To appear in the *Journal of Nonparametric Statistics*
Vol. 00, No. 00, Month 20XX, 1–23

An alternative local polynomial estimator for the error-in-variables problem

Xianzheng Huang[a]* and Haiming Zhou[b]

[a]*Department of Statistics, University of South Carolina, Columbia, South Carolina, U.S.A.;*
[b]*Division of Statistics, Northern Illinois University, DeKalb, Illinois, U.S.A.*

(*v4.0 released June 2015*)

We consider the problem of estimating a regression function when a covariate is measured with error. Using the local polynomial estimator of Delaigle, Fan, and Carroll (2009) as a benchmark, we propose an alternative way of solving the problem without transforming the kernel function. The asymptotic properties of the alternative estimator are rigorously studied. A detailed implementing algorithm and a computationally efficient bandwidth selection procedure are also provided. The proposed estimator is compared with the existing local polynomial estimator via extensive simulations and an application to the motorcycle crash data. The results show that the new estimator can be less biased than the existing estimator and is numerically more stable.

Keywords: convolution; deconvolution; Fourier transform; measurement error.

AMS Subject Classification: 62G05; 62G08; 62G20

1. Introduction

The error-in-covariates problem has received great attention among researchers who study nonparametric inference for regression functions over the past two decades. Schennach (2004a,b) proposed an estimator of the regression function when the error-prone covariate is measured twice. Her estimator does not require a known measurement error distribution. Zwanzig (2007) proposed a local least square estimator of the regression function, assuming a uniformly distributed error-prone covariate with normal measurement error. Many more existing methods are developed under the assumption of a known measurement error distribution and an unknown true covariate distribution. Among these works, many follow the theme of deconvolution kernel pioneered in the density estimation problem in the presence of measurement error (Carroll and Hall 1988; Stefanski and Carroll 1990). In particular, starting from the well-known Nadaraya-Watson kernel estimator developed for error-free case (Nadaraya 1964; Watson 1964), Fan and Truong (1993) formulated the local constant estimator of a regression function using the deconvolution kernel technique. Generalization of this estimator to local polynomial estimators of higher orders was achieved by Delaigle et al. (2009) via introducing a complex transform of the kernel function. This transform is the key step that allows for the extension from the zero-order to a higher-order local polynomial estimator in error-in-variables problems.

*Corresponding author. Email: huang@stat.sc.edu

In this study, we propose a new estimator motivated by an identity that relates the Fourier transform of the functions to be estimated to the Fourier transform of the counterpart naive functions. Here, a naive estimate refers to an estimate that results from replacing the unobserved true covariate one would use in the absence of measurement error with the error-contaminated observed covariate. This identity and the new estimator are presented in Section 2, following a brief review of the estimator in Delaigle et al. (2009), which we refer to as the DFC estimator henceforth. Sections 3, 4, and 5 are devoted to studying the asymptotic distribution of the new estimator. The finite sample performance of our estimator is demonstrated in comparison with the DFC estimator in Section 6. We summarize our contribution and findings in Section 7. All appendices referenced in this article are provided in the Supplementary Materials.

2. Existing and proposed estimators

Denote by $\{(Y_j, W_j),\, j = 1, \ldots, n\}$ a random sample of size n from a regression model with additive measurement error in the covariate specified as follow,

$$E(Y_j|X_j) = m(X_j),\ W_j = X_j + U_j, \tag{1}$$

where X_j is the unobserved true covariate following a distribution with probability density function (pdf) $f_X(x)$, U_j is the measurement error, assumed to be independent of (X_j, Y_j) and follow a known distribution with pdf $f_U(u)$, W_j is the error-contaminated observed covariate following a distribution with pdf $f_W(w)$, for $j = 1, \ldots, n$. The problem of interest in this study is to estimate the regression function, $m(x)$, based on the observed data. The index j is often suppressed in the sequel when a generic observation or random variable is referenced.

2.1. *The DFC estimator*

In the absence of measurement error, the well-known local polynomial estimator of order p for $m(x)$ is given by (Fan and Gijbels 1996, Chapter 3)

$$\hat{m}(x) = \boldsymbol{e}_1^{\mathrm{T}} \mathbf{S}_n^{-1} \mathbf{T}_n, \tag{2}$$

where $\boldsymbol{e}_1$ is a $(p+1) \times 1$ vector with 1 in the first entry and 0 in the remaining p entries,

$$\mathbf{S}_n = \begin{bmatrix} S_{n,0}(x) & \ldots & S_{n,p}(x) \\ \vdots & \ddots & \vdots \\ S_{n,p}(x) & \ldots & S_{n,2p}(x) \end{bmatrix},$$

and $\mathbf{T}_n = (T_{n,0}(x), \ldots, T_{n,p}(x))^{\mathrm{T}}$, in which

$$\begin{cases} S_{n,\ell}(x) = n^{-1} \sum_{j=1}^{n} \left(\dfrac{X_j - x}{h} \right)^{\ell} K_h(X_j - x), \text{ for } \ell = 0, 1, \ldots, 2p, \\ T_{n,\ell}(x) = n^{-1} \sum_{j=1}^{n} Y_j \left(\dfrac{X_j - x}{h} \right)^{\ell} K_h(X_j - x), \text{ for } \ell = 0, 1, \ldots, p, \end{cases} \tag{3}$$

and $K_h(x) = h^{-1}K(x/h)$ with $K(\cdot)$ being a symmetric kernel function and h being the bandwidth.

