Bernoulli **29**(4), 2023, 3059–3092
https://doi.org/10.3150/22-BEJ1575

Additive regression with parametric help

HYERIM HONG[1,a], YOUNG KYUNG LEE[2,c] and BYEONG U. PARK[1,b]

[1]*Seoul National University, Seoul, South Korea,* [a]*hhong@snu.ac.kr,* [b]*bupark@snu.ac.kr*
[2]*Kangwon National University, Chuncheon, South Korea,* [c]*youngklee@kangwon.ac.kr*

Additive models have been studied as a way of overcoming theoretical and practical difficulties in estimating a multivariate nonparametric regression function. Several methods have been proposed that ensure the optimal univariate rate one can achieve in estimating univariate nonparametric functions. In this paper a new method is proposed which reduces the constant factor in the first-order approximation of the average squared error of the most successful existing method. The new estimator is based on an orthogonal decomposition of the underlying regression function, with an arbitrarily chosen parametric family, under a special inner product structure arising from the bias formula of the estimator. It is shown that the proposed method entails reduction in the constant factor of the leading bias of the existing method while it retains the same first-order variance. These theoretical findings are confirmed in Monte Carlo experiments.

Keywords: Additive model; bias reduction; local linear smoothing; parametric help; smooth backfitting

1. Introduction

In this paper, we propose a new method of estimating the additive regression model

$$Y = m_1(X_1) + \cdots + m_d(X_d) + \varepsilon, \tag{1.1}$$

where m_j are unknown component functions and $\mathrm{E}(\varepsilon|X_1,\ldots,X_d) = 0$. The model (1.1) is the simplest form of structured nonparametric models that successfully deal with the curse of dimensionality. Several kernel-based methods of fitting (1.1) have been proposed and studied, which include ordinary backfitting (Buja, Hastie and Tibshirani, 1989, Febrero-Bande and González-Manteiga, 2013, Opsomer, 2000, Opsomer and Ruppert, 1997, Sperlich, Linton and Härdle, 1999), marginal integration (Boente and Martínez, 2017, Lee, 2004, Linton and Nielsen, 1995, Sperlich, Linton and Härdle, 1999) and smooth backfitting (Huang and Yu, 2019, Jeon and Park, 2020, Jeon, Park and Van Keilegom, 2021a, Jeon et al., 2022, Mammen, Linton and Nielsen, 1999, Mammen and Park, 2006, Nielsen and Sperlich, 2005). It is widely accepted that smooth backfitting (SBF) has both theoretical and practical advantages, not shared by others. The idea of SBF has been further developed for other structured regression problems (Han, Müller and Park, 2018, 2020, Han and Park, 2018, Jeon, Park and Van Keilegom, 2021b, Lee, Mammen and Park, 2010, 2012, Lee et al., 2022, Linton, Sperlich and Van Keilegom, 2008, Park et al., 2015, 2018, Yang and Park, 2014, Yu, Mammen and Park, 2011, Yu, Park and Mammen, 2008, Zhang, Park and Wang, 2013).

The main idea of the new proposal is to consider a parametric family $\{g(\cdot,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^D\}$ with a known $g : \mathbb{R}^d \times \mathbb{R}^D \to \mathbb{R}$ such that $g(\mathbf{x},\boldsymbol{\theta}) = g_1(x_1,\boldsymbol{\theta}) + \cdots + g_d(x_d,\boldsymbol{\theta})$, and then apply the SBF technique to the *pseudo-response* $Y - g(\mathbf{X},\boldsymbol{\theta})$. This would give an estimator, say $\hat{m}_{Y-g(\mathbf{X},\boldsymbol{\theta})}$, of the *additive* function $m - g(\cdot,\boldsymbol{\theta})$, where $m(\mathbf{x}) = m_1(x_1) + \cdots + m_d(x_d)$, and thus produce a class of estimators of m: $\{\hat{m}_{Y-g(\mathbf{X},\boldsymbol{\theta})}(\cdot) + g(\cdot,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^D\}$. It is worthwhile to note here that $\hat{m}_{Y-g(\mathbf{X},\boldsymbol{\theta})} + g(\cdot,\boldsymbol{\theta}) \neq \hat{m}_Y$, where $\hat{m}_Y$ is the result of applying the SBF technique to the genuine response Y, unless $g(\cdot,\boldsymbol{\theta}) \equiv 0$ because of the effect of smoothing over $g(\mathbf{X}_i,\boldsymbol{\theta})$. It turns out that the variances of the two are the same in the first-order while their biases differ. To take the full advantage of the parametric family, we think of the

(unknown) parameter vector, say $\boldsymbol{\theta}_0$, that minimizes the bias of $\hat{m}_{Y-g(\mathbf{X},\boldsymbol{\theta})} + g(\cdot,\boldsymbol{\theta})$, or equivalently, the bias of $\hat{m}_{Y-g(\mathbf{X},\boldsymbol{\theta})}$ as an estimator of $m - g(\cdot,\boldsymbol{\theta})$. If g is chosen so that there exists $\boldsymbol{\theta}_{\text{null}} \in \mathbb{R}^D$ such that $g(\cdot,\boldsymbol{\theta}_{\text{null}}) \equiv 0$, then our approach with the help of $g(\cdot,\boldsymbol{\theta}_0)$ would improve the bias of the original SBF estimator $\hat{m}_Y$ since the class $\{\hat{m}_{Y-g(\mathbf{X},\boldsymbol{\theta})} + g(\cdot,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^D\}$ contains $\hat{m}_Y$.

To demonstrate the real advantages of the above approach in theory and in practice, we consider the local linear SBF and focus on the parametric family with $D = d$ and

$$g(\mathbf{x},\boldsymbol{\theta}) := \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{x}) = \theta_1 \cdot g_1(x_1) + \cdots + \theta_d \cdot g_d(x_d), \tag{1.2}$$

where and throughout this paper $\mathbf{g}(\mathbf{x}) := (g_1(x_1),\ldots,g_d(x_d))^\top$. We consider such a family since it is simple and easy to implement the idea of *parametric help* with. Since $\boldsymbol{\theta}_0$ is unknown, we replace $\boldsymbol{\theta}_0$ by a suitable estimator, say $\hat{\boldsymbol{\theta}}_0$. Then, the resulting parametrically-helped estimator of m is given by

$$\hat{m} := \hat{m}_{Y-\hat{\boldsymbol{\theta}}_0^\top \mathbf{g}(\mathbf{X})} + \hat{\boldsymbol{\theta}}_0^\top \mathbf{g}.$$

We show that $\hat{m}$ with $\hat{\boldsymbol{\theta}}_0$ such that $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0 + o_p(1)$ always improves the bias of the *direct* SBF estimator $\hat{m}_Y$, regardless of the choice of $\mathbf{g}$, while guaranteeing the same asymptotic variance as $\hat{m}_Y$ has.

We also implement the idea of parametric help for the estimation of the individual component functions m_j. For the latter problem, to identify m_j that we estimate, we put the constraints $\mathrm{E}\, m_j(X_j) = 0$. We find that the parameter vector $\boldsymbol{\theta}_0$ in estimating the additive function m is different from the one whose jth component minimizes the bias of the respective estimator of the individual centered m_j. We show that our proposals, now say $\hat{m}_j$, with an estimator of the latter $\boldsymbol{\theta}_0$ such that $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0 + o_p(1)$, also improve the biases of the corresponding estimators of m_j based on the application of the direct SBF, again regardless of the choice of $\mathbf{g}$. Both for estimating the additive regression function m and for estimating its components ($m_j : 1 \le j \le d$), we devise simple consistent estimators of the corresponding best parameter vectors $\boldsymbol{\theta}_0$.

We remark that the idea of parametric help has been implemented for the estimation of univariate regression functions by Lee et al. (2020). Although the present work is closely related to Lee et al. (2020), it is not straightforward from the latter work. Indeed, dealing with the additive structure of the model (1.1) and the backfitting procedure for estimating it make the problem far different from the development of the idea for estimating a single univariate function. It turns out that simply following the procedure of Lee et al. (2020) for each individual component m_j is not a right way for estimating the additive regression function m. This leads us to proposing two separate schemes for estimating m and m_j (Sections 4.1 and 4.2). Moreover, the method of estimating the optimal parameter value in the univariate regression of Lee et al. (2020) is found to be not applicable to the parametric help for estimating the individual m_j in the additive model (1.1). Furthermore, in the local linear SBF estimation of the individual components, the constraints to be put on the estimators $\hat{m}_j$ involve the estimators of the derivatives m_j'. Thus, the present problem requires a suitable way of implementing the parametric help for the estimators of m_j' as well.

We also remark that the idea of using a parametric family, which we develop for estimating the additive model (1.1), is related to the well-known two-step procedure (Fan, Wu and Feng, 2009, Glad, 1998, Gozalo and Linton, 2000, Hjort and Glad, 1995, Talamakrouni, El Ghouch and Van Keilegom, 2015, Talamakrouni, Van Keilegom and El Ghouch, 2016). However, the two approaches are quite different since the latter considers estimating $\boldsymbol{\theta}_0$ such that the underlying regression function is best approximated by $g(\cdot,\boldsymbol{\theta}_0)$. It is widely known that the success of the two-step procedure depends highly on the choice of g, contrary to the technique of the parametric help that we study here, see e.g. Lee et al. (2020). For this reason, we do not consider the two-step procedure for additive regression although the problem has never been studied in the literature.

In the next section we collect some general features of SBF additive regression that we use to motivate our proposal and to develop the associated theory. We present the main idea of the parametric help in Section 3, and discuss its realization in Section 4. We report the results of Monte Carlo experiments in Section 5, and give brief remarks on some extensions in Section 6. We defer all technical proofs to the Appendix.

2. Smooth backfitting operation

Throughout this paper, we assume that the covariate vector $\mathbf{X} \equiv (X_1, \ldots, X_d)$ has a density p supported on $[0,1]^d$, and each X_j has p_j on $[0,1]$, with respect to the corresponding Lebesgue measures. For a general random variable W, we let m_W denote $\mathrm{E}(W|\mathbf{X} = \cdot)$. For various specifications of W for which m_W is an *additive function*, i.e., $m_W(\mathbf{x}) = m_{W,1}(x_1) + \cdots + m_{W,d}(x_d)$, we consider the *local linear* SBF estimator of m_W, which we denote by $\hat{m}_W$. An initial theory for the local linear SBF method was studied by Mammen, Linton and Nielsen (1999), and recently an extensive theory for the local polynomial SBF method and for Hilbertian responses has been developed by Jeon et al. (2022). In this section we introduce some of the theory that are essential for motivating our proposals and for developing their theory to be presented in Sections 3 and 4.

2.1. Local linear smooth backfitting

We describe the local linear SBF method of estimating the additive function $m_W = m_{W,1} + \cdots + m_{W,d}$ and its (scaled) partial derivatives. Throughout this paper, we use the convention of expressing an additive function $f : [0,1]^d \to \mathbb{R}$ with $f(\mathbf{x}) = f_1(x_1) + \cdots + f_d(x_d)$, simply by $f = f_1 + \cdots + f_d$. In the latter expression, we interpret $f_j : [0,1] \to \mathbb{R}$ as $f_j^{\mathrm{E}} : [0,1]^d \to \mathbb{R}$ where $f_j^{\mathrm{E}}(\mathbf{x}) = f_j(x_j)$.

Since the local linear SBF involves derivative estimation as well, we describe the method in terms of estimating a $(d+1)$-tuple of functions. This is in contrast with the case of the Nadaraya-Watson SBF method, where the corresponding SBF technique is to estimate a single additive function. Let $m_{W,1,j}(u) = h \cdot m'_{W,j}(u)$, where η' for a univariate function η denotes its first derivative and h is the bandwidth we use in kernel smoothing, see below. We consider the $(d+1)$-tuple of functions

$$\mathbf{m}_W^{\mathrm{tp}}(\mathbf{x}) := \big(m_W(\mathbf{x}), m_{W,1,1}(x_1), \ldots, m_{W,1,d}(x_d)\big)^\top,$$
$$\mathbf{m}_{W,j}^{\mathrm{tp}}(x_j) := \big(m_{W,j}(x_j), 0, \ldots, 0, m_{W,1,j}(x_j), 0, \ldots, 0\big)^\top,$$

where $m_{W,1,j}(x_j)$ appears as the $(j+1)$th entry. Then, $\mathbf{m}_W^{\mathrm{tp}}(\mathbf{x}) = \mathbf{m}_{W,1}^{\mathrm{tp}}(x_1) + \cdots + \mathbf{m}_{W,d}^{\mathrm{tp}}(x_d)$. There is one-to-one correspondence between the $(d+1)$-tuple $\mathbf{m}_{W,j}^{\mathrm{tp}}$ and the 2-tuple $(m_{W,j}, m_{W,1,j})^\top$ via

$$\big(m_{W,j}(x_j), m_{W,1,j}(x_j)\big)^\top = \mathbf{U}_j \mathbf{m}_{W,j}^{\mathrm{tp}}(x_j), \quad 1 \le j \le d$$
$$\mathbf{m}_{W,j}^{\mathrm{tp}}(x_j) = \mathbf{U}_j^\top \big(m_{W,j}(x_j), m_{W,1,j}(x_j)\big)^\top, \quad 1 \le j \le d,$$

where $\mathbf{U}_j := (\mathbf{u}_1, \mathbf{0}, \ldots, \mathbf{0}, \mathbf{u}_2, \mathbf{0}, \ldots, \mathbf{0})$ are $2 \times (d+1)$ matrices with $\mathbf{u}_1 = (1,0)^\top$ as the first column and $\mathbf{u}_2 = (0,1)^\top$ as the $(j+1)$th column.

For a baseline kernel function K supported on $[-1,1]$, define $K_h(u) := h^{-1}K(u/h)$ and the normalized kernel

$$K_h(u,v) := \frac{K_h(u-v)}{\int_0^1 K_h(t-v)\,dt} \cdot I_{[0,1]^2}(u,v).$$

The normalized kernel has been used in the smooth backfitting literature, e.g., Jeon and Park (2020), Jeon et al. (2022), Mammen, Linton and Nielsen (1999), Yu, Park and Mammen (2008). It holds that

$$\begin{aligned} &K_h(u,v) = K_h(u-v), \quad (u,v) \in [2h, 1-2h] \times [0,1], \\ &\int_0^1 K_h(u,v)\,du = 1, \quad v \in [0,1]. \end{aligned} \tag{2.1}$$

Because of the first identity in (2.1) and the fact that $\int_0^1 K_h(u-v)\,dv = \int_{-1}^1 K(t)\,dt$ for all $u \in [h, 1-h]$, the interval $[2h, 1-2h]$ is called the *interior* region of the support $[0,1]$ of p_j in SBF estimation. Writing $\mathbf{vec}(u) := (1,u)^\top$, define 2×2 matrices of functions

$$\hat{\mathbf{M}}_{jj}(x_j) := n^{-1} \sum_{i=1}^n \mathbf{vec}\left(\frac{X_{ij}-x_j}{h}\right) \mathbf{vec}\left(\frac{X_{ij}-x_j}{h}\right)^\top \cdot K_h(x_j, X_{ij}),$$

$$\hat{\mathbf{M}}_{jk}(x_j, x_k) := n^{-1} \sum_{i=1}^n \mathbf{vec}\left(\frac{X_{ij}-x_j}{h}\right) \mathbf{vec}\left(\frac{X_{ik}-x_k}{h}\right)^\top \cdot K_h(x_j, X_{ij}) K_h(x_k, X_{ik}).$$

Also, define $\hat{p}_j$ and $\hat{p}_{1,j}$ by

$$\left(\hat{p}_j(x_j), \hat{p}_{1,j}(x_j)\right)^\top := n^{-1} \sum_{i=1}^n \mathbf{vec}\left(\frac{X_{ij}-x_j}{h}\right) K_h(x_j, X_{ij}), \quad 1 \le j \le d.$$

As an estimator of the 2-tuple of the *marginal* regression function $\mathrm{E}(W|X_j = x_j)$ and its first derivative multiplied by h, i.e., $h \cdot (d/dx_j)\mathrm{E}(W|X_j = x_j)$, define

$$\tilde{\mathbf{m}}_{W,j}(x_j) := \hat{\mathbf{M}}_{jj}(x_j)^{-1} \cdot n^{-1} \sum_{i=1}^n \mathbf{vec}\left(\frac{X_{ij}-x_j}{h}\right) \cdot K_h(x_j, X_{ij}) \cdot W_i. \tag{2.2}$$

Now, we define the local linear SBF estimator $\hat{\mathbf{m}}_W^{\mathrm{tp}}(\mathbf{x}) \equiv (\hat{m}_W, \hat{m}_{W,1,1}, \ldots, \hat{m}_{W,1,d})^\top$ of $\mathbf{m}_W^{\mathrm{tp}}$, where $\hat{m}_W$ is an additive (stochastic) function taking the form $\hat{m}_W = \hat{m}_{W,1} + \cdots + \hat{m}_{W,d}$ for some univariate (stochastic) functions $\hat{m}_{W,j}$. It is defined as $\hat{\mathbf{m}}_W^{\mathrm{tp}} := \hat{\mathbf{m}}_{W,1}^{\mathrm{tp}} + \cdots + \hat{\mathbf{m}}_{W,d}^{\mathrm{tp}}$ with

$$\hat{\mathbf{m}}_{W,j}^{\mathrm{tp}}(x_j) = \left(\hat{m}_{W,j}(x_j), 0, \ldots, 0, \hat{m}_{W,1,j}(x_j), 0, \ldots, 0\right)^\top,$$

where $(\hat{\mathbf{m}}_{W,1}^{\mathrm{tp}}, \ldots, \hat{\mathbf{m}}_{W,d}^{\mathrm{tp}})$ solves the following system of SBF equations: for $1 \le j \le d$,

$$\hat{\mathbf{m}}_{W,j}^{\mathrm{tp}}(x_j) = \mathbf{U}_j^\top \tilde{\mathbf{m}}_{W,j}(x_j) - \sum_{k=1, \neq j}^d \int_0^1 \mathbf{U}_j^\top \hat{\mathbf{M}}_{jj}(x_j)^{-1} \hat{\mathbf{M}}_{jk}(x_j, x_k) \mathbf{U}_k \hat{\mathbf{m}}_{W,k}^{\mathrm{tp}}(x_k)\,dx_k. \tag{2.3}$$

We note that the SBF system (2.3) gives an estimator of $\mathbf{m}_W^{\mathrm{tp}}$, not those of the individual $(d+1)$-tuples $\mathbf{m}_{W,j}^{\mathrm{tp}}$ for $1 \le j \le d$, or equivalently those of the individual 2-tuples $(m_{W,j}(x_j), m_{W,1,j}(x_j))$. In fact, the individual functions $m_{W,j}$ summing up to m_W, themselves, are not identifiable.

We note that the SBF system (2.3), expressed in terms of $(d+1)$-tuples of functions $\hat{\mathbf{m}}_{W,j}^{\mathrm{tp}}$, is convenient later in our theoretical development. Observing that $\mathbf{U}_j \hat{\mathbf{m}}_{W,j}^{\mathrm{tp}}(x_j) = (\hat{m}_{W,j}(x_j), \hat{m}_{W,1,j}(x_j))^\top$ and

$\mathbf{U}_j\mathbf{U}_j^\top$ is the two-dimensional identity matrix, we see that (2.3) is equivalent to the following equation, now expressed in terms of 2-tuples of functions: for $1 \le j \le d$,

$$\begin{pmatrix} \hat{m}_{W,j}(x_j) \\ \hat{m}_{W,1,j}(x_j) \end{pmatrix} = \tilde{\mathbf{m}}_{W,j}(x_j) - \sum_{k=1,\neq j}^{d} \int_0^1 \hat{\mathbf{M}}_{jj}(x_j)^{-1} \hat{\mathbf{M}}_{jk}(x_j, x_k) \begin{pmatrix} \hat{m}_{W,k}(x_k) \\ \hat{m}_{W,1,k}(x_k) \end{pmatrix} dx_k. \tag{2.4}$$

The above SBF equation is more convenient than (2.3) in practical implementation.

