*The Annals of Statistics*
2018, Vol. 46, No. 5, 2216–2250
https://doi.org/10.1214/17-AOS1617


# SMOOTH BACKFITTING FOR ERRORS-IN-VARIABLES ADDITIVE MODELS[1]

BY KYUNGHEE HAN AND BYEONG U. PARK

*Seoul National University*

In this work, we develop a new smooth backfitting method and theory for estimating additive nonparametric regression models when the covariates are contaminated by measurement errors. For this, we devise a new kernel function that suitably deconvolutes the bias due to measurement errors as well as renders a projection interpretation to the resulting estimator in the space of additive functions. The deconvolution property and the projection interpretation are essential for a successful solution of the problem. We prove that the method based on the new kernel weighting scheme achieves the optimal rate of convergence in one-dimensional deconvolution problems when the smoothness of measurement error distribution is less than a threshold value. We find that the speed of convergence is slower than the univariate rate when the smoothness of measurement error distribution is above the threshold, but it is still much faster than the optimal rate in multivariate deconvolution problems. The theory requires a deliberate analysis of the nonnegligible effects of measurement errors being propagated to other additive components through backfitting operation. We present the finite sample performance of the deconvolution smooth backfitting estimators that confirms our theoretical findings.

**1. Introduction.** We study the estimation of the additive regression model

$$Y = f_1(X_1) + \cdots + f_d(X_d) + \varepsilon, \tag{1.1}$$

where $\mathbb{E}(\varepsilon \mid \mathbf{X}) = 0$, $Y$ is a response variable, $\mathbf{X} = (X_1, \ldots, X_d)$ is a $d$-dimensional covariate vector and $f_j$ are unknown univariate component functions. The model has been popular since [25] and is known to be one of the structured nonparametric models that one may estimate with univariate accuracy. Indeed, [20] proposed and studied a powerful kernel smoothing technique that fits (1.1) based on the observations of $(\mathbf{X}, Y)$, called smooth backfitting (SBF). They proved that the method yields an estimator that converges to the true regression function $f(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$ at a univariate rate, and thus avoids the curse of dimensionality. In this paper, we consider the case where one does not observe $X_j$, but observes noisy variables $Z_j = X_j + U_j$ contaminated by measurement errors $U_j$. For this

Received November 2016; revised July 2017.

[1]Supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIP) (No. NRF-2014R1A4A1007895 and No. NRF-2015R1A2A1A05001753).

*MSC2010 subject classifications.* Primary 62G08; secondary 62G20.

*Key words and phrases.* Nonparametric additive regression, smooth backfitting, errors-in-variables models, deconvolution, kernel smoothing.

errors-in-variables additive regression problem, we establish a suitably modified SBF method and develop its theory. To the best of our knowledge, this is the first attempt dealing with errors-in-variables in structured nonparametric regression.

The SBF method is based on local smoothing. The idea of local smoothing is that, at each point in the domain where the target function sits, it gives more weights to observations that are closer to the point, the weights being determined by a kernel function. The very core of the matter is that, for noisy covariate values that are close to a point of interest, the corresponding unobserved true covariate values may be far away from the point due to measurement errors, and thus the concomitant observations of $Y$ may not have relevant information about the regression function at the point. Therefore, one should use a nonconventional kernel function that deconvolutes efficiently the effects of measurement errors in local smoothing. In the standard errors-in-variable nonparametric regression, [24] introduced the so-called deconvolution kernel for this purpose. Indeed, [10] and [6] proved that nonparametric regression estimators based on the deconvolution kernel have the same biases as in the case of no measurement errors.

In our modification of the SBF method, we introduce a novel kernel weighting scheme that accounts for this problem in additive regression. Our new kernel function is different from the deconvolution kernel in that it is normalized so that its integral along the line of local smoothing equals one. The normalization property of the kernel function is required for a projection interpretation of the SBF estimator, which is a key element in the theoretical development of the method; see [20]. One may normalize the deconvolution kernel by scaling it up or down, but then the resulting normalized kernel loses the property of deconvoluting the effects of measurement errors.

In the standard errors-in-variables nonparametric regression, [10] and [6] showed that the variances of the regression estimators are inflated by an order of magnitude that depends on the smoothness of the measurement error distribution. The inflated variances are basically from stochastic terms arising from deconvoluted kernel smoothing. In SBF additive regression, the analysis of stochastic terms leading to the inflated variance is much more complex than the standard nonparametric regression.

To give an idea of the difficulty involved, let $(\mathbf{Z}^i, Y^i)$ be the observations of $(\mathbf{Z}, Y)$ and $\tilde{K}_h^\star$, for a bandwidth $h > 0$, denote our new kernel scheme that assigns the weights $\tilde{K}_h^\star(x_j, z_j)$ to noisy covariate values $z_j$ of $Z_j$ in local smoothing around each point $x_j$. Let $m_j = \mathbb{E}(Y \mid X_j = \cdot)$ denote the marginal regression functions and define their estimators by

$$\hat{m}_j(x_j) = \left( n^{-1} \sum_{i=1}^{n} \tilde{K}_h^\star(x_j, Z_j^i) \right)^{-1} n^{-1} \sum_{i=1}^{n} \tilde{K}_h^\star(x_j, Z_j^i) Y^i.$$

Then the SBF estimator $\hat{f}(\mathbf{x}) = \hat{f}_1(x_1) + \cdots + \hat{f}_d(x_d)$ of the additive function $f(\mathbf{x}) = f_1(x_1) + \cdots + f_d(x_d)$ satisfies the equation

$$\hat{f} - f = \hat{T}(\hat{f} - f) + \hat{\delta}_\oplus,$$

where $\hat{T}$ is a linear operator that maps the space of additive functions to itself and $\hat{\delta}_{\oplus}$ is an additive function. The linear operator $\hat{T}$ involves the estimators of the one- and two-dimensional density functions of the unobserved covariates $X_j$, obtained by using the new kernel $\tilde{K}_h^{\star}$, and $\hat{\delta}_{\oplus}$ is constructed from the errors of $\hat{m}_j$; see Section 2.2 and Section 3.2 for the definitions of $\hat{T}$ and $\hat{\delta}_{\oplus}$, respectively. It can be shown that $\hat{T}$ is a contraction operator, so that one has

$$\hat{f} - f = \sum_{j=0}^{\infty} \hat{T}^j \hat{\delta}_{\oplus}. \tag{1.2}$$

The equation at (1.2) demonstrates that there are two types of errors we need to analyze. One type is for the errors of the one- and two-dimensional density estimators involved in $\hat{T}$, and the other is for those in $\hat{m}_j$ as estimators of $m_j$. The errors of $\hat{m}_j$ not only affect $\hat{f}_j$, but also are spread into the errors of other component function estimators $\hat{f}_k$, $k \neq j$, and interrelated with those of $\hat{T}$ through the backfitting operation, which is particularly difficult to analyze since the disseminated errors of $\hat{m}_j$ involve nonnegligible effects of measurement errors.

In this paper, we show that the SBF estimator based on the new kernel scheme $\tilde{K}_h^{\star}$ has the same bias expansion as in the case of no measurement errors. But, for the variance part, we find that it has two additional terms due to measurement errors. In terms of estimating a specific component function $f_j$, one is a stochastic term solely from measurement errors, which one would also have additionally in a stochastic expansion of the "oracle" estimator of $f_j$. The oracle estimator of $f_j$ is the one-dimensional regression estimator obtained by regressing $Y^i - \sum_{k \neq j} f_k(X_k^i)$ on $X_j^i$. The other is a nonnegligible stochastic term arising from the backfitting in conjunction with measurement errors. The latter term is found to be of the same magnitude as the stochastic terms of the oracle estimator when the distributions of the measurement errors $U_j$ is less smooth, but it dominates all other stochastic terms in the expansions of $\hat{f}_j - f_j$, otherwise. It turns out that our SBF estimators of $f_j$ have the optimal univariate rate in errors-in-variables problems when the smoothness of the measurement error distributions is less than a threshold value, and in such a case they have the same rate as oracle estimators. For smoother measurement error distributions, however, one would get a rate that is slower than the univariate rate, although it is much faster than the optimal multivariate rate.

There have been a plenty of studies on structured nonparametric models. Examples include [21], [27], [19], [15], [23], [16] and [17]. But, none of these has treated the type of the errors-in-variables problem considered in this paper. Some earlier works on partially linear models or varying coefficient models by [18], [28] and [26], among others, are of completely different nature and simpler than the current study. In their frameworks, the contaminated covariates enter the linear part, not the nonparametric part of the models, or they enter the nonparametric part that does not have any structure. The additive model is bottommost in structured

nonparametric regression. The methodology and theory we develop in this paper may provide the basic building blocks for studying errors-in-variables problems in more complex structured nonparametric models.

**2. Methodology.** We assume we observe i.i.d. copies of $(\mathbf{Z}, Y)$, denoted by $(\mathbf{Z}^i, Y^i)$, $1 \le i \le n$, where $\mathbf{Z}^i = \mathbf{X}^i + \mathbf{U}^i$ and $\mathbf{U}^i$ are independent of the true unobserved covariate vectors $\mathbf{X}^i$. We also assume that the true covariate vector $\mathbf{X}$ is supported on a compact set in $\mathbb{R}^d$, say $[0, 1]^d$ without loss of generality. We rewrite the model (1.1) as

$$\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = f_0 + f_1(x_1) + \cdots + f_d(x_d), \tag{2.1}$$

where now each component function has the constraint $\mathbb{E} f_j(X_j) = 0$ for identifiability and $f_0 = \mathbb{E}Y$. Taking the conditional expectation of both sides of (2.1) given $X_j = x_j$ for each $1 \le j \le d$, we get

$$f_j(x_j) = m_j(x_j) - f_0 - \sum_{k \ne j} \int_0^1 f_k(x_k) \frac{p_{jk}(x_j, x_k)}{p_j(x_j)}\, dx_k, \qquad 1 \le j \le d, \tag{2.2}$$

where $p_j$ and $p_{jk}$ are the marginal densities of $X_j$ and $(X_j, X_k)$, respectively, and $m_j = \mathbb{E}(Y \mid X_j = \cdot)$. The SBF method is to estimate the unknown functions $m_j$, $p_j$ and $p_{jk}$ in the estimating equation (2.2) and then solve the resulting system of estimated integral equations. The name SBF comes from the fact that the resulting estimator of each component function $f_j$ is the minimizer of a "smoothed" version of the objective function for the corresponding (ordinary) backfitting estimator; see the related discussion in Section 3 of [15].

2.1. *Deconvolution normalized kernels.* As we mentioned in Section 1, we need to employ a normalized kernel function that also has the ability of deconvoluting the measurement errors $U_j^i$ in the covariates $Z_j^i$ in the estimation of $m_j$, $p_j$ and $p_{jk}$ in (2.2). Below we present a novel kernel weighting scheme that has both of these properties.

Let $K$ be a baseline kernel function that is nonnegative, symmetric and supported on $[-1, 1]$, and $h > 0$ be the bandwidth in local smoothing. Let $X$, $U$ and $Z$ denote generic random variables representing $X_j$, $U_j$ and $Z_j = X_j + U_j$, respectively. We assume that the densities of $U_j$, and thus their Fourier transforms, are known since otherwise the densities of $X_j$ are not identified by those of the observed $Z_j$. In case the densities of $U_j$ are unknown but there are available repeated measurements of $X_j$ (with errors) for each subject, then one may be able to estimate the densities of $U_j$ based on the repeated measurements or a preliminary sample, and thus those of $X_j$; see [7], [4], [13] and [5] for the methodology.

The classical deconvolution kernel function, introduced by [24], is

$$K^D(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itu} \frac{\phi_K(t)}{\phi_U(t/h)}\, dt, \tag{2.3}$$

where $\phi_K$ and $\phi_U$ are the Fourier transforms of $K$ and the density of $U$, respectively. Here, for simplicity we suppress the dependence of $K^D$ on $h$ through the argument in $\phi_U$. In this definition, we basically assume that $\phi_U$ is not vanishing on $(-\infty, \infty)$. With this kernel, one assigns the weights $K^D(\frac{x-z}{h})$ to an observation $Z = z$ in local smoothing around $x$. The salient feature of $K^D$ is the "unbiased scoring" property,

$$\mathbb{E}\left[K^D\left(\frac{x-Z}{h}\right) \,\middle|\, X\right] = K\left(\frac{x-X}{h}\right), \qquad x \in [0,1]. \tag{2.4}$$

The property (2.4) entails that the bias of the kernel estimators based on $K^D$ and $Z_j^i$ is the same as that of the conventional kernel estimators based on the baseline kernel $K$ and the original $X_j^i$. Thus, the standard kernel smoothing theory for bias in the case of no measurement errors applies in errors-in-variables problems.

Now, in SBF additive regression when the true covariates $X_j$ are observed, a normalized kernel function, denoted by $\tilde{K}_h$, is constructed from $K$ in the following way:

$$\tilde{K}_h(x,u) = \left[\int_0^1 K_h(v-u)\,dv\right]^{-1} K_h(x-u)\cdot \mathbb{I}_{[0,1]^2}(x,u), \tag{2.5}$$

where and throughout the paper $K_h(v) = K(v/h)/h$. The normalized kernel $\tilde{K}_h(x,u)$ equals the conventional kernel $K_h(x-u)$ for all $u \in [0,1]$ if $x$ is in the "interior" region $I_0 \equiv [2h, 1-2h]$. For the standard kernel $K_h(\cdot)$, the interior region is $[h, 1-h]$ since $\int_0^1 K_h(x-v)\,dv = 1$ for $x \in [h, 1-h]$. However, for the normalized kernel $\tilde{K}_h(\cdot,\cdot)$ we note that

$$\int_0^1 \tilde{K}_h(x,u)\,du = \int_0^1 \frac{K_h(x-u)}{\int_0^1 K_h(v-u)\,dv}\,du = 1$$

for $x \in [2h, 1-2h]$. The reason is that, for $x \in [2h, 1-2h]$, all $u$ for which $K_h(x-u) \neq 0$ belong to $[h, 1-h]$, so that $\int_0^1 K_h(v-u)\,dv = \int_0^1 K_h(u-v)\,dv = 1$ for all $u$ such that $K_h(x-u) \neq 0$. With this normalized kernel, one assigns the weight $\tilde{K}_h(x, X^i)$ to the $i$th covariate value $X^i$ in local smoothing around $x \in [0,1]$. This kernel satisfies

$$\int_0^1 \tilde{K}_h(x,u)\,dx = 1 \qquad \text{for all } u \in [0,1]. \tag{2.6}$$

In this regard, the normalized kernel $\tilde{K}_h$ is different from the typical one for boundary correction, say $K_h(x,u)$. For the latter, one normalizes $K_h(x-\cdot)$ for each $x \in [0,1]$, not $K_h(\cdot - u)$ for each $u \in [0,1]$, so that $\int_0^1 K_h(x,u)\,du = 1$ for all $x \in [0,1]$.

