Proc. Int. Cong. of Math. – 2018
Rio de Janeiro, Vol. 4 (3013–3036)

NONPARAMETRIC ADDITIVE REGRESSION

Byeong U. Park

Abstract

In this article we discuss statistical methods of estimating structured nonparametric regression models. Our discussion is mainly on the additive models where the regression function (map) is expressed as a sum of unknown univariate functions (maps), but it also covers some other non- and semi-parametric models. We present the state of the art in the subject area with the prospect of an extension to non-Euclidean data objects.

1 Introduction

Let Y be a scalar random variable and $\mathbf{X} \equiv (X_1, \ldots, X_d)$ be a d-dimensional random vector. Suppose that one has observations $(\mathbf{X}_i, Y_i)$, $1 \leq i \leq n$, that are independent and identically distributed copies of $(\mathbf{X}, Y)$. The regression problem in statistics is to estimate the conditional mean $f(\mathbf{x}) \equiv \mathrm{E}(Y \mid \mathbf{X} = \mathbf{x})$ using the observations $(\mathbf{X}_i, Y_i)$. The parametric approach to this problem is to assume that the true regression function f belongs to a finite-dimensional model $\mathfrak{F}$. The simplest example of $\mathfrak{F}$ is a linear model $\mathfrak{F} = \{f(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^{d+1}\}$, where $f(\mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d$. This is certainly restrictive excluding many important realities. The nonparametric approach, on the contrary, is to allow the unknown f to lie in an infinite-dimensional function space. The problem is clearly 'ill-posed' since one is given only a finite number of observations $(\mathbf{X}_i, Y_i)$. One way, called method of sieves, is to reduce $\mathfrak{F}$ to a subspace $\mathfrak{F}_n$ in such a way that the sequence of sieve spaces $\mathfrak{F}_n$ grows as n increases and one searches for an estimator among functions in $\mathfrak{F}_n$. Another way of solving the ill-posed inverse problem is through penalization, putting more penalties for functions that are more complex to enforce smoothness

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2015R1A2A1A05001753).
MSC2010: primary 62G08; secondary 62G20.
Keywords: Additive models, smooth backfitting, varying coefficient models, partially linear models, errors-in variables, Hilbertian responses, Bochner integral.

for the resulting estimator. The approach termed as ‘kernel smoothing’ has quite a different nature and is based on localization. It basically converts the infinite-dimensional problem to solving locally finite-dimensional problems with the localization being finer for larger sample size n. In this paper we discuss nonparametric regression, focusing on kernel smoothing.

There is another problem of dimensionality. When the dimension d of $\mathbf{X}$ gets high, all nonparametric estimation techniques fail theoretically. For instance, if $\mathcal{F}$ is a class of functions with two continuous (partial) derivatives, then one cannot get an estimator $\hat{f}$ that has a rate faster than $n^{-2/(d+4)}$ for $\|\hat{f} - f\|_2$. Nonparametric methods fail practically as well when d is high. In the case of local kernel smoothing one basically takes $\mathcal{X}_h(\mathbf{x}) \equiv \{(\mathbf{X}_i, Y_i) : \mathbf{X}_i \text{ are within distance } h \text{ from } \mathbf{x}\}$ for each $\mathbf{x}$, where $h > 0$ is termed as ‘window width’ or ‘bandwidth’, and then estimate $f(\mathbf{x})$ using those $(\mathbf{X}_i, Y_i) \in \mathcal{X}_h(\mathbf{x})$. The practical difficulty one encounters here is that one cannot choose h small enough for a fine local approximation of f since the number of $(\mathbf{X}_i, Y_i)$ in $\mathcal{X}_h(\mathbf{x})$, which is asymptotic to nh^d, gets smaller very fast as h decreases when d is high. Note that one needs $nh^d \geq \ell$ for the corresponding locally ℓ-dimensional problem to be well-posed. This phenomenon, referred to as ‘the curse of dimensionality’, is present in other nonparametric methods such as sieves and penalization techniques.

Structured nonparametric models have been studied to circumvent the curse of dimensionality. A structured nonparametric model is defined as a known function of lower-dimensional unknown underlying functions, see Mammen and J. P. Nielsen [2003] for discussion on generalized structured models. They typically allow reliable estimation when a full nonparametric model does not work. The simplest example is the additive model

$$\text{(1-1)} \qquad \mathrm{E}(Y \mid \mathbf{X} = \mathbf{x}) = f_1(x_1) + \cdots + f_d(x_d),$$

where f_j are unknown univariate smooth functions. This model was first introduced by Friedman and Stuetzle [1981]. Various nonparametric regression problems reduce to the estimation of this model. Examples include nonparametric regression with time series errors or with repeated measurements, panels with individual effects and semiparametric GARCH models, see Mammen, Park, and Schienle [2014].

Three main techniques of fitting the model (1-1) are ordinary backfitting (Buja, Hastie, and Tibshirani [1989]), marginal integration (Linton and J. P. Nielsen [1995]) and smooth backfitting (SBF, Mammen, Linton, and J. Nielsen [1999]). A difficulty with the ordinary backfitting technique is that the estimator of (1-1) is defined only when the backfitting iteration converges, as its limit. It is known that the backfitting iteration converges under rather strong conditions on the joint distribution of the covariates, see Opsomer and Ruppert [1997] and Opsomer [2000]. For marginal integration, the main drawback is that it does not resolve the dimensionality issue since it requires consistent estimation of the full-dimensional density of $\mathbf{X}$, see Y. K. Lee [2004]. Smooth backfitting, on the other hand, is not subject to these difficulties. The method gives a well-defined estimator of the model and the iterative algorithm converges always under weak conditions. Furthermore, it has been shown for many structured nonparametric models that smooth backfitting estimators have univariate rates of convergence regardless of the dimension d.

In this paper, we revisit the theory of smooth backfitting for the additive regression model (1-1). We discuss some important extensions that include varying coefficient models, the case of errors-in-variables, some structured models for functional response and/or predictors and a general framework with Hilbertian response. Our discussion is primarily on the i.i.d. case where $(\mathbf{X}_i, Y_i)$ are independent across $1 \leq i \leq n$ and identically distributed, and for Nadaraya-Watson (locally constant) kernel smoothing since the theory is best understood under this setting.

2 Additive regression models

Let the distributions of X_j have densities p_j with respect to the Lebesgue measure on $\mathbb{R}$, and $\mathbf{X}$ have a joint density p with respect to the Lebesgue measure on $\mathbb{R}^d$. We assume that p_j are commonly supported on the unit interval $[0,1]$, for simplicity. In the original theory of Mammen, Linton, and J. Nielsen [1999], it is assumed that the joint density p is bounded away from zero on $[0,1]^d$. Here, we relax this condition to requiring only that each marginal density p_j is bounded away from zero on $[0,1]$.

2.1 SBF estimation. Let p_{jk} denote the two-dimensional joint densities of (X_j, X_k) for $1 \leq j \neq k \leq d$. From the model (1-1) we get a system of d integral equations,

$$(2\text{-}1) \qquad f_j(x_j) = \mathrm{E}(Y|X_j = x_j) - \sum_{k \neq j}^{d} \int_0^1 f_k(x_k) \frac{p_{jk}(x_j, x_k)}{p_j(x_j)}\, dx_k, \quad 1 \leq j \leq d.$$

The smooth backfitting method is nothing else than to replace the unknown marginal regression functions $m_j \equiv \mathrm{E}(Y|X_j = \cdot)$ and the marginal and joint densities p_j and p_{jk} by suitable estimators, and then to solve the resulting system of estimated integral equations. It is worthwhile to note here that the system of equations (2-1) only identifies $f_+(\mathbf{x}) \equiv \sum_{j=1}^d f_j(x_j)$, not the individual component functions f_j. We discuss the estimation of f_j later in Section 2.3.

For simplicity, we consider Nadaraya-Watson type estimators of m_j, p_j and p_{jk}. For a projection interpretation of SBF estimation, we use a normalized kernel scheme as described below. The projection interpretation is crucial for the success of SBF estimation.

Let K be a baseline symmetric, bounded and nonnegative kernel function supported on $[-1,1]$ such that $\int K = 1$. The conventional kernel weight scheme for the variable X_j based on K is to give the weight $K_{h_j}(x-u) \equiv h_j^{-1} K((x-u)/h_j)$ to an observed value u of X_j locally at each point $x_j \in [0,1]$, where $h_j > 0$ is called the bandwidth and determines the degree of localization for X_j. The normalized kernel function based on K is defined by

$$K_{h_j}(x_j,u) = \left[\int_0^1 K_{h_j}(v-u)\,dv\right]^{-1} K_{h_j}(x_j-u), \quad 0 \le x_j, u \le 1. \tag{2-2}$$

Then, it holds that $K_{h_j}(x_j,u) = K_{h_j}(x_j-u)$ for all $(x_j,u) \in [2h_j, 1-2h_j] \times [0,1]$ or $(x_j,u) \in [0,1] \times [h_j, 1-h_j]$. Furthermore,

$$\begin{aligned} &\int_0^1 K_{h_j}(x_j,u)\,dx_j = 1, \qquad \text{for all } u \in [0,1], \\ &\mu_{j,\ell}(x_j) = \int_{-1}^1 t^\ell K(t)\,dt, \qquad \text{for all } x_j \in [2h_j, 1-2h_j], \\ &|\mu_{j,\ell}(x_j)| \le 2\int_{-1}^1 |t|^\ell K(t)\,dt, \qquad \text{for all } x_j \in [0,1], \end{aligned} \tag{2-3}$$

where and below $\mu_{j,\ell}(x_j) = \int_0^1 h_j^{-\ell}(x_j-u)^\ell K_{h_j}(x_j,u)\,du$. We set $I_j = [2h_j, 1-2h_j]$ and refer to them as interior regions.

We write X_{ij} for the jth entry of $\mathbf{X}_i$. With the normalized kernel function $K_{h_j}(\cdot,\cdot)$ we estimate the marginal and joint densities by

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

Also, by Nadaraya-Watson smoothing we estimate m_j by

$$\hat{m}_j(x_j) = \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n K_{h_j}(x_j,X_{ij})Y_i.$$

Plugging these estimators into (2-1) gives the following system of backfitting equations to solve for $\hat{f} : \hat{f}(\mathbf{x}) \equiv \sum_{j=1}^d \hat{f}_j(x_j)$.

$$\hat{f}_j(x_j) = \hat{m}_j(x_j) - \sum_{k\neq j}^d \int_0^1 \hat{f}_k(x_k)\frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)}\,dx_k, \quad 1 \le j \le d. \tag{2-4}$$

We call it *smooth backfitting equation*. The system of equations (2-4) can identify only the sum function $\hat{f}(\mathbf{x}) = \sum_{j=1}^d \hat{f}_j(x_j)$ as we discuss in Section 2.2.