We now turn to the estimation of the individual $m_{W,j}$ and $m_{W,1,j}$ for $1 \le j \le d$. We note that $m_{W,1,j}$ and $\hat{m}_{W,1,j}$ are identifiable, but $m_{W,j}$ and $\hat{m}_{W,j}$ are determined only up to a constant, see Lemma 3 in the Appendix C. As identifiable components, we consider the *centered* versions $m^c_{W,j}$ such that $\mathrm{E}(m^c_{W,j}(X_j)) = 0$. For their estimators $\hat{m}^c_{W,j}$, we put the following constraints:

$$\int_0^1 \left(\hat{m}^c_{W,j}(u)\hat{p}_j(u) + \hat{m}_{W,1,j}(u)\hat{p}_{1,j}(u)\right) du = 0, \quad 1 \le j \le d. \tag{2.5}$$

For any $(d+1)$-tuple $(\hat{m}_{W,1}, \ldots, \hat{m}_{W,d})$, which comprises $\hat{m}_W = \hat{m}_{W,1} + \cdots + \hat{m}_{W,d}$, the centered versions $\hat{m}^c_{W,j}$ are uniquely determined by

$$\hat{m}^c_{W,j}(x_j) := \hat{m}_{W,j}(x_j) - \int_0^1 \left(\hat{m}_{W,j}(u)\hat{p}_j(u) + \hat{m}_{W,1,j}(u)\hat{p}_{1,j}(u)\right) du. \tag{2.6}$$

The constraints at (2.5) are well motivated in Jeon et al. (2022).

2.2. Main theory of local linear SBF technique

We base our theory for the SBF method on the following basic assumptions on the density p of $\mathbf{X}$, the baseline kernel K and the bandwidth h.

- (A1) The density function p of $\mathbf{X}$ is bounded away from zero and infinity on $[0,1]^d$, and the two-dimensional joint densities p_{jk} of (X_j, X_k) for $1 \le j \ne k \le d$ are continuous on $[0,1]^2$.
- (A2) The baseline kernel $K \ge 0$ is symmetric, Lipschitz continuous, vanishes on $\mathbb{R} \setminus [-1,1]$, and $0 < \int_{-1}^1 K(u)\,du < \infty$. Without loss of generality, we assume $\int_{-1}^1 K(u)\,du = 1$.
- (A3) The bandwidth h is asymptotic to $n^{-1/5}$.

The above assumptions are standard in additive kernel regression. We may lift the symmetry assumption on K with more sophisticated expression for the asymptotic bias terms of our estimators, but we assume it just for simplicity. Without symmetry, we need to assume that K is supported on a nontrivial interval in both of $[0,1]$ and $[-1,0]$.

In the following proposition, the first part is the specialization of Theorem 1 in Jeon et al. (2022) to local linear SBF. It demonstrates that the SBF equation (2.3) determines $\hat{\mathbf{m}}^{\mathrm{tp}}_W$ uniquely, where the uniqueness means that, if $\mathbf{f}^{\mathrm{tp}}$ and $\tilde{\mathbf{f}}^{\mathrm{tp}}$ solves the equation (2.3), then $\mathbf{f}^{\mathrm{tp}} = \tilde{\mathbf{f}}^{\mathrm{tp}}$ a.e. on $[0,1]^d$. For such tuples, if both are continuous on $[0,1]^d$, then $\mathbf{f}^{\mathrm{tp}} \equiv \tilde{\mathbf{f}}^{\mathrm{tp}}$. The second part shows that $\hat{\mathbf{m}}^{\mathrm{tp}}_W$ is linear in the response variable W. It helps to motivate an optimal way of utilizing the parametric model at (1.2) in additive regression, and provides an easy way of implementing the parametrically-helped estimators, see Section 4.

Proposition 1. *Assume (A1)–(A3) and that* $\mathrm{E}(|\varepsilon_W|^\alpha) < \infty$ *for some* $\alpha > 2$, *where* $\varepsilon_W := W - \mathrm{E}(W|\mathbf{X})$. *Then, with probability tending to one, the SBF equation (2.3) has a unique solution in the space of*

additive functions. Furthermore, the solution is linear in W: for any constants c_j and random variables W_j, it holds that $\hat{\mathbf{m}}^{\mathrm{tp}}_{c_1W_1+c_2W_2} = c_1 \cdot \hat{\mathbf{m}}^{\mathrm{tp}}_{W_1} + c_2 \cdot \hat{\mathbf{m}}^{\mathrm{tp}}_{W_2}$ with probability tending to one.

The next proposition presents a stochastic expansion of $\hat{\mathbf{m}}^{\mathrm{tp}}_{W}$, whose proof is given in the Appendix. To state the proposition, put

$$\tilde{\mathbf{m}}^{A}_{W,j}(x_j) := \hat{\mathbf{M}}_{jj}(x_j)^{-1} \cdot n^{-1} \sum_{i=1}^{n} \mathbf{vec}\left(\frac{X_{ij} - x_j}{h}\right) \cdot K_h(x_j, X_{ij}) \cdot \varepsilon_{W,i},$$

where $\varepsilon_{W,i} := W_i - \mathrm{E}(W_i|\mathbf{X}_i)$. Define

$$\mathbf{N}(u) := \int_0^1 \mathbf{vec}\left(\frac{v-u}{h}\right)\mathbf{vec}\left(\frac{v-u}{h}\right)^{\top} K_h(u,v)\,dv,$$

$$\boldsymbol{\gamma}(u) := \int_0^1 \left(\frac{v-u}{h}\right)^2 \mathbf{vec}\left(\frac{v-u}{h}\right) K_h(u,v)\,dv.$$

Proposition 2. *Assume (A1)–(A3) and that $\mathrm{E}(|\varepsilon_W|^{\alpha}) < \infty$ for some $\alpha > 5/2$ and $\mathrm{E}(\varepsilon_W^2|\mathbf{X} = \cdot)$ is bounded on $[0,1]^d$. If $m_{W,j}$ for $1 \le j \le d$ are twice continuously differentiable on $[0,1]$, then*

$$\begin{aligned}&\hat{\mathbf{m}}^{\mathrm{tp}}_{W}(\mathbf{x}) - \left(m_W(\mathbf{x}), m_{W,1,1}(x_1), \ldots, m_{W,1,d}(x_d)\right)^{\top} \\ &\quad = \sum_{j=1}^{d} \mathbf{U}_j^{\top} \tilde{\mathbf{m}}^{A}_{W,j}(x_j) + \frac{h^2}{2} \sum_{j=1}^{d} \mathbf{U}_j^{\top} \mathbf{N}(x_j)^{-1} \boldsymbol{\gamma}(x_j) \cdot m''_{W,j}(x_j) + o_p\left(n^{-2/5}\right)\end{aligned}$$

uniformly for $\mathbf{x} \in [0,1]^d$.

Next, we demonstrate the stochastic expansions of the individual component tuples. Since $\mathbf{m}^{\mathrm{tp}}_{W,j}$ are not identifiable, we consider their centered versions. With $m^{c}_{W,j}$ such that $\mathrm{E}(m^{c}_{W,j}(X_j)) = 0$ and the centered $\hat{m}^{c}_{W,j}$ defined at (2.6), let

$$\mathbf{m}^{c,\mathrm{tp}}_{W,j}(x_j) := \left(m^{c}_{W,j}(x_j), 0, \ldots, 0, m_{W,1,j}(x_j), 0, \ldots, 0\right)^{\top},$$

$$\hat{\mathbf{m}}^{c,\mathrm{tp}}_{W,j}(x_j) := \left(\hat{m}^{c}_{W,j}(x_j), 0, \ldots, 0, \hat{m}_{W,1,j}(x_j), 0, \ldots, 0\right)^{\top}.$$

Proposition 3. *Assume the conditions of Proposition 2. If $m_{W,j}$ for $1 \le j \le d$ are twice continuously differentiable on $[0,1]$, then*

$$\hat{\mathbf{m}}^{c,\mathrm{tp}}_{W,j}(x_j) - \mathbf{m}^{c,\mathrm{tp}}_{W,j}(x_j) = \mathbf{U}_j^{\top} \tilde{\mathbf{m}}^{A}_{W,j}(x_j) + \frac{h^2}{2} \cdot \mathbf{U}_j^{\top} \mathbf{N}(x_j)^{-1} \boldsymbol{\gamma}(x_j) \cdot m''_{W,j}(x_j) + o_p\left(n^{-2/5}\right)$$

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

The proof of the above proposition is given in the Appendix. We note that Proposition 3 is not direct from Proposition 2. We need to make it clear how the constraints (2.5) for $\hat{m}^{c}_{W,j}$ and $\mathrm{E}(m^{c}_{W,j}(X_j)) = 0$ for $m^{c}_{W,j}$ affect the stochastic expansion of $\hat{m}^{c}_{W,j} - m^{c}_{W,j}$.

The following corollary is for a special case where W itself is an additive function of $\mathbf{X}$, say $\eta(\mathbf{X}) := \eta_1(X_1) + \cdots + \eta_d(X_d)$. Let $\hat{\mathbf{m}}^{\mathrm{tp}}_{\eta(\mathbf{X})}$ be the solution of the SBF equation at (2.3) with W_i being replaced

by $\eta(\mathbf{X}_i)$. Even though $\eta : [0,1]^d \to \mathbb{R}$ is additive and the response observations $\eta(\mathbf{X}_i)$ do not contain error since $\varepsilon_{\eta(\mathbf{X})} \equiv 0$, $\hat{m}_{\eta(\mathbf{X})}(\mathbf{x})$ is not equal to $\eta(\mathbf{x})$ because of smoothing over $\mathbf{X}_i$, where $\hat{m}_{\eta(\mathbf{X})}(\mathbf{x})$ is the first entry of $\hat{\mathbf{m}}^{\mathrm{tp}}_{\eta(\mathbf{X})}(\mathbf{x})$. Let $\eta_{1,j}(x_j) := h\eta_j'(x_j)$. Consider the centered version of η_j, denoted by η_j^c, such that $\mathrm{E}(\eta_j^c(X_j)) = 0$. Also, let $\hat{m}^c_{\eta(\mathbf{X}),j}$ be the centered components of $\hat{m}_{\eta(\mathbf{X})}$ such that

$$\int_0^1 \left(\hat{m}^c_{\eta(\mathbf{X}),j}(x_j)\hat{p}_j(x_j) + \hat{m}_{\eta(\mathbf{X}),1,j}(x_j)\hat{p}_{1,j}(x_j)\right) dx_j = 0.$$

Noting that $\varepsilon_{\eta(\mathbf{X})} \equiv 0$, the following corollary is immediate from Propositions 2 and 3.

Corollary 1. *Assume (A1)–(A3). Suppose that η_j for $1 \le j \le d$ are twice continuously differentiable on $[0,1]$. Then,*

$$\hat{\mathbf{m}}^{c,\mathrm{tp}}_{\eta(\mathbf{X}),j}(x_j) - \mathbf{U}_j^\top \left(\eta_j^c(x_j), \eta_{1,j}(x_j)\right)^\top = \frac{h^2}{2} \cdot \mathbf{U}_j^\top \mathbf{N}(x_j)^{-1} \boldsymbol{\gamma}(x_j) \cdot \eta_j''(x_j) + o_p\left(n^{-2/5}\right)$$

uniformly for $x_j \in [0,1]$. Furthermore, uniformly for $\mathbf{x} \in [0,1]^d$,

$$\hat{\mathbf{m}}^{\mathrm{tp}}_{\eta(\mathbf{X})}(\mathbf{x}) - \left(\eta(\mathbf{x}), \eta_{1,1}(x_1), \ldots, \eta_{1,d}(x_d)\right)^\top = \frac{h^2}{2} \sum_{j=1}^{d} \mathbf{U}_j^\top \mathbf{N}(x_j)^{-1} \boldsymbol{\gamma}(x_j) \cdot \eta_j''(x_j) + o_p\left(n^{-2/5}\right).$$

With the second part of Proposition 1, the above corollary plays an important role in motivating the idea of the parametric help. It also serves as a basic ingredient in developing the theory for the parametrically-helped estimators in Section 4.

3. Parametric help in additive regression

Let the regression function be denoted by $m = \mathrm{E}(Y|\mathbf{X} = \cdot) : [0,1]^d \to \mathbb{R}$. Assume $\mathrm{E}(m(\mathbf{X})^2) < \infty$ and

$$m(\mathbf{x}) = m_1(x_1) + \cdots + m_d(x_d) \tag{3.1}$$

for some univariate functions m_j. We note that the individual components m_j are determined under some constraints. In this section we discuss a new approach to additive regression, i.e., the estimation of m and its components m_j. The new scheme takes the advantages of the parametric help, which we describe below.

We choose a d-tuple of twice differentiable univariate functions, $(g_1, \ldots, g_d)$, and let $\mathbf{g} : [0,1]^d \to \mathbb{R}^d$ defined by $\mathbf{g}(\mathbf{x}) := (g_1(x_1), \ldots, g_d(x_d))^\top$. Our approach is first to apply the SBF technique, described in Section 2, to $W = Y - \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X})$ for an arbitrary $\boldsymbol{\theta}$ to get $\hat{m}_{Y-\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X})}$ and form

$$\hat{m}(\mathbf{x}, \boldsymbol{\theta}) := \hat{m}_{Y-\boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X})}(\mathbf{x}) + \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{x})$$

as a candidate estimator of m. We consider some $\boldsymbol{\theta}_0$ that gives the largest bias reduction to $\hat{m}(\cdot, \boldsymbol{\theta}_0)$ in comparison with $\hat{m}_Y$, or to its additive components $\hat{m}_j(\cdot, \boldsymbol{\theta}_0)$ in comparison with $\hat{m}_{Y,j}$. Then, we find a suitable estimator $\hat{\boldsymbol{\theta}}_0$ of $\boldsymbol{\theta}_0$, and investigate whether the largest bias reduction achieved by $\hat{m}(\cdot, \boldsymbol{\theta}_0)$ or by $\hat{m}_j(\cdot, \boldsymbol{\theta}_0)$ is retained by $\hat{m}(\cdot, \hat{\boldsymbol{\theta}}_0)$ or by $\hat{m}_j(\cdot, \hat{\boldsymbol{\theta}}_0)$, respectively. In this section, we develop this idea with a general $\boldsymbol{\theta}_0$ and a consistent estimator $\hat{\boldsymbol{\theta}}_0$ of $\boldsymbol{\theta}_0$. Our development in this section is a foundation to the method and theory in Section 4 where we specialize $\boldsymbol{\theta}_0$ for the estimation of the regression function m and for the estimation of its identifiable components m_j, to achieve the best bias properties.

3.1. Estimation of additive regression function

Let $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ be a general unknown vector that depends on the unknown distribution of $(\mathbf{X}, Y)$ as well as $\mathbf{g}$, and $\hat{\boldsymbol{\theta}}_0$ be its estimator. We consider the following estimator of m:

$$\hat{m} := \hat{m}_{Y-\hat{\boldsymbol{\theta}}_0^\top \mathbf{g}(\mathbf{X})} + \hat{\boldsymbol{\theta}}_0^\top \mathbf{g}, \tag{3.2}$$

which we call the *local linear SBF estimator with parametric help*. Our theory for $\hat{m}$ at (3.2) is developed in conjunction with the estimation of the partial derivatives of m. Recall that the local linear smoothing technique involves its (partial) derivatives.

Let $m_{1,j}(x_j) := h \cdot m_j'(x_j)$ and put $\mathbf{m}^{\mathrm{tp}}(\mathbf{x}) := (m(\mathbf{x}), m_{1,1}(x_1), \ldots, m_{1,d}(x_d))^\top$. We note that $m_{1,j}$ are uniquely identified, contrary to m_j. Motivated by the observation that $\hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),1,j}(x_j)$ approximates $m_{1,j}(x_j) - h \cdot \theta_{0j} g_j'(x_j)$ (Proposition 3), we estimate the scaled derivatives $m_{1,j}$ by

$$\hat{m}_{1,j}(x_j) := \hat{m}_{Y-\hat{\boldsymbol{\theta}}_0^\top \mathbf{g}(\mathbf{X}),1,j}(x_j) + h\,\hat{\theta}_{0j}\, g_j'(x_j), \quad 1 \le j \le d. \tag{3.3}$$

Put $\hat{\mathbf{m}}^{\mathrm{tp}} := (\hat{m}(\mathbf{x}), \hat{m}_{1,1}(x_1), \ldots, \hat{m}_{1,d}(x_d))^\top$ as our estimator of the tuple $\mathbf{m}^{\mathrm{tp}}$. Define $\mathbf{g}_j^{\mathrm{tp}} : [0,1] \to \mathbb{R}^{d+1}$ by

$$\mathbf{g}_j^{\mathrm{tp}}(x_j) := \big(g_j(x_j), 0, \ldots, 0, h\, g_j'(x_j), 0, \ldots, 0\big)^\top,$$

where $h\, g_j'(x_j)$ appears at the $(j+1)$th position. Then, by (3.2) and (3.3) it holds that

$$\hat{\mathbf{m}}^{\mathrm{tp}}(\mathbf{x}) = \hat{\mathbf{m}}^{\mathrm{tp}}_{Y-\hat{\boldsymbol{\theta}}_0^\top \mathbf{g}(\mathbf{X})}(\mathbf{x}) + \sum_{j=1}^{d} \hat{\theta}_{0j}\, \mathbf{g}_j^{\mathrm{tp}}(x_j). \tag{3.4}$$

It is worthwhile to note that $\hat{\mathbf{m}}_Y^{\mathrm{tp}} = (\hat{m}_Y, \hat{m}_{Y,1,1}, \ldots, \hat{m}_{Y,1,d})^\top$ is the estimator of $\mathbf{m}^{\mathrm{tp}}$ obtained by applying the local linear SBF operation directly to the dataset $\{(\mathbf{X}_i, Y_i) : 1 \le i \le n\}$, *without* the parametric help.

To state our first main theorem, we make the following additional assumptions on g_j, m_j and $Y - m(\mathbf{X})$.

(A4) For $\varepsilon := Y - m(\mathbf{X})$, it holds that $\mathrm{E}(|\varepsilon|^\alpha) < \infty$ for some $\alpha > 5/2$ and $\mathrm{E}(\varepsilon^2 | \mathbf{X} = \cdot)$ is bounded.

(A5) The components g_j and m_j for $1 \le j \le d$ are twice continuously differentiable on $[0,1]$.

Theorem 1. *Assume (A1)–(A5). If $\hat{\boldsymbol{\theta}}_0 \to \boldsymbol{\theta}_0$ in probability, then*

$$\hat{\mathbf{m}}^{\mathrm{tp}}(\mathbf{x}) - \mathbf{m}^{\mathrm{tp}}(\mathbf{x}) = \hat{\mathbf{m}}^{\mathrm{tp}}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X})}(\mathbf{x}) - \left(\mathbf{m}^{\mathrm{tp}}(\mathbf{x}) - \sum_{j=1}^{d} \theta_{0j} \mathbf{g}_j^{\mathrm{tp}}(x_j)\right) + o_p\big(n^{-2/5}\big)$$

uniformly for $\mathbf{x} \in [0,1]^d$.