To get a kernel that has both the deconvolution and normalization properties, one may easily think of normalizing the deconvolution kernel $K_h^D(x-z)$ as in (2.5) with $K_h^D \equiv h^{-1}K^D(\cdot/h)$ taking the role of $K_h$ there, where $K^D$ is defined

at (2.3). But, we find that the resulting kernel does not have the unbiased scoring property such as (2.4), so that its use fails to remove the bias at a point $x$ owing to those unobserved covariate values $X^i$ that are far from $x$. Another option is to reverse the order of normalization and deconvolution. For this, we observe from (2.3) that

$$\begin{aligned} K_h^D(x-z) &= \frac{1}{2\pi h}\int_{-\infty}^{\infty} e^{-it\cdot\frac{x-z}{h}}\frac{\phi_K(t)}{\phi_U(t/h)}\,dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itz}\frac{\phi_{K_h(x-\cdot)}(t)}{\phi_U(-t)}\,dt. \end{aligned} \tag{2.7}$$

Thus, one may replace the conventional kernel $K_h(x-\cdot)$ by the normalized kernel $\tilde{K}_h(x,\cdot)$ in the formula (2.7) to define another deconvolution kernel by

$$\tilde{K}_h^D(x,z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itz}\frac{\phi_{\tilde{K}_h(x,\cdot)}(t)}{\phi_U(-t)}\,dt, \tag{2.8}$$

whenever the integral is defined. The integral at (2.8) is well defined for $x$ in the interior $I_0 = [2h, 1-2h]$ under some regularity conditions on $\phi_K$ and $\phi_U$, to be specified below. This follows basically from the fact that, if $x \in I_0$, then $\tilde{K}_h(x,u) = K_h(x-u)$ for all $u \in [0,1]$ so that $\tilde{K}_h^D(x,z)$ equals the classical deconvolution kernel $K_h^D(x-z)$ for all $z \in \mathbb{R}$.

However, the integral is not well defined for $x$ in boundary regions. To see this and motivate our new kernel scheme, define

$$\phi_K(t;x) = \int_0^1 e^{it(x-u)/h}\tilde{K}_h(x,u)\,du,$$

suppressing its dependence on $h$. Then, from the fact that $\phi_{\tilde{K}_h(x,\cdot)}(t) = e^{itx}\times \phi_K(-ht;x)$ we obtain

$$\tilde{K}_h^D(x,z) = \frac{1}{2\pi h}\int_{-\infty}^{\infty} e^{-it(x-z)/h}\frac{\phi_K(t;x)}{\phi_U(t/h)}\,dt \tag{2.9}$$

whenever the integral is defined. Note that $\phi_K(t;x) \equiv \phi_K(t)$ when $x \in I_0$. It can be shown that $\phi_K(t;x)/\phi_U(t/h)$ is still integrable over $t \in \mathbb{R}$ for all $x$ in the extended interior region $[h, 1-h]$ so that the integral at (2.9) is well defined. The tail property of $\phi_K(t;x)$ for $x \in [0,h)\cup(1-h,1]$ is different from that for $x \in [h,1-h]$, however. For $x \in [h,1-h]$, one may show $|\phi_K(t;x)| \sim |t|^{-\alpha}$ as $|t| \to \infty$ with $\alpha > 0$ large enough if $K(u)$ is sufficiently smooth at $u = \pm 1$. But, for $x \in [0,h)\cup(1-h,1]$ we only have $|\phi_K(t;x)| \sim |t|^{-1}$ regardless of the smoothness of $K$. For example, if $x \in [0,h)$ and $h < 1/2$, then

$$\begin{aligned} |\phi_K(t;x)| &= \left|\int_{-1}^{x/h} e^{itu}\left(\int_{u-(x/h)}^{1} K(w)\,dw\right)^{-1} K(u)\,du\right| \\ &= 2|t|^{-1}K(x/h)\bigl(1+o(1)\bigr) \end{aligned} \tag{2.10}$$

as $|t| \to \infty$. The second equality in (2.10) is obtained from integration by parts. The reason we only get the slow decaying speed $|t|^{-1}$ is that the range of integration, $[-1, x/h]$, on the right-hand side of the first equation does not cover the full support $[-1, 1]$ of $K$ and $K(u)$ does not vanish at the end point $u = x/h$. Since $\phi_U(t)$ decays to zero as $|t| \to \infty$, it means that $|\phi_K(t; x)/\phi_U(t/h)|$ is not integrable over $t \in \mathbb{R}$.

The above investigation motivates our new kernel scheme. The idea is to replace $\phi_K(t; x)$ in (2.9) by $\phi_K(t; x) \cdot \phi_K(t)$. Since $|\phi_K(t)| \sim |t|^{-\alpha}$ as $|t| \to \infty$ with $\alpha > 0$ large enough if $K(u)$ is sufficiently smooth, $|\phi_K(t; x) \cdot \phi_K(t)/\phi_U(t/h)|$ is integrable for all $x \in [0, 1]$ under suitable conditions on $K$ and the measurement error $U$. We define our new kernel function as

$$\tilde{K}_h^\star(x, z) = \frac{1}{2\pi h} \int_{-\infty}^{\infty} e^{-it(x-z)/h} \frac{\phi_K(t; x)\phi_K(t)}{\phi_U(t/h)}\, dt. \tag{2.11}$$

A very attractive property of this kernel is that it can be expressed in the form of the conventional deconvolution kernel at (2.7) with $K_h(x - \cdot)$ there being replaced by a special kernel function. Indeed, from the facts that $\phi_{\tilde{K}_h(x,\cdot)}(t) = e^{itx}\phi_K(-ht; x)$ and that $\phi_{K_h}(t) = \phi_K(ht)$, we have

$$\begin{aligned} \tilde{K}_h^\star(x, z) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{it(x-z)} \frac{\phi_K(-ht; x)\phi_K(-ht)}{\phi_U(-t)}\, dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} \frac{\phi_{\tilde{K}_h(x,\cdot)*K_h}(t)}{\phi_U(-t)}\, dt, \end{aligned} \tag{2.12}$$

where $\tilde{K}_h(x, \cdot) * K_h(u) = \int_{-\infty}^{\infty} K_h(u - v)\tilde{K}_h(x, v)\, dv$. In fact, $\tilde{K}_h(x, \cdot) * K_h(u) = K_h * K_h(x - u)$ when $x \in I_0$. Also, for all $u \in [0, 1]$,

$$\int_0^1 \tilde{K}_h(x, \cdot) * K_h(u)\, dx = \int_0^1 \int_{-\infty}^{\infty} \tilde{K}_h(x, v) K_h(u - v)\, dv\, dx = 1. \tag{2.13}$$

Thus, in view of (2.7) and (2.8), $\tilde{K}_h^\star(x, z)$ is obtained by the usual procedure of deconvoluting a special normalized kernel $\tilde{K}_h(x, \cdot) * K_h$ with normalization being applied to one of $K_h$ in $K_h * K_h$. Figure 1 depicts our deconvolution normalized kernel $\tilde{K}_h^\star(x, \cdot)$ for $x = 0$ and $x = 0.5$. It is for the choice $h = 0.1$ and for $U$ having the Laplace distribution with density $e^{-|u|}/2$.

In the following theorem, we show that the new kernel scheme $\tilde{K}_h^\star$ has both the unbiased scoring and the normalization properties. To state the theorem, we assume

(D1) $c_1(1 + |t|)^{-\beta} \le |\phi_U(t)| \le c_2(1 + |t|)^{-\beta}$ for some constants $\beta \ge 0$ and $c_1, c_2 > 0$.

We note that $\beta = 0$ includes the case where $U \equiv 0$, that is, there is no measurement error in the corresponding covariate. For the smoothness of the baseline kernel function $K$, we assume

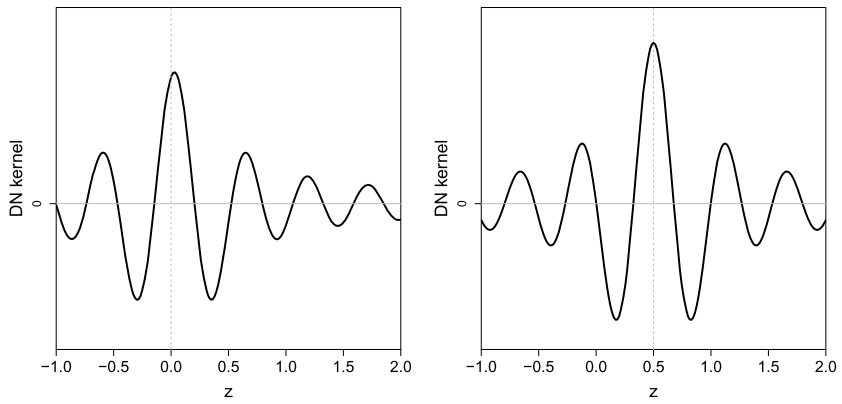


FIG. 1. *Shapes of* $\tilde{K}_h^\star(x,\cdot)$ *for* $x = 0$ (*left*) *and* $x = 0.5$ (*right*).

(K1) the baseline kernel $K$ is $\lfloor \gamma + 1 \rfloor$-times continuously differentiable for some $\gamma \geq \beta$, and $K^{(\ell)}(-1) = K^{(\ell)}(1) = 0$ for $0 \leq \ell \leq \lfloor \gamma \rfloor$, where $\lfloor \gamma \rfloor$ denotes the largest integer that is less than or equal to $\gamma$, and $K^{(\ell)}$ the $\ell$th derivative of $K$.

THEOREM 2.1. *Under the conditions* (D1) *and* (K1), *it holds that* $\tilde{K}_h^\star(x,z)$ *at* (2.11) *or* (2.12) *is well defined for all* $x \in [0,1]$ *and* $z \in \mathbb{R}$. *Furthermore*,

$$\begin{aligned} &\int_0^1 \tilde{K}_h^\star(x,z)\,dx = 1 \qquad \text{for all } z \in \mathbb{R}, \\ &\mathbb{E}\big(\tilde{K}_h^\star(x,Z) \mid X = u\big) = \tilde{K}_h(x,\cdot) * K_h(u) \qquad \text{for all } x,u \in [0,1]. \end{aligned} \tag{2.14}$$

The second property in (2.14) corresponds to what is called unbiased scoring. As we have shown in (2.13), $\tilde{K}_h(x,\cdot) * K_h(u)$ could be a normalized kernel that one can use in the standard SBF additive regression without measurement errors.

PROOF OF THEOREM 2.1. For any $a > 0$ and fixed $h > 0$, there exist constants $c, C > 0$ such that

$$\sup_{x\in[0,1]} \big|\phi_K(t;x)\big| \leq C|t|^{-1}, \qquad \big|\phi_K(t)\big| \leq C|t|^{-\lfloor\beta\rfloor-1}, \qquad \big|\phi_U(t/h)\big| \geq c|t|^{-\beta}$$

for all $|t| \geq a$. Thus, the integrand in (2.11) is bounded by $(C^2/c)|t|^{\beta-\lfloor\beta\rfloor-2}$, which is integrable over $t : |t| \geq a$ since $\beta - \lfloor\beta\rfloor - 2 < -1$. To prove the first part in (2.14), we note from (2.13) that

$$\begin{aligned} \int_0^1 \phi_{\tilde{K}_h(x,\cdot)*K_h}(t)\,dx &= \int_0^1 \int_{-\infty}^{\infty} e^{itu} \cdot \tilde{K}_h(x,\cdot) * K_h(u)\,du\,dx \\ &= \int_{-\infty}^{\infty} e^{itu}\,du = 2\pi\delta_0(t), \end{aligned}$$

where $\delta_0$ denotes the Dirac delta function at 0. Thus, we obtain from (2.12)

$$\int_0^1 \tilde{K}_h^\star(x,z)\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itz}\frac{2\pi\,\delta_0(t)}{\phi_U(-t)}\,dt = 1.$$

For the second part, if we denote the density of $U$ by $p_U$, then

$$\begin{aligned}\mathbb{E}\big(\tilde{K}_h^\star(x,Z)\mid X=u\big) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-it(u+v)}\frac{\phi_{\tilde{K}_h(x,\cdot)*K_h}(t)}{\phi_U(-t)}p_U(v)\,dt\,dv\\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itu}\phi_{\tilde{K}_h(x,\cdot)*K_h}(t)\,dt\\ &= \tilde{K}_h(x,\cdot)*K_h(u),\end{aligned}$$

the last equality is followed by the Fourier inversion theorem. □

2.2. *Smooth backfitting estimation with* $\tilde{K}_h^\star$. In the system of equations at (2.2), we estimate $p_j$, $p_{jk}$ and $m_j$ using the new kernel $\tilde{K}_h^\star$ introduced in Section 2.1. Namely, we estimate them, respectively, by

$$\begin{aligned}\hat{p}_j(x_j) &= n^{-1}\sum_{i=1}^n \tilde{K}_{h_j}^\star(x_j,Z_j^i),\\ \hat{p}_{jk}(x_j,x_k) &= n^{-1}\sum_{i=1}^n \tilde{K}_{h_j}^\star(x_j,Z_j^i)\tilde{K}_{h_k}^\star(x_k,Z_k^i),\\ \hat{m}_j(x_j) &= n^{-1}\sum_{i=1}^n \tilde{K}_{h_j}^\star(x_j,Z_j^i)Y^i/\hat{p}_j(x_j),\end{aligned} \tag{2.15}$$

where we allow the bandwidths $h_j$ for smoothing across different coordinates to be different from each other. Our SBF estimator of $(f_j : 1\le j\le d)$ is then defined to be the solution $(\hat{f}_j : 1\le j\le d)$ of the estimated system of integral equations,

$$\hat{f}_j(x_j) = \hat{m}_j(x_j) - \bar{Y} - \sum_{k\neq j}\int_0^1 \hat{f}_k(x_k)\frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)}\,dx_k, \qquad 1\le j\le d, \tag{2.16}$$

subject to the constraints

$$\int_0^1 \hat{f}_j(x_j)\hat{p}_j(x_j)\,dx_j = 0, \qquad 1\le j\le d. \tag{2.17}$$

We briefly discuss the existence of the solution of (2.16) subject to (2.17), borrowing the idea in the existing theory of SBF [20]. For this, we define a projection operator $\hat{\pi}_j : L_2(\mathbb{R}^d)\to\mathcal{H}_j$, where $\mathcal{H}_j = \{g\in L_2(\mathbb{R}^d) : g(\mathbf{x}) = g_j(x_j)$ for a univariate function $g_j\}$. Specifically,

$$\hat{\pi}_j g(x_j) = \int g(\mathbf{x})\frac{\hat{p}(\mathbf{x})}{\hat{p}_j(x_j)}\,d\mathbf{x}_{-j}, \tag{2.18}$$

where $\hat{p}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} \tilde{K}^{\star}_{h_1}(x_1, Z_1^i)\cdots\tilde{K}^{\star}_{h_d}(x_d, Z_d^i)$ and $\mathbf{x}_{-j}$ for a $d$-vector $\mathbf{x}$ equals the $(d-1)$-vector $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d)$. Then (2.16) can be rewritten as

$$\hat{f}_j = \hat{\pi}_j\left(\hat{m} - \bar{Y} - \sum_{k\neq j}^{d}\hat{f}_k\right), \qquad 1 \le j \le d,$$

where $\hat{m}(\mathbf{x}) = \hat{p}(\mathbf{x})^{-1}n^{-1}\sum_{i=1}^{n}\tilde{K}^{\star}_{h_1}(x_1, Z_1^i)\cdots\tilde{K}^{\star}_{h_d}(x_d, Z_d^i)Y^i$ is a full-dimensional estimator of $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$. In fact, $\hat{\pi}_j\hat{m} = \hat{m}_j$ due to the normalization property in (2.14). Let $\hat{T} = (I - \hat{\pi}_d)(I - \hat{\pi}_{d-1})\cdots(I - \hat{\pi}_1)$ be a linear operator that maps $\mathcal{H}$ to itself, where $\mathcal{H}$ is the space of additive functions $g$ of the form $g(\mathbf{x}) = g_1(x_1) + \cdots + g_d(x_d)$ with each $g_j \in \mathcal{H}_j$ and $I$ is the identity map. Also, for $\hat{m}_j^c = \hat{m}_j - \bar{Y}$ define $\hat{m}_{\oplus} = \hat{m}_d^c + (I - \hat{\pi}_d)\hat{m}_{d-1}^c + \cdots + (I - \hat{\pi}_d)\cdots(I - \hat{\pi}_2)\hat{m}_1^c$. Then one can deduce that the estimated additive function $\hat{f}(\mathbf{x}) = \hat{f}_1(x_1) + \cdots + \hat{f}_d(x_d)$ satisfies $\hat{f} = \hat{T}\hat{f} + \hat{m}_{\oplus}$. One can also prove that $\hat{T}$ is a contraction operator with probability tending to one, that is,

$$P\big(\|\hat{T}\| \le c\big) \to 1 \qquad \text{as } n \to \infty \tag{2.19}$$

for some constant $0 < c < 1$. The latter property follows basically from the uniform consistency of $\hat{p}_j$ and $\hat{p}_{jk}$ as estimators of $p_j$ and $p_{jk}$ on the interiors $[2h_j, 1 - 2h_j]$ and $[2h_j, 1 - 2h_j] \times [2h_k, 1 - 2h_k]$, respectively, which we establish in the next section. Indeed, if we define $T$ in the same way as $\hat{T}$ with $\hat{p}$ and $\hat{p}_j$ in its definition being replaced by the true densities $p$ and $p_j$, respectively, then one can argue that $\|T\| < 1$ using the results in Appendix A.4 of [2]. The existence and uniqueness of the solution of the equation $\hat{f} = \hat{T}\hat{f} + \hat{m}_{\oplus}$ now follows from (2.19). The solution is given by $\hat{f} = \sum_{j=0}^{\infty}\hat{T}^j\hat{m}_{\oplus}$. Furthermore, because of the constraints (2.17), each component function estimator $\hat{f}_j$ is uniquely determined.

