrm{tup},[r]}) \ominus \hat{\boldsymbol{\mu}}_{x,j+}(x_j; \hat{\mathbf{m}}_{x,j+}^{\mathrm{tup},[r-1]}), \\
&\qquad\qquad 1 \le j \le d_x, \\
\hat{\mathbf{m}}_{u,j}^{[r]}(u_j) &= \hat{\boldsymbol{\mu}}_{u,j}(u_j) \ominus \hat{\mathbf{m}}_0 \ominus \hat{\boldsymbol{\mu}}_{u,+j}(u_j; \hat{\mathbf{m}}_{u,+j}^{\mathrm{tup},[r]}) \ominus \hat{\boldsymbol{\mu}}_{u,j+}(u_j; \hat{\mathbf{m}}_{u,j+}^{\mathrm{tup},[r-1]}), \\
&\qquad\qquad 1 \le j \le d_u, \\
\hat{\mathbf{m}}_{v,j}^{[r]}(v_j) &= \hat{\boldsymbol{\mu}}_{v,j}(v_j) \ominus \hat{\mathbf{m}}_0 \ominus \hat{\boldsymbol{\mu}}_{v,+j}(v_j; \hat{\mathbf{m}}_{v,+j}^{\mathrm{tup},[r]}) \ominus \hat{\boldsymbol{\mu}}_{v,j+}(v_j; \hat{\mathbf{m}}_{v,j+}^{\mathrm{tup},[r-1]}), \\
&\qquad\qquad 1 \le j \le d_v.
\end{aligned} \tag{A.3}$$

A.3. Implementation and smoothing parameter selection

We may not evaluate the integrals in (A.3) with the usual numerical integration techniques since Bochner integrals are defined in an abstract way. To implement the gB-SBF algorithm, we adopt the following idea: for any measure space (S, Σ, λ) and for any integrable function $f : S \to \mathbb{R}$ it holds that

$$(\text{Bochner}) \int_S f(s) \odot \mathbf{b}\, d\lambda(s) = (\text{Lebesgue}) \int_S f(s) d\lambda(s) \odot \mathbf{b}, \tag{A.4}$$

where $\mathbf{b}$ is a constant in a Banach space. Because of this, it turns out that the original gB-SBF algorithm at (A.3) based on Bochner integrals may be implemented through a simple iteration scheme based on Lebesgue integrals. Specifically, suppose that we take

$$\begin{aligned}
\hat{\mathbf{m}}_{x,j}^{[0]}(x_j) &= n^{-1} \odot \bigoplus_{i=1}^{n} w_{x,ij}^{[0]}(x_j) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*), \\
\hat{\mathbf{m}}_{u,j}^{[0]}(u_j) &= n^{-1} \odot \bigoplus_{i=1}^{n} w_{u,ij}^{[0]}(u_j) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*), \\
\hat{\mathbf{m}}_{v,j}^{[0]}(v_j) &= n^{-1} \odot \bigoplus_{i=1}^{n} w_{v,ij}^{[0]}(v_j) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*),
\end{aligned} \tag{A.5}$$

with initial weight functions $w_{t,ij}^{[0]}$ for $t = x$, u and v that satisfy

$$\begin{aligned}
\int_0^1 w_{x,ij}^{[0]}(x_j) \hat{p}_{x,j}(x_j) dx_j &= \sum_{u_j} w_{u,ij}^{[0]}(u_j) \hat{p}_{u,j}(u_j) \\
&= \sum_{v_j} w_{v,ij}^{[0]}(v_j) \hat{p}_{v,j}(v_j) = 0,
\end{aligned} \tag{A.6}$$

and $\int_0^1 (w_{x,ij}^{[0]}(x_j))^2 \hat{p}_{x,j}(x_j) dx_j < \infty$. The constraints (A.6) are not restrictive since we may set all initial weights $w_{t,ij}^{[0]} \equiv 0$. Define

$$w_{x,ij}^{[r]}(x_j) = \frac{K_{h_j}(x_j, X_{ij})}{\hat{p}_{x,j}(x_j)} - 1 - \hat{\mu}_{x,+j}(x_j; w_{x,i,+j}^{\mathrm{tup},[r]}) - \hat{\mu}_{x,j+}(x_j; w_{x,i,j+}^{\mathrm{tup},[r-1]}),$$

$$w_{u,ij}^{[r]}(u_j) = \frac{L_{\lambda_j}(u_j, U_{ij})}{\hat{p}_{u,j}(u_j)} - 1 - \hat{\mu}_{u,+j}(u_j; w_{u,i,+j}^{\mathrm{tup},[r]}) - \hat{\mu}_{u,j+}(u_j; w_{u,i,j+}^{\mathrm{tup},[r-1]}),$$
$$w_{v,ij}^{[r]}(v_j) = \frac{W_{s_j}(v_j, V_{ij})}{\hat{p}_{v,j}(v_j)} - 1 - \hat{\mu}_{v,+j}(v_j; w_{v,i,+j}^{\mathrm{tup},[r]}) - \hat{\mu}_{v,j+}(v_j; w_{v,i,j+}^{\mathrm{tup},[r-1]}),$$

where $w_{t,i,+j}^{\mathrm{tup},[r]}$ and $w_{t,i,j+}^{\mathrm{tup},[r]}$ for $t = x$, u and v are defined as $\hat{\mathbf{m}}_{t,+j}^{\mathrm{tup},[r]}$ and $\hat{\mathbf{m}}_{t,j+}^{\mathrm{tup},[r]}$ in Section A.2, respectively, from the $(d_x + d_u + d_v)$-tuples $(w_{x,i1}^{[r]}, \ldots, w_{v,id_v}^{[r]})$ for each $1 \le i \le n$, and $\hat{\mu}_{t,+j}$ and $\hat{\mu}_{t,j+}$ as $\hat{\boldsymbol{\mu}}_{t,+j}$ and $\hat{\boldsymbol{\mu}}_{t,j+}$ with $\oplus$ and $\odot$ being replaced by $+$ and $\times$, respectively. Then, by making use of (A.4) we may verify

$$\begin{aligned}
\hat{\mathbf{m}}_{x,j}^{[r]}(x_j) &= n^{-1} \odot \bigoplus_{i=1}^{n} w_{x,ij}^{[r]}(x_j) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*), \quad 1 \le j \le d_x, \\
\hat{\mathbf{m}}_{u,j}^{[r]}(u_j) &= n^{-1} \odot \bigoplus_{i=1}^{n} w_{u,ij}^{[r]}(u_j) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*), \quad 1 \le j \le d_u, \\
\hat{\mathbf{m}}_{v,j}^{[r]}(v_j) &= n^{-1} \odot \bigoplus_{i=1}^{n} w_{v,ij}^{[r]}(v_j) \odot \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*), \quad 1 \le j \le d_v.
\end{aligned} \tag{A.7}$$

In the case where $\mathbb{H}$ is a space of compositional vectors or density functions, the $\oplus$ and $\odot$ operations in (A.4) and (A.6) may be performed by the usual Euclidean addition $+$ and scalar multiplication $\times$ via centered log-ratio (clr) transformations. In case $\mathbb{H} = \mathcal{S}^k$ for some $k \in \mathbb{N}$, the transformation $\mathrm{clr} : \mathcal{S}^k \to \mathbb{R}^k$ and its inverse clr^{-1} are given by

$$\begin{aligned}
\mathrm{clr}((a_1, \ldots, a_k)) &= \left(\log a_1 - \frac{1}{k} \sum_{j=1}^{k} \log a_j, \ \ldots, \ \log a_k - \frac{1}{k} \sum_{j=1}^{k} \log a_j \right) \\
\mathrm{clr}^{-1}((c_1, \ldots, c_k)) &= \left(\frac{\exp(c_1)}{\sum_{j=1}^{k} \exp(c_j)}, \ \ldots, \ \frac{\exp(c_k)}{\sum_{j=1}^{k} \exp(c_j)} \right).
\end{aligned}$$

In case $\mathbb{H} = \mathfrak{B}^2(S)$, the space of density functions $f(\cdot)$ supported on a Borel subset S of $\mathbb{R}^k$ with finite Lebesgue measure for some $k \in \mathbb{N}$ and satisfying $\int_S (\log f(\mathbf{s}))^2 \, d\mathbf{s} < \infty$, the transformation $\mathrm{clr} : \mathfrak{B}^2(S) \to L^2(S)$ and its inverse clr^{-1} are given by

$$\begin{aligned}
\mathrm{clr}(f(\cdot)) &= \log f(\cdot) - \frac{1}{\mathrm{Leb}_k(S)} \int_S \log f(\mathbf{s}) d\mathbf{s} \\
\mathrm{clr}^{-1}(g(\cdot)) &= \frac{\exp(g(\cdot))}{\int_S \exp(g(\mathbf{s})) d\mathbf{s}}.
\end{aligned}$$

In both cases, $\hat{\mathbf{m}}_{t,j}^{[r]}(t_j)$ for $t = x, u, v$ and $r \ge 0$ can be computed by

$$\mathrm{clr}^{-1} \left(n^{-1} \sum_{i=1}^{n} w_{t,ij}^{[r]}(t_j) \cdot \mathrm{clr}\big(\hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*)\big) \right).$$

In our numerical studies in Section 5, we used the Epanechnikov kernel $K(t) = (3/4)(1-t^2)I(|t|<1)$. We chose the initial weights as follows:

$$w_{x,ij}^{[0]}(x_j) = \frac{K_{h_j}(x_j, X_{ij})}{\hat{p}_{x,j}(x_j)} - 1,$$
$$w_{u,ij}^{[0]}(u_j) = \frac{L_{\lambda_j}(u_j, U_{ij})}{\hat{p}_{u,j}(u_j)} - 1,$$
$$w_{v,ij}^{[0]}(v_j) = \frac{W_{s_j}(v_j, V_{ij})}{\hat{p}_{v,j}(v_j)} - 1,$$

so that they are square integrable and satisfy (A.6). Instead of applying the stopping criteria described in the gB-SBF algorithm in Section 2.4, we simply stopped the iteration when the following criteria were all met:

$$\begin{aligned}
\max_{1\le j\le d_x} \int_0^1 \|\hat{\mathbf{m}}_{x,j}^{[r]}(x_j) \ominus \hat{\mathbf{m}}_{x,j}^{[r-1]}(x_j)\|^2 \hat{p}_{x,j}(x_j)dx_j < 10^{-4},\\
\max_{1\le j\le d_u} \sum_{u_j} \|\hat{\mathbf{m}}_{u,j}^{[r]}(u_j) \ominus \hat{\mathbf{m}}_{u,j}^{[r-1]}(u_j)\|^2 \hat{p}_{u,j}(u_j) < 10^{-4},\\
\max_{1\le j\le d_v} \sum_{v_j} \|\hat{\mathbf{m}}_{v,j}^{[r]}(v_j) \ominus \hat{\mathbf{m}}_{v,j}^{[r-1]}(v_j)\|^2 \hat{p}_{v,j}(v_j) < 10^{-4}.
\end{aligned} \tag{A.8}$$

Now, we discuss the selection of the smoothing parameters h_j, λ_j and s_j. We note that a full-dimensional grid search is not feasible when the number of predictors, $d_x + d_u + d_v$, is large. We propose a selection rule, named as "CSS"(Coordinate-wise Smoothing-parameter Selection). The same idea was employed in [15]. Let $\mathrm{CV}(h_1, \ldots, s_{d_v})$ denote a cross-validatory criterion for the tuple of smoothing parameters $(h_1, \ldots, s_{d_v})$.

CSS algorithm. Take grids $\mathcal{G}_t = \prod_{j=1}^{d_t} \{g_{t,j,1}, \ldots, g_{t,j,L_{t,j}}\}$ for $t = x$, u and v with $L_{t,j} \in \mathbb{N}$. Choose initial smoothing parameters $h_1^{(0)}, \ldots, h_{d_x}^{(0)}$, $\lambda_1^{(0)}, \ldots, \lambda_{d_u}^{(0)}$, $s_1^{(0)}, \ldots, s_{d_v}^{(0)}$ from the respective grids. For $l = 1, 2, \ldots$, find

$$h_j^{(l)} = \underset{g_j \in \{g_{x,j,1}, \ldots, g_{x,j,L_{x,j}}\}}{\arg\min} \mathrm{CV}(h_1^{(l)}, \ldots, h_{j-1}^{(l)}, g_j, h_{j+1}^{(l-1)}, \ldots, s_{d_v}^{(l-1)}), \quad 1 \le j \le d_x,$$
$$\lambda_j^{(l)} = \underset{g_j \in \{g_{u,j,1}, \ldots, g_{u,j,L_{u,j}}\}}{\arg\min} \mathrm{CV}(h_1^{(l)}, \ldots, \lambda_{j-1}^{(l)}, g_j, \lambda_{j+1}^{(l-1)}, \ldots, s_{d_v}^{(l-1)}), \quad 1 \le j \le d_u,$$
$$s_j^{(t)} = \underset{g_j \in \{g_{v,j,1}, \ldots, g_{v,j,L_{v,j}}\}}{\arg\min} \mathrm{CV}(h_1^{(l)}, \ldots, s_{j-1}^{(l)}, g_j, s_{j+1}^{(l-1)}, \ldots, s_{d_v}^{(l-1)}), \quad 1 \le j \le d_v.$$

Repeat the procedure until $(h_1^{(l)}, \ldots, s_{d_v}^{(l)}) = (h_1^{(l-1)}, \ldots, s_{d_v}^{(l-1)})$. □

We note that the CSS algorithm always ends in finite steps since the grid size is finite. Also, the selected vector of smoothing parameters achieves a coordinate-

wise minimum. In our numerical study, we used a 10-fold cross-validation. For the gB-SBF and for the methods of [45] and of [44], we chose $\mathcal{G}_x = \prod_{j=1}^{d_x}\{a_j + 0.01 \times k : k = 0, \ldots, 100\}$ for some small values a_j that satisfy (S1*). We also took the bandwidth grid $\{a + 0.01 \times k : k = 0, \ldots, 100\}$ for some small $a > 0$ for the methods of [10] and of [27], and took the nearest neighbor grid $\{1, \ldots, 50\}$ for the method of [21]. We chose $\mathcal{G}_u = \prod_{j=1}^{d_u}\{0.02 \times k : k = 0, \ldots, 50\}$ and $\mathcal{G}_v = \prod_{j=1}^{d_v}\{b_j/50 \times k : k = 0, \ldots, 50\}$, which we used for the gB-SBF method to fit the model (A.1) as well as (1.1), where $b_j \in [0, 1]$ are some small values satisfying $\sum_{v_j' \in \mathcal{V}_j : v_j' \neq v_j} b_j^{\delta_j(v_j, v_j')} \leq 1$ for all v_j.

A.4. Closedness of $\mathcal{S}^{\mathbb{H}}(\hat{p})$

In this subsection, we prove the closedness of $\mathcal{S}^{\mathbb{H}}(\hat{p})$ defined at (3.1). The proof is largely based on the projection theory of Hilbert spaces. The materials covered here is used in the proofs of other theoretical results. Recall the definition of the probability measure $\hat{P}\mathbf{Z}^{-1}$ introduced immediately below (2.9). Define an inner product $\langle \cdot, \cdot \rangle_{2,n}$ of $L^2((\mathcal{Z}, \mathscr{A}, \hat{P}\mathbf{Z}^{-1}), \mathbb{H})$ by

$$\langle \mathbf{f}, \mathbf{g} \rangle_{2,n} = \int_{\mathcal{Z}} \langle \mathbf{f}(\mathbf{z}), \mathbf{g}(\mathbf{z}) \rangle d\hat{P}\mathbf{Z}^{-1}(\mathbf{z}) = \int_{\mathcal{Z}} \langle \mathbf{f}(\mathbf{z}), \mathbf{g}(\mathbf{z}) \rangle \hat{p}(\mathbf{z}) d\nu(\mathbf{z}),$$

where $\langle \cdot, \cdot \rangle$ is an inner product of $\mathbb{H}$. Then, the $L^2((\mathcal{Z}, \mathscr{A}, \hat{P}\mathbf{Z}^{-1}), \mathbb{H})$ with the inner product $\langle \cdot, \cdot \rangle_{2,n}$ is a Hilbert space.

The closedness of $\mathcal{S}^{\mathbb{H}}(\hat{p})$ is essential for the proof of the existence of the gB-SBF estimators, since it is a part of a sufficient condition for the existence of a minimizer of the objective functional $\hat{F}$ defined in Section 3.2, see Lemma 4 in [5]. The closedness is also important for the convergence of the gB-SBF algorithm. To see why, define the linear operators $\hat{\pi}_j : L^2((\mathcal{Z}, \mathscr{A}, \hat{P}\mathbf{Z}^{-1}), \mathbb{H}) \to L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$ by

$$\hat{\pi}_j(\mathbf{f})(z_j) = \int_{\mathcal{Z}_{-j}} \mathbf{f}(\mathbf{z}) \odot (\hat{p}(\mathbf{z})/\hat{p}_j(z_j)) d\nu_{-j}(\mathbf{z}_{-j}), \quad \mathbf{f} \in L^2((\mathcal{Z}, \mathscr{A}, \hat{P}\mathbf{Z}^{-1}), \mathbb{H})$$

whenever the integral exists, and put $\hat{\pi}_j(\mathbf{f})(z_j) = \mathbf{0}$ otherwise. The following proposition shows that $\hat{\pi}_j$ are projection operators.

Proposition 1. *If $\hat{p}_j(z_j) > 0$ for all $z_j \in \mathcal{Z}_j$, then $\hat{\pi}_j$ is a projection operator.*

Proof. For $\mathbf{f} \in L^2((\mathcal{Z}, \mathscr{A}, \hat{P}\mathbf{Z}^{-1}), \mathbb{H})$, define

$$D_j(\mathbf{f}) = \{z_j \in \mathcal{Z}_j : \int_{\mathcal{Z}_{-j}} \|\mathbf{f}(\mathbf{z})\| \hat{p}(\mathbf{z}) d\nu_{-j}(\mathbf{z}_{-j}) < \infty\}.$$

We note that $\nu_j(\mathcal{Z}_j \setminus D_j(\mathbf{f})) = 0$. Then, for $\mathbf{f}_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$, it

holds that

$$
\begin{aligned}
&\langle \mathbf{f} \ominus \hat{\pi}_j(\mathbf{f}), \mathbf{f}_j \rangle_{2,n} \\
&= \int_{\mathcal{Z}} \left\langle \mathbf{f}(\mathbf{z}) \ominus \int_{\mathcal{Z}_{-j}} \mathbf{f}(\mathbf{z}) \odot \frac{\hat{p}(\mathbf{z})}{\hat{p}_j(z_j)} d\nu_{-j}(\mathbf{z}_{-j}), \mathbf{f}_j(z_j) \right\rangle \hat{p}(\mathbf{z}) d\nu(\mathbf{z}) \\
&= \int_{\mathcal{Z}_j} \int_{\mathcal{Z}_{-j}} \left\langle \left(\mathbf{f}(\mathbf{z}) \ominus \int_{\mathcal{Z}_{-j}} \mathbf{f}(\mathbf{z}) \odot \frac{\hat{p}(\mathbf{z})}{\hat{p}_j(z_j)} d\nu_{-j}(\mathbf{z}_{-j}) \right) \odot \hat{p}(\mathbf{z}), \mathbf{f}_j(z_j) \right\rangle \\
&\qquad\qquad d\nu_{-j}(\mathbf{z}_{-j}) d\nu_j(z_j) \\
&= \int_{\mathcal{Z}_j} \left\langle \int_{\mathcal{Z}_{-j}} \left(\mathbf{f}(\mathbf{z}) \ominus \int_{\mathcal{Z}_{-j}} \mathbf{f}(\mathbf{z}) \odot \frac{\hat{p}(\mathbf{z})}{\hat{p}_j(z_j)} d\nu_{-j}(\mathbf{z}_{-j}) \right) \odot \hat{p}(\mathbf{z}) d\nu_{-j}(\mathbf{z}_{-j}), \mathbf{f}_j(z_j) \right\rangle \\
&\qquad\qquad d\nu_j(z_j) \\
&= \mathbf{0}.
\end{aligned}
$$

This shows that $\hat{\pi}_j$ is a projection operator. □

Now, define a linear operator $\hat{T} : \mathcal{S}^{\mathbb{H}}(\hat{p}) \to \mathcal{S}^{\mathbb{H}}(\hat{p})$ by

$$\hat{T} = (I - \hat{\pi}_d) \circ \cdots \circ (I - \hat{\pi}_1), \tag{A.9}$$

where I is the identity operator. We note that $\hat{T}$ is an alternating projection operator. According to the projection theory (Lemma S.7 in [15]), the closedness of $\mathcal{S}^{\mathbb{H}}(\hat{p})$ implies that $\hat{T}$ is a contraction, i.e.,

$$\|\hat{T}\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))} := \sup\{\|\hat{T}(\mathbf{g})\|_{2,n} : \mathbf{g} \in \mathcal{S}^{\mathbb{H}}(\hat{p}),\ \|\mathbf{g}\|_{2,n} = 1\} < 1, \tag{A.10}$$

where $\|\cdot\|_{2,n}$ is the norm induced by $\langle \cdot, \cdot \rangle_{2,n}$. The property $\|\hat{T}\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))} < 1$ is essential for the convergence of the gB-SBF algorithm since the constant $\hat{\gamma}$ in Theorem 2 is in fact $\|\hat{T}\|^2_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))}$, see the proof of Theorem 2 given in Section A.6.2.