Taking the first elements of the tuples from both sides of the equation in the above theorem, we get

$$\hat{m}(\mathbf{x}) - m(\mathbf{x}) = \hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X})}(\mathbf{x}) - \big(m(\mathbf{x}) - \boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{x})\big) + o_p\big(n^{-2/5}\big) \tag{3.5}$$

uniformly for $\mathbf{x} \in [0,1]^d$. We also have

$$\hat{m}_{1,j}(x_j) - m_{1,j}(x_j) = \hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),1,j}(x_j) - \big(m_{1,j}(x_j) - \theta_{0j} g_{1,j}(x_j)\big) + o_p\big(n^{-2/5}\big)$$

uniformly for $x_j \in [0,1]$. Below, we derive a stochastic expansion of the leading term in the expansion (3.5). Put

$$\kappa_h(u,v) := \frac{\mu_2(u,K) - h^{-1}(v-u)\mu_1(u,K)}{\mu_0(u,K)\mu_2(u,K) - \mu_1(u,K)^2} \cdot K_h(u,v),$$

where $\mu_j(u,K) = \int_0^1 ((v-u)/h)^j K_h(u,v)\,dv$. We note that $\kappa_h(x_j, X_{ij})$ approximate the kernel weights applied to the first entry of the 2-vector $\tilde{\mathbf{m}}_{W,j}$ defined at (2.2). We also note that the first entry of the 2-vector $\mathbf{N}(x_j)^{-1}\boldsymbol{\gamma}(x_j)$ equals $\mu_2(x_j,\kappa)$. Define

$$\tilde{m}^*_{W,j}(x_j) := n^{-1}\sum_{i=1}^n \kappa_h(x_j, X_{ij}) \cdot W_i.$$

Looking at (3.5) in the application of Proposition 2 to $W = Y - \boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X})$, and noting that

$$\varepsilon_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X})} = \big(Y - \boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X})\big) - \mathrm{E}\big(Y - \boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X})|\mathbf{X}\big) = \varepsilon,$$

we get the following corollary.

Corollary 2. *Assume the conditions of Theorem 1. Then,*

$$\hat{m}(\mathbf{x}) - m(\mathbf{x}) = \sum_{j=1}^d \tilde{m}^*_{\varepsilon,j}(x_j) + \frac{h^2}{2}\sum_{j=1}^d \mu_2(x_j,\kappa)\big(m_j''(x_j) - \theta_{0j}\, g_j''(x_j)\big) + o_p\big(n^{-2/5}\big)$$

uniformly for $\mathbf{x} \in [0,1]^d$.

According to our notation in Section 2, $\hat{m}_Y$ is the direct local linear SBF estimator of m without parametric help. From Proposition 2 we may get

$$\hat{m}_Y(\mathbf{x}) - m(\mathbf{x}) = \sum_{j=1}^d \tilde{m}^*_{\varepsilon,j}(x_j) + \frac{h^2}{2}\sum_{j=1}^d \mu_2(x_j,\kappa)\cdot m_j''(x_j) + o_p\big(n^{-2/5}\big) \tag{3.6}$$

uniformly for $\mathbf{x} \in [0,1]^d$. Thus, the first-order bias of our proposal $\hat{m}(\mathbf{x})$ differs from that of $\hat{m}_Y(\mathbf{x})$ by $h^2\sum_{j=1}^d \theta_{0j}\,\mu_2(x_j,\kappa)\,g_j''(x_j)/2$.

3.2. Estimation of individual components

Recall that the individual components m_j in the model (3.1) are identifiable up to a constant. Put

$$m_j^c(x_j) := m_j(x_j) - \int_0^1 m_j(u)p_j(u)\,du, \quad 1 \le j \le d.$$

Then, m_j^c are uniquely determined and satisfy $\mathrm{E}(m_j^c(X_j)) = 0$. We may rewrite the underlying additive model (3.1) as $m(\mathbf{x}) = \mathrm{E}(Y) + \sum_{j=1}^d m_j^c(x_j)$. Recall also that we estimate $m_{1,j} = h\,m_j'$ by $\hat{m}_{1,j}$ at (3.3). The estimators $\hat{m}_{1,j}$ are used in constructing our estimators of the centered m_j^c. Indeed, our proposal

for the estimators of the centered components m_j^c are given by

$$\hat{m}_j^c(x_j) := \hat{m}_j(x_j) - \int_0^1 \left(\hat{m}_j(u)\hat{p}_j(u) + \hat{m}_{1,j}(u)\hat{p}_{1,j}(u)\right) du, \quad 1 \le j \le d, \tag{3.7}$$

where $(\hat{m}_1, \ldots, \hat{m}_d)$ is *any* tuple of univariate functions such that $\hat{m}_1 + \cdots + \hat{m}_d = \hat{m}$. We find that estimating $m_{1,j}$ simply by $\hat{m}_{Y,1,j}$ and using them in (3.7) brings disharmony with the parametrically-helped estimation of the additive function m at (3.2), which would produce non-negligible extra bias. We note that $\hat{m}_j^c$ satisfy

$$\int_0^1 \left(\hat{m}_j^c(u)\hat{p}_j(u) + \hat{m}_{1,j}(u)\hat{p}_{1,j}(u)\right) du = 0, \quad 1 \le j \le d. \tag{3.8}$$

According to Proposition 1 in Section 2.2, $\hat{\mathbf{m}}^{\mathrm{tp}}$ is unique. By Lemma 3 in the Appendix C, $\hat{m}_j$ in their sum $\hat{m}$ are identified up to constant, so that $\hat{m}_j - \int_0^1 \hat{m}_j(u)\hat{p}_j(u)\,du$ are uniquely determined. Since $\hat{m}_{1,j}$ are unique and from the definition of $\hat{m}_j^c$ at (3.7), $\hat{m}_j^c$ are also uniquely determined. Below, we present relevant stochastic expansions of $\hat{m}_j^c$, as estimators of m_j^c.

The expansion in Corollary 2 does not give directly relevant expansions for the individual centered components $\hat{m}_j^c(x_j) - m_j^c(x_j)$. Basically, each centered component in the sum function $\hat{m}(\mathbf{x}) - m(\mathbf{x})$ may not equal, up to $o_p(n^{-2/5})$, to the corresponding centered component in $\hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X})}(\mathbf{x}) - (m(\mathbf{x}) - \boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{x}))$. We need to investigate how the constraints (3.8) affect the expansions of $\hat{m}_j^c(x_j) - m_j^c(x_j)$. The following theorem gives a new insight on this, which motivates a different design for $\boldsymbol{\theta}_0$ from the one for estimating the additive regression function m, see Section 4.2 below.

Theorem 2. *Assume the conditions of Theorem 1. Then, it holds that, for all* $1 \le j \le d$,

$$\begin{aligned}&\hat{m}_j^c(x_j) - m_j^c(x_j)\\ &\quad = \tilde{m}^*_{\varepsilon,j}(x_j) + \frac{h^2}{2}\left(\mu_2(x_j,\kappa)m_j''(x_j) - \theta_{0j}\left[\mu_2(x_j,\kappa)g_j''(x_j) - \mathrm{E}\left(\mu_2(X_j,\kappa)g_j''(X_j)\right)\right]\right) + o_p\left(n^{-2/5}\right)\end{aligned}$$

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

According to our notation in Section 2, $\hat{m}^c_{Y,j}$ is the estimator of m_j^c obtained by the direct local linear SBF without parametric help. From Proposition 3 we get

$$\hat{m}^c_{Y,j}(x_j) - m_j^c(x_j) = \tilde{m}^*_{\varepsilon,j}(x_j) + \frac{h^2}{2}\,\mu_2(x_j,\kappa)\cdot m_j''(x_j) + o_p\left(n^{-2/5}\right) \tag{3.9}$$

uniformly for $x_j \in [0,1]$. Thus, the first-order bias of our proposal $\hat{m}_j^c(x_j)$ differs from that of $\hat{m}^c_{Y,j}$ by $h^2\theta_{0j}\left[\mu_2(x_j,\kappa)g_j''(x_j) - \mathrm{E}\left(\mu_2(X_j,\kappa)g_j''(X_j)\right)\right]/2$. For the rate of convergence of $\hat{m}_j^c$ as an estimator of m_j^c, we observe that

$$\sup_{x_j\in[0,1]} |\tilde{m}^*_{\varepsilon,j}(x_j)| = O_p\left(n^{-1/2}h^{-1/2}\sqrt{\log n}\right) = O_p\left(n^{-2/5}\sqrt{\log n}\right).$$

The following corollary is an immediate consequence of Theorem 2.

Corollary 3. *Assume (A1)–(A5). Then, it holds that*

$$\max_{1\le j\le d}\sup_{x_j\in[0,1]}|\hat{m}_j^c(x_j)-m_j^c(x_j)| = O_p\big(n^{-2/5}\sqrt{\log n}\big),$$

4. Parametric help in action

In this section we find $\boldsymbol{\theta}_0$ that allows for maximal gain from the parametric help with $\mathbf{g}$ in the local linear SBF estimation of m at (3.1), and the one in the estimation of its centered components m_j^c. We also present ways of obtaining consistent estimators of them. It turns out that the 'best' $\boldsymbol{\theta}_0$ for estimating m is different from the one for estimating $(m_j^c : 1\le j\le d)$.

4.1. Bias reduction in estimating *m*

We first note that, in the expansion in Corollary 2,

$$\mu_2(x_j,\kappa) = \int_0^1\left(\frac{v-u}{h}\right)^2\kappa_h(u,v)\,dv = \frac{\mu_2(x_j,K)^2-\mu_1(x_j,K)\mu_3(x_j,K)}{\mu_0(x_j,K)\mu_2(x_j,K)-\mu_1(x_j,K)^2}.$$

For x_j in the interior region $[2h,1-2h]$, the (incomplete) moments $\mu_j(x_j,K)$ reduce to the complete moments of K, $\mu_j(K) := \int u^j K(u)\,du$. From the symmetry of the baseline kernel K, we have then $\mu_2(x_j,\kappa)=\mu_2(K)$ for all $x_j\in[2h,1-2h]$. Put $m_+''(\mathbf{x}) := m_1''(x_1)+\cdots+m_d''(x_d)$ and $\mathbf{g}''(\mathbf{x}) = (g_1''(x_1),\ldots,g_d''(x_d))^\top$. Then, from Corollary 2, it turns out that the expected value of the squared asymptotic bias of $\hat{m}$ equals

$$\frac{h^4}{4}\,\mu_2(K)^2\,\mathrm{E}\big(m_+''(\mathbf{X})-\boldsymbol{\theta}_0^\top\mathbf{g}''(\mathbf{X})\big)^2. \tag{4.1}$$

We consider

$$\begin{aligned}\boldsymbol{\theta}_0 \equiv (\theta_{01},\ldots,\theta_{0d})^\top &:= \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d}\mathrm{E}\big(m_+''(\mathbf{X})-\boldsymbol{\theta}^\top\mathbf{g}''(\mathbf{X})\big)^2\\ &= \big[\mathrm{E}\big(\mathbf{g}''(\mathbf{X})\mathbf{g}''(\mathbf{X})^\top\big)\big]^{-1}\cdot\mathrm{E}\big(\mathbf{g}''(\mathbf{X})\cdot m_+''(\mathbf{X})\big).\end{aligned} \tag{4.2}$$

For (4.1) and (4.2), we make the following additional assumption throughout this section.

(A6) $\max_{1\le j\le d}\mathrm{E}(g_j''(X_j)^2)<\infty$, and $\mathrm{E}(\mathbf{g}''(\mathbf{X})\mathbf{g}''(\mathbf{X})^\top)$ is invertible.

We note that $\boldsymbol{\theta}_0^\top\mathbf{g}$ is nothing else than the projection of m onto $\{\boldsymbol{\theta}^\top\mathbf{g}:\boldsymbol{\theta}\in\mathbb{R}^d\}$ in the space of additive functions $\mathscr{H}_{\mathrm{add}}$ equipped with the inner product $\langle f,\eta\rangle = \mathrm{E}(f_+''(\mathbf{X})\eta_+''(\mathbf{X}))$, where $f_+''(\mathbf{x}) = f_1''(x_1)+\cdots+f_d''(x_d)$ for $f\in\mathscr{H}_{\mathrm{add}}$. According to (3.6), the expected value of the squared asymptotic bias of the direct local linear estimator $\hat{m}_Y$ is given by $h^4\,\mu_2(K)^2\,\mathrm{E}(m_+''(\mathbf{X}))^2/4$. Comparing this and (4.1) with the choice $\boldsymbol{\theta}_0$ at (4.2), we get

$$\begin{aligned}0 &\le \mathrm{E}\big(m_+''(\mathbf{X})-\boldsymbol{\theta}_0^\top\mathbf{g}''(\mathbf{X})\big)^2\\ &= \mathrm{E}\big(m_+''(\mathbf{X})\big)^2-\big[\mathrm{E}\big(\mathbf{g}''(\mathbf{X})\cdot m_+''(\mathbf{X})\big)\big]^\top\big[\mathrm{E}\big(\mathbf{g}''(\mathbf{X})\mathbf{g}''(\mathbf{X})^\top\big)\big]^{-1}\big[\mathrm{E}\big(\mathbf{g}''(\mathbf{X})\cdot m_+''(\mathbf{X})\big)\big]\\ &\le \mathrm{E}\big(m_+''(\mathbf{X})\big)^2\end{aligned}$$

with the equality holding only if $\mathrm{E}(\mathbf{g}''(\mathbf{X}) \cdot m_+''(\mathbf{X})) = \mathbf{0}$. This means that, as long as g_j are chosen so that $\mathrm{E}(\mathbf{g}''(\mathbf{X}) \cdot m_+''(\mathbf{X})) \neq \mathbf{0}$ and we use a consistent estimator $\hat{\boldsymbol{\theta}}_0$ in the construction of $\hat{m}$ at (3.2), $\hat{m}$ always improves the first-order bias of the direct local linear SBF $\hat{m}_Y$.

Now, we find a consistent estimator of $\boldsymbol{\theta}_0$ defined at (4.2). Put $Y^{\boldsymbol{\theta}} := Y - \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X})$ and $Y_i^{\boldsymbol{\theta}} := Y_i - \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X}_i)$, $1 \le i \le n$. We estimate $\boldsymbol{\theta}_0$ by

$$\hat{\boldsymbol{\theta}}_0 := \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} n^{-1} \sum_{i=1}^n \big(Y_i^{\boldsymbol{\theta}} - \hat{m}_{Y^{\boldsymbol{\theta}}}(\mathbf{X}_i)\big)^2. \tag{4.3}$$

For a generic random variable W, recall $\varepsilon_W = W - m_W(\mathbf{X})$. We let $\hat{\varepsilon}_{W,i} := W_i - \hat{m}_W(\mathbf{X}_i)$. Put

$$\hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i} := (\hat{\varepsilon}_{g_1(X_1),i}, \ldots, \hat{\varepsilon}_{g_d(X_d),i})^\top = \mathbf{g}(\mathbf{X}_i) - \hat{\mathbf{m}}_{\mathbf{g}(\mathbf{X})}(\mathbf{X}_i),$$

where $\hat{\mathbf{m}}_{\mathbf{g}(\mathbf{X})}(\mathbf{x}) := (\hat{m}_{g_1(X_1)}(\mathbf{x}), \ldots, \hat{m}_{g_d(X_d)}(\mathbf{x}))^\top$. Then, using the linearity of the SBF operation asserted by Proposition 1 in Section 2, we find that $\hat{\boldsymbol{\theta}}_0$ at (4.3) can be given explicitly. Indeed, it can be defined alternatively as

$$\hat{\boldsymbol{\theta}}_0 := \left(n^{-1} \sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i} \cdot \hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i}^\top \right)^{-1} \cdot n^{-1} \sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i} \cdot \hat{\varepsilon}_{Y,i}, \tag{4.4}$$

whenever $n^{-1} \sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i} \cdot \hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i}^\top$ is invertible.

The definition of the estimator $\hat{\boldsymbol{\theta}}_0$ at (4.3) is motivated by observing

$$n^{-1} \sum_{i=1}^n \big(Y_i^{\boldsymbol{\theta}} - \hat{m}_{Y^{\boldsymbol{\theta}}}(\mathbf{X}_i)\big)^2 \simeq c_K^2 \cdot \mathrm{E}\big[\big(m_+''(\mathbf{X}) - \boldsymbol{\theta}^\top \mathbf{g}''(\mathbf{X})\big)^2 \big] + (\text{irrelevant term}) \tag{4.5}$$

for some constant c_K depending solely on the baseline kernel K. To see the approximation at (4.5), we note that $Y_i^{\boldsymbol{\theta}} = m(\mathbf{X}_i) - \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X}_i) + \varepsilon_i$ so that

$$Y_i^{\boldsymbol{\theta}} - \hat{m}_{Y^{\boldsymbol{\theta}}}(\mathbf{X}_i) = \big(m(\mathbf{X}_i) - \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X}_i) - \hat{m}_{m(\mathbf{X}) - \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X})}(\mathbf{X}_i)\big) + \big(\varepsilon_i - \hat{m}_\varepsilon(\mathbf{X}_i)\big),$$

where we have used the linearity of $\hat{m}_W$ in W asserted in Proposition 1. The approximation at (4.5) then follows from the application of (B.2) in the Appendix to $\eta(\mathbf{X}) = m(\mathbf{X}) - \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X})$. Indeed, the left hand side of (4.5) is well approximated by

$$\begin{aligned} &n^{-1} \sum_{i=1}^n \big(m(\mathbf{X}_i) - \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X}_i) - \hat{m}_{m(\mathbf{X}) - \boldsymbol{\theta}^\top \mathbf{g}(\mathbf{X})}(\mathbf{X}_i)\big)^2 + n^{-1} \sum_{i=1}^n \big(\varepsilon_i - \hat{m}_\varepsilon(\mathbf{X}_i)\big)^2 \\ &\quad \simeq c_K^2 \cdot \mathrm{E}\big[\big(m_+''(\mathbf{X}) - \boldsymbol{\theta}^\top \mathbf{g}''(\mathbf{X})\big)^2 \big] + n^{-1} \sum_{i=1}^n \big(\varepsilon_i - \hat{m}_\varepsilon(\mathbf{X}_i)\big)^2. \end{aligned}$$

The first term on the right hand side of the approximation at (4.5) is what $\boldsymbol{\theta}_0$ minimizes, see (4.2), and the second term does not involve $\boldsymbol{\theta}$.

Theorem 3. *Assume (A1)–(A6). Let $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_0$ are defined as at (4.2) and (4.4), respectively. Then, $\hat{\boldsymbol{\theta}}_0 \to \boldsymbol{\theta}_0$ in probability.*

In the next theorem we present the asymptotic distribution of $\hat{m}(\mathbf{x})$ defined at (3.2) with $\hat{\boldsymbol{\theta}}_0$ given at (4.4). It is an immediate consequence of Corollary 2 and Theorem 3. Recall that $\mu_j(K) = \int u^j K(u)\,du$ and $\mu_j(K^2) = \int u^j K(u)^2\,du$ for $j \geq 0$. Define

$$\sigma_j^2(x_j) := \frac{1}{\tau} \cdot \mu_0\big(K^2\big) \cdot \frac{\mathrm{E}(\varepsilon^2 | X_j = x_j)}{p_j(x_j)},$$

where $\tau > 0$ is a constant such that $n^{1/5}h \to \tau$.