Naive implementation of the above estimation in the presence of measurement error is to replace X_j with W_j, for $j = 1, \ldots, n$, in (3). Clearly, the resulting naive estimator of $m(x)$, denoted by $\hat{m}^*(x)$, is merely a sensible estimator of the naive regression function $m^*(x) = E(Y|W = x)$. Following the rationale behind the corrected score method (Carroll, Ruppert, Stefanski, and Crainiceanu 2006, Section 7.4), Delaigle et al. (2009) sought some function, denoted by $L_\ell(\cdot)$, that satisfies

$$E\left\{(W_j - x)^\ell L_{\ell,h}(W_j - x)|X_j\right\} = (X_j - x)^\ell K_h(X_j - x), \text{ for } \ell = 0, 1, \ldots, 2p, \tag{4}$$

where $L_{\ell,h}(x) = h^{-1}L_\ell(x/h)$. The authors derived such function via solving the Fourier transform version of (4), and showed that $L_\ell(x) = x^{-\ell}K_{U,\ell}(x)$, where

$$K_{U,\ell}(x) = i^{-\ell}\frac{1}{2\pi}\int e^{-itx}\frac{\phi_K^{(\ell)}(t)}{\phi_U(-t/h)}dt, \text{ for } \ell = 0, 1, \ldots, 2p, \tag{5}$$

in which $i = \sqrt{-1}$, $\phi_K^{(\ell)}(t)$ is the ℓ-th derivative of $\phi_K(t) = \int e^{itx}K(x)dx$, and $\phi_U(x)$ is the characteristic function of U. Throughout this article, ϕ_g denotes the Fourier transform of a function g if g is a function, and it denotes the characteristic function of g if g is a random variable. All integrals in this article integrate over either the entire real line or a subset of it that guarantees the existence of relevant integrals, and we will make remarks on such subset whenever it is needed for clarity. Now, besides substituting X_j with W_j in (3), replacing $K_h(X_j - x)$ there with $L_{\ell,h}(W_j - x)$ gives the DFC estimator, $\hat{m}_{\text{DFC}}(x) = \boldsymbol{e}_1^{\text{T}}\hat{\mathbf{S}}_n^{-1}\hat{\mathbf{T}}_n$, where $\hat{\mathbf{S}}_n$ and $\hat{\mathbf{T}}_n$ are similarly defined as $\mathbf{S}_n$ and $\mathbf{T}_n$ in (2) but with the elements in the matrices given by

$$\begin{cases} \hat{S}_{n,\ell}(x) = n^{-1}\sum_{j=1}^{n}\left(\frac{W_j - x}{h}\right)^\ell L_{\ell,h}(W_j - x), \text{ for } \ell = 0, 1, \ldots, 2p, \\ \hat{T}_{n,\ell}(x) = n^{-1}\sum_{j=1}^{n} Y_j\left(\frac{W_j - x}{h}\right)^\ell L_{\ell,h}(W_j - x), \text{ for } \ell = 0, 1, \ldots, p. \end{cases}$$

The transform of K defined in (5) is a brilliant extension of the transform used in Stefanski and Carroll (1990), based on which a deconvolution density estimator in the presence of measurement error is constructed, and also in Fan and Truong (1993), where the local constant estimator of $m(x)$ under the setting of (1) is proposed. In particular, the estimator in Fan and Truong (1993) is a special case of the DFC estimator with $p = 0$.

2.2. *An alternative estimator*

Deviating from the theme of deconvolution kernel and its extension in (5), we propose a new estimator that more directly exploits the naive inference as a whole. This direct use of the naive inference is motivated by the following result proved in Delaigle (2014), $m^*(w)f_W(w) = (mf_X) * f_U(w)$, where $(mf_X) * f_U(w)$ is the convolution given by $\int m(x)f_X(x)f_U(w - x)dx$. Applying Fourier transform on both sides of this identity, one

has

$$\phi_{m^* f_W}(t) = \phi_{m f_X}(t)\phi_U(t), \tag{6}$$

where $\phi_{m^* f_W}(t)$ is the Fourier transform of $m^*(w)f_W(w)$ and $\phi_{m f_X}(t)$ is the Fourier transform of $m(x)f_X(x)$. Immediately following (6), by the Fourier inversion theorem, one has $m(x)f_X(x) = (2\pi)^{-1}\int e^{-itx}\phi_{m^* f_W}(t)/\phi_U(t)dt$. This motivates our local polynomial estimator of order p for $m(x)$ given by, assuming the relevant Fourier transforms well defined,

$$\hat{m}_{\text{HZ}}(x) = \left\{\hat{f}_X(x)\right\}^{-1}\frac{1}{2\pi}\int e^{-itx}\frac{\phi_{\hat{m}^*\hat{f}_W}(t)}{\phi_U(t)}dt, \tag{7}$$

where $\hat{f}_X(x)$ is the deconvolution kernel density estimator of $f_X(x)$ in Stefanski and Carroll (1990), and $\phi_{\hat{m}^*\hat{f}_W}(t)$ is the Fourier transform of $\hat{m}^*(w)\hat{f}_W(w)$, in which $\hat{m}^*(w)$ is the p-th order local polynomial estimator of $m^*(w)$, and $\hat{f}_W(w)$ is the regular kernel density estimator of $f_W(w)$ (Fan and Gijbels 1996, Section 2.7.1), i.e., the naive estimator of $f_X(\cdot)$.

By its appearance, the new estimator in (7) results from applying an integral transform similar to that in (5) on the naive product $\hat{m}^*(\cdot)\hat{f}_W(\cdot)$ rather than on K. It can be shown (via straightforward algebra omitted here) that, when $p = 0$, this new estimator is the same as the DFC estimator, both reducing to the local constant estimator in Fan and Truong (1993). Other than this special case, $\hat{m}_{\text{HZ}}(x)$ differs from $\hat{m}_{\text{DFC}}(x)$ in general. A natural question is how they compare in terms of large sample properties and finite sample performance. The upcoming three sections illustrate the asymptotic properties of $\hat{m}_{\text{HZ}}(x)$, and demonstration of the finite sample performance is deferred to Section 6.

