Statistics and Its Interface Volume 15 (2022) 129–149

Estimation of Hilbertian varying coefficient models*

Young Kyung Lee, Byeong U. Park†,
Hyerim Hong, and Dongwoo Kim

In this paper we discuss the estimation of a fairly general type of varying coefficient model. The model is for a response variable that takes values in a general Hilbert space and allows for various types of additive interaction terms in representing the effects of predictors. It also accommodates both continuous and discrete predictors. We develop a powerful technique of estimating the very general model. Our approach may be used in a variety of situations where one needs to analyze the relation between a set of predictors and a Hilbertian response. We prove the existence of the estimators of the model itself and of its components, and also the convergence of a backfitting algorithm that realizes the estimators. We derive the rates of convergence of the estimators and their asymptotic distributions. We also demonstrate via simulation study that our approach works efficiently, and illustrate its usefulness through a real data application.

Keywords and phrases: Hilbertian response, Varying coefficient model, Additive regression, Smooth backfitting, Compact operator.

1. INTRODUCTION

One of the main issues in nonparametric regression is the curse of dimensionality, which one faces when one deals with multivariate predictor. The difficulties are that, theoretically, the estimation accuracy deteriorates rapidly as the dimension of predictor increases and that, in practice, the implementation of the method gets easily in jeopardy due to sparsity of data points on the domain of the predictor. Structured nonparametric regression is a useful option of avoiding the curse of dimensionality. Among various structured models, the simplest one is additive model. A powerful technique of estimating additive model, called smooth backfitting, was proposed and studied by Mammen et al. (1999). The idea has been developed further for generalized additive model by Yu et al. (2008), for additive quantile model by Lee et al. (2010), and for errors-in-variables additive model by Han and Park (2018). These are all for Euclidean responses. Recently, analysis of non-Euclidean data has been one of the main focuses in statistics. Among others, Han et al. (2020) developed the idea of smooth backfitting for the case where the response variable is a random density. The space of density functions is an example of Hilbert space. Jeon and Park (2020) extended the work of Mammen et al. (1999), in full generality, to the case where the response variable takes values in a general Hilbert space.

Although additive modeling has a number of advantages, evidenced by Jeon and Park (2020) for Hilbertian response and by the aforementioned earlier works for Euclidean response, it may not accommodate discrete predictors. Missing important predictors results in inferior prediction accuracy. An example is given in Section 4.3, where we build up a varying coefficient model that predicts household electricity consumption trajectory for 24 hours based on the information of the two continuous (temperature, cloudiness) and one discrete (an indicator of weekday or weekend) predictors. Comparing with the additive regression approach by Jeon and Park (2020) that used only the information of the two continuous predictors, our approach improved the prediction accuracy by a factor of 13.5%. The improvement is demonstrated by Figure 1 that depicts the prediction results for six randomly chosen months.

In this paper, we study the estimation of a general type of varying coefficient models with *Hilbertian* responses, which includes as a special case the model used to predict electricity consumption. Varying coefficient model, proposed initially by Hastie and Tibshirani (1993), is known to be another important structured model with which one may deal with discrete predictors. Some of the main developments for this type of model with *Euclidean* response include Yang et al. (2006) and Lee et al. (2012a,b). For a comprehensive review, see Park et al. (2015) and the references therein. A varying coefficient model consists of additive terms that are of the form $X_k \cdot f_{j,k}(X_j)$ for some predictors and coefficient functions $f_{j,k}$. The simplest case is that the whole set of predictors is divided into two groups, one group for 'linear part' taking the role of X_k in $X_k \cdot f_{j,k}(X_j)$, and the other for 'nonlinear part' taking the role of X_j in $X_k \cdot f_{j,k}(X_j)$, and each predictor in the linear group is paired up with one in the nonlinear group to form an additive term. Specifically, with two types of predictors $\mathbf{X} = (X_1, \ldots, X_{d_0})$ and

*Research of Young Kyung Lee was supported by a research grant of Kangwon National University in 2020. Research of Byeong U. Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2019R1A2C3007355).
†Corresponding author.

$\mathbf{Z} = (Z_1, \ldots, Z_{d_0})$ it is to postulate the model $\mathrm{E}(Y|\mathbf{X}, \mathbf{Z}) = Z_1 \cdot f_1(X_1) + \cdots + Z_{d_0} \cdot f_{d_0}(X_{d_0})$. This is the model studied by Yang et al. (2006) and Lee et al. (2012a) in the case of Euclidean response. One of the main difficulties with this formulation is that the total number of predictors should be even and one needs to decide which predictors to put in the linear part and which in the nonlinear part. Another limitation is that it does not allow a predictor in an additive term to appear in the linear or nonlinear part of another additive term. This means that each predictor is paired up with only one other predictor, which requires us to determine a single set of pairs for the implementation of the model. The work of Lee et al. (2012b) completely removed these limitations in the Euclidean case by allowing a predictor to enter any number of additive terms, i.e., allowing X_j and X_k in an additive term $X_k \cdot f_{j,k}(X_j)$ to be identical either to $X_{j'}$ or to $X_{k'}$ in another additive term $X_{k'} \cdot f_{j',k'}(X_{j'})$. Thus, the general formulation includes the model studied by Yang et al. (2006) and Lee et al. (2012a) as a special case, and also the full model that involves all possible interaction terms $X_k \cdot f_{j,k}(X_j)$, $j \neq k$ for a given set of predictors $\{X_1, \ldots, X_d\}$.

In the present paper we consider the general formulation of varying coefficient model proposed in Lee et al. (2012b), but for general Hilbertian response. The general framework enables us to analyze the electricity consumption data with the discrete predictor indicating weekday or weekend. We note that the extension of the work, Lee et al. (2012a), to the Hilbertian case, i.e., the simplest varying coefficient model at (2.2), is rather straightforward. However, the extension in the general formulation of Lee et al. (2012b) is challenging since the theory in the identification and the estimation of individual component maps in conjunction with Hilbertian vector operation is more complex than the Euclidean case. In fact, the technical details of the theory for the Euclidean case are missing in Lee et al. (2012b). Furthermore, we discuss the estimation of the parametric part that arises from implementing a set of constraints on the component maps, and also give a full account of its influence on the accuracy of the estimators of the normalized nonparametric component maps, which were largely neglected in Lee et al. (2012b). Thus, the present work enhances the theory of Lee et al. (2012b) in several important respects.

In the next section, we describe the methodology. In Section 3 we present the theory, which includes the existence of the estimators, the convergence of a backfitting algorithm that realizes the estimators, the rates of convergence and the asymptotic distributions of the estimators. In Section 4 we report the results of a simulation study and a real data application. Appendices A and B are for technical details.

2. METHODOLOGY

2.1 Some terminologies

We denote by $\mathbb{H}$ the Hilbert space where the response variable $\mathbf{Y}$ takes values. Let $\oplus$ and $\odot$, respectively, be the addition operation and the scalar multiplication for $\mathbb{H}$. For some examples of Hilbert space and the associated vector operations, see Jeon and Park (2020). We denote the zero vector by $\mathbf{0}$.

We introduce several conventions in Hilbertian vector operations. This is for simplicity of presentation. For $c_j \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{H}$, we write $c_1 \cdot c_2 \odot \mathbf{v}$ for $c_1 \odot (c_2 \odot \mathbf{v})$, $c_1 \cdot c_2 \cdot c_3 \odot \mathbf{v}$ for $c_1 \odot (c_2 \odot (c_3 \odot \mathbf{v}))$ and so on. We let $\ominus$ denote the subtraction operation defined by $\mathbf{v}_1 \ominus \mathbf{v}_2 = \mathbf{v}_1 \oplus (-1 \odot \mathbf{v}_2)$ for $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{H}$. For a tuple of k Hilbertian values, $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_k)^\top \in \mathbb{H}^k$ and for $\mathbf{c} = (c_1, \ldots, c_k)^\top \in \mathbb{R}^k$, we write $\mathbf{c}^\top \odot \mathbf{v}$ for $\bigoplus_{j=1}^k (c_j \odot \mathbf{v}_j)$. Likewise, for an $l \times k$ real matrix $\mathbf{A} = (a_{ij})$, we write $\mathbf{A} \odot \mathbf{v}$ for $\left(\bigoplus_{j=1}^k a_{1j} \odot \mathbf{v}_j, \ldots, \bigoplus_{j=1}^k a_{lj} \odot \mathbf{v}_j\right)^\top$. We let $\mathbf{0}_k$ denote $(\mathbf{0}, \ldots, \mathbf{0})^\top \in \mathbb{H}^k$. For $\mathbf{u} \in \mathbb{H}$ and $\mathbf{c} = (c_1, \ldots, c_k)^\top \in \mathbb{R}^k$, we write $\mathbf{c} \odot \mathbf{u}$ to indicate $(c_1 \odot \mathbf{u}, \ldots, c_k \odot \mathbf{u})^\top \in \mathbb{H}^k$. For $c \in \mathbb{R}$ and $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_k)^\top$, we write $c \odot \mathbf{v}$ for $(c \odot \mathbf{v}_1, \ldots, c \odot \mathbf{v}_k)^\top$. W also extend the addition and subtraction operations to tuples of Hilbertian values. For example, for $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_k)^\top \in \mathbb{H}^k$ and $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_k)^\top \in \mathbb{H}^k$,

$$\mathbf{v} \oplus \mathbf{w} = (\mathbf{v}_1 \oplus \mathbf{w}_1, \ldots, \mathbf{v}_k \oplus \mathbf{w}_k)^\top,$$
$$\mathbf{v} \ominus \mathbf{w} = (\mathbf{v}_1 \ominus \mathbf{w}_1, \ldots, \mathbf{v}_k \ominus \mathbf{w}_k)^\top.$$

We let $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ be an inner product of $\mathbb{H}$. Let $\| \cdot \|_{\mathbb{H}}$ be the associated norm defined by $\|\mathbf{v}\|_{\mathbb{H}} = \langle \mathbf{v}, \mathbf{v} \rangle_{\mathbb{H}}$ for $\mathbf{v} \in \mathbb{H}$. For $\mathbb{H}^k$, we define $\langle \cdot, \cdot \rangle_{\mathbb{H}^k}$ by

$$\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{H}^k} = \sum_{j=1}^k \langle \mathbf{v}_j, \mathbf{w}_j \rangle_{\mathbb{H}}$$

for $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_k)^\top$, $\mathbf{w} = (\mathbf{w}_1, \ldots \mathbf{w}_k)^\top$ and the associated norm $\| \cdot \|_{\mathbb{H}^k}$ by $\|\mathbf{v}\|_{\mathbb{H}^k}^2 = \sum_{j=1}^k \|\mathbf{v}_j\|^2$. Our methodology involves integration of Hilbertian maps taking values in $\mathbb{H}^k$. We adopt the notion of *Bochner integral* that generalizes the conventional Lebesgue integral to functions that take values in a Hilbert space. The statistical properties of Bochner integral were well studied by Jeon and Park (2020), see also Cohn (2013). Below throughout this paper, we do not distinguish in notation between Lebesgue integral and Bochner integral. It should be understood as Bochner integral in case the integrand is a map taking values in $\mathbb{H}^k$ for some $k \geq 1$. Finally, for an index set $\mathcal{I}$, we let $|\mathcal{I}|$ denote its cardinality.

2.2 The model

Suppose that we have d predictors, $X_1, \ldots, X_d$. Let $\mathbf{Y}$ be a response that takes values in a separable Hilbert space $\mathbb{H}$. Write $\mathbf{X} = (X_1, \ldots, X_d)$. We assume that $\mathrm{E}(\mathbf{Y}|\mathbf{X})$ equals the sum of the terms, $X_k \odot \mathbf{f}_{j,k}(X_j)$, over (j,k) in a subset of $\{(j,k) : 1 \leq j \neq k \leq d\}$, where $\mathbf{f}_{j,k} : \mathbb{R} \to \mathbb{H}$ are unknown, termed (Hilbertian) *component maps*. Each additive term $X_k \odot \mathbf{f}_{j,k}(X_j)$ is an interaction of the 'linear' effect of X_k and the 'nonlinear' effect of X_j. We say that X_k and X_j

in $X_k \odot \mathbf{f}_{j,k}(X_j)$ are in linear and nonlinear parts, respectively. Among the d predictors, let $X_1, \ldots, X_{d_0}$ ($d_0 \le d$) be the collection of continuous-type predictors that enter some nonlinear parts. The case $d_0 = d$ means that all predictors are in nonlinear parts. We allow some of $X_1, \ldots, X_{d_0}$ to appear in linear parts as well. Without loss of generality, we assume that they are $X_{d_0-r+1}, \ldots, X_{d_0}$ with $0 \le r \le d_0$. Here, $r = 0$ means that there is no predictor that appears in both linear and nonlinear parts, which is the case with the real data example we discussed in Section 1 and will treat in Section 4.3. The case $r = d_0$ corresponds to the situation where all predictors in nonlinear parts also appear in linear parts. The remaining predictors, $X_{d_0+1}, \ldots, X_d$ appear only in linear parts and they are either continuous- or discrete-type. In this formulation, X_j ($1 \le j \le d_0$) are in nonlinear parts and of continuous-type, X_j ($d_0 - r + 1 \le j \le d$) are in linear parts, and X_j ($d_0 - r + 1 \le j \le d_0$) are in both.

For $1 \le j \le d_0$, let I_j be the set of indices k such that X_k is in a linear part and interacts with X_j. The index set I_j is a subset of $\{d_0 - r + 1, \ldots, d_0, d_0 + 1, \ldots, d\} \setminus \{j\}$. Define

$$\mathbf{Z}_j = (X_k : k \in I_j), \quad \mathbf{f}_j = (\mathbf{f}_{j,k} : k \in I_j), \quad 1 \le j \le d_0.$$

The model we study in this paper assumes

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \bigoplus_{j=1}^{d_0} \mathbf{Z}_j^\top \odot \mathbf{f}_j(X_j). \tag{2.1}$$

The above model is general to include all types of varying coefficient models. The case with $r = 0$, $I_j = \{d_0 + j\}$ and $d = 2d_0$ corresponds to the simplest model,

$$\begin{aligned}\mathrm{E}(\mathbf{Y}|\mathbf{X}) = {} & X_{d_0+1} \odot \mathbf{f}_{1,d_0+1}(X_1) \\ & \oplus \cdots \oplus X_{2d_0} \odot \mathbf{f}_{d_0,2d_0}(X_{d_0}),\end{aligned} \tag{2.2}$$

which was studied in Lee et al. (2012a) for scalar responses $\mathbf{Y} = Y$. Our general framework also includes the following model, which corresponds to the special case where $r = 0$ and $I_j \equiv \{d_0 + 1, \ldots, d\}$ for all $1 \le j \le d_0$:

$$\begin{aligned}\mathrm{E}(\mathbf{Y}|\mathbf{X}) = {} & X_{d_0+1} \odot \Big(\bigoplus_{j=1}^{d_0} \mathbf{f}_{j,d_0+1}(X_j) \Big) \\ & \oplus \cdots \oplus X_d \odot \Big(\bigoplus_{j=1}^{d_0} \mathbf{f}_{j,d}(X_j) \Big).\end{aligned} \tag{2.3}$$

In the above case, $\mathbf{Z}_j \equiv (X_{d_0+1}, \ldots, X_d)$ for all $1 \le j \le d_0$. We note that it is the model we used in the real data application discussed in Section 1 where all entries in $\mathbf{Z}_j$ are discrete-type random variables. In the case where $r = d_0$ and $I_j = \{1, \ldots, d\} \setminus \{j\}$ for $1 \le j \le d_0$, the model (2.1) is reduced to

$$\begin{aligned}\mathrm{E}(\mathbf{Y}|\mathbf{X}) = {} & \bigoplus_{k=1}^{d_0} X_k \odot \Big(\bigoplus_{j=1, \ne k}^{d_0} \mathbf{f}_{j,k}(X_j) \Big) \\ & \oplus \bigoplus_{k=d_0+1}^{d} X_k \odot \Big(\bigoplus_{j=1}^{d_0} \mathbf{f}_{j,k}(X_j) \Big).\end{aligned} \tag{2.4}$$

In the above case, $\mathbf{Z}_j = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_d)$ for $1 \le j \le d_0$.

The component maps $\mathbf{f}_{j,k}$ in the model (2.1) are not identifiable in case $\tilde{I}_k = \{j : 1 \le j \le d_0, I_j \ni k\}$ are not singletons for some $d_0 - r + 1 \le k \le d$. To see why, we rewrite the model (2.1) as

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \bigoplus_{k=d_0-r+1}^{d} X_k \odot \Big(\bigoplus_{j \in \tilde{I}_k} \mathbf{f}_{j,k}(X_j) \Big).$$

For each $d_0 - r + 1 \le k \le d$, the components $\mathbf{f}_{j,k}$ for $j \in \tilde{I}_k$ are not identified with $\mathbf{f}_{j,k} \oplus \mathbf{a}$ for a constant $\mathbf{a} \in \mathbb{H}$. For those k with $d_0 - r + 1 \le k \le d_0$, in particular, X_k in $X_k \odot \mathbf{f}_{j,k}(X_j)$ for some $j \in \tilde{I}_k$ may also appear in another additive term $X_j \odot \mathbf{f}_{k,j}(X_k)$. Note that $X_k \odot [\mathbf{f}_{j,k}(X_j) \oplus X_j \odot \mathbf{a}] \oplus X_j \odot [\mathbf{f}_{k,j}(X_k) \ominus X_k \odot \mathbf{a}] = X_k \odot \mathbf{f}_{j,k}(X_j) \oplus X_j \odot \mathbf{f}_{k,j}(X_j)$. Thus, for $d_0 - r + 1 \le j \ne k \le d_0$ such that $k \in I_j$ and $j \in I_k$, $\mathbf{f}_{j,k}$ and $\mathbf{f}_{k,j}$ in $X_k \odot \mathbf{f}_{j,k}(X_j)$ and $X_j \odot \mathbf{f}_{k,j}(X_k)$, respectively, are identifiable only up to an additive linear term. To identify the component maps and estimate them, we put the following constraints on $\mathbf{f}_{j,k}$.

$$\begin{aligned}&\int_0^1 w_j(x_j) \odot \mathbf{f}_{j,k}(x_j)\, dx_j = \mathbf{0}, \\ &\qquad j \in \tilde{I}_k,\ d_0 - r + 1 \le k \le d, \\ &\int_0^1 x_j w_j(x_j) \odot \mathbf{f}_{j,k}(x_j)\, dx_j = \mathbf{0}, \\ &\qquad j \in \tilde{I}_k,\ d_0 - r + 1 \le k \le d_0,\end{aligned} \tag{2.5}$$

where $w_j : \mathbb{R} \to [0, \infty)$ are some nonnegative weight functions and the integrals are in the Bochner sense.

Let $J = \{(j,k) : j \in \tilde{I}_k,\ d_0 - r + 1 \le j < k \le d_0\}$. With the constraints at (2.5) we may rewrite the model (2.1) as

$$\begin{aligned}\mathrm{E}(\mathbf{Y}|\mathbf{X}) = {} & \Big(\bigoplus_{k=d_0-r+1}^{d} X_k \odot \boldsymbol{\alpha}_{+,k}^0 \Big) \\ & \oplus \Big(\bigoplus_{(j,k) \in J} X_j X_k \odot \boldsymbol{\alpha}_{j,k}^1 \Big) \\ & \oplus \Big(\bigoplus_{j=1}^{d_0} \bigoplus_{k \in I_j} X_k \odot \mathbf{f}_{j,k}(X_j) \Big),\end{aligned} \tag{2.6}$$

where $\boldsymbol{\alpha}_{+,k}^0$ and $\boldsymbol{\alpha}_{j,k}^1$ are unknown constants in $\mathbb{H}$. Let $P_{\mathbf{X}}$ denote the distribution of $\mathbf{X}$ and P_{X_j} the marginal distribution of X_j. For continuous-type predictors X_j, let p_j be

the density of P_{X_j} with respect to Lebesgue measure. For discrete-type predictors X_j, we assume that P_{X_j} have finite support. Let

$$\|\boldsymbol{\alpha}\|_*^2 = \int \Big\| \bigoplus_{d_0-r+1\le k\le d} x_k \odot \boldsymbol{\alpha}^0_{+,k} \oplus \bigoplus_{(j,k)\in J} x_j x_k \odot \boldsymbol{\alpha}^1_{j,k} \Big\|_{\mathbb{H}}^2 P_{\mathbf{X}}(d\mathbf{x}).$$

Under the following conditions, the Hilbertian maps $\mathbf{f}_{j,k}$ in (2.6) are identifiable, as demonstrated in Proposition 1 below.