Theorem 4. *Assume (A1)–(A6). Assume also that* $\mathrm{E}(|\varepsilon|^\alpha | X_j = \cdot)$ *for all* $1 \leq j \leq d$ *are bounded on* $[0,1]$ *for some* $\alpha > 5/2$, *and* $\mathrm{E}(\varepsilon^2 | X_j = \cdot)$ *for* $1 \leq j \leq d$ *are continuous on* $[0,1]$. *Then, for each* $\mathbf{x} \in (0,1)^d$, *the distribution of* $n^{2/5}(\hat{m}(\mathbf{x}) - m(\mathbf{x}))$ *converges to the normal distribution with mean* $\tau^2 \mu_2(K)(m_+''(\mathbf{x}) - \boldsymbol{\theta}_0^\top \mathbf{g}''(\mathbf{x}))/2$ *and variance* $\sum_{j=1}^d \sigma_j^2(x_j)$, *where* $\boldsymbol{\theta}_0$ *is defined at (4.2).*

4.2. Bias reduction in estimating m_j^c

Since $\mu_2(x_j, \kappa) = \mu_2(K)$ for all $x_j \in [2h, 1-2h]$, the asymptotic bias of $\hat{m}_j^c(x_j)$ equals

$$h^2 \mu_2(K)\big[m_j''(x_j) - \theta_{0j}\big(g_j''(x_j) - \mathrm{E}\, g_j''(X_j)\big)\big]/2$$

for all $x_j \in [2h, 1-2h]$. There is an extra term $\mathrm{E}\, g_j''(X_j)$ in the expansion of the jth centered individual component estimator, which is absent in the summands in the expansion of $\hat{m}(\mathbf{x}) - m(\mathbf{x})$ given in Corollary 2. The expected value of the squared asymptotic bias of $\hat{m}_j^c$ equals $h^4\,\mu_2(K)^2\,\mathrm{E}(m_j''(X_j) - \theta_{0j}[g_j''(X_j) - \mathrm{E}\, g_j''(X_j)])^2/4$. For the local linear estimator $\hat{m}_{Y,j}^c$ without parametric help, the expected value turns out to be $h^4\,\mu_2(K)^2\,\mathrm{E}(m_j''(X_j))^2/4$. We note that the inequality $\mathrm{E}(m_+''(\mathbf{X}) - \boldsymbol{\theta}_0^\top \mathbf{g}''(\mathbf{X}))^2 \leq \mathrm{E}(m_+''(\mathbf{X}))^2$ for $\boldsymbol{\theta}_0$ at (4.2), which we observed in the estimation of m, implies *neither* $\mathrm{E}(m_j''(X_j) - \theta_{0j} g_j''(X_j))^2 \leq \mathrm{E}(m_j''(X_j))^2$ *nor*

$$\mathrm{E}\big(m_j''(X_j) - \theta_{0j}\big[g_j''(X_j) - \mathrm{E}\, g_j''(X_j)\big]\big)^2 \leq \mathrm{E}\big(m_j''(X_j)\big)^2$$

for any $1 \leq j \leq d$.

To take full benefit of parametric help for estimating individual components, we consider $\boldsymbol{\theta}_0$ that is *different from* the one defined at (4.2). With a slight abuse of notation, we continue to denote it by $\boldsymbol{\theta}_0$. For individual component estimation, we choose $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0d})^\top$ with each element θ_{0j} being defined by

$$\theta_{0j} := \operatorname*{argmin}_{\theta_j \in \mathbb{R}} \mathrm{E}\big(m_j''(X_j) - \theta_j\big[g_j''(X_j) - \mathrm{E}\, g_j''(X_j)\big]\big)^2 = \frac{\mathrm{Cov}(m_j''(X_j), g_j''(X_j))}{\mathrm{Var}(g_j''(X_j))} \tag{4.6}$$

For this definition, we assume

(A6′) $0 < \min_{1 \leq j \leq d} \mathrm{Var}(g_j''(X_j)^2) \leq \max_{1 \leq j \leq d} \mathrm{Var}(g_j''(X_j)^2) < \infty$.

A projection interpretation can be given to $\theta_{0j}[g_j'' - \mathrm{E}\, g_j''(X_j)]$ similarly as for $\boldsymbol{\theta}_0^\top \mathbf{g}$ in Section 4.1. It is clear that

$$\begin{aligned} \mathrm{E}\big(m_j''(X_j) - \theta_{0j}\big[g_j''(X_j) - \mathrm{E}\, g_j''(X_j)\big]\big)^2 &= \mathrm{E}\big(m_j''(X_j)\big)^2 - \frac{\mathrm{Cov}(m_j''(X_j), g_j''(X_j))^2}{\mathrm{Var}(g_j''(X_j))} \\ &\le \mathrm{E}\big(m_j''(X_j)\big)^2 \end{aligned} \tag{4.7}$$

with the equality holding if and only if $\mathrm{Cov}(m_j''(X_j), g_j''(X_j)) = 0$.

To achieve the benefit of parametric help in the individual component estimation we have discussed above, we need to find consistent estimators of θ_{0j}. Put

$$\omega_{j,h}(u,v) := \frac{\hat{\mu}_{j,2}(u,K) - h^{-1}(v-u)\hat{\mu}_{j,1}(u,K)}{\hat{\mu}_{j,0}(u,K)\hat{\mu}_{j,2}(u,K) - \hat{\mu}_{j,1}(u,K)^2} \cdot K_h(u,v),$$

where $\hat{\mu}_{j,l}(u,K) := n^{-1}\sum_{i=1}^n h^{-l}(X_{ij}-u)^l K_h(u,X_{ij})$. Let $\tilde{m}_{W,j}$ be the *marginal* local linear estimators with W_i as the response and X_{ij} as the covariate values. It is the first entry of $\tilde{\mathbf{m}}_{W,j}$ defined at (2.2). Then, $\tilde{m}_{W,j}(x_j) = n^{-1}\sum_{i=1}^n \omega_{j,h}(x_j, X_{ij})\cdot W_i$. Let $\hat{\delta}_i(j,W) := W_i - \tilde{m}_{W,j}(X_{ij})$ and

$$\hat{\delta}_i^{\mathrm{dev}}(j,W) := \hat{\delta}_i(j,W) - n^{-1}\sum_{i=1}^n \hat{\delta}_i(j,W).$$

Recall that $\hat{m}_{Y,j}^c$ are the centered versions of components $\hat{m}_{Y,j}$ that comprise $\hat{m}_Y = \hat{m}_{Y,1} + \cdots + \hat{m}_{Y,d}$, the latter being the first entry of the $(d+1)$-tuple $\hat{\mathbf{m}}_Y^{\mathrm{tp}}$ given as the solution of the local linear SBF equation (2.3) with $W = Y$. We propose

$$\hat{\theta}_{0j} := \left(n^{-1}\sum_{i=1}^n \big[\hat{\delta}_i^{\mathrm{dev}}\big(j, g_j(X_j)\big)\big]^2\right)^{-1} \cdot n^{-1}\sum_{i=1}^n \hat{\delta}_i^{\mathrm{dev}}\big(j, g_j(X_j)\big)\cdot \hat{\delta}_i\big(j, \hat{m}_{Y,j}^c(X_j)\big). \tag{4.8}$$

The definition (4.8) is motivated from the observation that, for a smooth function η, $\hat{\delta}_i(j,\eta(X_j))$ approximates well $-h^2\mu_2(X_{ij},\kappa)\cdot\eta''(X_{ij})/2$. If we apply the approximation to $\eta = g_j$ and to $\eta = m_j^c$, then we see that

$$\hat{\theta}_{0j}^* := \left(n^{-1}\sum_{i=1}^n \big[\hat{\delta}_i^{\mathrm{dev}}\big(j, g_j(X_j)\big)\big]^2\right)^{-1} \cdot n^{-1}\sum_{i=1}^n \hat{\delta}_i^{\mathrm{dev}}\big(j, g_j(X_j)\big)\cdot \hat{\delta}_i\big(j, m_j^c(X_j)\big)$$

approximates well θ_{0j}. Our proposal $\hat{\theta}_{0j}$ simply replaces $\hat{\delta}_i(j, m_j^c(X_j))$ in $\hat{\theta}_{0j}^*$ by $\hat{\delta}_i(j, \hat{m}_{Y,j}^c(X_j))$.

Theorem 5. *Assume (A1)–(A5) and (A6′). Let θ_{0j} and $\hat{\theta}_{0j}$ are defined as at (4.6) and (4.8), respectively. Then, $\hat{\theta}_{0j} \to \theta_{0j}$ in probability for all $1 \le j \le d$.*

Remark 1. One may be tempted to replace $\hat{\delta}_i^{\mathrm{dev}}(j, g_j(X_j))$ by $g_j''(X_{ij}) - \overline{g_j''(X_j)}$ in the definition of $\hat{\theta}_{0j}$, where $\overline{g_j''(X_j)} := n^{-1}\sum_{i=1}^n g_j''(X_{ij})$. But, this does not work since then the numerator of the resulting quantity would be of magnitude $O_p(h^2)$ while the denominator converges to $\mathrm{Var}(g_j''(X_j))$.

Remark 2. One may estimate θ_{0j} in a direct way if one is given a uniformly consistent estimator of m_j''. If this is the case and if we denote it by $\check{m}_j''$, then, the estimator defined by

$$\check{\theta}_{0j} := \frac{\sum_{i=1}^n \check{m}_j''(X_{ij})(g_j''(X_{ij}) - \overline{g_j''(X_j)})}{\sum_{i=1}^n (g_j''(X_{ij}) - \overline{g_j''(X_j)})^2}$$

is also consistent. For consistent estimation of m_j'' one needs to employ a higher-order local polynomial SBF and assume a higher-order smoothness of the component functions m_j, however.

Let $\hat{\boldsymbol{\theta}}_0 := (\hat{\theta}_{01}, \ldots, \hat{\theta}_{0d})^\top$. Since $\hat{\boldsymbol{\theta}}_0 \to \boldsymbol{\theta}_0$ in probability by Theorem 5, the conclusion of Theorem 2 with θ_{0j} now defined at (4.6) still applies to $\hat{m}_j^c$ defined at (3.7) with

$$\hat{m}_{1,j}(x_j) := \hat{m}_{Y - \hat{\boldsymbol{\theta}}_0^\top \mathbf{g}(\mathbf{X}),1,j}(x_j) + h\,\hat{\theta}_{0j}\,g_j'(x_j),$$

where $\hat{\boldsymbol{\theta}}_0$ and its components $\hat{\theta}_{0j}$ are defined through (4.8). Thus, the bias improvement in comparison with the direct local linear estimator $\hat{m}_{Y,j}^c$ is evidenced by (4.7). Furthermore, our proposed estimators $\hat{m}_j^c$ are jointly asymptotically normal, as demonstrated by the following theorem. To state the theorem, define

$$\beta_j(x_j) := \frac{\tau^2}{2} \cdot \mu_2(K) \cdot \left(m_j''(x_j) - \theta_{0j}\left(g_j''(x_j) - \mathrm{E}g_j''(X_j)\right)\right).$$

Recall $\tau = \lim_{n\to\infty} n^{1/5} h > 0$.

Theorem 6. *Assume the conditions of Theorem 4 with (A6) replaced by (A6′). Then, for each* $\mathbf{x} \in (0,1)^d$*, the joint distribution of* $n^{2/5}(\hat{m}_1^c(x_1) - m_1^c(x_1), \ldots, \hat{m}_d^c(x_d) - m_d^c(x_d))^\top$ *converges to the multivariate normal distribution with mean vector* $(\beta_1(x_1), \ldots, \beta_d(x_d))^\top$ *and variance matrix* $\mathrm{diag}(\sigma_j^2(x_j))$.

5. Monte Carlo experiments

We generated the response variable Y from the following additive model:

$$Y = \sin(2\pi X_1) + \rho_1\, X_1(X_1 - 0.5)^2(X_1 - 1) + \cos(2\pi X_2) + \rho_2\, X_2(X_2 - 0.5)(X_2 - 1) + \varepsilon, \tag{5.1}$$

where $\varepsilon \sim N(0, \sigma^2)$ with $\sigma = 0.5$, independently of X_j. We took $X_j = \Phi(Z_j)$ for $j = 1, 2$, where Φ is the distribution function of the standard normal distribution and

$$(Z_1, Z_2)^\top \sim N_2\left((0,0)^\top, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right).$$

Note that the generated X_j are marginally uniformly distributed on $[0,1]$ but they are dependent since $\mathrm{Corr}(X_1, X_2) \neq 0$. We made two choices for ρ_1 and ρ_2: $\rho_1 = 0{,}64$; $\rho_2 = 0{,}12\sqrt{3}$. We chose $\rho_1 = 64$ since $\max_{x_1 \in [0,1]} |x_1(x_1 - 0.5)^2(x_1 - 1)| = 1/64$, so that the two components $\sin(2\pi x_1)$ and $64\, x_1(x_1 - 0.5)^2(x_1 - 1)$ have the same amplitude from zero. The choice $\rho_2 = 12\sqrt{3}$ was made for the same reason. We investigated the following two scenarios for the working parametric family.

Case 1: $g_1(x_1) = \sin(2\pi x_1)$, $\quad g_2(x_2) = \cos(2\pi x_2)$,

Case 2: $g_1(x_1) = x_1(x_1 - 0.5)^2(x_1 - 1)$, $\quad g_2(x_2) = x_2(x_2 - 0.5)(x_2 - 1)$.

For estimating m, the formula at (4.2) gives $\theta_{01} = 1 - 3\rho_2 A_1/\pi^2$ and $\theta_{02} = 1 - 6\rho_1 A_2/\pi^2$ in Case 1, and $\theta_{01} = \rho_1 - 320\pi^2 A_2/7$ and $\theta_{02} = \rho_2 - 8\pi^2 A_1$ in Case 2, where $A_1 = \mathrm{E}(X_2 - 0.5)\sin(2\pi X_1)$ and $A_2 = \mathrm{E}(X_1 - 0.5)^2\cos(2\pi X_2)$. For estimating the individual components m_j^c, the formula for θ_{0j} at (4.6) gives $\theta_{0j} \equiv 1$ in Case 1, and $\theta_{0j} = \rho_j$, $j = 1,2$ in Case 2.

We discuss briefly the simulation model (5.1) in conjunction with the two choices of (g_1, g_2). We first consider the case where $\rho_1 = \rho_2 = 0$, i.e., $m(\mathbf{x}) = \sin(2\pi x_1) + \cos(2\pi x_2)$. In this setting, Case 1 corresponds to the case where one picks by chance a working parametric family that contains the true additive regression function m. In this case, the leading bias of $\hat{m}$ vanishes since $m_+'' - \boldsymbol{\theta}_0^\top \mathbf{g}'' \equiv 0$. From this experiment one may be able to see how the theoretical benefit of the perfect parametric help takes into effect in practice for estimating m. On the other hand, in Case 2, the leading bias of $\hat{m}$ is not zero and our theory tells that there is some bias reduction in comparison with the direct local liner SBF estimator $\hat{m}_Y$ since $\mathrm{E}(\mathbf{g}''(\mathbf{X}) \cdot m_+''(\mathbf{X})) \neq \mathbf{0}$. As for the estimation of the individual components, which are $m_1(x_1) = \sin(2\pi x_1)$ and $m_2(x_2) = \cos(2\pi x_2)$, Case 1 and Case 2 are truly opposite. From (4.7) and the correlation inequality, it follows that

$$\mathrm{E}\big(m_j''(X_j) - \theta_{0j}\big[g_j''(X_j) - \mathrm{E}\, g_j''(X_j)\big]\big)^2 \geq \mathrm{E}\big(m_j''(X_j)\big)^2 - \mathrm{Var}\big(m_j''(X_j)\big)$$

for all g_j. The lower bound is achieved by the (g_1, g_2) in Case 1 since then $\mathrm{Cov}(m_j''(X_j), g_j''(X_j)) = \mathrm{Var}(m_j''(X_j))$. On the contrary, with g_j in Case 2, g_j'' are perpendicular to the respective m_j'', i.e., $\mathrm{Cov}(m_j''(X_j), g_j''(X_j)) = \mathrm{E}(m_j''(X_j) g_j''(X_j)) = 0$, in which case, according to our theory, there is no improvement in the leading biases of the estimators of the individual components. From the experiments with $\rho_1 \neq 0$ or $\rho_2 \neq 0$ in (5.1), we may be able to assess the performance of the parametric help in the intermediate cases between the best and worst scenarios.

We first compared the performance of our proposal $\hat{m}$ with the direct local linear SBF estimator $\hat{m}_Y$. We used the Epanechnikov kernel, $K(u) = (3/4)(1 - u^2)I(|u| \leq 1)$, and the theoretical bandwidths h that minimize the respective asymptotic density-weighted mean integrated squared errors

$$\frac{2}{nh} \cdot \int K(u)^2\, du \cdot \mathrm{E}(\varepsilon^2) + \frac{h^4}{4}\left(\int u^2 K(u)\, du\right)^2 \cdot A, \tag{5.2}$$

where $A = \mathrm{E}m_+''(\mathbf{X})^2$ for $\hat{m}_Y$ and $A = \mathrm{E}(m_+''(\mathbf{X}) - \theta_{01} g_1''(X_1) - \theta_{02} g_2''(X_2))^2$ for $\hat{m}$. Note that the A for $\hat{m}$ vanishes in Case 1 with $\rho_1 = \rho_2 = 0$, in which case we used the theoretical bandwidth for $\hat{m}_Y$ in computing $\hat{m}$. As a measure of performance, we computed the Monte Carlo approximations of the integrated squared bias (ISB), the integrated variance (IV) and the mean integrated squared error (MISE) of $\bar{m} = \hat{m}$ and $\bar{m} = \hat{m}_Y$: $\mathrm{MISE}(\bar{m}) := \mathrm{ISB}(\bar{m}) + \mathrm{IV}(\bar{m})$,

$$\begin{aligned} \mathrm{ISB}(\bar{m}) &:= \int_{[0,1]^2} \left(M^{-1} \sum_{i=1}^{M} \bar{m}^{(i)}(\mathbf{x}) - m(\mathbf{x})\right)^2 d\mathbf{x}, \\ \mathrm{IV}(\bar{m}) &:= \int_{[0,1]^2} M^{-1} \sum_{i=1}^{M} \left(\bar{m}^{(i)}(\mathbf{x}) - M^{-1} \sum_{i=1}^{M} \bar{m}^{(i)}(\mathbf{x})\right)^2 d\mathbf{x}, \end{aligned} \tag{5.3}$$

where $\bar{m}^{(i)}$ is the estimate of m computed from the ith dataset and $M = 100$.

Table 1 reports the values of these measures. The results confirm our theoretical finding that $\hat{m}$ has smaller MISE than $\hat{m}_Y$, except in the worst scenario (Case 2 with $\rho_1 = \rho_2 = 0$) where m_j'' and g_j'' are perpendicular to each other for $j = 1,2$. We note that, in Case 2, the values of the MISE as well as those of the ISB and the IV for our proposal $\hat{m}$ are nearly the same for all combinations $(\rho_1, \rho_2) =$

Table 1. The values of ISB, IV and MISE, multiplied by 10^3, of the direct local linear SBF estimator ($\hat{m}_Y$) and of our proposal ($\hat{m}$).