In practice, the SBF equation is solved by an iteration algorithm. Let $\hat{f}_j^{[0]}$ denote the initial estimators and $\hat{f}_j^{[r]}$ the updates in the $r$th iteration step. Then

$$\begin{aligned}\hat{f}_j^{[r]}(x_j) = \hat{m}_j(x_j) - \bar{Y} &- \sum_{k=1}^{j-1}\int_0^1 \hat{f}_k^{[r]}(x_k)\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\,dx_k \\ &- \sum_{k=j+1}^{d}\int_0^1 \hat{f}_k^{[r-1]}(x_k)\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\,dx_k.\end{aligned} \tag{2.20}$$

If the initial estimators $\hat{f}_j^{[0]}$ are chosen so that they satisfy the constraints (2.17), then all subsequent updates $\hat{f}_j^{[r]}$ satisfy (2.17) automatically. The updating formula (2.20) can be also expressed in terms of $\hat{T}$, $\hat{m}_{\oplus}$ and $\hat{f}^{[r]}(\mathbf{x}) = \hat{f}_1^{[r]}(x_1) + \cdots + \hat{f}_d^{[r]}(x_d)$. Indeed, we have $\hat{f}^{[r]} = \hat{T}\hat{f}^{[r-1]} + \hat{m}_{\oplus}$, $r \ge 1$. Since $\hat{T}$ is a contraction with probability tending to one, $\hat{f}^{[r]} = \hat{T}^r\hat{f}^{[0]} + \sum_{j=0}^{r-1}\hat{T}^j\hat{m}_{\oplus}$ converges to $\hat{f} = \sum_{j=0}^{\infty}\hat{T}^j\hat{m}_{\oplus}$ and thus $\hat{f}_j^{[r]}$ to $\hat{f}_j$ as $r \to \infty$, with probability tending to one as $n$ tends to infinity.

**3. Theoretical properties.** In this section, we establish the theory for the SBF estimation introduced in Section 2. Without loss of generality, we assume $f_0 = 0$ and ignore $\bar{Y}$ in the discussion that follows, since $\bar{Y} = f_0 + O_p(n^{-1/2})$. For simplicity, we first consider the case where all the measurement error distributions have the same smoothness order $\beta \geq 0$ as in the condition (D1) in Section 2.1. The case where $\phi_{U_j}$ have different decaying speeds at tails will be treated at the end of this section. Some additional conditions we need are given below:

(K2) The joint density $p$ of $\mathbf{X}$ is bounded away from zero and infinity on $[0, 1]^d$ and partially continuously differentiable, and $p_j$ and $p_{jk}$ are also (partially) continuously differentiable.

(K3) $\mathbb{E}|Y|^\alpha < \infty$ for some $\alpha > 5/2$ and $\mathbb{E}(Y^2 \mid X_j = \cdot)$ are continuous on $[0, 1]$.

(K4) $f_j$ are twice continuously differentiable on $[0, 1]$.

(D2) $|t^\beta \phi_{U_j}(t)| \to c_\beta$ for some $c_\beta \neq 0$ and $|t^{\beta+1}\phi'_{U_j}(t)| = O(1)$ as $|t| \to \infty$, where $\beta$ is the nonnegative constant in (D1).

(D3) $\int |t^\gamma \phi_K(t)|\, dt < \infty$, where $\gamma$ is the nonnegative constant in (K1).

The above conditions are typically assumed in kernel smoothing theory and in standard deconvolution problems; see [20], [16] and [6], among others. The condition (D2) enables us to obtain an inequality enveloping $\tilde{K}_h^\star$ (Lemma 5.1). The condition (D3) is used to get uniform convergence rates of various $\tilde{K}_h^\star$-weighted quantities. The latter condition was also used [22], among others, for the uniform consistency of the deconvolution kernel density and regression estimators.

3.1. *Consistency of projection operators.* Define $\pi_j$ in the same way as $\hat{\pi}_j$ at (2.18) with $\hat{p}$ and $\hat{p}_j$ being replaced by $p$ and $p_j$, respectively. We establish the consistency of the projection operators $\hat{\pi}_j$ as estimators of $\pi_j$. This is a fundamental property that our SBF theory is built on. The consistency entails that $\|\hat{T} - T\|$ converges to zero in probability, so that $\hat{T}$ is a contraction with probability tending to one. It also implies that the solution of the backfitting equation (2.16) exists and unique and that the backfitting iteration at (2.20) converges to the solution of (2.16). The consistency of $\hat{\pi}_j$ relies on the following proposition.

PROPOSITION 3.1. *Under the conditions* (K1), (K2) *and* (D1)–(D3), *it follows that*

$$\begin{aligned} \sup_{x_j \in [0,1]} |\hat{p}_j(x_j) - \mathbb{E}\hat{p}_j(x_j)| &= O_p\left(\sqrt{\frac{\log n}{n h_j^{1+2\beta}}}\right), \\ \sup_{(x_j, x_k) \in [0,1]^2} |\hat{p}_{jk}(x_j, x_k) - \mathbb{E}\hat{p}_{jk}(x_j, x_k)| &= O_p\left(\sqrt{\frac{\log n}{n h_j^{1+2\beta} h_k^{1+2\beta}}}\right). \end{aligned} \tag{3.1}$$

*It also holds that*

$$\begin{aligned}\sup_{x_j\in I_{0j}}\bigl|\mathbb{E}\hat{p}_j(x_j)-p_j(x_j)\bigr|&\le C_j h_j,\\ \sup_{x_j\in I_{0j},x_k\in I_{0k}}\bigl|\mathbb{E}\hat{p}_{jk}(x_j,x_k)-p_{jk}(x_j,x_k)\bigr|&\le C_{jk}(h_j+h_k)\end{aligned}\tag{3.2}$$

*for some constants $C_j$ and $C_{jk}$, where $I_{0j}=[2h_j,1-2h_j]$.*

One may prove (3.1) in the above proposition using Lemma 5.1, the order of magnitude of $\mathbb{E}|\tilde{K}_h^\star(x,Z)|^2$ in the proof of Theorem 3.3, the unbiased scoring property in (2.14) and an exponential inequality in the standard theory of kernel smoothing. In the case of standard deconvolution kernel, similar results are provided by [22], for example. The second result (3.2) in Proposition 3.1 is due to the unbiased scoring property proved in Theorem 2.1. If we assume second (partial) derivatives of $p_j$ and $p_{jk}$, then the bias orders at (3.2) would be $h_j^2$ and $h_j^2+h_k^2$, respectively, instead of $h_j$ and $h_j+h_k$. In that case, we can achieve the one- and two-dimensional optimal rates $n^{-2/(5+2\beta)}$ and $n^{-2/(6+4\beta)}$ for the estimation of $p_j$ and $p_{jk}$, respectively, by choosing the bandwidth orders $h_j\asymp n^{-1/(5+2\beta)}$ for $\hat{p}_j$ and $h_j,h_k\asymp n^{-1/(6+4\beta)}$ for $\hat{p}_{jk}$.

From Proposition 3.1 it holds that $\inf_{x_j\in[0,1]}\hat{p}_j(x_j)>c$ with probability tending to one for some constant $c>0$. These together with (3.1), (3.2) and the facts that $[0,1]\cap I_{0j}^c$ and $[0,1]^2\cap(I_{0j}\times I_{0k})^c$ have Lebesgue measures $4h_j$ and $4(h_j+h_k)$, respectively, give the following proposition that demonstrates the consistency of $\hat{\pi}_j$. Let $\|\cdot\|_2$ denote the $L_2$-norm defined by $\|g\|_2^2=\int_0^1 g(u)^2\,du$.

PROPOSITION 3.2. *Assume that* (K1), (K2), (D1)–(D3) *hold and that* $n(h_j\times h_k)^{1+2\beta}/\log n\to\infty$ *as* $n\to\infty$. *Then*

$$\|\hat{\pi}_j-\pi_j\|\equiv\sup\bigl\{\|\hat{\pi}_j g-\pi_j g\|_2: g\in L_2\bigl(\mathbb{R}^d\bigr),\|g\|_2=1\bigr\}=o_p(1).$$

3.2. *Asymptotic properties of component estimators.* We now discuss the statistical properties of the component function estimators $\hat{f}_j$. The main innovation in the development of our theory, beyond the ones in standard SBF additive regression and in one-dimensional deconvolution, is a careful decomposition of the errors in $\hat{m}_j$. The decomposition is based on two attributes of the effects of the measurement errors $U_j^i$, one attribute contributing to the biases of $\hat{f}_j$ and the other to the variances.

From the equation (2.16) and the fact $\hat{\pi}_j f_j=f_j$ due to the normalization property in (2.14), it follows that

$$\hat{f}_j-f_j=\hat{m}_j-\hat{\pi}_j(f_1+\cdots+f_d)-\sum_{k\ne j}^{d}\hat{\pi}_j(\hat{f}_k-f_k).\tag{3.3}$$

Note that $f_1(x_1) + \cdots + f_d(x_d) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$, and thus $\pi_j(f_1 + \cdots + f_d) = m_j$. This means that $\hat{m}_j - \hat{\pi}_j(f_1 + \cdots + f_d)$ in (3.3) tell how $\hat{m}_j$ are accurate as estimators of $m_j$ neglecting the errors in $\hat{\pi}_j$ as estimators of $\pi_j$. We define

$$\hat{\delta}_j = \hat{m}_j - \hat{\pi}_j(f_1 + \cdots + f_d)$$

and $\hat{\delta}_\oplus = \hat{\delta}_d + (I - \hat{\pi}_d)\hat{\delta}_{d-1} + (I - \hat{\pi}_d)(I - \hat{\pi}_{d-1})\hat{\delta}_{d-2} + \cdots + (I - \hat{\pi}_d)\cdots(I - \hat{\pi}_2)\hat{\delta}_1$. As in Section 2.2, we get from (3.3) that

$$\hat{f} - f = \hat{T}(\hat{f} - f) + \hat{\delta}_\oplus = \sum_{j=0}^{\infty} \hat{T}^j \hat{\delta}_\oplus. \tag{3.4}$$

The expression at (3.4) reveals that the errors $\hat{\delta}_j$ of $\hat{m}_j$ are propagated into the SBF estimation error $\hat{f} - f$ through the backfitting operation.

We decompose $\hat{\delta}_j$ into five terms. Let $\varepsilon^i = Y^i - \mathbb{E}(Y^i \mid \mathbf{X}^i)$ and define

$$V_{nk}^i = \int_0^1 \tilde{K}_{h_k}^\star(x_k, Z_k^i)\big(f_k(X_k^i) - f_k(x_k)\big)\,dx_k.$$

Then it holds that $\hat{\delta}_j = \hat{\delta}_j^A + \hat{\delta}_j^B + \hat{\delta}_j^C$, where

$$\begin{aligned}
\hat{\delta}_j^A(x_j) &= \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n \tilde{K}_{h_j}^\star(x_j, Z_j^i)\varepsilon^i,\\
\hat{\delta}_j^B(x_j) &= \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n \tilde{K}_{h_j}^\star(x_j, Z_j^i)\big(f_j(X_j^i) - f_j(x_j)\big),\\
\hat{\delta}_j^C(x_j) &= \sum_{k\neq j}^d \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n \tilde{K}_{h_j}^\star(x_j, Z_j^i) V_{nk}^i.
\end{aligned}$$

We further decompose each of $\hat{\delta}_j^B$ and $\hat{\delta}_j^C$ into two terms: $\hat{\delta}_j^*(x_j) = \hat{\delta}_{1j}^*(x_j) + \hat{\delta}_{2j}^*(x_j)$ for $* = B$ or $C$, where

$$\begin{aligned}
\hat{\delta}_{1j}^B(x_j) &= \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n \mathbb{E}\big(\tilde{K}_{h_j}^\star(x_j, Z_j^i) \mid X_j\big)\big(f_j(X_j^i) - f_j(x_j)\big),\\
\hat{\delta}_{2j}^B(x_j) &= \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n \big[\tilde{K}_{h_j}^\star(x_j, Z_j^i) - \mathbb{E}\big(\tilde{K}_{h_j}^\star(x_j, Z_j^i) \mid X_j\big)\big]\\
&\quad \times \big(f_j(X_j^i) - f_j(x_j)\big),\\
\hat{\delta}_{1j}^C(x_j) &= \sum_{k\neq j}^d \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n \tilde{K}_{h_j}^\star(x_j, Z_j^i)\mathbb{E}\big(V_{nk}^i \mid X_k^i\big),
\end{aligned} \tag{3.5}$$

$$\hat{\delta}_{2j}^{C}(x_j) = \sum_{k\neq j}^{d} \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^{n} \tilde{K}_{h_j}^{\star}(x_j, Z_j^i)\big[V_{nk}^i - \mathbb{E}(V_{nk}^i \mid X_k^i)\big].$$

Recall that the unbiased scoring property in (2.14) entails $\mathbb{E}(\tilde{K}_{h_j}^{\star}(x_j, Z_j^i) \mid X_j) = \tilde{K}_{h_j}(x_j, \cdot) * K_{h_j}(X_j^i)$ and it equals $K_h * K_h(x_j - X_j^i) = (K*K)_{h_j}(x_j - X_j^i)$ when $x_j \in I_{0j} = [2h_j, 1-2h_j]$.