2.3. *Preamble for asymptotic analyses*

The majority of the theoretical development presented in Delaigle et al. (2009) revolve around properties of the transformed kernel, $K_{U,\ell}(x)$, which is not surprising as $K_{U,\ell}(x)$ is everywhere in the building blocks of their estimator. Because of the close tie between our proposed estimator and the naive estimators, much of our theoretical development builds upon well established results for kernel-based estimators of regression functions and density functions in the absence of measurement error. This can be better appreciated by interchanging the order of the two integrals in (7), assuming that $\phi_{\hat{m}^*\hat{f}_W}(t)$ is compactly supported on I_t (to allow the interchange), $\hat{m}_{\text{HZ}}(x)\hat{f}_X(x) = \int \hat{m}^*(w)\hat{f}_W(w)(2\pi)^{-1}\int_{I_t} e^{-it(x-w)}/\phi_U(t)\,dtdw$. This identity can be re-expressed more succinctly as

$$\mathcal{B}(x) = \int \mathcal{A}(w)D(x-w)\,dw = (\mathcal{A} * D)(x), \tag{8}$$

where $\mathcal{A}(w) = \hat{m}^*(w)\hat{f}_W(w)$, $\mathcal{B}(x) = \hat{m}_{\text{HZ}}(x)\hat{f}_X(x)$, and $D(s) = (2\pi)^{-1}\int_{I_t} e^{-its}/\phi_U(t)dt$. Note that, in (8), $\mathcal{A}(w)$ and $\mathcal{B}(x)$ are random processes, and $\mathcal{B}(x)$ results from convoluting $\mathcal{A}(w)$ and the non-random function $D(s)$. Because understanding the asymptotic properties of $\mathcal{B}(x) = \hat{m}_{\text{HZ}}(x)\hat{f}_X(x)$ brings one very close to understanding the properties of $\hat{m}_{\text{HZ}}(x)$, one may first study $\mathcal{B}(x)$ as an entirety. And (8) suggests that the question boils

down to, given the established results for the naive estimators in $\mathcal{A}(w)$, which are $\hat{m}^*(w)$ and $\hat{f}_W(w)$, what can be deduced for the counterpart non-naive estimator resulting from a convolution of $\mathcal{A}$ and D? More specifically, how do the moments of $\mathcal{A}$ compare with those of $\mathcal{B}$? And, if $\mathcal{A}(w)$ is a Gaussian process asymptotically, is $\mathcal{B}(x)$ also a Gaussian process asymptotically given this particular D? These questions about random process convolution are of mathematical interest in their own rights besides being the key to understanding $\hat{m}_{\text{HZ}}(x)$.

For clarity of exposition and comparison with the theoretical development and results for the DFC estimator, we dissect the asymptotic analyses of $\hat{m}_{\text{HZ}}(x)$ into three parts, first asymptotic bias, second asymptotic variance, and third asymptotic normality. Before starting the analyses, we shall provide two definitions of smoothness of a distribution (Fan 1991a; Fan, 1991b; Fan 1991c) and two sets of conditions to be referenced later, most of which are also stated in Delaigle et al. (2009).

Definition 1 The distribution of U is ordinary smooth of order b if

$$\lim_{t\to+\infty} t^b \phi_U(t) = c \text{ and } \lim_{t\to+\infty} t^{b+1}\phi_U'(t) = -cb$$

for some positive constants b and c.

Definition 2 The distribution of U is super smooth of order b if

$$d_0|t|^{b_0}\exp(-|t|^b/d_2) \le |\phi_U(t)| \le d_1|t|^{b_1}\exp(-|t|^b/d_2) \text{ as } |t|\to\infty$$

for some positive constants d_0, d_1, d_2, b, b_0 and b_1.

Condition O: For $\ell = 0,\ldots,2p+1$, $\|\phi_K^{(\ell)}(t)\|_\infty < \infty$ and $\int(|t|^b + |t|^{b-1})|\phi_K^{(\ell)}(t)|dt < \infty$. For $0 \le \ell_1, \ell_2 \le 2p$, $\int |t|^{2b}|\phi_K^{(\ell_1)}(t)||\phi_K^{(\ell_2)}(t)|dt < \infty$. And, $\|\phi_U'(t)\|_\infty < \infty$.

Condition S: For $\ell = 0,\ldots,2p$, $\|\phi_K^{(\ell)}(t)\|_\infty < \infty$, and $\phi_K(t)$ is supported on $[-1,1]$.

Besides the above conditions, we assume throughout the study that $f_X(x) > 0$ and $\phi_U(t)$ is an even function that never vanishes. We reach the convolution form in (8) under the assumption that $\phi_{\hat{m}^*\hat{f}_W}(t)$ is compactly supported on I_t, where I_t is a region that guarantees $D(s)$ well defined. This assumption can be easily satisfied by choosing a kernel of which the Fourier transform has a finite support. Even without this assumption the asymptotic properties presented in the following three sections still hold, although some of the proof need to be revised to use the estimator of its original form in (7). While acknowledging the overlap between the regularity conditions needed in our asymptotic analyses and those required for the DFC estimator, we also assume existence of the Fourier transform of $m^*(\cdot)f_W(\cdot)$ and that of $m(\cdot)f_X(\cdot)$ to reach (6). To satisfy this assumption, it suffices to have both products in L_1, which is not a practically stringent condition because, in practice, it is often a bounded support over which a regression function is of interest. Finally, when asymptotic is concerned, we let $n\to\infty$, and assume $h\to 0$ and $nh\to\infty$ as $n\to\infty$.