Proving the closedness of $\mathcal{S}^{\mathbb{H}}(\hat{p})$ requires an advanced theory of functional analysis. To describe this, let $\hat{\pi}_j | L^2((\mathcal{Z}_k, \mathscr{A}_k, \hat{P}Z_k^{-1}), \mathbb{H})$ denote the operator $\hat{\pi}_j$ restricted to $L^2((\mathcal{Z}_k, \mathscr{A}_k, \hat{P}Z_k^{-1}), \mathbb{H})$ for $k \neq j$. When $\mathbb{H} = \mathbb{R}$, $\mathcal{Z}_j = [0,1]$ and ν_j are the Lebesgue measure for all $1 \leq j \leq d$, the common approach in the existing SBF literature to establishing the closeness of $\mathcal{S}^{\mathbb{H}}(\hat{p})$ is to prove that $\hat{\pi}_j | L^2((\mathcal{Z}_k, \mathscr{A}_k, \hat{P}Z_k^{-1}), \mathbb{H})$ for all $1 \leq j \neq k \leq d$ are compact operators (e.g. [25]). Indeed, if $\hat{\pi}_j$ are projection operators and $\hat{\pi}_j | L^2((\mathcal{Z}_k, \mathscr{A}_k, \hat{P}Z_k^{-1}), \mathbb{H})$ are compact, then $\mathcal{S}^{\mathbb{H}}(\hat{p})$ is closed by Theorem 8.1 in [7]. Unfortunately, the restricted projection operators are not compact for infinite-dimensional $\mathbb{H}$, as we demonstrate it below.

Proposition 2. *Under the condition (S2), $\hat{\pi}_j | L^2((\mathcal{Z}_k, \mathscr{A}_k, \hat{P}Z_k^{-1}), \mathbb{H})$ are compact if and only if $\mathbb{H}$ is finite-dimensional.*

Proof. Let $\mathcal{L}(\mathbb{H})$ denote the space of all bounded and linear operators from $\mathbb{H}$ to itself. Then, under the condition (S2), $\hat{\pi}_j | L^2((\mathcal{Z}_k, \mathscr{A}_k, \hat{P}Z_k^{-1}), \mathbb{H})$ is an integral operator with the kernel $\hat{\mathbf{k}}_{jk} : \mathcal{Z} \times \mathcal{Z} \to \mathcal{L}(\mathbb{H})$ defined by $\hat{\mathbf{k}}_{jk}(\mathbf{z}, \mathbf{z}^*)(\mathbf{h}) = \mathbf{h} \odot (\hat{p}_{jk}(z_j, z_k^*)/(\hat{p}_j(z_j)\hat{p}_k(z_k^*)))$. This with Theorem 3.1 and Theorem 3.2 in [15] gives the proposition. □

To prove that $\mathcal{S}^{\mathbb{H}}(\hat{p})$ is closed, we use Proposition 1 and Lemma S.7 in [15], the latter of which tells that the closedness of $\mathcal{S}^{\mathbb{H}}(\hat{p})$ is equivalent to the conclusion of the next proposition.

Proposition 3. *Under the condition (S2), there exists a constant $\hat{c} > 0$ such that, for all $\mathbf{f} \in \mathcal{S}^{\mathbb{H}}(\hat{p})$, there exists a tuple $(\mathbf{f}_j : 1 \le j \le d)$ with $\mathbf{f}_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$ for all $1 \le j \le d$ satisfying $\bigoplus_{j=1}^{d} \mathbf{f}_j = \mathbf{f}$ a.e. with respect to $\hat{P}\mathbf{Z}^{-1}$ and $\sum_{j=1}^{d} \|\mathbf{f}_j\|_{2,n}^2 \le \hat{c} \cdot \|\mathbf{f}\|_{2,n}^2$.*

Proof. We note that Proposition 3 in specialization to finite-dimensional $\mathbb{H}$ follows from Theorem 8.1 in [7], Lemma S.7 in [15] and Proposition 2. Now, suppose that $\mathbb{H}$ is infinite-dimensional, and let $\{\mathbf{e}_k\}_{k=1}^{\infty}$ be an orthonormal basis of $\mathbb{H}$. For a given $\mathbf{f} \in \mathcal{S}^{\mathbb{H}}(\hat{p})$, it holds that $\mathbf{f}(\mathbf{z}) = \bigoplus_{k=1}^{\infty} \langle \mathbf{f}(\mathbf{z}), \mathbf{e}_k \rangle \odot \mathbf{e}_k$ and $\|\mathbf{f}(\mathbf{z})\|^2 = \sum_{k=1}^{\infty} \langle \mathbf{f}(\mathbf{z}), \mathbf{e}_k \rangle^2$ for all $\mathbf{z} \in \mathcal{Z}$. Thus, we have

$$\|\mathbf{f}\|_{2,n}^2 = \int_{\mathcal{Z}} \sum_{k=1}^{\infty} \langle \mathbf{f}(\mathbf{z}), \mathbf{e}_k \rangle^2 \hat{p}(\mathbf{z}) d\nu(\mathbf{z}) = \sum_{k=1}^{\infty} \|\langle \mathbf{f}(\cdot), \mathbf{e}_k \rangle\|_{2,n}^2,$$

where with slight abuse of the notation for $\|\cdot\|_{2,n}$, we write

$$\|g\|_{2,n}^2 = \int_{\mathcal{Z}} |g(\mathbf{z})|^2 \hat{p}(\mathbf{z}) d\nu(\mathbf{z})$$

for real-valued maps $g \in L^2((\mathcal{Z}, \mathscr{A}, \hat{P}\mathbf{Z}^{-1}), \mathbb{R})$ as well. Proposition 3 in specialization to $\mathbb{H} = \mathbb{R}$ implies that there exists a constant $\hat{c} > 0$ such that, for any $g \in \mathcal{S}^{\mathbb{R}}(\hat{p})$, there exist $g_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{R})$ for $1 \le j \le d$ satisfying $g = \sum_{j=1}^{d} g_j$ a.e. with respect to $\hat{P}\mathbf{Z}^{-1}$ and $\sum_{j=1}^{d} \|g_j\|_{2,n}^2 \le \hat{c} \cdot \|g\|_{2,n}^2$. Since $\langle \mathbf{f}(\cdot), \mathbf{e}_k \rangle \in \mathcal{S}^{\mathbb{R}}(\hat{p})$ for all $k \ge 1$, there exist $f_{kj} \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{R})$ for $1 \le j \le d$ satisfying $\langle \mathbf{f}(\cdot), \mathbf{e}_k \rangle = \sum_{j=1}^{d} f_{kj}$ a.e. with respect to $\hat{P}\mathbf{Z}^{-1}$ and $\sum_{j=1}^{d} \|f_{kj}\|_{2,n}^2 \le \hat{c} \cdot \|\langle \mathbf{f}(\cdot), \mathbf{e}_k \rangle\|_{2,n}^2$. Thus, it holds that

$$\sum_{j=1}^{d} \sum_{k=1}^{\infty} \|f_{kj}\|_{2,n}^2 \le \hat{c} \cdot \sum_{k=1}^{\infty} \|\langle \mathbf{f}(\cdot), \mathbf{e}_k \rangle\|_{2,n}^2 = \hat{c} \cdot \|\mathbf{f}\|_{2,n}^2 < \infty. \tag{A.11}$$

Now, (A.11) implies that, for each $1 \le j \le d$, the sequence $\{\bigoplus_{k=1}^{N} f_{kj}(\cdot) \odot \mathbf{e}_k\}_{N \ge 1}$ is Cauchy in $L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$ since

$$\begin{aligned} \Big\| \bigoplus_{k=m+1}^{N} f_{kj}(\cdot) \odot \mathbf{e}_k \Big\|_{2,n}^2 &= \int_{\mathcal{Z}} \sum_{k=m+1}^{N} \|f_{kj}(\mathbf{z}) \odot \mathbf{e}_k\|^2 \hat{p}(\mathbf{z}) d\nu(\mathbf{z}) \\ &= \sum_{k=m+1}^{N} \|f_{kj}\|_{2,n}^2 \\ &\to 0 \end{aligned}$$

as $N > m \to \infty$. Denote the limit of the Cauchy sequence by $\mathbf{f}_j$. Then, there exists a subsequence $\{\bigoplus_{k=1}^{N_{jl}} f_{kj}(\cdot) \odot \mathbf{e}_k\}_{l \geq 1}$ of $\{\bigoplus_{k=1}^{N} f_{kj}(\cdot) \odot \mathbf{e}_k\}_{N \geq 1}$ such that

$$\lim_{l \to \infty} \bigoplus_{k=1}^{N_{jl}} f_{kj}(\mathbf{z}) \odot \mathbf{e}_k = \mathbf{f}_j(\mathbf{z}) \text{ a.e. with respect to } \hat{P}\mathbf{Z}^{-1}.$$

Then, it holds that

$$\sum_{j=1}^{d} \|\mathbf{f}_j\|_{2,n}^2 = \sum_{j=1}^{d} \int_{\mathcal{Z}} \Big(\lim_{l \to \infty} \sum_{k=1}^{N_{jl}} f_{kj}^2(\mathbf{z}) \Big) \hat{p}(\mathbf{z}) d\nu(\mathbf{z}) = \sum_{j=1}^{d} \sum_{k=1}^{\infty} \|f_{kj}\|_{2,n}^2 \leq \hat{c} \|\mathbf{f}\|_{2,n}^2,$$

where the inequality follows from (A.11). Moreover, we get

$$\begin{aligned}
\bigoplus_{j=1}^{d} \mathbf{f}_j(\mathbf{z}) &= \bigoplus_{k=1}^{\infty} \Big(\sum_{j=1}^{d} \langle \lim_{l \to \infty} \bigoplus_{i=1}^{N_{jl}} f_{ij}(\mathbf{z}) \odot \mathbf{e}_i, \mathbf{e}_k \rangle \Big) \odot \mathbf{e}_k \\
&= \bigoplus_{k=1}^{\infty} \Big(\sum_{j=1}^{d} \lim_{l \to \infty} \langle \bigoplus_{i=1}^{N_{jl}} f_{ij}(\mathbf{z}) \odot \mathbf{e}_i, \mathbf{e}_k \rangle \Big) \odot \mathbf{e}_k \\
&= \bigoplus_{k=1}^{\infty} \Big(\sum_{j=1}^{d} f_{kj}(\mathbf{z}) \Big) \odot \mathbf{e}_k \\
&= \bigoplus_{k=1}^{\infty} \langle \mathbf{f}(\mathbf{z}), \mathbf{e}_k \rangle \odot \mathbf{e}_k = \mathbf{f}(\mathbf{z})
\end{aligned}$$

a.e. with respect to $\hat{P}\mathbf{Z}^{-1}$. This completes the proof. □

The following proposition now follows from Proposition 3.

Proposition 4. *If the condition (S2) holds, then $\mathcal{S}^{\mathbb{H}}(\hat{p})$ is a closed subspace of $L^2((\mathcal{Z}, \mathscr{A}, \hat{P}\mathbf{Z}^{-1}), \mathbb{H})$ and $\|\hat{T}\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))} < 1$.*

A.5. Some lemmas

For the below lemma, we write $P\mathbf{Z}_{-j}^{-1}$ for the distribution of $\mathbf{Z}_{-j}$ and write $p_{\mathbf{Z}_{-j}}$ for the density of $\mathbf{Z}_{-j}$ with respect to ν_{-j}. The lemma is used to prove Theorems 3, 4 and 5.

Lemma 1. *Assume that there is a constant $c > 0$ such that $p(\mathbf{z}) \geq c \cdot p_j(z_j) \cdot p_{\mathbf{Z}_{-j}}(\mathbf{z}_{-j})$ for all $1 \leq j \leq d$ and $\mathbf{z} \in \mathcal{Z}$. Let $\mathbf{f}_j : \mathcal{Z}_j \to \mathbb{H}$ be measurable maps for $1 \leq j \leq d$. If $\bigoplus_{j=1}^{d} \mathbf{f}_j(z_j) = \mathbf{0}$ a.e. with respect to $P\mathbf{Z}^{-1}$, then $\mathbf{f}_j(z_j) = \mathbf{c}_j$ a.e. with respect to PZ_j^{-1} for all $1 \leq j \leq d$, where $\mathbf{c}_j \in \mathbb{H}$ are some constants satisfying $\bigoplus_{j=1}^{d} \mathbf{c}_j = \mathbf{0}$.*

Proof. We only show that $\mathbf{f}_1(z_1) = \mathbf{c}_1$ a.e. with respect to PZ_1^{-1}, since for $j \geq 2$ we may simply exchange the roles of 1 and j. We claim that, if $p(\mathbf{z}) \geq c \cdot p_1(z_1) \cdot p_{\mathbf{Z}_{-1}}(\mathbf{z}_{-1})$ for all $\mathbf{z} \in \mathcal{Z}$, then

$$PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1} \ll P\mathbf{Z}^{-1}. \tag{A.12}$$

Let $N \in \mathscr{A}$ be a $P\mathbf{Z}^{-1}$-null set. Then,

$$0 = P\mathbf{Z}^{-1}(N) \geq c \int_{\mathcal{Z}} 1_N(\mathbf{z}) p_1(z_1) p_{\mathbf{Z}_{-1}}(\mathbf{z}_{-1}) d\nu(\mathbf{z}) = c \cdot PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1}(N) \geq 0.$$

This proves (A.12). Let $E = \{\mathbf{z} \in \mathcal{Z} : \mathbf{f}_1(z_1) = -1 \odot \bigoplus_{k=2}^d \mathbf{f}_k(z_k)\}$. We note that

$$1_E(\mathbf{z}) \odot \mathbf{f}_1(z_1) = (-1_E(\mathbf{z})) \odot \bigoplus_{k=2}^d \mathbf{f}_k(z_k), \quad \mathbf{z} \in \mathcal{Z}.$$

For $D \in \mathcal{B}(\mathbb{H})$, we may prove that

$$(1_E \odot \mathbf{f}_1)^{-1}(D) = \begin{cases} (A \times \mathcal{Z}_{-1}) \cap E, & \text{if } \mathbf{0} \notin D, \\ (A \times \mathcal{Z}_{-1}) \cup E^c, & \text{if } \mathbf{0} \in D, \end{cases}$$

$$\Big((-1_E) \odot \bigoplus_{k=2}^d \mathbf{f}_k\Big)^{-1}(D) = \begin{cases} (\mathcal{Z}_1 \times B) \cap E, & \text{if } \mathbf{0} \notin D, \\ (\mathcal{Z}_1 \times B) \cup E^c, & \text{if } \mathbf{0} \in D, \end{cases}$$

for some $A \in \mathscr{A}_1$ and $B \in \mathscr{A}_{-1}$.

First, consider the case $\mathbf{0} \notin D$. In this case,

$$\begin{aligned} (A \times \mathcal{Z}_{-1}) \cap E &= (\mathcal{Z}_1 \times B) \cap E \\ &= ((A \times \mathcal{Z}_{-1}) \cap E) \cap ((\mathcal{Z}_1 \times B) \cap E) \\ &= (A \times B) \cap E. \end{aligned} \tag{A.13}$$

Since $\mathbf{f}_1(z_1) = -1 \odot \bigoplus_{k=2}^d \mathbf{f}_k(z_k)$ a.e. with respect to $P\mathbf{Z}^{-1}$, we get $P\mathbf{Z}^{-1}(E) = 1$. Then, (A.12) implies that

$$PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1}(E) = 1. \tag{A.14}$$

From (A.13) and (A.14), it follows that

$$\begin{aligned} PZ_1^{-1}(A) &= PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1}(A \times \mathcal{Z}_{-1}) \\ &= PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1}((A \times \mathcal{Z}_{-1}) \cap E) \\ &= PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1}((A \times B) \cap E) \\ &= PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1}(A \times B) \\ &= PZ_1^{-1}(A) P\mathbf{Z}_{-1}^{-1}(B), \end{aligned} \tag{A.15}$$

$$\begin{aligned} P\mathbf{Z}_{-1}^{-1}(B) &= PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1}(\mathcal{Z}_1 \times B) \\ &= PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1}((\mathcal{Z}_1 \times B) \cap E) \\ &= PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1}((A \times B) \cap E) \\ &= PZ_1^{-1} \otimes P\mathbf{Z}_{-1}^{-1}(A \times B) \\ &= PZ_1^{-1}(A) P\mathbf{Z}_{-1}^{-1}(B). \end{aligned} \tag{A.16}$$

From (A.15) and (A.16), we have $PZ_1^{-1}(A) = 0$ or 1. When $\mathbf{0} \in D$, a similar argument shows that $PZ_1^{-1}(A) = 0$ or 1. Thus, $PZ_1^{-1}(\mathbf{f}_1^{-1}(D)) = 0$ or 1 for any $D \in \mathcal{B}(\mathbb{H})$. For a measure $PZ_1^{-1}\mathbf{f}_1^{-1}$ on $\mathcal{B}(\mathbb{H})$ defined by $PZ_1^{-1}\mathbf{f}_1^{-1}(D) = PZ_1^{-1}(\mathbf{f}_1^{-1}(D))$, it holds that $PZ_1^{-1}\mathbf{f}_1^{-1}(\mathbb{H}) = 1$. Also, if $PZ_1^{-1}\mathbf{f}_1^{-1}(D) = 1$, then $PZ_1^{-1}\mathbf{f}_1^{-1}(D^c) = 0$ for any $D \in \mathcal{B}(\mathbb{H})$. Hence, $\mathbb{H}$ is an atom of the measure $PZ_1^{-1}\mathbf{f}_1^{-1}$. Therefore, there exists a singleton $\{\mathbf{c}_1\} \in \mathcal{B}(\mathbb{H})$ such that $PZ_1^{-1}\mathbf{f}_1^{-1}(\{\mathbf{c}_1\}) > 0$ by Lemma 10.17 in [2]. Since $PZ_1^{-1}(\mathbf{f}_1^{-1}(\{\mathbf{c}_1\}))$ must be 0 or 1, $PZ_1^{-1}(\mathbf{f}_1^{-1}(\{\mathbf{c}_1\})) = 1$. This completes the proof. □

Lemma 2. *Assume the conditions (B1) and (B3), and that p is bounded on its support,* $\inf_n n^{c_1} \prod_{j=1}^{d_x} h_j > 0$ *for some* $c_1 < (\alpha-2)/\alpha$ *and* $\inf_n n^{c_2} \min_{1\le j\le d_x} h_j > 0$ *for some* $c_2 \in \mathbb{R}$. *Then for* $\mathbf{S}_n(\mathbf{w}) := n^{-1} \odot \bigoplus_{i=1}^n \kappa_i(\mathbf{w}) \odot \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)$, *it holds that*

$$\sup_{\mathbf{w}} \|\mathbf{S}_n(\mathbf{w}) \ominus \mathrm{E}(\mathbf{S}_n(\mathbf{w}))\| = O_p\left(\left(n \prod_{j=1}^{d_x} h_j\right)^{-1/2} \cdot \sqrt{\log n}\right).$$

Proof. Take $\delta \in (0, (\alpha/2 - c_1\alpha/2 - 1)/\alpha)$. This choice is possible since $c_1 < (\alpha-2)/\alpha$. Define

$$\begin{aligned} \boldsymbol{\eta}_{ni}(\mathbf{w}) = {} & \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*) \odot I\Big(\|\boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)\| \le n^{1/2-\delta} \prod_{j=1}^{d_x} h_j^{1/2}\Big)\kappa_i(\mathbf{w}) \\ & \ominus \mathrm{E}\left(\boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*) \odot I\Big(\|\boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)\| \le n^{1/2-\delta} \prod_{j=1}^{d_x} h_j^{1/2}\Big)\kappa_i(\mathbf{w})\right). \end{aligned}$$

By techniques of kernel smoothing theory (e.g. Theorem 2 in [28]), we get that, for sufficiently large $\gamma > 0$,

$$\begin{aligned} & \sup_{\mathbf{w}} \left\| n^{-1} \odot \bigoplus_{i=1}^n (\kappa_i(\mathbf{w}) \odot \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)) \ominus \mathrm{E}\,(\kappa_i(\mathbf{w}) \odot \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)) \right\| \\ & \quad = \sup_{\mathbf{x} \in I^{d_x}(n^{-\gamma}), \mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j, \mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j} \left\| n^{-1} \odot \bigoplus_{i=1}^n \boldsymbol{\eta}_{ni}(\mathbf{w}) \right\| + o_p\Big(n^{-1/2} \prod_{j=1}^{d_x} h_j^{1/2}\Big) \end{aligned}$$

where $I^{d_x}(n^{-\gamma}) = \prod_{j=1}^{d_x} \{0, n^{-\gamma}, \cdots, [n^\gamma] \cdot n^{-\gamma}, 1\}$. Thus, it suffices to show that

$$\lim_{C\to\infty} \limsup_{n\to\infty} n^{d_x \cdot \gamma} \cdot P\left(\left\| n^{-1} \odot \bigoplus_{i=1}^n \boldsymbol{\eta}_{ni}(\mathbf{w}) \right\| > C \sqrt{\frac{\log n}{n \prod_{j=1}^{d_x} h_j}}\right) = 0 \tag{A.17}$$

for all $\mathbf{w} \in [0,1]^{d_x} \times \prod_{j=1}^{d_u} \mathcal{U}_j \times \prod_{j=1}^{d_v} \mathcal{V}_j$. We note that $\mathrm{E}(\boldsymbol{\eta}_{ni}(\mathbf{w})) = \mathbf{0}$ and

$$\|\boldsymbol{\eta}_{ni}(\mathbf{w})\| \le 2n^{1/2-\delta} \prod_{j=1}^{d_x} h_j^{-1/2} \left(\sup_{t\in[-1,1]} K(t) \right)^{d_x},$$
$$n^{-1} \sum_{i=1}^{n} \mathrm{E}(\|\boldsymbol{\eta}_{ni}(\mathbf{w})\|^2) \le c \cdot \Big(\prod_{j=1}^{d_x} h_j \Big)^{-1}$$

for some constant $c > 0$. By applying Corollary 2.2 in [6], we get that, for sufficiently large n,

$$\begin{aligned} P\left(\Big\| n^{-1} \odot \bigoplus_{i=1}^{n} \boldsymbol{\eta}_{ni}(\mathbf{w}) \Big\| > C \sqrt{\frac{\log n}{n \prod_{j=1}^{d_x} h_j}} \right) & \\ \le 2 \exp\left(- \frac{3C^2 n^{\delta} \log n}{6cn^{\delta} + 4C\sqrt{\log n} (\sup_{t\in[-1,1]} K(t))^{d_x}} \right) & \\ \le 2n^{-C^2/(4c)}. & \end{aligned}$$

This implies (A.17). □

For the next lemma, we define

$$\hat{\boldsymbol{\mu}}^A_{x,j}(x_j) = (n\hat{p}_{x,j}(x_j))^{-1} \odot \bigoplus_{i=1}^{n} K_{h_j}(x_j, X_{ij}) \odot \boldsymbol{\epsilon}_{i+},$$
$$\hat{\boldsymbol{\mu}}^A_{u,j}(u_j) = (n\hat{p}_{u,j}(u_j))^{-1} \odot \bigoplus_{i=1}^{n} L_{\lambda_j}(u_j, U_{ij}) \odot \boldsymbol{\epsilon}_{i+},$$
$$\hat{\boldsymbol{\mu}}^A_{v,j}(v_j) = (n\hat{p}_{v,j}(v_j))^{-1} \odot \bigoplus_{i=1}^{n} W_{s_j}(v_j, V_{ij}) \odot \boldsymbol{\epsilon}_{i+},$$

where $\boldsymbol{\epsilon}_{i+}$ is the ith observation of $\boldsymbol{\epsilon}_+$ defined at the beginning of Section 4.3. Recall the definition of C_{j,x_j} given at (4.10). The following lemma is used to prove Theorem 5.