In the next two theorems, we demonstrate how the above five terms in the decomposition of the errors $\delta_j$ are propagated through the backfitting operation. Let

$$\tau_{n,j}(\beta) = \begin{cases} 1, & \beta < 1/2, \\ \sqrt{\log h_j^{-1}}, & \beta = 1/2, \\ h_j^{1/2-\beta}, & \beta > 1/2. \end{cases}$$

Below we first present the orders of magnitude of the three stochastic terms $\hat{\delta}_j^A$, $\hat{\delta}_{2j}^B$ and $\hat{\delta}_{2j}^C$. We find that $\hat{\delta}_j^A(x_j)$ and $\hat{\delta}_{2j}^B(x_j)$ have a typical stochastic error rate in one-dimensional deconvolution problems. We note that $\hat{\delta}_{2j}^B(x_j)$ vanishes when $U_j^i \equiv 0$ since $\tilde{K}_h^{\star}(x,u) = \tilde{K}_{h_j}(x_j, \cdot) * K_{h_j}(u)$ in that case. The fifth term $\hat{\delta}_{2j}^C(x_j)$, which is negligible in the case where $U_j^i \equiv 0$, dominates $\hat{\delta}_j^A(x_j)$ and $\hat{\delta}_{2j}^B(x_j)$ depending on the smoothness of the measurement error distributions.

THEOREM 3.3. *Assume that* (K1)–(K4) *and* (D1)–(D3) *hold. Then, uniformly for* $x_j \in [0,1]$,

$$\hat{\delta}_j^A(x_j) = O_p\Bigg(\sqrt{\frac{\log n}{n h_j^{1+2\beta}}}\Bigg), \qquad \hat{\delta}_{2j}^B(x_j) = O_p\Bigg(\sqrt{\frac{\log n}{n h_j^{1+2\beta}}}\Bigg),$$

$$\hat{\delta}_{2j}^C(x_j) = O_p\Bigg(\sqrt{\frac{\log n}{n h_j^{1+2\beta}}} \sum_{k\neq j} \tau_{n,k}(\beta)\Bigg).$$

The above theorem remains to hold even if (K4) is replaced by that $f_j$ are continuously differentiable. In the propagation of $\hat{\delta}_j^A$, $\hat{\delta}_{2j}^B$ and $\hat{\delta}_{2j}^C$ through the backfitting operation, their effects do not spread to other component function estimators $\hat{f}_k$, $k \neq j$, to the first-order, as is demonstrated in the following theorem. However, the other two terms, $\hat{\delta}_{1j}^B$ and $\hat{\delta}_{1j}^C$, affect the biases of other component function estimators through the backfitting operation and their effects are shaped into some deterministic bias terms as given in the theorem. These bias terms also appear in the SBF estimation without measurement errors, that is, with $U_j^i \equiv 0$.

To state the theorem, let $\mu_{\ell,j}(x_j) = h_j^{-\ell} \int_0^1 (v_j - x_j)^{\ell} \tilde{K}_{h_j}(x_j, \cdot) * K_{h_j}(v_j)\, dv_j$, where $\tilde{K}_{h_j}(\cdot,\cdot)$ is the normalized kernel defined in (2.5). Note that $\mu_{1,j}(x_j) \equiv 0$

and $\mu_{2,j}(x_j) \equiv \mu_2$ for $x \in I_{0j}$, where $\mu_2 = \int u^2 K * K(u)\,du$. Define $(\Delta_j : 1 \le j \le d)$ to be the minimizer of

$$\int_{[0,1]^d} \left( \Delta_{n,f}(\mathbf{x}) - \sum_{j=1}^{d} h_j^2 \Delta_j(x_j) \right)^2 p(\mathbf{x})\,d\mathbf{x}$$

subject to the constraints

$$\int_0^1 \Delta_j(x_j) p_j(x_j)\,dx_j = \mu_2 \int_0^1 f_j'(x_j) p_j'(x_j)\,dx_j, \qquad 1 \le j \le d, \tag{3.6}$$

where $\Delta_{n,f}(\mathbf{x}) = \mu_2 \sum_{j=1}^{d} h_j^2 f_j'(x_j) p_j^{(1)}(\mathbf{x})/p(\mathbf{x})$ is a full-dimensional deterministic function and $p_j^{(1)}(\mathbf{x}) = \partial p(\mathbf{x})/\partial x_j$. In fact, $(h_j^2 \Delta_j : 1 \le j \le d)$ is the solution of the system of equations

$$h_j^2 \Delta_j = \pi_j \left( \Delta_{n,f} - \sum_{k \ne j} h_k^2 \Delta_k \right), \qquad 1 \le j \le d, \tag{3.7}$$

subject to (3.6). In the following theorem, we assume that all bandwidth $h_j$ are of the same order of magnitude, $h_j \asymp h \asymp n^{-c}$ for some $c > 0$. Let $\tau_n$ be defined in the same way as $\tau_{n,j}$ with $h_j$ being replaced by a common bandwidth order. Also, let $r_{n,j}$ denote generic stochastic terms such that

$$\begin{aligned} \sup_{x_j \in I_{0j}} |r_{n,j}(x_j)| &= o_P\left( h^2 + \sqrt{\frac{\log n}{nh^{1+2\beta}}} \cdot \tau_n(\beta) \right), \\ \sup_{x_j \in [0,1]} |r_{n,j}(x_j)| &= O_P(h^2) + o_P\left( \sqrt{\frac{\log n}{nh^{1+2\beta}}} \cdot \tau_n(\beta) \right). \end{aligned} \tag{3.8}$$

THEOREM 3.4. *Under the conditions of Theorem 3.3, if $nh^{3+4\beta}/\log n$ is bounded away from zero, then*

$$\begin{aligned} \hat{f}_j(x_j) = f_j(x_j) &+ \hat{\delta}_j^A(x_j) + \hat{\delta}_{2j}^B(x_j) + \hat{\delta}_{2j}^C(x_j) \\ &+ \frac{\mu_{1,j}(x_j)}{\mu_{0,j}(x_j)} f_j'(x_j) h_j + \frac{1}{2}\mu_2 f_j''(x_j) h_j^2 + \Delta_j(x_j) h_j^2 + r_{n,j}(x_j), \\ &1 \le j \le d. \end{aligned}$$

It would be interesting to compare Theorem 3.4 with a stochastic expansion of the oracle estimator of $f_j$. The oracle estimator of $f_j$ is the one-dimensional deconvolution kernel estimator that utilizes the knowledge of all other component functions $f_k$, $k \ne j$. It is given by

$$\hat{f}_j^{\mathrm{ora}}(x_j) = \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^{n} \tilde{K}_{h_j}^{\star}(x_j, Z_j^i) \left[ Y^i - \sum_{k \ne j} f_k(X_k^i) \right].$$

Using the standard theory of kernel smoothing and that of nonparametric deconvolution, we may prove that

$$\hat{f}_j^{\mathrm{ora}}(x_j) = f_j(x_j) + \hat{\delta}_j^A(x_j) + \hat{\delta}_{2j}^B(x_j) + \frac{\mu_{1,j}(x_j)}{\mu_{0,j}(x_j)} f_j'(x_j)h_j + \frac{1}{2}\mu_2 f_j''(x_j)h_j^2 + o_P\left(h^2 + \sqrt{\frac{\log n}{nh^{1+2\beta}}}\right).$$

This demonstrates that our SBF estimators $\hat{f}_j$ have two additional terms, $\Delta_j(x_j)h_j^2$ and $\hat{\delta}_{2j}^C(x_j)$, in comparison with the expansion of the oracle estimators, both of which are from the backfitting operation.

According to Theorem 3.3, $\hat{\delta}_{2j}^C(x_j)$ has the same order of magnitude as $\hat{\delta}_j^A(x_j)$ and $\hat{\delta}_{2j}^B(x_j)$ in case $\beta < 1/2$, while it dominates them when $\beta \geq 1/2$. By combining Theorems 3.3 and Theorem 3.4, we get the following corollary.

COROLLARY 3.5. *Assume the conditions of Theorem* 3.3. (i) *When* $\beta < 1/2$,

$$\sup_{x_j \in I_{0j}} |\hat{f}_j(x_j) - f_j(x_j)| = O_P\big(n^{-2/(5+2\beta)}\sqrt{\log n}\big),$$

$$\sup_{x_j \in I_{0j}^c} |\hat{f}_j(x_j) - f_j(x_j)| = O_P\big(n^{-1/(5+2\beta)}\big)$$

*by choosing* $h_j \asymp n^{-1/(5+2\beta)}$. (ii) *When* $\beta = 1/2$,

$$\sup_{x_j \in I_{0j}} |\hat{f}_j(x_j) - f_j(x_j)| = O_P\big(n^{-1/3}\log n\big),$$

$$\sup_{x_j \in I_{0j}^c} |\hat{f}_j(x_j) - f_j(x_j)| = O_P\big(n^{-1/6}\big)$$

*by choosing* $h_j \asymp n^{-1/6}$. (iii) *When* $\beta > 1/2$,

$$\sup_{x_j \in I_{0j}} |\hat{f}_j(x_j) - f_j(x_j)| = O_P\big(n^{-1/(2+2\beta)}\sqrt{\log n}\big),$$

$$\sup_{x_j \in I_{0j}^c} |\hat{f}_j(x_j) - f_j(x_j)| = O_P\big(n^{-1/(4+4\beta)}\big)$$

*by choosing* $h_j \asymp n^{-1/(4+4\beta)}$.

For the above corollary, we have used the fact that $\mu_{1,j}(x_j) = 0$ for $x_j \in I_{0j}$. According to Corollary 3.5, $\hat{f}_j$ achieve the optimal rate that one can achieve in one-dimensional deconvolution problems [8], in the case $\beta < 1/2$. It is interesting to notice that one also gets different convergence rates, depending on whether $0 \leq \beta < 1/2$, $\beta = 1/2$ or $\beta > 1/2$, in the estimation of distribution function

and quantiles with contaminated data; see [11] and [5] for details. Although the rates when $\beta > 1/2$ are slower than the corresponding rates when $\beta \le 1/2$, they are much faster than the optimal $d$-variate rates. As shown by [9], the optimal rates in $d$-variate deconvolution problems are $n^{-2/(4+d+d2\beta)}$ in the interior and $n^{-1/(4+d+2d\beta)}$ on the boundary.

We close this section by briefly commenting on the case where not all $\beta_j$ are the same. In the latter case, in the statement of Theorem 3.3, we only need to replace $h_j^{1+2\beta}$ by $h_j^{1+2\beta_j}$ and $\tau_{n,k}(\beta)$ by $\tau_{n,k}(\beta_k)$. For an analog of Theorem 3.4, the remainders $r_{n,j}$ at (3.8) in the expansions of $\hat{f}_j$ now satisfy

$$\sup_{x_j \in I_{0j}} |r_{n,j}(x_j)| = o_P\Bigg(h^2 + \sqrt{\frac{\log n}{nh^{1+2\beta_*}}} \cdot \tau_n(\beta_*)\Bigg),$$

$$\sup_{x_j \in [0,1]} |r_{n,j}(x_j)| = O_P(h^2) + o_P\Bigg(\sqrt{\frac{\log n}{nh^{1+2\beta_*}}} \cdot \tau_n(\beta_*)\Bigg),$$

where $\beta_* = \max_{1\le j\le d} \beta_j$. Here, we present the slowest rate of convergence of $\hat{f}_j$ among all components. The slowest rate actually determines the rate of the convergence of $\hat{f} = \hat{f}_1 + \cdots + \hat{f}_d$ as an estimator of the regression function $f = f_1 + \cdots + f_d$. The slowest rate corresponds to the component that has the largest $\beta_j = \beta_*$. Let $j_* = \arg\max_{1\le j\le d} \beta_j$. Then, from the versions of Theorems 3.3 and 3.4 for different $\beta_j$ it follows that $\hat{f}_{j_*}(u) - f_{j_*}(u) = O_p(n^{-2/(5+2\beta_*)}\sqrt{\log n})$ and $O_p(n^{-1/3}\log n)$, respectively, when $\beta_* < 1/2$ and $\beta_* = 1/2$, uniformly for $u$ in the interior with the bandwidths $h_j \asymp n^{-1/(5+2\beta_*)}$. In case $\beta_* > 1/2$, we have $\hat{f}_{j_*}(u) - f_{j_*}(u) = O_p(n^{-1/(2+2\beta_*)}\sqrt{\log n})$ uniformly for $u$ in the interior with the bandwidths $h_j \asymp n^{-1/(4+4\beta_*)}$. The rates on the boundary are $O_p(n^{-1/(5+2\beta_*)})$ and $O_p(n^{-1/(4+4\beta_*)})$, respectively, when $\beta_* \le 1/2$ and $\beta_* > 1/2$.

**4. Finite sample performance.** We first examine how the use of our deconvolution normalization kernels improves the estimation accuracy in comparison with naive applications of the standard smooth backfitting technique that ignores the presence of measurement errors. Below in this section, we refer to them as D-SBF and N-SBF, respectively. For the D-SBF estimators, we used $\tilde{K}_h^\star$ introduced at (2.11), and for the N-SBF the conventional normalized kernel $\tilde{K}_h$ at (2.5). For both, we used $\hat{f}_j^{[0]} \equiv 0$ as the initial estimators in the updating equations at (2.20).

In our simulation, the response variable $Y$ was generated by

$$Y = g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4) + \varepsilon, \tag{4.1}$$

where $X_j$ and $\varepsilon \sim N(0, 0.25^2)$ are independent. We set $X_j = \Phi(W_j)$ where $\Phi$ is the cumulative distribution function of $N(0,1)$ and $\mathbf{W} = (W_1, \ldots, W_4)^\top$ has a multivariate normal distribution with mean zero and covariance $\Sigma$ with $\Sigma_{j,k} =$

$(1/2)^{|j-k|}$, $1 \le j,k \le 4$. We considered $g_1(x) = -2x$, $g_2(x) = \arctan(2\pi(x - 0.5))$, $g_3(x) = \mathrm{pdf}_{\Gamma(4,0.1)}(x)$ and $g_4(x) = \cos(3\pi x)$, defined on $[0,1]$, where $\Gamma(4, 0.1)$ stands for a Gamma distribution with shape and scale parameters 4 and 0.1, respectively. The function $g_3$ is asymmetric and $g_4$ is periodic. As for the distributions of the measurement errors $U_j$, we took a Laplace and a double gamma difference (DGD) distributions. Laplace distributions are often adopted in the literature as a representative of ordinary smooth distributions, and we note that $\beta$ in (D1) equals 2 for Laplace distributions. A DGD distribution [1] is obtained by taking difference of two independent Gamma random variables with the same scale parameter. Its smoothness equals $\beta$ if it is the difference of two $\Gamma(\beta/2, \theta)$ random variables. DGD distributions are known to form a family of symmetric variance gamma distributions; see [14]. We used a DGD distribution with $\beta = 0.4$. We generated $U_j^i$ independently across components $j$ as well as across subjects $i$, from a Laplace and a DGD distribution. The noise-to-signal ratio (NSR), defined by $\mathbb{V}ar(U_j)/\mathbb{V}ar(X_j)$, was set to be 0.1 or 0.2, and the scale parameters of the Laplace or DGD distributions were chosen to satisfy a given NSR for each simulation setting. We estimated the centered component functions $f_j = g_j - \int_0^1 g_j$ using a set of generated pseudo samples $\mathcal{X}_n = \{(Y^i, \mathbf{Z}^i) : 1 \le i \le n\}$ of sizes $n = 400$ and 1000 according to the model (4.1).