Define $\hat{p}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n}\prod_{j=1}^{d} K_{h_j}(x_j, X_{ij})$, the estimator of the joint density p. Let $\mathcal{H}(\hat{p})$ denote the space of additive functions $g \in L_2(\hat{p})$ of the form $g(\mathbf{x}) = g_1(x_1) + \cdots + g_d(x_d)$ where g_j are univariate functions. Endowed with the inner product $\langle g, \eta\rangle_n = \int g(\mathbf{x})\eta(\mathbf{x})\hat{p}(\mathbf{x})\,d\mathbf{x}$, it is a Hilbert space. By considering the Fréchet differentials of functionals defined on $\mathcal{H}(\hat{p})$ and from the first property of (2-3), we may show that

$$\hat{f} = \underset{g\in\mathcal{H}(\hat{p})}{\arg\min} \int_{[0,1]^d} n^{-1}\sum_{i=1}^{n}(Y_i - g(\mathbf{x}))^2 \prod_{j=1}^{d} K_{h_j}(x_j, X_{ij})\,d\mathbf{x} \tag{2-5}$$

whenever a solution $\hat{f}$ of (2-4) exists and is unique. To solve the system of equations (2-4) the following iterative scheme is employed. First, initialize $\hat{f}_j^{[0]}$ for $1 \le j \le d$. In the rth cycle of the iteration, update $\hat{f}_j^{[r-1]}$ successively for $1 \le j \le d$ by

$$\begin{aligned} \hat{f}_j^{[r]}(x_j) = \hat{m}_j(x_j) &- \sum_{1\le k\le j-1}\int_0^1 \hat{f}_k^{[r]}(x_k)\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\,dx_k \\ &- \sum_{j+1\le k\le d}\int_0^1 \hat{f}_k^{[r-1]}(x_k)\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\,dx_k. \end{aligned} \tag{2-6}$$

2.2 Convergence of SBF algorithm.

Here, we discuss the existence and uniqueness of the solution of the backfitting Equation (2-4), and also the convergence of the backfitting Equation (2-6).

Consider the subspaces of $L_2(\hat{p})$ defined by

$$L_2(\hat{p}_j) = \{g \in L_2(\hat{p}) : g(\mathbf{x}) = g_j(x_j) \text{ for some univariate function } g_j\}.$$

Let $\hat{\pi}_j : L_2(\hat{p}) \to L_2(\hat{p}_j)$ denote projection operators such that

$$\hat{\pi}_j(g) = \int_{[0,1]^{d-1}} g(\mathbf{x})\frac{\hat{p}(\mathbf{x})}{\hat{p}_j(x_j)}\,d\mathbf{x}_{-j}, \tag{2-7}$$

where $\mathbf{x}_{-j}$ for $\mathbf{x}$ equals $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)$. Then, the system of equations (2-4) can be written as

$$\hat{f} = (I - \hat{\pi}_j)\hat{f} + \hat{m}_j, \quad 1 \le j \le d, \tag{2-8}$$

where we have used the convention that $\hat{m}_j(\mathbf{x}) = \hat{m}_j(x_j)$. The equivalence between (2-4) and (2-8) follows from

$$(\hat{\pi}_j f_k)(\mathbf{x}) = \int_0^1 f_k(x_k)\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\,dx_k, \quad 1 \le j \ne k \le d,$$

which holds due to the first property of (2-3). Put

$$\hat{m}_\oplus = \hat{m}_d + (I - \hat{\pi}_d)\hat{m}_{d-1} + (I - \hat{\pi}_d)(I - \hat{\pi}_{d-1})\hat{m}_{d-2} + \cdots + (I - \hat{\pi}_d)\cdots(I - \hat{\pi}_2)\hat{m}_1$$

and $\hat{T} = (I - \hat{\pi}_d)\cdots(I - \hat{\pi}_1)$. Note that $\hat{T}$ is a linear operator that maps $\mathcal{H}(\hat{p})$ to itself. A successive application of (2-8) for $j = d, d-1, \dots, 2, 1$ gives

$$\hat{f} = \hat{T}\hat{f} + \hat{m}_\oplus. \tag{2-9}$$

If (2-9) has a solution $\hat{f} \in \mathcal{H}(\hat{p})$, then solving (2-9) is equivalent to solving (2-8) and thus $\hat{f}$ is also a solution of (2-8). To see this, consider a version of $\hat{T}$ for which the index j takes the role of the index d. Call it $\hat{T}_j$. Define a version of $\hat{m}_\oplus$ accordingly and call it $\hat{m}_{\oplus,j}$. Then, it holds that $\hat{\pi}_j\hat{T}_j = 0$ and $\hat{\pi}_j\hat{m}_{\oplus,j} = \hat{m}_j$. Suppose that there exists $\hat{f} \in \mathcal{H}(\hat{p})$ that satisfies (2-9). If we exchange the roles of j and d, then the solution also satisfies $\hat{f} = \hat{T}_j\hat{f} + \hat{m}_{\oplus,j}$. Since this holds for all $1 \le j \le d$, we may conclude

$$\hat{\pi}_j\hat{f} = \hat{\pi}_j\hat{T}_j\hat{f} + \hat{\pi}_j\hat{m}_{\oplus,j} = 0 + \hat{m}_j, \quad 1 \le j \le d,$$

which is equivalent to (2-8).

The existence and uniqueness of the solution of (2-9) now follows if the linear operator $\hat{T}$ is a contraction. An application of Proposition A.4.2 of Bickel, Klaassen, Ritov, and Wellner [1993] to the projection operators $\hat{\pi}_j$ gives that $\mathcal{H}(\hat{p})$ is a closed subspace of $L_2(\hat{p})$ and $\|\hat{T}\|_{\text{op}} < 1$, under the condition that

$$\int_{[0,1]^2}\left[\frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)\hat{p}_k(x_k)}\right]^2 \hat{p}_j(x_j)\hat{p}_k(x_k)\,dx_j\,dx_k < \infty \quad \text{for all } 1 \le j \ne k \le d. \tag{2-10}$$

An analogue of (2-9) for the backfitting Equation (2-6) is

$$\hat{f}^{[r]} = \hat{T}\hat{f}^{[r-1]} + \hat{m}_\oplus. \tag{2-11}$$

Assuming (2-10), we get from (2-9) that $\hat{f} = \sum_{j=1}^{\infty}\hat{T}^j\hat{m}_\oplus$. This and the fact that $\hat{T}\hat{f}^{[r-1]} + \hat{m}_\oplus = \hat{T}^r\hat{f}^{[0]} + \sum_{j=0}^{r-1}\hat{T}^j\hat{m}_\oplus$ give

$$\|\hat{f}^{[r]} - \hat{f}\|_{2,n} \le \|\hat{T}\|_{\text{op}}^r\left(\|\hat{f}^{[0]}\|_{2,n} + \frac{1}{1 - \|\hat{T}\|_{\text{op}}}\cdot\|\hat{m}_\oplus\|_{2,n}\right), \tag{2-12}$$

where $\|\cdot\|_{2,n}$ denote the induced norm of the inner product $\langle\cdot,\cdot\rangle_n$ defined earlier. The following theorem is a non-asymptotic version of Theorem 1 of Mammen, Linton, and J. Nielsen [1999].

THEOREM 2.1. *Assume the condition (2-10). Then, it holds that the solution of the system of equations (2-4) exists and is unique, and that the backfitting iteration (2-6) converges to the solution.*

The condition (2-10) holds with probability tending to one if p_j are continuous and bounded away from zero on $[0,1]$ and p_{jk} are continuous and bounded above on $[0,1]^2$. This follows since under these conditions there exists a constant $0 < C < \infty$ such that

$$\sup_{(x_j,x_k)\in[0,1]^2} \frac{\hat{p}_{jk}(x_j,x_k)^2}{\hat{p}_j(x_j)\hat{p}_k(x_k)} \le C$$

with probability tending to one. Thus, we can deduce that

$$P\left(\lim_{r\to\infty} \|\hat{f}^{[r]} - \hat{f}\|_{2,n} = 0\right) \to 1 \tag{2-13}$$

as $n \to \infty$. Below, we give a stronger result than (2-13) owing to Mammen, Linton, and J. Nielsen [ibid.]. We make the following assumptions to be used in the subsequent discussion.

(C1) The joint densities p_{jk} are partially continuously differentiable and p is bounded away from zero and infinity on $[0,1]^d$.

(C2) The bandwidths satisfy $h_j \to 0$ and $nh_jh_k/\log n \to \infty$ as $n \to \infty$ for all $1 \le j \ne k \le d$.

(C3) The baseline kernel function K is bounded, has compact support $[-1,1]$, is symmetric about zero and Lipschitz continuous.

Define an analogue of $\mathcal{H}(\hat{p})$ as

$$\mathcal{H}(p) \equiv \{g \in L_2(p) : g(\mathbf{x}) = g_1(x_1) + \cdots + g_d(x_d),\ g_j \text{ are univariate functions}\}$$

equipped with the inner product $\langle g, \eta\rangle = \int g(\mathbf{x})\eta(\mathbf{x})p(\mathbf{x})\,d\mathbf{x}$ and its induced norm $\|\cdot\|_2$. We note that $P\big(\mathcal{H}(\hat{p}) = \mathcal{H}(p)\big) \to 1$ under the condition (C1)–(C3). This follows since the conditions imply that there exist absolute constants $0 < c < C < \infty$ such that

$$c\|g_j\|_2 \le \|g_j\|_{2,n} \le C\|g_j\|_2$$

with probability tending to one. Now, define π_j as $\hat{\pi}_j$ with $\hat{p}$ and $\hat{p}_j$ being replaced by p and p_j, respectively. Let $T = (I - \pi_d)\cdots(I - \pi_1)$. From (C1) we get that, for all $1 \le j \ne k \le d$,

$$\int_{[0,1]^2} \left[\frac{p_{jk}(x_j,x_k)}{p_j(x_j)p_k(x_k)}\right]^2 p_j(x_j)p_k(x_k)\,dx_j\,dx_k < \infty,$$

so that T is also a contraction as a map from $\mathcal{H}(p)$ to itself. Furthermore, another application of Proposition A.4.2 of Bickel, Klaassen, Ritov, and Wellner [1993] gives that there is an absolute constant $0 < c < \infty$ such that for any $g \in \mathcal{H}(p)$ there exists a decomposition $g = g_1 + \cdots + g_d$ with

$$\|g\|_2 \geq c \sum_{j=1}^{d} \|g_j\|_2. \tag{2-14}$$

For such a decomposition and from successive applications of the Minkowski and Hölder inequalities, we get

$$\|(\hat{\pi}_j - \pi_j)g\|_2 \leq$$
$$\leq \sum_{k \neq j}^{d} \|g_k\|_2 \left(\int \left[\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)p_k(x_k)} - \frac{p_{jk}(x_j, x_k)}{p_j(x_j)p_k(x_k)} \right]^2 p_j(x_j)p_k(x_k)\, dx_j\, dx_k \right)^{1/2}.$$

Using this and (2-14), we may prove $\|\hat{\pi}_j - \pi_j\|_{\mathrm{op}} = o_p(1)$ for all $1 \leq j \leq d$ and thus $\|\hat{T} - T\|_{\mathrm{op}} = o_p(1)$. This proves that there exists a constant $0 < \gamma < 1$ such that $P(\|\hat{T}\|_{\mathrm{op}} < \gamma) \to 1$ as $n \to \infty$. The following theorem is an asymptotic version of Theorem 2.1.