3. Asymptotic bias

We provide in Appendix A derivations of the asymptotic bias of $\hat{m}_{\text{HZ}}(x)$ for $p \ge 0$. To better apprehend the distinction between our bias results and those of $\hat{m}_{\text{DFC}}(x)$, we

present in this section a brief derivation of the bias when $p = 1$.

3.1. *Dominating bias when* $p = 1$

Define $\mu_\ell = \int u^\ell K(u)\,du$, for $\ell = 0, 1, \ldots, 2p$. Let $A(w) = m^*(w)f_W(w)$ and $B(x) = m(x)f_X(x)$ be the non-random counterparts of $\mathcal{A}(w)$ and $\mathcal{B}(x)$ in (8), respectively. Then, like (8), we have $B(x) = (A * D)(x)$.

Under the conditions in Theorem 2.1 in Stefanski and Carroll (1990), the deconvolution density estimator $\hat{f}_X(x)$ is a consistent estimator of $f_X(x)$. Having $\hat{f}_X(x)/f_X(x)$ converge to one in probability, we study the dominating terms in the bias, $E\{\hat{m}_{\text{HZ}}(x) - m(x)|\mathbb{W}\}$, via elaborating $E[\{\hat{m}_{\text{HZ}}(x) - m(x)\}\hat{f}_X(x)/f_X(x)|\mathbb{W}]$, where $\mathbb{W} = (W_1, \ldots, W_n)$. Note that the latter expectation is equal to

$$\{f_X(x)\}^{-1}\left[E\left\{\mathcal{B}(x)|\mathbb{W}\right\} - m(x)\hat{f}_X(x)\right], \tag{9}$$

where, by equation (1.9) and Theorem 2.1 in Stefanski and Carroll (1990),

$$\hat{f}_X(x) = f_X(x) + \mu_2 h^2 f_X^{(2)}(x)/2 + o_P(h^2). \tag{10}$$

To derive $E\{\mathcal{B}(x)|\mathbb{W}\}$ appearing in (9), we invoke the following two results for kernel-based estimators in the absence of measurement error (Fan and Gijbels 1996, Chapter 3),

$$\begin{aligned} E\left\{\hat{m}^*(w)|\mathbb{W}\right\} &= m^*(w) + \mu_2 m^{*(2)}(w)h^2/2 + o_P(h^2), \\ \hat{f}_W(w) &= f_W(w) + \mu_2 f_W^{(2)}(w)h^2/2 + o_P(h^2). \end{aligned}$$

Following these two results, one can show that

$$E\left\{\mathcal{A}(w)|\mathbb{W}\right\} = A(w) + \mu_2 M(w)h^2/2 + o_P(h^2), \tag{11}$$

where $M(w) = m^*(w)f_W^{(2)}(w) + m^{*(2)}(w)f_W(w)$. Then, assuming interchangeability of expectation and integration, (8) and (11) imply

$$E\{\mathcal{B}(x)|\mathbb{W}\} = \{E(\mathcal{A}|\mathbb{W}) * D\}(x) = B(x) + \mu_2 h^2 (M * D)(x)/2 + o_P(h^2). \tag{12}$$

Finally, by (10) and (12), (9) reduces to

$$\frac{\mu_2 h^2}{2f_X(x)}\left\{(M * D)(x) - m(x)f_X^{(2)}(x)\right\} + o_P(h^2), \tag{13}$$

which reveals the dominating bias of $\hat{m}_{\text{HZ}}(x)$ of order h^2.

Different from the bias analysis in Delaigle et al. (2009), deriving asymptotic bias here mostly involve direct use of existing results associated with estimators in the absence of measurement error, with (8) as the bridge leading one back to our estimator.

3.2. *Comparison with the bias of DFC estimator*

By Theorem 3.2 in Delaigle et al. (2009), the dominating bias of $\hat{m}_{\text{DFC}}(x)$ is the same as that of $\hat{m}(x)$, which is $\mu_2 h^2 m^{(2)}(x)/2$ when $p = 1$. To make the comparison of dominating bias more tractable without being overly restrictive in the shape of regression functions, we consider regression functions in the form of a polynomial of order r, $m(x) = \sum_{k=0}^{r} \beta_k x^k$. Furthermore, suppose $X \sim N(0,1)$ and $U \sim N(0, \sigma_u^2)$, resulting in a reliability ratio (Carroll et al. 2006, Section 3.2.1) of $\lambda = 1/(1+\sigma_u^2)$.

Under this setting, the dominating bias in (13) can be derived explicitly for any non-negative integer r. Instead of comparing the dominating bias associated with the two estimators directly for an arbitrary r, we focus on studying the number of x's at which each dominating bias is zero. The answer to this question for $\hat{m}_{\text{DFC}}(x)$ reveals itself because now $m^{(2)}(x)$ is a polynomial of order $r-2$ provided that $r \geq 2$, and thus the dominating bias associated with $\hat{m}_{\text{DFC}}(x)$ is zero at no more than $r-2$ x's for $r \geq 2$. In contrast, we show in Appendix A that the dominating bias in (13) reduces to ($\mu_2 h^2 \times$ a polynomial of order r), suggesting that the dominating bias of $\hat{m}_{\text{HZ}}(x)$ can be zero at r x's. Suppose that the bias of each estimator is continuous in x, which is a realistic assumption in many applications. Then having two more roots to the equation, dominating bias $= 0$, for $\hat{m}_{\text{HZ}}(x)$ indicates that this alternative estimator can have two more regions in the support of $m(x)$ within which $\hat{m}_{\text{HZ}}(x)$ is less biased than $\hat{m}_{\text{DFC}}(x)$, where each region is a neighborhood of some root.