As means of comparison, we computed the mean integrated squared errors (MISE), the integrated squared bias (ISB) and the integrated variance (IV) for each component. Let $\mathcal{X}_n^{(m)}$ be the $m$th Monte Carlo sample. For $\hat{f}_j$ denoting either our D-SBF or the conventional N-SBF estimators and for $\hat{f}_{j,m}$ the estimates constructed from $\mathcal{X}_n^{(m)}$, we approximated the MISEs of $\hat{f}_j$ by the formula

$$\mathrm{MISE}(\hat{f}_j) \approx \frac{1}{M} \sum_{m=1}^{M} \int_0^1 \bigl(\hat{f}_{j,m}(x_j) - f_j(x_j)\bigr)^2 \, dx_j, \tag{4.2}$$

where we took $M = 200$. We estimated the ISB and the IV similarly. We used the Epanechnikov kernel as the baseline kernel function $K$. To the effect of comparing their respective optimal performances in regard to bandwidth choice, we chose the respective optimal bandwidths for the D-SBF and N-SBF and used them in computing the criterion values. For this, we generated $M_0$ samples separately from $\mathcal{X}_n^{(1)}, \ldots, \mathcal{X}_n^{(M)}$ and computed the approximate MISE values according to the formula (4.2) using the $M_0$ samples on the grid of bandwidth $\{0.05 + 0.01i : 0 \le i \le 15\}$. We tried several choices of $M_0$ and found that those $M_0 \ge 20$ gave the same optimal bandwidths. For instance, in the case of $\beta = 2$ and $\mathrm{NSR} = 0.1$, we found that for the N-SBF the best bandwidths were $h = 0.08, 0.06$ for $n = 400, 1000$, respectively, while for the D-SBF they were $h = 0.12, 0.09$, respectively. We noted that the optimal bandwidths were larger for larger $\beta$ or larger NSR, which is supported by our theoretical results in Section 3.

The results based on these optimal bandwidths are presented in Tables 1–4 and Figures 2 and 3. From the tables, we clearly see that the D-SBF is better than

TABLE 1

*Double gamma difference (DGD) measurement error distribution with* $\beta = 0.4$ *and* $\mathrm{NSR}(\sigma_U^2/\sigma_X^2) = 0.2$, *based on* 200 *MC samples*

| Sample size & criterion | | Naive SBF | | | | Deconvolution SBF | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $j=1$ | $j=2$ | $j=3$ | $j=4$ | $j=1$ | $j=2$ | $j=3$ | $j=4$ |
| 400 | MISE | 0.056 | 0.095 | 0.059 | 0.057 | 0.029 | 0.033 | 0.037 | 0.031 |
| | ISB | 0.043 | 0.082 | 0.048 | 0.044 | 0.022 | 0.022 | 0.019 | 0.013 |
| | IV | 0.013 | 0.013 | 0.011 | 0.013 | 0.007 | 0.011 | 0.018 | 0.018 |
| 1000 | MISE | 0.050 | 0.088 | 0.044 | 0.047 | 0.015 | 0.019 | 0.021 | 0.019 |
| | ISB | 0.044 | 0.080 | 0.038 | 0.040 | 0.008 | 0.011 | 0.007 | 0.007 |
| | IV | 0.006 | 0.008 | 0.006 | 0.007 | 0.007 | 0.008 | 0.014 | 0.012 |

TABLE 2

*Double gamma difference (DGD) measurement error distribution with* $\beta = 0.4$ *and* $\mathrm{NSR}(\sigma_U^2/\sigma_X^2) = 0.1$, *based on* 200 *MC samples*

| Sample size & criterion | | Naive SBF | | | | Deconvolution SBF | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $j=1$ | $j=2$ | $j=3$ | $j=4$ | $j=1$ | $j=2$ | $j=3$ | $j=4$ |
| 400 | MISE | 0.049 | 0.080 | 0.046 | 0.042 | 0.028 | 0.031 | 0.032 | 0.028 |
| | ISB | 0.038 | 0.069 | 0.037 | 0.032 | 0.023 | 0.023 | 0.019 | 0.014 |
| | IV | 0.011 | 0.011 | 0.009 | 0.010 | 0.005 | 0.008 | 0.013 | 0.014 |
| 1000 | MISE | 0.043 | 0.070 | 0.030 | 0.031 | 0.012 | 0.016 | 0.016 | 0.014 |
| | ISB | 0.038 | 0.064 | 0.026 | 0.026 | 0.008 | 0.011 | 0.007 | 0.007 |
| | IV | 0.005 | 0.006 | 0.004 | 0.005 | 0.004 | 0.005 | 0.009 | 0.007 |

TABLE 3

*Laplace measurement error distribution with* $\beta = 2$ *and* $\mathrm{NSR} = 0.2$, *based on* 200 *MC samples*

| Sample size & criterion | | Naive SBF | | | | Deconvolution SBF | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $j=1$ | $j=2$ | $j=3$ | $j=4$ | $j=1$ | $j=2$ | $j=3$ | $j=4$ |
| 400 | MISE | 0.064 | 0.114 | 0.115 | 0.133 | 0.047 | 0.047 | 0.094 | 0.058 |
| | ISB | 0.046 | 0.094 | 0.098 | 0.113 | 0.024 | 0.020 | 0.017 | 0.009 |
| | IV | 0.018 | 0.020 | 0.017 | 0.020 | 0.023 | 0.027 | 0.077 | 0.049 |
| 1000 | MISE | 0.056 | 0.094 | 0.092 | 0.114 | 0.021 | 0.025 | 0.046 | 0.033 |
| | ISB | 0.045 | 0.084 | 0.083 | 0.104 | 0.005 | 0.005 | 0.019 | 0.012 |
| | IV | 0.011 | 0.010 | 0.009 | 0.010 | 0.016 | 0.020 | 0.027 | 0.021 |

TABLE 4
*Laplace measurement error distribution with* $\beta = 2$ *and* NSR $= 0.1$, *based on* 200 *MC samples*

| Sample size & criterion | | Naive SBF | | | | Deconvolution SBF | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $j=1$ | $j=2$ | $j=3$ | $j=4$ | $j=1$ | $j=2$ | $j=3$ | $j=4$ |
| 400 | MISE | 0.048 | 0.078 | 0.068 | 0.074 | 0.024 | 0.027 | 0.062 | 0.038 |
| | ISB | 0.035 | 0.064 | 0.055 | 0.060 | 0.011 | 0.011 | 0.039 | 0.012 |
| | IV | 0.013 | 0.014 | 0.013 | 0.014 | 0.013 | 0.016 | 0.023 | 0.026 |
| 1000 | MISE | 0.044 | 0.069 | 0.055 | 0.063 | 0.016 | 0.021 | 0.033 | 0.025 |
| | ISB | 0.037 | 0.062 | 0.049 | 0.056 | 0.007 | 0.010 | 0.008 | 0.007 |
| | IV | 0.007 | 0.007 | 0.006 | 0.007 | 0.009 | 0.011 | 0.025 | 0.018 |

the naive SBF in all cases in terms of the MISE, and that the MISE of the D-SBF decreases faster than the naive SBF as the sample size increases. The smaller values of the MISE for the D-SBF are due to the much smaller values of the ISB. This demonstrates that the deconvolution procedure done by our new kernel $\tilde{K}_h^\star$ at (2.11) is successful in correcting the bias due to measurement errors. The biases of the N-SBF do not decrease substantially as the sample size increases. One further thing to note is that the margin of the improvement by the D-SBF is larger for smaller $\beta$. Figures 2 and 3 depict the bias and variance curves of the D-SBF and N-SBF estimators. Here, we only report the results for the case $\beta = 2.0$ and NSR $=$ 0.1 since those for other cases of $\beta$ and NSR gave similar lessons. The left panels of the figures suggest visually what we observe in the tables for the bias properties. The biases of the D-SBF estimators are improved as the sample size increases, but this is not the case with the N-SBF.

What if the measurement error distribution is misspecified in the construction of $\tilde{K}_h^\star$? Under the Laplace measurement error setting of Table 4, we used four DGD distributions with different smoothness $\beta$ but the same NSR $= 0.1$, to construct $\tilde{K}_h^\star$. We found $\mathrm{MISE}(\hat{f}_1) + \cdots + \mathrm{MISE}(\hat{f}_4) = 0.207, 0.191, 0.161, 0.187$ for $\beta = 0.5, 1.0, 4.0, 8.0$, respectively, when $n = 400$. Note that the corresponding sum of the four MISEs in the case the error distribution is correctly specified (the case $n = 400$ in Table 4) equals 0.151. The results suggest that our deconvolution method is not much sensitive to the misspecification of the error distribution. A similar phenomenon was also observed in [10]. We also examined what happens if we use the deconvolution kernel when the covariates are not actually contaminated. For this, we chose a Laplace and a DGD error distribution with $\beta = 2.0$ and 0.4, respectively, and used them in constructing $\tilde{K}_h^\star$. We found $\mathrm{MISE}(\hat{f}_1) + \cdots + \mathrm{MISE}(\hat{f}_4) = 0.094$ and 0.084 for the Laplace and DGD deconvolution kernels, respectively, while the N-SBF (no deconvolution) gave 0.047.

It is also of interest to see whether our proposed method still works for asymmetric measurement errors. For this, we tried a Gamma measurement error with

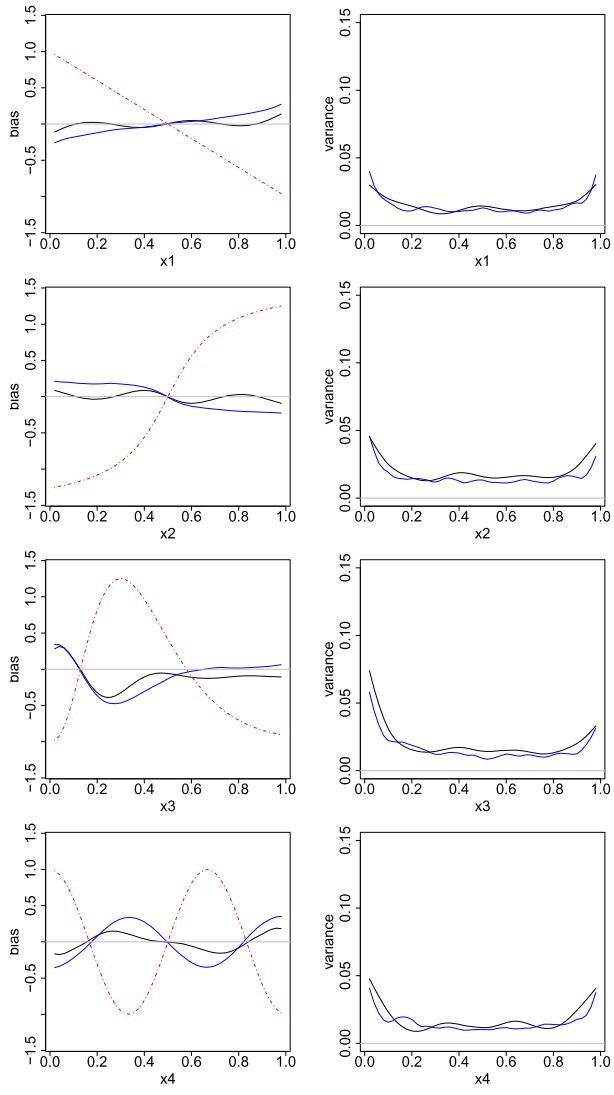


FIG. 2. *The case* $n = 400$, $\beta = 2.0$ *and* NSR $= 0.1$. *From the top, the bias (left) and variance curves (right) of the estimators of* $f_1$, $f_2$, $f_3$ *and* $f_4$, *based on* 200 *MC replications. The black curves are for the D-SBF and the blue for the N-SBF, with the true functions* $f_j$ *depicted as dot-dashed red curves on the left panels.*

$\beta = 1$ and NSR $= 0.1$. The results for the sample size $n = 400$ are contained in Table 5 and Figure 4. They demonstrate that the use of our proposed deconvolution kernel $\tilde{K}_h^\star$ corrects effectively the bias owing to the one-sided measurement error as well. An interesting phenomenon in Figure 4 that deserves particular attention and is more clearly visible in the two bottom panels is that the N-SBF curves are

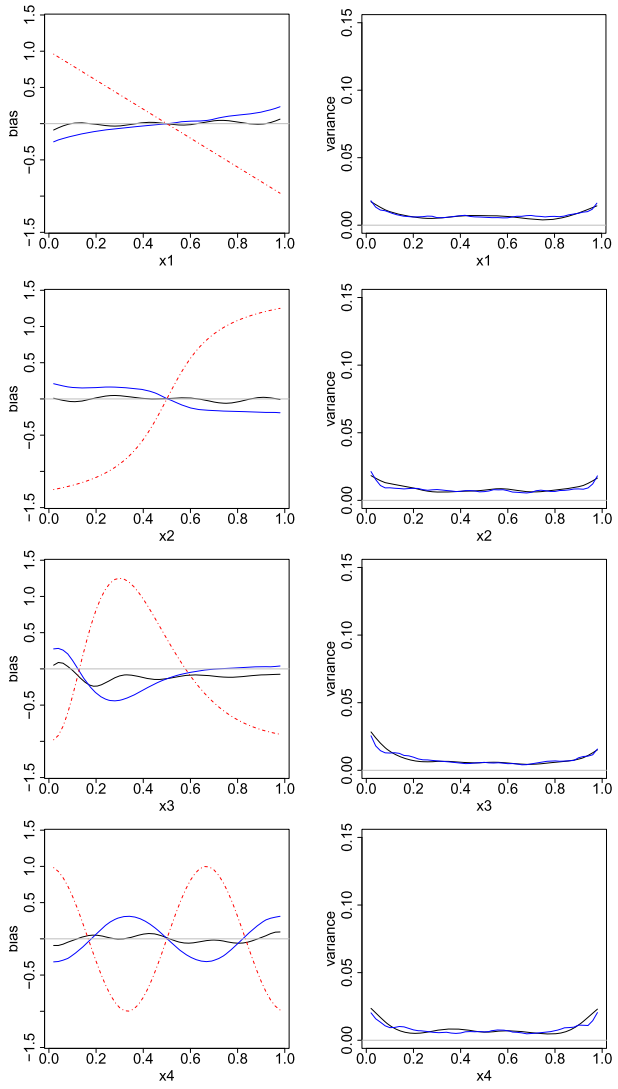


FIG. 3. *The case* $n = 1000$, $\beta = 2.0$ *and* NSR $= 0.1$.

shifted to the right of their targets. This is due to the uncorrected effects of positive measurement errors.

The results given above are for the case where the bandwidths $h_j \equiv h$ and $h$ is chosen optimally. The performance of the proposed deconvolution kernel estimators depends on the choice of bandwidth, as is demonstrated in Theorems 3.3 and 3.4. We also found this when we located in a grid search the optimal common bandwidths $h_{\text{opt}}$ that gave the results in Tables 1–5. For example, in the case where $\beta = 2.0$, NSR $= 0.1$ and $n = 400$, we found that the MISEs of the D-SBF estima-

TABLE 5
*Gamma measurement error distribution with* $\beta = 1.0$ *and* NSR $= 0.1$, *based on* 200 *MC samples of size* $n = 400$

| | Naive SBF | | | | Deconvolution SBF | | | |
|---|---|---|---|---|---|---|---|---|
| **Criterion** | $j=1$ | $j=2$ | $j=3$ | $j=4$ | $j=1$ | $j=2$ | $j=3$ | $j=4$ |
| MISE | 0.040 | 0.100 | 0.189 | 0.206 | 0.023 | 0.026 | 0.037 | 0.037 |
| ISB | 0.027 | 0.086 | 0.175 | 0.191 | 0.005 | 0.006 | 0.014 | 0.015 |
| IV | 0.013 | 0.014 | 0.014 | 0.015 | 0.018 | 0.020 | 0.023 | 0.022 |

tors of $f_j$ for $j = 1, 2, 3, 4$ were 0.044, 0.046, 0.054, 0.040, respectively, when $h_j = h_{\text{opt}}/2$, and 0.048, 0.116, 0.326, 0.489, respectively, when $h_j = 2h_{\text{opt}}$. In the practical implementation of our method, we may want to determine data-driven bandwidths and choose possibly different $\hat{h}_j$ for different component functions $f_j$. Below, we suggest a simple method based on a "bandwidth factor" scheme, and assess its performance.