Lemma 3. *Fix $\mathbf{x} \in (0,1)^{d_x}, \mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j$ and $\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j$. Assume that the conditions on h_j, λ_j and s_j in (D3) hold, that K is bounded, that $\mathrm{E}(\|\boldsymbol{\epsilon}_+\|^\alpha) < \infty$ for some $\alpha > 2$ and that, for all u_k, v_k and $1 \le j \le d_x$,* (a) *$\mathrm{E}(\|\boldsymbol{\epsilon}_+\|^\alpha | X_j = \cdot)$, $\mathrm{E}(\langle \boldsymbol{\epsilon}_+, \mathbf{e}_l \rangle \cdot \langle \boldsymbol{\epsilon}_+, \mathbf{e}_{l'} \rangle \,|\, X_j = \cdot, U_k = u_k)$ and $\mathrm{E}(\langle \boldsymbol{\epsilon}_+, \mathbf{e}_l \rangle \cdot \langle \boldsymbol{\epsilon}_+, \mathbf{e}_{l'} \rangle \,|\, X_j = \cdot, V_k = v_k)$ are bounded on a respective neighborhood of x_j, $\mathrm{E}(\langle \boldsymbol{\epsilon}_+, \mathbf{e}_l \rangle \cdot \langle \boldsymbol{\epsilon}_+, \mathbf{e}_{l'} \rangle \,|\, X_j = \cdot, X_k = \cdot)$ and $p_{xx,jk}$ are bounded on a respective neighborhood of (x_j, x_k), and $\mathrm{E}(\langle \boldsymbol{\epsilon}_+, \mathbf{e}_l \rangle \cdot \langle \boldsymbol{\epsilon}_+, \mathbf{e}_{l'} \rangle \,|\, X_j = \cdot)$ for all l and l', are continuous on a common neighborhood of x_j;* (b) *$p_{x,j}$ is continuous on a neighborhood of x_j and $p_{x,j}(x_j) > 0$. Then,*

$$n^{2/5} \odot (\hat{\boldsymbol{\mu}}^A_{x,1}(x_1), \ldots, \hat{\boldsymbol{\mu}}^A_{x,d_x}(x_{d_x}), \hat{\boldsymbol{\mu}}^A_{u,1}(u_1), \ldots, \hat{\boldsymbol{\mu}}^A_{v,d_v}(v_{d_v}))$$

$$\xrightarrow{d} (\mathbf{G}(\mathbf{0}, C_{1,x_1}), \ldots, \mathbf{G}(\mathbf{0}, C_{d_x,x_{d_x}}), \mathbf{0}, \ldots, \mathbf{0}).$$

Moreover, $\mathbf{G}(\mathbf{0}, C_{1,x_1}), \ldots, \mathbf{G}(\mathbf{0}, C_{d_x,x_{d_x}})$ *are independent.*

Proof. We first note that any fixed $\mathbf{x} \in (0,1)^{d_x}$ lies in $\prod_{j=1}^{d_x}[2h_j, 1-2h_j]$ for sufficiently large n, so that we may assume $\mathbf{x} \in \prod_{j=1}^{d_x}[2h_j, 1-2h_j]$. We denote $d_x + d_u + d_v$ by d and let $\mathbb{H}^d$ be the space of tuples $(\mathbf{h}_j : 1 \le j \le d)$ with $\mathbf{h}_j \in \mathbb{H}$. Let $\|\cdot\|_{\mathbb{H}^d}$ and $\langle\cdot,\cdot\rangle_{\mathbb{H}^d}$ denote the norm and inner product of $\mathbb{H}^d$, respectively, defined in the standard way. Let $\mathbf{e}_{jl} = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{e}_l, \mathbf{0}, \ldots, \mathbf{0}) \in \mathbb{H}^d$, where $\mathbf{e}_l$ is placed at the jth entry. Then, $(\mathbf{e}_{jl} : 1 \le j \le d,\, l \ge 1)$ forms an orthonormal basis of $\mathbb{H}^d$. Define

$$\boldsymbol{\eta}_{ni}(\mathbf{w}) = \Bigg(\frac{n^{2/5} K_{h_1}(x_1 - X_{i1})}{n p_1(x_1)} \odot \boldsymbol{\epsilon}_{i+}, \cdots, \frac{n^{2/5} K_{h_d}(x_d - X_{id})}{n p_d(x_d)} \odot \boldsymbol{\epsilon}_{i+},$$
$$\frac{n^{2/5} L_{\lambda_1}(u_1, U_{i1})}{n p_{u,1}(u_1)} \odot \boldsymbol{\epsilon}_{i+}, \cdots, \frac{n^{2/5} W_{s_{d_v}}(v_{d_v}, V_{id_v})}{n p_{v,d_v}(v_{d_v})} \odot \boldsymbol{\epsilon}_{i+} \Bigg) \in \mathbb{H}^d.$$

Note that $\mathrm{E}(\langle \boldsymbol{\eta}_{ni}(\mathbf{w}), \mathbf{e}_{jl} \rangle_{\mathbb{H}^d}) = 0$ and $\mathrm{E}(\|\boldsymbol{\eta}_{ni}(\mathbf{w})\|^2_{\mathbb{H}^d}) < \infty$. For $\mathbf{S}_n(\mathbf{w}) = \bigoplus_{i=1}^n \boldsymbol{\eta}_{ni}(\mathbf{w})$ and $1 \le j, k \le d_x$, it holds that

$$\mathrm{E}\left(\langle \mathbf{S}_n(\mathbf{w}), \mathbf{e}_{jl} \rangle_{\mathbb{H}^d} \cdot \langle \mathbf{S}_n(\mathbf{w}), \mathbf{e}_{km} \rangle_{\mathbb{H}^d} \right) \to a_{j,lm}(x_j) 1(j = k),$$

where

$$a_{j,lm}(x_j) = \alpha_j^{-1} p_{x,j}(x_j)^{-1} \int_{-1}^{1} K^2(t) dt \cdot \mathrm{E}(\langle \boldsymbol{\epsilon}_+, \mathbf{e}_l \rangle \cdot \langle \boldsymbol{\epsilon}_+, \mathbf{e}_m \rangle \,|\, X_j = x_j)$$

with α_j being the constants in the condition (D3). Also, if $d_x + 1 \le j \le d$ or $d_x + 1 \le k \le d$, then

$$\mathrm{E}\left(\langle \mathbf{S}_n(\mathbf{w}), \mathbf{e}_{jl} \rangle_{\mathbb{H}^d} \cdot \langle \mathbf{S}_n(\mathbf{w}), \mathbf{e}_{km} \rangle_{\mathbb{H}^d} \right) \to 0.$$

We also get

$$\lim_{n\to\infty} \sum_{j=1}^{d} \sum_{l} \mathrm{E}\left(\langle \mathbf{S}_n(\mathbf{w}), \mathbf{e}_{jl} \rangle_{\mathbb{H}^d} \cdot \langle \mathbf{S}_n(\mathbf{w}), \mathbf{e}_{jl} \rangle_{\mathbb{H}^d} \right)$$
$$= \sum_{j=1}^{d_x} \alpha_j^{-1} \frac{1}{p_{x,j}(x_j)} \int_{-1}^{1} K^2(t) dt \cdot \mathrm{E}(\, \|\boldsymbol{\epsilon}_+\|^2 \,|\, X_j = x_j).$$

In addition, for $0 < \delta \le \alpha - 2$,

$$\sum_{i=1}^{n} \mathrm{E}\left(\left\| \boldsymbol{\eta}_{ni}(\mathbf{w}) \right\|^{2+\delta}_{\mathbb{H}^d} \right)$$
$$\le n^{-(1+3\delta)/5} d^{1+\delta/2} \mathrm{E}\Bigg(\|\boldsymbol{\epsilon}_+\|^{2+\delta} \Bigg\{ \sum_{j=1}^{d_x} \left(\frac{1}{p_{x,j}(x_j) h_j} \right)^{2+\delta} K^{2+\delta}\left(\frac{x_j - X_j}{h_j} \right)$$

$$+\sum_{j=d_x+1}^{d_x+d_u}\left(\frac{1}{p_{u,j}(u_j)}\right)^{2+\delta}+\sum_{j=d_x+d_u+1}^{d}\left(\frac{1}{p_{v,j}(v_j)}\right)^{2+\delta}\Bigg\}\Bigg)$$

$= o(1)$.

Therefore, by applying Theorem 1.1 in [17] for infinite-dimensional $\mathbb{H}$ and Proposition S.2 in [15] for finite-dimensional $\mathbb{H}$, we obtain

$$\mathbf{S}_n(\mathbf{w}) \xrightarrow{d} \mathbf{G}(\mathbf{0}, C_{\mathbf{w}}),$$

where $C_{\mathbf{w}} : \mathbb{H}^d \to \mathbb{H}^d$ is a covariance operator such that

$$\langle C_{\mathbf{w}}(\mathbf{h}), \mathbf{e}_{jl}\rangle_{\mathbb{H}^d} = \begin{cases} \sum_m \langle \mathbf{h}_j, \mathbf{e}_m\rangle \cdot a_{j,lm}(x_j) & 1 \le j \le d_x,\ l \ge 1 \\ 0 & d_x + 1 \le j \le d,\ l \ge 1 \end{cases} \tag{A.18}$$

for all $\mathbf{h} = (\mathbf{h}_1, \ldots, \mathbf{h}_d) \in \mathbb{H}^d$. Since $\hat{p}_{t,j}(t_j)^{-1} - p_{t,j}(t_j)^{-1} = o_p(1)$ for all $t = x, u, v$ and j, we get

$$\begin{aligned} &\big(n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{x,1}(x_1), \ldots, n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{x,d_x}(x_{d_x}), \\ &\qquad n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{u,1}(u_1), \ldots, n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{v,d_v}(v_{d_v})\big) \\ &\quad \xrightarrow{d} \mathbf{G}(\mathbf{0}, C_{\mathbf{w}}). \end{aligned} \tag{A.19}$$

Let P_j denote the projection operator that maps $(\mathbf{h}_1, \ldots, \mathbf{h}_d) \in \mathbb{H}^d$ to $\mathbf{h}_j$. Then, its adjoint $P_j^* : \mathbb{H} \to \mathbb{H}^d$ is given by $P_j^*(\mathbf{g}) = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{g}, \mathbf{0}, \cdots, \mathbf{0})$ where $\mathbf{g}$ is placed at the jth entry. We note that the conclusions of Propositions 4.9–4.10 in [40] also hold for $\mathbb{H}$-valued Gaussian random elements. This implies $P_j(\mathbf{G}(\mathbf{0}, C_{\mathbf{w}})) = \mathbf{G}(\mathbf{0}, P_j \circ C_{\mathbf{w}} \circ P_j^*)$. Now, for $\mathbf{g} \in \mathbb{H}$ and $1 \le j \le d_x$,

$$\begin{aligned} \langle P_j \circ C_{\mathbf{w}} \circ P_j^*(\mathbf{g}), \mathbf{e}_l\rangle &= \langle C_{\mathbf{w}}(\mathbf{0}, \ldots, \mathbf{0}, \mathbf{g}, \mathbf{0}, \ldots, \mathbf{0}), P_j^*(\mathbf{e}_l)\rangle_{\mathbb{H}^d} \\ &= \langle C_{\mathbf{w}}(\mathbf{0}, \ldots, \mathbf{0}, \mathbf{g}, \mathbf{0}, \ldots, \mathbf{0}), \mathbf{e}_{jl}\rangle_{\mathbb{H}^d} \\ &= \sum_m \langle \mathbf{g}, \mathbf{e}_m\rangle \cdot a_{j,lm}(x_j), \end{aligned}$$

where the last equality follows from (A.18). This proves $P_j \circ C_{\mathbf{w}} \circ P_j^* = C_{j,x_j}$ for $1 \le j \le d_x$, which coupled with (A.19) implies

$$\begin{aligned} &P_j\big(n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{x,1}(x_1), \ldots, n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{x,d_x}(x_{d_x}), \\ &\qquad n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{u,1}(u_1), \ldots, n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{v,d_v}(v_{d_v})\big) \\ &\quad \xrightarrow{d} P_j(\mathbf{G}(\mathbf{0}, C_{\mathbf{w}})) = \mathbf{G}(\mathbf{0}, C_{j,x_j}). \end{aligned}$$

On the other hand, for $\mathbf{g} \in \mathbb{H}$ and $d_x + 1 \le j \le d$,

$$\langle P_j \circ C_{\mathbf{w}} \circ P_j^*(\mathbf{g}), \mathbf{e}_l\rangle = 0.$$

This proves $P_j(\mathbf{G}(\mathbf{0}, C_{\mathbf{w}})) = \mathbf{0}$ for $d_x + 1 \le j \le d$. The independence of $\mathbf{G}(\mathbf{0}, C_{j,x_j})$ for different $1 \le j \le d_x$ follows from Theorem 4.2 in [15]. □

A.6. Proofs for Section 3

A.6.1. Proof of Theorem 1

First, we show that $\hat{F}$ is a strictly convex and continuous functional satisfying $\hat{F}(\mathbf{f}) \to \infty$ as $\|\mathbf{f}\|_{2,n} \to \infty$. This together with Lemma 4 in [5] and Proposition 4 implies that there exists a minimizer of $\hat{F}$ in $\mathcal{S}^{\mathbb{H}}(\hat{p})$. Note that the strict convexity is trivial. For the continuity, we note that

$$|\hat{F}(\mathbf{f}) - \hat{F}(\mathbf{f}_k)| \leq \|\mathbf{f} \ominus \mathbf{f}_k\|_{2,n}(\|\mathbf{f} \ominus \mathbf{f}_k\|_{2,n} + 2\|\mathbf{f}\|_{2,n} + 2\sqrt{d} \cdot \sqrt{\hat{c}} \cdot \hat{M}),$$

where $\hat{c}$ and $\hat{M}$ are defined in (A.21). For the divergence, we note that

$$\hat{F}(\mathbf{f}) \geq \|\mathbf{f}\|_{2,n}^2 \Big(1 - \frac{2\sqrt{d} \cdot \sqrt{\hat{c}} \cdot \hat{M}}{\|\mathbf{f}\|_{2,n}}\Big).$$

Next, we prove that $\hat{F}$ is Gâteaux differentiable. For $\mathbf{f} \in \mathcal{S}^{\mathbb{H}}(\hat{p})$, define $D\hat{F}(\mathbf{f}) : \mathcal{S}^{\mathbb{H}}(\hat{p}) \to \mathbb{R}$ by

$$\begin{aligned} D\hat{F}(\mathbf{f})(\mathbf{g}) &= \lim_{\delta \to 0} \frac{\hat{F}(\mathbf{f} \oplus \delta\mathbf{g}) - \hat{F}(\mathbf{f})}{\delta} \\ &= -2 \int_{\mathcal{Z}} \big\langle \mathbf{g}(\mathbf{z}), \mathbf{f}(\mathbf{z}) \ominus \hat{\boldsymbol{\mu}}(\mathbf{z}) \big\rangle \hat{p}(\mathbf{z}) d\nu(\mathbf{z}). \end{aligned} \tag{A.20}$$

It is clear that $D\hat{F}(\mathbf{f})$ is a linear operator. Also, $D\hat{F}(\mathbf{f})$ is a bounded operator under the conditions (S1) and (S2) since

$$|D\hat{F}(\mathbf{f})(\mathbf{g})| \leq 2 \left(\|\mathbf{f}\|_{2,n} + \sqrt{d} \cdot \sqrt{\hat{c}} \cdot \hat{M}\right) \|\mathbf{g}\|_{2,n}, \tag{A.21}$$

where $\hat{c}$ is the constant in Proposition 3 and $\hat{M} = \max_{1 \leq j \leq d} \|\hat{\boldsymbol{\mu}}_j\|_{2,n}^2 < \infty$. The inequality (A.21) may be proved by applying the Hölder inequality and considering a decomposition $\bigoplus_{j=1}^d \mathbf{g}_j$ of $\mathbf{g}$ with $\mathbf{g}_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$ such that $\sum_{j=1}^d \|\mathbf{g}_j\|_{2,n}^2 \leq \hat{c}\|\mathbf{g}\|_{2,n}^2$ whose existence is guaranteed by Proposition 3. Hence, $D\hat{F}(\mathbf{f})$ is the Gâteaux derivative of $\hat{F}$ at $\mathbf{f}$. Thus, $\hat{F}$ is Gâteaux differentiable.

Now, $\hat{\mathbf{f}} \in \mathcal{S}^{\mathbb{H}}(\hat{p})$ being a minimizer of $\hat{F}$ is equivalent to $D\hat{F}(\hat{\mathbf{f}})(\mathbf{g}) = 0$ for all $\mathbf{g} \in \mathcal{S}^{\mathbb{H}}(\hat{p})$ by Theorem 5.3.19 in [4]. With the specification of $\mathbf{g} \in \mathcal{S}^{\mathbb{H}}(\hat{p})$ in $D\hat{F}(\hat{\mathbf{f}})(\mathbf{g}) = 0$ to $\mathbf{g}_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$ for each $1 \leq j \leq d$, the equation implies that

$$\int_{\mathcal{Z}_{-j}} (\hat{\mathbf{f}}(\mathbf{z}) \ominus \hat{\boldsymbol{\mu}}(\mathbf{z})) \odot \hat{p}(\mathbf{z}) d\nu_{-j}(\mathbf{z}_{-j}) = \mathbf{0} \tag{A.22}$$

a.e. with respect to ν_j, for all $1 \leq j \leq d$.

Let $\hat{\mathbf{f}} = \hat{\mathbf{f}}_0 \oplus \bigoplus_{j=1}^d \hat{\mathbf{f}}_j$ be a decomposition of $\hat{\mathbf{f}}$ with $\hat{\mathbf{f}}_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$ such that $\int_{\mathcal{Z}_j} \hat{\mathbf{f}}_j(z_j) \odot \hat{p}_j(z_j) d\nu_j(z_j) = \mathbf{0}$ for all $1 \leq j \leq d$. Plugging the decomposition into the left hand side of (A.22) and using (2.9), we see that $\hat{\mathbf{f}}_0 = \hat{\mathbf{m}}_0$

and $(\hat{\mathbf{f}}_j : 1 \le j \le d)$ satisfies

$$\hat{\mathbf{f}}_j(z_j) = \hat{\boldsymbol{\mu}}_j(z_j) \ominus \hat{\mathbf{m}}_0 \ominus \bigoplus_{k \neq j} \int_{\mathcal{Z}_k} \hat{\mathbf{f}}_k(z_k) \odot \frac{\hat{p}_{jk}(z_j, z_k)}{\hat{p}_j(z_j)} d\nu_k(z_k) \tag{A.23}$$

a.e. with respect to ν_j, for all $1 \le j \le d$. We define the right hand side of (A.23) by $\hat{\mathbf{m}}_j(z_j)$ for all $z_j \in \mathcal{Z}_j$. Then, $(\hat{\mathbf{m}}_j : 1 \le j \le d) \in \prod_{j=1}^d L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$ and it satisfies (2.15) and (2.16).