Theorem 2.2. (Mammen, Linton, and J. Nielsen [1999]). *Assume the conditions (C1)–(C3). Then, with probability tending to one, the solution of the system of equations (2-4) exists and is unique. Furthermore, there exists a constant $0 < \gamma < 1$ such that*

$$\lim_{n \to \infty} P\left(\|\hat{f}^{[r]} - \hat{f}\|_2 \leq \gamma^r (\|\hat{f}^{[0]}\|_2 + (1-\gamma)^{-1} \|\hat{m}_{\oplus}\|_2) \right) = 1.$$

2.3 Estimation of individual component functions.

The component functions f_j in the model (1-1) are not identified, but only their sum f is. We need put constraints on f_j to identify them. There may be various constraints. We consider the constraints

$$\int_0^1 f_j(x_j)p_j(x_j)\, dx_j = 0, \quad 1 \leq j \leq d. \tag{2-15}$$

With the constraints at (2-15) the model (1-1) is rewritten as

$$f(\mathbf{x}) = \mu + f_1(x_1) + \cdots + f_d(x_d), \tag{2-16}$$

for $\mu = \mathrm{E}(Y)$, and each f_j is uniquely determined. The latter follows from (2-14) and the fact that, for $c_j = \int_0^1 g_j(x_j)p_j(x_j)\, dx_j$, we get

$$\|g_j\|_2^2 = \|g_j - c_j\|_2^2 + |c_j|^2 \geq \|g_j - c_j\|_2^2.$$

For the estimators of f_j we consider the following constraint.

$$\int_0^1 \hat{f}_j(x_j)\hat{p}_j(x_j)\,dx_j = 0, \quad 1 \le j \le d. \tag{2-17}$$

For the estimation of f_j that satisfy (2-15), the backfitting Equation (2-4) and the backfitting Equation (2-6) are modified by simply putting $\hat{m}_j - \bar{Y}$ in the place of $\hat{m}_j$, where $\bar{Y}$ is used as an estimator of μ. Then, we may prove that, with probability tending to one, there exists a solution $(\hat{f}_j : 1 \le j \le d)$ of the resulting backfitting equation that satisfies the constraint (2-17). In this section, we discuss the asymptotic properties of the estimators $\hat{f}_j$. The error of $\bar{Y}$ as an estimator of μ is of magnitude $O_p(n^{-1/2})$, which is negligible compared to nonparametric rates. In the subsequent discussion in this section, we assume $\mu = 0$ and ignore $\bar{Y}$ in the backfitting equation, for simplicity.

Put $\varepsilon_i = Y_i - \sum_{j=1}^d f_j(X_{ij})$ and

$$\hat{m}_j^A(x_j) = \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n K_{h_j}(x_j, X_{ij})\varepsilon_i,$$

$$\hat{m}_j^B(x_j) = \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n K_{h_j}(x_j, X_{ij})\left[f_j(X_{ij}) - f_j(x_j)\right],$$

$$\hat{m}_{jk}^C(x_j) = n^{-1} \sum_{i=1}^n \int_0^1 \left[f_k(X_{ik}) - f_k(x_k)\right] K_{h_j}(x_j, X_{ij}) K_{h_k}(x_k, X_{ik})\,dx_k.$$

Then, from the backfitting Equation (2-4) we get

(2-18)
$$\hat{f}_j(x_j) - f_j(x_j) = \hat{m}_j^A(x_j) + \hat{m}_j^B(x_j) + \hat{p}_j(x_j)^{-1} \sum_{k \ne j} \hat{m}_{jk}^C(x_j) - \sum_{k \ne j} \int_0^1 \left[\hat{f}_k(x_k) - f_k(x_k)\right] \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\,dx_k, \quad 1 \le j \le d.$$

The above equation is a key to deriving stochastic expansions of $\hat{f}_j$. To analyze the three terms $\hat{m}_j^A$, $\hat{m}_j^B$ and $\hat{m}_{jk}^C$ in (2-18), we make the following assumptions.

(C4) The component functions f_j are twice continuously differentiable.

(C5) $\mathrm{E}|Y|^\alpha < \infty$ for $\alpha > 5/2$ and $\mathrm{var}(Y|X_j = \cdot)$ are continuous on $[0, 1]$

We also assume that h_j are of order $n^{-1/5}$, which is known to be optimal in univariate smoothing.

We write $\mu_\ell = \int_{-1}^{1} t^\ell K(t)\,dt$ and recall the definition of $\mu_{j,\ell}$ given immediately after (2-3). Let r_j denote a generic sequence of stochastic terms corresponding to $\hat{m}_j$ such that

$$\sup_{x_j\in[2h_j,1-2h_j]} |r_j(x_j)| = o_p(n^{-2/5}), \qquad \sup_{x_j\in[0,1]} |r_j(x_j)| = O_p(n^{-2/5}). \tag{2-19}$$

Using the conditions (C1), (C3) and (C4) we may verify that, for $1 \le j \ne k \le d$,

(2-20)

$$\begin{aligned}
\hat{m}_j^B(x_j) &= h_j \frac{\mu_{j,1}(x_j)}{\mu_{j,0}(x_j)} f_j'(x_j) + h_j^2 \mu_2 f_j'(x_j) \frac{p_j'(x_j)}{p_j(x_j)} + \frac{1}{2} h_j^2 \mu_2 f_j''(x_j) + r_j(x_j),\\
\frac{\hat{m}_{jk}^C(x_j)}{\hat{p}_j(x_j)} &= h_k^2 \mu_2 \int_0^1 f_k'(x_k) \frac{\partial p_{jk}(x_j,x_k)/\partial x_k}{p_j(x_j)}\,dx_k\\
&\quad + \int_0^1 \left[h_k \frac{\mu_{k,1}(x_k)}{\mu_{k,0}(x_k)} f_k'(x_k) + \frac{1}{2} h_k^2 \mu_2 f_k''(x_k) \right] \frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)}\,dx_k + r_j(x_j).
\end{aligned}$$

Define

(2-21)

$$\begin{aligned}
\tilde{\Delta}_j(x_j) &= h_j^2 \mu_2 f_j'(x_j) \frac{p_j'(x_j)}{p_j(x_j)} + \sum_{k\ne j} h_k^2 \mu_2 \int_0^1 f_k'(x_k) \frac{\partial p_{jk}(x_j,x_k)/\partial x_k}{p_j(x_j)}\,dx_k,\\
\hat{\Delta}_j(x_j) &= \hat{f}_j(x_j) - f_j(x_j) - \hat{m}_j^A(x_j) - h_j \frac{\mu_{j,1}(x_j)}{\mu_{j,0}(x_j)} f_j'(x_j) - \frac{1}{2} h_j^2 \mu_2 f_j''(x_j).
\end{aligned}$$

Then, the equations at (2-18) and the expansions at (2-20) give

$$\hat{\Delta}_j(x_j) = \tilde{\Delta}_j(x_j) - \sum_{k\ne j} \int \hat{\Delta}_k(x_k) \frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)}\,dx_k + r_j(x_j), \tag{2-22}$$

where we have used

$$\int \hat{m}_k^A(x_k) \frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)}\,dx_k = o_p(n^{-2/5})$$

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

Now, we consider a system of equations for $\hat{D} \in \mathcal{H}(\hat{p})$,

$$\hat{D}_j(x_j) = \tilde{\Delta}_j(x_j) - \sum_{k\ne j} \int \hat{D}_k(x_k) \frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)}\,dx_k, \quad 1 \le j \le d. \tag{2-23}$$

Arguing as in Section 2.2, solving this is equivalent to solving $\hat{D} = \hat{T}\hat{D} + \tilde{\Delta}_\oplus$, where $\tilde{\Delta}_\oplus$ is defined as $\hat{m}_\oplus$ with $\tilde{\Delta}_j$ taking the roles of $\hat{m}_j$. Similarly, solving (2-22) is equivalent to solving for $\hat{\Delta} \in \mathcal{H}(\hat{p})$ such that $\hat{\Delta} = \hat{T}\hat{\Delta} + \tilde{\Delta}_\oplus + r_\oplus$ with $r_\oplus$ being defined accordingly. Then, under the condition (2-10) or with probability tending to one under the condition (C1) it holds that $\hat{\Delta} = \hat{D} + \sum_{r=0}^\infty \hat{T}^r r_\oplus$. Since $\hat{\pi}_j r_k = o_p(n^{-2/5})$ uniformly over $[0,1]$, we get $r_\oplus = r_+$ with the generic r_+ such that $r_+(\mathbf{x}) = \sum_{j=1}^d r_j(x_j)$ for some r_j that satisfy (2-19). Also, from the observation that $(I-\hat{\pi}_j)\cdots(I-\hat{\pi}_1)r_j = o_p(n^{-2/5})$ uniformly over $[0,1]$ for all $1 \le j \le d$, we have

$$\sum_{r=0}^\infty \hat{T}^r r_\oplus = r_\oplus + \sum_{r=1}^\infty \hat{T}^r r_\oplus = r_+.$$

This proves

$$\hat{\Delta} = \hat{D} + r_+. \tag{2-24}$$

To identify the limit of $\hat{D}$ we consider the system of integral equations for $\Delta \in \mathcal{H}(p)$,

$$\Delta_j(x_j) = \tilde{\Delta}_j(x_j) - \sum_{k \neq j} \int \Delta_k(x_k) \frac{p_{jk}(x_j, x_k)}{p_j(x_j)}\, dx_k, \quad 1 \le j \le d. \tag{2-25}$$

Again, arguing as in Section 2.2, solving (2-25) is equivalent to solving $\Delta = T\Delta + \Delta_\oplus$, where $\Delta_\oplus$ is defined as $\tilde{\Delta}_\oplus$ but with $\hat{\pi}_j$ being replaced by π_j. Since $\|T\|_{\mathrm{op}} < 1$ under (C1), the latter equation has a unique solution $\Delta = \sum_{r=0}^\infty T^r \Delta_\oplus$. A careful analysis of the operators T and $\hat{T}$ gives that $(\hat{T} - T)\sum_{r=1}^\infty \hat{T}^{r-1}\tilde{\Delta}_\oplus = r_+$ and that

$$T \sum_{r=2}^\infty \sum_{j=0}^{r-2} T^j (\hat{T} - T)\hat{T}^{r-2-j}\tilde{\Delta}_\oplus = o_p(n^{-2/5}),$$

$$\sum_{r=1}^\infty T^r (\tilde{\Delta}_\oplus - \Delta_\oplus) = o_p(n^{-2/5})$$

uniformly over $[0,1]^d$. From these calculations it follows that $\hat{D} = \Delta + r_+$. This with (2-24) entails

$$\hat{\Delta} = \Delta + r_+. \tag{2-26}$$

To get expansions for each component $\hat{f}_j$ satisfying the constraint (2-17), we put the following constraints on Δ_j.

$$\int \Delta_j(x_j) p_j(x_j)\, dx_j = \mu_2 h_j^2 \int f_j'(x_j) p_j'(x_j)\, dx_j, \quad 1 \le j \le d. \tag{2-27}$$

Then, using (2-17) and (2-26) with the definition of $\hat{\Delta}_j$ at (2-21), we may prove $\hat{\Delta}_j = \Delta_j + r_j$ for $1 \le j \le d$, establishing the following theorem.