			Case 1				Case 2			
			$\rho_1 = 0$		$\rho_1 = 64$		$\rho_1 = 0$		$\rho_1 = 64$	
	n		$\hat{m}_Y$	$\hat{m}$	$\hat{m}_Y$	$\hat{m}$	$\hat{m}_Y$	$\hat{m}$	$\hat{m}_Y$	$\hat{m}$
$\rho_2 = 0$	100	ISB	4.27	0.23	7.73	3.96	4.27	6.56	7.73	6.56
		IV	35.5	36.1	40.6	44.1	35.5	34.0	40.6	34.0
		MISE	39.8	36.3	48.3	48.1	39.8	40.6	48.3	40.6
	400	ISB	1.42	0.09	2.79	1.49	1.42	2.24	2.79	2.24
		IV	9.02	9.09	10.0	10.8	9.02	8.54	10.0	8.54
		MISE	10.4	9.18	12.8	12.3	10.4	10.8	12.8	10.8
	1,000	ISB	0.72	0.04	1.46	0.77	0.72	1.12	1.46	1.12
		IV	4.34	4.38	4.81	5.20	4.34	4.04	4.81	4.04
		MISE	5.06	4.42	6.27	5.97	5.06	5.16	6.27	5.16
$\rho_2 = 12\sqrt{3}$	100	ISB	4.41	4.40	7.18	5.26	4.41	6.72	7.18	6.72
		IV	41.5	35.6	44.7	46.5	41.5	34.2	44.7	34.2
		MISE	45.9	40.0	51.9	51.8	45.9	40.9	51.9	40.9
	400	ISB	1.23	1.48	2.29	1.77	1.23	2.19	2.29	2.19
		IV	10.4	8.58	11.0	11.0	10.4	8.58	11.0	8.58
		MISE	11.6	10.1	13.3	13.8	11.6	10.7	13.3	10.7
	1,000	ISB	0.73	0.84	1.27	0.95	0.73	1.19	1.27	1.19
		IV	5.00	4.26	5.27	5.46	5.00	4.07	5.27	4.07
		MISE	5.73	5.10	6.54	6.41	5.73	5.26	6.54	5.26

$(0,0),(0,12\sqrt{3}),(64,0)$ and $(64,12\sqrt{3})$. This is expected. Indeed, the theoretical value of $A = \mathrm{E}(m''_+(\mathbf{X}) - \theta_{01}g''_1(X_1) - \theta_{02}g''_2(X_2))^2$ in (5.2) for $\hat{m}$ in Case 2 does not depend on ρ_j for $j = 1,2$. Another thing to be commented is that, in Cases 1 and 2 with $(\rho_1,\rho_2) = (0,12\sqrt{3})$, the values of the ISB for $\hat{m}$ are actually larger than those for $\hat{m}_Y$. This can be explained as follows. For $\hat{m}$, the reduced bias $\mathrm{E}(m''_+(\mathbf{X}) - \theta_{01}g''_1(X_1) - \theta_{02}g''_2(X_2))^2 < \mathrm{E}m''_+(\mathbf{X})^2$ led to a larger bandwidth than for the direct local linear SBF $\hat{m}_Y$, as a consequence of taking more care of the variance part. The larger bandwidth then decreased the variance significantly while increasing the bias slightly, and thus gave the smaller values of the MISE for $\hat{m}$ than for $\hat{m}_Y$. To assess the performance of the individual component estimators $\hat{m}_j^c$ and $\hat{m}_{Y,j}^c$, we also computed the ISB, IV and MISE for each component, replacing $\bar{m}$ in (5.3) by $\hat{m}_j^c$ and $\hat{m}_{Y,j}^c$, and m by m_j^c, for $j = 1,2$. We call them for the jth component, ISB_j, IV_j and MISE_j. Table 2 reports the values of $\mathrm{MISE}_1 + \mathrm{MISE}_2$ together with $\mathrm{ISB}_1 + \mathrm{ISB}_2$ and $\mathrm{IV}_1 + \mathrm{IV}_2$. Basically, the lessons are the same as those from Table 1. One thing to note is that the asymptotic mean integrated squared errors

$$\frac{1}{nh}\cdot\int K(u)^2\,du\cdot\mathrm{E}\big(\varepsilon^2\big) + \frac{h^4}{4}\bigg(\int u^2K(u)\,du\bigg)^2\cdot\mathrm{E}\big(m''_j(X_j) - \theta_{0j}\big(g''_j(X_j) - \mathrm{E}g''_j(X_j)\big)\big)^2$$

for $\hat{m}_j^c$ in Case 2 now depend on ρ_j, contrary to the case of $\hat{m}$. Thus, the theoretical values corresponding to those in Table 2 for PH_{SBF} in Case 2 are all different for different combinations $(\rho_1,\rho_2) = (0,0),(0,12\sqrt{3}),(64,0)$ and $(64,12\sqrt{3})$, contrary to those corresponding to Table 1.

Table 2. The values of $\mathrm{ISB}_+ = \mathrm{ISB}_1 + \mathrm{ISB}_2$, $\mathrm{IV}_+ = \mathrm{IV}_1 + \mathrm{IV}_2$ and $\mathrm{MISE}_+ = \mathrm{MISE}_1 + \mathrm{MISE}_2$, multiplied by 10^3, for the set of the direct local linear SBF component estimators 'SBF' $(\hat{m}^c_{Y,1}, \hat{m}^c_{Y,2})$ and for our proposal 'PH_{SBF}' $(\hat{m}^c_1, \hat{m}^c_2)$.

			Case 1				Case 2			
			$\rho_1 = 0$		$\rho_1 = 64$		$\rho_1 = 0$		$\rho_1 = 64$	
	n		SBF	PH_{SBF}	SBF	PH_{SBF}	SBF	PH_{SBF}	SBF	PH_{SBF}
$\rho_2 = 0$	100	ISB_+	4.16	0.28	7.44	3.73	4.16	6.38	7.44	7.59
		IV_+	43.5	33.9	49.8	43.7	43.5	41.7	49.8	42.4
		MISE_+	47.6	24.2	57.2	47.4	47.6	48.2	57.2	50.0
	400	ISB_+	1.42	0.09	2.92	1.49	1.42	2.19	2.92	2.46
		IV_+	10.9	8.59	12.3	10.7	10.9	10.3	12.3	10.4
		MISE_+	12.3	8.68	15.2	12.2	12.3	12.5	15.2	12.8
	1,000	ISB_+	0.72	0.04	1.50	0.80	0.72	1.02	1.50	1.22
		IV_+	5.16	4.14	5.71	5.07	5.16	4.85	5.71	4.85
		MISE_+	5.88	4.18	7.21	5.87	5.88	5.97	7.21	6.07
$\rho_2 = 12\sqrt{3}$	100	ISB_+	4.38	4.17	6.99	4.82	4.38	6.71	6.99	6.94
		IV_+	53.6	38.4	58.2	50.9	53.6	41.9	58.2	42.4
		MISE_+	58.0	42.6	65.2	55.9	58.0	48.6	65.2	50.3
	400	ISB_+	1.24	1.16	2.44	1.59	1.24	2.18	2.44	2.45
		IV_+	13.6	9.57	14.5	12.4	13.6	10.4	14.5	10.4
		MISE_+	14.8	10.7	17.0	14.0	14.8	12.6	17.0	12.9
	1,000	ISB_+	0.73	0.74	1.32	0.93	0.73	1.13	1.32	1.24
		IV_+	6.40	2.61	6.74	5.92	6.40	4.86	6.74	4.86
		MISE_+	7.13	5.35	8.06	6.85	7.13	5.99	8.06	6.10

6. Concluding remarks

We did not study the idea of the parametric help applied to the local constant SBF. One can imagine that it is more complicated than the application to the local linear SBF since the local constant SBF has boundary effects and its bias involves the joint density p as well, in an implicit way, see Mammen, Linton and Nielsen (1999). Above all, it is inferior to the local linear option, practically as well as theoretically.

The methodology and theory developed in this paper for real-valued responses Y may be extended to responses taking values in separable Hilbert spaces $\mathbb{H}$. In this general setting, the regression function $m : [0,1]^d \to \mathbb{H}$ is structured as $m = m_1 \oplus \cdots \oplus m_d$ with $m_j : [0,1] \to \mathbb{H}$, where $\oplus$ is the operation of addition on $\mathbb{H}$. Also, the parametric family that helps additive regression is given by $\{(\theta_1 \odot g_1) \oplus \cdots \oplus (\theta_d \odot g_d) : \theta_j \in \mathbb{R}, 1 \le j \le d\}$ with $g_j : [0,1] \to \mathbb{H}$, where $\odot$ is the operation of scalar multiplication. Let $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and $\| \cdot \|_{\mathbb{H}}$ denote the inner product and the associated norm of $\mathbb{H}$, respectively. Then, the parametric help with $\mathbf{g} = (g_1, \ldots, g_d)^\top$ is maximized by using a consistent estimator of

$$\boldsymbol{\theta}_0 := \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathrm{E} \left\| m_1''(X_1) \oplus \cdots \oplus m_d''(X_d) \ominus \left(\left(\theta_1 \odot g_1''(X_1)\right) \oplus \cdots \oplus \left(\theta_d \odot g_d''(X_d)\right) \right) \right\|_{\mathbb{H}}^2.$$

Here, $a \ominus b = a \oplus (-1) \odot b$ for $a, b \in \mathbb{H}$, and m_j'' and g_j'' are the second-order Fréchet derivatives of m_j and g_j, respectively. For the estimation of the individual components, the targets in the corresponding

parametric help are

$$\theta_{0j} := \frac{\mathrm{E}\left(\langle m_j''(X_j), g_j''(X_j) \ominus \mathrm{E} g_j''(X_j)\rangle_{\mathbb{H}}\right)}{\mathrm{E}\left(\|g_j''(X_j) \ominus \mathrm{E} g_j''(X_j)\|_{\mathbb{H}}^2\right)}, \quad 1 \le j \le d.$$

It is straightforward to think of versions of $\hat{\boldsymbol{\theta}}$ at (4.4) and $\hat{\theta}_{0j}$ at (4.8) in this Hilbertian setting. One may also get the corresponding analogues of the theorems and corollaries in Sections 3 and 4 using the analogues of Propositions 1, 2 and 3 that can be derived from Jeon et al. (2022).

Another extension that is immediate is to use different bandwidths h_j for different covariates X_j, instead of a common h for all X_j. In this case, the bias part in Corollary 2 is modified to

$$\frac{1}{2}\sum_{j=1}^{d} \mu_2(x_j,\kappa)\, h_j^2 \left(m_j''(x_j) - \theta_{0j}\, g_j''(x_j)\right).$$

Thus, the largest gain in the estimation of m with the parametric help is attained by using a consistent estimator of

$$\boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}_0(h_1,\ldots,h_d) := \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathrm{E}\left[\left(\sum_{j=1}^{d} h_j^2\big(m_j''(x_j) - \theta_j\, g_j''(x_j)\big)\right)^2\right],$$

which depends on the bandwidths h_j, contrary to the case of using a common bandwidth. However, for the estimation of the individual component functions, the targets θ_{0j} remain the same as in Section 4.2. A further extension of the present study is to consider a *nonlinear* parametric family where $g(\mathbf{x},\boldsymbol{\theta})$ is nonlinear in $\boldsymbol{\theta}$ while it is additive in modeling the effects of $\mathbf{X}$. This would only involve additional but straightforward complication in the corresponding methodology and theory.

Although we studied the advantage of the parametric help in the case where the underlying model is additive, one may be also interested in what happens when the true model is not additive. Here, we discuss the estimation of the multivariate regression function m only since it is hard to think of individual components of a non-additive function. In the non-additive case, our proposal with the parametric help is not actually considered as an estimator of the regression function $m = \mathrm{E}(Y|\mathbf{X} = \cdot)$ itself, but as an estimator of its *projection* onto the space of additive functions $\mathscr{H}_{\mathrm{add}}$. Call the projection m_+. We believe that one still benefits from the parametric help in terms of estimating m_+. To investigate this, one first needs to derive versions of Propositions 1 and 2 for a general response variable W with non-additive $m_W = \mathrm{E}(W|\mathbf{X} = \cdot)$. Let $m_{W,+}$ be the *projection* of m_W onto $\mathscr{H}_{\mathrm{add}}$. We continue to let $\hat{m}_W$ denote the solution of the SBF equation (2.4). Then, the versions of Propositions 1 and 2 basically replace m_W by $m_{W,+}$ and $m_{W,j}''$ by the second derivatives of the components of $m_{W,+}$. In the proofs of these versions, $\varepsilon_{W,+} := W - m_{W,+}(\mathbf{X})$ takes the role of $\varepsilon_W := W - m_W(\mathbf{X})$. One notable difference from the additive case is that one cannot use $\mathrm{E}(\varepsilon_{W,+}|\mathbf{X}) = 0$ since it is no longer valid in case m_W is not additive. Instead, one has

$$\mathrm{E}(\varepsilon_{W,+}|X_j) = 0, \quad 1 \le j \le d \tag{6.1}$$

since $\mathrm{E}(\varepsilon_{W,+}|\mathbf{X} = \cdot) \equiv m_W - m_{W,+}$ is perpendicular to $\mathscr{H}_{\mathrm{add}}$ and thus

$$0 = \mathrm{E}\big[\mathrm{E}(\varepsilon_{W,+}|\mathbf{X})\eta(X_j)\big] = \mathrm{E}\big[\mathrm{E}(\varepsilon_{W,+}|X_j)\eta(X_j)\big]$$

for all square integrable functions η. With (6.1), one may be able to prove the versions of Propositions 1 and 2. However, there are several hurdles one needs to overcome in going through other details with

$\varepsilon_+ := Y - m_+(\mathbf{X})$ replacing ε. For example, for the consistency of $\hat{\boldsymbol{\theta}}_0$ (Theorem 3), one needs to prove, in the non-additive case, that

$$n^{-1}\sum_{i=1}^{n}\hat{D}_j(X_{ij})\cdot\varepsilon_{+,i} = o_p\big(n^{-4/5}\big), \tag{6.2}$$

where $\hat{D}_j$ is defined at (B.4) and $\varepsilon_{+,i} := Y_i - m_+(\mathbf{X}_i)$. Compare (6.2) with the equations at (B.6). We note that the second equation at (B.6) is crucially dependent upon $\mathrm{E}(\varepsilon_i|\mathbf{X}_i) = 0$. However, one has only $\mathrm{E}(\varepsilon_+|X_j) = 0$ from (6.1), and the second equation is no longer valid for $\varepsilon_{+,i}$ replacing ε_i. The main reason is that $\hat{D}_j(X_{ij})$ involves not only $\{X_{1j},\ldots,X_{nj}\}$ but also $\{X_{1k},\ldots,X_{nk}\}$ for all $k \neq j$. We note that (6.2) follows if one proves

$$\sup_{D_j\in\mathcal{G}_n}|n^{-1}\sum_{i=1}^{n}D_j(X_{ij})\cdot\varepsilon_{+,i}| = o_p\big(n^{-4/5}\big) \tag{6.3}$$

for some class of functions $\mathcal{G}_n$ where $\hat{D}_j$ belongs with probability tending to one. Such a class would be for those D_j such that $\sup_{x_j\in[0,1]}|D_j(x_j)| \le C_1 n^{-2/5}$ and $|D_j(x_j) - D_j(x_j')| \le C_2 n^{-1/5}|x_j - x_j'|$ for all $x_j, x_j' \in [0,1]$. Then, one may employ a maximal inequality from the empirical process theory to prove (6.3), where we believe one does not need $\mathrm{E}(\varepsilon_+|\mathbf{X}) = 0$ but $\mathrm{E}(\varepsilon_+|X_j) = 0$ would be enough. For this approach to work, however, the bandwidth, say b, in the construction of $\hat{\boldsymbol{\theta}}_0$ might need to be of larger magnitude than $n^{-1/5}$. If this is the case, then one needs to prove (6.2) and (6.3) with $o_p(b^4)$ replacing $o_p(n^{-4/5})$, and take $\mathcal{G}_n$ accordingly, which in turn requires versions of Propositions 1 and 2 for a flexible range of bandwidth. We think this is an interesting topic for future study.

Appendix A: Proofs of theorems in Section 3

A.1. Proof of Theorem 1

Using the linearity of the SBF operation asserted in Proposition 1, it holds that

$$\begin{aligned}
&\hat{\mathbf{m}}^{\mathrm{tp}}(\mathbf{x}) - \mathbf{m}^{\mathrm{tp}}(\mathbf{x})\\
&\quad= \hat{\mathbf{m}}^{\mathrm{tp}}_{Y}(\mathbf{x}) - \sum_{j=1}^{d}\hat{\theta}_{0j}\hat{\mathbf{m}}^{\mathrm{tp}}_{g_j(X_j)}(\mathbf{x}) + \sum_{j=1}^{d}\hat{\theta}_{0j}\mathbf{g}^{\mathrm{tp}}_j(x_j) - \mathbf{m}^{\mathrm{tp}}(\mathbf{x})\\
&\quad= \hat{\mathbf{m}}_{Y-\boldsymbol{\theta}_0^\top\mathbf{g}(\mathbf{X})}(\mathbf{x}) - \left(\mathbf{m}^{\mathrm{tp}}(\mathbf{x}) - \sum_{j=1}^{d}\theta_{0j}\mathbf{g}^{\mathrm{tp}}_j(x_j)\right) - \sum_{j=1}^{d}(\hat{\theta}_{0j}-\theta_{0j})\big(\hat{\mathbf{m}}^{\mathrm{tp}}_{g_j(X_j)}(\mathbf{x}) - \mathbf{g}^{\mathrm{tp}}_j(x_j)\big).
\end{aligned}$$

An application of Proposition 2 to $W \equiv g_j(X_j)$ gives that $\hat{\mathbf{m}}^{\mathrm{tp}}_{g_j(X_j)}(\mathbf{x}) - \mathbf{g}^{\mathrm{tp}}_j(x_j) = O_p(n^{-2/5})$ uniformly for $\mathbf{x}\in[0,1]^d$. This and the assumed consistency of $\hat{\boldsymbol{\theta}}_0$ imply the theorem.

A.2. Proof of Theorem 2

Let $(\hat{m}_1,\ldots,\hat{m}_d)$ be a tuple such that $\hat{m} = \hat{m}_1 + \cdots + \hat{m}_d$. Then, from Theorem 1 and Lemma 3 in the Appendix C, it holds that, for some $\hat{c}_j = O_p(1)$,

$$\hat{m}_j(x_j) = \hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),j}(x_j) + \theta_{0j}\, g_j(x_j) + \hat{c}_j + o_p\left(n^{-2/5}\right) \tag{A.1}$$

uniformly for $x_j \in [0,1]$. Also, we get that, uniformly for $x_j \in [0,1]$,

$$\hat{m}_{1,j}(x_j) = \hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),1,j}(x_j) + h\,\theta_{0j}\, g_j'(x_j) + o_p\left(n^{-2/5}\right). \tag{A.2}$$

Now, for a tuple $(\hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),1},\ldots,\hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),d})$ comprising

$$\hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X})}(\mathbf{x}) = \hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),1}(x_1) + \cdots + \hat{m}_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),d}(x_d),$$

let $\hat{m}^c_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),j}$ be their centered versions satisfying the constraints (2.5) with $W = Y - \boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X})$. Then, by (A.1) and (A.2)

$$\begin{aligned}\hat{m}^c_j(x_j) = \hat{m}^c_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),j}(x_j) + \theta_{0j}\left[g_j(x_j) - \int_0^1 \left(g_j(u)\hat{p}_j(u) + h g_j'(u)\hat{p}_{1,j}(u)\right) du\right] \\ + o_p\left(n^{-2/5}\right),\end{aligned} \tag{A.3}$$

uniformly for $x_j \in [0,1]$. Note that the $\hat{c}_j$ in (A.1) are canceled out in the above equation for the centered versions.