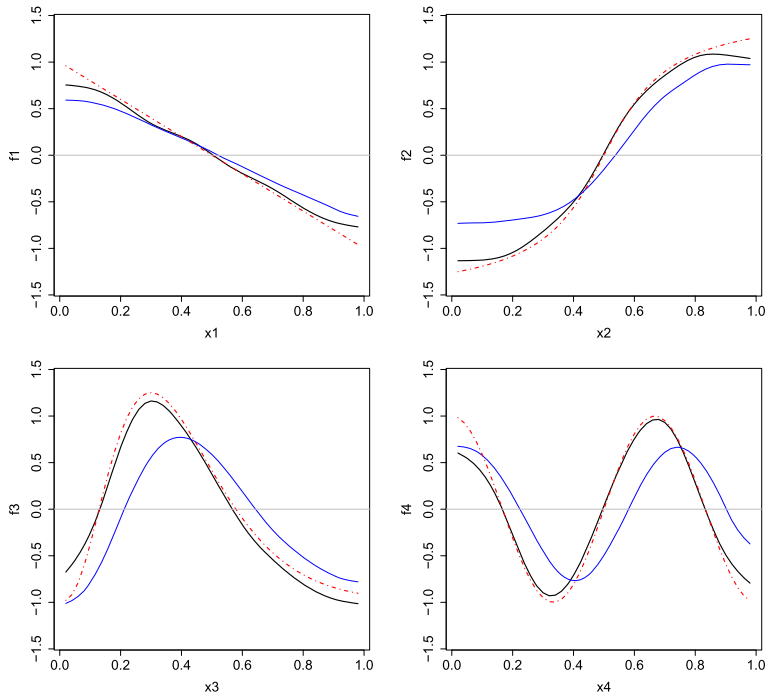


FIG. 4. *The case of a Gamma measurement error (asymmetric) with* $\beta = 1.0$ *and* NSR $= 0.1$. *The D-SBF estimates (black) and N-SBF (blue) averaged over* 200 *MC replications, with the true functions* $f_j$ *(red). The sample size* $n = 400$.

TABLE 6
*Same setting as Table 2 ($n = 400$), but with data-driven bandwidth selectors*

| | Naive SBF | | | | Deconvolution SBF | | | |
|---|---|---|---|---|---|---|---|---|
| **Criterion** | $j=1$ | $j=2$ | $j=3$ | $j=4$ | $j=1$ | $j=2$ | $j=3$ | $j=4$ |
| MISE | 0.046 | 0.076 | 0.044 | 0.042 | 0.018 | 0.021 | 0.034 | 0.028 |
| ISB | 0.036 | 0.065 | 0.031 | 0.031 | 0.009 | 0.008 | 0.004 | 0.005 |
| IV | 0.010 | 0.011 | 0.013 | 0.011 | 0.009 | 0.013 | 0.030 | 0.022 |

Let $\hat{m}_j(\cdot; h_j)$ denote the marginal regression estimators as defined in (2.15). First, for each $j = 1, \ldots, d$, choose the bandwidth $h_j$ that minimizes the marginal goodness-of-fit $\sum_{i=1}^n (Y^i - \hat{m}_j(Z_j^i; h_j))^2$. For each given common factor $\alpha > 0$, compute the smooth backfitting goodness-of-fit $\mathrm{GOF}(\alpha) = \sum_{i=1}^n (Y^i - \bar{Y} - \sum_{j=1}^d \hat{f}_j(Z_j^i; \alpha h_j))^2$. Then, find $\hat{\alpha} = \arg\min_{\alpha>0} \mathrm{GOF}(\alpha)$ and finally choose $\hat{h}_j = \hat{\alpha} h_j$. One could use a cross-validatory version of the GOF criterion but it would be computationally expensive since our method requires the Fourier and inverse-Fourier transformations for deconvolution, the normalization step and the iterative backfitting, for every different choice of bandwidth. In fact, we applied the above simple bandwidth selection method to both the N-SBF and the D-SBF estimators under our simulation settings, and found it worked very well. The results for the case $\beta = 0.4$, $\mathrm{NSR} = 0.1$ and $n = 400$ are presented in Table 6. A comparison with Table 2 suggests that the D-SBF estimators with the data-driven bandwidth selectors have comparable with or even better performance than those that use an optimally chosen universal bandwidth.

**5. Proofs of theorems.** In this section, we provide the proofs of Theorems 3.3 and 3.4. We begin by a lemma that gives an enveloping inequality for the deconvolution and normalized kernel function $\tilde{K}_h^\star$. The lemma is used frequently in our asymptotic analysis and its proof is given in the Supplementary Material [12]. Proofs of other technical details are also found there.

LEMMA 5.1. *Assume the conditions* (K1), (D1) *and* (D2). *Then there exists a constant $C > 0$ such that*

$$h^{1+\beta} |\tilde{K}_h^\star(x, z)| \le C \cdot \frac{h}{h + |x - z|}$$

*for all $x \in [0, 1]$ and $z \in \mathbb{R}$.*

5.1. *Proof of Theorem 3.3.* We first note that $\hat{\delta}_j^A$, $\hat{\delta}_{2j}^B$ and $\hat{\delta}_{2j}^C$ have mean zero. For the proof of the first part of the theorem, we compute $\mathbb{E}|\tilde{K}_{h_j}^\star(x_j, Z_j)|^2$. Define

$$J_\beta(v; x) = \frac{1}{2\pi c_\beta} \int_{-\infty}^{\infty} e^{-itv} t^\beta \phi_K(t; x) \phi_K(t)\, dt,$$

where $c_\beta$ is the constant that appears in the condition (D2). As a function of the first argument, $J_\beta(\cdot; x)$ is square integrable. Indeed, it holds that

$$\begin{aligned}\int_{-\infty}^{\infty} |J_\beta(v;x)|^2\,dv &= \frac{1}{2\pi c_\beta^2}\int_{-\infty}^{\infty} t^{2\beta}|\phi_K(t;x)|^2\phi_K(t)^2\,dt \\ &\le (\text{const})\cdot\int_0^{\infty}(1+t)^{2(\beta-\lfloor\beta\rfloor-2)}\,dt<\infty.\end{aligned} \tag{5.1}$$

The equality in (5.1) follows from the Plancherel identity and the fact that $2\pi c_\beta J_\beta(\cdot; x)$ is the Fourier transform of the function $g$ defined by $g(t) = t^\beta\phi_K(\cdot; x)\phi_K(\cdot)$. The inequality in (5.1) holds due to the condition (K1). Let $p_{U_j}$ and $p_{Z_j}$ denote the densities of $U_j$ and $Z_j$, respectively. Then it follows that

$$\begin{aligned}\mathbb{E}|\tilde{K}^\star_{h_j}(x_j, Z_j)|^2 &= h_j^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itv}\frac{\phi_K(t;x_j)\phi_K(t)}{\phi_{U_j}(t/h_j)}\,dt\right|^2 \\ &\quad\times p_j(x_j-u-h_jv)p_{U_j}(u)\,dv\,du \\ &= h_j^{-1-2\beta}p_{Z_j}(x_j)\int_{-\infty}^{\infty}|J_\beta(v;x_j)|^2\,dv + o\big(h_j^{-1-2\beta}\big),\end{aligned} \tag{5.2}$$

uniformly for $x_j \in [0, 1]$. The first part of the theorem now follows from an exponential inequality for sums of independent mean zero random variables, applied to $\hat{\delta}_j^A$ conditioning on $\{Z_j^i : 1 \le i \le n\}$.

For $\hat{\delta}_{2j}^B$, we apply an exponential inequality conditioning on $\{X_j^i : 1 \le i \le n\}$. For this, we need to approximate $n^{-1}\sum_{i=1}^n \mathbb{E}(|\tilde{K}^\star_{h_j}(x_j, Z_j^i)(f_j(X_j^i) - f_j(x_j))|^2 \mid X_j^i)$. Similarly, as in the case of $\hat{\delta}_j^A$, we obtain

$$\begin{aligned}&\mathbb{E}|\tilde{K}^\star_{h_j}(x_j, Z_j)\big(f_j(X_j) - f_j(x_j)\big)|^2 \\ &\quad= h_j^{-1-2\beta}p_{Z_j}(x_j)\mathbb{E}\big[\big(f_j(X_j) - f_j(x_j)\big)^2 \mid Z_j = x_j\big] \\ &\qquad\times\int_{-\infty}^{\infty}|J_\beta(v;x_j)|^2\,dv + o\big(h_j^{-1-2\beta}\big),\end{aligned}$$

uniformly for $x_j \in [0, 1]$. This implies the second part of the theorem.

The calculation of the magnitude of $\hat{\delta}_{2j}^C \equiv \sum_{k\ne j}\hat{\delta}_{k,2j}^C$ is more involved. We treat $\hat{\delta}_{k,2j}^C$ for each $k \ne j$. We apply an exponential inequality to $\hat{\delta}_{k,2j}^C$, now conditioning on $\{X_k^i : 1 \le i \le n\}$. The order of magnitude of $\hat{\delta}_{k,2j}^C(x_j)$ is determined by that of

$$\begin{aligned}&n^{-1}\sum_{i=1}^n \mathbb{E}\big(|\tilde{K}^\star_{h_j}(x_j, Z_j^i)\big(V_{nk}^i - \mathbb{E}(V_{nk}^i \mid X_k^i)\big)|^2 \mid X_k^i\big) \\ &\quad= n^{-1}\sum_{i=1}^n \mathbb{E}\big(\mathbb{E}\big(|\tilde{K}^\star_{h_j}(x_j, Z_j^i)\big)|^2 \mid X_j^i, X_k^i\big)\cdot\big|V_{nk}^i - \mathbb{E}(V_{nk}^i \mid X_k^i)\big|^2 \mid X_k^i\big) \\ &\quad\le (\text{const})\cdot h_j^{-1-2\beta}n^{-1}\sum_{i=1}^n \mathbb{E}\big(\big|V_{nk}^i - \mathbb{E}(V_{nk}^i \mid X_k^i)\big|^2 \mid X_k^i\big).\end{aligned} \tag{5.3}$$

The equality in (5.3) follows since $U_j$'s are independent among themselves and also independent of $X_j$'s so that $\mathbb{E}(|\tilde{K}^{\star}_{h_j}(x_j, Z_j)|^2 \mid X_j, X_k, U_k) = \mathbb{E}(|\tilde{K}^{\star}_{h_j}(x_j, Z_j)|^2 \mid X_j, X_k)$. The inequality in (5.3) holds since

$$\begin{aligned}&\mathbb{E}\big(|\tilde{K}^{\star}_{h_j}(x_j, Z_j)|^2 \mid X_j = \xi\big)\\ &\quad = h_j^{-1-2\beta} p_{U_j}(x_j - \xi) \int_{-\infty}^{\infty} |J_\beta(v; x_j)|^2 \, dv + o\big(h_j^{-1-2\beta}\big),\end{aligned} \tag{5.4}$$

which is bounded by $Ch_j^{-1-2\beta}$ uniformly for $x_j$ and $\xi$ for some constant $C > 0$. The result (5.4) may be obtained by using similar arguments as in deriving (5.2). The right-hand side of the inequality (5.3) is of the same magnitude as

$$h_j^{-1-2\beta} \cdot \mathbb{E}|V_{nk} - \mathbb{E}(V_{nk} \mid X_k)|^2 \le h_j^{-1-2\beta} \cdot \mathbb{E}|V_{nk}|^2.$$

We calculate the magnitude of $\mathbb{E}|V_{nk}|^2$. Let $D_k(u) = (f_k(X_k) - f_k(u))I_{[0,1]}(u)$ and $\phi_{D_k}$ denote its Fourier transform. We first note that $V_{nk} = V_{nk,1} + V_{nk,2}$, where

$$\begin{aligned}V_{nk,1} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itZ_k} \frac{\phi_K(h_k t)^2 \phi_{D_k}(-t)}{\phi_{U_k}(t)} \, dt,\\ V_{nk,2} &= \frac{1}{2\pi} \int_{I_{0k}^c} \int_{-\infty}^{\infty} e^{-it(x_k - Z_k)} \frac{\phi_K(h_k t)[\phi_K(h_k t; x_k) - \phi_K(h_k t)]}{\phi_{U_k}(t)}\\ &\quad \times \big(f_k(X_k) - f_k(x_k)\big) \, dt \, dx_k.\end{aligned}$$

From the Plancherel equality and the fact that $|\phi_{D_k}(t)| \le (\text{const}) \cdot (1 + |t|)^{-1}$, we find that

$$\begin{aligned}\mathbb{E}|V_{nk,1}|^2 &\le \frac{\|p_{Z_k}\|_\infty}{4\pi^2} \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} e^{itz} \frac{\phi_K(h_k t)^2 \phi_{D_k}(-t)}{\phi_{U_k}(t)} \, dt \right|^2 dz\\ &= \frac{\|p_{Z_k}\|_\infty}{2\pi} \int_{-\infty}^{\infty} \left| \frac{\phi_K(h_k t)^2 \phi_{D_k}(-t)}{\phi_{U_k}(t)} \right|^2 dt\\ &\le C_1 h_k^{1-2\beta} \int_0^{\infty} (h_k + t)^{2\beta-2} \phi_K(t)^4 \, dt\end{aligned} \tag{5.5}$$

for some constant $C_1 > 0$. The right-hand side of (5.5) has different magnitudes for different ranges of $\beta$. In case $\beta < 1/2$, we may prove that

$$\kappa_n \equiv \int_0^{\infty} (h_k + t)^{2\beta-2} \phi_K(t)^4 \, dt \le C_2 h_k^{2\beta-1}$$

for some constant $C_2 > 0$, using the fact that $|\phi_K(t)| \le (1 + |t|)^{-\lfloor \beta \rfloor - 1}$. We may also show that $\kappa_n \le C_3$ when $\beta > 1/2$ and $\kappa_n \le C_4 \log h_k^{-1}$ when $\beta = 1/2$, for some positive constants $C_3$ and $C_4$. This establishes $\mathbb{E}|V_{nk,1}|^2 = O(\tau_{n,k}(\beta)^2)$.

Now, we compute $\mathbb{E}|V_{nk,2}|^2$. For this, we note that

$$\left|\int_{I_{0k}^c} e^{-itx_k}\big(\phi_K(h_kt;x_k)-\phi_K(h_kt)\big)\big(f_k(X_k)-f_k(x_k)\big)\,dx_k\right| \le C_5h_k$$

for some constant $C_5>0$. This follows from the facts that $|\phi_K(h_kt;x_k)|$, $|\phi_K(h_kt)|$ and $|f_k(X_k)-f_k(x_k)|$ are bounded and that the length of $I_{0k}^c$ is $4h_k$. Thus, by the Plancherel equality again, we obtain

$$\begin{aligned}\mathbb{E}|V_{nk,2}|^2 &\le \frac{C_5^2h_k^2\|pZ_k\|_\infty}{4\pi^2}\int_{-\infty}^{\infty}\left|\int_{-\infty}^{\infty}e^{itz}\frac{\phi_K(h_kt)}{\phi_{U_k}(t)}\,dt\right|^2dz\\ &= \frac{C_5^2h_k^2\|pZ_k\|_\infty}{2\pi}\int_{-\infty}^{\infty}\left|\frac{\phi_K(h_kt)}{\phi_{U_k}(t)}\right|^2dt \qquad (5.6)\\ &\le C_6h_k^{1-2\beta}\int_0^{\infty}(h_k+t)^{2\beta}(1+t)^{-2\lfloor\beta\rfloor-2}\,dt.\end{aligned}$$

The integral on the right-hand side of (5.6) is bounded by some constant for all $\beta\ge 0$. This verifies $\mathbb{E}|V_{nk,2}|^2=O(h_k^{1-2\beta})$, and thus completes the proof of Theorem 3.3.