THEOREM 2.3. *Assume that the conditions (C1)–(C5) and that the bandwidths h_j are asymptotic to $n^{-1/5}$. Then,*

$$\hat{f}_j(x_j) = f_j(x_j) + \hat{m}_j^A(x_j) + h_j \frac{\mu_{j,1}(x_j)}{\mu_{j,0}(x_j)} f_j'(x_j) + \frac{1}{2} h_j^2 \mu_2 f_j''(x_j) + \Delta_j(x_j) + r_j(x_j),$$

where r_j satisfy (2-19).

For fixed $x_j \in (0,1)$, all $\mu_{j,1}(x_j) = 0$ for sufficiently large n, and $(nh_j)^{1/2}\hat{m}_j^A(x_j)$ are asymptotically normal with mean zero and variance

$$\mathrm{var}(Y|X_j = x_j) p_j(x_j)^{-1} \int K^2(u)\, du$$

. Thus, the asymptotic distributions of $(nh_j)^{1/2}(\hat{f}_j(x_j) - f_j(x_j))$ of $\hat{f}_j$ are readily obtained from the stochastic expansion in the above theorem.

Although we have not discussed here, Mammen, Linton, and J. Nielsen [1999] also developed a local linear version of the smooth backfitting technique. However, the original proposal does not have easy interpretation as the Nadaraya-Watson estimator that we have discussed, and its implementation is more complex than the latter. Mammen and Park [2006] suggested a new smooth backfitting estimator that has the simple structure of the Nadaraya-Watson estimator while maintaining the nice asymptotic properties of the local linear smooth backfitting estimator.

2.4 Bandwidth selection and related models. In nonparametric function estimation, selection of smoothing parameters is essential for the accuracy of the estimation. It is well known that one should not choose these tuning parameters by minimizing a measure of fit, such as the residual sum of squares $n^{-1}\sum_{i=1}^n (Y_i - \hat{f}(\mathbf{X}_i))^2$, since it tends to choose h_j that give 'overfitting'. Mammen and Park [2005] tackled this problem by deriving higher-order stochastic expansions of the residual sum of squares and proposed a penalized least squares method of choosing h_j. They also proposed two plug-in bandwidth selectors that rely on expansions of the average square errors $n^{-1}\sum_{i=1}^n (\hat{f}(\mathbf{X}_i) - f(\mathbf{X}_i))^2$. J. P. Nielsen and Sperlich [2005] considered a cross-validated bandwidth selector and discussed some other practical aspects of the smooth backfitting algorithm.

A very important extension of the additive mean regression model at (1-1) or (2-16) is to a generalized additive model,

$$g(\mathrm{E}(Y|\mathbf{X} = \mathbf{x})) = f_1(x_1) + \cdots + f_d(x_d), \tag{2-28}$$

where g is a known link function. This model accommodates discrete-type responses Y such as Bernoulli and Poisson random variables. Yu, Park, and Mammen [2008] extended the idea of smooth backfitting to generalized additive models. The estimation of the additive function $f = f_1 + \cdots + f_d$ based on observations of $(\mathbf{X}, Y)$ involves a nonlinear optimization problem due to the presence of the link g. To resolve the difficulty, Yu, Park, and Mammen [ibid.] introduced the so called 'smoothed likelihood' and studied an innovative idea of double iteration to maximize the smoothed likelihood. They proved that the double iteration algorithm converges and developed a complete theory for the smooth backfitting likelihood estimators of f_j.

Varying coefficient models are another important class of structured nonparametric regression models. The models arise in many real applications, see Hastie and Tibshirani [1993], Yang, Park, Xue, and Härdle [2006] and Park, Mammen, Y. K. Lee, and E. R. Lee [2015]. Their structure is similar to classical linear models, but they are more flexible since the regression coefficients are allowed to be functions of other predictors. There are two types of varying coefficient models that have been studied most. One type is to let all regression coefficients depend on a single predictor, say Z: $\mathrm{E}(Y|\mathbf{X} = \mathbf{x}, Z = z) = f_1(z)x_1 + \cdots + f_d(z)x_d$. The estimation of this type of models is straightforward. For each given z, we may estimate $\mathbf{f}(z) \equiv (f_1(z), \ldots, f_d(z))$ by

$$\mathbf{f}(z) = \underset{(\theta_1,\ldots,\theta_d)\in\mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{d} \theta_j X_{ij} \right)^2 K_h(z, Z_i).$$

There have been a large body of literature on this model, see Fan and W. Zhang [1999] and Fan and W. Zhang [2000], for example. The second type is to let different regression coefficients be functions of different predictors, say $\mathbf{Z} \equiv (Z_1, \ldots, Z_d)$:

$$\mathrm{E}(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = f_1(z_1)x_1 + \cdots + f_d(z_d)x_d. \tag{2-29}$$

Fitting the model (2-29) is completely different from fitting the first type. The standard kernel smoothing that minimizes

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{d} \theta_j X_{ij} \right)^2 \prod_{j=1}^{d} K_{h_j}(z_j, Z_{ij})$$

for each $\mathbf{z}$ would give multivariate function estimators of $f_j(z_j)$ that also depend on other values of predictors z_k for $k \neq j$. Yang, Park, Xue, and Härdle [2006] studied the estimation of the latter model based on the marginal integration technique. Later, Y. K. Lee, Mammen, and Park [2012b] extended the idea of smooth backfitting to estimating the model.

Two limitations in the application of the model (2-29) are that the number of predictors X_j should be the same as that of Z_j and that in a modeling stage it is rather difficult to determine which predictors we choose to be the ‘smoothing variables’ Z_j and which to be ‘regressors’ X_j. Y. K. Lee, Mammen, and Park [2012a] removed the limitations completely by studying a very general form of varying coefficient models. With a link function g and a given set of d predictors, they introduced the model

$$g(\mathrm{E}(Y|\mathbf{X}=\mathbf{x})) = x_1 \left(\sum_{k \in I_1} f_{1k}(x_k) \right) + \cdots + x_q \left(\sum_{k \in I_q} f_{qk}(x_k) \right), \tag{2-30}$$

where $q \leq d$ and the index sets I_j are known subsets of $\{1, \ldots, d\}$ and allowed to overlap with each other, but not to include j. If each I_j consists of a single index different from each other, then (2-30) reduces to the model (2-29), while taking $X_1 \equiv 1$, $q = 1$ and $I_1 = \{2, \ldots, d\}$ gives the generalized additive model (2-28). Y. K. Lee, Mammen, and Park [ibid.] proved that the component functions f_{jk} are identifiable under weak conditions, developed a powerful technique of fitting the model and presented its theory.

Other related works include Y. K. Lee, Mammen, and Park [2010], Y. K. Lee, Mammen, and Park [2014], Yu, Mammen, and Park [2011] and Y. K. Lee [2017], to list a few. Among them, Y. K. Lee, Mammen, and Park [2010] considered the estimation of additive quantile models, $Y = f_1(X_1) + \cdots + f_d(X_d) + \varepsilon$, where ε satisfies $P(\varepsilon \leq 0|\mathbf{X}) = \alpha$ for $0 < \alpha < 1$. They successfully explored the theory for both the ordinary and smooth backfitting by devising a theoretical mean regression model under which the least squares ordinary and smooth backfitting estimators are asymptotically equivalent to the corresponding quantile estimators under the original model. Y. K. Lee, Mammen, and Park [2014] further extended the idea to the estimation of varying coefficient quantile models. Yu, Mammen, and Park [2011] considered a partially linear additive model. They derived the semiparametric efficiency bound in the estimation of the parametric part of the model and proposed a semiparametric efficient estimator based on smooth backfitting estimation of the additive nonparametric part. Finally, Y. K. Lee [2017] studied the estimation of bivariate additive regression models based on the idea of smooth backfitting.

3 Errors-in-variable additive models

In this section we consider the situation where the predictors X_j are not directly observed in the additive model (1-1), but contaminated $Z_j = X_j + U_j$ with measurement errors U_j are. Many people worked on errors-in-variables problems in nonparametric density and regression estimation. A few notable examples include Carroll and Hall [1988], Stefanski and Carroll [1990], Fan and Truong [1993], Delaigle, Hall, and Meister [2008], Delaigle, Fan, and Carroll [2009], Delaigle and Hall [2016] and Han and Park [2018]. Among them, Han and Park [ibid.] is considered as the first attempt dealing with errors-in-variables in structured nonparametric regression. In this section we outline the work of Han and Park [ibid.] on the model (1-1) and discuss its extensions.

3.1 Normalized deconvolution kernel. Suppose that we observe $Z_{ij} = X_{ij} + U_{ij}$ for $1 \le i \le n$ and $1 \le j \le d$, where we assume that $\mathbf{U}_i \equiv (U_{i1}, \ldots, U_{id})$ are independent of $\mathbf{X}_i$. We also assume that U_{ij} are independent and have known densities. Write $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{id})$. The task is to estimate the mean regression function $f = \mathrm{E}(Y|\mathbf{X} = \cdot)$ with the additive structure $f(\mathbf{x}) = f_1(x_1) + \cdots + f_d(x_d)$ using the contaminated data $(\mathbf{Z}_i, Y_i), 1 \le i \le n$. The very core of the difficulty is that the observed responses Y_i for $\mathbf{Z}_i$ near a point of interest, say $\mathbf{x}$, may not contain relevant information about the true function $f(\mathbf{x})$ because of the measurement errors $\mathbf{U}_i \equiv (U_{i1}, \ldots, U_{id})$. Thus, local smoothing of Y_i with a conventional kernel weighting scheme that acts on $\mathbf{Z}_i$ fails.