For example, when $r = 2$, clearly the dominating bias of $\hat{m}_{\text{DFC}}(x)$ can never be zero. It is shown in Appendix A that, the dominating bias of $\hat{m}_{\text{HZ}}(x)$ is zero at the roots of the equation $2(\lambda-1)\beta_2 x^2 + (\lambda-1)\beta_1 x + (2\lambda^2 - 2\lambda + 1)\beta_2 = 0$. With $\lambda \in (0,1)$, one can easily show that this quadratic equation has two roots. This is an interesting phenomenon because $\hat{m}_{\text{DFC}}(x)$ has the same dominating bias as that of $\hat{m}(x)$, thus the finding here implies that $\hat{m}_{\text{HZ}}(x)$ can be less biased than the local polynomial estimator obtained from error-free data, at least asymptotically if not for a given finite sample. As counterintuitive as it sounds, plenty empirical evidence from our extensive simulation experiments, some presented in Section 6, suggest that $\hat{m}_{\text{HZ}}(x)$ can outperform $\hat{m}_{\text{DFC}}(x)$ in terms of bias. Moreover, we also observe in simulation study impressive gain in accuracy from the new estimator well beyond the regression function/covariate/measurement error configuration considered in this subsection.

4. Asymptotic variance

Because

$$\text{Var}\{\hat{m}_{\text{HZ}}(x)|\mathbb{W}\} = \text{Var}\left\{\mathcal{B}(x)|\mathbb{W}\right\} f_X^{-2}(x)\left\{1 + o_P(1)\right\}, \tag{14}$$

we focus on deriving $\text{Var}\{\mathcal{B}(x)|\mathbb{W}\}$ in order to study the asymptotic variance of $\hat{m}_{\text{HZ}}(x)$. Detailed derivations are provided in Appendix B, which consist of five steps. In what follows, we provide a sketch of the derivations, where we highlight the connection between our results and the counterpart results in the absence of measurement error, and how our derivations differ from and relate to those in Delaigle et al. (2009).

4.1. *Derivations of Var$\{\mathcal{B}(x)|\mathbb{W}\}$*

First, we deduce from (8) that $\mathrm{Var}\{\mathcal{B}(x)|\mathbb{W}\}$ can be formulated as an iterative convolution of the covariance of $\mathcal{A}(w)$ as follows,

$$\mathrm{Var}\{\mathcal{B}(x)|\mathbb{W}\} = \int D(x-w_1) \int D(x-w_2)\mathrm{Cov}\left\{\mathcal{A}(w_1), \mathcal{A}(w_2)|\mathbb{W}\right\} dw_2 dw_1. \tag{15}$$

Here, since $\hat{f}_W(w)/f_W(w)$ converges to 1 in probability under regularity conditions,

$$\mathrm{Cov}\left\{\mathcal{A}(w_1),\, \mathcal{A}(w_2)|\mathbb{W}\right\} = \mathrm{Cov}\{\hat{m}^*(w_1),\, \hat{m}^*(w_2)|\mathbb{W}\} f_W(w_1) f_W(w_2)\{1+o_P(1)\}. \tag{16}$$

This leads us to the next step, where we study $\mathrm{Cov}\{\hat{m}^*(w_1),\, \hat{m}^*(w_2)|\mathbb{W}\}$.

In the second step, we view $\hat{m}^*(w)$ as a weighted least squares estimator (Fan and Gijbels 1996, page 58), and show that

$$\mathrm{Cov}\{\hat{m}^*(w_1),\, \hat{m}^*(w_2)|\mathbb{W}\} = \boldsymbol{e}_1^{\mathrm{T}}(\mathbf{G}_1^{\mathrm{T}}\mathbf{W}_1\mathbf{G}_1)^{-1}(\mathbf{G}_1^{\mathrm{T}}\boldsymbol{\Sigma}_{12}\mathbf{G}_2)(\mathbf{G}_2^{\mathrm{T}}\mathbf{W}_2\mathbf{G}_2)^{-1}\boldsymbol{e}_1, \tag{17}$$

where $\boldsymbol{\Sigma}_{12} = \mathrm{diag}\{K_h(W_1-w_1)K_h(W_1-w_2)\nu^2(W_1), \ldots, K_h(W_n-w_1)K_h(W_n-w_2)\nu^2(W_n)\}$, $\nu^2(w) = \mathrm{Var}(Y|W=w)$, and, for $k=1,2$, $\mathbf{W}_k = \mathrm{diag}\{K_h(W_1-w_k), \ldots, K_h(W_n-w_k)\}$,

$$\mathbf{G}_k = \begin{bmatrix} 1 & (W_1-w_k) & \ldots & (W_1-w_k)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (W_n-w_k) & \ldots & (W_n-w_k)^p \end{bmatrix}.$$