(A0) The product measure $P_{X_1} \times \cdots \times P_{X_d}$ has a density with respect to the joint probability measure $P_{\mathbf{X}}$ and the density is bounded away from zero and infinity on the support of $P_{\mathbf{X}}$. The marginal distributions are absolutely continuous with respect to Lebesgue measure and their densities are supported on bounded intervals, or they are discrete measure with finite support. The marginal densities p_j for $1 \le j \le d_0$ satisfy that w_j/p_j are bounded away from zero and infinity on the support of the respective p_j.

(A1) The smallest eigenvalues of $\mathrm{E}(\mathbf{Z}_j\mathbf{Z}_j^\top \mid X_j = x_j)\cdot p_j(x_j)$ for $1 \le j \le d_0$ are bounded away from zero on the support of the respective p_j.

Proposition 1. *Assume that the conditions (A0) and (A1) hold. For a set of Hilbertian constants* $\mathbf{a}^0_{+,k}, \mathbf{a}^1_{j,k} \in \mathbb{H}$ *and Hilbertian maps* $\mathbf{g}_{j,k} : \mathbb{R} \to \mathbb{H}$ *satisfying the constraints (2.5), let*

$$\boldsymbol{\mu}(\mathbf{x}) = \Big(\bigoplus_{k=d_0-r+1}^{d} x_k \odot \mathbf{a}^0_{+,k} \Big) \oplus \Big(\bigoplus_{(j,k)\in J} x_j x_k \odot \mathbf{a}^1_{j,k} \Big) \oplus \Big(\bigoplus_{j=1}^{d_0} \bigoplus_{k\in I_j} x_k \odot \mathbf{g}_{j,k}(x_j) \Big).$$

Then, there exist constants $0 < c < C < \infty$ *such that*

$$c \int \|\boldsymbol{\mu}(\mathbf{x})\|_{\mathbb{H}}^2 P_{\mathbf{X}}(d\mathbf{x}) \le \|\mathbf{a}\|_*^2 + \sum_{j=1}^{d_0} \sum_{k\in I_j} \|\mathbf{g}_{j,k}(x_j)\|_{\mathbb{H}}^2 p_j(x_j)\, dx_j \le C \int \|\boldsymbol{\mu}(\mathbf{x})\|_{\mathbb{H}}^2 P_{\mathbf{X}}(d\mathbf{x}).$$

To see that the above proposition implies the identification of $\mathbf{f}_{j,k}$, suppose that the true regression map $\mathbf{m} = \mathrm{E}(\mathbf{Y}|\mathbf{X} = \cdot)$ admits two representations as at (2.6), one with $\boldsymbol{\alpha}^0_{+,k}, \boldsymbol{\alpha}^1_{j,k}, \mathbf{f}_{j,k}$ and the other with $\tilde{\boldsymbol{\alpha}}^0_{+,k}, \tilde{\boldsymbol{\alpha}}^1_{j,k}, \tilde{\mathbf{f}}_{j,k}$. Then, Proposition 1 ensures

$$\|\boldsymbol{\alpha} \ominus \tilde{\boldsymbol{\alpha}}\|_*^2 + \sum_{j=1}^{d_0} \sum_{k\in I_j} \|\mathbf{f}_{j,k}(x_j) \ominus \tilde{\mathbf{f}}_{j,k}(x_j)\|_{\mathbb{H}}^2 p_j(x_j)\, dx_j \le C \int \|\mathbf{0}\|_{\mathbb{H}}^2 P_{\mathbf{X}}(d\mathbf{x}) = 0,$$

which implies that $\mathbf{f}_{j,k} \equiv \tilde{\mathbf{f}}_{j,k}$ for all (j,k) with $k \in I_j$ and $1 \le j \le d_0$.

2.3 Estimation of the model

Throughout this paper, we assume that all continuous-type predictors in the nonlinear parts (X_j with $1 \le j \le d_0$) have compact supports. Without loss of generality, we assume that they are supported on $[0,1]$. We note that the constraints (2.5) on $\mathbf{f}_{j,k}$ are needed only in case we want to identify and estimate $\mathbf{f}_{j,k}$. They are not needed for estimating the *regression map* $\mathbf{m} := \mathrm{E}(\mathbf{Y}|\mathbf{X} = \cdot)$. In the description of the methodology here and its theory in Section 3 we neglect the parametric part for simplicity of presentation. In practice, one may simply replace $\mathbf{Y}_i$ in the description given below by

$$\mathbf{Y}_i(\hat{\boldsymbol{\alpha}}) = \mathbf{Y}_i \ominus \Big(\bigoplus_{k=d_0-r+1}^{d} X_{ik} \odot \hat{\boldsymbol{\alpha}}^0_{+,k} \Big) \ominus \Big(\bigoplus_{(j,k)\in J} X_{ij}X_{ik} \odot \hat{\boldsymbol{\alpha}}^1_{j,k} \Big)$$

and applies (2.5) in case one estimates $\mathbf{f}_{j,k}$, where $\hat{\boldsymbol{\alpha}}$'s are suitable estimators of $\boldsymbol{\alpha}$'s. In fact, the effect of estimating the parameters $\boldsymbol{\alpha}$'s on the estimation of normalized $\mathbf{f}_{j,k}$ can be made negligible. In the Appendix B, we give a full account of the issues in estimating the parametric part. Below, we first describe our method of estimating $\mathbf{m}$ and then present a way of implementing the constraints (2.5) that gives estimators of the individual $\mathbf{f}_{j,k}$. In the representation $\mathbf{m}(\mathbf{x}) = \bigoplus_{j=1}^{d_0} \mathbf{z}_j^\top \odot \mathbf{f}_j(x_j)$, we suppress the dependence of $\mathbf{z}_j = (x_k : k \in I_j)$ on $\mathbf{x}$, and likewise the dependence of $\mathbf{Z}_{ij} = (X_{ik} : k \in I_j)$ on $\mathbf{X}_i$ as well, here and throughout the paper. Recall $\mathbf{f}_j = \big(\mathbf{f}_{j,k} : k \in I_j\big)^\top$, where $\mathbf{f}_{j,k}$ in $\mathbf{f}_j$ is enumerated in the same order as x_k in $\mathbf{z}_j$.

We use normalized kernel functions of the form

$$K_h(u,v) = \frac{K_h(u-v)}{\int_0^1 K_h(w-v)\,dw}, \quad u, v \in [0,1], \tag{2.7}$$

where $K_h(u-v) = h^{-1}K((u-v)/h)$, h is a bandwidth and $K : \mathbb{R} \to [0,\infty)$ is a baseline kernel function. This type of normalized kernels has been used in the smooth backfitting (SBF) literature. Throughout this paper we assume that K is bounded, symmetric, Lipschitz continuous, vanishing on $\mathbb{R} \setminus (-1,1)$ and positive on $(-1,1)$. The normalized kernels have the property that $\int_0^1 K_h(u,v)\,du = 1$ for all $v \in [0,1]$.

We estimate $\mathbf{m}(\mathbf{x})$ by $\hat{\mathbf{m}}(\mathbf{x}) = \bigoplus_{j=1}^{d_0} \mathbf{z}_j^\top \odot \hat{\mathbf{f}}_j(x_j)$ where the tuple $(\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_{d_0})$ minimizes

$$\int_{[0,1]^{d_0}} n^{-1} \sum_{i=1}^{n} \Big\| \mathbf{Y}_i \ominus \bigoplus_{j=1}^{d_0} \mathbf{Z}_{ij}^\top \odot \mathbf{g}_j(x_j) \Big\|_{\mathbb{H}}^2 \cdot \prod_{j=1}^{d_0} K_{h_j}(x_j, X_{ij})\, d\mathbf{x}_{d_0} \tag{2.8}$$

over $(\mathbf{g}_1, \ldots, \mathbf{g}_{d_0})$ in an appropriate function class, where $\mathbf{x}_{d_0} = (x_1, \ldots, x_{d_0})$. The function class over which we minimize (2.8) is the space of tuples of Hilbertian maps $(\mathbf{g}_1, \ldots, \mathbf{g}_{d_0})$ with $\mathbf{g}_j : [0,1] \to \mathbb{H}^{|I_j|}$.

By considering the Fréchet derivative of the objective functional at (2.8) on the function class, we may see that the minimizer of (2.8) satisfies

$$\begin{aligned} &\int_{[0,1]^{d_0-1}} \bigoplus_{i=1}^{n} n^{-1} \cdot \mathbf{Z}_{ij} \cdot \prod_{l=1}^{d_0} K_{h_l}(x_l, X_{il}) \\ &\qquad \odot \Big(\mathbf{Y}_i \ominus \bigoplus_{l=1}^{d_0} \mathbf{Z}_{il}^\top \odot \hat{\mathbf{f}}_l(x_l) \Big) d\mathbf{x}_{-j} \\ &= \mathbf{0}_{|I_j|} \quad \text{for } x_j \text{ a.e. on } [0,1], \quad 1 \le j \le d_0, \end{aligned} \tag{2.9}$$

where $\mathbf{x}_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{d_0})$. In the above integration, the integrand is a Hilbertian map: $[0,1]^{d_0-1} \to \mathbb{H}^{|I_j|}$ and thus the integral is in the Bochner sense. Define $|I_j| \times |I_j|$ and $|I_j| \times |I_l|$ real matrices

$$\begin{aligned} \hat{\mathbf{M}}_{jj}(x_j) &= n^{-1} \sum_{i=1}^{n} \mathbf{Z}_{ij} \mathbf{Z}_{ij}^\top K_{h_j}(x_j, X_{ij}), \\ &\qquad 1 \le j \le d_0, \\ \hat{\mathbf{M}}_{jl}(x_j, x_l) &= n^{-1} \sum_{i=1}^{n} \mathbf{Z}_{ij} \mathbf{Z}_{il}^\top K_{h_j}(x_j, X_{ij}) \\ &\qquad \times K_{h_l}(x_l, X_{il}), \ 1 \le j \ne l \le d_0, \end{aligned} \tag{2.10}$$

respectively. Also, define

$$\hat{\boldsymbol{\nu}}_j(x_j) = \hat{\mathbf{M}}_{jj}(x_j)^{-1} \cdot n^{-1} \odot \Big(\bigoplus_{i=1}^{n} \mathbf{Z}_{ij} \cdot K_{h_j}(x_j, X_{ij}) \odot \mathbf{Y}_i \Big), \quad 1 \le j \le d_0. \tag{2.11}$$

The above $\hat{\boldsymbol{\nu}}_j : [0,1] \to \mathbb{H}^{|I_j|}$ is nothing else than an estimator of the *marginal regression map* $\boldsymbol{\nu}_j(\cdot) := \mathrm{E}(\mathbf{Z}_j \mathbf{Z}_j^\top \,|\, X_j = \cdot)^{-1} \odot \mathrm{E}(\mathbf{Z}_j \odot \mathbf{Y} \,|\, X_j = \cdot)$, which minimizes $\mathrm{E}\|\mathbf{Y} \ominus \mathbf{Z}_j^\top \odot \boldsymbol{\nu}_j(X_j)\|_{\mathbb{H}}^2$. From the normalization property that $\int_0^1 K_h(u, v)\, du = 1$ for all $v \in [0,1]$, the system of equations (2.9) is equivalent to

$$\hat{\mathbf{f}}_j(x_j) = \hat{\boldsymbol{\nu}}_j(x_j) \ominus \bigoplus_{l=1:\ne j}^{d_0} \int_0^1 \hat{\mathbf{M}}_{jj}(x_j)^{-1} \cdot \hat{\mathbf{M}}_{jl}(x_j, x_l) \odot \hat{\mathbf{f}}_l(x_l)\, dx_l, \quad 1 \le j \le d_0. \tag{2.12}$$

We note that the system of equations at (2.12) defines only an estimator of the form $\hat{\mathbf{m}}(\mathbf{x}) = \bigoplus_{j=1}^{d_0} \mathbf{z}_j^\top \odot \hat{\mathbf{f}}_j(x_j)$ without specifying individual components $\hat{\mathbf{f}}_{j,k}$. In Section 3.2 we will show that a solution $\hat{\mathbf{m}}$ of the system of equations (2.12) exists and is unique under mild conditions. To identify the individual components satisfying the constraints (2.5), define

$$\begin{aligned} \hat{\boldsymbol{\delta}}_{j,k}^0 &= \Big(\int_0^1 w_j(x_j)\, dx_j \Big)^{-1} \odot \int_0^1 w_j(x_j) \odot \hat{\mathbf{f}}_{j,k}(x_j)\, dx_j, \\ &\qquad k \in I_j, \ d_0 + 1 \le k \le d, \\ \begin{pmatrix} \hat{\boldsymbol{\delta}}_{j,k}^0 \\ \hat{\boldsymbol{\delta}}_{j,k}^1 \end{pmatrix} &= \begin{pmatrix} \int_0^1 w_j(x_j)\, dx_j & \int_0^1 x_j w_j(x_j)\, dx_j \\ \int_0^1 x_j w_j(x_j)\, dx_j & \int_0^1 x_j^2 w_j(x_j)\, dx_j \end{pmatrix}^{-1} \\ &\qquad \odot \begin{pmatrix} \int_0^1 w_j(x_j) \odot \hat{\mathbf{f}}_{j,k}(x_j)\, dx_j \\ \int_0^1 x_j w_j(x_j) \odot \hat{\mathbf{f}}_{j,k}(x_j)\, dx_j \end{pmatrix}, \\ &\qquad k \in I_j, \ d_0 - r + 1 \le k \le d_0. \end{aligned}$$

Then, the normalized component maps are given by

$$\begin{aligned} &\hat{\mathbf{f}}_{j,k}(x_j) \ominus \hat{\boldsymbol{\delta}}_{j,k}^0, \quad k \in I_j, \ d_0 + 1 \le k \le d, \\ &\hat{\mathbf{f}}_{j,k}(x_j) \ominus \hat{\boldsymbol{\delta}}_{j,k}^0 \ominus x_j \odot \hat{\boldsymbol{\delta}}_{j,k}^1, \\ &\qquad k \in I_j, \ d_0 - r + 1 \le k \le d_0. \end{aligned} \tag{2.13}$$

2.4 Backfitting algorithm

The solution of the system of equations at (2.12) is not explicit. We present an iterative algorithm which is easy to implement. Suppose that we are given initial estimates $\hat{\mathbf{f}}_j^{[0]}$. Let $\hat{\mathbf{f}}_j^{[r]}$ denote the update for the jth component map in the rth iteration step. For a concise presentation define $\hat{\mathbf{f}}_{+j}^{[r]}$ for $2 \le j \le d_0$ and $\hat{\mathbf{f}}_{j+}^{[r]}$ for $1 \le j \le d_0 - 1$ by

$$\begin{aligned} \hat{\mathbf{f}}_{+j}^{[r]}(x_j) &= \bigoplus_{l \le j-1} \int_0^1 \hat{\mathbf{M}}_{jj}(x_j)^{-1} \cdot \hat{\mathbf{M}}_{jl}(x_j, x_l) \\ &\qquad \odot \hat{\mathbf{f}}_l^{[r]}(x_l)\, dx_l, \\ \hat{\mathbf{f}}_{j+}^{[r]}(x_j) &= \bigoplus_{l \ge j+1} \int_0^1 \hat{\mathbf{M}}_{jj}(x_j)^{-1} \cdot \hat{\mathbf{M}}_{jl}(x_j, x_l) \\ &\qquad \odot \hat{\mathbf{f}}_l^{[r]}(x_l)\, dx_l. \end{aligned} \tag{2.14}$$

We put $\hat{\mathbf{f}}_{+1}^{[r]} \equiv \mathbf{0}_{|I_j|}$ and $\hat{\mathbf{f}}_{d_0+}^{[r]} \equiv \mathbf{0}_{|I_j|}$. Then, the Hilbertian SBF iterative algorithm that is driven from (2.12) is given

by

$$\hat{\mathbf{f}}_j^{[r]}(x_j) = \hat{\boldsymbol{\nu}}_j(x_j) \ominus \hat{\mathbf{f}}_{+j}^{[r]}(x_j) \ominus \hat{\mathbf{f}}_{j+}^{[r-1]}(x_j), \quad 1 \le j \le d_0. \tag{2.15}$$

We note that the integrals in (2.15) are in the Bochner sense and thus they are abstractly defined. We present a useful way of computing these Bochner integrals based on Lebesgue integration. The main idea is based on the fact that, for any (Lebesgue) integrable $\mathbf{g}_j : [0,1] \to \mathbb{R}^{|I_j|}$ and for any constant $\mathbf{b}$ in a Banach space, the Bochner integral of the map $\mathbf{g}_j \odot \mathbf{b} : [0,1] \to \mathbb{B}^{|I_j|}$ over $[0,1]$ equals $\int_0^1 \mathbf{g}_j(x_j)\,dx_j \odot \mathbf{b}$. Note that the integration $\int_0^1 \mathbf{g}_j(x_j)\,dx_j$ is in the Lebesgue sense. To implement (2.15) using this idea, we take initial estimates $\hat{\mathbf{f}}_j^{[0]}$ which are linear in $\mathbf{Y}_i$:

$$\hat{\mathbf{f}}_j^{[0]}(x_j) = n^{-1} \odot \bigoplus_{i=1}^{n} \mathbf{r}_{ij}^{[0]}(x_j) \odot \mathbf{Y}_i, \quad 1 \le j \le d_0, \tag{2.16}$$

where $\mathbf{r}_{ij}^{[0]} : [0,1] \to \mathbb{R}^{|I_j|}$ are vectors of *real-valued* functions depending solely on $\{X_{ij} : 1 \le i \le n, 1 \le j \le d\}$ and do not involve Hilbertian responses $\mathbf{Y}_i$. For a choice of $\mathbf{r}_{ij}^{[0]}$ one may take *zero* functions, which corresponds to the choice $\hat{\mathbf{f}}_j^{[0]} \equiv \mathbf{0}_{|I_j|}$. Now, given $\mathbf{r}_{ij}^{[r]}$, $r \ge 0$ define $\mathbf{r}_{i,+j}^{[r]}$ and $\mathbf{r}_{i,j+}^{[r]}$, respectively, just like $\hat{\mathbf{f}}_{+j}^{[r]}$ and $\hat{\mathbf{f}}_{j+}^{[r]}$ are defined from $(\hat{\mathbf{f}}_1^{[r]}, \ldots, \hat{\mathbf{f}}_d^{[r]})$ at (2.14). From the initial $\mathbf{r}_{ij}^{[0]}$, define $\mathbf{r}_{ij}^{[r]}$ for $r \ge 1$ and for $1 \le i \le n$ recursively by

$$\mathbf{r}_{ij}^{[r]}(x_j) = \hat{\mathbf{M}}_{jj}^{-1}(x_j)\mathbf{Z}_{ij}K_{h_j}(x_j, X_{ij}) - \mathbf{r}_{i,+j}^{[r]}(x_j) - \mathbf{r}_{i,j+}^{[r-1]}(x_j), \quad 1 \le j \le d_0.$$

Then, we may express the rth updates $\hat{\mathbf{f}}_j^{[r]}$ as

$$\hat{\mathbf{f}}_j^{[r]}(x_j) = n^{-1} \odot \bigoplus_{i=1}^{n} \mathbf{r}_{ij}^{[r]}(x_j) \odot \mathbf{Y}_i, \quad 1 \le j \le d_0. \tag{2.17}$$

In Section 3.2 we will show that the backfitting algorithm (2.15) converges to the solution of the system of the backfitting equations at (2.12).

The implementation of the SBF algorithm (2.15) via (2.17) requires only updating the weights $\mathbf{r}_{ij}^{[r]}(x_j) \in \mathbb{R}^{|I_j|}$ for $1 \le i \le n$ and $1 \le j \le d_0$. For functional responses $\mathbf{Y}_i \equiv Y_i(\cdot)$ sitting on a time domain $\mathcal{T}$, this means that the rth updates $\hat{\mathbf{f}}_j^{[r]}(x_j) \equiv \hat{\mathbf{f}}_j^{[r]}(x_j)(\cdot)$ for each x_j are obtained on the entire $\mathcal{T}$ all at once after the weights are updated. The computation does not need to be done for each point t on a fine grid of $\mathcal{T}$, so that the computation is fast. In case $Y_i(\cdot)$ are not observed on the entire domain $\mathcal{T}$ but only on a discrete subset of $\mathcal{T}$, one may still apply our method after a pre-smoothing step where one can use popular nonparametric smoothing techniques. For example, in the real data example to be discussed in Section 4.3 we used the local linear smoothing technique. Our theory given in the next section assumes that $\mathbf{Y}_i$ are completely observed. It does not take into account the errors in the reconstruction of $\mathbf{Y}_i$ from incomplete observations.

3. THEORY

Here, we discuss the theoretical properties of the Hilbertian SBF estimation that we described in Section 2. For this we first introduce some relevant spaces of Hilbertian maps and associated linear operators.