In the estimation of a density p_0 of a random variable X taking values in $\mathbb{R}$, one uses a special kernel scheme to effectively deconvolute irrelevant information contained in the contaminated $Z = X + U$. For a baseline kernel function $K \ge 0$, define

$$\tilde{K}_h(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itu} \frac{\phi_K(ht)}{\phi_U(t)}\, dt, \tag{3-1}$$

where ϕ_W for a random variable W denotes the characteristic function of W. This kernel has the so called ‘unbiased scoring’ property that

$$\mathrm{E}\left(\tilde{K}_h(x - Z)|X\right) = K_h(x - X). \tag{3-2}$$

The property (3-2) basically tells that the bias of the deconvolution kernel density estimator $\hat{p}_0(x) = n^{-1} \sum_{i=1}^n \tilde{K}_h(x - Z_i)$ is the same as the ‘oracle’ estimator $\hat{p}_{0,\mathrm{ora}}(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$ that is based on unobservable X_i and the conventional kernel scheme K_h.

In Section 2 we have seen that the first property of (2-3) is essential in the estimation of the additive model (1-1). One may think of normalizing the deconvolution kernel as defined at (3-1) as in (2-2) with $\tilde{K}_{h_j}(x_j - u)$ taking the role of $K_{h_j}(x_j - u)$. But, it turns out that the resulting kernel violates the corresponding version of the unbiased scoring property (3-2). Han and Park [ibid.] noted that

$$\tilde{K}_{h_j}(x_j - z) = \frac{1}{2\pi h_j} \int_{-\infty}^{\infty} e^{-it(x_j - z)/h_j} \frac{\varphi_K(t, x_j, h_j)}{\phi_{U_j}(t/h_j)}\, dt,$$

where $\varphi_K(t, x_j; h_j) = \int_0^1 e^{it(x_j - v)/h_j} K_{h_j}(x_j - v)\, dv$. The basic idea was then to replace $K_{h_j}(x_j - \cdot)$ in the definition of $\varphi_K(t, x_j; h_j)$ by the normalized kernel $K_{h_j}(x_j, \cdot)$ as defined at (2-2). The resulting kernel is not well-defined, however, for x_j on the boundary region $[0, h_j) \cup (1-h_j, 1]$. To remedy this, Han and Park [2018] proposed a new kernel scheme $\tilde{K}^{\star}_{h_j}$ defined by

$$\tilde{K}^{\star}_{h_j}(x_j, z) = \frac{1}{2\pi h_j} \int_{-\infty}^{\infty} e^{-it(x_j - z)/h_j} \frac{\phi_K(t, x_j; h_j)\phi_K(t)}{\phi_{U_j}(t/h_j)}\, dt, \tag{3-3}$$

where $\phi_K(t, x_j; h_j) = \int_0^1 e^{it(x_j - v)/h_j} K_{h_j}(x_j, v)\, dv$. Han and Park [ibid.] proved that $\tilde{K}^{\star}_{h_j}$ has both the properties of normalization and unbiased scoring under the following condition (A1). Let $\lfloor \gamma \rfloor$ denote the largest integer that is less than or equal to γ, and $K^{(\ell)}$ the ℓ-th derivative of K.

(A1) There exist constants $\beta \geq 0$ and $0 < c < C < \infty$ such that $c(1+|t|)^{-\beta} \leq |\phi_{U_j}(t)| \leq C(1+|t|)^{-\beta}$ for all $t \in \mathbb{R}$ and for all $1 \leq j \leq d$. For such constant β the baseline kernel K is $\lfloor \beta + 1 \rfloor$-times continuously differentiable and $K^{(\ell)}(-1) = K^{(\ell)}(1) = 0$ for $0 \leq \ell \leq \lfloor \beta \rfloor$.

THEOREM 3.1. (Han and Park [ibid.]). *Under the conditions (A1) and (C3), the integral in (3-3) exists for all $x_j \in [0,1]$ and $z \in \mathbb{R}$. Furthermore, $\int_0^1 \tilde{K}^{\star}_{h_j}(x_j, z)\, dx_j = 1$ for all $z \in \mathbb{R}$ and*

$$\mathrm{E}\left(\tilde{K}^{\star}_{h_j}(x_j, Z_j) \,\middle|\, X_j = u_j\right) = K_{h_j}(x_j, \cdot) * K_{h_j}(u_j) \quad \textit{for all } x_j, u_j \in [0,1].$$

3.2 Theory of smooth backfitting. With the normalized and smoothed deconvolution kernel $\tilde{K}^{\star}_{h_j}$ introduced in Section 3.1, we simply replace $\hat{p}_j$, $\hat{p}_{jk}$ and $\hat{m}_j$ in (2-4), respectively, by

$$\hat{p}^{\star}_j(x_j) = n^{-1} \sum_{i=1}^{n} \tilde{K}^{\star}_{h_j}(x_j, Z_{ij}),$$

$$\hat{p}^{\star}_{jk}(x_j, x_k) = n^{-1} \sum_{i=1}^{n} \tilde{K}^{\star}_{h_j}(x_j, Z_{ij}) \tilde{K}^{\star}_{h_k}(x_k, Z_{ik}),$$

$$\hat{m}^{\star}_j(x_j) = \hat{p}^{\star}_j(x_j)^{-1} n^{-1} \sum_{i=1}^{n} \tilde{K}^{\star}_{h_j}(x_j, Z_{ij}) Y_i.$$

Define $\hat{p}^{\star}(\mathbf{x}) = n^{-1} \sum_{i=1}^{n} \prod_{j=1}^{d} \tilde{K}^{\star}_{h_j}(x_j, X_{ij})$ and $\hat{\pi}^{\star}_j$ as $\hat{\pi}_j$ at (2-7) with $\hat{p}$ and $\hat{p}_j$ being replaced by $\hat{p}^{\star}$ and $\hat{p}^{\star}_j$, respectively. Let $\hat{T}^{\star} = (I - \hat{\pi}^{\star}_d) \cdots (I - \hat{\pi}^{\star}_1)$. We can express the resulting backfitting equation as equations

$$\hat{f}^{\star} = (I - \hat{\pi}^{\star}_j)\hat{f}^{\star} + \hat{m}^{\star}_j, \quad 1 \leq j \leq d. \tag{3-4}$$

As we argued in Section 2.2, solving this system of equations is equivalent to solving $\hat{f}^\star = \hat{T}^\star \hat{f}^\star + \hat{m}_\oplus^\star$, where $\hat{m}_\oplus^\star$ is defined as $\hat{m}_\oplus$ with $\hat{\pi}_j$ and $\hat{m}_j$ being replaced by $\hat{\pi}_j^\star$ and $\hat{m}_j^\star$, respectively. The corresponding version of the backfitting algorithm as at (2-11) is given by $\hat{f}^{\star[r]} = \hat{T}^\star \hat{f}^{\star[r-1]} + \hat{m}_\oplus^\star$, $r \geq 1$. It holds that $\hat{f}^{\star[r]}$ converges to $\hat{f}^\star$ as $r \to \infty$ under the condition that

(3-5)
$$\int_{[0,1]^2} \left[\frac{\hat{p}_{jk}^\star(x_j, x_k)}{\hat{p}_j^\star(x_j)\hat{p}_k^\star(x_k)} \right]^2 \hat{p}_j^\star(x_j)\hat{p}_k^\star(x_k)\, dx_j\, dx_k < \infty \quad \text{for all } 1 \leq j \neq k \leq d.$$

An analogue of Theorem 2.2 also holds. We make the following additional assumptions for this.

(A2) For the constant $\beta \geq 0$ in the condition (A1), $|t^{\beta+1}\phi_{U_j}'(t)| = O(1)$ as $|t| \to \infty$ and $\int |t^\beta \phi_K(t)|\, dt < \infty$.

(A3) For the constant $\beta \geq 0$ in the condition (A1), $h_j \to 0$ and $n(h_j h_k)^{1+2\beta}/\log n \to \infty$ as $n \to \infty$ for all $1 \leq j \neq k \leq d$.

THEOREM 3.2. (Han and Park [ibid.]). *Assume the conditions (C1), (C3) and (A1)–(A3). Then, with probability tending to one, the solution of the system of equations (3-4) exists and is unique. Furthermore, there exists a constant $0 < \gamma < 1$ such that*

$$\lim_{n\to\infty} P\Big(\|\hat{f}^{\star[r]} - \hat{f}^\star\|_2 \leq \gamma^r (\|\hat{f}^{\star[0]}\|_2 + (1-\gamma)^{-1}\|\hat{m}_\oplus^\star\|_2) \Big) = 1.$$

Now we discuss the asymptotic properties of $\hat{f}^\star$ and its components. To identify the individual components $\hat{f}_j^*$, we use the constraints

$$\int_0^1 \hat{f}_j^\star(x_j)\hat{p}_j^\star(x_j)\, dx_j = 0, \quad 1 \leq j \leq d. \tag{3-6}$$

As in Section 2, we assume $\mathrm{E}Y = 0$ for simplicity so that $f(\mathbf{x}) = f_1(x_1) + \cdots + f_d(x_d)$ with f_j satisfying the constraints (2-15). We also set $h_j \asymp h$.