For the uniqueness of $\hat{\boldsymbol{\mu}}_+$, suppose that there exists another solution $(\hat{\mathbf{m}}_j^\star : 1 \le j \le d)$ in $\prod_{j=1}^d L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$ of (2.15). Recall the definition of $\hat{T}$ given at (A.9) and define

$$\begin{aligned} \tilde{\mathbf{m}} &= \hat{\boldsymbol{\mu}}_d \ominus \hat{\mathbf{m}}_0 \oplus (I - \hat{\pi}_d)(\hat{\boldsymbol{\mu}}_{d-1}) \oplus \cdots \oplus (I - \hat{\pi}_d) \circ \cdots \circ (I - \hat{\pi}_2)(\hat{\boldsymbol{\mu}}_1) \\ &\quad \in \mathcal{S}^{\mathbb{H}}(\hat{p}). \end{aligned} \tag{A.24}$$

Since $\hat{\pi}_j(\hat{\mathbf{m}}_\oplus) = \hat{\boldsymbol{\mu}}_j \ominus \hat{\mathbf{m}}_0$ for all $1 \le j \le d$ from (2.15), we get

$$\hat{\mathbf{m}}_\oplus = (I - \hat{\pi}_j)(\hat{\mathbf{m}}_\oplus) \oplus (\hat{\boldsymbol{\mu}}_j \ominus \hat{\mathbf{m}}_0), \quad 1 \le j \le d. \tag{A.25}$$

Applying (A.25) from $j = d$ to $j = 1$ successively gives

$$\begin{aligned} \hat{\mathbf{m}}_\oplus &= (I - \hat{\pi}_d)(\hat{\mathbf{m}}_\oplus) \oplus (\hat{\boldsymbol{\mu}}_d \ominus \hat{\mathbf{m}}_0) \\ &= (I - \hat{\pi}_d)\left((I - \hat{\pi}_{d-1})(\hat{\mathbf{m}}_\oplus) \oplus (\hat{\boldsymbol{\mu}}_{d-1} \ominus \hat{\mathbf{m}}_0)\right) \oplus (\hat{\boldsymbol{\mu}}_d \ominus \hat{\mathbf{m}}_0) \\ &= (I - \hat{\pi}_d) \circ (I - \hat{\pi}_{d-1})(\hat{\mathbf{m}}_\oplus) \oplus (I - \hat{\pi}_d)(\hat{\boldsymbol{\mu}}_{d-1}) \oplus (\hat{\boldsymbol{\mu}}_d \ominus \hat{\mathbf{m}}_0) \\ &\cdots \\ &= \hat{T}(\hat{\mathbf{m}}_\oplus) \oplus \tilde{\mathbf{m}}. \end{aligned} \tag{A.26}$$

Similarly, for $\hat{\mathbf{m}}_\oplus^\star = \bigoplus_{j=1}^d \hat{\mathbf{m}}_j^\star$, we have

$$\hat{\mathbf{m}}_\oplus^\star = \hat{T}(\hat{\mathbf{m}}_\oplus^\star) \oplus \tilde{\mathbf{m}}.$$

Since $\hat{\mathbf{m}}_\oplus \ominus \hat{\mathbf{m}}_\oplus^\star = \hat{T}(\hat{\mathbf{m}}_\oplus \ominus \hat{\mathbf{m}}_\oplus^\star)$ and $\|\hat{T}\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))} < 1$ by Proposition 4, we conclude that $\hat{\mathbf{m}}_\oplus = \hat{\mathbf{m}}_\oplus^\star$ a.e. with respect to $\hat{P}\mathbf{Z}^{-1}$. This proves the first part of the theorem.

For the proof of the second part, suppose that $\bigoplus_{j=1}^d \hat{\mathbf{g}}_j(z_j) = \mathbf{0}$ a.e. with respect to $\hat{P}\mathbf{Z}^{-1}$ with $\hat{\mathbf{g}}_j$ satisfying (2.16). Since $\hat{p} > 0$ on $\mathcal{Z}$ by the assumption, this implies $\bigoplus_{j=1}^d \hat{\mathbf{g}}_j(z_j) = \mathbf{0}$ a.e. with respect to ν, so that, for any map $\boldsymbol{\eta}_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$, we get

$$\left\langle \bigoplus_{k=1}^d \hat{\mathbf{g}}_k(z_k) \odot \hat{p}_{\mathbf{Z}_{-j}}(\mathbf{z}_{-j}),\, \boldsymbol{\eta}_j(z_j) \right\rangle = 0 \text{ a.e. with respect to } \nu. \tag{A.27}$$

Because of the marginalization property $\int_{\mathcal{Z}_{-jk}} \hat{p}_{\mathbf{Z}_{-j}}(\mathbf{z}_{-j}) d\nu_{-jk}(\mathbf{z}_{-jk}) = \hat{p}_k(z_k)$ and the constraints (2.16), the equation (A.27) implies that

$$\begin{aligned} 0 &= \sum_{k=1}^{d} \int_{\mathcal{Z}} \langle \hat{\mathbf{g}}_k(z_k) \odot \hat{p}_{\mathbf{Z}_{-j}}(\mathbf{z}_{-j}),\, \boldsymbol{\eta}_j(z_j) \rangle d\nu(\mathbf{z}) \\ &= \sum_{k \neq j}^{d} \int_{\mathcal{Z}_j} \left\langle \int_{\mathcal{Z}_k} \hat{\mathbf{g}}_k(z_k) \odot \hat{p}_k(z_k) d\nu_k(z_k),\, \boldsymbol{\eta}_j(z_j) \right\rangle d\nu_j(z_j) \\ &\quad + \int_{\mathcal{Z}_j} \langle \hat{\mathbf{g}}_j(z_j),\, \boldsymbol{\eta}_j(z_j) \rangle d\nu_j(z_j) \\ &= \int_{\mathcal{Z}_j} \langle \hat{\mathbf{g}}_j(z_j),\, \boldsymbol{\eta}_j(z_j) \rangle d\nu_j(z_j) \end{aligned}$$

for all $\boldsymbol{\eta}_j \in L^2((\mathcal{Z}_j, \mathscr{A}_j, \hat{P}Z_j^{-1}), \mathbb{H})$. This implies $\hat{\mathbf{g}}_j(z_j) = \mathbf{0}$ a.e. with respect to ν_j. This proves the second part of the theorem.

A.6.2. Proof of Theorem 2

Let $\hat{\mathbf{m}}_{\oplus} = \bigoplus_{j=1}^{d} \hat{\mathbf{m}}_j$ and $\hat{\mathbf{m}}_{\oplus}^{[r]} = \bigoplus_{j=1}^{d} \hat{\mathbf{m}}_j^{[r]}$ for $r \geq 0$. From (A.26), we have $\hat{\mathbf{m}}_{\oplus} = \hat{T}(\hat{\mathbf{m}}_{\oplus}) \oplus \tilde{\mathbf{m}}$, where $\tilde{\mathbf{m}}$ is defined at (A.24). One may similarly prove that $\hat{\mathbf{m}}_{\oplus}^{[r]} = \hat{T}(\hat{\mathbf{m}}_{\oplus}^{[r-1]}) \oplus \tilde{\mathbf{m}}$ from the gB-SBF algorithm in Section 2.4. Since $\|\hat{T}\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))} < 1$ by Proposition 4, it holds that $\hat{\mathbf{m}}_{\oplus}^{[\infty]} := \bigoplus_{k=0}^{\infty} \hat{T}^k(\tilde{\mathbf{m}})$ exists in $\mathcal{S}^{\mathbb{H}}(\hat{p})$, $\hat{\mathbf{m}}_{\oplus}^{[\infty]} = \hat{T}(\hat{\mathbf{m}}_{\oplus}^{[\infty]}) \oplus \tilde{\mathbf{m}}$ a.e. with respect to $\hat{P}\mathbf{Z}^{-1}$ and thus $\hat{\mathbf{m}}_{\oplus}^{[\infty]} = \hat{\mathbf{m}}_{\oplus}$ a.e. with respect to $\hat{P}\mathbf{Z}^{-1}$. This entails

$$\|\hat{\mathbf{m}}_{\oplus}^{[r]} \ominus \hat{\mathbf{m}}_{\oplus}\|_{2,n} \leq \frac{\|\hat{T}\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))}^{r}}{1 - \|\hat{T}\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))}} \left(\|\hat{\mathbf{m}}_{\oplus}^{[0]}\|_{2,n} + \|\tilde{\mathbf{m}}\|_{2,n} \right). \tag{A.28}$$

The inequality (A.28) gives the theorem with the choices $\hat{c}^* = \left(\|\hat{\mathbf{m}}_{\oplus}^{[0]}\|_{2,n} + \|\tilde{\mathbf{m}}\|_{2,n}\right)^2 / (1 - \|\hat{T}\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))})^2$ and $\hat{\gamma} = \|\hat{T}\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))}^2$.

A.6.3. Proof of Theorem 3

Let $\mathcal{E}_1$ denote the event where (2.8) and the first property at (2.10) hold. For a given constant $C > 0$, let $\mathcal{E}_2(C)$ denote the event where

$$\max_{1 \leq j \leq d} \sup_{z_j \in \mathcal{Z}_j} \hat{p}_j(z_j)^{-1} \leq C, \quad \max_{1 \leq j \neq k \leq d} \sup_{z_j \in \mathcal{Z}_j, z_k \in \mathcal{Z}_k} \hat{p}_{jk}(z_j, z_k) \leq C,$$

$$\max_{1 \leq j \leq d} \sup_{z_j \in \mathcal{Z}_j} \|\hat{\boldsymbol{\mu}}_j(z_j)\| \leq C, \quad \max_{1 \leq j \leq d} \int_{\mathcal{Z}_j} \|\hat{\mathbf{m}}_j^{[0]}(z_j)\|^2 d\nu_j(z_j) < C.$$

Also, for a given $\delta > 0$, let $\mathcal{E}_3(\delta)$ denote the event where

$$\begin{aligned} \max_{1\le j\le d} \int_{\mathcal{Z}_j} \frac{(\hat{p}_j(z_j) - p_j(z_j))^2}{p_j(z_j)} d\nu_j(z_j) \le \delta, \\ \max_{1\le j\ne k\le d} \int_{\mathcal{Z}_j\times\mathcal{Z}_k} \left(\frac{\hat{p}_{jk}(z_j,z_k)}{\hat{p}_j(z_j)} - \frac{p_{jk}(z_j,z_k)}{p_j(z_j)} \right)^2 \\ \times \frac{p_j(z_j)}{p_k(z_k)} d\nu_j \otimes \nu_k(z_j,z_k) \le \delta^2. \end{aligned} \tag{A.29}$$

Put $\mathcal{E}(C,\delta) = \mathcal{E}_1 \cap \mathcal{E}_2(C) \cap \mathcal{E}_3(\delta)$. By the assumptions of the theorem, there exists a constant $C > 0$ such that $P(\mathcal{E}(C,\delta)) \to 1$ for any $\delta > 0$. Henceforth, suppose that $\mathcal{E}(C,\delta)$ occurs for such C and δ.

Define the operators $\pi_j : L^2((\mathcal{Z},\mathscr{A},P\mathbf{Z}^{-1}),\mathbb{H}) \to L^2((\mathcal{Z}_j,\mathscr{A}_j,PZ_j^{-1}),\mathbb{H})$ in the same way as $\hat{\pi}_j$ with $\hat{p}$ and $\hat{p}_j$ being replaced by p and p_j, respectively. Also define the operator $T : \mathcal{S}^{\mathbb{H}}(p) \to \mathcal{S}^{\mathbb{H}}(p)$ in the same way as $\hat{T}$ with $\hat{\pi}_j$ being replaced by the respective π_j. Here,

$$\begin{aligned} \mathcal{S}^{\mathbb{H}}(p) := \left\{ \bigoplus_{j=1}^d \mathbf{f}_j : \mathbf{f}_j \in L^2((\mathcal{Z}_j,\mathscr{A}_j,PZ_j^{-1}),\mathbb{H}),\ 1\le j\le d \right\} \\ \subset L^2((\mathcal{Z},\mathscr{A},P\mathbf{Z}^{-1}),\mathbb{H}). \end{aligned}$$

Finally, define the norm $\|\cdot\|_2$ on $L^2((\mathcal{Z},\mathscr{A},P\mathbf{Z}^{-1}),\mathbb{H})$ by

$$\|\mathbf{f}\|_2^2 = \int_{\mathcal{Z}} \|\mathbf{f}(\mathbf{z})\|^2 dP\mathbf{Z}^{-1}(\mathbf{z}) = \int_{\mathcal{Z}} \|\mathbf{f}(\mathbf{z})\|^2 p(\mathbf{z}) d\nu(\mathbf{z}),$$

and the operator norm $\|\cdot\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(p))}$ in the same way as $\|\cdot\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(\hat{p}))}$ with $\hat{p}$ and $\|\cdot\|_{2,n}$ being replaced by p and $\|\cdot\|_2$, respectively. Then, similarly as in the proof of Proposition 4, it holds that $\mathcal{S}^{\mathbb{H}}(p)$ is a closed subspace of $L^2((\mathcal{Z},\mathscr{A},P\mathbf{Z}^{-1}),\mathbb{H})$ and $\|T\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(p))} < 1$, under the condition (P). Also, similarly as in the proof of Proposition 3, there exists a constant $c > 0$ such that, for any $\mathbf{f} \in \mathcal{S}^{\mathbb{H}}(p)$, there exist a decomposition $\bigoplus_{j=1}^d \mathbf{f}_j$ of $\mathbf{f}$ with $\mathbf{f}_j \in L^2((\mathcal{Z}_j,\mathscr{A}_j,PZ_j^{-1}),\mathbb{H})$ satisfying $\max\{\|\mathbf{f}_1\|_2,\dots,\|\mathbf{f}_d\|_2\} \le c\|\mathbf{f}\|_2$, under the condition (P). For such decomposition of $\mathbf{f} \in \mathcal{S}^{\mathbb{H}}(p)$, we get

$$\begin{aligned} &\|(\hat{\pi}_j - \pi_j)(\mathbf{f})\|_2 \\ &\le \sum_{k\ne j}^d \|\mathbf{f}_k\|_2 \left(\int_{\mathcal{Z}_j\times\mathcal{Z}_k} \left(\frac{\hat{p}_{jk}(z_j,z_k)}{\hat{p}_j(z_j)} - \frac{p_{jk}(z_j,z_k)}{p_j(z_j)} \right)^2 \frac{p_j(z_j)}{p_k(z_k)} d\nu_j \otimes \nu_k(z_j,z_k) \right)^{1/2} \\ &\le c(d-1)\delta \cdot \|\mathbf{f}\|_2. \end{aligned}$$

This implies $\|\hat{\pi}_j - \pi_j\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(p))} \le c(d-1)\delta$ and thus $\|\hat{T} - T\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(p))} \le d \cdot 2^{d-1} \cdot c(d-1)\delta$. We choose $0 < \delta \le (1 - \|T\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(p))})/(d \cdot 2^d \cdot c(d-1))$. Then,

$$\|\hat{T}\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(p))} \le (1 + \|T\|_{\mathcal{L}(\mathcal{S}^{\mathbb{H}}(p))})/2 =: \tau < 1.$$

As in the derivation of (A.28) we may prove that there exists an absolute constant $c_0 > 0$ such that

$$\|\hat{\mathbf{m}}_{\oplus}^{[r]} \ominus \hat{\mathbf{m}}_{\oplus}\|_2 \leq c_0 \cdot \tau^r \quad \text{for all } r \geq 0. \tag{A.30}$$

Now, let $\bigoplus_{j=1}^{d} \hat{\mathbf{f}}_j^{[r]}$ be a decomposition of $\hat{\mathbf{m}}_{\oplus}^{[r]} \ominus \hat{\mathbf{m}}_{\oplus}$ satisfying

$$\max\{\|\hat{\mathbf{f}}_1^{[r]}\|_2, \ldots, \|\hat{\mathbf{f}}_d^{[r]}\|_2\} \leq c\,\|\hat{\mathbf{m}}_{\oplus}^{[r]} \ominus \hat{\mathbf{m}}_{\oplus}\|_2.$$

Put $\hat{\mathbf{c}}_j^{[r]} = \int_{\mathcal{Z}_j} \hat{\mathbf{f}}_j^{[r]}(z_j) \odot \hat{p}_j(z_j) d\nu_j(z_j)$. Then,

$$\begin{aligned}
\|\hat{\mathbf{f}}_j^{[r]}\|_2^2 &\geq \|\hat{\mathbf{f}}_j^{[r]} \ominus \hat{\mathbf{c}}_j^{[r]}\|_2^2 + \|\hat{\mathbf{c}}_j^{[r]}\|^2 - 2\,\|\hat{\mathbf{c}}_j^{[r]}\| \cdot \|\hat{\mathbf{f}}_j^{[r]} \ominus \hat{\mathbf{c}}_j^{[r]}\|_2 \cdot \delta^{1/2} \\
&\geq (1-\delta)\|\hat{\mathbf{f}}_j^{[r]} \ominus \hat{\mathbf{c}}_j^{[r]}\|_2^2 \\
&= (1-\delta)\|\hat{\mathbf{m}}_j^{[r]} \ominus \hat{\mathbf{m}}_j\|_2^2.
\end{aligned} \tag{A.31}$$

The first inequality in (A.31) follows from an application of Hölder's inequality and the first bound in (A.29). The equality in (A.31) holds due to Lemma 1 and the fact that both $\hat{\mathbf{f}}_j^{[r]} \ominus \hat{\mathbf{c}}_j^{[r]}$ and $\hat{\mathbf{m}}_j^{[r]} \ominus \hat{\mathbf{m}}_j$ satisfy the constraints (2.16). This with (A.30) gives the theorem with the choices $c^{**} = (c \cdot c_0)^2/(1-\delta)$ and $\gamma = \tau^2$.

A.7. Terminologies and proofs for Section 4

A.7.1. Terminologies for Section 4.4

Here, we give the definitions of $\boldsymbol{\Delta}_{t,j}$ for $t = x$, u and v that appear in the asymptotic distribution of $(\hat{\mathbf{m}}_{x,1}, \ldots, \hat{\mathbf{m}}_{x,d_x}; \hat{\mathbf{m}}_{u,1}, \ldots, \hat{\mathbf{m}}_{u,d_u}; \hat{\mathbf{m}}_{v,1}, \ldots, \hat{\mathbf{m}}_{v,d_v})$ in Section 4.4. Define

$$\begin{aligned}
\boldsymbol{\delta}_{x,j}(x_j) &= \left(\frac{dp_{x,j}(x_j)/dx_j}{p_{x,j}(x_j)} \cdot \int_{-1}^{1} t^2 K(t) dt\right) \odot D\mathbf{m}_{x,j}(x_j)(1), \\
\boldsymbol{\delta}_{xx,jk}(x_j, x_k) &= \left(\frac{\partial p_{xx,jk}(x_j, x_k)/\partial x_k}{p_{xx,jk}(x_j, x_k)} \cdot \int_{-1}^{1} t^2 K(t) dt\right) \odot D\mathbf{m}_{x,k}(x_k)(1), \\
\boldsymbol{\delta}_{ux,jk}(u_j, x_k) &= \left(\frac{\partial p_{ux,jk}(u_j, x_k)/\partial x_k}{p_{ux,jk}(u_j, x_k)} \cdot \int_{-1}^{1} t^2 K(t) dt\right) \odot D\mathbf{m}_{x,k}(x_k)(1), \\
\boldsymbol{\delta}_{vx,jk}(v_j, x_k) &= \left(\frac{\partial p_{vx,jk}(v_j, x_k)/\partial x_k}{p_{vx,jk}(v_j, x_k)} \cdot \int_{-1}^{1} t^2 K(t) dt\right) \odot D\mathbf{m}_{x,k}(x_k)(1),
\end{aligned}$$

where the second one arises only when $d_x \geq 2$. Recall the definition of δ_j^* in Section 4.3 as given by $\delta_j^* = \min\{\delta_j(v_j, v_j') : v_j, v_j' \in \mathcal{V}_j, v_j \neq v_j'\}$. For the constants α_j, β_j and γ_j in the condition (D3) and c_j denoting the cardinality

of $\mathcal{U}_j$, define

$$
\begin{aligned}
&\tilde{\boldsymbol{\Delta}}_{x,j}(x_j) \\
&= \alpha_j^2 \odot \boldsymbol{\delta}_{x,j}(x_j) \oplus \bigoplus_{k\neq j}^{d_x} \int_0^1 \boldsymbol{\delta}_{xx,jk}(x_j, x_k) \odot \left(\alpha_k^2 \cdot \frac{p_{xx,jk}(x_j, x_k)}{p_{x,j}(x_j)} \right) dx_k \\
&\quad \oplus \bigoplus_{k=1}^{d_u} \bigoplus_{u_k \in \mathcal{U}_k} \bigoplus_{u'_k \in \mathcal{U}_k : u'_k \neq u_k} (\mathbf{m}_{u,k}(u'_k) \ominus \mathbf{m}_{u,k}(u_k)) \odot \left(\frac{\beta_k}{c_k - 1} \cdot \frac{p_{xu,jk}(x_j, u'_k)}{p_{x,j}(x_j)} \right) \\
&\quad \oplus \bigoplus_{k=1}^{d_v} \bigoplus_{v_k \in \mathcal{V}_k} \bigoplus_{v'_k \in \mathcal{V}_k : \delta_k(v_k, v'_k) = \delta_k^*} (\mathbf{m}_{v,k}(v'_k) \ominus \mathbf{m}_{v,k}(v_k)) \odot \left(\gamma_k \cdot \frac{p_{xv,jk}(x_j, v'_k)}{p_{x,j}(x_j)} \right),
\end{aligned}
$$