3.1 Projection onto model space

Let $\mathcal{X}_j$ be the support of the predictor X_j in the linear parts ($d_0+1 \le j \le d$). We let $\mathcal{X} := [0,1]^{d_0} \times \prod_{j=d_0+1}^{d} \mathcal{X}_j$ denote the support of the distribution P_X. The space of functions that embodies the true regression map is the collection of $\boldsymbol{\mu} : \mathcal{X} \to \mathbb{H}$ with $\|\boldsymbol{\mu}\|_2^2 := \int \|\boldsymbol{\mu}(\mathbf{x})\|_{\mathbb{H}}^2\, dP_{\mathbf{X}}(\mathbf{x}) < \infty$. We denote the function class by $\mathscr{M}$. Clearly, $\|\cdot\|_2$ is a norm and $\mathscr{M}$ is a Hilbert space. Now, let $\mathscr{M}_{\mathrm{vc}}$ be a subspace of $\mathscr{M}$ such that its elements are of the form $\boldsymbol{\mu}(\mathbf{x}) = \bigoplus_{j=1}^{d_0} \mathbf{z}_j^\top \odot \mathbf{g}_j(x_j)$ with $\mathbf{g}_j : [0,1] \to \mathbb{H}^{|I_j|}$ for $1 \le j \le d_0$. The space $\mathscr{M}_{\mathrm{vc}}$ embodies the true regression map $\mathbf{m}(\cdot) = \mathrm{E}(\mathbf{Y}|\mathbf{X} = \cdot)$ under the model (2.1). Also, let $\mathscr{M}_j$ denote subspaces of $\mathscr{M}_{\mathrm{vc}}$ (and $\mathscr{M}$) such that $\boldsymbol{\mu}(\mathbf{x}) = \mathbf{z}_j^\top \odot \mathbf{g}_j(x_j)$, which embodies each additive interaction term in the model (2.1). Thus, $\mathscr{M}_{\mathrm{vc}} = \mathscr{M}_1 + \cdots + \mathscr{M}_{d_0}$. The Hilbertian SBF estimation that we introduced in Section 2.3 is expressed in terms of linear operators, $\hat{\pi}_j$ defined at (3.4) below, that map $\mathscr{M}_{\mathrm{vc}}$ to $\mathscr{M}_j$. The existence of an estimator $\hat{\mathbf{m}}$ of $\mathbf{m}$ that takes the form $\hat{\mathbf{m}}(\mathbf{x}) = \bigoplus_{j=1}^{d_0} \mathbf{z}_j^\top \hat{\mathbf{f}}_j(x_j)$ with $\hat{\mathbf{f}}_j$ satisfying (2.12) and the convergence of the backfitting algorithm (2.15), depend on the contraction property of a linear operator that maps $\mathscr{M}_{\mathrm{vc}}$ to itself. The latter linear operator is constructed from $\hat{\pi}_j$, see $\hat{T}$ defined at (3.7) below.

We first introduce the population versions of $\hat{\pi}_j$. Let $\langle \cdot, \cdot \rangle_2 : \mathscr{M} \times \mathscr{M} \to [0, \infty)$ denote the inner product of $\mathscr{M}$, which is defined by $\langle \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \rangle_2 = \int \langle \boldsymbol{\mu}_1(\mathbf{x}), \boldsymbol{\mu}_2(\mathbf{x}) \rangle_{\mathbb{H}}\, dP_{\mathbf{X}}(\mathbf{x})$. Let $\pi_j : \mathscr{M} \to \mathscr{M}_j$ with $\pi_j(\boldsymbol{\mu})(\mathbf{x}) = \mathbf{z}_j^\top \odot \mathbf{g}_j^*(x_j)$ be the projection operator characterized by

$$\int \langle \boldsymbol{\mu}(\mathbf{x}) \ominus \pi_j(\boldsymbol{\mu})(\mathbf{x}),\, \mathbf{z}_j \odot \mathbf{g}_j(x_j) \rangle_{\mathbb{H}}\, dP_{\mathbf{X}}(\mathbf{x}) = 0 \quad \text{for all } \mathbf{g}_j : \mathbf{g}_j : [0,1] \to \mathbb{H}^{|I_j|}. \tag{3.1}$$

Define $|I_j| \times |I_j|$ and $|I_j| \times |I_l|$ real matrices, respectively, by

$$\mathbf{M}_{jj}(x_j) = \mathrm{E}(\mathbf{Z}_j\mathbf{Z}_j^\top \,|\, X_j = x_j) \cdot p_j(x_j), \quad 1 \le j \le d_0$$
$$\mathbf{M}_{jl}(x_j, x_l) = \mathrm{E}(\mathbf{Z}_j\mathbf{Z}_l^\top \,|\, X_j = x_j, X_l = x_l) \cdot p_{jl}(x_j, x_l), \quad 1 \le j \ne l \le d_0,$$

where p_{jl} are the two-dimensional joint densities of (X_j, X_l). These are the population versions of the empirical $\hat{\mathbf{M}}_{jj}$ and

$\hat{\mathbf{M}}_{jl}$, respectively, introduced in Section 2.3. Then, we may derive from (3.1) that

$$\pi_j(\boldsymbol{\mu})(\mathbf{x}) = \mathbf{z}_j^\top \cdot \mathbf{M}_{jj}(x_j)^{-1} \cdot p_j(x_j) \odot \mathrm{E}(\mathbf{Z}_j \odot \boldsymbol{\mu}(\mathbf{X}) \,|\, X_j = x_j), \quad 1 \le j \le d_0. \tag{3.2}$$

The projection of $\boldsymbol{\mu} \in \mathscr{M}_{\mathrm{vc}}$ onto $\mathscr{M}_j$ is a restriction of π_j to the domain space $\mathscr{M}_{\mathrm{vc}}$. For this let $\boldsymbol{\mu}(\mathbf{x}) = \boldsymbol{\mu}_1(\mathbf{x}) \oplus \cdots \oplus \boldsymbol{\mu}_{d_0}(\mathbf{x})$ with $\boldsymbol{\mu}_l(\mathbf{x}) = \mathbf{z}_l^\top \odot \mathbf{g}_l(x_l)$. Applying (3.2) to $\boldsymbol{\mu}_l$ for $l \neq j$, we get

$$\pi_j(\boldsymbol{\mu}_l)(\mathbf{x}) = \mathbf{z}_j^\top \cdot \mathbf{M}_{jj}(x_j)^{-1} \odot \int_0^1 \mathbf{M}_{jl}(x_j, x_l) \odot \mathbf{g}_l(x_l)\, dx_l.$$

This gives that, for $\boldsymbol{\mu} \in \mathscr{M}_{\mathrm{vc}}$ of the form $\boldsymbol{\mu}(\mathbf{x}) = \bigoplus_{l=1}^{d_0} \mathbf{z}_l^\top \odot \mathbf{g}_l(x_l)$,

$$\pi_j(\boldsymbol{\mu})(\mathbf{x}) = \mathbf{z}_j^\top \odot \Bigg(\mathbf{g}_j(x_j) \oplus \bigoplus_{l=1:\neq j}^{d_0} \mathbf{M}_{jj}(x_j)^{-1} \odot \int_0^1 \mathbf{M}_{jl}(x_j, x_l) \odot \mathbf{g}_l(x_l)\, dx_l \Bigg). \tag{3.3}$$

The empirical versions of π_j restricted to $\mathscr{M}_{\mathrm{vc}}$ are obtained if we replace $\mathbf{M}_{jj}(x_j)$ and $\mathbf{M}_{jl}(x_j, x_l)$, respectively, by $\hat{\mathbf{M}}_{jj}(x_j)$ and $\hat{\mathbf{M}}_{jl}(x_j, x_l)$ at (2.10). Indeed, we get

$$\hat{\pi}_j(\boldsymbol{\mu}_l)(\mathbf{x}) = \mathbf{z}_j^\top \cdot \hat{\mathbf{M}}_{jj}(x_j)^{-1} \odot \int_0^1 \hat{\mathbf{M}}_{jl}(x_j, x_l) \odot \mathbf{g}_l(x_l)\, dx_l$$

for $\boldsymbol{\mu}_l$ $(l \neq j)$ of the form $\boldsymbol{\mu}_l(\mathbf{x}) = \mathbf{z}_l^\top \odot \mathbf{g}_l(x_l)$. We define $\hat{\pi}_j : \mathscr{M}_{\mathrm{vc}} \to \mathscr{M}_j$ by

$$\hat{\pi}_j(\boldsymbol{\mu})(\mathbf{x}) = \mathbf{z}_j^\top \odot \Bigg(\mathbf{g}_j(x_j) \oplus \bigoplus_{l=1:\neq j}^{d_0} \hat{\mathbf{M}}_{jj}(x_j)^{-1} \odot \int_0^1 \hat{\mathbf{M}}_{jl}(x_j, x_l) \odot \mathbf{g}_l(x_l)\, dx_l \Bigg), \tag{3.4}$$

where $\boldsymbol{\mu}(\mathbf{x}) = \bigoplus_{l=1}^{d_0} \mathbf{z}_l^\top \odot \mathbf{g}_l(x_l)$.

We note that, under the condition (A1), $\mathbf{M}_{jj}(x_j)$ are invertible for all $x_j \in [0,1]$ and for all $1 \le j \le d_0$. Define $\mu_{l,j}(x_j) = \int_0^1 ((v - x_j)/h_j)^l K_{h_j}(x_j, v)\, dv$. Then,

$$\hat{\mathbf{M}}_{jj}(x_j) = \mu_{0,j}(x_j) \cdot \mathbf{M}_{jj}(x_j) + o_p(1) \tag{3.5}$$

uniformly for $x_j \in [0,1]$ under some mild conditions, see the proof of Lemma A.1 in Section A. Since K is symmetric, if $h_j \le 1/2$, then

$$\int_0^1 K(u)\, du = \int_{-1}^0 K(u)\, du \le \int_0^1 K_h(u - v)\, du \le \int_{-1}^1 K(u)\, du \quad \text{for all } v \in [0,1].$$

This entails that $\inf_{x_j \in [0,1]} \mu_{0,j}(x_j) \ge 1/2$, which with the uniform convergence at (3.5) implies that $\hat{\mathbf{M}}_{jj}(x_j)$ is invertible for all $x_j \in [0,1]$ with probability tending to one.

We express the backfitting equation at (2.12) using the linear operators $\hat{\pi}_j$. For concise representation, we define $\hat{\mathbf{f}}_{j,\mathrm{vc}} : \mathcal{X} \to \mathbb{H}$ by $\hat{\mathbf{f}}_{j,\mathrm{vc}}(\mathbf{x}) = \mathbf{z}_j^\top \odot \hat{\mathbf{f}}_j(x_j)$. Similarly, define $\mathbf{f}_{j,\mathrm{vc}}$ with $\mathbf{f}_j$ taking the role of $\hat{\mathbf{f}}_j$. Also, let $\hat{\mathbf{m}}_j(\mathbf{x}) = \mathbf{z}_j^\top \odot \hat{\boldsymbol{\nu}}_j(x_j)$. Then, we may rewrite (2.12) as

$$\hat{\mathbf{f}}_{j,\mathrm{vc}}(\mathbf{x}) = \hat{\mathbf{m}}_j(\mathbf{x}) \ominus \bigoplus_{l=1,\neq j}^{d_0} \hat{\pi}_j(\hat{\mathbf{f}}_{l,\mathrm{vc}})(\mathbf{x}), \quad 1 \le j \le d_0.$$

This gives

$$\hat{\mathbf{m}} = \hat{\mathbf{m}}_j \oplus (I - \hat{\pi}_j)(\hat{\mathbf{m}}), \quad 1 \le j \le d_0, \tag{3.6}$$

where $\hat{\mathbf{m}}(\mathbf{x}) = \bigoplus_{j=1}^{d_0} \mathbf{z}_j^\top \odot \hat{\mathbf{f}}_j(x_j) = \bigoplus_{j=1}^{d_0} \hat{\mathbf{f}}_{j,\mathrm{vc}}(\mathbf{x})$ and I is the identity map. Define $\hat{T} : \mathscr{M}_{\mathrm{vc}} \to \mathscr{M}_{\mathrm{vc}}$ by

$$\hat{T} = (I - \hat{\pi}_{d_0}) \circ \cdots \circ (I - \hat{\pi}_1). \tag{3.7}$$

Applying (3.6) successively from $j = d_0$ to $j = 1$, we obtain

$$\hat{\mathbf{m}} = \hat{\mathbf{m}}_\oplus \oplus \hat{T}(\hat{\mathbf{m}}), \tag{3.8}$$

where $\hat{\mathbf{m}}_\oplus = \hat{\mathbf{m}}_{d_0} \oplus \hat{\pi}_{d_0}^\perp(\hat{\mathbf{m}}_{d_0-1}) \oplus (\hat{\pi}_{d_0}^\perp \circ \hat{\pi}_{d_0-1}^\perp)(\hat{\mathbf{m}}_{d_0-2}) \oplus \cdots \oplus (\hat{\pi}_{d_0}^\perp \circ \cdots \circ \hat{\pi}_2^\perp)(\hat{\mathbf{m}}_1)$ and $\hat{\pi}_j^\perp = I - \hat{\pi}_j$. Note that here we continue to use $\oplus$ to denote the map-addition operation, i.e., $(\boldsymbol{\mu}_1 \oplus \boldsymbol{\mu}_2)(\mathbf{x}) = \boldsymbol{\mu}_1(\mathbf{x}) \oplus \boldsymbol{\mu}_2(\mathbf{x})$. We also get an analogue of (3.8) for the backfitting iteration (2.15) as

$$\hat{\mathbf{m}}^{[r]} = \hat{\mathbf{m}}_\oplus \oplus \hat{T}(\hat{\mathbf{m}}^{[r-1]}), \tag{3.9}$$

where $\hat{\mathbf{m}}^{[r]}(\mathbf{x}) = \bigoplus_{j=1}^{d_0} \mathbf{z}_j^\top \odot \hat{\mathbf{f}}_j^{[r]}(x_j)$ for $r \ge 0$.

3.2 Existence of estimator and convergence of algorithm

Let T be the population version of $\hat{T}$ at (3.7) defined by $(I - \pi_{d_0}) \circ \cdots \circ (I - \pi_1)$. They are bounded linear operators that map $\mathscr{M}$ to $\mathscr{M}$. Here, we consider their restrictions on $\mathscr{M}_{\mathrm{vc}}$. Let $\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})$ denote the space of all bounded linear operators mapping $\mathscr{M}_{\mathrm{vc}}$ to itself. We endow $\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})$ with the operator norm $\|\cdot\|_{\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})}$ defined by

$$\|L\|_{\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})} = \sup\{\|L(\boldsymbol{\mu})\|_2 : \boldsymbol{\mu} \in \mathscr{M}_{\mathrm{vc}} \text{ with } \|\boldsymbol{\mu}\|_2 = 1\}.$$

According to the theory developed in Jeon and Park (2020), the equation (3.8) has a unique solution $\hat{\mathbf{m}} \in \mathscr{M}_{\mathrm{vc}}$ and the iteration of $\hat{\mathbf{m}}^{[r]}$ at (3.9) converges to the solution if $\|\hat{T}\|_{\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})} < 1$, see Section 3.2 of the aforementioned paper. In Lemma A.1 in Section A we show that $\|\hat{T} - T\|_{\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})}$ converges to zero in probability as $n \to \infty$. Thus, if $\|T\|_{\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})} < 1$, then $\|\hat{T}\|_{\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})} < \gamma$ with probability tending to one for some $0 < \gamma < 1$. According to

Theorem 4.6 in Xu and Zikatanov (2002) and Lemma 2.1 in Blot and Cieutat (2016), the statement $\|T\|_{\mathcal{L}(\mathscr{M}_{\mathrm{vc}},\mathscr{M}_{\mathrm{vc}})} < 1$ is equivalent to the following one:

$$\begin{aligned}&\text{There exists a constant } 0 < c < \infty \text{ such that}\\ &\text{each } \boldsymbol{\mu} \in \mathscr{M}_{\mathrm{vc}} \text{ admits a decomposition}\\ &\boldsymbol{\mu} = \bigoplus_{j=1}^{d_0} \mathbf{g}_{j,\mathrm{vc}} \text{ with } \mathbf{g}_{j,\mathrm{vc}} \in \mathscr{M}_j \text{ and}\\ &\sum_{j=1}^{d_0} \|\mathbf{g}_{j,\mathrm{vc}}\|_2 \le c\|\boldsymbol{\mu}\|_2.\end{aligned} \tag{3.10}$$

In Lemma A.2 in Section A we prove that (3.10) holds. In the lemma we allow $\mathbb{H}$ to be any separable Hilbert space of finite- or infinite-dimension.

The foregoing discussion gives the following Theorem 1. Before stating the theorem we collect some additional assumptions we need for the theorem. The following conditions are used to prove the consistency of $\hat{\mathbf{M}}_{jj}$ and $\hat{\mathbf{M}}_{jj'}$ as estimators of $\mathbf{M}_{jj}$ and $\mathbf{M}_{jj'}$, respectively, see Lemma A.1 in the Appendix A. In particular, the restriction $c_j + c_{j'} < 1$ in (A4) makes the variances of $\hat{\mathbf{M}}_{jj'}$, which is of magnitude $(nh_jh_{j'})^{-1}$, converge to zero.

(A2) The joint density p is bounded away from zero and infinity on $[0,1]^{d_0}$, and $\mathbf{M}_{jj}$ and $\mathbf{M}_{jj'}$ are continuous on $[0,1]$ and $[0,1]^2$, respectively, for $1 \le j \ne j' \le d_0$.

(A3) There exists a constant $\alpha > 2$ such that the followings hold: (i) for all $1 \le j \le d_0$, $\mathrm{E}(|X_lX_m|^\alpha) < \infty$ and $\mathrm{E}(|X_lX_m|^2 \,|X_j = \cdot)$ are bounded on $[0,1]$ for all $l, m \in I_j$; (ii) for all $1 \le j \ne j' \le d_0$, $\mathrm{E}(|X_lX_m|^\alpha) < \infty$ and $\mathrm{E}(|X_lX_m|^2 \,|X_j = \cdot, X_{j'} = \cdot)$ are bounded on $[0,1] \times [0,1]$ for all $l \in I_j$ and $m \in I_{j'}$.

(A4) The bandwidths h_j for $1 \le j \le d_0$ satisfy that $h_j = o(1)$ as $n \to \infty$ and $n^{c_j}h_j$ are bounded away from zero for some $0 < c_j < (\alpha - 2)/2$ with $c_j + c_{j'} < 1$, where α is the constant in (A3).

Theorem 1. *Assume that the conditions (A1)–(A4) hold. Then, there exists a unique solution of the Hilbertian backfitting equation at (3.8). Moreover, the Hilbertian backfitting algorithm at (3.9) converges to the solution in the following sense: with probability tending to one*

$$\|\hat{\mathbf{m}}^{[r]} \ominus \hat{\mathbf{m}}\|_2 \le c \cdot \gamma^r\big(\|\hat{\mathbf{m}}_\oplus\|_2 + \|\hat{\mathbf{m}}^{[0]}\|_2\big), \quad r \ge 1,$$

where $0 < c < \infty$ and $0 < \gamma < 1$ are absolute constants.

3.3 Rates of convergence of backfitting estimators

With a slight abuse of notation we continue to let $\hat{\mathbf{f}}_{j,k}$ denote the normalized component estimators satisfying the constraints (2.5). The construction of the normalized versions is described in (2.13). We also continue to write $\mathbf{f}_{j,k}$ for the normalized versions of the true component maps. Let $\hat{\mathbf{f}}_j = (\hat{\mathbf{f}}_{j,k} : k \in I_j)$. Likewise, put $\mathbf{f}_j = (\mathbf{f}_{j,k} : k \in I_j)$. Here, we derive the rates of convergence of $\hat{\mathbf{f}}_j$ in various modes. For a sequence of random elements $\{\mathbf{Z}_n : n \ge 1\}$ taking values in $\mathbb{H}^k$ for some $k \ge 1$, we write $\mathbf{Z}_n = o_p(1)$ if $P(\|\mathbf{Z}_n\|_{\mathbb{H}^k} > \epsilon) \to 0$ as $n \to \infty$ for any $\epsilon > 0$. We also write $\mathbf{Z}_n = O_p(1)$ if $\lim_{C\to\infty} \limsup_{n\to\infty} P(\|\mathbf{Z}_n\|_{\mathbb{H}^k} > C) = 0$.