Then we approximate the random quantities on the right hand side of (17) to establish that

$$\begin{aligned} &\mathrm{Cov}\{\hat{m}^*(w_1),\, \hat{m}^*(w_2)|\mathbb{W}\} \\ = \;& \frac{\nu^2\left\{(w_1+w_2)/2\right\} f_W\left\{(w_1+w_2)/2\right\}}{nhf_W(w_1)f_W(w_2)} \boldsymbol{e}_1^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{S}^*_{W,h}\mathbf{S}^{-1}\boldsymbol{e}_1 \left\{1+o_P\left(\frac{1}{nh}\right)\right\}, \end{aligned} \tag{18}$$

where $\mathbf{S} = (\mu_{\ell_1+\ell_2})_{0\le\ell_1,\ell_2\le p}$ and $\mathbf{S}^*_{W,h} = (\xi_{\ell_1,\ell_2}((w_1-w_2)/2, h))_{0\le\ell_1,\ell_2\le p}$, in which, for $\ell_1, \ell_2 = 0, 1, \ldots, p$,

$$\xi_{\ell_1,\ell_2}(w,h) = \int (u-w/h)^{\ell_1}(u+w/h)^{\ell_2}K(u-w/h)K(u+w/h)du. \tag{19}$$

The result in (18) is a counterpart result for $\mathrm{Var}\{\hat{m}(x)|\mathbb{X}\}$, where $\mathbb{X} = (X_1, \ldots, X_n)$ (Fan and Gijbels 1996, equation (3.7)).

Third, substituting (18) in (16) gives

$$\mathrm{Cov}\left\{\mathcal{A}(w_1),\, \mathcal{A}(w_2)|\mathbb{W}\right\} = \frac{\gamma\left\{(w_1+w_2)/2\right\}}{nh} \boldsymbol{e}_1^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{S}^*_{W,h}\mathbf{S}^{-1}\boldsymbol{e}_1 \left\{1+o_P\left(\frac{1}{nh}\right)\right\}, \tag{20}$$

where $\gamma(w) = \nu^2(w) f_W(w)$. And plugging (20) in (15) yields

$$\begin{aligned}\mathrm{Var}\left\{\mathcal{B}(x)|\mathbb{W}\right\} &= \int D(x-w_1)\int D(x-w_2)\times \\ &\left[\frac{\gamma\left\{(w_1+w_2)/2\right\}}{nh}\boldsymbol{e}_1^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{S}_{W,h}^{*}\mathbf{S}^{-1}\boldsymbol{e}_1\left\{1+o_P\left(\frac{1}{nh}\right)\right\}\right]dw_2dw_1.\end{aligned} \tag{21}$$

Note that, among the matrices in (21), only $\mathbf{S}_{W,h}^{*}$ depends on w_1 and w_2, of which the entries are $\xi_{\ell_1,\ell_2}(w,h)$ in (19). This leads us to the fourth step of the derivations, where we derive

$$\int D(x-w_1)\int D(x-w_2)\gamma\left(\frac{w_1+w_2}{2}\right)\xi_{\ell_1,\ell_2}\left(\frac{w_1-w_2}{2},h\right)dw_2dw_1. \tag{22}$$

The fourth step entails solving and approximating a multi-dimensional integral, by the end of which we show that (22) is equal to

$$\left\{\gamma(x)+O(h)\right\}\int K_{U,\ell_1}(v)K_{U,\ell_2}(v)\,dv. \tag{23}$$

Define $\kappa_{\ell_1,\ell_2}(h) = \int K_{U,\ell_1}(v)K_{U,\ell_2}(v)\,dv$ to highlight the dependence of this integral on h (since $K_{U,\ell}(v)$ depends on h according to (5)), and define matrix $\mathbf{K}(h) = (\kappa_{\ell_1,\ell_2}(h))_{0\le\ell_1,\ell_2\le p}$. To this end, we can conclude that, by (21) and (23),

$$\mathrm{Var}\left\{\mathcal{B}(x)|\mathbb{W}\right\} = \frac{\gamma(x)}{nh}\boldsymbol{e}_1^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{K}(h)\mathbf{S}^{-1}\boldsymbol{e}_1\left\{1+o_P\left(\frac{1}{nh}\right)\right\}. \tag{24}$$

This is where the path of our derivations meets that of Delaigle et al. (2009), as now we need to incorporate the properties of $\kappa_{\ell_1,\ell_2}(h)$ as $n\to\infty$ (and thus $h\to 0$), for an ordinary smooth U and for a super smooth U, respectively, to move forward from (24). These properties are thoroughly studied in Delaigle et al. (2009) and summarized in their Lemmas B.4, B.6, B.9, which are restated in Appendix B for completeness. Equipped with these lemmas, we are ready to move on to the fifth step of the derivations.

By Lemma B.4, for an ordinary smooth U, under Condition O, $\kappa_{\ell_1,\ell_2}(h) = h^{-2b}\eta_{\ell_1,\ell_2} + o\left(h^{-2b}\right)$ as $n\to\infty$, where

$$\eta_{\ell_1,\ell_2} = i^{-\ell_1-\ell_2}(-1)^{-\ell_2}c^{-2}(2\pi)^{-1}\int |t|^{2b}\phi_K^{(\ell_1)}(t)\phi_K^{(\ell_2)}(t)\,dt,$$

in which b and c are constants in Definition 1. Define $\mathbf{S}^{*} = (\eta_{\ell_1,\ell_2})_{0\le\ell_1,\ell_2\le p}$, then $\mathbf{K}(h) = h^{-2b}\mathbf{S}^{*} + o\left(h^{-2b}\right)$, and thus (24) implies (25) in Theorem 4.1 below. For a super smooth U, by Lemma B.9, under Condition S, $|\kappa_{\ell_1,\ell_2}(h)| \le Ch^{2b_2}\exp(2h^{-b}/d_2)$, where $b_3 = b_0 I(b_0 < 0.5)$, b_0, b and d_2 are constants in Definition 2, and C is some generic non-negative finite constant appearing in Lemma B.8 in Delaigle et al. (2009). This leads to (26) in Theorem 4.1 below, which serves as a recap of our findings in this subsection.