$$
\begin{aligned}
&\tilde{\boldsymbol{\Delta}}_{u,j}(u_j) \\
&= \bigoplus_{k=1}^{d_x} \int_0^1 \boldsymbol{\delta}_{ux,jk}(u_j, x_k) \odot \left(\alpha_k^2 \cdot \frac{p_{ux,jk}(u_j, x_k)}{p_{u,j}(u_j)} \right) dx_k \\
&\quad \oplus \bigoplus_{u'_j \in \mathcal{U}_j : u'_j \neq u_j} (\mathbf{m}_{u,j}(u'_j) \ominus \mathbf{m}_{u,j}(u_j)) \odot \left(\frac{\beta_j}{c_j - 1} \cdot \frac{p_{u,j}(u'_j)}{p_{u,j}(u_j)} \right) \\
&\quad \oplus \bigoplus_{k\neq j}^{d_u} \bigoplus_{u_k \in \mathcal{U}_k} \bigoplus_{u'_k \in \mathcal{U}_k : u'_k \neq u_k} (\mathbf{m}_{u,k}(u'_k) \ominus \mathbf{m}_{u,k}(u_k)) \odot \left(\frac{\beta_k}{c_k - 1} \cdot \frac{p_{uu,jk}(u_j, u'_k)}{p_{u,j}(u_j)} \right) \\
&\quad \oplus \bigoplus_{k=1}^{d_v} \bigoplus_{v_k \in \mathcal{V}_k} \bigoplus_{v'_k \in \mathcal{V}_k : \delta_k(v_k, v'_k) = \delta_k^*} (\mathbf{m}_{v,k}(v'_k) \ominus \mathbf{m}_{v,k}(v_k)) \odot \left(\gamma_k \cdot \frac{p_{uv,jk}(u_j, v'_k)}{p_{u,j}(u_j)} \right),
\end{aligned}
$$

$$
\begin{aligned}
&\tilde{\boldsymbol{\Delta}}_{v,j}(v_j) \\
&= \bigoplus_{k=1}^{d_x} \int_0^1 \boldsymbol{\delta}_{vx,jk}(v_j, x_k) \odot \left(\alpha_k^2 \cdot \frac{p_{vx,jk}(v_j, x_k)}{p_{v,j}(v_j)} \right) dx_k \\
&\quad \oplus \bigoplus_{k=1}^{d_u} \bigoplus_{u_k \in \mathcal{U}_k} \bigoplus_{u'_k \in \mathcal{U}_k : u'_k \neq u_k} (\mathbf{m}_{u,k}(u'_k) \ominus \mathbf{m}_{u,k}(u_k)) \odot \left(\frac{\beta_k}{c_k - 1} \cdot \frac{p_{vu,jk}(v_j, u'_k)}{p_{v,j}(v_j)} \right) \\
&\quad \oplus \bigoplus_{v'_j \in \mathcal{V}_j : \delta_j(v_j, v'_j) = \delta_j^*} (\mathbf{m}_{v,j}(v'_j) \ominus \mathbf{m}_{v,j}(v_j)) \odot \left(\gamma_j \cdot \frac{p_{v,j}(v'_j)}{p_{v,j}(v_j)} \right) \\
&\quad \oplus \bigoplus_{k=1}^{d_v} \bigoplus_{v_k \in \mathcal{V}_k} \bigoplus_{v'_k \in \mathcal{V}_k : \delta_k(v_k, v'_k) = \delta_k^*} (\mathbf{m}_{v,k}(v'_k) \ominus \mathbf{m}_{v,k}(v_k)) \odot \left(\gamma_k \cdot \frac{p_{vv,jk}(v_j, v'_k)}{p_{v,j}(v_j)} \right).
\end{aligned}
$$

Let $\boldsymbol{\Delta}^{\text{tup}}_{x,+j}$ denote the $(j-1)$-tuple of maps obtained by taking the first $(j-1)$ maps from $\boldsymbol{\Delta}^{\text{tup}} \equiv (\boldsymbol{\Delta}_{x,1}, \ldots, \boldsymbol{\Delta}_{x,d_x}; \boldsymbol{\Delta}_{u,1}, \ldots, \boldsymbol{\Delta}_{u,d_u}; \boldsymbol{\Delta}_{v,1}, \ldots, \boldsymbol{\Delta}_{v,d_v})$,

and $\boldsymbol{\Delta}^{\mathrm{tup}}_{x,j+}$ the tuple consisting of those from $\boldsymbol{\Delta}_{x,j+1}$ to $\boldsymbol{\Delta}_{v,d_v}$. Similarly, let $\boldsymbol{\Delta}^{\mathrm{tup}}_{u,+j} = (\boldsymbol{\Delta}_{x,1}, \ldots, \boldsymbol{\Delta}_{u,j-1})$ denote the $(d_x + j - 1)$-tuple and let $\boldsymbol{\Delta}^{\mathrm{tup}}_{u,j+} = (\boldsymbol{\Delta}_{u,j+1}, \ldots, \boldsymbol{\Delta}_{v,d_v})$. Also, let $\boldsymbol{\Delta}^{\mathrm{tup}}_{v,+j} = (\boldsymbol{\Delta}_{x,1}, \ldots, \boldsymbol{\Delta}_{v,j-1})$ be the $(d_x + d_u + j - 1)$-tuple and $\boldsymbol{\Delta}^{\mathrm{tup}}_{v,j+} = (\boldsymbol{\Delta}_{v,j+1}, \ldots, \boldsymbol{\Delta}_{v,d_v})$. For $1 \le j \le d_x$, define $\boldsymbol{\mu}_{x,+j}(\cdot;\cdot)$ and $\boldsymbol{\mu}_{x,j+}(\cdot;\cdot)$ by

$$
\begin{aligned}
\boldsymbol{\mu}_{x,+j}(x_j; \boldsymbol{\Delta}^{\mathrm{tup}}_{x,+j}) &= \bigoplus_{k \le j-1} \int_0^1 \boldsymbol{\Delta}_{x,k}(x_k) \odot \frac{p_{xx,jk}(x_j, x_k)}{p_{x,j}(x_j)} dx_k, \\
\boldsymbol{\mu}_{x,j+}(x_j; \boldsymbol{\Delta}^{\mathrm{tup}}_{x,j+}) &= \Bigg(\bigoplus_{k \ge j+1} \int_0^1 \boldsymbol{\Delta}_{x,k}(x_k) \odot \frac{p_{xx,jk}(x_j, x_k)}{p_{x,j}(x_j)} dx_k \Bigg) \\
&\quad \oplus \Bigg(\bigoplus_{k=1}^{d_u} \bigoplus_{u_k \in \mathcal{U}_k} \boldsymbol{\Delta}_{u,k}(u_k) \odot \frac{p_{xu,jk}(x_j, u_k)}{p_{x,j}(x_j)} \Bigg) \\
&\quad \oplus \Bigg(\bigoplus_{k=1}^{d_v} \bigoplus_{v_k \in \mathcal{V}_k} \boldsymbol{\Delta}_{v,k}(v_k) \odot \frac{p_{xv,jk}(x_j, v_k)}{p_{x,j}(x_j)} \Bigg).
\end{aligned}
$$

Likewise, define $\boldsymbol{\mu}_{t,+j}(\cdot;\cdot)$ and $\boldsymbol{\mu}_{t,j+}(\cdot;\cdot)$ for $t = u$ and v. For example,

$$
\begin{aligned}
\boldsymbol{\mu}_{u,+j}(u_j; \boldsymbol{\Delta}^{\mathrm{tup}}_{u,+j}) &= \Bigg(\bigoplus_{k=1}^{d_x} \int_0^1 \boldsymbol{\Delta}_{x,k}(x_k) \odot \frac{p_{ux,jk}(u_j, x_k)}{p_{u,j}(u_j)} dx_k \Bigg) \\
&\quad \oplus \Bigg(\bigoplus_{k \le j-1} \bigoplus_{u_k \in \mathcal{U}_k} \boldsymbol{\Delta}_{u,k}(u_k) \odot \frac{p_{uu,jk}(u_j, u_k)}{p_{u,j}(u_j)} \Bigg), \\
\boldsymbol{\mu}_{u,j+}(u_j; \boldsymbol{\Delta}^{\mathrm{tup}}_{u,j+}) &= \Bigg(\bigoplus_{k \ge j+1} \bigoplus_{u_k \in \mathcal{U}_k} \boldsymbol{\Delta}_{u,k}(u_k) \odot \frac{p_{uu,jk}(u_j, u_k)}{p_{u,j}(u_j)} \Bigg) \\
&\quad \oplus \Bigg(\bigoplus_{k=1}^{d_v} \bigoplus_{v_k \in \mathcal{V}_k} \boldsymbol{\Delta}_{v,k}(v_k) \odot \frac{p_{uv,jk}(u_j, v_k)}{p_{u,j}(u_j)} \Bigg).
\end{aligned}
$$

Then, $\boldsymbol{\Delta}^{\mathrm{tup}}$ is defined as a solution of the following system of equations

$$
\begin{aligned}
\boldsymbol{\Delta}_{x,j}(x_j) &= \tilde{\boldsymbol{\Delta}}_{x,j}(x_j) \ominus \boldsymbol{\mu}_{x,+j}(x_j; \boldsymbol{\Delta}^{\mathrm{tup}}_{x,+j}) \ominus \boldsymbol{\mu}_{x,j+}(x_j; \boldsymbol{\Delta}^{\mathrm{tup}}_{x,j+}), \\
&\qquad\qquad 1 \le j \le d_x, \\
\boldsymbol{\Delta}_{u,j}(u_j) &= \tilde{\boldsymbol{\Delta}}_{u,j}(u_j) \ominus \boldsymbol{\mu}_{u,+j}(u_j; \boldsymbol{\Delta}^{\mathrm{tup}}_{u,+j}) \ominus \boldsymbol{\mu}_{u,j+}(u_j; \boldsymbol{\Delta}^{\mathrm{tup}}_{u,j+}), \\
&\qquad\qquad 1 \le j \le d_u, \\
\boldsymbol{\Delta}_{v,j}(v_j) &= \tilde{\boldsymbol{\Delta}}_{v,j}(v_j) \ominus \boldsymbol{\mu}_{v,+j}(v_j; \boldsymbol{\Delta}^{\mathrm{tup}}_{v,+j}) \ominus \boldsymbol{\mu}_{v,j+}(v_j; \boldsymbol{\Delta}^{\mathrm{tup}}_{v,j+}), \\
&\qquad\qquad 1 \le j \le d_v,
\end{aligned}
\tag{A.32}
$$

subject to the constraints

$$\begin{aligned}
&\int_0^1 \boldsymbol{\Delta}_{x,j}(x_j) \odot p_{x,j}(x_j) dx_j \\
&\quad = \int_0^1 \boldsymbol{\delta}_{x,j}(x_j) \odot (\alpha_j^2 \cdot p_{x,j}(x_j)) dx_j, \quad 1 \le j \le d_x, \\
&\bigoplus_{u_j \in \mathcal{U}_j} \boldsymbol{\Delta}_{u,j}(u_j) \odot p_{u,j}(u_j) \\
&\quad = \bigoplus_{u_j \in \mathcal{U}_j} \bigoplus_{u_j' \in \mathcal{U}_j : u_j' \neq u_j} (\mathbf{m}_{u,j}(u_j') \ominus \mathbf{m}_{u,j}(u_j)) \\
&\qquad\qquad \odot \left(\frac{\beta_j}{c_j - 1} \cdot p_{u,j}(u_j') \right), \quad 1 \le j \le d_u, \\
&\bigoplus_{v_j \in \mathcal{V}_j} \boldsymbol{\Delta}_{v,j}(v_j) \odot p_{v,j}(v_j) \\
&\quad = \bigoplus_{v_j \in \mathcal{V}_j} \bigoplus_{v_j' \in \mathcal{V}_j : \delta_j(v_j, v_j') = \delta_j^*} (\mathbf{m}_{v,j}(v_j') \ominus \mathbf{m}_{v,j}(v_j)) \\
&\qquad\qquad \odot (\gamma_j \cdot p_{v,j}(v_j')), \quad 1 \le j \le d_v.
\end{aligned} \tag{A.33}$$

Below in Section A.7.4, we prove that there exists a unique tuple $\boldsymbol{\Delta}^{\text{tup}}$ that solves the system of equations (A.32) subject to the constraints (A.33).

A.7.2. Proof of Corollary 3

We first note that (2.8) always holds due to the normalization properties (4.2), (4.3) and (4.4). Also, the first property at (2.10) holds with probability tending to one, since

$$\begin{aligned}
&\int_{[0,1]^{d_x}} \bigoplus_{\mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j} \bigoplus_{\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j} \|\hat{\boldsymbol{\mu}}(\mathbf{w})\| \hat{p}(\mathbf{w}) d\mathbf{x} \\
&\quad \le n^{-1} \sum_{i=1}^n \|\hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*) - \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)\| + n^{-1} \sum_{i=1}^n \|\boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)\| \\
&\quad \le M + \mathrm{E}(\|\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)\|) + \delta
\end{aligned}$$

with probability tending to one, where M is the constant given at (B5) and $\delta > 0$ is any constant. We only prove

$$\sum_{u_k \in \mathcal{U}_k} \int_0^1 \left(\frac{\hat{p}_{xu,jk}(x_j, u_k)}{\hat{p}_{x,j}(x_j)} - \frac{p_{xu,jk}(x_j, u_k)}{p_{x,j}(x_j)} \right)^2 \frac{p_{x,j}(x_j)}{p_{u,k}(u_k)} dx_j = o_p(1), \tag{A.34}$$

since the proofs for the other parts follow similarly. We note that the left hand side of (A.34) is bounded by

$$\max_{u_k \in \mathcal{U}_k} \frac{1}{p_{u,k}(u_k)} \sup_{x_j \in [0,1]} \frac{p_{x,j}(x_j)}{(\hat{p}_{x,j}(x_j) p_{x,j}(x_j))^2}$$

$$\times \sum_{u_k \in \mathcal{U}_k} \int_0^1 \left[\hat{p}_{xu,jk}(x_j, u_k)p_{x,j}(x_j) - p_{xu,jk}(x_j, u_k)\hat{p}_{x,j}(x_j)\right]^2 dx_j.$$

We decompose the integrand of the above integral as

$$\begin{aligned}
&\hat{p}_{xu,jk}(x_j, u_k)p_{x,j}(x_j) - p_{xu,jk}(x_j, u_k)\hat{p}_{x,j}(x_j)) \\
&\quad = \left[\hat{p}_{xu,jk}(x_j, u_k) - \int_0^1 K_{h_j}(x_j, x_j')dx_j' \cdot p_{xu,jk}(x_j, u_k)\right] p_{x,j}(x_j) \\
&\qquad + p_{xu,jk}(x_j, u_k)\left[\int_0^1 K_{h_j}(x_j, x_j')dx_j' \cdot p_{x,j}(x_j) - \hat{p}_{x,j}(x_j)\right]
\end{aligned} \tag{A.35}$$

For the first term on the right hand side of (A.35), we note that

$$\begin{aligned}
&\sup_{x_j \in [0,1]} \left|\hat{p}_{xu,jk}(x_j, u_k) - \int_0^1 K_{h_j}(x_j, x_j')dx_j' \cdot p_{xu,jk}(x_j, u_k)\right| \\
&\quad \le \sup_{x_j \in [0,1]} |\hat{p}_{xu,jk}(x_j, u_k) - \mathrm{E}(\hat{p}_{xu,jk}(x_j, u_k))| \\
&\qquad + \sup_{x_j \in [0,1]} \left|\mathrm{E}(\hat{p}_{xu,jk}(x_j, u_k)) - \int_0^1 K_{h_j}(x_j, x_j')dx_j' \cdot p_{xu,jk}(x_j, u_k)\right|.
\end{aligned} \tag{A.36}$$

Lemma 2 implies that the first term on the right hand side of (A.36) is $o_p(1)$. For the second term, we observe

$$\begin{aligned}
\mathrm{E}(\hat{p}_{xu,jk}(x_j, u_k)) &= \mathrm{E}(K_{h_j}(x_j, X_j)L_{\lambda_k}(u_k, U_k)) \\
&= (1 - \lambda_k)\int_0^1 K_{h_j}(x_j, x_j')p_{xu,jk}(x_j', u_k)dx_j' \\
&\qquad + \sum_{u_k' \in \mathcal{U}_k} \lambda_k \int_0^1 K_{h_j}(x_j, x_j')p_{xu,jk}(x_j', u_k')dx_j' \\
&= \int_0^1 K_{h_j}(x_j, x_j')p_{xu,jk}(x_j', u_k)dx_j' + o(1) \\
&= \int_0^1 K_{h_j}(x_j, x_j')dx_j' \cdot p_{xu,jk}(x_j, u_k) + o(1)
\end{aligned}$$

uniformly for $x_j \in [0, 1]$. Hence, the first term on the right hand side of (A.35) is $o_p(1)$ uniformly for $x_j \in [0, 1]$. Similarly, one may prove that the second term on the right hand side of (A.35) is $o_p(1)$ uniformly for $x_j \in [0, 1]$. Since $\hat{p}_{x,j}(x_j) \ge c$ with probability tending to one for some constant $c > 0$, we obtain (A.34).

A.7.3. Proof of Theorem 4

We only sketch the proof since a full proof is too long. Hereafter, we denote $\bigoplus_{u_j \in \mathcal{U}_j}$ and $\bigoplus_{v_j \in \mathcal{V}_j}$ by $\bigoplus_{u_j}$ and $\bigoplus_{v_j}$, respectively. Let $\mathfrak{D}_{t,k}(z_k, z_k') =$

$\hat{\mathbf{m}}_{t,k}(z_k) \ominus \mathbf{m}_{t,k}(z_k')$ for $t = x$, u and v and for z_k, $z_k' \in [0,1]$, $\mathcal{U}_k$ and $\mathcal{V}_k$, respectively, and $\mathfrak{R}_i = \hat{\boldsymbol{\psi}}(\mathbf{W}_i, \mathbf{Y}_i^*) \ominus \boldsymbol{\psi}(\mathbf{W}_i, \mathbf{Y}_i^*)$. Recall the definitions of $\hat{\boldsymbol{\mu}}_{x,j}^A(x_j)$, $\hat{\boldsymbol{\mu}}_{u,j}^A(u_j)$ and $\hat{\boldsymbol{\mu}}_{v,j}^A(v_j)$ immediately before Lemma 3. Define

$$\hat{\boldsymbol{\mu}}_{x,j}^B(x_j) = (n\hat{p}_{x,j}(x_j))^{-1} \odot \bigoplus_{i=1}^{n} K_{h_j}(x_j, X_{ij}) \odot (\mathbf{m}_{x,j}(X_{ij}) \ominus \mathbf{m}_{x,j}(x_j)),$$

and likewise $\hat{\boldsymbol{\mu}}_{u,j}^B(u_j)$ and $\hat{\boldsymbol{\mu}}_{v,j}^B(v_j)$ with $\mathbf{m}_{x,j}(X_{ij}) \ominus \mathbf{m}_{x,j}(x_j)$ being replaced by $\mathbf{m}_{u,j}(U_{ij}) \ominus \mathbf{m}_{u,j}(u_j)$ and $\mathbf{m}_{v,j}(V_{ij}) \ominus \mathbf{m}_{v,j}(v_j)$, respectively. Then, we may write the gB-SBF system of equations for the mixed predictor case, as

$$\begin{aligned}
\hat{\mathbf{m}}_{x,j}(x_j) = \mathbf{m}_{x,j}(x_j) &\oplus \hat{\boldsymbol{\mu}}_{x,j}^A(x_j) \oplus \hat{\boldsymbol{\mu}}_{x,j}^B(x_j) \oplus \mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)) \ominus \hat{\mathbf{m}}_0 \\
&\ominus \frac{1}{n\hat{p}_{x,j}(x_j)} \odot \Bigg[\bigoplus_{i=1}^{n} \bigoplus_{k \neq j} \int_0^1 \mathfrak{D}_{x,k}(x_k, X_{ik}) \odot (K_{h_j}(x_j, X_{ij}) \\
&\quad \times K_{h_k}(x_k, X_{ik}))dx_k \oplus \bigoplus_{i=1}^{n} \bigoplus_{k=1}^{d_u} \bigoplus_{u_k} \mathfrak{D}_{u,k}(u_k, U_{ik}) \\
&\quad \odot (K_{h_j}(x_j, X_{ij}) L_{\lambda_k}(u_k, U_{ik})) \oplus \bigoplus_{i=1}^{n} \bigoplus_{k=1}^{d_u} \bigoplus_{v_k} \mathfrak{D}_{v,k}(v_k, V_{ik}) \\
&\quad \odot (K_{h_j}(x_j, X_{ij}) W_{s_k}(v_k, V_{ik})) \ominus \bigoplus_{i=1}^{n} K_{h_j}(x_j, X_{ij}) \odot \mathfrak{R}_i \Bigg], \\
&\qquad\qquad 1 \leq j \leq d_x,
\end{aligned} \tag{A.37}$$

$$\begin{aligned}
\hat{\mathbf{m}}_{u,j}(u_j) = \mathbf{m}_{u,j}(u_j) &\oplus \hat{\boldsymbol{\mu}}_{u,j}^A(u_j) \oplus \hat{\boldsymbol{\mu}}_{u,j}^B(u_j) \oplus \mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)) \ominus \hat{\mathbf{m}}_0 \\
&\ominus \frac{1}{n\hat{p}_{u,j}(u_j)} \odot \Bigg[\bigoplus_{i=1}^{n} \bigoplus_{k=1}^{d_x} \int_0^1 \mathfrak{D}_{x,k}(x_k, X_{ik}) \odot (L_{\lambda_j}(u_j, U_{ij}) \\
&\quad \times K_{h_k}(x_k, X_{ik}))dx_k \oplus \bigoplus_{i=1}^{n} \bigoplus_{k \neq j} \bigoplus_{u_k} \mathfrak{D}_{u,k}(u_k, U_{ik}) \\
&\quad \odot (L_{\lambda_j}(u_j, U_{ij}) L_{\lambda_k}(u_k, U_{ik})) \oplus \bigoplus_{i=1}^{n} \bigoplus_{k=1}^{d_v} \bigoplus_{v_k} \mathfrak{D}_{v,k}(v_k, V_{ik}) \\
&\quad \odot (L_{\lambda_j}(u_j, U_{ij}) W_{s_k}(v_k, V_{ik})) \ominus \bigoplus_{i=1}^{n} L_{\lambda_j}(u_j, U_{ij}) \odot \mathfrak{R}_i \Bigg], \\
&\qquad\qquad 1 \leq j \leq d_u,
\end{aligned} \tag{A.38}$$

$$
\begin{aligned}
\hat{\mathbf{m}}_{v,j}(v_j) = \mathbf{m}_{v,j}(v_j) \oplus \hat{\boldsymbol{\mu}}^A_{v,j}(v_j) \oplus \hat{\boldsymbol{\mu}}^B_{v,j}(v_j) \oplus \mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)) \ominus \hat{\mathbf{m}}_0 \\
\ominus \frac{1}{n\hat{p}_{v,j}(v_j)} \odot \Bigg[\bigoplus_{i=1}^{n} \bigoplus_{k=1}^{d_x} \int_0^1 \mathfrak{D}_{x,k}(x_k, X_{ik}) \odot (W_{s_j}(v_j, V_{ij}) \\
\times K_{h_k}(x_k, X_{ik}))dx_k \oplus \bigoplus_{i=1}^{n} \bigoplus_{k=1}^{d_u} \bigoplus_{u_k} \mathfrak{D}_{u,k}(u_k, U_{ik}) \\
\odot (W_{s_j}(v_j, V_{ij}) L_{\lambda_k}(u_k, U_{ik})) \oplus \bigoplus_{i=1}^{n} \bigoplus_{k \neq j} \bigoplus_{v_k} \mathfrak{D}_{v,k}(v_k, V_{ik}) \\
\odot (W_{s_j}(u_j, U_{ij}) W_{s_k}(v_k, V_{ik})) \ominus \bigoplus_{i=1}^{n} W_{s_j}(v_j, V_{ij}) \odot \mathfrak{R}_i \Bigg] \\
, \qquad\qquad 1 \le j \le d_v.
\end{aligned}
\tag{A.39}
$$