For each $\mathbf{f}_{j,k} : [0,1] \to \mathbb{H}$ constituting $\mathbf{f}_j$, its first Fréchet derivative $D\mathbf{f}_{j,k}(x_j)$ at $x_j \in [0,1]$ is a bounded linear map from $\mathbb{R}$ to $\mathbb{H}$ such that $D\mathbf{f}_{j,k}(x_j)(w) = w \odot D\mathbf{f}_{j,k}(x_j)(1)$ for $w \in \mathbb{R}$, where $D\mathbf{f}_{j,k}(x_j)(1)$ is defined by

$$\lim_{|\epsilon|\to 0} \frac{1}{|\epsilon|} \cdot \big\|\mathbf{f}_{j,k}(x_j + \epsilon) \ominus \mathbf{f}_{j,k}(x_j) \ominus \big(\epsilon \odot D\mathbf{f}_{j,k}(x_j)(1)\big)\big\|_{\mathbb{H}} = 0.$$

The second Fréchet derivative of $\mathbf{f}_{j,k} : [0,1] \to \mathbb{H}$, which we denote by $D^2\mathbf{f}_{j,k}$, is a map from $[0,1]$ to $\mathcal{L}(\mathbb{R}\times\mathbb{R}, \mathbb{H})$, so that $D^2\mathbf{f}_{j,k}(x_j)$ for $x_j \in [0,1]$ is a bounded linear map from $\mathbb{R}\times\mathbb{R}$ to $\mathbb{H}$. Specifically, $D^2\mathbf{f}_{j,k}(x_j)(w, w') = w' \odot D^2\mathbf{f}_{j,k}(x_j)(w, 1)$, where $D^2\mathbf{f}_{j,k}(x_j)(w, 1)$ is defined by

$$\begin{aligned}\lim_{|\epsilon|\to 0} \frac{1}{|\epsilon|} \cdot \|&D\mathbf{f}_{j,k}(x_j + \epsilon)(w) \ominus D\mathbf{f}_{j,k}(x_j)(w)\\ &\ominus \big(\epsilon \odot D^2\mathbf{f}_{j,k}(x_j)(w, 1)\big)\big\|_{\mathbb{H}} = 0\end{aligned}$$

for $w \in \mathbb{R}$. It holds that $D^2\mathbf{f}_{j,k}(x_j)(w, w') = w \cdot w' \odot D^2\mathbf{f}_{j,k}(x_j)(1, 1)$.

To derive the rates of convergence of $\hat{\mathbf{f}}_j$ to $\mathbf{f}_j$, we make the following additional assumptions. Let $\boldsymbol{\varepsilon} = \mathbf{Y} \ominus \bigoplus_{j=1}^{d_0} \mathbf{Z}_j^\top \odot \mathbf{f}_j(X_j)$.

(A5) $\mathrm{E}(\|\boldsymbol{\varepsilon}\|_{\mathbb{H}}^\alpha) < \infty$ for some $\alpha > 5/2$ and $\mathrm{E}(\|\boldsymbol{\varepsilon}\|_{\mathbb{H}}^2 \,|\, X_j = \cdot)$ for $1 \le j \le d_0$ are bounded on $[0,1]$.

(A6) Each entry of $\mathbf{f}_j$ for $1 \le j \le d_0$ are twice continuously Fréchet differentiable on $[0,1]$.

(A7) Each entry of $\mathbf{M}_{jj'}$, for $1 \le j \ne j' \le d_0$, are continuously differentiable on $[0,1]^2$.

(A8) For all $1 \le j \le d_0$, it holds that $n^{1/5}h_j \to c_j$ for some constants $0 < c_j < \infty$.

Theorem 2. *Assume (A0)–(A3) and (A5)–(A8). Let $(\mathbf{f}_j : 1 \le j \le d_0)$ and $(\hat{\mathbf{f}}_j : 1 \le j \le d_0)$ are the tuple of the true component maps and of their estimators, respectively, that satisfy the constraints (2.5). Then, for all $1 \le j \le d_0$, it holds that*

(i) (pointwise convergence)

$$\begin{aligned}&\hat{\mathbf{f}}_j(x_j) \ominus \mathbf{f}_j(x_j) = O_p(n^{-2/5}), \quad x_j \in [2h_j, 1 - 2h_j],\\ &\hat{\mathbf{f}}_j(x_j) \ominus \mathbf{f}_j(x_j) = O_p(n^{-1/5}), \quad x_j \in [0,1] \setminus [2h_j, 1 - 2h_j];\end{aligned}$$

(ii) (L_2 convergence)

$$\begin{aligned}&\int_{2h_j}^{1-2h_j} \|\hat{\mathbf{f}}_j(x_j) \ominus \mathbf{f}_j(x_j)\|_{\mathbb{H}^{|I_j|}}^2 \cdot p_j(x_j)\,dx_j = O_p(n^{-4/5}),\\ &\int_{[0,1]\setminus[2h_j,1-2h_j]} \|\hat{\mathbf{f}}_j(x_j) \ominus \mathbf{f}_j(x_j)\|_{\mathbb{H}^{|I_j|}}^2 \cdot p_j(x_j)\,dx_j\\ &\qquad = O_p(n^{-3/5});\end{aligned}$$

(iii) (uniform convergence)

$$\sup_{x_j\in[2h_j,1-2h_j]} \|\hat{\mathbf{f}}_j(x_j)\ominus \mathbf{f}_j(x_j)\|_{\mathbb{H}^{|I_j|}} = O_p(n^{-2/5}\sqrt{\log n}),$$
$$\sup_{x_j\in[0,1]\setminus[2h_j,1-2h_j]} \|\hat{\mathbf{f}}_j(x_j)\ominus \mathbf{f}_j(x_j)\|_{\mathbb{H}^{|I_j|}} = O_p(n^{-1/5}).$$

3.4 Asymptotic distributions of backfitting estimators

Here, we present the joint asymptotic distribution of $\big(\hat{\mathbf{f}}_j(x_j) : 1\le j\le d_0\big)$. Let $\{\mathbf{e}_l : 1\le l\le L\}$ be an orthonormal basis of $\mathbb{H}$, where we allow $L=\infty$. We make the following additional assumptions.

(A9) For all l,l' and $1\le j\ne j'\le d_0$, it holds that $\mathrm{E}\big(\langle\boldsymbol{\varepsilon},\mathbf{e}_l\rangle\cdot\langle\boldsymbol{\varepsilon},\mathbf{e}_{l'}\rangle_{\mathbb{H}}\,|\,X_j=\cdot\big)$ is continuous on $[0,1]$, and for the constant α in (A5) the real-valued functions $\mathrm{E}(\|\boldsymbol{\varepsilon}\|_{\mathbb{H}}^{\alpha}\,|\,X_j=\cdot)$ and $\mathrm{E}\big(\langle\boldsymbol{\varepsilon},\mathbf{e}_l\rangle\cdot\langle\boldsymbol{\varepsilon},\mathbf{e}_{l'}\rangle_{\mathbb{H}}\,|\,X_j=\cdot, X_{j'}=\cdot\big)$ are bounded on $[0,1]$ and $[0,1]^2$, respectively.

Define $\mathbf{e}_{k,l}^{(j)} = (\mathbf{0},\dots,\mathbf{0},\mathbf{e}_l,\mathbf{0},\dots,\mathbf{0})^\top\in\mathbb{H}^{|I_j|}$, where $\mathbf{e}_l$ appears at the same position as X_k in $\mathbf{Z}_j=(X_k : k\in I_j)$. We note that $\{\mathbf{e}_{k,l}^{(j)} : k\in I_j, 1\le l\le L\}$ constitutes an orthonormal basis of $\mathbb{H}^{|I_j|}$. For each $1\le j\le d_0$ and $x_j\in[0,1]$, define a linear operator $\mathcal{C}_j(\cdot,x_j):\mathbb{H}^{|I_j|}\to\mathbb{H}^{|I_j|}$ characterized by

$$\langle\mathcal{C}_j(\mathbf{e}_{k,l}^{(j)},x_j),\mathbf{e}_{k',l'}^{(j)}\rangle_{\mathbb{H}^{|I_j|}} = c_j^{-1}\cdot\big(\mathbf{M}_{jj}(x_j)^{-1}\big)_{k,k'} \cdot\mathrm{E}\big(\langle\boldsymbol{\varepsilon},\mathbf{e}_l\rangle\cdot\langle\boldsymbol{\varepsilon},\mathbf{e}_{l'}\rangle_{\mathbb{H}}\,|\,X_j=x_j\big)\cdot\int K(t)^2\,dt,$$

where $\big(\mathbf{M}_{jj}(x_j)^{-1}\big)_{k,k'}$ denotes the entry of the $|I_j|\times|I_j|$ matrix $\mathbf{M}_{jj}(x_j)^{-1}$ at the same position as $\mathrm{E}(X_kX_{k'}\,|\,X_j=x_j)\cdot p_j(x_j)$ in $\mathbf{M}_{jj}(x_j)$. Let $\mathbf{W}_j(x_j)\equiv\mathbf{G}_j(\mathbf{0}_{|I_j|},\mathcal{C}_j(\cdot,x_j))$ denote a Gaussian random element taking values in $\mathbb{H}^{|I_j|}$, with mean $\mathbf{0}_{|I_j|}$ and covariance operator $\mathcal{C}_j(\cdot,x_j):\mathbb{H}^{|I_j|}\to\mathbb{H}^{|I_j|}$. This means that, for each $\boldsymbol{\eta}_j\in\mathbb{H}^{|I_j|}$, the real-valued random variable $\langle\mathbf{W}_j(x_j),\boldsymbol{\eta}_j\rangle_{\mathbb{H}^{|I_j|}}$ is normally distributed with mean zero and variance $\mathrm{E}\big(\langle\mathbf{W}_j(x_j),\boldsymbol{\eta}_j\rangle^2_{\mathbb{H}^{|I_j|}}\big)$, and that

$$\mathrm{E}\big(\langle\mathbf{W}_j(x_j),\boldsymbol{\eta}_j\rangle_{\mathbb{H}^{|I_j|}}\cdot\langle\mathbf{W}_j(x_j),\boldsymbol{\eta}_j'\rangle_{\mathbb{H}^{|I_j|}}\big) = \langle\mathcal{C}_j(\boldsymbol{\eta}_j,x_j),\boldsymbol{\eta}_j'\rangle_{\mathbb{H}^{|I_j|}},\quad \boldsymbol{\eta},\boldsymbol{\eta}'\in\mathbb{H}^{|I_j|}.$$

For the constants c_j in (A8), define the non-stochastic terms

$$\tilde{\boldsymbol{\beta}}_j(x_j) = \mu_2\cdot c_j^2\cdot\mathbf{M}_{jj}(x_j)^{-1}\cdot\Big(\frac{\partial}{\partial x_j}\mathbf{M}_{jj}(x_j)\Big)\odot D\mathbf{f}_j(x_j)(1) \oplus\bigoplus_{k=1:\ne j}^{d_0}\int_0^1\mu_2\cdot c_k^2\cdot\mathbf{M}_{jj}(x_j)^{-1}\cdot\Big(\frac{\partial}{\partial x_k}\mathbf{M}_{jk}(x_j,x_k)\Big)\odot D\mathbf{f}_k(x_k)(1)\,dx_k,$$

where $\mu_2=\int t^2K(t)\,dt$. Consider the following system of equations in $(\boldsymbol{\xi}_j : 1\le j\le d_0)$ with $\boldsymbol{\xi}_j:[0,1]\to\mathbb{H}^{|I_j|}$:

$$\boldsymbol{\xi}_j(x_j) = \tilde{\boldsymbol{\beta}}_j(x_j)\ominus\bigoplus_{k=1:\ne j}^{d_0}\int_0^1\mathbf{M}_{jj}(x_j)^{-1}\cdot\mathbf{M}_{jk}(x_j,x_k)\odot\boldsymbol{\xi}_k(x_k)\,dx_k,\quad 1\le j\le d_0. \tag{3.11}$$

An equivalent version of (3.11) in terms of $\boldsymbol{\xi}_{j,\mathrm{vc}}\in\mathscr{M}_j\subset\mathscr{M}_{\mathrm{vc}}$ is

$$\boldsymbol{\xi}_{j,\mathrm{vc}}(\mathbf{x}) = \tilde{\boldsymbol{\beta}}_{j,\mathrm{vc}}(\mathbf{x})\ominus\bigoplus_{k=1:\ne j}^{d_0}\pi_j(\boldsymbol{\xi}_{k,\mathrm{vc}})(\mathbf{x}),\quad 1\le j\le d_0,$$

where $\tilde{\boldsymbol{\beta}}_{j,\mathrm{vc}}(\mathbf{x})=\mathbf{z}_j^\top\odot\tilde{\boldsymbol{\beta}}_j(x_j)$. The latter system of equations defines a unique map in $\mathscr{M}_{\mathrm{vc}}$. Call it $\boldsymbol{\beta}_{\mathrm{vc}}$, which is represented as $\boldsymbol{\beta}_{\mathrm{vc}}(\mathbf{x})=\bigoplus_{j=1}^{d_0}\mathbf{z}_j^\top\odot\boldsymbol{\xi}_j(x_j)$ for some $\boldsymbol{\xi}_j:[0,1]\to\mathbb{H}^{|I_j|}$. Recall that each component map $\boldsymbol{\xi}_j$ is identified only up to an additive constant or up to an additive linear term. Let $(\boldsymbol{\beta}_j : 1\le j\le d_0)$ denote a version of $(\boldsymbol{\xi}_j : 1\le j\le d_0)$ that satisfies

$$\int_0^1 w_j(x_j)\odot\boldsymbol{\beta}_{j,k}(x_j)\,dx_j = (-\mu_2c_j^2/2)\odot\int_0^1 w_j(x_j)\odot D^2\mathbf{f}_{j,k}(x_j)(1,1)\,dx_j,\quad j\in\tilde{I}_k,\ d_0-r+1\le k\le d,$$
$$\int_0^1 x_jw_j(x_j)\odot\boldsymbol{\beta}_{j,k}(x_j)\,dx_j = (-\mu_2c_j^2/2)\odot\int_0^1 x_jw_j(x_j)\odot D^2\mathbf{f}_{j,k}(x_j)(1,1)\,dx_j,\quad j\in\tilde{I}_k,\ d_0-r+1\le k\le d_0.$$

Here, we write $\boldsymbol{\beta}_j(x_j)=(\boldsymbol{\beta}_{j,k}(x_j) : k\in I_j)$ and $D^2\mathbf{f}_j(x_j)(1,1)=\big(D^2\mathbf{f}_{j,k}(x_j)(1,1) : k\in I_j\big)$. Define

$$\boldsymbol{\theta}_j(x_j)=\boldsymbol{\beta}_j(x_j)\oplus\frac{1}{2}\mu_2\cdot c_j^2\odot D^2\mathbf{f}_j(x_j)(1,1).$$

We note that $\boldsymbol{\theta}_j$ satisfy the constraints (2.5).

Theorem 3. *Assume that the conditions (A0)–(A3) and (A5)–(A9) hold. Then, for each $\mathbf{x}^c=(x_1,\dots,x_{d_0})^\top\in(0,1)^{d_0}$, the joint distribution of $\big(n^{2/5}\odot(\hat{\mathbf{f}}_j(x_j)\ominus\mathbf{f}_j(x_j)) : 1\le j\le d_0\big)$ converges to $\big(\boldsymbol{\theta}_j(x_j)\oplus\mathbf{G}_j(\mathbf{0}_{|I_j|},\mathcal{C}_j(\cdot,x_j)) : 1\le j\le d_0\big)$, where $\mathbf{G}_j(\mathbf{0}_{|I_j|},\mathcal{C}_j(\cdot,x_j))$ are independent. Moreover, for each $\mathbf{x}\in(0,1)^{d_0}\times\prod_{j=d_0+1}^{d}\mathcal{X}_j$, the distribution of $n^{2/5}\odot\big(\hat{\mathbf{m}}(\mathbf{x})\ominus\mathbf{m}(\mathbf{x})\big)$ converges to $\bigoplus_{j=1}^{d_0}\mathbf{z}_j^\top\odot\big[\boldsymbol{\theta}_j(x_j)\oplus\mathbf{G}_j(\mathbf{0}_{|I_j|},\mathcal{C}_j(\cdot,x_j))\big]$.*

Remark 1. *If we choose h_j going to zero faster than $n^{-1/5}$, then the asymptotic biases $\boldsymbol{\theta}_j(x_j)$ are of negligible magnitudes. Specifically, if $h_j=o(n^{-1/5})$ for all $1\le j\le d_0$, then we may show that the joint distribution of*

$\big(n^{1/2}h_j^{1/2} \odot (\hat{\mathbf{f}}_j(x_j) \ominus \mathbf{f}_j(x_j)) : 1 \le j \le d_0\big)$ *converges to* $\big(\mathbf{G}_j(\mathbf{0}_{|I_j|}, \mathcal{C}_j^*(\cdot, x_j)) : 1 \le j \le d_0\big)$, *where* $\mathbf{G}_j(\mathbf{0}_{|I_j|}, \mathcal{C}_j^*(\cdot, x_j))$ *are independent and* $\mathcal{C}_j^*(\cdot, x_j)$ *are defined as* $\mathcal{C}_j(\cdot, x_j)$ *with* c_j *in the definition being replaced by* 1. *Furthermore, if all* h_j *are of the same magnitude such that* $h_j = \tilde{c}_j n^{-a}$ *for some constants* $a > 1/5$ *and* $0 < \tilde{c}_j < \infty$, *then the distribution of* $n^{(1-a)/2} \odot \big(\hat{\mathbf{m}}(\mathbf{x}) \ominus \mathbf{m}(\mathbf{x})\big)$ *converges to* $\bigoplus_{j=1}^{d_0} \tilde{c}_j^{-1/2} \mathbf{z}_j^\top \odot \mathbf{G}_j(\mathbf{0}_{|I_j|}, \mathcal{C}_j^*(\cdot, x_j))$.

4. NUMERICAL PROPERTIES

4.1 Density responses

Since we deal with random densities as response variables in the simulation study and real data example in the next two subsections, we briefly introduce the associated vector operations and the inner product acting on the space of probability density functions that makes the space be a Hilbert space. Let U be a subset of $\mathbb{R}$ with finite Lebesgue measure. Consider the space of density functions $\mathbf{y} \equiv y(\cdot)$, with respect to the Lebesgue measure on $\mathbb{R}$, supported on U such that $\int_U \log^2(y(u))\,du < \infty$. For this space, the zero vector is the constant density $\mathbf{0} = \mathrm{Leb}(U)^{-1}$, where Leb denotes the Lebesgue measure. For a scalar $c \in \mathbb{R}$ and for density functions $\mathbf{y}_1 \equiv y_1(\cdot)$ and $\mathbf{y}_2 \equiv y_2(\cdot)$, the vector addition $\mathbf{y}_1 \oplus \mathbf{y}_2$ and scalar multiplication $c \odot \mathbf{y}$ are defined by

$$\mathbf{y}_1 \oplus \mathbf{y}_2 = \frac{y_1(\cdot) \cdot y_2(\cdot)}{\int_U y_1(u) \cdot y_2(u)\,du}, \quad c \odot \mathbf{y} = \frac{y(\cdot)^c}{\int_U y(u)^c\,du}.$$

The inner product and norm are

$$\langle \mathbf{y}_1, \mathbf{y}_2 \rangle_{\mathbb{H}} = \frac{1}{2\mathrm{Leb}(U)} \int_{U^2} \log\left(\frac{y_1(u)}{y_1(u')}\right) \cdot \log\left(\frac{y_2(u)}{y_2(u')}\right) du\,du', \tag{4.1}$$

$$\|\mathbf{y}\|_{\mathbb{H}} = \left(\frac{1}{2\mathrm{Leb}(U)} \int_{U^2} \left[\log\left(\frac{y(u)}{y(u')}\right)\right]^2 du\,du'\right)^{1/2}.$$

The space of density functions with the above vector operations and inner product is then a separable Hilbert space, see van den Boogaart et al. (2014).

4.2 Simulation study

We considered one varying coefficient (VC) model and two non-VC models for each of the dimensions $d_0 = 2$ and 3. The inclusion of the non-VC models was to see the sensitivity of our approach to model violation. The response variables $\mathbf{Y} = Y(\cdot)$ were random densities.