On the other hand, applying Proposition 3 to $W = Y - \boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X})$, we obtain

$$\begin{aligned}\hat{m}^c_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),j}(x_j) = m^c_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),j}(x_j) + \tilde{m}^A_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),j}(x_j) \\ + \frac{h^2}{2}\,\mu_2(x_j,\kappa)\left[m_j''(x_j) - \theta_{0j} g_j''(x_j)\right] + o_p\left(n^{-2/5}\right)\end{aligned} \tag{A.4}$$

uniformly for $x_j \in [0,1]$, where $\tilde{m}^A_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),j}$ is the first element of the 2-vector $\tilde{\mathbf{m}}^A_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),j}$. Since

$$m^c_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),j}(x_j) = m^c_j(x_j) - \theta_{0j}\left[g_j(x_j) - \int_0^1 g_j(u)p_j(u)\,du\right],$$

the approximations (A.3) and (A.4) imply that, for all $1 \le j \le d$,

$$\begin{aligned}\hat{m}^c_j(x_j) = m^c_j(x_j) + \tilde{m}^A_{Y-\boldsymbol{\theta}_0^\top \mathbf{g}(\mathbf{X}),j}(x_j) + \frac{h^2}{2}\,\mu_2(x_j,\kappa)\left[m_j''(x_j) - \theta_{0j} g_j''(x_j)\right] \\ + \theta_{0j}\left[\int_0^1 g_j(u)p_j(u)\,du - \int_0^1 \left(g_j(u)\hat{p}_j(u) + h g_j'(u)\hat{p}_{1,j}(u)\right) du\right] + o_p\left(n^{-2/5}\right)\end{aligned} \tag{A.5}$$

uniformly for $x_j \in [0,1]$. We elaborate on the integrals in the last term in (A.5). We observe that

$$\begin{aligned}
&\int_0^1 g_j(u)p_j(u)\,du - \int_0^1 \left(g_j(u)\hat{p}_j(u) + hg_j'(u)\hat{p}_{1,j}(u)\right)\,du \\
&\quad = \int_{[0,1]^2} \left[g_j(v) - g_j(u) - (v-u)g_j'(u)\right] p_j(v)K_h(u,v)\,dv\,du + O_p\left(n^{-1/2}\right) \\
&\quad = \frac{h^2}{2}\int_0^1 \mu_2(u,K)g_j''(u)p_j(u)\,du + o_p\left(n^{-2/5}\right) \\
&\quad = \frac{h^2}{2}\int_0^1 \mu_2(u,\kappa)g_j''(u)p_j(u)\,du + o_p\left(n^{-2/5}\right).
\end{aligned} \tag{A.6}$$

In (A.6), the first equation follows from the normalization property of the kernel $K_h(\cdot,\cdot)$ at (2.1) and the fact that

$$\begin{aligned}
&\int_0^1 \left(g_j(u)\hat{p}_j(u) + hg_j'(u)\hat{p}_{1,j}(u)\right)\,du \\
&\quad = \mathrm{E}\left(\int_0^1 \left(g_j(u)\hat{p}_j(u) + hg_j'(u)\hat{p}_{1,j}(u)\right)\,du\right) + O_p\left(n^{-1/2}\right) \\
&\quad = \int_{[0,1]^2} \left[g_j(u) + (v-u)g_j'(u)\right] p_j(v)K_h(u,v)\,dv\,du + O_p\left(n^{-1/2}\right).
\end{aligned}$$

The approximation at (A.6) with (A.5) gives the theorem.

Appendix B: Proofs of theorems in Section 4

B.1. Proof of Theorem 3

Recall $\hat{\varepsilon}_{W,i} = W_i - \hat{m}_W(\mathbf{X}_i)$. Put $m^{\ominus}(\mathbf{x}) := m(\mathbf{x}) - \boldsymbol{\theta}_0^{\top}\mathbf{g}(\mathbf{x})$. We claim

$$n^{-1}\sum_{i=1}^n \hat{\varepsilon}_{m^{\ominus}(\mathbf{X}),i}\cdot\hat{\varepsilon}_{g_j(X_j),i} = o_p\left(n^{-4/5}\right), \quad 1 \le j \le d. \tag{B.1}$$

It also holds that, for any additive function η,

$$n^{-1}\sum_{i=1}^n \hat{\varepsilon}_{\varepsilon,i}\cdot\hat{\varepsilon}_{\eta(\mathbf{X}),i} = o_p\left(n^{-4/5}\right), \quad 1 \le j \le d. \tag{B.2}$$

The two claims give the theorem. To see this, we note that

$$Y_i^{\boldsymbol{\theta}} - \hat{m}_{Y^{\boldsymbol{\theta}}}(\mathbf{X}_i) = \hat{\varepsilon}_{m^{\ominus}(\mathbf{X}),i} + \hat{\varepsilon}_{\varepsilon,i} - (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\top}\hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i}.$$

From the definition of $\hat{\boldsymbol{\theta}}_0$ at (4.4), we may write

$$\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0 + \left(n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i}\cdot\hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i}^{\top}\right)^{-1}\left(n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i}\cdot\hat{\varepsilon}_{\varepsilon,i} + n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i}\cdot\hat{\varepsilon}_{m^{\ominus}(\mathbf{X}),i}\right).$$

Now, applying the second part of Corollary 1 with $\eta(\mathbf{x}) = g_j(x_j)$ for each $1 \le j \le d$, we get

$$\hat{m}_{g_j(X_j)}(\mathbf{x}) - g_j(x_j) = \frac{h^2}{2}\,\mu_2(x_j,\kappa)\,g_j''(x_j) + o_p\big(n^{-2/5}\big)$$

uniformly for $x_j \in [0,1]$. This gives that

$$n^{-1}\sum_{i=1}^{n} \hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i} \cdot \hat{\boldsymbol{\varepsilon}}_{\mathbf{g}(\mathbf{X}),i}^{\top} = \frac{h^4}{4}\,\mu_2(K)^2\,\mathrm{E}\left(\mathbf{g}''(\mathbf{X})\mathbf{g}''(\mathbf{X})^{\top}\right) + o_p\big(n^{-4/5}\big).$$

This and the two claims (B.1) and (B.2) entail $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0 + o_p(1)$.

It remains to prove the two claims at (B.1) and (B.2). For (B.1), recall that $\mu_2(x_j,\kappa)$ is the first entry of the 2-vector $\mathbf{N}(x_j)^{-1}\boldsymbol{\gamma}(x_j)$. We apply the second part of Corollary 1 twice to $\eta(\mathbf{x}) = m^{\ominus}(\mathbf{x})$ and to $\eta(\mathbf{x}) = g_j(x_j)$. By taking the first entries of the respective tuples in the resulting expansions, we obtain

$$\max_{1\le i\le n} \big|\,\hat{\varepsilon}_{m^{\ominus}(\mathbf{X}),i} - \frac{h^2}{2}\sum_{k=1}^{d}\mu_2(X_{ik},\kappa)\big(m_k''(X_{ik}) - \theta_{0k}\,g_k''(X_{ik})\big)\,\big| = o_p\big(n^{-2/5}\big),$$

$$\max_{1\le i\le n} \big|\,\hat{\varepsilon}_{g_j(X_j),i} - \frac{h^2}{2}\,\mu_2(X_{ij},\kappa)\,g_j''(X_{ij})\big| = o_p\big(n^{-2/5}\big).$$

The above two uniform expansions entail that, for each $1 \le j \le d$,

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n} \hat{\varepsilon}_{m^{\ominus}(\mathbf{X}),i}\cdot\hat{\varepsilon}_{g_j(X_j),i}\\
&\quad= \frac{h^4}{4n}\sum_{k=1}^{d}\sum_{i=1}^{n}\mu_2(X_{ik},\kappa)\mu_2(X_{ij},\kappa)\big(m_k''(X_{ik}) - \theta_{0k}\,g_k''(X_{ik})\big)\,g_j''(X_{ij}) + o_p\big(n^{-4/5}\big)\\
&\quad= \frac{h^4}{4}\,\mu_2(K)^2\,\mathrm{E}\big(\big[m_+''(\mathbf{X}) - \boldsymbol{\theta}_0^{\top}\mathbf{g}''(\mathbf{X})\big]\cdot g_j(X_j)\big) + o_p\big(n^{-4/5}\big)\\
&\quad= o_p\big(n^{-4/5}\big),
\end{aligned}$$

where the last equality follows from the definition of $\boldsymbol{\theta}_0$ at (4.2).

Next, we prove (B.2). Applying Proposition 2 with $W = \varepsilon$ and the second part of Corollary 1, and since $\mathrm{E}(\varepsilon|\mathbf{X}) = 0$, we get

$$\begin{aligned}
n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_{\varepsilon,i}\cdot\hat{\varepsilon}_{\eta(\mathbf{X}),i} &= n^{-1}\sum_{i=1}^{n}\varepsilon_i\big(\eta(\mathbf{X}_i) - \hat{m}_{\eta(\mathbf{X})}(\mathbf{X}_i)\big)\\
&\quad - \sum_{j=1}^{d} n^{-1}\sum_{i=1}^{n}\tilde{m}_{\varepsilon,j}(X_{ij})\big(\eta(\mathbf{X}_i) - \hat{m}_{\eta(\mathbf{X})}(\mathbf{X}_i)\big) + o_p\big(n^{-4/5}\big),
\end{aligned} \tag{B.3}$$

where we have also used the simplification that $\tilde{m}_{\varepsilon,j}^{A}$ reduces to $\tilde{m}_{\varepsilon,j}$. We prove that both terms on the right hand side of (B.3) are of magnitude $o_p(n^{-4/5})$. We give a proof for the second term only since the

first one is simpler. Put

$$\hat{D}_j(x_j) := n^{-1} \sum_{i=1}^{n} \omega_{j,h}(X_{ij}, x_j)\big(\eta(\mathbf{X}_i) - \hat{m}_{\eta(\mathbf{X})}(\mathbf{X}_i)\big). \tag{B.4}$$

From the second part of Corollary 1, it follows that

$$\sup_{x_j \in [0,1]} |\hat{D}_j(x_j)| = O_p\big(n^{-2/5}\big). \tag{B.5}$$

Since $\hat{D}_j(x_j)$ involves only $\mathbf{X}_i$, $1 \le i \le n$, and $\mathrm{E}(\varepsilon_i|\mathbf{X}_i) = 0$, we find from (B.5)

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} \tilde{m}_{\varepsilon,j}(X_{ij})\big(\eta(\mathbf{X}_i) - \hat{m}_{\eta(\mathbf{X})}(\mathbf{X}_i)\big) &= n^{-1} \sum_{i=1}^{n} \hat{D}_j(X_{ij}) \cdot \varepsilon_i \\ &= O_p\left(\sqrt{n^{-2} \sum_{i=1}^{n} \hat{D}_j(X_{ij})^2 \,\mathrm{E}\big(\varepsilon_i^2 | \mathbf{X}_1, \ldots, \mathbf{X}_n\big)}\right) \\ &= O_p\big(n^{-9/10}\big). \end{aligned} \tag{B.6}$$

This completes the proof of (B.2).

B.2. Proof of Theorem 5

First, we note that $\tilde{m}^A_{Y,j} = \tilde{m}_{\varepsilon,j}$. By applying Proposition 3 to $W = Y$, we get

$$\max_{1 \le i \le n} |\,\hat{m}^c_{Y,j}(X_{ij}) - m^c_j(X_{ij}) - \frac{h^2}{2}\mu_2(X_{ij}, \kappa)\, m''_j(X_{ij}) + \tilde{m}_{\varepsilon,j}(X_{ij})\,| = o_p\big(n^{-2/5}\big).$$

This gives

$$\max_{1 \le i \le n} |\,\hat{\delta}_i\big(j, \hat{m}^c_{Y,j}(X_j)\big) + \frac{h^2}{2}\mu_2(X_{ij}, \kappa)\, m''_j(X_{ij}) - \hat{\delta}_i\big(j, \tilde{m}_{\varepsilon,j}(X_j)\big)\,| = o_p\big(n^{-2/5}\big).$$

Also, we note that

$$\max_{1 \le i \le n} |\,\hat{\delta}_i\big(j, g_j(X_j)\big) + \frac{h^2}{2}\mu_2(X_{ij}, \kappa)\, g''_j(X_{ij})\,| = o_p\big(n^{-2/5}\big).$$

It suffices to prove then

$$n^{-1} \sum_{i=1}^{n} \hat{\delta}^{\mathrm{dev}}_i\big(j, g_j(X_j)\big) \cdot \hat{\delta}_i\big(j, \tilde{m}_{\varepsilon,j}(X_j)\big) = o_p\big(n^{-4/5}\big). \tag{B.7}$$

To prove (B.7), put

$$\hat{\Delta}_{i,j} := n^{-1} \sum_{i'=1}^{n} \left[\hat{\delta}^{\mathrm{dev}}_{i'}\big(j, g_j(X_j)\big) - n^{-1} \sum_{i''=1}^{n} \omega_{j,h}(X_{i''j}, X_{i'j})\hat{\delta}^{\mathrm{dev}}_{i''}\big(j, g_j(X_j)\big)\right] \omega_{j,h}(X_{i'j}, X_{ij}).$$

Then, the left hand side of (B.7) equals $T := n^{-1}\sum_{i=1}^{n}\hat{\Delta}_{i,j}\cdot\varepsilon_i$. We claim

$$n^{-1}\sum_{i=1}^{n}\hat{\Delta}_{i,j}^2 = O_p\big(n^{-4/5}\big). \tag{B.8}$$

Since $\hat{\Delta}_{i,j}$ for all i depend on solely on $X_{1j},\ldots X_{nj}$, we get $\mathrm{E}(T|X_{1j},\ldots X_{nj}) = 0$. From (B.8) we also have

$$\mathrm{Var}(T|X_{1j},\ldots X_{nj}) = n^{-2}\sum_{i=1}^{n}\hat{\Delta}_{i,j}^2\cdot\mathrm{E}\big(\varepsilon_i^2|X_{ij}\big) = O_p\big(n^{-9/5}\big).$$

This proves (B.7).

It remains to prove the claim (B.8). Define

$$A_j(x_j) := n^{-1}\sum_{i=1}^{n}\omega_{j,h}(X_{ij},x_j)\cdot\hat{\delta}_i^{\mathrm{dev}}\big(j,g_j(X_j)\big),$$

$$B_j(x_j) := n^{-1}\sum_{i=1}^{n}\omega_{j,h}(X_{ij},x_j)\cdot A_j(X_{ij}).$$

Then, $\hat{\Delta}_{i,j} = A_j(X_{ij}) - B_j(X_{ij})$. It suffices to prove

$$\sup_{x_j\in[0,1]}|A_j(x_j)| = O_p\big(h^2\big),\qquad \sup_{x_j\in[0,1]}|B_j(x_j)| = O_p\big(h^2\big). \tag{B.9}$$

To prove the first assertion in (B.9), we note that

$$\begin{aligned}\sup_{1\le i\le n}|\hat{\delta}_i^{\mathrm{dev}}\big(j,g_j(X_j)\big)| &\le \sup_{1\le i\le n}|\hat{\delta}_i\big(j,g_j(X_j)\big)| + n^{-1}\sum_{i=1}^{n}\sup_{1\le i\le n}|\hat{\delta}_i\big(j,g_j(X_j)\big)|\\ &\le 2\sup_{1\le i\le n}|\hat{\delta}_i\big(j,g_j(X_j)\big)|.\end{aligned}$$

Recall that $\hat{\delta}_i(j,g_j(X_j)) = -\frac{h^2}{2}\mu_2(X_{ij},\kappa)\,g_j''(X_{ij}) + o_p(h^2)$ uniformly for $1\le i\le n$. Since the assumption (A2) on the baseline kernel K ensures that $\sup_{u\in[0,1]}|\mu_2(u,\kappa)|\le C_1$ for some absolute constant $0 < C_1 < \infty$, we get $\sup_{1\le i\le n}|\hat{\delta}_i(j,g_j(X_j))| = O_p(h^2)$ and thus $\sup_{1\le i\le n}|\hat{\delta}_i^{\mathrm{dev}}(j,g_j(X_j))| = O_p(h^2)$. Since there exists an absolute constant $0 < C_2 < \infty$ such that

$$n^{-1}\sum_{i=1}^{n}|\omega_{j,h}(X_{ij},x_j)| \le C_2\cdot n^{-1}\sum_{i=1}^{n}K_h(X_{ij},x_j), \tag{B.10}$$

this completes the proof of the first assertion in (B.9). The second assertion in (B.9) follows immediately from (B.10) and the first assertion.

Appendix C: Further technical details

Here, we bring in relevant function spaces in our theoretical development in Sections 2–4, together with some linear operators mapping $\mathscr{M}$ to $\mathscr{M}_j$ and $\mathscr{M}_{\mathrm{add}}$ to itself, which pertain to the SBF operation that we described in Section 2. We also present the proofs of Propositions 1–3 in Section 2.

C.1. Function spaces

Let $\mathfrak{vec}(\mathbf{u}) := (1, u_1, \ldots, u_d)^\top$ for $\mathbf{u} = (u_1, \ldots, u_d) \in \mathbb{R}^d$ and define

$$\mathfrak{M}(\mathbf{x}) := p(\mathbf{x}) \cdot \int \mathfrak{vec}(\mathbf{u})\mathfrak{vec}(\mathbf{u})^\top \prod_{k=1}^{d} K(u_k)\, d\mathbf{u}.$$

The space that embodies all function tuples considered in our paper is the collection all $(d+1)$-tuples of real-valued functions

$$\mathscr{M} := \Bigg\{ \mathbf{f}^{\mathrm{tp}} : \mathbf{f}^{\mathrm{tp}}(\mathbf{x}) = \left(f^0(\mathbf{x}), f^{1,1}(\mathbf{x}), \ldots, f^{1,d}(\mathbf{x})\right)^\top \text{ for real-valued functions } f^0 \text{ and } f^{1,j} \text{ defined on } [0,1]^d, \int_{[0,1]^d} \mathbf{f}^{\mathrm{tp}}(\mathbf{x})^\top \mathfrak{M}(\mathbf{x}) \mathbf{f}^{\mathrm{tp}}(\mathbf{x})\, d\mathbf{x} < \infty \Bigg\}.$$

We endow $\mathscr{M}$ with the inner product

$$\left\langle \mathbf{f}^{\mathrm{tp}}, \mathbf{g}^{\mathrm{tp}} \right\rangle_2 := \int_{[0,1]^d} \mathbf{f}^{\mathrm{tp}}(\mathbf{x})^\top \mathfrak{M}(\mathbf{x}) \mathbf{g}^{\mathrm{tp}}(\mathbf{x})\, d\mathbf{x}, \quad \mathbf{f}^{\mathrm{tp}}, \mathbf{g}^{\mathrm{tp}} \in \mathscr{M},$$

and let $\|\cdot\|_2$ be the induced norm.