We note that

$$
\begin{aligned}
\left\| (n\hat{p}_{x,j}(x_j))^{-1} \odot \bigoplus_{i=1}^{n} K_{h_j}(x_j, X_{ij}) \odot \mathfrak{R}_i \right\| = O_p(a_n), \\
\left\| (n\hat{p}_{u,j}(u_j))^{-1} \odot \bigoplus_{i=1}^{n} L_{\lambda_j}(u_j, U_{ij}) \odot \mathfrak{R}_i \right\| = O_p(a_n), \\
\left\| (n\hat{p}_{v,j}(v_j))^{-1} \odot \bigoplus_{i=1}^{n} W_{s_j}(v_j, V_{ij}) \odot \mathfrak{R}_i \right\| = O_p(a_n), \\
\| \mathrm{E}(\boldsymbol{\psi}(\mathbf{W}, \mathbf{Y}^*)) \ominus \hat{\mathbf{m}}_0 \| = O_p(n^{-1/2} + a_n).
\end{aligned}
\tag{A.40}
$$

We first approximate the right hand side of (A.37). By the standard kernel smoothing theory and using Lemma 2, we may prove that

$$
\begin{aligned}
\frac{1}{n\hat{p}_{x,j}(x_j)} \odot \bigoplus_{i=1}^{n} \bigoplus_{k=1}^{d_u} \bigoplus_{u_k} (\mathbf{m}_{u,k}(U_{ik}) \ominus \mathbf{m}_{u,k}(u_k)) \\
\odot (K_{h_j}(x_j, X_{ij}) L_{\lambda_k}(u_k, U_{ik})) \\
= \bigoplus_{k=1}^{d_u} \bigoplus_{u_k} \bigoplus_{u'_k} (\mathbf{m}_{u,k}(u'_k) \ominus \mathbf{m}_{u,k}(u_k)) \odot \left(L_{\lambda_k}(u_k, u'_k) \cdot \frac{p_{xu,jk}(x_j, u'_k)}{p_{x,j}(x_j)} \right) \\
\oplus o_p(h_j \cdot \lambda_*) \oplus O_p(n^{-2/5} \sqrt{\log n} \cdot \lambda_*), \\
\frac{1}{n\hat{p}_{x,j}(x_j)} \odot \bigoplus_{i=1}^{n} \bigoplus_{k=1}^{d_v} \bigoplus_{v_k} (\mathbf{m}_{v,k}(V_{ik}) \ominus \mathbf{m}_{v,k}(v_k)) \\
\odot (K_{h_j}(x_j, X_{ij}) W_{s_k}(v_k, V_{ik})) \\
= \bigoplus_{k=1}^{d_v} \bigoplus_{v_k} \bigoplus_{v'_k} (\mathbf{m}_{v,k}(v'_k) \ominus \mathbf{m}_{v,k}(v_k)) \odot \left(W_{s_k}(v_k, v'_k) \cdot \frac{p_{xv,jk}(x_j, v'_k)}{p_{x,j}(x_j)} \right) \\
\oplus o_p(h_j \cdot s_*) \oplus O_p(n^{-2/5} \sqrt{\log n} \cdot s_*)
\end{aligned}
\tag{A.41}
$$

uniformly for $x_j \in [0,1]$. We may also show that

$$\begin{aligned}\sup_{x_j\in[0,1]}\left\|\bigoplus_{u_k}\hat{\boldsymbol{\mu}}^A_{u,k}(u_k)\odot\frac{\hat{p}_{xu,jk}(x_j,u_k)}{\hat{p}_{x,j}(x_j)}\right\| &= O_p(n^{-1/2}),\\ \sup_{x_j\in[0,1]}\left\|\bigoplus_{v_k}\hat{\boldsymbol{\mu}}^A_{v,k}(v_k)\odot\frac{\hat{p}_{xv,jk}(x_j,v_k)}{\hat{p}_{x,j}(x_j)}\right\| &= O_p(n^{-1/2}).\end{aligned} \tag{A.42}$$

Define

$$\begin{aligned}\mathbf{a}_j(x_j) &= \int_0^1\left(\frac{x_j'-x_j}{h_j}\right)K_{h_j}(x_j,x_j')dx_j'\odot D\mathbf{m}_{x,j}(x_j)(1),\\ \boldsymbol{\Delta}^\dagger_{x,j}(x_j) &= h_j^2\odot\boldsymbol{\delta}_{x,j}(x_j)\oplus\bigoplus_{k\neq j}\int_0^1\boldsymbol{\delta}_{xx,jk}(x_j,x_k)\odot\left(h_k^2\cdot\frac{p_{xx,jk}(x_j,x_k)}{p_{x,j}(x_j)}\right)dx_k\\ &\quad\oplus\bigoplus_{k=1}^{d_u}\bigoplus_{u_k}\bigoplus_{u_k'}(\mathbf{m}_{u,k}(u_k')\ominus\mathbf{m}_{u,k}(u_k))\\ &\qquad\odot\left(L_{\lambda_k}(u_k,u_k')\cdot\frac{p_{xu,jk}(x_j,u_k')}{p_{x,j}(x_j)}\right)\\ &\quad\oplus\bigoplus_{k=1}^{d_v}\bigoplus_{v_k}\bigoplus_{v_k'}(\mathbf{m}_{v,k}(v_k')\ominus\mathbf{m}_{v,k}(v_k))\\ &\qquad\odot\left(W_{s_k}(v_k,v_k')\cdot\frac{p_{xv,jk}(x_j,v_k')}{p_{x,j}(x_j)}\right).\end{aligned}$$

Then, from (A.40), (A.41), (A.42) and Lemma S.9 in [15] it follows that

$$\begin{aligned}&\hat{\mathbf{m}}_{x,j}(x_j)\ominus\mathbf{m}_{x,j}(x_j)\ominus\hat{\boldsymbol{\mu}}^A_{x,j}(x_j)\ominus\frac{h_j}{\int_0^1K_{h_j}(x_j,x_j')dx_j'}\odot\mathbf{a}_j(x_j)\\ &\qquad\ominus h_j^2\odot\mathbf{c}_j(x_j)\\ &=\boldsymbol{\Delta}^\dagger_{x,j}(x_j)\ominus\bigoplus_{k\neq j}\int_0^1\left[\hat{\mathbf{m}}_{x,k}(x_k)\ominus\mathbf{m}_{x,k}(x_k)\ominus\hat{\boldsymbol{\mu}}^A_{x,k}(x_k)\right.\\ &\quad\left.\ominus\frac{h_k}{\int_0^1K_{h_k}(x_k,x_k')dx_k'}\odot\mathbf{a}_k(x_k)\ominus h_k^2\odot\mathbf{c}_k(x_k)\right]\odot\frac{\hat{p}_{xx,jk}(x_j,x_k)}{\hat{p}_{x,j}(x_j)}dx_k\\ &\quad\ominus\bigoplus_{k=1}^{d_u}\bigoplus_{u_k}\left[\hat{\mathbf{m}}_{u,k}(u_k)\ominus\mathbf{m}_{u,k}(u_k)\ominus\hat{\boldsymbol{\mu}}^A_{u,k}(u_k)\right]\odot\frac{\hat{p}_{xu,jk}(x_j,u_k)}{\hat{p}_{x,j}(x_j)}\\ &\quad\ominus\bigoplus_{k=1}^{d_v}\bigoplus_{v_k}\left[\hat{\mathbf{m}}_{v,k}(v_k)\ominus\mathbf{m}_{v,k}(v_k)\ominus\hat{\boldsymbol{\mu}}^A_{v,k}(v_k)\right]\odot\frac{\hat{p}_{xv,jk}(x_j,v_k)}{\hat{p}_{x,j}(x_j)}\\ &\quad\oplus\mathbf{r}_{x,j}(x_j),\end{aligned} \tag{A.43}$$

where $\mathbf{r}_{x,j}:[0,1]\to\mathbb{H}$ are generic stochastic maps satisfying

$$\sup_{x_j\in I_j}\|\mathbf{r}_{x,j}(x_j)\| = o_p(n^{-2/5})+O_p(n^{-2/5}\sqrt{\log n}\cdot(\lambda_*+s_*)+a_n),$$

$$\sup_{x_j\in[0,1]} \|\mathbf{r}_{x,j}(x_j)\| = O_p(n^{-2/5} + n^{-2/5}\sqrt{\log n}\cdot(\lambda_* + s_*) + a_n).$$

We now approximate the right hand side of (A.38). We get that

$$\begin{aligned}
&\frac{1}{n\hat{p}_{u,j}(u_j)} \odot \bigoplus_{i=1}^{n}\bigoplus_{k=1}^{d_x}\int_0^1 [\mathbf{m}_{x,k}(X_{ik}) \ominus \mathbf{m}_{x,k}(x_k)] \\
&\qquad\qquad \odot (L_{\lambda_j}(u_j, U_{ij})K_{h_k}(x_k, X_{ik}))dx_k \\
&= \bigoplus_{k=1}^{d_x}\Bigg[\int_0^1 \left(\frac{h_k}{\int_0^1 K_{h_k}(x_k, x_k^*)dx_k^*}\odot \mathbf{a}_k(x_k) \oplus h_k^2\mathbf{c}_k(x_k)\right) \\
&\qquad\qquad \odot \frac{\hat{p}_{ux,jk}(u_j, x_k)}{\hat{p}_{u,j}(u_j)}dx_k \\
&\qquad \oplus \int_0^1 \boldsymbol{\delta}_{ux,jk}(u_j, x_k)\odot\left(h_k^2\cdot\frac{p_{ux,jk}(u_j,x_k)}{p_{u,j}(u_j)}\right)dx_k\Bigg] \\
&\quad \oplus o_p(n^{-2/5}), \\
&\frac{1}{n\hat{p}_{u,j}(u_j)} \odot \bigoplus_{i=1}^{n}\bigoplus_{k\neq j}\bigoplus_{u_k}[\mathbf{m}_{u,k}(U_{ik}) \ominus \mathbf{m}_{u,k}(u_k)] \\
&\qquad\qquad \odot (L_{\lambda_j}(u_j, U_{ij})L_{\lambda_k}(u_k, U_{ik})) \\
&= \bigoplus_{k\neq j}\bigoplus_{u_k}\bigoplus_{u_k'}[\mathbf{m}_{u,k}(u_k') \ominus \mathbf{m}_{u,k}(u_k)] \\
&\qquad\qquad \odot \left(L_{\lambda_k}(u_k, u_k')\cdot\frac{p_{uu,jk}(u_j,u_k')}{p_{u,j}(u_j)}\right) \\
&\quad \oplus O_p(\lambda_*\cdot(\lambda_j + n^{-1/2})), \\
&\frac{1}{n\hat{p}_{u,j}(u_j)} \odot \bigoplus_{i=1}^{n}\bigoplus_{k=1}^{d_v}\bigoplus_{v_k}[\mathbf{m}_{v,k}(V_{ik}) \ominus \mathbf{m}_{v,k}(v_k)] \\
&\qquad\qquad \odot (L_{\lambda_j}(u_j, U_{ij})W_{s_k}(v_k, V_{ik})) \\
&= \bigoplus_{k=1}^{d_v}\bigoplus_{v_k}\bigoplus_{v_k'}[\mathbf{m}_{v,k}(v_k') \ominus \mathbf{m}_{v,k}(v_k)] \\
&\qquad\qquad \odot \left(W_{s_k}(v_k, v_k')\cdot\frac{p_{uv,jk}(u_j,v_k')}{p_{u,j}(u_j)}\right) \\
&\quad \oplus O_p(s_*\cdot(\lambda_j + n^{-1/2})).
\end{aligned} \tag{A.44}$$

We may also prove that

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_{u,j}^B(u_j) = \bigoplus_{u_j'}(\mathbf{m}_{u,j}(u_j') \ominus \mathbf{m}_{u,j}(u_j))\odot\left(L_{\lambda_j}(u_j, u_j')\cdot\frac{p_{u,j}(u_j')}{p_{u,j}(u_j)}\right) \\
\oplus O_p(\lambda_j\cdot(\lambda_j + n^{-1/2})).
\end{aligned} \tag{A.45}$$

Furthermore,

$$\begin{aligned}\int_0^1 \hat{\boldsymbol{\mu}}^A_{x,k}(x_k) \odot \frac{\hat{p}_{ux,jk}(u_j, x_k)}{\hat{p}_{u,j}(u_j)} dx_k &= O_p(n^{-1/2}),\\ \bigoplus_{u_k} \hat{\boldsymbol{\mu}}^A_{u,k}(u_k) \odot \frac{\hat{p}_{uu,jk}(u_j, u_k)}{\hat{p}_{u,j}(u_j)} &= O_p(n^{-1/2}),\\ \bigoplus_{v_k} \hat{\boldsymbol{\mu}}^A_{v,k}(v_k) \odot \frac{\hat{p}_{uv,jk}(u_j, v_k)}{\hat{p}_{u,j}(u_j)} &= O_p(n^{-1/2}).\end{aligned} \tag{A.46}$$

Define

$$\begin{aligned}\boldsymbol{\Delta}^{\dagger}_{u,j}(u_j) = \int_0^1 \boldsymbol{\delta}_{ux,jk}(u_j, x_k) \odot \left(h_k^2 \cdot \frac{p_{ux,jk}(u_j, x_k)}{p_{u,j}(u_j)} \right) dx_k \\ \oplus \bigoplus_{u'_j} (\mathbf{m}_{u,j}(u'_j) \ominus \mathbf{m}_{u,j}(u_j)) \odot \left(L_{\lambda_j}(u_j, u'_j) \cdot \frac{p_{u,j}(u'_j)}{p_{u,j}(u_j)} \right) \\ \oplus \bigoplus_{k \neq j} \bigoplus_{u_k} \bigoplus_{u'_k} (\mathbf{m}_{u,k}(u'_k) \ominus \mathbf{m}_{u,k}(u_k)) \\ \odot \left(L_{\lambda_k}(u_k, u'_k) \cdot \frac{p_{uu,jk}(u_j, u'_k)}{p_{u,j}(u_j)} \right) \\ \oplus \bigoplus_{k=1}^{d_v} \bigoplus_{v_k} \bigoplus_{v'_k} (\mathbf{m}_{v,k}(v'_k) \ominus \mathbf{m}_{v,k}(v_k)) \\ \odot \left(W_{s_k}(v_k, v'_k) \cdot \frac{p_{uv,jk}(u_j, v'_k)}{p_{u,j}(u_j)} \right).\end{aligned}$$

Let $\mathbf{r}_{u,j} : \mathcal{U}_j \to \mathbb{H}$ denote generic stochastic maps satisfying

$$\max_{u_j \in \mathcal{U}_j} \|\mathbf{r}_{u,j}(u_j)\| = o_p(n^{-2/5}) + O_p(\lambda_j \cdot (\lambda_* + s_*) + a_n).$$

Then, from (A.44), (A.45) and (A.46) we have

$$\begin{aligned}&\hat{\mathbf{m}}_{u,j}(u_j) \ominus \mathbf{m}_{u,j}(u_j) \ominus \hat{\boldsymbol{\mu}}^A_{u,j}(u_j) \\ &= \boldsymbol{\Delta}^{\dagger}_{u,j}(u_j) \ominus \bigoplus_{k=1}^{d_x} \int_0^1 \left[\hat{\mathbf{m}}_{x,k}(x_k) \ominus \mathbf{m}_{x,k}(x_k) \ominus \hat{\boldsymbol{\mu}}^A_{x,k}(x_k) \right. \\ &\quad \left. \ominus \left(\frac{h_k}{\int_0^1 K_{h_k}(x_k, x'_k) dx'_k} \right) \odot \mathbf{a}_k(x_k) \ominus h_k^2 \odot \mathbf{c}_k(x_k) \right] \\ &\qquad \odot \frac{\hat{p}_{ux,jk}(u_j, x_k)}{\hat{p}_{u,j}(u_j)} dx_k \\ &\quad \ominus \bigoplus_{k \neq j} \bigoplus_{u_k} \left[\hat{\mathbf{m}}_{u,k}(u_k) \ominus \mathbf{m}_{u,k}(u_k) \ominus \hat{\boldsymbol{\mu}}^A_{u,k}(u_k) \right] \odot \frac{\hat{p}_{uu,jk}(u_j, u_k)}{\hat{p}_{u,j}(u_j)}\end{aligned} \tag{A.47}$$

$$\ominus \bigoplus_{k=1}^{d_v} \bigoplus_{v_k} \Big[\hat{\mathbf{m}}_{v,k}(v_k) \ominus \mathbf{m}_{v,k}(v_k) \ominus \hat{\boldsymbol{\mu}}^A_{v,k}(v_k)\Big] \odot \frac{\hat{p}_{uv,jk}(u_j, v_k)}{\hat{p}_{u,j}(u_j)}$$
$$\oplus\ \mathbf{r}_{u,j}(u_j).$$

Similarly, from (A.39) we also get

$$\begin{aligned}
&\hat{\mathbf{m}}_{v,j}(v_j) \ominus \mathbf{m}_{v,j}(v_j) \ominus \hat{\boldsymbol{\mu}}^A_{v,j}(v_j) \\
&= \boldsymbol{\Delta}^{\dagger}_{v,j}(v_j) \ominus \bigoplus_{k=1}^{d_x} \int_0^1 \Bigg[\hat{\mathbf{m}}_{x,k}(x_k) \ominus \mathbf{m}_{x,k}(x_k) \ominus \hat{\boldsymbol{\mu}}^A_{x,k}(x_k) \\
&\qquad\qquad \ominus \left(\frac{h_k}{\int_0^1 K_{h_k}(x_k, x'_k)dx'_k}\right) \odot \mathbf{a}_k(x_k) \ominus h_k^2 \odot \mathbf{c}_k(x_k)\Bigg] \\
&\qquad\qquad \odot \frac{\hat{p}_{vx,jk}(v_j, x_k)}{\hat{p}_{v,j}(v_j)} dx_k \qquad \text{(A.48)} \\
&\ominus \bigoplus_{k=1}^{d_u} \bigoplus_{u_k} \Big[\hat{\mathbf{m}}_{u,k}(u_k) \ominus \mathbf{m}_{u,k}(u_k) \ominus \hat{\boldsymbol{\mu}}^A_{u,k}(u_k)\Big] \odot \frac{\hat{p}_{vu,jk}(v_j, u_k)}{\hat{p}_{v,j}(v_j)} \\
&\ominus \bigoplus_{k\neq j} \bigoplus_{v_k} \Big[\hat{\mathbf{m}}_{v,k}(v_k) \ominus \mathbf{m}_{v,k}(v_k) \ominus \hat{\boldsymbol{\mu}}^A_{v,k}(v_k)\Big] \odot \frac{\hat{p}_{vv,jk}(v_j, v_k)}{\hat{p}_{v,j}(v_j)} \\
&\oplus\ \mathbf{r}_{v,j}(v_j),
\end{aligned}$$

where

$$\begin{aligned}
&\boldsymbol{\Delta}^{\dagger}_{v,j}(v_j) \\
&= \int_0^1 \boldsymbol{\delta}_{vx,jk}(v_j, x_k) \odot \left(h_k^2 \cdot \frac{p_{vx,jk}(v_j, x_k)}{p_{v,j}(v_j)}\right) dx_k \\
&\quad \oplus \bigoplus_{v'_j} (\mathbf{m}_{v,j}(v'_j) \ominus \mathbf{m}_{v,j}(v_j)) \odot \left(W_{s_j}(v_j, v'_j) \cdot \frac{p_{v,j}(v'_j)}{p_{v,j}(v_j)}\right) \\
&\quad \oplus \bigoplus_{k=1}^{d_u} \bigoplus_{u_k} \bigoplus_{u'_k} (\mathbf{m}_{u,k}(u'_k) \ominus \mathbf{m}_{u,k}(u_k)) \odot \left(L_{\lambda_k}(u_k, u'_k) \cdot \frac{p_{vu,jk}(v_j, u'_k)}{p_{v,j}(v_j)}\right) \\
&\quad \oplus \bigoplus_{k\neq j} \bigoplus_{v_k} \bigoplus_{v'_k} (\mathbf{m}_{v,k}(v'_k) \ominus \mathbf{m}_{v,k}(v_k)) \odot \left(W_{s_k}(v_k, v'_k) \cdot \frac{p_{vv,jk}(v_j, v'_k)}{p_{v,j}(v_j)}\right),
\end{aligned}$$

and $\mathbf{r}_{v,j} : \mathcal{V}_j \to \mathbb{H}$ are generic stochastic maps satisfying

$$\max_{v_j \in \mathcal{V}_j} \|\mathbf{r}_{v,j}(v_j)\| = o_p(n^{-2/5}) + O_p(s_j^{\delta_j^*} \cdot (\lambda_* + s_*) + a_n).$$