The VC models were of the form

$$Y(\cdot) = \left(\int_U \prod_{j=1}^{d_0} (f_j(X_j)(u))^{Z_j}\,\varepsilon(u)\,du\right)^{-1} \cdot \prod_{j=1}^{d_0} (f_j(X_j)(\cdot))^{Z_j}\,\varepsilon(\cdot), \tag{4.2}$$

where $U = [-1/2, 1/2]$ and we write $f_j(x_j)(\cdot)$ for $\mathbf{f}_j(x_j)$ and $\varepsilon(\cdot)$ for $\boldsymbol{\varepsilon}$. The model falls into the type of model given at (2.2) with Z_j here corresponding to X_{d_0+j} there and $\mathbf{f}_j$ here to $\mathbf{f}_{j,d_0+j}$ there, so that we do not need to implement the constraints (2.5) since $r = 0$ and all $\tilde{I}_{d_0+j} = \{j\}$ are singletons. We took $f_j(x_j)(u) = \exp(-jx_j^j u^j)$ for $1 \le j \le d_0$ and $\varepsilon(u) = \exp(-Wu^4)$ with W being a uniform $[-1, 1]$ random variable. The predictors $(X_1, \ldots, X_{d_0})$ and $(Z_1, \ldots, Z_{d_0})$ were chosen to be independent. Specifically, $Z_1 \equiv 1$ and $Z_2 \sim N(0, 1)$ for the case $d_0 = 2$, and $Z_1 \equiv 1$ and (Z_2, Z_3) is bivariate normal with mean $(0, 0)$ and variance-covariance $(1, 1, 0.5)$.

The first non-VC scenario was

$$Y(u) = \frac{\prod_{j=1}^2 (f_j(X_j)(u))^{Z_j} f_{12}(X_1, X_2)(u)\varepsilon(u)}{\int_U \prod_{j=1}^2 (f_j(X_j)(u))^{Z_j} f_{12}(X_1, X_2)(u)\varepsilon(u)\,du} \quad (d_0 = 2),$$

$$Y(u) = \frac{\prod_{j=1}^3 (f_j(X_j)(u))^{Z_j} f_{123}(X_1, X_2, X_3)(u)\varepsilon(u)}{\int_U \prod_{j=1}^3 (f_j(X_j)(u))^{Z_j} f_{123}(X_1, X_2, X_3)(u)\varepsilon(u)\,du} \quad (d_0 = 3).$$

where $f_j(x_j)$'s and ε are as defined in the VC scenario at (4.2) and $f_{12}(x_1, x_2)(u) = \exp(-x_1 x_2 u^2)$ and $f_{123}(x_1, x_2, x_3)(u) = \exp(-(x_1x_2 + x_1x_3 + x_2x_3)u^2)$. The second non-VC scenario was

$$Y(u) = \frac{\log((X_1 + X_2)u/2 + 2)^{(Z_1+Z_2)/2}\varepsilon(u)}{\int_{-1/2}^{1/2} \log((X_1 + X_2)u/2 + 2)^{(Z_1+Z_2)/2}\epsilon(u)\,du} \quad (d_0 = 2),$$

$$Y(u) = \frac{\log((X_1 + X_2 + X_3)u/2 + 2)^{(Z_1+Z_2+Z_3)/2}\varepsilon(u)}{\int_{-1/2}^{1/2} \log((X_1 + X_2 + X_3)u/2 + 2)^{(Z_1+Z_2+Z_3)/2}\epsilon(u)\,du} \quad (d_0 = 3).$$

We generated training samples of sizes $n = 100$ and 400 and a test sample of size $N = 100$, repeatedly for $M = 100$ times. We used the Epanechnikov kernel $K(u) = 3/4(1-u^2)I(|u| < 1)$.

We chose the initial estimators with weights $r_{ij}^{[0]}(x_j) \equiv 0$ for $1 \le i \le n, 1 \le j \le d_0$, i.e., $\hat{\mathbf{f}}_j^{[0]}(x_j) = \hat{f}_j^{[0]}(x_j)(\cdot) = \mathbf{0}$, which means that $\hat{f}_j^{[0]}(x_j)(u) \equiv 1$ for all $u \in [-1/2, 1/2]$. We set the convergence criterion of the backfitting algorithm as

$$\max_{1 \le j \le d_0} \int_0^1 \|\hat{f}_j^{[r]}(x_j)(\cdot) \ominus \hat{f}_j^{[r-1]}(x_j)(\cdot)\|_{\mathbb{H}}^2\,dx_j < 10^{-4}.$$

We adopted the coordinate-wise bandwidth selection (CBS) scheme proposed by Jeon and Park (2020) to choose the bandwidths h_j. It is an iterative method of updating h_j in such a way that at the ℓth iteration step one performs the *one-dimensional* minimization of an objective function $L(h_1^{(\ell)}, \ldots, h_{j-1}^{(\ell)}, g_j, h_{j+1}^{(\ell-1)}, \ldots, h_{d_0}^{(\ell-1)})$ with respect to g_j successively from $j = 1$ to $j = d_0$. We chose a 10-fold cross-validation criterion for L. The grid over which we minimized L was $H = \prod_{j=1}^{d_0} \{a_j + 0.01 \times \ell : \ell = 0, \ldots, 20\}$, where $a_j = \max\{(X_{(i+1),j} - X_{(i),j})/2 : 0 \le i \le n\} + 0.001$ with

$$\begin{aligned} -X_{(1),j} =: & X_{(0),j} < X_{(1),j} < \cdots \\ & < X_{(n),j} < X_{(n+1),j} := 2 - X_{(n),j}, \end{aligned}$$

so that $\inf_{x_j \in [0,1]} \max_{1 \le i \le n} K_{h_j}(x_j, X_{ij}) > 0$ for $h_j \in H$.

We compared our approach with the full-dimensional methods: the functional Nadaraya-Watson (Ferraty et al., 2012) and the kernel-based functional k-NN (Lian, 2011, 2012), in terms of prediction accuracy. For the Nadaraya-Watson estimator, we used Epanechnikov kernel and tuned the bandwidth on $\{b + 0.001 \times \ell : 1 \le \ell \le 200\}$ for some small b. For the k-NN estimator, we selected k from $\{1, 2, \ldots, 30\}$. Both the bandwidth and k were determined by 10-fold cross-validation. We measured prediction accuracy by the mean squared prediction error (MSPE),

$$\mathrm{MSPE} = M^{-1} \sum_{m=1}^{M} N^{-1} \sum_{i=1}^{N} \left\| Y_i^{\mathrm{test}(m)}(\cdot) \ominus \hat{Y}_i^{\mathrm{test}(m)}(\cdot) \right\|_{\mathbb{H}}^2.$$

Here, $Y_i^{\mathrm{test}(m)}(\cdot)$ denotes the ith response in the mth test sample and

$$\hat{Y}_i^{\mathrm{test}(m)}(\cdot) = \bigoplus_{j=1}^{d_0} Z_{ij}^{\mathrm{test}(m)} \odot \hat{f}_j(X_{ij}^{\mathrm{test}(m)})(\cdot),$$

where $\big((Z_{ij}^{\mathrm{test}(m)}, X_{ij}^{\mathrm{test}(m)}) : 1 \le j \le d_0\big)$ is the ith predictor vector in the mth test sample and $(\hat{f}_j : 1 \le j \le d_0)$ is the tuple of component estimators we computed using the mth training sample.

Table 1 reports the values of the MSPE. It is observed that our method dominates the full-dimensional approaches except for non-VC (II). In the non-VC (II) scenario, the functional k-NN shows the best performance, but it becomes comparable to our approach in the higher dimension. The performance of our approach does not change much as the dimension increases in the VC-scenario, but the full-dimensional methods deteriorate very fast.

For the VC scenario and for our methods where individual component estimators are available, we calculated the Monte Carlo approximations of the integrated squared bias (ISB), the integrated variance (IV) and the mean integrated squared error (MISE) as follows:

$$\mathrm{ISB}_j(\hat{\mathbf{f}}_j) = \int_0^1 \left\| M^{-1} \odot \bigoplus_{\ell=1}^{M} \hat{\mathbf{f}}_j^{(\ell)}(x_j) \ominus \mathbf{f}_j(x_j) \right\|_{\mathbb{H}}^2 dx_j,$$

$$\begin{aligned} &\mathrm{IV}_j(\hat{\mathbf{f}}_j) \\ &\quad = M^{-1} \sum_{\ell=1}^{M} \int_0^1 \left\| \hat{\mathbf{f}}_j^{(\ell)}(x_j) \ominus M^{-1} \odot \bigoplus_{\ell'=1}^{M} \hat{\mathbf{f}}_j^{(\ell')}(x_j) \right\|_{\mathbb{H}}^2 dx_j, \\ &\mathrm{MISE}_j(\hat{\mathbf{f}}_j) = \mathrm{ISB}_j(\hat{\mathbf{f}}_j) + \mathrm{IV}_j(\hat{\mathbf{f}}_j), \end{aligned}$$

where $\hat{\mathbf{f}}_j^{(\ell)}$ denotes the estimate of $\mathbf{f}_j$ based on the ℓth Monte Carlo training sample. Table 2 presents the results. It suggests that our approach does not suffer from the dimensionality problem and is quickly stabilized as the sample size increases.

4.3 Real data example

Prediction of electricity load over time is very important for efficient power supply. There has been a recent study on an application of additive regression to predicting electricity consumption, done by Jeon and Park (2020). In the application, the response ($\mathbf{Y}$) was the trajectory of household electricity consumption on time domain (0–24 hour) and the predictors were temperature (X_1) and cloudiness (X_2). They used the data from January 2008 to December 2016, collected from KOSIS (Korean Statistical Information Service, http://kosis.kr/statHtml/statHtml.do?orgId=310&tblId=DT_3664N_2008) and the Korea meteorological administration (https://data.kma.go.kr/cmmn/main.do). However, they did not use the important predictor indicating whether it is weekday or weekend. In the present application, we added it as a new predictor Z.

To apply our approach with the additional discrete predictor, we obtained the monthly averages of the consumption trajectories, temperatures and amounts of cloud, for week days and for weekends. This means that for each month we had two observations of $(Z, X_1, X_2, \mathbf{Y})$, one for weekdays and one for weekend. We set $Z_i = 1/2$ if i corresponds to an index for weekdays and $Z_i = -1/2$ otherwise. The sample size was thus $n = 2 \times 12 \times 9 = 216$. Then, we applied the model at (2.3) with $d_0 = 2$, $d = 4$, $X_3 \equiv 1$ and $X_4 = Z$, that is,

$$\begin{aligned} \mathrm{E}(\mathbf{Y}|\mathbf{X}) = & \mathbf{f}_{1,3}(X_1) \oplus \mathbf{f}_{2,3}(X_2) \\ & \oplus Z \odot \Big(\mathbf{f}_{1,4}(X_1) \oplus \mathbf{f}_{2,4}(X_2) \Big). \end{aligned} \tag{4.3}$$

We describe the response variable $\mathbf{Y} \equiv Y(\cdot)$ in more details. The original data were given in the form of hourly trajectories of *relative* electricity loads measured at each hour. Specifically, the original data were $Y_i^{\mathrm{unsmth}}(\cdot)$ observed on the discrete time domain $\{1, 2, \ldots, 24\}$ such that

$$Y_i^{\mathrm{unsmth}}(t) = 1{,}000 \times \frac{Z_i(t)}{\sum_{s=1}^{24} Z_i(s)/24},$$

where $Z_i(s)$ denotes the electricity consumption during the one hour time interval $[s-1, s]$ averaged for weekdays or weekend days in the month corresponding to the index i. Thus, at the outset we did not know the absolute

Table 1. The values of the mean squared prediction error (MSPE), multiplied by 10^3, for our approach, the functional Nadaraya-Watson, and the kernel-based functional k-NN

Scenario	d_0	n	Our approach	Functional Nadaraya-Watson	Kernel-based functional k-NN
VC models	2	100	0.1340	3.7001	1.9280
		400	0.1023	2.5293	0.7632
	3	100	0.1612	8.0108	5.6271
		400	0.1102	6.5733	3.2092
Non-VC models (I)	2	100	0.3230	3.9033	1.9901
		400	0.2686	2.6356	0.7813
	3	100	1.4707	9.0964	6.3030
		400	1.1995	7.4041	3.5241
Non-VC models (II)	2	100	0.7040	0.7398	0.4531
		400	0.6353	0.4580	0.2343
	3	100	2.7435	4.9497	2.9849
		400	2.3269	3.1537	1.6614

Table 2. The values of the ISB, IV and MISE, multiplied by 10^3, of the varying coefficient estimators for the VC scenario. The last column contains the averages of ISB, of IV and of MISE

d_0	n		$\mathbf{f}_1$	$\mathbf{f}_2$	$\mathbf{f}_3$	average
2	100	ISB	0.00138	0.01562		0.00850
		IV	1.81114	2.17359		1.99237
		MISE	1.81252	2.18921		2.00087
	400	ISB	0.00013	0.00330		0.00172
		IV	0.45556	0.45745		0.45651
		MISE	0.45569	0.46075		0.45822
3	100	ISB	0.00115	0.01216	0.00098	0.00476
		IV	2.02386	2.13758	2.61169	2.25771
		MISE	2.02501	2.14974	2.61267	2.26247
	400	ISB	0.00032	0.00205	0.00028	0.00088
		IV	0.48699	0.50092	0.53160	0.50650
		MISE	0.48731	0.50296	0.53188	0.50739

magnitudes of daily electricity consumptions. We obtained smoothed trajectories $Y_i^{\mathrm{smth}}(\cdot)$ from these un-smoothed observations $Y_i^{\mathrm{unsmth}}(\cdot)$ by employing the local linear smoothing technique: $Y_i^{\mathrm{smth}}(s) = a(s)$ for $(a(s), b(s))$ that minimized

$$\sum_{t=1}^{24} \phi\left(\frac{s-t}{h}\right)\left(Y_i^{\mathrm{unsmth}}(t) - a(s) - b(s)(t-s)\right)^2,$$

where ϕ is the standard Gaussian density function and $h = 1/2$. Then, we normalized the smoothed $Y_i^{\mathrm{smth}}(\cdot)$ as follows:

$$Y_i(t) = \left(\int_0^{24} Y_i^{\mathrm{smth}}(s)\,ds\right)^{-1} \cdot Y_i^{\mathrm{smth}}(t), \quad t \in [0, 24].$$

The smoothed and normalized $Y_i(\cdot)$ are considered as observed values of a random density, which was the response variable $\mathbf{Y}$ in the working model (4.3).

For comparison, we also applied the additive model that involves only X_1 and X_2, which was analyzed by Jeon and Park (2020). Specifically, we considered

$$\mathrm{E}(\mathbf{Y}|X_1, X_2) = \mathbf{f}_1(X_1) \oplus \mathbf{f}_2(X_2). \tag{4.4}$$

and applied the Bochner smooth backfitting technique they developed. In this application and also for our method, we used the Epanechnikov kernel as the baseline kernel K and used the CBS algorithm that we described in Section 4.2. As a measure of performance, we used the leave-one-curve-out average squared prediction error (ASPE) defined by

$$\mathrm{ASPE} = n^{-1}\sum_{i=1}^{n} \|Y_i(\cdot) \ominus \hat{Y}_i^{(-i)}(\cdot)\|_{\mathbb{H}}^2,$$

where $\|\cdot\|_{\mathbb{H}}$ here is as defined at (4.1) with $U = [0, 24]$.

We found that the value of ASPE for our approach based on the model (4.3) was 0.0052, while the value for the additive regression based on the model (4.4) was 0.0385. Figure 1 depicts the predicted electricity consumption curves $\hat{Y}_i^{(-i)}(\cdot)$ with the actual trajectories $Y_i(\cdot)$ for six randomly chosen months. We note that both $\hat{Y}_i^{(-i)}(\cdot)$ and $Y_i(\cdot)$ are densities supported on $[0, 24]$. They clearly show the superior prediction performance done by our method. These results demonstrate that there is a marked difference in the usage of electricity between weekday and weekend, and that our approach improves prediction accuracy greatly by taking into account the difference very efficiently.

To see how different the effects of temperature and cloudiness are between weekday and weekend, we computed the estimates of the component maps in the model (4.3). For this, we used the constraints given at (2.5) with $w_j \equiv 1$. In the estimation of $\mathbf{f}_{j,k}$, we actually do not need to estimate the parametric part, which is $\boldsymbol{\alpha}_3$ and $\boldsymbol{\alpha}_4$ in $\boldsymbol{\alpha}_3 \oplus Z_i \odot \boldsymbol{\alpha}_4$ in this example, but may apply the method described in

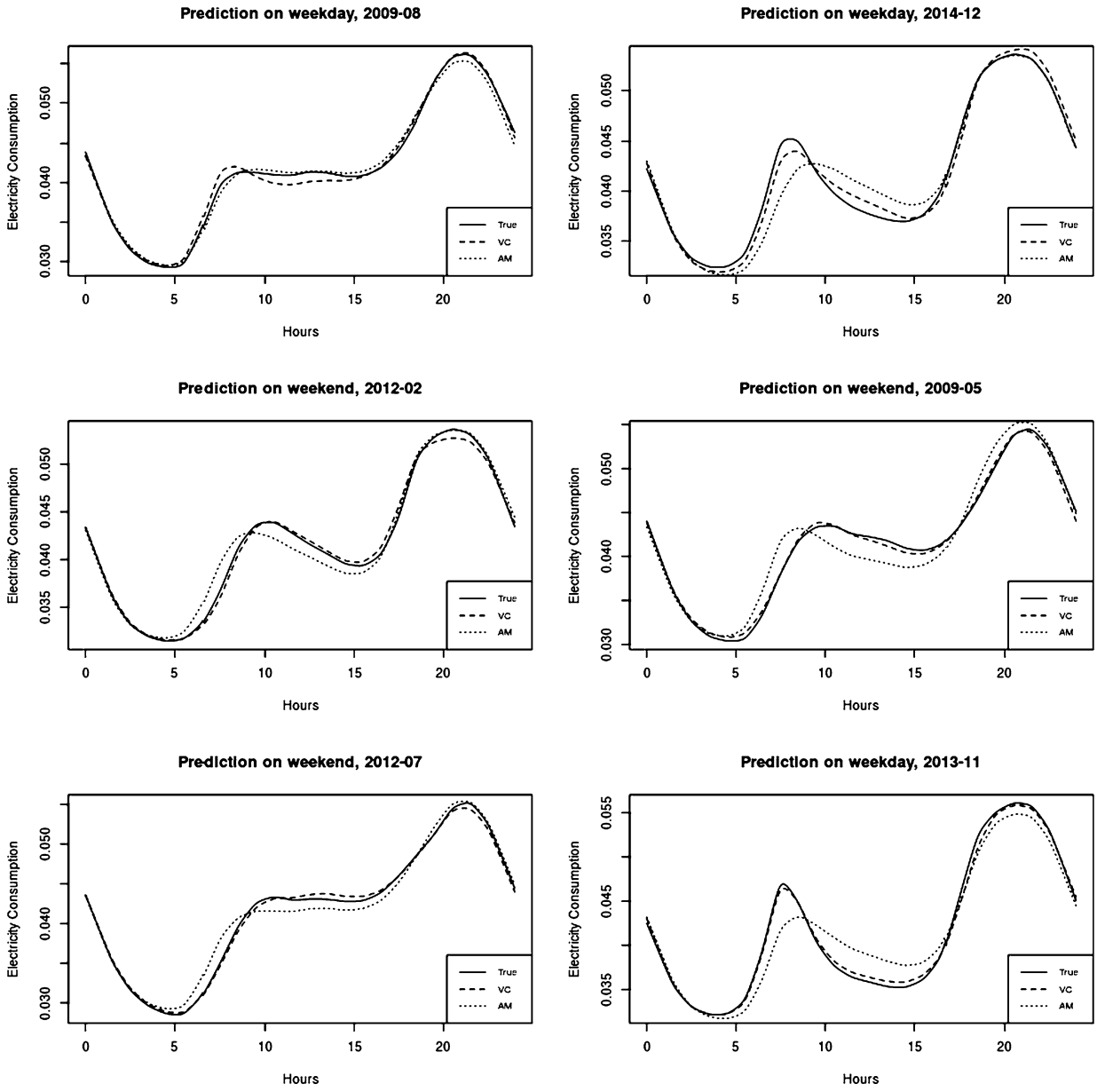
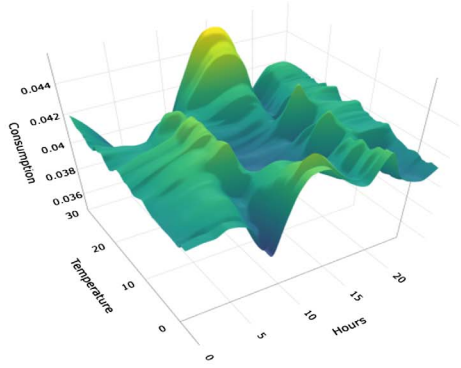


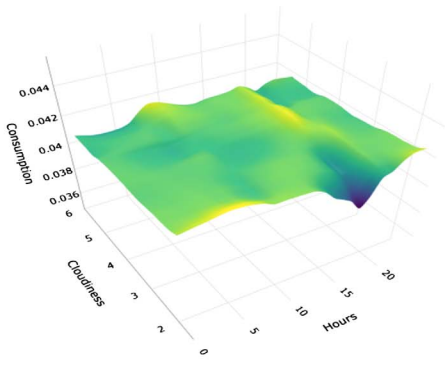
Figure 1. Prediction results for six randomly chosen months: solid (actual trajectory); dashed (prediction based on (4.3), the VC model); dotted (prediction based on (4.4), the additive model).