5.2. *Proof of Theorem 3.4.* We first present two lemmas. The proofs of these lemmas and some other technical details can be found in the online Supplementary Material [12]. Let $d_{n,j}=\pi_j\Delta_{n,f}\in\mathcal{H}_j$ for $1\le j\le d$. Also, let $a_j(u)=\mu_{1,j}(u)f_j'(u)$ and $c_j(u)=\mu_2f_j''(u)/2$.

LEMMA 5.2. *Under the conditions of Theorem 3.4, it follows that*

$$\hat\delta_{1j}^B+\hat\delta_{1j}^C=d_{n,j}+\sum_{k=1}^{d}\hat\pi_j\left(h_k\frac{a_k}{\mu_{0,k}}+h_k^2c_k\right)+r_{n,j}.$$

LEMMA 5.3. *Let $(h_j^2\hat\Delta_j:1\le j\le d)$ with each $\hat\Delta_j\in\mathcal{H}_j$ be the solution of the system of equations*

$$h_j^2\hat\Delta_j=d_{n,j}-\hat\pi_j\left(\sum_{k\ne j}h_k^2\hat\Delta_k\right),\qquad 1\le j\le d, \tag{5.7}$$

*subject to the constraints*

$$\begin{aligned}h_j^2\int_0^1\hat\Delta_j(x_j)\hat p_j(x_j)\,dx_j=&-\int_0^1\left(f_j(x_j)+h_j\frac{a_j(x_j)}{\mu_{0,j}(x_j)}+h_j^2c_j(x_j)\right)\\ &\times\hat p_j(x_j)\,dx_j.\end{aligned} \tag{5.8}$$

*Under the conditions of Theorem 3.4, it holds that*

$$h_j^2\hat\Delta_j=h_j^2\Delta_j+r_{n,j}. \tag{5.9}$$

Now we prove Theorem 3.4. From (3.3) and Lemma 5.2, we get

$$\begin{aligned}\hat{f}_j - f_j &= \hat{\delta}_j^A + \hat{\delta}_{2j}^B + \hat{\delta}_{2j}^C + h_j \frac{a_j}{\mu_{0j}} + h_j^2 c_j + d_{n,j} \\ &\quad - \sum_{k\neq j}^{d} \hat{\pi}_j \left( \hat{f}_k - f_k - h_k \frac{a_k}{\mu_{0,k}} - h_k^2 c_k \right) + r_{n,j}.\end{aligned} \tag{5.10}$$

Define $\hat{D}_j = \hat{f}_j - f_j - h_j \frac{a_j}{\mu_{0,j}} - h_j^2 c_j \in \mathcal{H}_j$. Then (5.10) is equivalent to

$$\hat{D}_j = \hat{\delta}_j^A + \hat{\delta}_{2j}^B + \hat{\delta}_{2j}^C + d_{n,j} - \sum_{k\neq j}^{d} \hat{\pi}_j \hat{D}_k + r_{n,j}, \qquad 1 \le j \le d. \tag{5.11}$$

Let $\hat{D}_+ = \hat{D}_1 + \cdots + \hat{D}_d \in \mathcal{H}$. Define $\hat{\delta}_\oplus^A$, $\hat{\delta}_{2\oplus}^B$, $\hat{\delta}_{2\oplus}^C$ and $d_{n,\oplus}$ in the same way as $\hat{\delta}_\oplus$ with $\hat{\delta}_j$ being replaced by $\hat{\delta}_j^A$, $\hat{\delta}_{2j}^B$, $\hat{\delta}_{2j}^C$ and $d_{n,j}$, respectively. As we get (3.4) from (3.3), we obtain from (5.11)

$$\hat{D}_+ = \hat{T}\hat{D}_+ + \hat{\delta}_\oplus^A + \hat{\delta}_{2\oplus}^B + \hat{\delta}_{2\oplus}^C + d_{n,\oplus} = \sum_{j=0}^{\infty} \hat{T}^j (\hat{\delta}_\oplus^A + \hat{\delta}_{2\oplus}^B + \hat{\delta}_{2\oplus}^C + d_{n,\oplus}). \tag{5.12}$$

Define $\hat{\delta}_+^A = \hat{\delta}_1^A + \cdots + \hat{\delta}_d^A$. Likewise, define $\hat{\delta}_{2,+}^B$ and $\hat{\delta}_{2,+}^C$ from $\hat{\delta}_{2j}^B$ and $\hat{\delta}_{2j}^C$, respectively.

We claim that

$$\sum_{j=0}^{\infty} \hat{T}^j \hat{\delta}_\oplus^A = \hat{\delta}_+^A + r_{n,+}, \qquad \sum_{j=0}^{\infty} \hat{T}^j \hat{\delta}_{2\oplus}^* = \hat{\delta}_{2,+}^* + r_{n,+} \tag{5.13}$$

for $* = B$ and $C$, where $r_{n,+}$ denote generic terms such that $r_{n,+} = r_{n,1} + \cdots + r_{n,d}$. We note that $\hat{\Delta}_{n,+} \equiv \sum_{j=0}^{\infty} \hat{T}^j d_{n,\oplus}$ solves the equation $\hat{\Delta}_{n,+} = \hat{T}\hat{\Delta}_{n,+} + d_{n,\oplus}$. This means that

$$\hat{\Delta}_{n,+} = h_1^2 \hat{\Delta}_1 + \cdots + h_d^2 \hat{\Delta}_d, \tag{5.14}$$

where $(h_j^2 \hat{\Delta}_j : 1 \le j \le d)$ is the solution of the system of equations (5.7) in Lemma 5.3. The $j$th component (function) of $\hat{\Delta}_{n,+}$ may differ from $h_j^2 \hat{\Delta}_j(\cdot)$ only by a constant function (possibly random), say $c_{n,j}(\cdot) \equiv c_{n,j}$, such that $\sum_{j=1}^{d} c_{n,j} = 0$. Since $\hat{T}$ is linear, (5.12)–(5.14) imply

$$\hat{D}_j = \hat{\delta}_j^A + \hat{\delta}_{2j}^B + \hat{\delta}_{2j}^C + h_j^2 \hat{\Delta}_j + r_{n,j} + c_{n,j}^*, \qquad 1 \le j \le d, \tag{5.15}$$

for some constant functions $c^*_{n,j}$ (possibly random) such that $\sum_{j=1}^d c^*_{n,j} = 0$. We note that

$$\begin{aligned}
\int_0^1 \hat{\delta}_j^A(x_j)\hat{p}_j(x_j)\,dx_j &= n^{-1}\sum_{i=1}^n \varepsilon^i = O_p\bigl(n^{-1/2}\bigr),\\
\int_0^1 \hat{\delta}_{2j}^B(x_j)\hat{p}_j(x_j)\,dx_j &= n^{-1}\sum_{i=1}^n \bigl(V_{nj}^i - \mathbb{E}\bigl(V_{nj}^i \mid X_j^i\bigr)\bigr)\\
&= O_p\bigl(n^{-1/2}\tau_{n,j}(\beta_j)\bigr),\\
\int_0^1 \hat{\delta}_{k,2j}^C(x_j)\hat{p}_j(x_j)\,dx_j &= n^{-1}\sum_{i=1}^n \bigl(V_{nk}^i - \mathbb{E}\bigl(V_{nk}^i \mid X_k^i\bigr)\bigr)\\
&= O_p\bigl(n^{-1/2}\tau_{n,k}(\beta_k)\bigr).
\end{aligned} \tag{5.16}$$

We may prove the above results using the normalization property of $\tilde{K}^\star_{h_j}$ and the fact that $\mathbb{E}|V_{nj}|^2 = O(\tau_{n,j}(\beta_j)^2)$. Due to the constraints on $\hat{\Delta}_j$ at (5.8) and the results (5.16), we may set $c^*_{n,j} \equiv 0$ for all $1 \le j \le d$ in (5.15). From Lemma 5.3, we establish that

$$\hat{D}_j = \hat{\delta}_j^A + \hat{\delta}_{2j}^B + \hat{\delta}_{2j}^C + h_j^2\Delta_j + r_{n,j}, \qquad 1 \le j \le d.$$

This completes the proof of Theorem 3.4.

We prove the claim (5.13). The claim follows if we prove that, for all $k \ne j$,

$$\begin{aligned}
\hat{\pi}_k\hat{\delta}_j^A(u) &= o_p\bigl(n^{-1/2}h_j^{-1/2-\beta_j}\sqrt{\log n}\bigr) = \hat{\pi}_k\hat{\delta}_{2j}^B(u),\\
\hat{\pi}_k\hat{\delta}_{2j}^C(u) &= o_p\left(n^{-1/2}h_j^{-1/2-\beta_j}\sqrt{\log n}\sum_{l\ne j}^d \tau_{n,l}(\beta_l)\right),
\end{aligned} \tag{5.17}$$

uniformly for $u \in [0,1]$. We prove the first and third parts of (5.17). The proof of the second part is similar to that of the third part. Below in our presentation, we fix a pair $(j,k)$ such that $j \ne k$ and suppress $(j,k)$ in some places. Define

$$Q(x_j,x_k) = \frac{p_{jk}(x_j,x_k)}{p_j(x_j)p_k(x_k)}, \qquad \hat{Q}(x_j,x_k) = \frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)\hat{p}_k(x_k)}.$$

Then we may write

$$\bigl(\hat{\pi}_k\hat{\delta}_j^A\bigr)(u) = n^{-1}\sum_{i=1}^n \varepsilon^i\bigl(W_1^i(u) + W_2^i(u)\bigr),$$

where $W_1^i(u) = \int_0^1 \tilde{K}^\star_{h_j}(x_j,Z_j^i)Q(x_j,u)\,dx_j$, $W_2^i(u) = \int_0^1 \tilde{K}^\star_{h_j}(x_j,Z_j^i)(\hat{Q}(x_j,u) - Q(x_j,u))\,dx_j$. Using the arguments in calculating the magnitude of $\mathbb{E}|V_{nk}|^2$ in the proof of Theorem 3.3 and the fact that $Q$ is bounded, we can show that

$\mathbb{E}|W_1(u)|^2 = O(\tau_{n,j}(\beta_j)^2)$ uniformly for $u \in [0,1]$. Since $h_j^{-1/2-\beta_j}/\tau_{n,j}(\beta_j) \to \infty$ as $n \to \infty$, this establishes

$$\sup_{u\in[0,1]} \left| n^{-1} \sum_{i=1}^{n} \varepsilon^i W_1^i(u) \right| = o_p\bigl(n^{-1/2} h_j^{-1/2-\beta_j} \sqrt{\log n}\bigr). \tag{5.18}$$

Now, for the second term of $\hat{\pi}_k \hat{\delta}_j^A$, we note that $W_2^i(u)$ involves $Z_k^i$ as well as $Z_j^i$. We use an exponential inequality, now conditioning on $\{(Z_j^i, Z_k^i) : 1 \le i \le n\}$. Since

$$\begin{aligned} &\sup_{u\in[0,1]} n^{-1} \sum_{i=1}^{n} \bigl|W_2^i(u)\bigr|^2 \\ &\quad \le \max_{1\le i\le n} \sup_{u\in[0,1]} \left| \int_0^1 \tilde{K}_{h_j}^{\star}(x_j, Z_j^i) \bigl(\hat{Q}(x_j,u) - Q(x_j,u)\bigr)\, dx_j \right|^2 \\ &\quad \le (\text{const}) h_j^{-2\beta_j} \|\hat{Q} - Q\|_\infty^2 \\ &\quad = o_p\bigl(h_j^{-1-2\beta_j}\bigr), \end{aligned} \tag{5.19}$$

we conclude that

$$\sup_{u\in[0,1]} \left| n^{-1} \sum_{i=1}^{n} \varepsilon^i W_2^i(u) \right| = o_p\bigl(n^{-1/2} h_j^{-1/2-\beta_j} \sqrt{\log n}\bigr). \tag{5.20}$$

The second inequality in (5.19) is due to (5.24) below. The first part of (5.17) now follows from (5.18) and (5.20).

The proof of the third part is more involved and needs refined arguments. We prove it using some probability bounds in empirical process theory. Recall that

$$\begin{aligned} \hat{\delta}_{l,2j}^C(x_j) = \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^{n} \int_0^1 \tilde{K}_{h_j}^{\star}(x_j, Z_j^i) \bigl[\tilde{K}_{h_l}^{\star}(x_l, Z_l^i) - \mathbb{E}\bigl(\tilde{K}_{h_l}^{\star}(x_l, Z_l^i) \mid X_l^i\bigr)\bigr] \\ \times \bigl[f_l(X_l^i) - f_l(x_l)\bigr]\, dx_l \end{aligned}$$

and $\hat{\delta}_{2j}^C(x_j) = \sum_{l\ne j} \hat{\delta}_{l,2j}^C(x_j)$. We fix $l \ne j$ and analyze $\hat{\pi}_k \hat{\delta}_{l,2j}^C$. Note that $l$ may equal $k$ although both of them are different from $j$. For a bivariate function $G$, we define

$$A_{nj}^i(u, G) = \int_0^1 \tilde{K}_{h_j}^{\star}(x_j, Z_j^i) G(x_j, u)\, dx_j,$$

and let $\tilde{V}_{nl}^i = V_{nl}^i - \mathbb{E}(V_{nl}^i \mid X_l^i)$ for simplicity of notation. Then

$$(\hat{\pi}_k \hat{\delta}_{l,2j}^C)(u) = n^{-1} \sum_{i=1}^{n} A_{nj}^i(u, \hat{Q}) \tilde{V}_{nl}^i.$$

Take an arbitrarily small $\epsilon > 0$ and put $\varepsilon_n \equiv \varepsilon_n(\epsilon) = \epsilon \cdot n^{-1/2} h_j^{-1/2-\beta_j} \sqrt{\log n} \cdot \tau_{n,l}(\beta_l)$. Then, using the techniques in the proof of Theorem 3.3, we may prove

$$P\left(\left|n^{-1}\sum_{i=1}^{n} A_{nj}^{i}(u,G)\tilde{V}_{nl}^{i}\right| > \varepsilon_n/2\right) \to 0 \tag{5.21}$$

for each fixed $u \in [0,1]$ and bounded $G$. In fact,

$$n^{-1}\sum_{i=1}^{n} A_{nj}^{i}(u,G)\tilde{V}_{nl}^{i} = O_p\big(n^{-1/2} h_j^{-\beta_j}\sqrt{\log n}\cdot \tau_{n,l}(\beta_l)\big). \tag{5.22}$$

The proof of (5.22) can be found in the online Supplementary Material [12]. From (5.21) and an application of the symmetrization technique, it follows that

$$\begin{aligned} &P\left(\sup_{u\in[0,1]}\sup_{G\in\mathcal{G}_n}\left|n^{-1}\sum_{i=1}^{n} A_{nj}^{i}(u,G)\tilde{V}_{nl}^{i}\right| > \varepsilon_n\right) \\ &\quad\le 4P\left(\sup_{u\in[0,1]}\sup_{G\in\mathcal{G}_n}\left|n^{-1}\sum_{i=1}^{n} R^i A_{nj}^{i}(u,G)\tilde{V}_{nl}^{i}\right| > \varepsilon_n/4\right) \end{aligned} \tag{5.23}$$

for any function class $\mathcal{G}_n$, where $R^i$ are i.i.d. Rademacher sequence, that is, $P(R^i = 1) = P(R^i = -1) = 1/2$, independent of $T^i \equiv (X_j^i, U_j^i, X_l^i, U_l^i, X_k^i, U_k^i)$.