The asymptotic analysis of $\hat{f}_j^\star$ is much more complex than in the case of no measurement error. To explain the main technical challenges, we note that

$$\hat{f}_j^\star - f_j = \hat{\delta}_j - \sum_{k\neq j} \hat{\pi}_j^\star(\hat{f}_k^\star - f_k), \quad 1 \leq j \leq d, \tag{3-7}$$

where $\hat{\delta}_j = \hat{m}_j^\star - \hat{\pi}_j^\star(f)$. Since

$$\pi_j(f) = \int_{[0,1]^{d-1}} \mathrm{E}(Y|\mathbf{X} = \mathbf{x}) \frac{p(\mathbf{x})}{p_j(x_j)}\, d\mathbf{x}_{-j} = \mathrm{E}(Y|X_j = x_j) = m_j,$$

$\hat{\delta}_j$ basically represent the errors of $\hat{m}_j^\star$ as an estimator of m_j. Each $\hat{\delta}_j$ corresponds to $\hat{m}_j^A + \hat{m}_j^B + \hat{p}_j^{-1} \sum_{k \neq j} \hat{m}_{jk}^C$ in the no measurement error case. Consider the same decomposition $\hat{\delta}_j = \hat{m}_j^{\star A} + \hat{m}_j^{\star B} + \hat{p}_j^{\star -1} \sum_{k \neq j} \hat{m}_{jk}^{\star C}$, where $\hat{m}_j^{\star A}$, $\hat{m}_j^{\star B}$ and $\hat{p}_j^{\star -1} \sum_{k \neq j} \hat{m}_{jk}^{\star C}$ are defined in the same way as $\hat{m}_j^A$, $\hat{m}_j^B$ and $\hat{p}_j^{-1} \sum_{k \neq j} \hat{m}_{jk}^C$, respectively, with $K_{h_j}(x_j, X_{ij})$ and $K_{h_k}(x_k, X_{ik})$ being replaced by $\tilde{K}_{h_j}^\star(x_j, Z_{ij})$ and $\tilde{K}_{h_k}^\star(x_k, Z_{ik})$, respectively. We have seen in Section 2 that the error components $\hat{m}_j^B$ and $\hat{m}_{jk}^C$ of $\hat{m}_j - \hat{\pi}_j(f)$ are spread, through the backfitting operation, into the errors of the other component function estimators $\hat{f}_k, k \neq j$, to the first order. In the present case, the errors are of two types. One type is for the replacement of K_{h_j} with X_{ij} by $\tilde{K}_{h_j}^\star$ with contaminated Z_{ij}, and the other is for those one would have when one uses K_{h_j} with X_{ij} in the estimation of f. The analysis of the first type is more involved. It has an additional complexity that we need to analyze whether the first type of errors in $\hat{m}_j^{\star B}$ and $\hat{m}_{jk}^{\star C}$ are spread into the errors of $\hat{f}_k^\star$ for $k \neq j$, through the backfitting operation.

Han and Park [2018] solved this problem and proved the following theorem. To state the theorem, let

$$\tau_n(\beta) = \begin{cases} 1 & \beta < 1/2 \\ \sqrt{\log h^{-1}} & \beta = 1/2 \\ h^{1/2-\beta} & \beta > 1/2. \end{cases}$$

Also, let $r_j^\star$ be generic stochastic terms such that

$$\sup_{x_j \in [2h_j, 1-2h_j]} |r_j^\star(x_j)| = o_p(h^2), \qquad \sup_{x_j \in [0,1]} |r_j^\star(x_j)| = O_p(h^2).$$

THEOREM 3.3. (Han and Park [ibid.]). *Assume the conditions (C1), (C3)–(C5), (A1) and (A2). Assume also that $nh^{3+4\beta}/\log n$ is bounded away from zero. Then, uniformly for $x_j \in [0,1]$,*

$$\hat{f}_j^\star(x_j) = f_j(x_j) + h_j \frac{\mu_{1,j}(x_j)}{\mu_{0,j}(x_j)} f_j'(x_j) + \frac{1}{2} h_j^2 \mu_2 f_j''(x_j) + \Delta_j(x_j)$$
$$+ r_j^\star(x_j) + O_p\left(\sqrt{\frac{\log n}{nh^{1+2\beta}}} \cdot \tau_n(\beta)\right), \quad 1 \leq j \leq d,$$

where Δ_j are the same as those in Theorem 2.3.

The rates of convergence of $\hat{f}_j^\star$ to their targets f_j are readily obtained from Theorem 3.3. For example, in case $\beta < 1/2$ we may get

$$\sup_{x_j \in [2h_j, 1-2h_j]} |\hat{f}_j^\star(x_j) - f_j(x_j)| = O_p\left(n^{-2/(5+2\beta)}\sqrt{\log n}\right),$$
$$\sup_{x_j \in [0,1]} |\hat{f}_j^\star(x_j) - f_j(x_j)| = O_p\left(n^{-1/(5+2\beta)}\right)$$

by choosing $h \asymp n^{-1/(5+2\beta)}$. The uniform rate in the interior is known to be the optimal rate that one can achieve in one-dimensional deconvolution problems, see Fan [1991]. For other cases where $\beta \geq 1/2$, see Corollary 3.5 of Han and Park [2018].

3.3 Extension to partially linear additive models.

In this subsection we consider the model

$$Y = \boldsymbol{\theta}^\top \mathbf{X} + f_1(Z_1) + \cdots + f_d(Z_d) + \varepsilon, \tag{3-8}$$

where ε is independent of the predictor vectors $\mathbf{X} \equiv (X_1, \ldots, X_p)^\top$ and $\mathbf{Z} \equiv (Z_1, \ldots, Z_d)^\top$, $\boldsymbol{\theta}$ are unknown and f_j are unknown univariate functions. We do not observe $\mathbf{X}$ and $\mathbf{Z}$, but the contaminated $\mathbf{X}^* = \mathbf{X}+\mathbf{U}$ and $\mathbf{Z}^* = \mathbf{Z}+\mathbf{V}$ for measurement error vectors $\mathbf{U}$ and $\mathbf{V}$. We assume that ε is also independent of $(\mathbf{U}, \mathbf{V})$, $\mathbf{U}$ has mean zero and a known variance $\boldsymbol{\Sigma}_\mathbf{U}$ and is independent of $\mathbf{V}$, V_j are independent across j and have known densities, and $(\mathbf{U}, \mathbf{V})$ is independent of $\mathbf{X}$ and $\mathbf{Z}$. Below, we outline the work of E. R. Lee, Han, and Park [2018] that studies the estimation of $\boldsymbol{\theta}$ and f_j in the model (3-8) based on independent and identically distributed observations $(\mathbf{X}_i^*, \mathbf{Z}_i^*, Y_i)$, $1 \leq i \leq n$.

Let $\mathcal{H}$ be the space of square integrable functions $g : \mathbb{R}^d \to \mathbb{R}$ such that $g(\mathbf{z}) = g_1(z_1) + \cdots + g_d(z_d)$. Let $\Pi(\cdot|\mathcal{H})$ denote the projection operator onto $\mathcal{H}$. Define $\xi = \Pi(E(Y|\mathbf{Z} = \cdot)|\mathcal{H})$ and $\eta_j = \Pi(E(X_j|\mathbf{Z} = \cdot)|\mathcal{H})$, $1 \leq j \leq p$. We write $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^\top$. Under the condition that $\mathbf{D} := \mathrm{E}(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z}))(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z}))^\top$ is positive definite, it holds that

$$\boldsymbol{\theta} = \mathbf{D}^{-1}\mathrm{E}(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z}))(Y - \xi(\mathbf{Z})) \overset{\text{let}}{=} \mathbf{D}^{-1}\mathbf{c}. \tag{3-9}$$

If we observe $\mathbf{X}_i$ and $\mathbf{Z}_i$, then the estimation of $\boldsymbol{\theta}$ is straightforward from the Equation (3-9). If we observe $\mathbf{Z}_i$ and $\mathbf{X}_i^*$ but not $\mathbf{X}_i$, then we may employ the standard technique that corrects 'attenuation effect' due to the measurement errors $\mathbf{U}_i$ in the estimation of $\mathbf{D}$, see Liang, Härdle, and Carroll [1999].

In our setting where both $\mathbf{X}_i$ and $\mathbf{Z}_i$ are not available, we may estimate $\boldsymbol{\eta}$ and ξ by the technique we have discussed in Section 3.1 with the normalized deconvolution kernel scheme. Call them $\hat{\boldsymbol{\eta}}^\star$ and $\hat{\xi}^\star$, respectively. A further complication here is that we may not use $\hat{\eta}^\star(\mathbf{Z}_i)$ and $\hat{\xi}^\star(\mathbf{Z}_i)$ in a formula for estimating $\boldsymbol{\theta}$ that basically replaces the expectations in (3-9) by the corresponding sample average, since $\mathbf{Z}_i$ are not available. E. R. Lee, Han, and Park [2018] successfully addressed this problem by observing the following identities.

$$\mathbf{D} = \int_{[0,1]^d} \mathrm{E}\left((\mathbf{X}^* - \boldsymbol{\eta}(\mathbf{z}))(\mathbf{X}^* - \boldsymbol{\eta}(\mathbf{z}))^\top \middle| \mathbf{Z} = \mathbf{z} \right) p_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} - \boldsymbol{\Sigma}_{\mathbf{U}},$$

$$\mathbf{c} = \int_{[0,1]^d} \mathrm{E}\left((\mathbf{X}^* - \boldsymbol{\eta}(\mathbf{z}))(Y - \xi(\mathbf{z}))^\top \middle| \mathbf{Z} = \mathbf{z} \right) p_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z},$$

where $p_{\mathbf{Z}}$ denote the joint density of $\mathbf{Z}$. Using the normalized deconvolution kernel function $\tilde{K}^\star_{b_j}$ introduced in Section 3.1 with bandwidth b_j being possibly different from h_j that are used to estimate $\boldsymbol{\eta}$ and ξ, we may estimate $\mathbf{D}$ and $\mathbf{c}$ by

$$\hat{\mathbf{D}} = n^{-1} \sum_{i=1}^n \int_{[0,1]^d} (\mathbf{X}_i^* - \hat{\boldsymbol{\eta}}^\star(\mathbf{z}))(\mathbf{X}_i^* - \hat{\boldsymbol{\eta}}^\star(\mathbf{z}))^\top \prod_{j=1}^d \tilde{K}^\star_{b_j}(z_j, Z_{ij}^*)\, d\mathbf{z} - \boldsymbol{\Sigma}_{\mathbf{U}},$$

$$\hat{\mathbf{c}} = n^{-1} \sum_{i=1}^n \int_{[0,1]^d} (\mathbf{X}_i^* - \hat{\boldsymbol{\eta}}^\star(\mathbf{z}))(Y_i - \hat{\xi}^\star(\mathbf{z})) \prod_{j=1}^d \tilde{K}^\star_{b_j}(z_j, Z_{ij}^*)\, d\mathbf{z}.$$

These gives an estimator $\hat{\boldsymbol{\theta}} = \hat{\mathbf{D}}^{-1}\hat{\mathbf{c}}$ of $\boldsymbol{\theta}$.

We may then estimate the additive function $f = f_1 + \cdots + f_d$ and its component f_j by applying the technique discussed in Section 3. In this application we takes $Y_i - \hat{\boldsymbol{\theta}}^\top \mathbf{X}_i^*$ as responses and $\mathbf{Z}_i^*$ as the contaminated predictor values. Since the rate of convergence of the parametric estimator $\hat{\boldsymbol{\theta}}$ is faster than the nonparametric rate, as we will see in the following theorem, the resulting estimators of f and its components f_j have the same first-order asymptotic properties as the corresponding oracle estimators that use $Y_i - \boldsymbol{\theta}^\top \mathbf{X}_i^*$ as responses. The asymptotic properties of the oracle estimators are the same as in Theorem 3.3. Theorem 3.4 below demonstrates the best possible rates that $\hat{\boldsymbol{\theta}}$ can achieves in the three ranges of β, the index for the smoothness of measurement error distribution in the condition (A1). To state the theorem for $\hat{\boldsymbol{\theta}}$, we make the following additional assumptions.