Section 2.3 directly to $\mathbf{Y}_i$. This is because the model (4.3) corresponds to the case $r = 0$, so that the estimators of $\mathbf{f}_{j,k}$ based on $\mathbf{Y}_i$, prior to implementing the constraints, differ from those based on $\tilde{\mathbf{Y}}_i = \mathbf{Y}_i \ominus \big(\hat{\boldsymbol{\alpha}}_3 \oplus Z_i \odot \hat{\boldsymbol{\alpha}}_4\big)$, only by (random) constants, see the Appendix B.1. The differences vanish when one implements the constraints on the first line of (2.5).

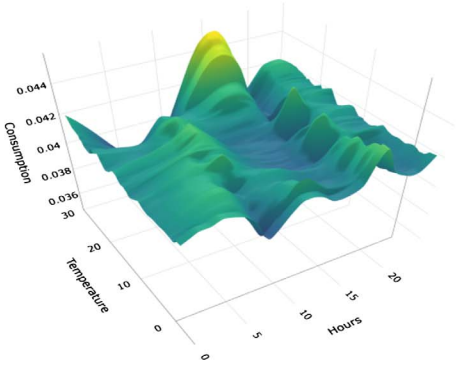
Figure 2 depicts the estimated $\hat{\mathbf{f}}_{j,k}$. We note that $\hat{\mathbf{f}}_{j,k}(x_j)$ for each fixed x_j and fixed (j,k) is a density function supported on $[0, 24]$. The three left panels are for the effect of temperature and the three right for the effect of cloudiness. Comparing them, we see that the effect of cloudiness is relatively weaker than that of temperature. Temperature has quite prominent effect on daily electricity usage. The overall picture of electricity usage as a map on the (temperature)×(hour) domain for weekday looks similar to that for weekend, but the joint effect of temperature and hour is stronger for weekday than for weekend. There are marked differences in strength for mid-level temperature and for the time period 8am–3pm, which is well reflected on the left-bottom panel that depicts (weekday effect) $\ominus$ (weekend effect). We note that the substraction operation at each temperature level also gives a density over the time domain.



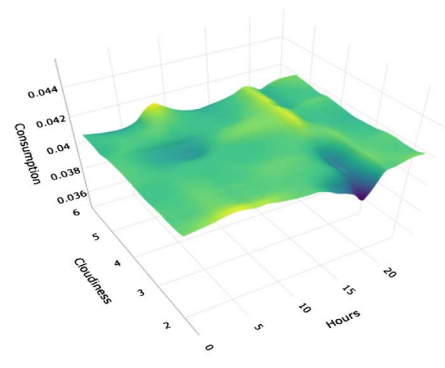
(a) Effect of temperature: weekday



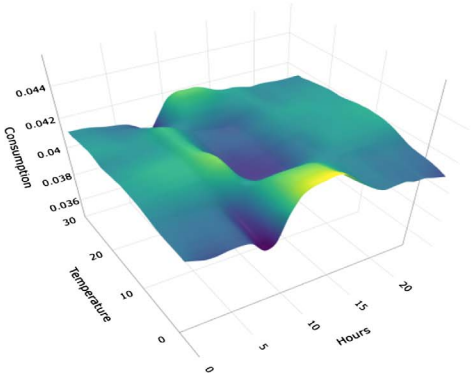
(b) Effect of cloudiness: weekday



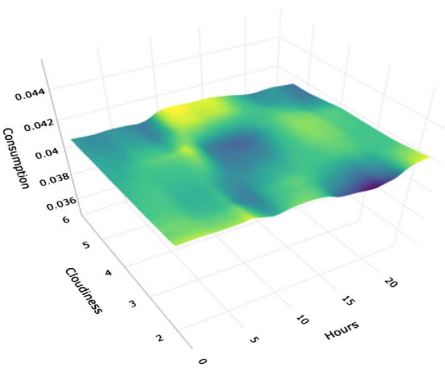
(c) Effect of temperature: weekend



(d) Effect of cloudiness: weekend



(e) Effect of temperature: difference



(f) Effect of cloudiness: difference

Figure 2. The estimated component maps. The two panels on the bottom depict (weekday effect) $\ominus$ (weekend effect) for temperature (left) and cloudiness (right).

5. A CONCLUDING REMARK

The present paper considered the estimation of the general varying coefficient model (2.1) with given index sets I_j for $1 \leq j \leq d_0$. In reality one may not know I_j. Thus, at the beginning one may fit the full model (2.4), identify I_j by searching for significant component maps $\mathbf{f}_{j,k}$ in the full model, and then refit a reduced varying coefficient model of the form at (2.1) with the chosen index sets I_j. This might be too much dependent on subjective judgement, however. An alternative but better way is to apply a penalized approach to the selection of significant components (such as LASSO and SCAD). To the best of our knowledge, there

has not been any attempt for developing a methodology of this sort in structured nonparametric regression based on kernel smoothing, even for real-valued responses. This is an important and challenging topic for future study. Another topic that might be of interest is bandwidth selection. In this paper we employed the CBS scheme with a 10-fold cross-validation criterion in our numerical studies. It is well admitted that cross-validation is subject to large sampling variability. For smooth backfitting additive regression, Mammen and Park (2005) proposed and studied three fully automatic bandwidth selection methods. One may extend the work to the current setting, which could be also a topic for future study.

APPENDIX A. PROOF OF THEOREMS

The proof of Proposition 1 is essentially the same as the proof of Lemma 1 in Lee et al. (2012b), thus is omitted. Below, we sketch the proofs of Theorems 1–3.

A.1 Proof of Theorem 1

Theorem 1 is a consequence of the following two lemmas.

Lemma A.1. *Under the conditions of Theorem 1, it holds that* $\|\hat{T} - T\|_{\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})} \to 0$ *in probability.*

Lemma A.2. *Under the conditions (A1) and (A2), the statement (3.10) holds.*

Proof of Lemma A.1. Under the conditions (A2)–(A4), we may prove

$$\begin{aligned} \hat{\mathbf{M}}_{jj}(x_j) &= \mu_{0,j}(x_j) \cdot \mathbf{M}_{jj}(x_j) + o_p(1), \\ &\qquad 1 \le j \le d_0, \\ \hat{\mathbf{M}}_{jk}(x_j, x_k) &= \mu_{0,j}(x_j) \cdot \mu_{0,k}(x_k) \cdot \mathbf{M}_{jk}(x_j, x_k) \\ &\quad + o_p(1), \quad 1 \le j \ne k \le d_0, \end{aligned} \tag{A.1}$$

uniformly for $x_j \in [0,1]$ and $(x_j, x_k) \in [0,1]^2$. Recall that $\mu_{0,j}(x_j) = 1$ for all $x_j \in [2h_j, 1-2h_j]$ and $\mu_{0,j}(x_j) \ge 1/2$ for all $x_j \in [0,1] \setminus [2h_j, 1-2h_j]$. Now, according to Lemma A.2, each $\boldsymbol{\mu} \in \mathscr{M}_{\mathrm{vc}}$ admits a decomposition $\boldsymbol{\mu} = \bigoplus_{j=1}^{d_0} \mathbf{g}_{j,\mathrm{vc}}$ with $\mathbf{g}_{j,\mathrm{vc}}(\mathbf{x}) = \mathbf{z}_j^\top \odot \mathbf{g}_j(x_j)$ such that $\max\{\|\mathbf{g}_{j,\mathrm{vc}}\|_2 : 1 \le j \le d_0\} \le c\|\boldsymbol{\mu}\|_2$, where $0 < c < \infty$ is universal for all $\boldsymbol{\mu}$. This and an application of the Hölder inequality with (A.1) give that, for all $\boldsymbol{\mu} \in \mathscr{M}_{\mathrm{vc}}$,

$$\begin{aligned} \|(\hat{\pi}_j - \pi_j)(\boldsymbol{\mu})\|_2 &\le o_p(1) \cdot \bigoplus_{k=1:\ne j}^{d_0} \|\mathbf{g}_{j,\mathrm{vc}}\|_2 \\ &\le o_p(1) \cdot \|\boldsymbol{\mu}\|_2. \end{aligned} \tag{A.2}$$

This proves $\|\hat{\pi}_j - \pi_j\|_{\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})} = o_p(1)$, which concludes the proof of the lemma. □

Proof of Lemma A.2. We first consider the case of infinite-dimensional $\mathbb{H}$. In this case, the lemma may be proved along the lines of the proof of Theorem 3.3 in Jeon and Park (2020), with $\mathscr{M}_{\mathrm{vc}}$ taking the role of their additive map space $S^{\mathbb{H}}(p)$ and $\mathcal{X}$ taking the role of their $[0,1]^d$. Below, we sketch the proof for the case of finite-dimensional $\mathbb{H}$. We prove that the projection operators π_j restricted to $\mathscr{M}_k$ for $k \ne j$ are all compact. According to Proposition A.4.2 in Bickel et al. (1993), this implies the lemma.

Recall that π_j restricted to $\mathscr{M}_k$ for $k \ne j$ is given by

$$\begin{aligned} \pi_j(\boldsymbol{\mu}_k)(\mathbf{x}) &= \mathbf{z}_j(\mathbf{x})^\top \cdot \mathbf{M}_{jj}(x_j)^{-1} \\ &\quad \odot \int_0^1 \mathbf{M}_{jk}(x_j, u_k) \odot \mathbf{g}_k(u_k)\, du_k \end{aligned} \tag{A.3}$$

for $\boldsymbol{\mu}_k(\mathbf{u}) = \mathbf{z}_k(\mathbf{u})^\top \odot \mathbf{g}_k(u_k)$, where and in this proof we restore the dependence of $\mathbf{z}_j$ and $\mathbf{z}_k$ on the points in $\mathcal{X} = [0,1]^{d_0} \times \prod_{j=d_0+1}^{d} \mathcal{X}_j$. Define $\mathbb{K}_{jk} : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathbb{H}, \mathbb{H})$ by

$$\begin{aligned} \mathbb{K}_{jk}(\mathbf{x}, \mathbf{u})(\mathbf{g}) &= \mathbf{z}_j(\mathbf{x})^\top \mathbf{M}_{jj}(x_j)^{-1} \mathbf{M}_{jk}(x_j, u_k) \\ &\quad \cdot \mathbf{M}_{kk}(u_k)^{-1} \mathbf{z}_k(\mathbf{u}) \odot \mathbf{g}, \quad \mathbf{g} \in \mathbb{H}. \end{aligned}$$

Then, from (A.3) we may see that

$$\pi_j(\boldsymbol{\mu}_k)(\mathbf{x}) = \int_{\mathcal{X}} \mathbb{K}_{jk}(\mathbf{x}, \mathbf{u})\big(\boldsymbol{\mu}_k(\mathbf{u})\big)\, dP_{\mathbf{X}}(\mathbf{u}). \tag{A.4}$$

Clearly, under the conditions (A1) and (A2), $\mathbb{K}_{j,k}(\mathbf{x}, \mathbf{u}) : \mathbb{H} \to \mathbb{H}$ is a compact operator for all $\mathbf{x}, \mathbf{u} \in \mathcal{X}$. We note that the integrals at (A.3) and (A.4) are Bochner integrals. Due to Theorem 3.1 in Jeon and Park (2020), which is an extended version of Proposition 4.7 in Conway (1985) for Lebesgue integrals, we conclude that π_j restricted to $\mathscr{M}_k$ for $k \ne j$ are all compact. This completes the proof of the lemma. □

A.2 Proof of Theorem 2

Let $\hat{\boldsymbol{\nu}}_j^A(x_j) = \bigoplus_{i=1}^n \boldsymbol{\kappa}_{ij}(x_j) \odot \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\kappa}_{ij}(x_j) = n^{-1} \cdot \hat{\mathbf{M}}_{jj}(x_j)^{-1} \mathbf{Z}_{ij} K_{h_j}(x_j, X_{ij})$ and $\boldsymbol{\varepsilon}_i := \mathbf{Y}_i \ominus \bigoplus_{k=1}^{d_0} \mathbf{Z}_{ik}^\top \odot \mathbf{f}_k(X_{ik})$. Also, define

$$\begin{aligned} \hat{\boldsymbol{\nu}}_j^B(x_j) &= \bigoplus_{i=1}^n \boldsymbol{\kappa}_{ij}(x_j) \cdot \mathbf{Z}_{ij}^\top \odot (\mathbf{f}_j(X_{ij}) - \mathbf{f}_j(x_j)), \\ \hat{\boldsymbol{\nu}}_{jk}^C(x_j) &= \bigoplus_{i=1}^n \boldsymbol{\kappa}_{ij}(x_j) \cdot \mathbf{Z}_{ik}^\top \\ &\quad \odot \int_0^1 K_{h_k}(x_k, X_{ik}) \odot (\mathbf{f}_k(X_{ik}) \ominus \mathbf{f}_k(x_k))\, dx_k. \end{aligned}$$

Due to the normalization property of the kernel $K_{h_k}(x_k, X_{ik})$ we get that

$$\begin{aligned} &\bigoplus_{i=1}^n \boldsymbol{\kappa}_{ij}(x_j) \cdot \mathbf{Z}_{ik}^\top \odot \mathbf{f}_k(X_{ik}) \\ &\quad = \hat{\boldsymbol{\nu}}_{jk}^C(x_j) \oplus \bigoplus_{i=1}^n \boldsymbol{\kappa}_{ij}(x_j) \cdot \mathbf{Z}_{ik}^\top \end{aligned}$$

$$\cdot \int_0^1 K_{h_k}(x_k, X_{ik}) \odot \mathbf{f}_k(x_k)\, dx_k.$$

This with (2.12) and the fact that $\sum_{i=1}^n \boldsymbol{\kappa}_{ij}(x_j)\mathbf{Z}_{ij}^\top$ equals the identity real matrix, gives

$$\begin{aligned}\hat{\mathbf{f}}_j(x_j) = \mathbf{f}_j(x_j) \oplus \hat{\boldsymbol{\nu}}_j^A(x_j) \oplus \hat{\boldsymbol{\nu}}_j^B(x_j) \oplus \bigoplus_{k=1:\neq j}^{d_0} \hat{\boldsymbol{\nu}}_{jk}^C(x_j) \\ \ominus \bigoplus_{k=1:\neq j}^{d_0} \int_0^1 \hat{\mathbf{M}}_{jj}(x_j)^{-1} \cdot \hat{\mathbf{M}}_{jk}(x_j, x_k) \\ \odot (\hat{\mathbf{f}}_k(x_k) \ominus \mathbf{f}_k(x_k))\, dx_k\end{aligned}$$

for $1 \le j \le d_0$.

We approximate $\hat{\boldsymbol{\nu}}_j^B$ and $\hat{\boldsymbol{\nu}}_{jk}^C$. Recall the definition of $\mu_{l,j}(x_j)$ given in Section 3.1. Let $\mu_l = \int_{-1}^1 t^l K(t)\, dt$. Note that $\mu_{l,j}(x_j) = \mu_l$ for $x_j \in \mathrm{Int}_j := [2h_j, 1-2h_j]$. Using the standard theory of kernel smoothing and the approximation $\mathbf{f}_j(v) - \mathbf{f}_j(x_j) \simeq (v - x_j) \odot D\mathbf{f}_j(x_j)(1) \oplus (1/2)(v-x_j)^2 \odot D^2\mathbf{f}_j(x_j)(1,1)$, we may prove

$$\begin{aligned}\hat{\boldsymbol{\nu}}_j^B(x_j) = h_j \cdot \frac{\mu_{1,j}(x_j)}{\mu_{0,j}(x_j)} \odot D\mathbf{f}_j(x_j)(1) \\ \oplus \mu_2 \cdot h_j^2 \cdot \mathbf{M}_{jj}(x_j)^{-1} \\ \cdot \Big(\frac{\partial}{\partial x_j}\mathbf{M}_{jj}(x_j)\Big) \odot D\mathbf{f}_j(x_j)(1) \\ \oplus \frac{1}{2} \cdot \mu_2 \cdot h_j^2 \odot D^2\mathbf{f}_j(x_j)(1,1) \oplus \mathbf{r}_j(x_j).\end{aligned} \tag{A.5}$$

Here and below, we let $\mathbf{r}_j : [0,1] \to \mathbb{H}^{|I_j|}$ denote a generic stochastic term with the following properties.

$$\begin{aligned}\sup_{x_j \in \mathrm{Int}_j} \|\mathbf{r}_j(x_j)\|_{\mathbb{H}^{|I_j|}} = o_p(n^{-2/5}), \\ \sup_{x_j \in \mathrm{Int}_j^c} \|\mathbf{r}_j(x_j)\|_{\mathbb{H}^{|I_j|}} = O_p(n^{-2/5}),\end{aligned}$$

where $\mathrm{Int}_j^c = [0,1] \setminus [2h_j, 1-2h_j]$. Furthermore, we get

$$\begin{aligned}\hat{\boldsymbol{\nu}}_{jk}^C(x_j) = \int_0^1 \hat{\mathbf{M}}_{jj}(x_j)^{-1}\hat{\mathbf{M}}_{jk}(x_j, x_k) \\ \odot \Big(h_k \cdot \frac{\mu_{1,k}(x_k)}{\mu_{0,k}(x_k)} \odot D\mathbf{f}_k(x_k)(1) \\ \oplus \frac{1}{2} \cdot \mu_2 \cdot h_k^2 \odot D^2\mathbf{f}_k(x_k)(1,1)\Big)\, dx_k \\ \oplus \int_0^1 \mu_2 \cdot h_k^2 \cdot \hat{\mathbf{M}}_{jj}(x_j)^{-1} \\ \cdot \Big(\frac{\partial}{\partial x_k}\mathbf{M}_{jk}(x_j, x_k)\Big) \odot D\mathbf{f}_k(x_k)(1) dx_k \\ \oplus o_p(n^{-2/5})\end{aligned} \tag{A.6}$$

uniformly for $x_j \in [0,1]$. Define

$$\begin{aligned}\hat{\boldsymbol{\Delta}}_j(x_j) = \hat{\mathbf{f}}_j(x_j) \ominus \mathbf{f}_j(x_j) \ominus \hat{\boldsymbol{\nu}}_j^A(x_j) \\ \ominus h_j \cdot \frac{\mu_{1,j}(x_j)}{\mu_{0,j}(x_j)} \odot D\mathbf{f}_j(x_j)(1) \\ \ominus \frac{1}{2} \cdot \mu_2 \cdot h_j^2 \odot D^2\mathbf{f}_j(x_j)(1,1) \oplus \mathbf{r}_j(x_j), \\ 1 \le j \le d_0.\end{aligned}$$

Also, define $\mathbf{b}_j(x_j)$ as we defined $\tilde{\boldsymbol{\beta}}_j(x_j)$ in Section 3.4 with c_j there being replaced by h_j. They are non-stochastic terms. Then, the expansions (A.5) and (A.6) entail

$$\begin{aligned}\hat{\boldsymbol{\Delta}}_j(x_j) = \mathbf{b}_j(x_j) \ominus \bigoplus_{k=1:\neq j}^{d_0} \int_0^1 \hat{\mathbf{M}}_{jj}(x_j)^{-1} \\ \cdot \hat{\mathbf{M}}_{jk}(x_j, x_k) \odot \hat{\boldsymbol{\Delta}}_k(x_k)\, dx_k \oplus o_p(n^{-2/5})\end{aligned} \tag{A.7}$$

uniformly for $x_j \in [0,1]$ for all $1 \le j \le d_0$.