Now, we introduce function spaces pertaining to additive functions and individual component functions. Let $\mathscr{M}_{\mathrm{add}}$ be the subspace of $\mathscr{M}$ such that, for $\mathbf{f}^{\mathrm{tp}} \in \mathscr{M}_{\mathrm{add}}$,

$$f^0(\mathbf{x}) = f_1(x_1) + \cdots + f_d(x_d), \quad f^{1,j}(\mathbf{x}) = f_{1,j}(x_j)$$

for some univariate functions f_j and $f_{1,j}$. Now, let $\mathscr{M}_j$ be the subspaces of $\mathscr{M}_{\mathrm{add}} \subset \mathscr{M}$ containing $\mathbf{f}^{\mathrm{tp}}$ such that $f^0(\mathbf{x}) = f_j(x_j)$, $f^{1,k}(\mathbf{x}) \equiv 0$ for all $k \neq j$ and $f^{1,j}(\mathbf{x}) = f_{1,j}(x_j)$ for some univariate functions f_j and $f_{1,j}$. Below throughout this paper, when we say $f = \tilde{f}$ for $f, \tilde{f} \in \mathscr{M}$, we mean that $\|f - \tilde{f}\|_2 = 0$. Since not all points in a rectangle with non-empty interior lie on a common hyperplane in $\mathbb{R}^d$, $\|\mathbf{f}^{\mathrm{tp}}\|_2 = 0$ is equivalent to that $\mathbf{f}^{\mathrm{tp}}(\mathbf{x}) = \mathbf{0}$ a.e. for $\mathbf{x} \in [0,1]^d$, provided that p is bounded away from zero.

C.2. Smooth backfitting linear operators

Recall $\mathbf{vec}(u) = (1, u)^\top \in \mathbb{R}^2$ and define

$$\mathbf{M}_{jj}(x_j) := p_j(x_j) \cdot \int \mathbf{vec}(u)\mathbf{vec}(u)^\top K(u)\, du.$$

The $\mathbf{M}_{jj}(x_j)$ are considered as approximations of $\hat{\mathbf{M}}_{jj}(x_j)$. Define $\pi_j : \mathscr{M} \to \mathscr{M}_j$ by

$$\pi_j\left(\mathbf{f}^{\mathrm{tp}}\right)(x_j) := \mathbf{U}_j^\top \mathbf{M}_{jj}(x_j)^{-1} \mathbf{U}_j \int_{[0,1]^{d-1}} \mathfrak{M}(\mathbf{x}) \mathbf{f}^{\mathrm{tp}}(\mathbf{x})\, d\mathbf{x}_{-j}, \quad \mathbf{f}^{\mathrm{tp}} \in \mathscr{M}.$$

It can be shown that each $\pi_j : \mathscr{M} \to \mathscr{M}_j$ is a projection operator, provided that $p_j(x_j) > 0$ for all $x_j \in [0,1]$. Put $T = (I - \pi_d) \circ (I - \pi_{d-1}) \circ \cdots \circ (I - \pi_1) : \mathscr{M}_{\mathrm{add}} \to \mathscr{M}_{\mathrm{add}}$.

To introduce sample versions of the projection operators, let

$$\hat{\mathfrak{M}}(\mathbf{x}) := n^{-1} \sum_{i=1}^{n} \mathfrak{vec}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right) \mathfrak{vec}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right)^\top \prod_{k=1}^{d} K_h(x_k, X_{ik}).$$

Define $\hat{\pi}_j : \mathscr{M} \to \mathscr{M}_j$ by

$$\hat{\pi}_j\big(\mathbf{f}^{\mathrm{tp}}\big)(x_j) := \mathbf{U}_j^\top \hat{\mathbf{M}}_{jj}(x_j)^{-1}\mathbf{U}_j \int_{[0,1]^{d-1}} \hat{\mathfrak{M}}(\mathbf{x})\mathbf{f}^{\mathrm{tp}}(\mathbf{x})\, d\mathbf{x}_{-j}, \quad \mathbf{f}^{\mathrm{tp}} \in \mathscr{M}.$$

Put $\hat{T} := (I - \hat{\pi}_d) \circ (I - \hat{\pi}_{d-1}) \circ \cdots \circ (I - \hat{\pi}_1) : \mathscr{M}_{\mathrm{add}} \to \mathscr{M}_{\mathrm{add}}$, where $I : \mathscr{M} \to \mathscr{M}$ is the identity map. Then, the SBF equation at (2.3) can be expressed in terms of $\hat{\mathbf{m}}_W^{\mathrm{tp}} = (\hat{m}_{W,1} + \cdots + \hat{m}_{W,d}, \hat{m}_{W,1,j}, \ldots, \hat{m}_{W,1,d})^\top$ and

$$\hat{\mathbf{r}}_{W,+} := \sum_{j=1}^{d} (I - \hat{\pi}_d) \circ \cdots \circ (I - \hat{\pi}_{j+1})\mathbf{U}_j^\top \tilde{\mathbf{m}}_{W,j}, \tag{C.1}$$

where we interpret $(I - \hat{\pi}_d) \circ \cdots \circ (I - \hat{\pi}_{j+1}) = I$ for $j = d$. Indeed, the SBF equation at (2.3) is equivalent to

$$\hat{\mathbf{m}}_W^{\mathrm{tp}} = \hat{T}\big(\hat{\mathbf{m}}_W^{\mathrm{tp}}\big) + \hat{\mathbf{r}}_{W,+}. \tag{C.2}$$

For an operator $\mathcal{L} : \mathscr{M}_{\mathrm{add}} \to \mathscr{M}_{\mathrm{add}}$, define its operator norm by

$$\|\mathcal{L}\|_{\mathrm{op}} := \sup\big\{\|\mathcal{L}\big(\mathbf{f}^{\mathrm{tp}}\big)\|_2 : \mathbf{f}^{\mathrm{tp}} \in \mathscr{M}_{\mathrm{add}} \text{ with } \|\mathbf{f}^{\mathrm{tp}}\|_2 = 1\big\}.$$

C.3. Basic lemmas and their proofs

Lemma 1. *Assume that p is bounded away from zero and infinity on $[0,1]^d$. Then, there exists a constant $0 < c < \infty$ depending only on d such that, for all $\mathbf{f}^{\mathrm{tp}} \in \mathscr{M}_{\mathrm{add}}$, there exists $\mathbf{f}_j^{\mathrm{tp}} \in \mathscr{M}_j$, $1 \le j \le d$, such that $\mathbf{f}^{\mathrm{tp}} = \mathbf{f}_1^{\mathrm{tp}} + \cdots + \mathbf{f}_d^{\mathrm{tp}}$ and $\max_{1\le j\le d} \|\mathbf{f}_j^{\mathrm{tp}}\|_2 \le c\|\mathbf{f}^{\mathrm{tp}}\|_2$. Furthermore, $\|T\|_{\mathrm{op}} < 1$.*

Proof. The projection operators $\pi_j : \mathscr{M}_k \to \mathscr{M}_j$ restricted to $\mathscr{M}_k$ for $k \ne j$ are compact since they are Hilbert-Schmidt. According to Proposition A.4.2 in Bickel et al. (1993), $\mathscr{M}_{\mathrm{add}}$ is closed in $\mathscr{M}$. The first part of the lemma is then an immediate consequence of applying Theorem 3.1 in Blot and Cieutat (2016). The second part follows from an application of Theorem 4.6 in Xu and Zikatanov (2002). □

Lemma 2. *Assume that p_j is continuous and bounded away from zero and infinity on $[0,1]$. Let $\mathbf{f}_j^{\mathrm{tp}}(\mathbf{x}) \equiv \mathbf{f}_j^{\mathrm{tp}}(x_j) = (f_j(x_j), 0, \ldots, 0, f_{1,j}(x_j), 0, \ldots, 0)^\top$ for $\mathbf{f}_j^{\mathrm{tp}} \in \mathscr{M}_j$. Likewise, let $\mathbf{f}_j^{c,\mathrm{tp}}$ be defined by $\mathbf{f}_j^{c,\mathrm{tp}}(x_j) := (f_j^c(x_j), 0, \ldots, 0, f_{1,j}, 0, \ldots, 0)^\top \in \mathscr{M}_j$, where $f_j^c(x_j) := f_j(x_j) - \int_0^1 (f_j(u)\hat{p}_j(u) + f_{1,j}(u)\hat{p}_{1,j}(u))\, du$. Then, for any $\epsilon > 0$ it holds that $\|\mathbf{f}_j^{\mathrm{tp}}\|_2^2 \ge (1-\epsilon)\|\mathbf{f}_j^{c,\mathrm{tp}}\|_2^2$ with probability tending to one.*

Proof. Let $c_j = \int_0^1 (f_j(u)\hat{p}_j(u) + f_{1,j}(u)\hat{p}_{1,j}(u))\, du$. Then, $\|(c_j, 0, \ldots, 0)^\top\|_2^2$ for the norm defined in Section C.1 equals c_j^2. This gives

$$\|\mathbf{f}_j^{\mathrm{tp}}\|_2^2 = \|\mathbf{f}_j^{c,\mathrm{tp}}\|_2^2 + c_j^2 + 2c_j \int_0^1 \big(\text{first row of } \hat{\mathfrak{M}}(\mathbf{x})\big) \cdot \mathbf{f}_j^{c,\mathrm{tp}}(\mathbf{x})\, d\mathbf{x}. \tag{C.3}$$

By the definition of $\mathbf{f}_j^{c,\mathrm{tp}}$ and noting that $\int_0^1 (\text{first row of } \hat{\mathfrak{M}}(\mathbf{x})) \cdot \mathbf{f}_j^{\mathrm{tp}}(\mathbf{x})\, d\mathbf{x}$ is nothing else than c_j, we get

$$\int_0^1 \big(\text{first row of } \hat{\mathfrak{M}}(\mathbf{x})\big) \cdot \mathbf{f}_j^{c,\mathrm{tp}}(\mathbf{x})\, d\mathbf{x} = 0.$$

Thus, applying Hölder's inequality we obtain

$$\begin{aligned}
&\int_0^1 \big(\text{first row of } \mathfrak{M}(\mathbf{x})\big)\cdot \mathbf{f}_j^{c,\mathrm{tp}}(\mathbf{x})\,d\mathbf{x} \\
&\quad= \int_0^1 \big(\text{first row of } \big[\mathfrak{M}(\mathbf{x})-\hat{\mathfrak{M}}(\mathbf{x})\big]\big)\cdot \mathbf{f}_j^{c,\mathrm{tp}}(\mathbf{x})\,d\mathbf{x} \\
&\quad\ge -\left(\int_0^1 \frac{(\hat{p}_j(x_j)-p_j(x_j))^2}{p_j(x_j)}\,dx_j + \int_0^1 \frac{\hat{p}_{1,j}(x_j)^2}{p_j(x_j)}\,dx_j\right)^{1/2} \\
&\qquad\qquad \cdot\left(\int_0^1 \big(f_j^c(x_j)^2 + f_{1,j}(x_j)^2\big)\,p_j(x_j)\,dx_j\right)^{1/2} \\
&\quad\ge -\hat{d}_{nj}\cdot\|\mathbf{f}_j^{c,\mathrm{tp}}\|_2
\end{aligned} \tag{C.4}$$

for some stochastic sequence $\{\hat{d}_{nj}\}$ such that $0\le\hat{d}_{nj}=o_p(1)$. For the second inequality in (C.4) we have used the facts that $\hat{p}_j(x_j)=\mu_0(x_j,K)p_j(x_j)+o_p(1)$ and $\hat{p}_{1,j}=\mu_1(x_j,K)p_j(x_j)+o_p(1)$ uniformly for $x_j\in[0,1]$ and that $\mu_0(x_j,K)=1$ and $\mu_1(x_j,K)=0$ for all $x_j\in[2h,1-2h]$ and they are bounded on $[0,1]^d$. From (C.3) and (C.4) it follows that

$$\begin{aligned}
\|\mathbf{f}_j^{\mathrm{tp}}\|_2^2 &= \|\mathbf{f}_j^{c,\mathrm{tp}}\|_2^2\cdot\big(1-\hat{d}_{nj}^2\big) + \big(\hat{d}_{nj}\cdot\|\mathbf{f}_j^{c,\mathrm{tp}}\|_2 - |c_j|\big)^2 \\
&\ge \|\mathbf{f}_j^{c,\mathrm{tp}}\|_2^2\cdot\big(1-\hat{d}_{nj}^2\big).
\end{aligned}$$

This completes the proof of the lemma. □

Lemma 3. *Let $f_j:[0,1]\to\mathbb{R}$ for $1\le j\le d$. If $\sum_{j=1}^d f_j(x_j)=0$ for a.e. $\mathbf{x}\in[0,1]^d$, then $f_j(x_j)=c_j$ a.e. for $x_j\in[0,1]$, where c_j are constants such that $\sum_{j=1}^d c_j=0$. If in addition each f_j satisfies $\mathrm{E}(f_j(X_j))=0$, then $f_j(x_j)=0$ a.e. for $x_j\in[0,1]$ for all $1\le j\le d$.*

Proof. We prove only the first part of the lemma since the second part is immediate from the first one. For the first part, we prove $f_1(x_1)=c_1$ for some constant c_1 a.e. for $x_1\in[0,1]$. Put $S:=\{\mathbf{x}\in[0,1]^d : f_1(x_1)+f_{-1}(\mathbf{x}_{-1})=0\}$, where $f_{-1}(\mathbf{x}_{-1}):=\sum_{j=2}^d f_j(x_j)$ and $\mathbf{x}_{-1}:=(x_2,\dots,x_d)$ for $\mathbf{x}=(x_1,\dots,x_d)$.

For an arbitrary measurable set $D\subset\mathbb{R}$, put

$$\begin{aligned}
A\equiv A(D) &:= f_1^{-1}(D)=\big\{x_1\in[0,1] : f_1(x_1)\in D\big\}, \\
B\equiv B(D) &:= f_{-1}^{-1}(D)=\big\{\mathbf{x}_{-1}\in[0,1]^{d-1} : -f_{-1}(\mathbf{x}_{-1})\in D\big\}.
\end{aligned}$$

Since $f_1(x_1)+f_{-1}(\mathbf{x}_{-1})=0$ on S, we get

$$\begin{aligned}
\big(A\times[0,1]^{d-1}\big)\cap S &= \big([0,1]\times B\big)\cap S \\
&= \big[\big(A\times[0,1]^{d-1}\big)\cap S\big]\cap\big[\big([0,1]\times B\big)\cap S\big] \\
&= (A\times B)\cap S.
\end{aligned} \tag{C.5}$$

Let Leb_q denote the q-dimensional Lebesgue measure for $q \geq 1$. Then, from (C.3) and since $\mathrm{Leb}_d(S) = 1$, we obtain

$$\begin{aligned} \mathrm{Leb}_1(A) &= \mathrm{Leb}_d\big(A \times [0,1]^{d-1}\big) \\ &= \mathrm{Leb}_d(A \times B) \\ &= \mathrm{Leb}_1(A) \cdot \mathrm{Leb}_{d-1}(B). \end{aligned} \tag{C.6}$$

Similarly, we get $\mathrm{Leb}_{d-1}(B) = \mathrm{Leb}_1(A) \cdot \mathrm{Leb}_{d-1}(B)$. This and (C.6) entail that

$$\mathrm{Leb}_1(A) = \mathrm{Leb}_1\big(f_1^{-1}(D)\big) = 0 \text{ or } 1$$

for any measurable set $D \subset \mathbb{R}$. Since $\mathrm{Leb}_1(f_1^{-1}(\mathbb{R})) = 1$ and it cannot happen that $\mathrm{Leb}_1(f_1^{-1}(D)) > 0$ and $\mathrm{Leb}_1(f_1^{-1}(D^c)) > 0$, the set $\mathbb{R}$ is an atom of the measure μ defined by $\mu(E) := \mathrm{Leb}_1(f_1^{-1}(E))$. According to Lemma 10.17 in Aliprantis and Border (2006), there exists a singleton $\{c_1\} \subset \mathbb{R}$ such that $\mathrm{Leb}_1(f_1^{-1}(\{c_1\})) > 0$. Since $\mathrm{Leb}_1(f_1^{-1}(\{c_1\}))$ must be either 0 or 1, we conclude $\mathrm{Leb}_1(f_1^{-1}(\{c_1\})) = 1$, which means $f_1(x_1) = c_1$ a.e. for $x_1 \in [0,1]$. This concludes the proof of the lemma. □

C.4. Proof of Proposition 1

We only outline the proof. It can be shown that $\|\hat{\pi}_j - \pi_j\|_{\mathrm{op}} = o_p(1)$. This implies $\|\hat{T} - T\|_{\mathrm{op}} = o_p(1)$. By the second part of Lemma 1 we get that

$$\|\hat{T}\|_{\mathrm{op}} < \tau \text{ with probability tending to one for some } 0 < \tau < 1. \tag{C.7}$$

We note that $\hat{\mathbf{r}}_{W,+}$ defined at (C.1) belongs to $\mathscr{M}_{\mathrm{add}}$ with probability tending to one. Indeed, we may prove that there exists a constant $0 < C < \infty$ such that $\|\hat{\mathbf{r}}_{W,+}\|_2 < C$ with probability tending to one. From the SBF equation at (C.2) and by (C.7), it holds that

$$\hat{\mathbf{m}}_W^{\mathrm{tp}} = \sum_{r=0}^{\infty} \hat{T}^r(\hat{\mathbf{r}}_{W,+}) \in \mathscr{M}_{\mathrm{add}} \tag{C.8}$$

with probability tending to one, where the convergence of the series is in $\|\cdot\|_2$. This proves the first part of the proposition.

To prove the second part, observe from the definition (2.2) that $\tilde{\mathbf{m}}_{c_1 W_1 + c_2 W_2, j} = c_1 \cdot \tilde{\mathbf{m}}_{W_1, j} + c_2 \cdot \tilde{\mathbf{m}}_{W_2, j}$. From the definition (C.1), $\hat{\mathbf{r}}_{W,+}$ is also linear in W. From (C.8) and the fact that $\hat{T}$ is a linear operator, so are $\hat{T}^r$ for all $r \geq 2$ as well, we may conclude that $\hat{\mathbf{m}}_W^{\mathrm{tp}}$ is linear in W.