(B1) $\mathrm{E}(X_j^2 \mid \mathbf{Z} = \cdot)$ are bounded on $[0,1]^d$.

(B2) For $1 \le j \le p$, the component functions of the additive function η_j are twice continuously differentiable on $[0,1]$.

(B3) $\mathrm{E}(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z}))(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z}))^\top$ is positive definite.

(B4) There exist constants $C > 0$ such that $\mathrm{E}e^{uW} \le \exp(Cu^2/2)$ for all u, for $W = U_j$, X_j and ε.

THEOREM 3.4. (E. R. Lee, Han, and Park [ibid.]). *Assume the condition (C1) holds for the marginal and joint densities of Z_j and Z_{jk} for all $1 \le j \ne k \le d$. Also, assume the conditions (C3), (A1), (A2) and (B1)–(B4) hold. Then, (i) $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{-1/2})$ when $\beta < 1/2$ if $h_j \asymp n^{-\alpha_1}$ and $b_j \asymp n^{-\alpha_2}$ with $1/4 \le \alpha_2 < \alpha_1/(2\beta)$ and $\max\{1/6, \beta/2\} < \alpha_1 < 1/(3+2\beta)$; (ii) $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{-1/2}\log n)$ when $\beta = 1/2$ if $h_j \asymp b_j \asymp n^{-1/4}\sqrt{\log n}$; (iii) $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{-1/(1+2\beta)}\sqrt{\log n})$ when $\beta > 1/2$ if $h_j \asymp b_j \asymp n^{-1/(2\beta+4)}(\log n)^{1/4}$.*

4 Hilbertian additive models

The analysis of non-Euclidean data objects is an emerging area in modern statistics. A well-known and most studied example is functional data analysis. There have been a few attempts for nonparametric models in this area. These include Dabo-Niang and Rhomari [2009], Ferraty, Laksaci, Tadj, and Vieu [2011] and Ferraty, Van Keilegom, and Vieu [2012]. They studied Nadaraya-Watson estimation of the full-dimensional regression function $\mathrm{E}(Y|\mathbf{X} = \cdot)$ without any structure when the response is in a separable Hilbert or Banach space. The full-dimensional estimator suffers from the curse of dimensionality. More recently, X. Zhang, Park, and Wang [2013], Han, Müller, and Park [2018] and Park, Chen, Tao, and Müller [2018] considered the estimation of structured nonparametric models for functional data, but their studies were either for SBF methods applied to $Y(t)$ for each t or for models based on finite number of functional principal/singular components of predictors and responses. Thus, the structured nonparametric models and the methods of estimating them were actually for finite-dimensional Euclidean variables.

In this section we introduce an additive model with response taking values in a Hilbert space and discuss briefly some statistical notions that lay the foundations for estimating the model. This discussion is largely based on the recent work in progress by Jeon [2018]. Let $\mathbf{Y}$ be a random element taking values in a separable Hilbert space $\mathbb{H}$. We confine our discussion to the case where the predictor $\mathbf{X} = (X_1, \cdots, X_d)^\top$ takes values in $[0,1]^d$, however. This is mainly because SBF methods discussed in the previous sections require the marginal and joint densities of X_j and (X_j, X_k), which generally do not exist in infinite-dimensional non-Euclidean cases. For the case where the predictors do not have densities, one may employ 'surrogate probability density functions' as discussed in Delaigle and Hall [2010]. Let us denote a vector addition and a real-scalar multiplication by $\oplus$ and $\odot$, respectively. For Borel measurable maps $\mathbf{f}_j : [0,1] \to \mathbb{H}$ as additive components, an additive model for $\mathrm{E}(\mathbf{Y}|\mathbf{X})$ may be written as

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{f}_1(X_1) \oplus \cdots \oplus \mathbf{f}_d(X_d). \tag{4-1}$$

Below we introduce the notion of Bochner integral, and then discuss briefly its applications to some important statistical notions for the SBF theory.

4.1 Bochner integration.

Bochner integral is defined for Banach space-valued maps. We start with the classical definition. Let $(\mathcal{Z}, \mathfrak{A}, \mu)$ be a measure space and $\mathbb{B}$ be a Banach space with a norm denoted by $\|\cdot\|$. We say a map $\mathbf{m} : \mathcal{Z} \to \mathbb{B}$ is μ-simple if $\mathbf{m} = \bigoplus_{i=1}^n \mathbf{b}_i \odot 1_{A_i}$ for $\mathbf{b}_i \in \mathbb{B}$ and disjoint $A_i \in \mathfrak{A}$ with $\mu(A_i) < \infty$. In this case, the Bochner integral of $\mathbf{m}$ is defined by

$$\int \mathbf{m}\, d\mu = \bigoplus_{i=1}^{n} \mathbf{b}_i \odot \mu(A_i).$$

A map $\mathbf{m} : \mathcal{Z} \to \mathbb{B}$ is called μ-measurable if $\mathbf{m}$ is an almost everywhere limit of μ-simple maps. A μ-measurable map $\mathbf{m}$ is called Bochner integrable if $\int \|\mathbf{m}\| d\mu < \infty$. In this case, the Bochner integral of $\mathbf{m}$ is defined by

$$\int \mathbf{m}\, d\mu = \lim_{n\to\infty} \int \mathbf{m}_n\, d\mu$$

for a sequence of μ-simple maps $\mathbf{m}_n$ such that $\mathbf{m}_n \to \mathbf{m}$ $a.e.$ $[\mu]$.

In statistical applications of Bochner integrals, the measure μ in a measure space $(\mathcal{Z}, \mathfrak{A}, \mu)$ is the distribution of a random variable. In the case of the additive maps $\mathbf{f}_j$ in (4-1), μ corresponds to PX_j^{-1} where P is the probability measure of the probability space $(\Omega, \mathfrak{F}, P)$ where X_j is defined. The classical definition given above for μ-measurable maps is not appropriate since PX_j^{-1}-measurability of $\mathbf{f}_j$ is not equivalent to Borel-measurability of $\mathbf{f}_j$. In the model (4-1), we implicitly assume that $\mathbf{f}_j(X_j)$ are random elements, i.e., Borel-measurable with respect to $\mathfrak{F}$, as is usual in all statistical problems. For this reason we assume in the model (4-1) that each $\mathbf{f}_j$ is Borel-measurable with respect to the Borel σ-field of $[0, 1]$.

The notion of Bochner integral may be extended to Borel-measurable maps. We introduce it briefly here. We refer to Cohn [2013] for more details. For a Banach space $\mathbb{B}$, a map $\mathbf{m} : \mathcal{Z} \to \mathbb{B}$ is called simple if $\mathbf{m}$ takes only finitely many values. A map $\mathbf{m} : \mathcal{Z} \to \mathbb{B}$ is called strongly measurable if $\mathbf{m}$ is Borel-measurable and $\mathbf{m}(\mathcal{Z})$ is separable. A map $\mathbf{m} : \mathcal{Z} \to \mathbb{B}$ is called strongly integrable if $\mathbf{m}$ is strongly measurable and $\int_{\mathcal{Z}} \|\mathbf{m}\|\, d\mu < \infty$. If $\mathbf{m}$ is strongly integrable, then there exists a Cauchy sequence of strongly integrable simple maps $\mathbf{m}_n$ such that $\lim_{n,m\to\infty} \int \|\mathbf{m}_n - \mathbf{m}_m\|\, d\mu \to 0$ and $\lim_{n\to\infty} \mathbf{m}_n(z) = \mathbf{m}(z)$ for all $z \in \mathcal{Z}$. In this case, $\int \mathbf{m}\, d\mu$ is defined as $\lim_{n\to\infty} \int \mathbf{m}_n\, d\mu$.

4.2 Statistical properties of Bochner integrals.

Since the notion of Bochner integral is new in statistics, statistical properties of this integral have been rarely studied. There are many statistical notions and properties one needs to define and derive to develop relevant theory for estimating the model (4-1). It was only recent that Jeon [2018] studied such basic ingredients. Below, we present two formulas regarding the notions of expectation and of conditional expectation that are essential in developing further theory for the SBF estimation of (4-1).

Let $\mathbb{B}$ be a separable Banach space. Let $\mathbf{Z}$ and $\mathbf{W}$ be random elements taking values in σ-finite measure spaces $(\mathcal{Z}, \mathfrak{A}, \mu)$ and $(\mathcal{W}, \mathfrak{B}, \nu)$, respectively. We assume $P\mathbf{Z}^{-1} \ll \mu$, $P\mathbf{W}^{-1} \ll \nu$ and $P(\mathbf{Z}, \mathbf{W})^{-1} \ll \mu \otimes \nu$, where $P\mathbf{Z}^{-1}$, $P\mathbf{W}^{-1}$ and $P(\mathbf{Z}, \mathbf{W})^{-1}$ are the probability distributions of $\mathbf{Z}$, $\mathbf{W}$ and $(\mathbf{Z}, \mathbf{W})$, respectively, so that there exist densities of $\mathbf{Z}$, $\mathbf{W}$ and $(\mathbf{Z}, \mathbf{W})$, denoted by $p_{\mathbf{Z}}$, $p_{\mathbf{W}}$ and $p_{\mathbf{Z},\mathbf{W}}$, respectively. We first introduce a general expectation formula, and then a conditional expectation formula, in terms of the densities of $\mathbf{Z}$, $\mathbf{W}$ and $(\mathbf{Z}, \mathbf{W})$.

PROPOSITION 4.1. (Jeon [ibid.]) *Assume that* $\mathbf{f} : \mathcal{Z} \to \mathbb{B}$ *is a strongly measurable map such that* $E(\|\mathbf{f}(\mathbf{Z})\|) < \infty$. *Then,* $E(\mathbf{f}(\mathbf{Z})) = \int_{\mathcal{Z}} \mathbf{f}(\mathbf{z}) \odot p_{\mathbf{Z}}(\mathbf{z})\, d\mu$.