Now, define the varying coefficient terms $\hat{\boldsymbol{\Delta}}_{j,\mathrm{vc}} : \mathcal{X} \to \mathbb{H}$, corresponding to $\hat{\boldsymbol{\Delta}}_j$, by

$$\hat{\boldsymbol{\Delta}}_{j,\mathrm{vc}}(\mathbf{x}) = \mathbf{z}_j^\top \odot \hat{\boldsymbol{\Delta}}_j(x_j),$$

where, as before, $\mathbf{z}_j \equiv \mathbf{z}_j(\mathbf{x})$ denotes the column real vector $(x_k : k \in I_j)$ for $\mathbf{x} = (x_1, \ldots, x_d) \in \mathcal{X}$. Let $\hat{\boldsymbol{\Delta}}_{\mathrm{vc}} = \bigoplus_{j=1}^{d_0} \hat{\boldsymbol{\Delta}}_{j,\mathrm{vc}}$. Then, (A.7) implies that $\hat{\boldsymbol{\Delta}}_{\mathrm{vc}} = \mathbf{b}_{\oplus,\mathrm{vc}} \oplus \hat{T}(\hat{\boldsymbol{\Delta}}_{\mathrm{vc}})$, where $\mathbf{b}_{\oplus,\mathrm{vc}} = \mathbf{b}_{d_0,\mathrm{vc}} \oplus \hat{\pi}_{d_0}^\perp(\mathbf{b}_{d_0-1,\mathrm{vc}}) \oplus (\hat{\pi}_{d_0}^\perp \circ \hat{\pi}_{d_0-1}^\perp)(\mathbf{b}_{d_0-2,\mathrm{vc}}) \oplus \cdots \oplus (\hat{\pi}_{d_0}^\perp \circ \cdots \circ \hat{\pi}_2^\perp)(\mathbf{b}_{1,\mathrm{vc}})$ and $\mathbf{b}_{j,\mathrm{vc}}(\mathbf{x}) = \mathbf{z}_j^\top \odot \mathbf{b}_j(x_j)$. We observe that $\|\mathbf{b}_{\oplus,\mathrm{vc}}\|_2 = O_p(n^{-2/5})$. Also, from Lemmas A.1 and A.2 we get that $\|\hat{T}\|_{\mathcal{L}(\mathscr{M}_{\mathrm{vc}}, \mathscr{M}_{\mathrm{vc}})} < 1$ with probability tending to one. This implies $\|\hat{\boldsymbol{\Delta}}_{\mathrm{vc}}\|_2 = O_p(n^{-2/5})$. Furthermore, Lemma A.2 allows us apply (3.10) to $\boldsymbol{\mu} = \hat{\boldsymbol{\Delta}}_{\mathrm{vc}}$. Thus, there exist $(\tilde{\boldsymbol{\Delta}}_{j,\mathrm{vc}} : 1 \le j \le d_0)$ such that $\hat{\boldsymbol{\Delta}}_{\mathrm{vc}} = \bigoplus_{j=1}^{d_0} \tilde{\boldsymbol{\Delta}}_{j,\mathrm{vc}}$ and $\max\{\|\tilde{\boldsymbol{\Delta}}_{j,\mathrm{vc}}\|_2 : 1 \le j \le d_0\} \le c \cdot \|\hat{\boldsymbol{\Delta}}_{\mathrm{vc}}\|_2$ for some absolute constant $0 < c < \infty$. This shows that $\|\tilde{\boldsymbol{\Delta}}_{j,\mathrm{vc}}\|_2 = O_p(n^{-2/5})$ for all $1 \le j \le d_0$. Because of the conditions (A1), we have

$$\begin{aligned}\|\tilde{\boldsymbol{\Delta}}_{j,\mathrm{vc}}\|_2^2 = \int_0^1 \langle \tilde{\boldsymbol{\Delta}}_j(x_j), \mathrm{E}(\mathbf{Z}_j\mathbf{Z}_j^\top \mid X_j = x_j) \odot \tilde{\boldsymbol{\Delta}}_j(x_j) \rangle_{\mathbb{H}^{|I_j|}} \\ \cdot p_j(x_j)\, dx_j \ge c_0 \int_0^1 \|\tilde{\boldsymbol{\Delta}}_j(x_j)\|_{\mathbb{H}^{|I_j|}}^2 p_j(x_j)\, dx_j\end{aligned}$$

for some constant $0 < c_0 < \infty$. Thus,

$$\int_0^1 \|\tilde{\boldsymbol{\Delta}}_j(x_j)\|_{\mathbb{H}^{|I_j|}}^2 p_j(x_j)\, dx_j = O_p(n^{-4/5})$$

for all $1 \le j \le d_0$. We claim

$$\int_0^1 \|\hat{\boldsymbol{\Delta}}_j(x_j) \ominus \tilde{\boldsymbol{\Delta}}_j(x_j)\|_{\mathbb{H}^{|I_j|}}^2 p_j(x_j)\, dx_j = O_p(n^{-4/5}) \tag{A.8}$$

for all $1 \le j \le d_0$. This establishes

$$\int_0^1 \|\hat{\boldsymbol{\Delta}}_j(x_j)\|^2_{\mathbb{H}^{|I_j|}} p_j(x_j)\,dx_j = O_p(n^{-4/5}) \tag{A.9}$$

for all $1 \le j \le d_0$. By applying the Hölder inequality to (A.7), we get

$$\begin{aligned}
&\sup_{x_j\in[0,1]} \|\hat{\boldsymbol{\Delta}}_j(x_j)\|_{\mathbb{H}^{|I_j|}} \\
&\quad\le \sup_{x_j\in[0,1]} \|\mathbf{b}_j(x_j)\|_{\mathbb{H}^{|I_j|}} \\
&\qquad + \sum_{k=1:\neq j}^{d_0} \left(\int_0^1 \|\hat{\boldsymbol{\Delta}}_k(x_k)\|^2_{\mathbb{H}^{|I_k|}} p_k(x_k)\,dx_k\right)^{1/2} \cdot O_p(1) \\
&\qquad + o_p(n^{-2/5}).
\end{aligned}$$

This with (A.9) and the fact that

$$\sup_{x_j\in[0,1]} \|\mathbf{b}_j(x_j)\|_{\mathbb{H}^{|I_j|}} = O(n^{-2/5})$$

implies

$$\sup_{x_j\in[0,1]} \|\hat{\boldsymbol{\Delta}}_j(x_j)\|_{\mathbb{H}^{|I_j|}} = O_p(n^{-2/5}) \tag{A.10}$$

for all $1 \le j \le d_0$. The theorem now follows from (A.10), the fact that $\mu_{1,j}(x_j) = 0$ for $x_j \in \mathrm{Int}_j$ and

$$\begin{aligned}
&\sup_{x_j\in[0,1]} \|\hat{\boldsymbol{\nu}}_j^A(x_j)\|_{\mathbb{H}^{|I_j|}} = O_p(n^{-2/5}\sqrt{\log n}), \\
&\int_0^1 \|\hat{\boldsymbol{\nu}}_j^A(x_j)\|^2_{\mathbb{H}^{|I_j|}} p_j(x_j)\,dx_j = O_p(n^{-4/5}).
\end{aligned}$$

It remains to prove (A.8). Note that $\hat{\boldsymbol{\Delta}}_{j,\mathrm{vc}}(\mathbf{x}) \ominus \tilde{\boldsymbol{\Delta}}_{j,\mathrm{vc}}(\mathbf{x}) = \mathbf{z}_j^\top \odot [\hat{\boldsymbol{\Delta}}_j(x_j) \ominus \tilde{\boldsymbol{\Delta}}_j(x_j)] = \bigoplus_{k\in I_j} x_k \odot [\hat{\boldsymbol{\Delta}}_{j,k}(x_j) \ominus \tilde{\boldsymbol{\Delta}}_{j,k}(x_j)]$. Since $\bigoplus_{j=1}^{d_0}(\hat{\boldsymbol{\Delta}}_{j,\mathrm{vc}} \ominus \tilde{\boldsymbol{\Delta}}_{j,\mathrm{vc}}) = \mathbf{0}$, the differences $\hat{\boldsymbol{\Delta}}_{j,k}(x_j) \ominus \tilde{\boldsymbol{\Delta}}_{j,k}(x_j)$ are constant for $d_0+1 \le k \le d$ and $j \in \tilde{I}_k$, and are linear in x_j for $d_0 - r + 1 \le k \le d_0$ and $j \in \tilde{I}_k$. We let

$$\begin{aligned}
\hat{\boldsymbol{\Delta}}_{j,k}(x_j) \ominus \tilde{\boldsymbol{\Delta}}_{j,k}(x_j) &= \mathbf{d}_{j,k}^{(0)}, \\
&\qquad j \in \tilde{I}_k, d_0+1 \le k \le d, \\
\hat{\boldsymbol{\Delta}}_{j,k}(x_j) \ominus \tilde{\boldsymbol{\Delta}}_{j,k}(x_j) &= \mathbf{d}_{j,k}^{(0)} \oplus x_j \odot \mathbf{d}_{j,k}^{(1)}, \\
&\qquad j \in \tilde{I}_k,\ d_0 - r + 1 \le k \le d_0
\end{aligned} \tag{A.11}$$

for some Hilbertian constants $\mathbf{d}_{j,k}^{(0)}$ and $\mathbf{d}_{j,k}^{(1)}$. From the definition of $\hat{\boldsymbol{\Delta}}_j$, the constraints (2.5) for both $\hat{\mathbf{f}}_j$ and $\mathbf{f}_j$ and $\int_0^1 \|\tilde{\boldsymbol{\Delta}}_j(x_j)\|^2_{\mathbb{H}^{|I_j|}} p_j(x_j)\,dx_j = O_p(n^{-4/5})$, we get that, for (j,k) with $k \in I_j$ and $d_0 - r + 1 \le k \le d_0$,

$$\begin{aligned}
&\int_0^1 w_j(x_j) \odot \big(\hat{\boldsymbol{\Delta}}_{j,k}(x_j) \ominus \tilde{\boldsymbol{\Delta}}_{j,k}(x_j)\big)\,dx_j \\
&\qquad = O_p(n^{-2/5}), \\
&\int_0^1 x_j w_j(x_j) \odot \big(\hat{\boldsymbol{\Delta}}_{j,k}(x_j) \ominus \tilde{\boldsymbol{\Delta}}_{j,k}(x_j)\big)\,dx_j \\
&\qquad = O_p(n^{-2/5}).
\end{aligned} \tag{A.12}$$

Also, for (j,k) with $k \in I_j$ and $d_0 + 1 \le k \le d$,

$$\begin{aligned}
&\int_0^1 w_j(x_j) \odot \big(\hat{\boldsymbol{\Delta}}_{j,k}(x_j) \ominus \tilde{\boldsymbol{\Delta}}_{j,k}(x_j)\big)\,dx_j \\
&\qquad = O_p(n^{-2/5}).
\end{aligned} \tag{A.13}$$

From the results at (A.12) and by multiplying $w_j(x_j)$ and $x_j w_j(x_j)$ to the right hand side of the second equation at (A.11) and then integrating them, we obtain that, for (j,k) with $k \in I_j$ and $d_0 - r + 1 \le k \le d_0$,

$$\begin{pmatrix} \int_0^1 w_j(x_j)\,dx_j & \int_0^1 x_j w_j(x_j)\,dx_j \\ \int_0^1 x_j w_j(x_j)\,dx_j & \int_0^1 x_j^2 w_j(x_j)\,dx_j \end{pmatrix} \odot \begin{pmatrix} \mathbf{d}_{j,k}^{(0)} \\ \mathbf{d}_{j,k}^{(1)} \end{pmatrix} = O_p(n^{-2/5})$$

This gives $\|\mathbf{d}_{j,k}^{(0)}\|_{\mathbb{H}} = O_p(n^{-2/5}) = \|\mathbf{d}_{j,k}^{(1)}\|_{\mathbb{H}}$ for all (j,k) with $k \in I_j$ and $d_0 - r + 1 \le k \le d_0$. Similarly, now using (A.13) and the first equation at (A.11), we may verify $\|\mathbf{d}_{j,k}^{(0)}\|_{\mathbb{H}} = O_p(n^{-2/5})$ for all (j,k) with $k \in I_j$ and $d_0+1 \le k \le d$. These and (A.11) imply that $\|\hat{\boldsymbol{\Delta}}_j(x_j) \ominus \tilde{\boldsymbol{\Delta}}_j(x_j)\|_{\mathbb{H}^{|I_j|}} = O_p(n^{-2/5})$ uniformly for $x_j \in [0,1]$, concluding the claim (A.8).

A.3 Proof of Theorem 3

We first note that the stochastic term $\hat{\boldsymbol{\nu}}_j^A(x_j)$ has the same asymptotic distribution as

$$\mathbf{S}_{n,j}(x_j) := n^{-3/5} \odot \bigoplus_{i=1}^n \mathbf{M}_{jj}(x_j)^{-1} \mathbf{Z}_{ij} K_{h_j}(x_j, X_{ij}) \odot \boldsymbol{\varepsilon}_i.$$

We write $\mathbf{x}^{\mathrm{N}} = (x_1, \ldots, x_{d_0})$ for $\mathbf{x} = (x_1, \ldots, x_d)$. Let $\mathbf{S}_n(\mathbf{x}^{\mathrm{N}}) = \big(\mathbf{S}_{n,1}(x_1)^\top, \ldots, \mathbf{S}_{n,d_0}(x_{d_0})\big)^\top$. Note that $\mathbf{S}_{n,j}$ is a map from $[0,1]$ to $\mathbb{H}^{|I_j|}$, and thus $\mathbf{S}_n$ is a map from $[0,1]^{d_0}$ to $\mathbb{H}^{|I_1|+\cdots+|I_{d_0}|}$. To characterize the asymptotic distribution of $\mathbf{S}_n(\mathbf{x}^{\mathrm{N}})$, we take the collection of $\mathbb{H}^{|I_1|+\cdots+|I_{d_0}|}$-vectors

$$\begin{aligned}
\mathbf{e}_{jk,l} := \big(&\mathbf{0}_{|I_1|+\cdots+|I_{j-1}|}^\top, \mathbf{e}_{k,l}^{(j)\top}, \mathbf{0}_{|I_{j+1}|+\cdots+|I_{d_0}|}^\top\big)^\top, \\
&1 \le j \le d_0,\ k \in I_j,\ l \ge 1
\end{aligned}$$

as an orthonormal basis of $\mathbb{H}^{|I_1|+\cdots+|I_{d_0}|}$. Using the standard kernel smoothing theory, we may show that

$$\begin{aligned}
&\mathrm{E}\big(\langle \mathbf{S}_n(\mathbf{x}^{\mathrm{N}}), \mathbf{e}_{jk,l}\rangle_{\mathbb{H}^*} \cdot \langle \mathbf{S}_n(\mathbf{x}^{\mathrm{N}}), \mathbf{e}_{j'k',l'}\rangle_{\mathbb{H}^*}\big) \to 0, \\
&\quad j \ne j',\ k \in I_j,\ k' \in I_{j'},\ l, l' \ge 1,
\end{aligned}$$

where we wrote $\mathbb{H}^*$ for $\mathbb{H}^{|I_1|+\cdots+|I_{d_0}|}$ for brevity. In the case where $j = j'$, we get that, for all $k, k' \in I_j$ and $l, l' \geq 1$,

$$\begin{aligned} &\mathrm{E}\big(\langle \mathbf{S}_n(\mathbf{x}^{\mathrm{N}}), \mathbf{e}_{jk,l}\rangle_{\mathbb{H}^*} \cdot \langle \mathbf{S}_n(\mathbf{x}^{\mathrm{N}}), \mathbf{e}_{jk',l'}\rangle_{\mathbb{H}^*}\big) \\ &= \mathrm{E}\big(\langle \mathbf{S}_{n,j}(x_j), \mathbf{e}^{(j)}_{k,l}\rangle_{\mathbb{H}^{|I_j|}} \cdot \langle \mathbf{S}_{n,j}(x_j), \mathbf{e}^{(j)}_{k',l'}\rangle_{\mathbb{H}^{|I_j|}}\big) \\ &\to c_j^{-1} \cdot \int K(t)^2\, dt \cdot \big(\mathbf{M}_{jj}(x_j)^{-1}\big)_{k,k'} \\ &\qquad \cdot \mathrm{E}\big(\langle \boldsymbol{\varepsilon}, \mathbf{e}_l\rangle \cdot \langle \boldsymbol{\varepsilon}, \mathbf{e}_{l'}\rangle_{\mathbb{H}} \,|\, X_j = x_j\big). \end{aligned} \tag{A.14}$$

Arguing as in the proof of Theorem 4.2 in Jeon and Park (2020) using (A.14) and Theorem 1.1 in Kundu et al. (2000), we may show that $\mathbf{S}_n(\mathbf{x}^{\mathrm{N}})$ converges to $\big(\mathbf{G}_j(\mathbf{0}_{|I_j|}, \mathcal{C}_j(\cdot, x_j)) : 1 \leq j \leq d_0\big)$. Then, we may proceed as in the proof of Theorem 4.3 in Jeon and Park (2020) to complete the proof of our Theorem 3.

APPENDIX B. ESTIMATION OF PARAMETRIC PART

Here, we show that the effect of estimating the parametric part in the model specification at (2.6) is negligible if the estimators of $\boldsymbol{\alpha}$'s converge to the corresponding $\boldsymbol{\alpha}$'s at a certain rate. We also present a construction of such estimators.

B.1 Effect of estimating parametric part

For a given set of Hilbertian constants $\mathbf{a}^0_{+,k}$ for $d_0 - r + 1 \leq k \leq d$ and $\mathbf{a}^1_{j,k}$ for $(j,k) \in J$, define

$$\mathbf{Y}_i(\mathbf{a}) = \mathbf{Y}_i \ominus \bigoplus_{k=d_0-r+1}^{d} X_{ik} \odot \mathbf{a}^0_{+,k} \ominus \bigoplus_{(j,k)\in J} X_{ij}X_{ik} \odot \mathbf{a}^1_{j,k}.$$

Let $\mathbf{a}^0_j$ and $\mathbf{a}^1_j$ be $|I_j|$-vectors such that

$$\begin{aligned} \bigoplus_{k=d_0-r+1}^{d} X_{ik} \odot \mathbf{a}^0_{+,k} &= \bigoplus_{j=1}^{d_0} \mathbf{Z}_{ij}^\top \odot \mathbf{a}^0_j, \\ \bigoplus_{(j,k)\in J} X_{ij} \cdot X_{ik} \odot \mathbf{a}^1_{j,k} &= \bigoplus_{j=1}^{d_0} X_{ij} \cdot \mathbf{Z}_{ij}^\top \odot \mathbf{a}^1_j. \end{aligned} \tag{B.1}$$

In fact, $\mathbf{a}^0_j = (\mathbf{a}^0_{j,k} : k \in I_j)^\top$ is the $|I_j|$-vector for some Hilbertian constants $\mathbf{a}^0_{j,k}$ such that $\bigoplus_{j\in\tilde{I}_k} \mathbf{a}^0_{j,k} = \mathbf{a}^0_{+,k}$. We may also identify $(\mathbf{a}^1_j : 1 \leq j \leq d_0)$ from $(\mathbf{a}^1_{j,k} : (j,k) \in J)$ based on the configuration of I_j. Define $\hat{\boldsymbol{\nu}}_j(\cdot, \mathbf{a})$ as $\hat{\boldsymbol{\nu}}_j$ at (2.11) with $\mathbf{Y}_i$ being replaced by $\mathbf{Y}_i(\mathbf{a})$. Let $(\hat{\mathbf{f}}_j(\cdot, \mathbf{a}) : 1 \leq j \leq d_0)$ be a tuple of component maps that satisfies

$$\begin{aligned} \hat{\mathbf{f}}_j(x_j, \mathbf{a}) &= \hat{\boldsymbol{\nu}}_j(x_j, \mathbf{a}) \\ &\quad \ominus \bigoplus_{l=1:\neq j}^{d_0} \int_0^1 \hat{\mathbf{M}}_{jj}(x_j)^{-1} \cdot \hat{\mathbf{M}}_{jl}(x_j, x_l) \\ &\qquad \odot \hat{\mathbf{f}}_l(x_l, \mathbf{a})\, dx_l, \quad 1 \leq j \leq d_0. \end{aligned} \tag{B.2}$$

Let $\hat{\boldsymbol{\alpha}}^0_j$ and $\hat{\boldsymbol{\alpha}}^1_j$ be some estimators of $\boldsymbol{\alpha}^0_j$ and $\boldsymbol{\alpha}^1_j$, respectively.

Theorem B.1. *Assume that the conditions (A1)–(A3) and (A8) hold. If $\|\hat{\boldsymbol{\alpha}}^1_j \ominus \boldsymbol{\alpha}^1_j\|_{\mathbb{H}^{|I_j|}} = o_p(n^{-1/5})$ for all $1 \leq j \leq d_0$, then the normalized version of $\hat{\mathbf{f}}_j(x_j, \hat{\boldsymbol{\alpha}})$ for each $1 \leq j \leq d_0$ according to the constraints (2.5) differs by $o_p(n^{-2/5})$, uniformly for $x_j \in [0,1]$, from the respective normalized version of $\hat{\mathbf{f}}_j(x_j, \boldsymbol{\alpha})$.*

The above theorem has an important implication that, in general, the effects of the errors in estimating the parametric part are negligible if their magnitudes are $o_p(n^{-1/5})$. Note that in the present case the errors of $\hat{\boldsymbol{\alpha}}$'s as the estimators of $\boldsymbol{\alpha}$'s are directly transmitted to the errors in the 'adjusted' responses $\mathbf{Y}_i(\hat{\boldsymbol{\alpha}})$ as the estimators of $\mathbf{Y}_i(\boldsymbol{\alpha})$. According to the standard theory in the smooth backfitting literature, the effects of errors in responses are negligible if the errors are of smaller order than the nonparametric rate $n^{-2/5}$, see Park et al. (2018) and Han et al. (2020), for example. The present case under study makes an example where the condition on errors in responses can be relaxed.