We prove the right-hand side of (5.23) converges to zero, as $n$ tends to infinity, for $\mathcal{G}_n$ where $\hat{Q}$ belongs with a high probability. Specifically, we set $\mathcal{G}_n \equiv \mathcal{G}_n(D_1, D_2)$ to be a class of bivariate functions $G$ such that:

$$\text{(i)} \quad \sup_{u_1,u_2\in[0,1]} \big|G(u_1,u_2)\big| \le D_1,$$

$$\text{(ii)} \quad \sup_{u_1\in[0,1]} \big|G(u_1,u_2) - G\big(u_1,u_2'\big)\big| \le D_2 \cdot \big|u_2 - u_2'\big| \cdot h_j^{-1/2}.$$

By taking $D_1$ and $D_2$ sufficiently large, we may make the probability $P(\hat{Q} \in \mathcal{G}_n)$ sufficiently large. For this, we need $nh_j^{2\beta_j} h_k^{3+2\beta_k}/\log n$ is bounded away from zero. Let $\mathcal{I}_n(2^{-\ell})$ and $\mathcal{G}_n(2^{-\ell})$ denote $2^{-\ell}$- and $D_1 2^{-\ell}$-covering sets of $[0,1]$ and $\mathcal{G}_n$, respectively. Their entropies equal $2^\ell/D_1$ and $\ell \cdot \log 2$, respectively. For each $(u,G) \in [0,1] \times \mathcal{G}_n$ and $\ell \ge 0$, we choose $(u_\ell, G_\ell) \in \mathcal{I}_n(2^{-\ell}) \times \mathcal{G}_n(2^{-\ell})$ such that $|u_\ell - u| \le 2^{-\ell}$ and $\|G_\ell - G\|_\infty \le D_1 2^{-\ell}$. We may choose $(u_0, G_0) = (0,0)$. Note that $A_{nj}^i(0,0) = 0$. Here, we suppress the dependency of the choice $\{(u_j, G_j) : j \ge 0\}$ on $(u,G)$. We can prove

$$\max_{1\le i\le n}\sup_{u\in[0,1]} \big|A_{nj}^{i}(u,G)\big| \le c_1 h_j^{-\beta_j}\|G\|_\infty, \qquad \max_{1\le i\le n}\big|\tilde{V}_{nl}^{i}\big| \le c_2 h_l^{-\beta_l} \tag{5.24}$$

for some absolute constant $c_1, c_2 > 0$. The proof of (5.24) can be found in the online Supplementary Material [12]. For such $c_1$ and $c_2$ and for $D_2$ in the definition

of $\mathcal{G}_n$, we define

$$J_n = \min\left\{\ell \geq 1 : 2^{-\ell} \leq \frac{\varepsilon_n h_j^{1/2+\beta_j} h_l^{\beta_l}}{16 \cdot c_1 c_2 D_2}\right\}.$$

Then, since $|A^i_{nj}(u, G)V^i_{nl} - A^i_{nj}(u_{J_n}, G_{J_n})V^i_{nl}| \leq \varepsilon_n/8$, we have

$$\sup_{u\in[0,1]}\sup_{G\in\mathcal{G}_n}\left|n^{-1}\sum_{i=1}^{n}\left(A^i_{nj}(u, G)\tilde{V}^i_{nl} - A^i_{nj}(u_{J_n}, G_{J_n})\tilde{V}^i_{nl}\right)\right| \leq \varepsilon_n/8. \tag{5.25}$$

The bound at (5.25) let us consider the supremum in (5.23) restricted to those $(u, G)$ in $\mathcal{I}_n(2^{-J_n}) \times \mathcal{G}_n(2^{-J_n})$. Let $\mathcal{T}_n$ denote the set of $T^i$, $1 \leq i \leq n$. Take $\eta_\ell > 0$ such that $\sum_{\ell=1}^{\infty}\eta_\ell \leq 1$. Then it follows that

$$\begin{aligned}
&P\left(\sup_{u\in\mathcal{I}_n(2^{-J_n})}\sup_{G\in\mathcal{G}_n(2^{-J_n})}\left|n^{-1}\sum_{i=1}^{n}R^i A^i_{nj}(u, G)\tilde{V}^i_{nl}\right| > \varepsilon_n/8 \mid \mathcal{T}_n\right)\\
&\quad\leq \sum_{\ell=1}^{J_n}\exp(D_1^{-1}2^\ell + \ell\log 2)\\
&\quad\quad\times \overset{*}{\sup} P\left(\left|n^{-1}\sum_{i=1}^{n}R^i\left[A^i_{nj}(u_\ell, G_\ell) - A^i_{nj}(u_{\ell-1}, G_{\ell-1})\right]\tilde{V}^i_{nl}\right|\right.\\
&\quad\quad\left.> \eta_\ell\cdot\varepsilon_n/8 \mid \mathcal{T}_n\right),
\end{aligned} \tag{5.26}$$

where sup* runs over all $(u_\ell, G_\ell) \in \mathcal{I}_n(2^{-\ell}) \times \mathcal{G}_n(2^{-\ell})$ and $(u_{\ell-1}, G_{\ell-1}) \in \mathcal{I}_n(2^{-\ell+1}) \times \mathcal{G}_n(2^{-\ell+1})$ with $|u_\ell - u_{\ell-1}| \leq 2^{-\ell+1}$ and $\|G_\ell - G_{\ell-1}\|_\infty \leq 2^{-\ell+1}$. We may make the probability of

$$\begin{aligned}
\Gamma_n &\equiv n^{-1}\sum_{i=1}^{n}|\tilde{V}^i_{nl}|^2 \cdot \left|A^i_{nj}(u_\ell, G_\ell) - A^i_{nj}(u_{\ell-1}, G_{\ell-1})\right|^2\\
&\leq c_3 D_2^2 2^{-2\ell} h_j^{-1-2\beta_j}\tau_{n,l}(\beta_l)^2
\end{aligned} \tag{5.27}$$

sufficiently large by choosing a constant $c_3 > 0$ sufficiently large. Applying the Höffding inequality and choosing $\eta_\ell = 2^{-\ell/2}\sqrt{\ell}/9$, we find that, on the event where the inequality (5.27) holds, each summand on the right-hand side of (5.26) is bounded by

$$\exp\left(D_1^{-1}2^\ell + \ell\log 2 - \frac{n\varepsilon_n^2\eta_\ell^2}{128\cdot\Gamma_n}\right) \leq \exp\left(-\ell\cdot\frac{\epsilon^2\log n}{128\cdot 9^2\cdot c_3\cdot D_2^2}\right).$$

This proves

$$P\left(\sup_{u\in\mathcal{I}_n(2^{-J_n})}\sup_{G\in\mathcal{G}_n(2^{-J_n})}\left|n^{-1}\sum_{i=1}^{n}R^i A^i_{nj}(u,G)\tilde{V}^i_{nl}\right|>\varepsilon_n/8\right)$$
$$\le 2\exp\left(-\frac{\epsilon^2\log n}{128\cdot 9^2\cdot c_3\cdot D_2^2}\right)\to 0.$$

This completes the proof of Theorem 3.4.

**6. Concluding remarks.** In this paper, we proposed a way of constructing normalized kernels with the unbiased scoring property that are suitable for errors-in-variables additive regression. We studied how the smooth backfitting method based on the proposed kernel scheme works theoretically and empirically. One challenging extension of our study is a treatment in case $\phi_{U_j}$ are unknown. In that case, one may estimate them from repeated observations $Z^i_{j\ell}=X^i_j+U^i_{j\ell}$ for each subject $i$, where $U_{j\ell}$'s are identically distributed as $U_j$. An example is given in [7] for symmetric $U_j$. In case the densities of $U_j$ are believed to belong to a parametric family, one may get estimators of $\phi_{U_j}$ that converge at the parametric rate; see the related discussion in [6]. With estimators $\hat{\phi}_{U_j}$ of $\phi_{U_j}$, one can basically plug them into the definition (2.11) to construct the corresponding versions, say $\bar{K}^{\star}_{h_j}$, of $\tilde{K}^{\star}_{h_j}$. One may then use them in the estimation of $p_j$, $p_{jk}$ and $m_j$, as in (2.15), and use the estimated functions $\hat{m}_j$, $\hat{p}_j$ and $\hat{p}_{jk}$ in the backfitting equations (2.16). The resulting kernels $\bar{K}^{\star}_{h_j}$ satisfy the normalization property at (2.14) if $\hat{\phi}_{U_j}(0)=1$. However, the unbiased scoring property at (2.14) does not hold in general even though $\hat{\phi}_{U_j}$ are constructed from separate independent observations. It holds only asymptotically if $\hat{\phi}_{U_j}$ converge to the respective $\phi_{U_j}$ at some rates. The asymptotic properties of the estimators of $f_j$ based on $\bar{K}^{\star}_{h_j}$ depend on how fast $\hat{\phi}_{U_j}$ converge to $\phi_{U_j}$, the investigation of which we think is a challenging topic one may pursue in a future study.

One may be also interested in an extension of our theory to the super smooth measurement error case. The distribution of the measurement error $U$ is called super smooth if its Fourier transform $\phi_U$ satisfies

$$d_1|t|^{\alpha_1}\exp\bigl(-|t|^{\beta}/\gamma\bigr)\le\bigl|\phi_U(t)\bigr|\le d_2|t|^{\alpha_2}\exp\bigl(-|t|^{\beta}/\gamma\bigr)$$

for some positive constants $d_1$, $d_2$, $\beta$ and $\gamma$ and for some constants $\alpha_1$ and $\alpha_2$. In the definition of our $\tilde{K}^{\star}_h$ at (2.11) as well as in that for the classical deconvolution kernel $K^D$ at (2.3), $\phi_U$ appears in the denominator of the integrand. To make $K^D$ and $\tilde{K}^{\star}_h$ well defined, one needs a strong regularity for $\phi_K$. Usually in the literature for $K^D$, $\phi_K$ is assumed to be compactly supported; see [10], [7], [6] and [5], for example. Meanwhile, our theory as well as the existing smooth backfitting theory is developed for compactly supported $K$. A natural question is then whether there exists a compactly supported $K$ whose Fourier transform $\phi_K$ is also compactly

supported. The answer is "no" except the trivial choice $K \equiv 0$, in which case both the supports of $K$ and $\phi_K$ are an empty set. It can be shown that $K \equiv 0$ is the only function whose Fourier transform is compactly supported (empty set). In general, the faster $|\phi_K|$ decays at tails, the more the support of $K$ stretches out. Thus, to deal with the super smooth case in the errors-in-variables additive regression problem, one needs to develop new smooth backfitting theory for noncompactly supported kernels, which we think is another challenging topic for future study.

Other extensions one may be interested in include those to generalized additive models, varying coefficient models and partially linear additive models. Also, an extension to the Berkson errors-in-variables case, and to the prediction problem as studied in [3] are topics for future study.

## SUPPLEMENTARY MATERIAL

**Supplement to "Smooth backfitting for errors-in-variables additive models"** (DOI: 10.1214/17-AOS1617SUPP; .pdf). The supplement contains proofs of Lemmas 5.1, 5.2 and 5.3. It also gives proofs of (3.1), (5.22) and (5.24).

## REFERENCES


[1] AUGUSTYNIAK, M. and DORAY, L. G. (2012). Inference for a leptokurtic symmetric family of distributions represented by the difference of two gamma variates. *J. Stat. Comput. Simul.* **82** 1621–1634.

[2] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York.

[3] CARROLL, R. J., DELAIGLE, A. and HALL, P. (2009). Nonparametric prediction in measurement error models. *J. Amer. Statist. Assoc.* **104** 993–1003. MR2562002

[4] COMTE, F. and LACOUR, C. (2011). Data-driven density estimation in the presence of additive noise with unknown distribution. *J. Roy. Statist. Soc. Ser. B* **73** 601–627. MR2853732

[5] DATTNER, I., REISS, M. and TRABS, M. (2016). Adaptive quantile estimation in deconvolution with unknown error distribution. *Bernoulli* **78** 231–252. MR3449779

[6] DELAIGLE, A., FAN, J. and CARROLL, R. J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *J. Amer. Statist. Assoc.* **104** 348–359. MR2504382

[7] DELAIGLE, A., HALL, P. and MEISTER, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.* **36** 665–685. MR2396811

[8] FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272. DOI:https://doi.org/10.1214/aos/1176348248. MR1126324

[9] FAN, J. and MASRY, E. (1992). Multivariate regression estimation with errors-in-variables: Asymptotic normality for mixing processes. *J. Multivariate Anal.* **43** 237–2711. MR1193614

[10] FAN, J. and TRUONG, Y. K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21** 1900–1925. MR1245773

[11] HALL, P. and LAHIRI, S. N. (2008). Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist.* **36** 2110–2134. MR2458181

[12] HAN, K. and PARK, B. U. (2018). Supplement to "Smooth backfitting for errors-in-variables additive models." DOI:10.1214/17-AOS1617SUPP.

[13] KAPPUS, J. and MABON, G. (2014). Adaptive density estimation in deconvolution problems with unknown error distribution. *Electron. J. Stat.* **8** 2879–2904.
[14] KLAR, B. (2015). A note on gamma difference distributions. *J. Stat. Comput. Simul.* **85** 3708–3715. MR3401104
[15] LEE, Y. K., MAMMEN, E. and PARK, B. U. (2010). Backfitting and smooth backfitting for additive quantile models. *Ann. Statist.* **38** 2857–2883. MR2722458
[16] LEE, Y. K., MAMMEN, E. and PARK, B. U. (2012). Flexible generalized varying coefficient regression models. *Ann. Statist.* **40** 1906–1933. MR3015048
[17] LEE, Y. K., MAMMEN, E. and PARK, B. U. (2012). Projection-type estimation for varying coefficient regression models. *Bernoulli* **18** 177–205. MR2888703
[18] LIANG, H., HÄRDLE, W. and CARROLL, R. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *Ann. Statist.* **27** 1519–1535. MR1742498
[19] LINTON, O., SPERLICH, S. and VAN KEILEGOM, I. (2008). Estimation of a semiparametric transformation model. *Ann. Statist.* **36** 686–718. MR2396812
[20] 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
[21] MAMMEN, E. and NIELSEN, J. P. (2003). Generalised structured models. *Biometrika* **90** 551–566.
[22] MASRY, E. (1993). Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stochastic Process. Appl.* **47** 53–74.
[23] ROCA-PARDINAS, J. and SPERLICH, S. (2010). Feasible estimation in generalized structured models. *Stat. Comput.* **20** 367–379.
[24] STEFANSKI, L. A. and CARROLL, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21** 169–184.
[25] STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566
[26] YOU, J., ZHOU, Y. and CHEN, G. (2006). Corrected local polynomial estimation in varying-coefficient models with measurement errors. *Ann. Statist.* **34** 391–410. MR2328551
[27] YU, K., PARK, B. U. and MAMMEN, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36** 228–260. MR2387970
[28] ZHU, L. and CUI, H. (2003). A semi-parametric regression model with errors in variables. *Scand. J. Stat.* **30** 429–442. MR1983135

DEPARTMENT OF STATISTICS
SEOUL NATIONAL UNIVERSITY
1 GWANAK-RO, GWANAK-GU
SEOUL 08826
REPUBLIC OF KOREA
E-MAIL: kyunghee.stat@gmail.com
bupark@stats.snu.ac.kr