PROPOSITION 4.2. (Jeon [ibid.]) *Assume that* $p_{\mathbf{W}} \in (0, \infty)$ *on* $\mathcal{W}$ *and that* $\mathbf{f} : \mathcal{Z} \to \mathbb{B}$ *is a strongly measurable map such that* $E(\|\mathbf{f}(\mathbf{Z})\|) < \infty$. *Let* $\mathbf{g} : \mathcal{W} \to \mathbb{B}$ *be a map defined by*

$$\mathbf{g}(\mathbf{w}) = \begin{cases} \int_{\mathcal{Z}} \mathbf{f}(\mathbf{z}) \odot \frac{p_{\mathbf{Z},\mathbf{W}}(\mathbf{z},\mathbf{w})}{p_{\mathbf{W}}(\mathbf{w})}\, d\mu, & \textit{if } \mathbf{w} \in D_{\mathcal{W}} \\ \mathbf{g}_0(\mathbf{w}), & \textit{otherwise} \end{cases}$$

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

4.3 Discussion. The additive regression model (4-1) for Hilbertian response have many important applications. Non-Euclidean data objects often take values in Hilbert spaces. Examples include functions, images, probability densities and simplices. Among them, the latter two data objects have certain constraints. A density is non-negative and its integral over the corresponding domain where it is defined equals 1. A simplex data object, $(v_1, \cdots, v_D)^\top$ with $v_k > 0$ for $1 \le k \le D$ and $\sum_{k=1}^{D} v_k = 1$, has similar constraints. Analyzing such data objects with standard Euclidean regression techniques would give estimates that are off the space where the data objects take values. The approach based on the model (4-1) with the corresponding Hilbertian operations $\oplus$ and $\odot$ would give a proper estimate of the regression map that forces its values lie in the space of the data objects. It also avoids the curse of dimensionality when d is high. This way would lead us to a powerful nonparametric technique that unifies various statistical methods for analyzing non-Euclidean data objects.

References

E. J. Beltrami (1967). "On infinite-dimensional convex programs". *J. Comput. System Sci.* 1, pp. 323–329. MR: 0232603.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, p. 560. MR: 1245941 (cit. on pp. 3018, 3020).

Andreas Buja, Trevor Hastie, and Robert Tibshirani (1989). "Linear smoothers and additive models". *Ann. Statist.* 17.2, pp. 453–555. MR: 994249 (cit. on p. 3014).

Raymond J. Carroll and Peter Hall (1988). "Optimal rates of convergence for deconvolving a density". *J. Amer. Statist. Assoc.* 83.404, pp. 1184–1186. MR: 997599 (cit. on p. 3026).

Donald L. Cohn (2013). *Measure theory*. Second. Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks]. Birkhäuser/Springer, New York, pp. xxi+457. MR: 3098996 (cit. on p. 3034).

Sophie Dabo-Niang and Noureddine Rhomari (2009). "Kernel regression estimation in a Banach space". *J. Statist. Plann. Inference* 139.4, pp. 1421–1434. MR: 2485136 (cit. on p. 3033).

Aurore Delaigle, Jianqing Fan, and Raymond J. Carroll (2009). "A design-adaptive local polynomial estimator for the errors-in-variables problem". *J. Amer. Statist. Assoc.* 104.485, pp. 348–359. MR: 2504382 (cit. on p. 3026).

Aurore Delaigle and Peter Hall (2010). "Defining probability density for a distribution of random functions". *Ann. Statist.* 38.2, pp. 1171–1193. MR: 2604709 (cit. on p. 3033).

– (2016). "Methodology for non-parametric deconvolution when the error distribution is unknown". *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 78.1, pp. 231–252. MR: 3453654 (cit. on p. 3026).

Aurore Delaigle, Peter Hall, and Alexander Meister (2008). "On deconvolution with repeated measurements". *Ann. Statist.* 36.2, pp. 665–685. MR: 2396811 (cit. on p. 3026).

Jianqing Fan (1991). "On the optimal rates of convergence for nonparametric deconvolution problems". *Ann. Statist.* 19.3, pp. 1257–1272. MR: 1126324 (cit. on p. 3031).

Jianqing Fan and Young K. Truong (1993). "Nonparametric regression with errors in variables". *Ann. Statist.* 21.4, pp. 1900–1925. MR: 1245773 (cit. on p. 3026).

Jianqing Fan and Wenyang Zhang (1999). "Statistical estimation in varying coefficient models". *Ann. Statist.* 27.5, pp. 1491–1518. MR: 1742497 (cit. on p. 3025).

– (2000). "Simultaneous confidence bands and hypothesis testing in varying-coefficient models". *Scand. J. Statist.* 27.4, pp. 715–731. MR: 1804172 (cit. on p. 3025).

F. Ferraty, I. Van Keilegom, and P. Vieu (2012). "Regression when both response and predictor are functions". *J. Multivariate Anal.* 109, pp. 10–28. MR: 2922850 (cit. on p. 3033).

Frédéric Ferraty, Ali Laksaci, Amel Tadj, and Philippe Vieu (2011). "Kernel regression with functional response". *Electron. J. Stat.* 5, pp. 159–171. MR: 2786486 (cit. on p. 3033).

Jerome H. Friedman and Werner Stuetzle (1981). "Projection pursuit regression". *J. Amer. Statist. Assoc.* 76.376, pp. 817–823. MR: 650892 (cit. on p. 3014).

K. Han and Byeong U. Park (2018). "Smooth backfitting for errors-in-variables additive models". To appear in *A*nnal of Statistics (cit. on pp. 3026–3031).

Kyunghee Han, Hans-Georg Müller, and Byeong U. Park (2018). "Smooth backfitting for additive modeling with small errors-in-variables, with an application to additive functional regression for multiple predictor functions". *Bernoulli* 24.2, pp. 1233–1265. MR: 3706793 (cit. on p. 3033).

Trevor Hastie and Robert Tibshirani (1993). "Varying-coefficient models". *J. Roy. Statist. Soc. Ser. B* 55.4. With discussion and a reply by the authors, pp. 757–796. MR: 1229881 (cit. on p. 3025).

J. M. Jeon (2018). "Additive regression with Hilbertian responses". PhD thesis. Seoul National University (cit. on pp. 3033–3035).

E. R. Lee, K. Han, and Byeong U. Park (2018). "Estimation of errors-in-variables partially linear additive models". To appera in *Statistica Sinica* (cit. on pp. 3031–3033).

Young K. Lee, Enno Mammen, and Byeong U. Park (2012a). "Flexible generalized varying coefficient regression models". *The Annals of Statistics* 40.3, pp. 1906–1933. MR: 3015048 (cit. on p. 3026).

– (2012b). "Projection-type estimation for varying coefficient regression models". *Bernoulli* 18.1, pp. 177–205. MR: 2888703 (cit. on p. 3025).

– (2014). "Backfitting and smooth backfitting in varying coefficient quantile regression". *Econom. J.* 17.2, S20–S38. MR: 3219148 (cit. on p. 3026).

Young Kyung Lee (2004). "On marginal integration method in nonparametric regression". *J. Korean Statist. Soc.* 33.4, pp. 435–447. MR: 2126371 (cit. on p. 3015).

– (2017). "Nonparametric estimation of bivariate additive models". *J. Korean Statist. Soc.* 46.3, pp. 339–348. MR: 3685573 (cit. on p. 3026).

Young Kyung Lee, Enno Mammen, and Byeong U. Park (2010). "Backfitting and smooth backfitting for additive quantile models". *Ann. Statist.* 38.5, pp. 2857–2883. MR: 2722458 (cit. on p. 3026).

Hua Liang, Wolfgang Härdle, and Raymond J. Carroll (1999). "Estimation in a semiparametric partially linear errors-in-variables model". *Ann. Statist.* 27.5, pp. 1519–1535. MR: 1742498 (cit. on p. 3031).

Oliver Linton and Jens Perch Nielsen (1995). "A kernel method of estimating structured nonparametric regression based on marginal integration". *Biometrika* 82.1, pp. 93–100. MR: 1332841 (cit. on p. 3014).

E. Mammen, O. Linton, and J. Nielsen (1999). "The existence and asymptotic properties of a backfitting projection algorithm under weak conditions". *Ann. Statist.* 27.5, pp. 1443–1490. MR: 1742496 (cit. on pp. 3014, 3015, 3018–3020, 3024).

Enno Mammen and Jens Perch Nielsen (2003). "Generalised structured models". *Biometrika* 90.3, pp. 551–566. MR: 2006834 (cit. on p. 3014).

Enno Mammen and Byeong U. Park (2005). "Bandwidth selection for smooth backfitting in additive models". *Ann. Statist.* 33.3, pp. 1260–1294. MR: 2195635 (cit. on p. 3024).

– (2006). "A simple smooth backfitting method for additive models". *Ann. Statist.* 34.5, pp. 2252–2271. MR: 2291499 (cit. on p. 3024).

Enno Mammen, Byeong U. Park, and Melanie Schienle (2014). "Additive models: extensions and related models". In: *The Oxford handbook of applied nonparametric and semiparametric econometrics and statistics*. Oxford Univ. Press, Oxford, pp. 176–211. MR: 3306926 (cit. on p. 3014).

Jens Perch Nielsen and Stefan Sperlich (2005). "Smooth backfitting in practice". *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67.1, pp. 43–61. MR: 2136638 (cit. on p. 3024).

Jean D. Opsomer (2000). "Asymptotic properties of backfitting estimators". *J. Multivariate Anal.* 73.2, pp. 166–179. MR: 1763322 (cit. on p. 3014).

Jean D. Opsomer and David Ruppert (1997). "Fitting a bivariate additive model by local polynomial regression". *Ann. Statist.* 25.1, pp. 186–211. MR: 1429922 (cit. on p. 3014).

Byeong U. Park, C.-J. Chen, W. Tao, and H.-G. Müller (2018). "Singular additive models for function to function regression". To appear in *Statistica Sinica* (cit. on p. 3033).

Byeong U. Park, Enno Mammen, Young K. Lee, and Eun Ryung Lee (2015). "Varying coefficient regression models: a review and new developments". *Int. Stat. Rev.* 83.1, pp. 36–64. MR: 3341079 (cit. on p. 3025).

Leonard Stefanski and Raymond J. Carroll (1990). "Deconvoluting kernel density estimators". *Statistics* 21.2, pp. 169–184. MR: 1054861 (cit. on p. 3026).

Lijian Yang, Byeong U. Park, Lan Xue, and Wolfgang Härdle (2006). "Estimation and testing for varying coefficients in additive models with marginal integration". *J. Amer. Statist. Assoc.* 101.475, pp. 1212–1227. MR: 2328308 (cit. on p. 3025).

Kyusang Yu, Enno Mammen, and Byeong U. Park (2011). "Semi-parametric regression: efficiency gains from modeling the nonparametric part". *Bernoulli* 17.2, pp. 736–748. MR: 2787613 (cit. on p. 3026).

Kyusang Yu, Byeong U. Park, and Enno Mammen (2008). "Smooth backfitting in generalized additive models". *Ann. Statist.* 36.1, pp. 228–260. MR: 2387970 (cit. on pp. 3024, 3025).

Xiaoke Zhang, Byeong U. Park, and Jane-Ling Wang (2013). "Time-varying additive models for longitudinal data". *J. Amer. Statist. Assoc.* 108.503, pp. 983–998. MR: 3174678 (cit. on p. 3033).

Received 2017-12-06.

Byeong U. Park
bupark@stats.snu.ac.kr