Now, define

$$\hat{\boldsymbol{\Delta}}_{x,j}(x_j) = \hat{\mathbf{m}}_{x,j}(x_j) \ominus \mathbf{m}_{x,j}(x_j) \ominus \hat{\boldsymbol{\mu}}^A_{x,j}(x_j) \ominus \left(\frac{h_j}{\int_0^1 K_{h_j}(x_j, x'_j)dx'_j}\right) \odot \mathbf{a}_j(x_j)$$

$$\ominus\, h_j^2 \odot \mathbf{c}_j(x_j) \ominus \mathbf{r}_{x,j}(x_j),$$
$$\hat{\mathbf{\Delta}}_{u,j}(u_j) = \hat{\mathbf{m}}_{u,j}(u_j) \ominus \mathbf{m}_{u,j}(u_j) \ominus \hat{\boldsymbol{\mu}}^A_{u,j}(u_j) \ominus \mathbf{r}_{u,j}(u_j),$$
$$\hat{\mathbf{\Delta}}_{v,j}(v_j) = \hat{\mathbf{m}}_{v,j}(v_j) \ominus \mathbf{m}_{v,j}(v_j) \ominus \hat{\boldsymbol{\mu}}^A_{v,j}(v_j) \ominus \mathbf{r}_{v,j}(v_j).$$

Then, from (A.43), (A.47) and (A.48), we have

$$\begin{aligned}
\hat{\mathbf{\Delta}}_{x,j}(x_j) &= \mathbf{\Delta}^\dagger_{x,j}(x_j) \ominus \hat{\boldsymbol{\mu}}_{x,+j}(x_j; \hat{\mathbf{\Delta}}^{\text{tup}}_{x,+j}) \ominus \hat{\boldsymbol{\mu}}_{x,j+}(x_j; \hat{\mathbf{\Delta}}^{\text{tup}}_{x,j+}) \\
&\qquad \oplus \tilde{\mathbf{r}}_{x,j}(x_j), \\
\hat{\mathbf{\Delta}}_{u,j}(u_j) &= \mathbf{\Delta}^\dagger_{u,j}(u_j) \ominus \hat{\boldsymbol{\mu}}_{u,+j}(u_j; \hat{\mathbf{\Delta}}^{\text{tup}}_{u,+j}) \ominus \hat{\boldsymbol{\mu}}_{u,j+}(u_j; \hat{\mathbf{\Delta}}^{\text{tup}}_{u,j+}) \\
&\qquad \oplus \tilde{\mathbf{r}}_{u,j}(u_j), \\
\hat{\mathbf{\Delta}}_{v,j}(v_j) &= \mathbf{\Delta}^\dagger_{v,j}(v_j) \ominus \hat{\boldsymbol{\mu}}_{v,+j}(v_j; \hat{\mathbf{\Delta}}^{\text{tup}}_{v,+j}) \ominus \hat{\boldsymbol{\mu}}_{v,j+}(u_j; \hat{\mathbf{\Delta}}^{\text{tup}}_{v,j+}) \\
&\qquad \oplus \tilde{\mathbf{r}}_{v,j}(v_j),
\end{aligned} \tag{A.49}$$

where $\hat{\boldsymbol{\mu}}_{t,+j}(t_j; \hat{\mathbf{\Delta}}^{\text{tup}}_{t,+j})$ and $\hat{\boldsymbol{\mu}}_{t,j+}(t_j; \hat{\mathbf{\Delta}}^{\text{tup}}_{t,j+})$ are defined as $\boldsymbol{\mu}_{t,+j}(t_j; \mathbf{\Delta}^{\text{tup}}_{t,+j})$ and $\boldsymbol{\mu}_{t,j+}(t_j; \mathbf{\Delta}^{\text{tup}}_{t,j+})$ with $\mathbf{\Delta}_{t,k}, p_{t,k}$ and $p_{tt',kk'}$ being replaced by $\hat{\mathbf{\Delta}}_{t,k}, \hat{p}_{t,k}$ and $\hat{p}_{tt',kk'}$ for all $t, t' = x, u, v$ and k, k', and $\tilde{\mathbf{r}}_{x,j}$, $\tilde{\mathbf{r}}_{u,j}$ and $\tilde{\mathbf{r}}_{v,j}$ are $\mathbb{H}$-valued stochastic maps satisfying

$$\begin{aligned}
&\sup_{x_j \in [0,1]} \|\tilde{\mathbf{r}}_{x,j}(x_j)\|, \max_{u_j \in \mathcal{U}_j} \|\tilde{\mathbf{r}}_{u,j}(u_j)\|, \max_{v_j \in \mathcal{V}_j} \|\tilde{\mathbf{r}}_{v,j}(v_j)\| \\
&\quad = o_p(n^{-2/5}) + O_p(n^{-2/5}\sqrt{\log n} \cdot (\lambda_* + s_*) + \lambda_*^2 + s_*^2 + a_n).
\end{aligned} \tag{A.50}$$

We can also show that

$$\begin{aligned}
&\sup_{x_j \in [0,1]} \|\mathbf{\Delta}^\dagger_{x,j}(x_j)\|, \max_{u_j \in \mathcal{U}_j} \|\mathbf{\Delta}^\dagger_{u,j}(u_j)\|, \max_{v_j \in \mathcal{V}_j} \|\mathbf{\Delta}^\dagger_{v,j}(v_j)\| \\
&\quad = O(n^{-2/5} + \lambda_* + s_*).
\end{aligned} \tag{A.51}$$

Define $\hat{\mathbf{\Delta}}_\oplus = \bigoplus_{j=1}^{d_x} \hat{\mathbf{\Delta}}_{x,j} \oplus \bigoplus_{j=1}^{d_u} \hat{\mathbf{\Delta}}_{u,j} \oplus \bigoplus_{j=1}^{d_v} \hat{\mathbf{\Delta}}_{v,j}$. Then, (A.50) and (A.51) yield

$$\|\hat{\mathbf{\Delta}}_\oplus\|_2 = O_p(n^{-2/5} + \lambda_* + s_* + a_n), \tag{A.52}$$

where the norm $\|\cdot\|_2$ for a square integrable map $\mathbf{f} : \mathcal{W} \to \mathbb{H}$ is defined by

$$\|\mathbf{f}\|_2^2 = \sum_{\mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j} \sum_{\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j} \int_{[0,1]^{d_x}} \|\mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{v})\|^2 p(\mathbf{x}, \mathbf{u}, \mathbf{v}) d\mathbf{x}. \tag{A.53}$$

A version of Proposition 3 implies that there exist a decomposition

$$\hat{\mathbf{\Delta}}_\oplus = \bigoplus_{j=1}^{d_x} \hat{\mathbf{\Delta}}^*_{x,j} \oplus \bigoplus_{j=1}^{d_u} \hat{\mathbf{\Delta}}^*_{u,j} \oplus \bigoplus_{j=1}^{d_v} \hat{\mathbf{\Delta}}^*_{v,j}$$

and a constant $c > 0$ such that

$$\max\left\{ \max_{1 \le j \le d_x} \|\hat{\mathbf{\Delta}}^*_{x,j}\|_2, \max_{1 \le j \le d_u} \|\hat{\mathbf{\Delta}}^*_{u,j}\|_2, \max_{1 \le j \le d_v} \|\hat{\mathbf{\Delta}}^*_{v,j}\|_2 \right\} \le c \cdot \|\hat{\mathbf{\Delta}}_\oplus\|_2. \tag{A.54}$$

Then, by Lemma 1 it holds that

$$\begin{aligned}
\hat{\boldsymbol{\Delta}}_{x,j}(x_j) &= \hat{\boldsymbol{\Delta}}^*_{x,j}(x_j) \oplus \hat{\mathbf{c}}_{x,j} \text{ a.e. with respect to Leb,} \\
\hat{\boldsymbol{\Delta}}_{u,j}(u_j) &= \hat{\boldsymbol{\Delta}}^*_{u,j}(u_j) \oplus \hat{\mathbf{c}}_{u,j} \text{ for all } u_j, \\
\hat{\boldsymbol{\Delta}}_{v,j}(v_j) &= \hat{\boldsymbol{\Delta}}^*_{v,j}(v_j) \oplus \hat{\mathbf{c}}_{v,j} \text{ for all } v_j
\end{aligned} \tag{A.55}$$

for some stochastic Hilbertian constants $\hat{\mathbf{c}}_{x,j}, \hat{\mathbf{c}}_{u,j}$ and $\hat{\mathbf{c}}_{v,j}$ satisfying $\bigoplus_{j=1}^{d_x} \hat{\mathbf{c}}_{x,j} \oplus \bigoplus_{j=1}^{d_u} \hat{\mathbf{c}}_{u,j} \oplus \bigoplus_{j=1}^{d_v} \hat{\mathbf{c}}_{v,j} = \mathbf{0}$. Expanding $\int_0^1 (\hat{\boldsymbol{\Delta}}_{x,j}(x_j) \ominus \hat{\boldsymbol{\Delta}}^*_{x,j}(x_j)) \odot \hat{p}_{x,j}(x_j) dx_j$, $\bigoplus_{u_j} (\hat{\boldsymbol{\Delta}}_{u,j}(u_j) \ominus \hat{\boldsymbol{\Delta}}^*_{u,j}(u_j)) \odot \hat{p}_{u,j}(u_j)$ and $\bigoplus_{v_j} (\hat{\boldsymbol{\Delta}}_{v,j}(v_j) \ominus \hat{\boldsymbol{\Delta}}^*_{v,j}(v_j)) \odot \hat{p}_{v,j}(v_j)$ and using the constraints

$$\begin{aligned}
\int_0^1 \mathbf{m}_{x,j}(x_j) \odot p_{x,j}(x_j) dx_j &= \int_0^1 \hat{\mathbf{m}}_{x,j}(x_j) \odot \hat{p}_{x,j}(x_j) dx_j = \mathbf{0}, \quad 1 \le j \le d_x, \\
\bigoplus_{u_j} \mathbf{m}_{u,j}(u_j) \odot p_{u,j}(u_j) &= \bigoplus_{u_j} \hat{\mathbf{m}}_{u,j}(u_j) \odot \hat{p}_{u,j}(u_j) = \mathbf{0}, \quad 1 \le j \le d_u, \\
\bigoplus_{v_j} \mathbf{m}_{v,j}(v_j) \odot p_{v,j}(v_j) &= \bigoplus_{v_j} \hat{\mathbf{m}}_{v,j}(v_j) \odot \hat{p}_{v,j}(v_j) = \mathbf{0}, \quad 1 \le j \le d_v,
\end{aligned}$$

we may prove that

$$\hat{\mathbf{c}}_{x,j},\ \hat{\mathbf{c}}_{u,j},\ \hat{\mathbf{c}}_{v,j} = O_p(n^{-2/5} + \lambda_* + s_* + a_n). \tag{A.56}$$

From (A.52), (A.54), (A.55) and (A.56), we have

$$\|\hat{\boldsymbol{\Delta}}_{x,j}\|_2,\ \|\hat{\boldsymbol{\Delta}}_{u,j}\|_2,\ \|\hat{\boldsymbol{\Delta}}_{v,j}\|_2 = O_p(n^{-2/5} + \lambda_* + s_* + a_n). \tag{A.57}$$

Now, using (A.49), (A.50), (A.51), (A.57) and the standard kernel smoothing theory we may establish the theorem.

A.7.4. Proof of Theorem 5

We also sketch the proof. We first prove that there exists $(\boldsymbol{\Delta}_{x,1}, \ldots, \boldsymbol{\Delta}_{v,d_v})$ that solves the system of equations (A.32) subject to the constraints (A.33). Let

$$\mathcal{S}^{\mathbb{H}}(p) = \left\{ \bigoplus_{j=1}^{d_x} \mathbf{f}_{x,j} \oplus \bigoplus_{j=1}^{d_u} \mathbf{f}_{u,j} \oplus \bigoplus_{j=1}^{d_v} \mathbf{f}_{v,j} \,\middle|\, \mathbf{f}_{x,j} : [0,1] \to \mathbb{H} \text{ are square integrable} \right\}.$$

Define $F_n : \mathcal{S}^{\mathbb{H}}(p) \to \mathbb{R}$ by

$$\begin{aligned}
F_n(\boldsymbol{\beta}) = \int_{[0,1]^d} & \sum_{\mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j} \sum_{\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j} \\
& \left\| \bigoplus_{j=1}^{d_x} D\mathbf{m}_{x,j}(x_j)(1) \odot \left(h_j^2 \int_0^1 t^2 K(t) dt \cdot \frac{\partial p(\mathbf{x}, \mathbf{u}, \mathbf{v}) / \partial x_j}{p(\mathbf{x}, \mathbf{u}, \mathbf{v})} \right) \right.
\end{aligned}$$

$$\oplus \bigoplus_{\mathbf{u}' \in \prod_{j=1}^{d_u} \mathcal{U}_j} (\mathbf{m}_{u,j}(u_j') \ominus \mathbf{m}_{u,j}(u_j)) \odot \left(L_{\lambda_j}(u_j, u_j') \cdot \frac{p(\mathbf{x}, \mathbf{u}, \mathbf{v})|_{u_j = u_j'}}{p(\mathbf{x}, \mathbf{u}, \mathbf{v})} \right)$$
$$\oplus \bigoplus_{\mathbf{v}' \in \prod_{j=1}^{d_v} \mathcal{V}_j} (\mathbf{m}_{v,j}(v_j') \ominus \mathbf{m}_{v,j}(v_j)) \odot \left(W_{s_j}(v_j, v_j') \cdot \frac{p(\mathbf{x}, \mathbf{u}, \mathbf{v})|_{v_j = v_j'}}{p(\mathbf{x}, \mathbf{u}, \mathbf{v})} \right)$$
$$\ominus \boldsymbol{\beta}(\mathbf{x}, \mathbf{u}, \mathbf{v}) \Big\|^2 p(\mathbf{x}, \mathbf{u}, \mathbf{v}) d\mathbf{x}.$$

Then, one may show that F_n is a strictly convex, continuous and Gateaux differentiable functional satisfying $F_n(\boldsymbol{\beta}) \to \infty$ as $\|\boldsymbol{\beta}\|_2 \to \infty$. Using this functional and arguing as in the proof of Theorem 1, we can conclude that there exists a tuple $(\breve{\boldsymbol{\Delta}}_{x,1}, \ldots, \breve{\boldsymbol{\Delta}}_{v,d_v})$ of $\mathbb{H}$-valued maps satisfying the system of equations

$$\begin{aligned}
\breve{\boldsymbol{\Delta}}_{x,j}(x_j) &= \boldsymbol{\Delta}_{x,j}^{\dagger}(x_j) \ominus \boldsymbol{\mu}_{x,+j}(x_j; \breve{\boldsymbol{\Delta}}_{x,+j}^{\mathrm{tup}}) \ominus \boldsymbol{\mu}_{x,j+}(x_j; \breve{\boldsymbol{\Delta}}_{x,j+}^{\mathrm{tup}}), \\
&\qquad 1 \le j \le d_x, \\
\breve{\boldsymbol{\Delta}}_{u,j}(u_j) &= \boldsymbol{\Delta}_{u,j}^{\dagger}(u_j) \ominus \boldsymbol{\mu}_{u,+j}(u_j; \breve{\boldsymbol{\Delta}}_{u,+j}^{\mathrm{tup}}) \ominus \boldsymbol{\mu}_{u,j+}(u_j; \breve{\boldsymbol{\Delta}}_{u,j+}^{\mathrm{tup}}), \\
&\qquad 1 \le j \le d_u, \\
\breve{\boldsymbol{\Delta}}_{v,j}(v_j) &= \boldsymbol{\Delta}_{v,j}^{\dagger}(v_j) \ominus \boldsymbol{\mu}_{v,+j}(v_j; \breve{\boldsymbol{\Delta}}_{v,+j}^{\mathrm{tup}}) \ominus \boldsymbol{\mu}_{v,j+}(v_j; \breve{\boldsymbol{\Delta}}_{v,j+}^{\mathrm{tup}}), \\
&\qquad 1 \le j \le d_v
\end{aligned} \tag{A.58}$$

and the constraints

$$\begin{aligned}
\int_0^1 \breve{\boldsymbol{\Delta}}_{x,j}(x_j) \odot p_{x,j}(x_j) dx_j &= \int_0^1 \boldsymbol{\delta}_{x,j}(x_j) \\
&\qquad \odot (h_j^2 \cdot p_{x,j}(x_j)) dx_j, \quad 1 \le j \le d_x, \\
\bigoplus_{u_j} \breve{\boldsymbol{\Delta}}_{u,j}(u_j) \odot p_{u,j}(u_j) &= \bigoplus_{u_j} \bigoplus_{u_j'} (\mathbf{m}_{u,j}(u_j') \ominus \mathbf{m}_{u,j}(u_j)) \\
&\qquad \odot (L_{\lambda_j}(u_j, u_j') \cdot p_{u,j}(u_j')), \quad 1 \le j \le d_u, \\
\bigoplus_{v_j} \breve{\boldsymbol{\Delta}}_{v,j}(v_j) \odot p_{v,j}(v_j) &= \bigoplus_{v_j} \bigoplus_{v_j'} (\mathbf{m}_{v,j}(v_j') \ominus \mathbf{m}_{u,j}(u_j)) \\
&\qquad \odot (W_{s_j}(v_j, v_j') \cdot p_{v,j}(v_j')), \quad 1 \le j \le d_v,
\end{aligned} \tag{A.59}$$

where $\boldsymbol{\mu}_{t,+j}(t_j; \breve{\boldsymbol{\Delta}}_{t,+j}^{\mathrm{tup}})$ and $\boldsymbol{\mu}_{t,j+}(t_j; \breve{\boldsymbol{\Delta}}_{t,j+}^{\mathrm{tup}})$ are defined as $\boldsymbol{\mu}_{t,+j}(t_j; \boldsymbol{\Delta}_{t,+j}^{\mathrm{tup}})$ and $\boldsymbol{\mu}_{t,j+}(t_j; \boldsymbol{\Delta}_{t,j+}^{\mathrm{tup}})$ with $\boldsymbol{\Delta}_{t,k}$ being replaced by $\breve{\boldsymbol{\Delta}}_{t,k}$ for all $t = x, u, v$ and k. Then by arguing as in the proof of Theorem 4, we get

$$\sup_{x_j \in [0,1]} \|\breve{\boldsymbol{\Delta}}_{x,j}\|, \max_{u_j} \|\breve{\boldsymbol{\Delta}}_{u,j}\|, \max_{v_j} \|\breve{\boldsymbol{\Delta}}_{v,j}\| = O(n^{-2/5}). \tag{A.60}$$

Recall the definitions of $\hat{\boldsymbol{\Delta}}_{x,j}, \hat{\boldsymbol{\Delta}}_{u,j}$ and $\hat{\boldsymbol{\Delta}}_{v,j}$ given in the proof of Theorem 4. Then, we have the following lemma.

Lemma 4. *Under the conditions of Theorem 5,* $\hat{\mathbf{\Delta}}_{x,j}(x_j) \ominus \breve{\mathbf{\Delta}}_{x,j}(x_j) = \mathbf{R}_{x,j}(x_j)$ *a.e. with respect to* Leb, $\hat{\mathbf{\Delta}}_{u,j}(u_j) \ominus \breve{\mathbf{\Delta}}_{u,j}(u_j) = \mathbf{R}_{u,j}(u_j)$ *for all* $u_j \in \mathcal{U}_j$ *and* $\hat{\mathbf{\Delta}}_{v,j}(v_j) \ominus \breve{\mathbf{\Delta}}_{v,j}(v_j) = \mathbf{R}_{v,j}(v_j)$ *for all* $v_j \in \mathcal{V}_j$, *where* $\mathbf{R}_{x,j}, \mathbf{R}_{u,j}$ *and* $\mathbf{R}_{v,j}$ *are maps satisfying* $\sup_{x\in[0,1]} \|\mathbf{R}_{x,j}(x_j)\| = o_p(n^{-2/5})$, $\max_{u_j} \|\mathbf{R}_{u,j}(u_j)\| = o_p(n^{-2/5})$ *and* $\max_{v_j} \|\mathbf{R}_{v,j}(v_j)\| = o_p(n^{-2/5})$.