THEOREM 4.1 *When U is ordinary smooth of order b, under Condition O, if $nh^{2b+1}\to$*

∞, *then*

$$Var\{\hat{m}_{HZ}(x)|\mathbb{W}\} = \boldsymbol{e}_1^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{S}^{*}\mathbf{S}^{-1}\boldsymbol{e}_1\frac{\gamma(x)}{f_X^2(x)nh^{2b+1}} + o_P\left(\frac{1}{nh^{2b+1}}\right). \tag{25}$$

When U is super smooth of order b, under Condition S, if $n\exp(2h^b/d_2)h^{1-2b_3} \to \infty$, then $Var\{\hat{m}_{HZ}(x)|\mathbb{W}\}$ is bounded from above by

$$\boldsymbol{e}_1^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{S}^{-1}\boldsymbol{e}_1\frac{C\gamma(x)h^{2b_3-1}}{f_X^2(x)n\exp(2h^b/d_2)} + o_P\left\{\frac{h^{2b_3-1}}{n\exp(2h^b/d_2)}\right\}. \tag{26}$$

4.2. *Comparison with the variance of DFC estimator*

By Theorem 3.1 in Delaigle et al. (2009), when the distribution of U is ordinary smooth, under Condition O, if $nh^{2b+1} \to \infty$, then

$$\mathrm{Var}\left\{\hat{m}_{\mathrm{DFC}}(x)\right\} = \boldsymbol{e}_1^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{S}^{*}\mathbf{S}^{-1}\boldsymbol{e}_1\frac{(\tau^2 f_X) * f_U(x)}{f_X^2(x)nh^{2b+1}} + o\left(\frac{1}{nh^{2b+1}}\right), \tag{27}$$

where $\tau^2(x) = \mathrm{Var}(Y|X = x)$. One may notice that the asymptotic variance results in Theorem 4.1, as well as the asymptotic bias results in Section 3, are conditional on $\mathbb{W}$ whereas (27) is an unconditional variance. The conditional arguments in our moment analysis originate from the direct use of asymptotic moments of the local polynomial estimator of a regression function in the absence of measurement error, which are conditional moments given $\mathbb{X}$ (Ruppert and Wand 1994). As pointed out in Ruppert and Wand (1994, Remark 1, page 1351), because the dominating terms in these conditional moments are free of $\mathbb{W}$, they still have the interpretation of unconditional dominating moments. Once this is clear, one can see that the difference between the dominating variance in (27) and that in (25) lies in the distinction between $(\tau^2 f_X) * f_U(x)$ and $\gamma(x)$. It is shown in Appendix B that $\gamma(x) = (\tau^2 f_X) * f_U(x) + f_W(x)\mathrm{Var}\{m(X)|W = x\} \geq (\tau^2 f_X) * f_U(x)$. Hence, for an ordinary smooth U, the dominating variance of $\hat{m}_{\mathrm{HZ}}(x)$ is greater than or equal to that of $\hat{m}_{\mathrm{DFC}}(x)$.

In Section 6, we will investigate via simulations to what extent the implication of this large sample comparison take effect in the comparison of finite sample variances associated with the two estimators. Also to be monitored in simulations are mean squared errors of the two estimators, which is very much of interest now that the asymptotic bias and variance analyses do not lead to one estimator that wins in both aspects.

5. Asymptotic normality

Under the conditions stated in Theorem 4.1, we show the asymptotic normality of $\hat{m}_{\mathrm{HZ}}(x)$ in Appendix C. The logic behind the proof is similar to that in Delaigle et al. (2009). More specifically, we first approximate $\mathcal{B}(x) - B(x)$ via an average, $n^{-1}\sum_{j=1}^{n}\tilde{U}_{n,j}(x)$, where $\{\tilde{U}_{n,j}(x)\}_{j=1}^{n}$ is a set of independent and identically distributed (i.i.d.) random variables at each fixed x. Then we show that, for some positive constant η,

$$\lim_{n\to\infty}\frac{E|\tilde{U}_{n,1}|^{2+\eta}}{n^{\eta/2}\{E(\tilde{U}_{n,1}^2)\}^{(2+\eta)/2}} = 0,$$

which is a sufficient condition for

$$\frac{\sum_{j=1}^{n} \tilde{U}_{n,j} - nE(\tilde{U}_{n,j})}{\sqrt{n\mathrm{Var}(\tilde{U}_{n,j})}} \xrightarrow{\mathcal{L}} N(0,1).$$

This in turn leads to the asymptotic normality of $\mathcal{B}(x) - B(x)$, and further suggests the asymptotic normality of $\hat{m}_{\mathrm{HZ}}(x)$.

To this end, we have answered the questions raised in Section 2.3 regarding the properties of a random process $\mathcal{B}(x)$ resulting from the convolution of another random process $\mathcal{A}(w)$ and the non-random function $D(s)$. We now see that the first two moments of $\mathcal{B}(x)$ are closely related to the the first two moments of $\mathcal{A}(w)$ via similar convolutions. Also, if $\mathcal{A}(w)$ is asymptotically Gaussian, then under mild regularity conditions, $\mathcal{B}(x)$ is also asymptotically Gaussian, and many of these conditions can be satisfied by choosing an appropriate kernel function in $\mathcal{A}(w)$.

6. Implementation and finite sample performance

After a thorough investigation of asymptotic properties of the proposed estimator, we are now in the position to look into its finite sample performance. By the construction of $\hat{m}_{\mathrm{HZ}}(x)$, computing the estimate requires evaluating continuous Fourier transforms (CFT) and an inverse CFT. In this section, we first describe the algorithm for these evaluations, then we discuss bandwidth selection. Finally, we present experiments where we compare our estimator with the DFC estimator under four settings where we simulate data from the true models with our design of $m(x)$, and one setting where error-prone data are simulated from a motorcycle-crash data set with the underlying $m(x)$ unknown.