C.5. Proof of Proposition 3

Define $\hat{\boldsymbol{\Delta}}_j^{c,\mathrm{tp}}$, which takes values in $\mathscr{M}_j$, by

$$\hat{\boldsymbol{\Delta}}_j^{c,\mathrm{tp}}(x_j) := \hat{\mathbf{m}}_{W,j}^{c,\mathrm{tp}}(x_j) - \mathbf{m}_{W,j}^{c,\mathrm{tp}}(x_j) - \mathbf{U}_j^{\top} \tilde{\mathbf{m}}_{W,j}^{A}(x_j) - \frac{h^2}{2} \cdot \mathbf{U}_j^{\top} \mathbf{N}(x_j)^{-1} \boldsymbol{\gamma}(x_j) \cdot m''_{W,j}(x_j). \tag{C.9}$$

Let $\hat{\Delta}_j^c$ and $\hat{\Delta}_{1,j}$ be the first and the $(j+1)$th entries of $\hat{\boldsymbol{\Delta}}_j^{c,\mathrm{tp}}$. Below, we prove

$$\sup_{x_j \in [0,1]} |\hat{\boldsymbol{\Delta}}_j^{c,\mathrm{tp}}(x_j)| = o_p\big(n^{-2/5}\big), \quad 1 \leq j \leq d. \tag{C.10}$$

Put $\hat{\mathbf{\Delta}}_+^{c,\mathrm{tp}}(\mathbf{x}) := \sum_{j=1}^d \hat{\mathbf{\Delta}}_j^{c,\mathrm{tp}}(x_j) = (\hat{\Delta}_1^c + \cdots + \hat{\Delta}_d^c, \hat{\Delta}_{1,1}, \ldots, \hat{\Delta}_{1,d})^\top$. Along the lines of the proof of Theorem 4.1 in Jeon and Park (2020), we may prove that, uniformly for $x_j \in [0,1]$,

$$\hat{\mathbf{\Delta}}_j^{c,\mathrm{tp}}(x_j) + \sum_{k=1,\neq j}^{d} \int_0^1 \mathbf{U}_j^\top \hat{\mathbf{M}}_{jj}(x_j)^{-1} \hat{\mathbf{M}}_{jk}(x_j, x_k) \mathbf{U}_k \cdot \hat{\mathbf{\Delta}}_k^{c,\mathrm{tp}}(x_k)\, dx_k \;=\; o_p\big(n^{-2/5}\big), \tag{C.11}$$

or equivalently $\hat{\pi}_j(\hat{\mathbf{\Delta}}_+^{c,\mathrm{tp}})(x_j) = o_p(n^{-2/5})$, for all $1 \le j \le d$. This gives

$$\hat{\mathbf{\Delta}}_+^{c,\mathrm{tp}}(\mathbf{x}) = \hat{T}\big(\hat{\mathbf{\Delta}}_+^{c,\mathrm{tp}}\big)(\mathbf{x}) + o_p\big(n^{-2/5}\big) \tag{C.12}$$

uniformly for $\mathbf{x} \in [0,1]^d$. Since $\|\hat{T}\|_{\mathrm{op}} \le \tau$ with probability tending to one for some $0 < \tau < 1$, as we have seen in the proof of Proposition 1, we get

$$\|\hat{\mathbf{\Delta}}_+^{c,\mathrm{tp}}\|_2 = o_p\big(n^{-2/5}\big). \tag{C.13}$$

Now, according to Lemma 1, there exists a constant $0 < c < \infty$ depending only on the dimension d and $\tilde{\mathbf{\Delta}}_j^{\mathrm{tp}} = (\tilde{\Delta}_j, 0, \ldots, 0, \tilde{\Delta}_{1,j}, 0, \ldots, 0)^\top \in \mathscr{M}_j$ such that $\hat{\mathbf{\Delta}}_+^{c,\mathrm{tp}} = \tilde{\mathbf{\Delta}}_1^{\mathrm{tp}} + \cdots + \tilde{\mathbf{\Delta}}_d^{\mathrm{tp}}$ and $\max_{1\le j\le d} \|\tilde{\mathbf{\Delta}}_j^{\mathrm{tp}}\|_2 \le c\|\hat{\mathbf{\Delta}}_+^{c,\mathrm{tp}}\|_2$. Furthermore, by Lemma 2 it holds that, for any $\epsilon > 0$,

$$\max_{1\le j\le d} \|\tilde{\mathbf{\Delta}}_j^{c,\mathrm{tp}}\|_2 \le (1-\epsilon)^{-1} \cdot \max_{1\le j\le d} \|\tilde{\mathbf{\Delta}}_j^{\mathrm{tp}}\|_2 \tag{C.14}$$

with probability tending to one, where $\tilde{\mathbf{\Delta}}_j^{c,\mathrm{tp}} = (\tilde{\Delta}_j^c, 0, \ldots, 0, \tilde{\Delta}_{1,j}, 0, \ldots, 0)^\top$ with

$$\tilde{\Delta}_j^c(x_j) = \tilde{\Delta}_j(x_j) - \int_0^1 \big(\tilde{\Delta}_j(u)\hat{p}_j(u) + \tilde{\Delta}_{1,j}(u)\hat{p}_{1,j}(u)\big)\, du.$$

We note that

$$0 = \int_0^1 \big(\tilde{\Delta}_j^c(x_j)\hat{p}_j(x_j) + \tilde{\Delta}_{1,j}(x_j)\hat{p}_{1,j}(x_j)\big)\, dx_j = \int_0^1 \big(\hat{p}_j(x_j), \hat{p}_{1,j}(x_j)\big) \cdot \mathbf{U}_j \tilde{\mathbf{\Delta}}_j^{c,\mathrm{tp}}(x_j)\, dx_j, \tag{C.15}$$

satisfying the constraints (2.5) for $\tilde{\Delta}_j^c$. Since $\hat{\mathbf{\Delta}}_+^{c,\mathrm{tp}} = \tilde{\mathbf{\Delta}}_1^{\mathrm{tp}} + \cdots + \tilde{\mathbf{\Delta}}_d^{\mathrm{tp}}$ and $p(\mathbf{x}) > 0$ on $[0,1]^d$, we get from Lemma 3 that $\hat{\mathbf{\Delta}}_j^{c,\mathrm{tp}} - \tilde{\mathbf{\Delta}}_j^{\mathrm{tp}} \equiv (c_j, 0, \ldots, 0)^\top \in \mathbb{R}^{d+1}$ on $[0,1]$ for some constant c_j. Likewise, $\hat{\mathbf{\Delta}}_j^{c,\mathrm{tp}} - \tilde{\mathbf{\Delta}}_j^{c,\mathrm{tp}} \equiv (c_j', 0, \ldots, 0)^\top \in \mathbb{R}^{d+1}$ on $[0,1]$ for some constant c_j'. We claim

$$\int_0^1 \big(\hat{p}_j(x_j), \hat{p}_{1,j}(x_j)\big) \cdot \mathbf{U}_j \hat{\mathbf{\Delta}}_j^{c,\mathrm{tp}}(x_j)\, dx_j = o_p\big(n^{-2/5}\big). \tag{C.16}$$

This with (C.15) entails $c_j' = o_p(n^{-2/5})$ since $\int_0^1 \hat{p}_j(x_j)\, dx_j = 1$, so that $\|\hat{\mathbf{\Delta}}_j^{c,\mathrm{tp}} - \tilde{\mathbf{\Delta}}_j^{c,\mathrm{tp}}\|_2 = o_p(n^{-2/5})$. From this with (C.13) and (C.14) we get

$$\|\hat{\mathbf{\Delta}}_j^{c,\mathrm{tp}}\|_2 \le \|\hat{\mathbf{\Delta}}_j^{c,\mathrm{tp}} - \tilde{\mathbf{\Delta}}_j^{c,\mathrm{tp}}\|_2 + \|\tilde{\mathbf{\Delta}}_j^{c,\mathrm{tp}}\|_2 = o_p\big(n^{-2/5}\big), \quad 1 \le j \le d.$$

Now, from (C.11) and an application of Hölder's inequality, we may conclude (C.10) since

$$\sup_{x_j\in[0,1]} |\hat{\mathbf{\Delta}}_j^{c,\mathrm{tp}}(x_j)| \le \sum_{k=1,\neq j} \|\hat{\mathbf{\Delta}}_k^{c,\mathrm{tp}}\|_2 \cdot O_p(1) + o_p\big(n^{-2/5}\big) = o_p\big(n^{-2/5}\big).$$

It remains to prove the claim (C.16). We multiply the row vector $(\hat{p}_j(x_j), \hat{p}_{1,j}(x_j))\mathbf{U}_j$ on both sides of the defining equation (C.9) for $\hat{\boldsymbol{\Delta}}_j^{c,\mathrm{tp}}$ and then integrate them. We first note that, due to the constraints (2.5),

$$\int_0^1 \big(\hat{p}_j(x_j), \hat{p}_{1,j}(x_j)\big)\,\mathbf{U}_j \cdot \hat{\mathbf{m}}_{W,j}^{c,\mathrm{tp}}(x_j)\,dx_j = 0. \tag{C.17}$$

Next, for the targets $\mathbf{m}_{W,j}^{c,\mathrm{tp}}$ we use $\mathrm{E}(m_{W,j}^c(X_j)) = 0$ to obtain

$$\begin{aligned}
&\int_0^1 \big(\hat{p}_j(x_j), \hat{p}_{1,j}(x_j)\big)\,\mathbf{U}_j \cdot \mathbf{m}_{W,j}^{c,\mathrm{tp}}(x_j)\,dx_j \\
&\quad = n^{-1}\sum_{i=1}^n \int_0^1 \big(m_{W,j}^c(x_j) + (X_{ij} - x_j)m'_{W,j}(x_j) - m_{W,j}^c(X_{ij})\big)\,K_h(x_j, X_{ij})\,dx_j \\
&\qquad + O_p\big(n^{-1/2}\big) \\
&\quad = -\int_{[0,1]^2} \big(m_{W,j}^c(v) - m_{W,j}^c(x_j) - (v - x_j)m'_{W,j}(x_j)\big)\,K_h(x_j, v)p_j(v)\,dv\,dx_j \\
&\qquad + O_p\big(n^{-1/2}\big) \\
&\quad = -\frac{h^2}{2}\int u^2K(u)\,du \cdot \int_0^1 p_j(x_j)m''_{W,j}(x_j)\,dx_j + o_p\big(n^{-2/5}\big).
\end{aligned} \tag{C.18}$$

It can be also shown that

$$\begin{aligned}
&\int_0^1 \big(\hat{p}_j(x_j), \hat{p}_{1,j}(x_j)\big)\,\mathbf{U}_j \cdot \mathbf{U}_j^\top \mathbf{N}(x_j)\boldsymbol{\gamma}(x_j) \cdot m''_{W,j}(x_j)\,dx_j \\
&\qquad = \int u^2K(u)\,du \cdot \int_0^1 p_j(x_j)m''_{W,j}(x_j)\,dx_j + o_p(1).
\end{aligned} \tag{C.19}$$

Furthermore, for the stochastic term $\mathbf{U}_j^\top \tilde{\mathbf{m}}_{W,j}^A$, the standard arguments in kernel smoothing give

$$\int_0^1 \big(\hat{p}_j(x_j), \hat{p}_{1,j}(x_j)\big)\,\mathbf{U}_j \cdot \mathbf{U}_j^\top \tilde{\mathbf{m}}_{W,j}^A(x_j)\,dx_j = O_p\big(n^{-1/2}\big). \tag{C.20}$$

The results (C.17)–(C.20) conclude (C.16).

C.6. Proof of Proposition 2

Recall the definition of the centered components $\hat{m}_{W,j}^c$ at (2.6). Also, note that $m_{W,j}^c(x_j) = m_{W,j}(x_j) - \mathrm{E}(m_{W,j}(X_j))$, where $(m_{W,j} : 1 \le j \le d)$ is any tuple that comprises $m_W = m_{W,1} + \cdots + m_{W,d}$. It holds that

$$m_W(\mathbf{x}) = \sum_{j=1}^d m_{W,j}^c(x_j) + \mathrm{E}(W),$$

$$\hat{m}_W(\mathbf{x}) = \sum_{j=1}^{d} \hat{m}^c_{W,j}(x_j) + \sum_{j=1}^{d} \int_0^1 \left(\hat{m}_{W,j}(u)\hat{p}_j(u) + \hat{m}_{W,1,j}(u)\hat{p}_{1,j}(u)\right) du.$$

The proposition follows from Proposition 3 if we prove

$$\sum_{j=1}^{d} \int_0^1 \left(\hat{m}_{W,j}(u)\hat{p}_j(u) + \hat{m}_{W,1,j}(u)\hat{p}_{1,j}(u)\right) du = \mathrm{E}(W) + o_p\left(n^{-2/5}\right). \tag{C.21}$$

To prove (C.21), we observe that the first row of $\hat{\mathbf{M}}_{jj}(x_j)$ equals $(\hat{p}_j(x_j), \hat{p}_{1,j}(x_j))$. Also, by the normalization property of the kernel $K_h(\cdot,\cdot)$, the first row of $\int_0^1 \hat{\mathbf{M}}_{jk}(x_j, x_k)\, dx_j$ reduces to $(\hat{p}_k(x_k), \hat{p}_{1,k}(x_k))$. Thus, by multiplying $\hat{\mathbf{M}}_{jj}(x_j)$ on both sides of the SBF equation (2.4), then integrating both sides with respect to x_j and comparing the first entries of the resulting quantities, we get that

$$\begin{aligned}
\sum_{j=1}^{d} \int_0^1 \left(\hat{m}_{W,j}(x_j)\hat{p}_j(x_j) + \hat{m}_{W,1,j}(x_j)\hat{p}_{1,j}(x_j)\right) dx_j &= n^{-1} \sum_{i=1}^{n} \int_0^1 K_h(x_j, X_{ij})\, dx_j \cdot W_i \\
&= n^{-1} \sum_{i=1}^{n} W_i \\
&= \mathrm{E}(W) + O_p\left(n^{-1/2}\right).
\end{aligned}$$

Here, we have also used the normalization property of the kernel that $\int_0^1 K_h(u,v)\, du = 1$ for all $v \in [0,1]$. This completes the proof of the proposition.

Acknowledgments

The authors would like to thank an associate editor and two referees for helpful and constructive comments.

Funding

The research of Young Kyung Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2021R1A2C1003920). Byeong U. Park's research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2019R1A2C3007355).

References

Aliprantis, C.D. and Border, K.C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd ed. Berlin: Springer. MR2378491

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Baltimore, MD: Johns Hopkins Univ. 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 https://doi.org/10.1177/003754977201900208

Boente, G. and Martínez, A. (2017). Marginal integration M-estimators for additive models. *TEST* **26** 231–260. MR3650526 https://doi.org/10.1007/s11749-016-0508-0

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555. MR0994249 https://doi.org/10.1214/aos/1176347115

Fan, J., Wu, Y. and Feng, Y. (2009). Local quasi-likelihood with a parametric guide. *Ann. Statist.* **37** 4153–4183. MR2572456 https://doi.org/10.1214/09-AOS713

Febrero-Bande, M. and González-Manteiga, W. (2013). Generalized additive models for functional data. *TEST* **22** 278–292. MR3062258 https://doi.org/10.1007/s11749-012-0308-0

Glad, I.K. (1998). Parametrically guided non-parametric regression. *Scand. J. Stat.* **25** 649–668. MR1666776 https://doi.org/10.1111/1467-9469.00127

Gozalo, P. and Linton, O. (2000). Local nonlinear least squares: Using parametric information in nonparametric regression. *J. Econometrics* **99** 63–106. MR1793389 https://doi.org/10.1016/S0304-4076(00)00031-2

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

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 https://doi.org/10.1080/01621459.2019.1604365

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

Hjort, N.L. and Glad, I.K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23** 882–904. MR1345205 https://doi.org/10.1214/aos/1176324627

Huang, L.-S. and Yu, C.-H. (2019). Classical backfitting for smooth-backfitting additive models. *J. Comput. Graph. Statist.* **28** 386–400. MR3974888 https://doi.org/10.1080/10618600.2018.1530120

Jeon, J.M. and Park, B.U. (2020). Additive regression with Hilbertian responses. *Ann. Statist.* **48** 2671–2697. MR4152117 https://doi.org/10.1214/19-AOS1902

Jeon, J.M., Park, B.U. and Van Keilegom, I. (2021a). Additive regression for non-Euclidean responses and predictors. *Ann. Statist.* **49** 2611–2641. MR4338377 https://doi.org/10.1214/21-aos2048

Jeon, J.M., Park, B.U. and Van Keilegom, I. (2021b). Additive regression for predictors of various natures and possibly incomplete Hilbertian responses. *Electron. J. Stat.* **15** 1473–1548. MR4255309 https://doi.org/10.1214/21-ejs1823

Jeon, J.M., Lee, Y.K., Mammen, E. and Park, B.U. (2022). Locally polynomial Hilbertian additive regression. *Bernoulli* **28** 2034–2066. MR4411521 https://doi.org/10.3150/21-bej1410

Lee, Y.K. (2004). On marginal integration method in nonparametric regression. *J. Korean Statist. Soc.* **33** 435–447. MR2126371

Lee, Y.K., Mammen, E. and Park, B.U. (2010). Backfitting and smooth backfitting for additive quantile models. *Ann. Statist.* **38** 2857–2883. MR2722458 https://doi.org/10.1214/10-AOS808

Lee, Y.K., Mammen, E. and Park, B.U. (2012). Flexible generalized varying coefficient regression models. *Ann. Statist.* **40** 1906–1933. MR3015048 https://doi.org/10.1214/12-AOS1026

Lee, Y.K., Mammen, E., Nielsen, J.P. and Park, B.U. (2020). Nonparametric regression with parametric help. *Electron. J. Stat.* **14** 3845–3868. MR4164866 https://doi.org/10.1214/20-EJS1760

Lee, Y.K., Park, B.U., Hong, H. and Kim, D. (2022). Estimation of Hilbertian varying coefficient models. *Stat. Interface* **15** 129–149. MR4363348

Linton, O. and Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93–100. MR1332841 https://doi.org/10.1093/biomet/82.1.93

Linton, O., Sperlich, S. and Van Keilegom, I. (2008). Estimation of a semiparametric transformation model. *Ann. Statist.* **36** 686–718. MR2396812 https://doi.org/10.1214/009053607000000848

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

Mammen, E. and Park, B.U. (2006). A simple smooth backfitting method for additive models. *Ann. Statist.* **34** 2252–2271. MR2291499 https://doi.org/10.1214/009053606000000696

Nielsen, J.P. and Sperlich, S. (2005). Smooth backfitting in practice. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 43–61. MR2136638 https://doi.org/10.1111/j.1467-9868.2005.00487.x

Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. *J. Multivariate Anal.* **73** 166–179. MR1763322 https://doi.org/10.1006/jmva.1999.1868

Opsomer, J.D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25** 186–211. MR1429922 https://doi.org/10.1214/aos/1034276626

Park, B.U., Mammen, E., Lee, Y.K. and Lee, E.R. (2015). Varying coefficient regression models: A review and new developments. *Int. Stat. Rev.* **83** 36–64. MR3341079 https://doi.org/10.1111/insr.12029

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

Sperlich, S., Linton, O. and Härdle, W. (1999). Integration and backfitting methods in additive models – Finite sample properties and comparison. *TEST* **8** 419–458.

Talamakrouni, M., El Ghouch, A. and Van Keilegom, I. (2015). Guided censored regression. *Scand. J. Stat.* **42** 214–233. MR3318033 https://doi.org/10.1111/sjos.12103

Talamakrouni, M., Van Keilegom, I. and El Ghouch, A. (2016). Parametrically guided nonparametric density and hazard estimation with censored data. *Comput. Statist. Data Anal.* **93** 308–323. MR3406214 https://doi.org/10.1016/j.csda.2015.01.009

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

Yang, S.J. and Park, B.U. (2014). Efficient estimation for partially linear varying coefficient models when coefficient functions have different smoothing variables. *J. Multivariate Anal.* **126** 100–113. MR3173084 https://doi.org/10.1016/j.jmva.2014.01.004

Yu, K., Mammen, E. and Park, B.U. (2011). Semi-parametric regression: Efficiency gains from modeling the nonparametric part. *Bernoulli* **17** 736–748. MR2787613 https://doi.org/10.3150/10-BEJ296

Yu, K., Park, B.U. and Mammen, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36** 228–260. MR2387970 https://doi.org/10.1214/009053607000000596

Zhang, X., Park, B.U. and Wang, J.-L. (2013). Time-varying additive models for longitudinal data. *J. Amer. Statist. Assoc.* **108** 983–998. MR3174678 https://doi.org/10.1080/01621459.2013.778776

Received March 2022 and revised November 2022