Lemma 4 yields that, for a.e. fixed $x_j \in (0,1)$ with respect to Leb and for all u_j and v_j,

$$\begin{aligned} n^{2/5} \odot (\hat{\mathbf{m}}_{x,j}(x_j) \ominus \mathbf{m}_{x,j}(x_j)) &= n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{x,j}(x_j) \oplus (n^{2/5} h_j^2) \odot \mathbf{c}_j(x_j) \\ &\quad \oplus n^{2/5} \odot \breve{\mathbf{\Delta}}_{x,j}(x_j) \oplus o_p(1), \\ n^{2/5} \odot (\hat{\mathbf{m}}_{u,j}(u_j) \ominus \mathbf{m}_{u,j}(u_j)) &= n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{u,j}(u_j) \oplus n^{2/5} \odot \breve{\mathbf{\Delta}}_{u,j}(u_j) \\ &\quad \oplus o_p(1), \\ n^{2/5} \odot (\hat{\mathbf{m}}_{v,j}(v_j) \ominus \mathbf{m}_{v,j}(v_j)) &= n^{2/5} \odot \hat{\boldsymbol{\mu}}^A_{v,j}(v_j) \oplus n^{2/5} \odot \breve{\mathbf{\Delta}}_{v,j}(v_j) \\ &\quad \oplus o_p(1). \end{aligned} \tag{A.61}$$

We take the limits of both sides of equations at (A.58) and at (A.59) after multiplying them by $n^{2/5}$. Then, by using (A.60) and Proposition E.6 in [8], we may prove that

$$(\mathbf{\Delta}_{x,1}, \ldots, \mathbf{\Delta}_{v,d_v}) := (\lim_{n\to\infty} n^{2/5} \odot \breve{\mathbf{\Delta}}_{x,1}, \ldots, \lim_{n\to\infty} n^{2/5} \odot \breve{\mathbf{\Delta}}_{v,d_v})$$

satisfies (A.32) and (A.33). The uniqueness of the solution of (A.32) subject to (A.33) follows by arguing as in the proof of Theorem 1 and by Lemma 1. The desired asymptotic distributions now follow from (A.61), Lemma 3 and a version of Proposition 4.8 in [40] for strongly measuarble Gaussian elements.

A.7.5. Proof of Lemma 4

Here, we again give a sketch for the proof. Recall the definitions of $\tilde{\mathbf{r}}_{x,j}$, $\tilde{\mathbf{r}}_{u,j}$, $\tilde{\mathbf{r}}_{v,j}$, $\mathbf{\Delta}^\dagger_{x,j}$, $\mathbf{\Delta}^\dagger_{u,j}$ and $\mathbf{\Delta}^\dagger_{v,j}$ given in the proof of Theorem 4. Then by (A.50) and (A.51), it holds that

$$\begin{aligned} &\sup_{x_j\in[0,1]} \|\tilde{\mathbf{r}}_{x,j}(x_j)\|, \ \max_{u_j} \|\tilde{\mathbf{r}}_{u,j}(u_j)\|, \ \max_{v_j} \|\tilde{\mathbf{r}}_{v,j}(v_j)\| = o_p(n^{-2/5}), \\ &\sup_{x_j\in[0,1]} \|\mathbf{\Delta}^\dagger_{x,j}(x_j)\|, \ \max_{u_j} \|\mathbf{\Delta}^\dagger_{u,j}(u_j)\|, \ \max_{v_j} \|\mathbf{\Delta}^\dagger_{v,j}(v_j)\| = O_p(n^{-2/5}). \end{aligned} \tag{A.62}$$

Define the integral operators $\pi_{x,j}, \pi_{u,j}$ and $\pi_{v,j}$ by

$$\begin{aligned} &\pi_{x,j}(\mathbf{f})(\mathbf{x},\mathbf{u},\mathbf{v}) \\ &\quad = \int_{[0,1]^{d_x-1}} \bigoplus_{\mathbf{u}\in\prod_{j=1}^{d_u}\mathcal{U}_j} \bigoplus_{\mathbf{v}\in\prod_{j=1}^{d_v}\mathcal{V}_j} \mathbf{f}(\mathbf{x},\mathbf{u},\mathbf{v}) \odot (p(\mathbf{x},\mathbf{u},\mathbf{v})/p_{x,j}(x_j)) d\mathbf{x}_{-j}, \\ &\pi_{u,j}(\mathbf{f})(\mathbf{x},\mathbf{u},\mathbf{v}) \end{aligned}$$

$$= \int_{[0,1]^{d_x}} \bigoplus_{\mathbf{u}_{-j} \in \prod_{k \neq j} \mathcal{U}_k} \bigoplus_{\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j} \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{v}) \odot (p(\mathbf{x}, \mathbf{u}, \mathbf{v}) / p_{u,j}(u_j)) d\mathbf{x},$$

$$\pi_{v,j}(\mathbf{f})(\mathbf{x}, \mathbf{u}, \mathbf{v})$$

$$= \int_{[0,1]^{d_x}} \bigoplus_{\mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j} \bigoplus_{\mathbf{v}_{-j} \in \prod_{k \neq j} \mathcal{V}_k} \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{v}) \odot (p(\mathbf{x}, \mathbf{u}, \mathbf{v}) / p_{v,j}(v_j)) d\mathbf{x}.$$

Likewise, define the operators $\hat{\pi}_{x,j}, \hat{\pi}_{u,j}$ and $\hat{\pi}_{v,j}$ as $\pi_{x,j}, \pi_{u,j}$ and $\pi_{v,j}$ with $p_{x,j}, p_{u,j}, p_{v,j}$ and p being replaced by $\hat{p}_{x,j}, \hat{p}_{u,j}, \hat{p}_{v,j}$ and $\hat{p}$, respectively. Let $T = (I - \pi_{v,d_v}) \circ \cdots \circ (I - \pi_{x,1})$ and $\hat{T} = (I - \hat{\pi}_{v,d_v}) \circ \cdots \circ (I - \hat{\pi}_{x,1})$. Also, define

$$\begin{aligned}
\boldsymbol{\tau} &= \boldsymbol{\Delta}^{\dagger}_{v,d_v} \oplus (I - \pi_{v,d_v})(\boldsymbol{\Delta}^{\dagger}_{v,d_v - 1}) \oplus \cdots \oplus (I - \pi_{v,d_v}) \circ \cdots \circ (I - \pi_{x,2})(\boldsymbol{\Delta}^{\dagger}_{x,1}), \\
\hat{\boldsymbol{\tau}} &= \boldsymbol{\Delta}^{\dagger}_{v,d_v} \oplus (I - \hat{\pi}_{v,d_v})(\boldsymbol{\Delta}^{\dagger}_{v,d_v - 1}) \oplus \cdots \oplus (I - \hat{\pi}_{v,d_v}) \circ \cdots \circ (I - \hat{\pi}_{x,2})(\boldsymbol{\Delta}^{\dagger}_{x,1}), \\
\hat{\boldsymbol{\xi}} &= \boldsymbol{\Delta}^{\dagger}_{v,d_v} \oplus (I - \hat{\pi}_{v,d_v})(\tilde{\mathbf{r}}_{v,d_v - 1}) \oplus \cdots \oplus (I - \hat{\pi}_{v,d_v}) \circ \cdots \circ (I - \hat{\pi}_{x,2})(\tilde{\mathbf{r}}_{x,1}).
\end{aligned}$$

Recall the definition of $\hat{\boldsymbol{\Delta}}_{\oplus}$ given in the proof of Theorem 4. Define

$$\breve{\boldsymbol{\Delta}}_{\oplus} = \bigoplus_{j=1}^{d_x} \breve{\boldsymbol{\Delta}}_{x,j} \oplus \bigoplus_{j=1}^{d_u} \breve{\boldsymbol{\Delta}}_{u,j} \oplus \bigoplus_{j=1}^{d_v} \breve{\boldsymbol{\Delta}}_{v,j}.$$

Then, we have

$$\begin{aligned}
&\hat{\boldsymbol{\Delta}}_{\oplus} \ominus \breve{\boldsymbol{\Delta}}_{\oplus} \\
&= \bigoplus_{l=0}^{\infty} \hat{T}^l(\hat{\boldsymbol{\tau}} \oplus \hat{\boldsymbol{\xi}}) \ominus \bigoplus_{l=0}^{\infty} T^l(\boldsymbol{\tau}) \\
&= \hat{T}\Big(\bigoplus_{l=0}^{\infty} \hat{T}^l(\hat{\boldsymbol{\xi}})\Big) \oplus T\Big(\bigoplus_{l=0}^{\infty} T^l(\hat{\boldsymbol{\tau}} \ominus \boldsymbol{\tau})\Big) \oplus (\hat{T} - T)\Big(\bigoplus_{l=0}^{\infty} \hat{T}^l(\hat{\boldsymbol{\tau}})\Big) \qquad \text{(A.63)} \\
&\qquad \oplus T\Big(\bigoplus_{l=2}^{\infty} \bigoplus_{j=0}^{l-2} T^j \circ (\hat{T} - T) \circ \hat{T}^{l-2-j}(\hat{\boldsymbol{\tau}})\Big) \oplus \hat{\boldsymbol{\xi}} \oplus \hat{\boldsymbol{\tau}} \ominus \boldsymbol{\tau} \\
&=: \hat{T}(\hat{\boldsymbol{\eta}}_1) \oplus T(\hat{\boldsymbol{\eta}}_2) \oplus (\hat{T} - T)(\hat{\boldsymbol{\eta}}_3) \oplus T(\hat{\boldsymbol{\eta}}_4) \oplus \hat{\boldsymbol{\xi}} \oplus \hat{\boldsymbol{\tau}} \ominus \boldsymbol{\tau}
\end{aligned}$$

a.e. with respect to $P\mathbf{W}^{-1}$, where $\mathbf{W} = (\mathbf{X}, \mathbf{U}, \mathbf{V})$. Using (A.62) we may prove that

$$\begin{aligned}
\|\hat{\boldsymbol{\eta}}_1\|_2, \|\hat{\boldsymbol{\eta}}_2\|_2, \|\hat{\boldsymbol{\eta}}_4\|_2 &= o_p(n^{-2/5}), \\
\|\hat{\boldsymbol{\eta}}_3\|_2 &= O_p(n^{-2/5}), \\
\sup_{\mathbf{x},\mathbf{u},\mathbf{v}} \|\hat{\boldsymbol{\xi}}(\mathbf{x}, \mathbf{u}, \mathbf{v})\| &= o_p(n^{-2/5}), \qquad \text{(A.64)} \\
\sup_{\mathbf{x},\mathbf{u},\mathbf{v}} \|\hat{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{u}, \mathbf{v}) \ominus \boldsymbol{\tau}(\mathbf{x}, \mathbf{u}, \mathbf{v})\| &= o_p(n^{-2/5}).
\end{aligned}$$

Note that, for a given sum map $\boldsymbol{\eta} := \bigoplus_{j=1}^{d_x} \boldsymbol{\eta}_{x,j} \oplus \bigoplus_{j=1}^{d_u} \boldsymbol{\eta}_{u,j} \oplus \bigoplus_{j=1}^{d_v} \boldsymbol{\eta}_{v,j}$ with $\|\boldsymbol{\eta}\|_2 \leq 1$, there exist a constant $C^{\boldsymbol{\eta}} > 0$ and a map $\mathbf{r}^{\boldsymbol{\eta}} := \bigoplus_{j=1}^{d_x} \mathbf{r}^{\boldsymbol{\eta}}_{x,j} \oplus \bigoplus_{j=1}^{d_u} \mathbf{r}^{\boldsymbol{\eta}}_{u,j} \oplus \bigoplus_{j=1}^{d_v} \mathbf{r}^{\boldsymbol{\eta}}_{v,j}$ depending on $\boldsymbol{\eta}$ such that

$$\begin{aligned} \sup_{\mathbf{w}} \|T(\boldsymbol{\eta})(\mathbf{x},\mathbf{u},\mathbf{v})\| &\leq C^{\boldsymbol{\eta}}, \\ \|(\hat{T}-T)(\boldsymbol{\eta})(\mathbf{x},\mathbf{u},\mathbf{v})\| &\leq \|\mathbf{r}^{\boldsymbol{\eta}}(\mathbf{x},\mathbf{u},\mathbf{v})\|, \\ \sup_{x_j \in [0,1]} \|\mathbf{r}^{\boldsymbol{\eta}}_{x,j}(x_j)\| &= o_p(1), \\ \max_{u_j} \|\mathbf{r}^{\boldsymbol{\eta}}_{u,j}(u_j)\| &= o_p(1), \\ \max_{v_j} \|\mathbf{r}^{\boldsymbol{\eta}}_{v,j}(v_j)\| &= o_p(1). \end{aligned} \tag{A.65}$$

Combining (A.63), (A.64) and (A.65) gives that $\hat{\boldsymbol{\Delta}}_{\oplus} \ominus \breve{\boldsymbol{\Delta}}_{\oplus} = \mathbf{R}_{\oplus}$ a.e. with respect to $P\mathbf{W}^{-1}$, where $\mathbf{R}_{\oplus}$ is a stochastic map satisfying

$$\sup_{\mathbf{x},\mathbf{u},\mathbf{v}} \|\mathbf{R}_{\oplus}(\mathbf{x},\mathbf{u},\mathbf{v})\| = o_p(n^{-2/5}). \tag{A.66}$$

Considering the terms

$$\begin{aligned} &\int_{[0,1]^{d_x-1}} \bigoplus_{\mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j} \bigoplus_{\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j} \hat{\boldsymbol{\Delta}}_{\oplus}(\mathbf{x},\mathbf{u},\mathbf{v}) \ominus \breve{\boldsymbol{\Delta}}_{\oplus}(\mathbf{x},\mathbf{u},\mathbf{v}) d\mathbf{x}_{-j}, \\ &\int_{[0,1]^{d_x}} \bigoplus_{\mathbf{u}_{-j} \in \prod_{k \neq j} \mathcal{U}_k} \bigoplus_{\mathbf{v} \in \prod_{j=1}^{d_v} \mathcal{V}_j} \hat{\boldsymbol{\Delta}}_{\oplus}(\mathbf{x},\mathbf{u},\mathbf{v}) \ominus \breve{\boldsymbol{\Delta}}_{\oplus}(\mathbf{x},\mathbf{u},\mathbf{v}) d\mathbf{x}, \\ &\int_{[0,1]^{d_x}} \bigoplus_{\mathbf{u} \in \prod_{j=1}^{d_u} \mathcal{U}_j} \bigoplus_{\mathbf{v}_{-j} \in \prod_{k \neq j} \mathcal{V}_k} \hat{\boldsymbol{\Delta}}_{\oplus}(\mathbf{x},\mathbf{u},\mathbf{v}) \ominus \breve{\boldsymbol{\Delta}}_{\oplus}(\mathbf{x},\mathbf{u},\mathbf{v}) d\mathbf{x}, \end{aligned}$$

using (A.66) and applying a similar argument used for (A.56), we may establish the desired result.

A.8. Proof of Theorem 6

The theorem follows along the lines of the proof of Theorem 4.

Acknowledgments

The authors thank the editor and an associate editor for handling the paper nicely and three referees for giving constructive comments on the earlier version of the paper. The authors also thank Dr. Kyusang Yu for giving an advice on the implementation of the method of [45].

References

[1] Aitchison, J. and Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413-420. MR0443222

[2] Aliprantis, C. D. and Border, K. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag Berlin Heidelberg. MR2378491

[3] Ansley, C. F. and Kohn, R. (1994). Convergence of the backfitting algorithm for additive models. *Journal of the Australian Mathematical Society (Series A)*, **57**, 316-329. MR1297006

[4] Atkinson, K. and Han, W. (2009). *Theoretical Numerical Analysis*. Springer-Verlag New York. MR2511061

[5] Beltrami, E. J. (1967). On infinite-dimensional convex programs. *Journal of Computer and System Sciences*, **1**, 323-329. MR0232603

[6] Bosq, D. (2000). *Linear Processes in Function Spaces*. Springer-Verlag New York. MR1783138

[7] Buja, A. (1996). What criterion for a power algorithm? In: Rieder, H. (eds.) *Robust Statistics, Data Analysis, and Computer Intensive Methods*. Springer-Verlag New York. MR1491396

[8] Cohn, D. L. (2013). *Measure Theory*. Birkhäuser Basel. MR3098996

[9] Egozcue, J. J. and Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its structure. *Test*, **28**, 599-638. MR3992128

[10] Ferraty, F., Van Keilegom, I. and Vieu, P. (2012). Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, **109**, 10-28. MR2922850

[11] Han, K., Müller, H.-G. and Park, B. U. (2020). Additive functional regression for densities as responses. *Journal of the American Statistical Association*, **115**, 997-1010. MR4107695

[12] Han, K. and Park, B. U. (2018). Smooth backfitting for error-in-variables additive models. *Annals of Statistics*, **46**, 2216-2250. MR3845016

[13] Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **55**, 757–796. MR1229881

[14] Hron, K., Menafoglio, A., Templ, M., Hruzová, K. and Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics and Data Analysis*, **94**, 330–350. MR3412829

[15] Jeon, J. M. and Park, B. U. (2020). Additive regression with Hilbertian responses. *Annals of Statistics*, **48**, 2671-2697. MR4152117

[16] Koul, H., Susarla, V., and Van Ryzin J. (1981). Regression analysis with randomly right-censored data. *Annals of Statistics*, **9**, 1276-1288. MR0630110

[17] Kundu, S., Majumdar, S. and Mukherjee, K. (2000). Central limit theorems revisited. *Statistics and Probability Letters*, **47**, 265-275. MR1747487

[18] Lee, Y. K., Mammen, E. and Park, B. U. (2010). Backfitting and smooth backfitting for additive quantile models. *Annals of Statistics*, **38**, 2857-2883. MR2722458

[19] Lee, Y. K., Mammen, E. and Park, B. U. (2012a). Projection-type estimation for varying coefficient regression models. *Bernoulli*, **18**, 177-205. MR2888703

[20] Lee, Y. K., Mammen, E. and Park, B. U. (2012b). Flexible generalized varying coefficient regression models. *Annals of Statistics*, **40**, 1906-1933. MR3015048

[21] Lian, H. (2011). Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics*, **5**, 31-40. MR2773606

[22] Linton, O. B. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93–100. MR1332841

[23] Linton, O. B., Sperich, S., and Van Keilegon, I. (2008). Estimation of a semiparametric transformation model. *Annals of Statistics*, **36**, 686-718. MR2396812

[24] Lo, S.-H. and Singh, K. (1986). The product-limit estimator and the bootstrap: some asymptotic representations. *Probability Theory and Related Fields*, **71**, 455-465. MR0824714

[25] Mammen, E., Linton, O. B. and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, **27**, 1443-1490. MR1742496

[26] Mammen, E. and Park, B. U. (2006). A simple smooth backfitting method for additive models. *Annals of Statistics*, **34**, 2252-2271. MR2291499

[27] Marzio, M. D., Panzera, A. and Venieri, C. (2015). Non-parametric regression for compositional data. *Statistical Modelling*, **15**, 113-133. MR3325749

[28] Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, **17**, 571-599. MR1424907

[29] Nielsen, J. P. and Sperlich, S. (2005). Smooth backfitting in practices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 43-61. MR2136638

[30] Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, **73**, 166-179. MR1763322

[31] Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, **25**, 186-211. MR1429922

[32] Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications*, John Wiley and Sons, Ltd., Chichester. MR2920574

[33] Petersen, A. and Müller, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics*, **44**, 183-218. MR3449766

[34] Petersen, A. and Müller, H.-G. (2019). Wasserstein covariance for multiple random densities. *Biometrika*, **106**, 339-351. MR3949307

[35] Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer New York. MR1910407

[36] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **50**, 413-436. MR0970977

[37] Talská, R., Menafoglio, A., Machalová, J., Hron, K. and Fišerová, E. (2018). Compositional regression with functional response. *Computuational Statistics and Data Analysis*, **123**, 66-85. MR3777086

[38] Tsagris, M. (2015). Regression analysis with compositional data containing zero values. *Chilean Journal of Statistics*, **6**, 47-57. MR3407274

[39] van den Boogaart, K. G., Egozcue, J. J. and Pawlowsky-Glahn, V. (2014). Bayes Hilbert spaces. *Australian and New Zealand Journal of Statistics*, **56**, 171-194. MR3226435

[40] Van Neerven, J. (2008). *Stochastic evolution equations*. Lecture Notes of the 11th Internet Seminar, TU Delft OpenCourseWare, http://ocw.tudelft.nl.

[41] Wang, J.-G. (1987). A note on the uniform consistency of the Kaplan-Meier estimator. *Annals of Statistics*, **15**, 1313-1316. MR0902260

[42] Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257-295.

[43] Xia, Y. (2009). A note on the backfitting estimation of additive models. *Bernoulli*, **73**, 1148-1153. MR2597586

[44] Yang, S. J., El Ghouch, A. and Van Keilegom, I. (2014). Varying coefficient models having different smoothing variables with randomly censored data. *Electronic Journal of Statistics*, **8**, 226-252. MR3189554

[45] Yu, K., Mammen, E. and Park, B. U. (2011). Semi-parametric regression: Efficiency gains from modeling the nonparametric part. *Bernoulli*, **17**, 736-748. MR2787613

[46] Yu, K., Park, B. U. and Mammen, E. (2008). Smooth backfitting in generalized additive models. *Annals of Statistics*, **36**, 228-260. MR2387970