Theorem B.1 also indicates that, in case $r = 0$, i.e., no predictor appears both in linear and nonlinear parts so that $\boldsymbol{\alpha}^1_{j,k}$ do not appear in the model specification at (2.6), there is no need to estimate the parametric part at all. Thus, when $r = 0$, one can simply apply the method described in Section 2.3 directly to $\mathbf{Y}_i$. This would give the same normalized components as those obtained from $\mathbf{Y}_i(\boldsymbol{\alpha})$.

Proof of Theorem B.1. First, we observe from the representations at (B.1) that, for $1 \leq j \leq d_0$,

$$\begin{aligned} &\hat{\mathbf{M}}_{jj}(x_j) \odot \big(\hat{\boldsymbol{\nu}}_j(\hat{\boldsymbol{\alpha}}) \ominus \hat{\boldsymbol{\nu}}_j(\boldsymbol{\alpha})\big) \\ &\quad = \bigoplus_{i=1}^{n} n^{-1} \cdot \mathbf{Z}_{ij} \cdot K_{h_j}(x_j, X_{ij}) \odot \Bigg(\bigoplus_{l=1}^{d_0} \mathbf{Z}_{il}^\top \odot (\boldsymbol{\alpha}^0_l \ominus \hat{\boldsymbol{\alpha}}^0_l) \\ &\qquad \oplus \bigoplus_{l=1}^{d_0} X_{il} \cdot \mathbf{Z}_{il}^\top \odot (\boldsymbol{\alpha}^1_l \ominus \hat{\boldsymbol{\alpha}}^1_l) \Bigg). \end{aligned}$$

Using this and the property of the kernel that $\int_0^1 K_{h_j}(x_j, v)\, dx_j = 1$ for all $v \in [0,1]$, we may express the system of equations at (B.2) for $\mathbf{a} = \hat{\boldsymbol{\alpha}}$ as

$$\begin{aligned} &\hat{\mathbf{M}}_{jj}(x_j) \odot \Big[\hat{\mathbf{f}}_j(x_j, \hat{\boldsymbol{\alpha}}) \ominus (\boldsymbol{\alpha}^0_j \ominus \hat{\boldsymbol{\alpha}}^0_j) \\ &\qquad\qquad \ominus x_j \odot (\boldsymbol{\alpha}^1_j \ominus \hat{\boldsymbol{\alpha}}^1_j)\Big] \\ &\quad = \left[n^{-1} \sum_{i=1}^{n} \mathbf{Z}_{ij}\mathbf{Z}_{ij}^\top (X_{ij} - x_j) \cdot K_{h_j}(x_j, X_{ij})\right] \\ &\qquad\qquad \odot (\boldsymbol{\alpha}^1_j \ominus \hat{\boldsymbol{\alpha}}^1_j) \\ &\qquad \oplus \bigoplus_{l=1:\neq j}^{d_0} \left[n^{-1} \sum_{i=1}^{n} \int_0^1 \mathbf{Z}_{ij}\mathbf{Z}_{il}^\top (X_{il} - x_l) \right. \end{aligned}$$

(B.3)
$$\cdot K_{h_j}(x_j, X_{ij}) K_{h_l}(x_l, X_{il}) dx_l \Big]$$
$$\odot (\boldsymbol{\alpha}_j^1 \ominus \hat{\boldsymbol{\alpha}}_j^1)$$
$$\oplus \hat{\mathbf{M}}_{jj}(x_j) \odot \hat{\boldsymbol{\nu}}_j(\boldsymbol{\alpha})$$
$$\ominus \bigoplus_{l=1:\neq j}^{d_0} \int_0^1 \hat{\mathbf{M}}_{jl}(x_j, x_l) \odot \Big[\hat{\mathbf{f}}_l(x_l, \hat{\boldsymbol{\alpha}})$$
$$\ominus (\boldsymbol{\alpha}_l^0 \ominus \hat{\boldsymbol{\alpha}}_l^0) \ominus x_l \odot (\boldsymbol{\alpha}_l^1 \ominus \hat{\boldsymbol{\alpha}}_l^1)\Big] dx_l,$$
$$1 \le j \le d_0.$$

The sums in the brackets in the first two terms on the right hand side of (B.3) are of size $O_p(n^{-1/5})$ uniformly for $x_j \in [0,1]$ since

$$\sup_{x_l \in [0,1]} \max\{|X_{il} - x_l| : K_{h_l}(x_l, X_{il}) > 0,\ 1 \le i \le n\} \le h_l,$$
$$1 \le l \le d_0.$$

Thus, we get

$$\hat{\mathbf{f}}_j(x_j, \hat{\boldsymbol{\alpha}}) \ominus (\boldsymbol{\alpha}_j^0 \ominus \hat{\boldsymbol{\alpha}}_j^0) \ominus x_j \odot (\boldsymbol{\alpha}_j^1 \ominus \hat{\boldsymbol{\alpha}}_j^1)$$
$$= \hat{\boldsymbol{\nu}}_j(\boldsymbol{\alpha}) \ominus \bigoplus_{l=1:\neq j}^{d_0} \int_0^1 \hat{\mathbf{M}}_{jl}(x_j, x_l) \odot \Big[\hat{\mathbf{f}}_l(x_l, \hat{\boldsymbol{\alpha}})$$
$$\ominus (\boldsymbol{\alpha}_l^0 \ominus \hat{\boldsymbol{\alpha}}_l^0) \ominus x_l \odot (\boldsymbol{\alpha}_l^1 \ominus \hat{\boldsymbol{\alpha}}_l^1)\Big] dx_l \oplus o_p(n^{-2/5}),$$
$$1 \le j \le d_0$$

uniformly for $x_j \in [0,1]$. This implies that $\hat{\mathbf{f}}_j(\cdot, \hat{\boldsymbol{\alpha}})$ differ from the corresponding $\hat{\mathbf{f}}_j(\cdot, \boldsymbol{\alpha})$ by linear terms up to $o_p(n^{-2/5})$. The differences by linear terms vanish if we normalize the component functions according to (2.13). This completes the proof of the theorem. □

B.2 Construction of estimators

We first note that the index set $\{(j,k) : k \in I_j, 1 \le j \le d_0\}$ can be expressed as $\{(j,k) : j \in \tilde{I}_k, d_0 - r + 1 \le k \le d\}$. Let $\mathbf{X}^{\mathrm{N}} = (X_1, \ldots, X_{d_0})^\top$ be the vector of all predictors in the nonlinear part, and $\mathbf{X}^{\mathrm{R}} = (X_{d_0+1}, \ldots, X_d)^\top$ be the vector of the remaining predictors. Recall that X_k for $d_0 - r + 1 \le k \le d_0$ also appear in the linear part, and that some of X_k constituting $\mathbf{X}^{\mathrm{R}}$ may be of continuous-type. Let $\tilde{P}_{\mathbf{X}}$ denote the product measure such that $d\tilde{P}_{\mathbf{X}}(\mathbf{x}) = \prod_{j=1}^{d_0} w_j(x_j) dw_j(x_j) \times dP_{\mathbf{X}^{\mathrm{R}}}(\mathbf{x}^{\mathrm{R}})$. Because of the constraints (2.5), we obtain

$$\int x_l \odot \bigoplus_{j=1}^{d_0} \bigoplus_{k \in I_j} x_k \odot \mathbf{f}_{j,k}(x_j)\, d\tilde{P}_{\mathbf{X}}(\mathbf{x}) = 0, \quad d_0 - r + 1 \le l \le d, \tag{B.4}$$

$$\int x_l \cdot x_{l'} \odot \bigoplus_{j=1}^{d_0} \bigoplus_{k \in I_j} x_k \odot \mathbf{f}_{j,k}(x_j)\, d\tilde{P}_{\mathbf{X}}(\mathbf{x}) = 0, \quad (l, l') \in J.$$

For concise representation, let $\mathbf{vec}(\boldsymbol{\alpha})$ denote the vector of $\boldsymbol{\alpha}_{+,k}^0$ for $d_0 - r + 1 \le k \le d$ and $\boldsymbol{\alpha}_{j,k}^1$ for $(j,k) \in J$, enumerated in a certain order. Also, let $\mathbf{vec}(\mathbf{x})$ be the resulting vector of x_k for $d_0 - r + 1 \le k \le d$ and $x_j \cdot x_k$ for $(j,k) \in J$, enumerated in the same way as $\boldsymbol{\alpha}$'s. Then, (B.4) gives

$$\int_{[0,1]^{d_0}} \mathbf{vec}(\mathbf{x}) \odot \mathbf{m}(\mathbf{x})\, d\tilde{P}_{\mathbf{X}}(\mathbf{x}) = \left(\int_{[0,1]^{d_0}} \mathbf{vec}(\mathbf{x}) \cdot \mathbf{vec}(\mathbf{x})^\top\, d\tilde{P}_{\mathbf{X}}(\mathbf{x}) \right) \odot \mathbf{vec}(\boldsymbol{\alpha}). \tag{B.5}$$

Now, define

$$\mathbf{A}(\mathbf{x}^{\mathrm{R}}) = \int_{[0,1]^{d_0}} \mathbf{vec}(\mathbf{x}^{\mathrm{N}}, \mathbf{x}^{\mathrm{R}}) \cdot \mathbf{vec}(\mathbf{x}^{\mathrm{N}}, \mathbf{x}^{\mathrm{R}})^\top \cdot \prod_{j=1}^{d_0} w_j(x_j)\, dw_j(x_j),$$
$$\mathbf{a}(\mathbf{x}^{\mathrm{R}}) = \int_{[0,1]^{d_0}} \mathbf{vec}(\mathbf{x}^{\mathrm{N}}, \mathbf{x}^{\mathrm{R}}) \odot \mathrm{E}(\mathbf{Y} | \mathbf{X}^{\mathrm{N}} = \mathbf{x}^{\mathrm{N}}, \mathbf{X}^{\mathrm{R}} = \mathbf{x}^{\mathrm{R}}) \cdot \prod_{j=1}^{d_0} w_j(x_j)\, dw_j(x_j).$$

Then, the equation (B.5) may be written as

$$\mathrm{E}\, \mathbf{a}(\mathbf{X}^{\mathrm{R}}) = \mathrm{E}\, \mathbf{A}(\mathbf{X}^{\mathrm{R}}) \odot \mathbf{vec}(\boldsymbol{\alpha}). \tag{B.6}$$

We may prove that $\mathfrak{A} \equiv \mathrm{E}\, \mathbf{A}(\mathbf{X}^{\mathrm{R}})$ is positive definite under the assumptions (A0) and (A1). Let $\hat{\mathfrak{A}} = n^{-1} \sum_{i=1}^n \mathbf{A}(\mathbf{X}_i^{\mathrm{R}})$. Then, $\|\hat{\mathfrak{A}} - \mathfrak{A}\| = O_p(n^{-1/2})$ where $\|\cdot\|$ is either the Frobenius or the spectral norm. We may also find an estimator $\hat{\mathfrak{a}}$ of $\mathfrak{a} \equiv \mathrm{E}\, \mathbf{a}(\mathbf{X}^{\mathrm{R}})$ such that

$$\hat{\mathfrak{a}} \ominus \mathfrak{a} = o_p(n^{-1/5}). \tag{B.7}$$

By (B.7) we mean that the Hilbert norm $\|\cdot\|_{\mathbb{H}}$ of each entry of $\hat{\mathfrak{a}} \ominus \mathfrak{a}$ is $o_p(n^{-1/5})$. Then, the estimator $\hat{\mathfrak{A}}^{-1} \odot \hat{\mathfrak{a}}$ of $\mathbf{vec}(\boldsymbol{\alpha}) = \mathfrak{A}^{-1} \odot \mathfrak{a}$ satisfies $\hat{\mathfrak{A}}^{-1} \odot \hat{\mathfrak{a}} \ominus \mathbf{vec}(\boldsymbol{\alpha}) = o_p(n^{-1/5})$. Below, we construct an estimator $\hat{\mathfrak{a}}$ that fulfills (B.7).

We start with a full-dimensional nonparametric estimator $\tilde{\mathbf{m}}(\mathbf{x})$ of $\mathbf{m}(\mathbf{x})$. Suppose that we adopt the approach studied in Park et al. (2017) for real-valued responses with suitable modification for Hilbertian responses. The latter is based on a product kernel with smoothing parameters, say b_j, for the continuous-type predictors and those, say s_j, for the discrete-type predictors. Let $\hat{\mathbf{a}}(\cdot)$ be an estimator of $\mathbf{a}(\cdot)$

defined by

$$\hat{\mathbf{a}}(\mathbf{x}^{\mathrm{R}}) = \int_{[0,1]^{d_0}} \mathbf{vec}(\mathbf{x}^{\mathrm{N}}, \mathbf{x}^{\mathrm{R}}) \odot \tilde{\mathbf{m}}(\mathbf{x}^{\mathrm{N}}, \mathbf{x}^{\mathrm{R}}) \prod_{j=1}^{d_0} w_j(x_j)\, dw_j(x_j). \tag{B.8}$$

Define

$$\hat{\mathfrak{a}} = n^{-1} \odot \bigoplus_{i=1}^{n} \hat{\mathbf{a}}(\mathbf{X}_i^{\mathrm{R}}).$$

The above estimator $\hat{\mathfrak{a}}$ satisfies (B.7) if we let h_j and s_j converge to zero fast enough. To see this, we decompose $\tilde{\mathbf{m}}$ into $\tilde{\mathbf{m}} = \tilde{\mathbf{m}}^A + \tilde{\mathbf{m}}^B$, where $\tilde{\mathbf{m}}^A$ is a local smoother of $\mathbf{Y}_i - \mathbf{m}(\mathbf{X}_i)$ and $\tilde{\mathbf{m}}^B$ is the corresponding local smoother of $\mathbf{m}(\mathbf{X}_i)$. The stochastic part $\tilde{\mathbf{m}}^A(\mathbf{x}^{\mathrm{N}}, \mathbf{x}^{\mathrm{R}})$ is of the magnitude $n^{-1/2} \prod_{j \in \mathcal{I}^{\mathrm{c}}} b_j^{-1/2}$ for each point $(\mathbf{x}^{\mathrm{N}}, \mathbf{x}^{\mathrm{R}})$, where $\mathcal{I}^{\mathrm{c}}$ is the index set for the continuous-type predictors X_j. We note that its magnitude does not depend on s_j (Park et al., 2017). Now, define $\hat{\mathbf{a}}^A$ and $\hat{\mathbf{a}}^B$ as $\hat{\mathbf{a}}$ at (B.8) with $\tilde{\mathbf{m}}$ being replaced by $\tilde{\mathbf{m}}^A$ and $\tilde{\mathbf{m}}^B$, respectively. Then, the standard kernel smoothing theory shows that $\hat{\mathbf{a}}^A(\mathbf{x}^{\mathrm{R}})$ is of the magnitude $n^{-1/2} \prod_{j \in \mathcal{I}^{\mathrm{R,c}}} b_j^{-1/2}$ for each point $\mathbf{x}^{\mathrm{R}}$, where $\mathcal{I}^{\mathrm{R,c}} = \{j \in \mathcal{I}^{\mathrm{c}} : d_0 + 1 \le j \le d\}$. This is due to the integration over $\mathbf{x}^{\mathrm{N}} \in [0,1]^{d_0}$, see Lee et al. (2017), e.g., for a detailed argument in a similar but more involved problem. The bandwidth effect on the magnitude of the stochastic function $\hat{\mathbf{a}}^A(\cdot)$ is completely removed by averaging its values at $\mathbf{X}_i^{\mathrm{R}}$. Indeed, we may prove

$$n^{-1} \odot \bigoplus_{i=1}^{n} \hat{\mathbf{a}}^A(\mathbf{X}_i^{\mathrm{R}}) = O_p(n^{-1/2}). \tag{B.9}$$

On the other hand, the effects of the smoothing parameters b_j and s_j on the deterministic part $\tilde{\mathbf{m}}^B$ do not vanish by integration or averaging. However, we can make it negligible by choosing proper speeds of $b_j \to 0$ and $s_j \to 0$. In fact, if we take $b_j = o(n^{-1/10})$ and $s_j = o(n^{-1/5})$, then

$$n^{-1} \odot \bigoplus_{i=1}^{n} \hat{\mathbf{a}}^B(\mathbf{X}_i^{\mathrm{R}}) \ominus \mathfrak{a} = o_p(n^{-1/5}). \tag{B.10}$$

The properties (B.9) and (B.10) give (B.7).

Received 4 March 2020

REFERENCES

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press. MR1245941

Blot, J. and Cieutat, P. (2016). Completeness of sums of subspaces of bounded functions and applications. *Commun. Math. Anal.* **19** 43–61. MR3501515

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

Conway, J. B. (1985). *A Course in Functional Analysis*. Springer-Verlag, New York. MR0768926

Ferraty, F., Van Keilegom, I. and Vieu, P. (2012). Regression when both response and predictor are functions. *J. Multivariate Anal.* **109** 10–28. MR2922850

Han, K., Müller, H.-G. and Park, B. U. (2020). Additive functional regression for densities as responses. *J. Amer. Statist. Assoc.* **115** 997–1010. MR4107695

Han, K. and Park, B. U. (2018). Smooth backfitting for error-in-variables additive models. *Ann. Statist.* **46** 2216–2250. MR3845016

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *J. R. Statist. Soc. B* **55** 757–796. MR1229881

Jeon, J. M. and Park, B. U. (2020). Additive regression with Hilbertian responses. *Ann. Statist.* **48** 2671–2697. MR4152117

Kundu, S., Majumdar, S. and Mukherjee, K. (2000). Central limit theorems revisited. *Statist. Prob. Lett.* **47** 265–275. MR1747487

Lee, Y. K., Mammen, E., Nielsen, J. P. and Park, B. U. (2017). Operational time and in-sample density forecasting. *Ann. Statist.* **45** 1312–1341. MR3662456

Lee, Y. K., Mammen, E. and Park, B. U. (2010). Backfitting and smooth backfitting for additive quantile models. *Ann. Statist.* **38** 2857–2883. MR2722458

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

Lee, Y. K., Mammen, E. and Park, B. U. (2012b). Flexible generalized varying coefficient regression models. *Ann. Statist.* **40** 1906–1933. MR3015048

Lian, H. (2011). Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electron. J. Statist.* **5** 31–40. MR2773606

Lian, H. (2012). Convergence of nonparametric functional regression estimates with functional responses. *Electron. J. Statist.* **6** 1373–1391. MR2988451

Mammen, E., Linton, O. B. and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490. MR1742496

Mammen, E. and Park, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *Ann. Statist.* **33** 1260–1294. MR2195635

Park, B. U., Chen, C.-J., Tao, W. and Müller, H.-G. (2018). Singular additive models for function to function regression. *Stat. Sin.* **28** 2497–2520. MR3839871

Park, B. U., Simar, L. and Zelenyuk, V. (2017). Nonparametric estimation of dynamic discrete choice models for time series data. *Comput. Stat. Data An.* **108** 97–120. MR3589904

Park, B. U., Mammen, E., Lee, Y. K. and Lee, E. R. (2015). Varying coefficient regression models: A review and new developments (with discussion). *Int. Stat. Rev.* **83** 36–76. MR3341079

van den Boogaart, K. G., Egozcue, J. J. and Pawlowsky-Glahn, V. (2014). Bayes Hilbert spaces. *Aust. N. Z. J. Statist.* **56** 171–194. MR3226435

Xu, J. and Zikatanov, L. (2002). The method of alternating projections and the method of subspace corrections in Hilbert space. *J. Amer. Math. Soc.* **15** 573–597. MR1896233

Yang, L., Park, B. U., Xu, L. and Härdle, W. (2006). Estimation and testing for varying coefficients in additive models with marginal integration *J. Amer. Statist. Assoc.* **101** 1212–1227. MR2328308

Yu, K., Park, B. U. and Mammen, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36** 228–260. MR2387970

Young Kyung Lee
Department of Information Statistics
Kangwon National University
South Korea
E-mail address: youngklee@kangwon.ac.kr

Byeong U. Park
Department of Statistics
Seoul National University
South Korea
E-mail address: bupark@stats.snu.ac.kr

Hyerim Hong
Department of Statistics
Seoul National University
South Korea
E-mail address: hhong@snu.ac.kr

Dongwoo Kim
Department of Statistics
Seoul National University
South Korea
E-mail address: kdu91@snu.ac.kr