6.1. *Numerical evaluations*

For an integrable function that maps the real line onto the complex space, $f : \mathbb{R} \to \mathbb{C}$, define the CFT of f as

$$\mathcal{F}[f](t) = \int_{-\infty}^{\infty} f(s)e^{-its}ds, \quad \forall t \in \mathbb{R}. \tag{28}$$

In our study, we first approximate the CFT via a discrete Fourier transform (DFT), then we use the fast Fourier transform algorithm (FFT, Bailey and Swarztrauber 1994) to evaluate the corresponding DFT. For a sequence of G complex values $\boldsymbol{z} = \{z_0, \ldots, z_{G-1}\}$, the DFT is defined as $D_k[\boldsymbol{z}] = \sum_{g=0}^{G-1} z_g e^{-i2\pi kg/G}$, for $k = 0, \ldots, G-1$, which can be easily evaluated using FFT in standard statistical software. The approximation of CFT using DFT is sketched next.

To prepare for the approximation, one first specifies a sequence of input values and then specifies a sequence of output values accordingly. More specifically, let $\{s_g = (g - G/2)\alpha_1,\ g = 0, 1, \ldots, G-1\}$ be the input values for the CFT, where $G/2$ is an even integer, $\alpha_1 = a/G$ is the increment, and a is chosen such that (28) can be well approximated by $\int_{-a/2}^{a/2} f(s)e^{-its}ds$. With the input values specified, the corresponding output values are $\{t_k = (k - G/2)\alpha_2,\ k = 0, 1, \ldots, G-1\}$, where $\alpha_2 = 2\pi/(G\alpha_1)$. With the input

and output values ready, we approximate the CFT as follows, for $k = 0, 1, \ldots, G-1$,

$$\begin{aligned}
\mathcal{F}[f](t_k) &\approx \int_{-a/2}^{a/2} f(s)e^{-it_k s}ds \\
&\approx \sum_{g=0}^{G-1} f(s_g)e^{-it_k s_g}\alpha_1 \\
&= \alpha_1 \sum_{g=0}^{G-1} f(s_g)e^{-i(k-G/2)\alpha_2(g-G/2)\alpha_1} \\
&= \alpha_1 e^{i(k-G/2)\pi} \sum_{g=0}^{G-1} f(s_g)e^{i\pi g}e^{-i2\pi kg/G} \\
&= \alpha_1(-1)^k D_k\left[\{(-1)^g f(s_g)\}\right].
\end{aligned}$$

This step function approximation converges to the truth very rapidly provided that the Fourier coefficients of f rapidly decrease (Davis and Rabinowitz 1984). The values of α_1 and α_2 determine the resolution of the input and output results, respectively. Comparable resolutions in s and t are typically desired, which can be achieved by setting $\alpha_1 = \alpha_2 = \sqrt{2\pi/G}$. A larger G tends to yield a more accurate approximation of the CFT. Bailey and Swarztrauber (1994) computed the CFT of the standard normal density function using $G = 2^{16}$ and achieved the root-mean-squared error of order 10^{-16}. In the simulations presented in this article, we set $G = 2^{16}$, resulting in $\alpha_1 = \alpha_2 \approx 0.01$, and both input and output values within $[-320.8, 320.8]$. In additional simulation studies where we used a larger G, we found the results essentially unchanged. This algorithm can be similarly applied to approximate the inverse CFT.

6.2. *Bandwidth selection*

It has been well acknowledged that the choice of bandwidth is crucial in kernel-based nonparametric estimation. In our study, we adopt the method of cross-validation (CV) in conjunction with simulation extrapolation (SIMEX, Carroll et al. 2006, Chapter 5) as proposed by Delaigle and Hall (2008). To implement this method, one first randomly divides the observed data, $\{(Y_j, W_j)\}_{j=1}^n$, into δ subsamples of (nearly) equal size. Denote by $\mathcal{D}_k$ the kth subsample, and I_k the set of subject indices corresponding to the observations in $\mathcal{D}_k$, for $k = 1, \ldots, \delta$. Then one carries out two rounds of δ-fold cross validation using further contaminated data. In the first round, one generates further contaminated data according to $W_{b,j}^* = W_j + U_{b,j}^*$, for $b = 1, \ldots, B$ and $j = 1, \ldots, n$, where $\{U_{b,j}^*, b = 1, \ldots, B\}_{j=1}^n$ are i.i.d. according to $f_U(u)$. Viewing $\mathbb{W}$ as the "unobserved true" covariate values, and $m^*(x) = E(Y|W = x)$ as the target regression function to be estimated using the "observed" data, $\{(Y_j, W_{b,j}^*)\}_{j=1}^n$, for $b = 1, \ldots, B$, one may use the proposed method to estimate $m^*(x)$. Denote this estimator by $\hat{m}_{\text{HZ}}^*(x)$. Now one carries out the δ-fold cross validation to choose a bandwidth for estimating $m^*(x)$ that minimizes

$$\text{CV}_1(h) = \frac{1}{nB}\sum_{b=1}^{B}\sum_{k=1}^{\delta}\sum_{j\in I_k}\left\{Y_j - \hat{m}_{\text{HZ},b}^{*(-k)}(W_j)\right\}^2 w(W_{i,j}),$$

where $\hat{m}^{*(-k)}_{\mathrm{HZ},b}(x)$ is the estimate $\hat{m}^{*}_{\mathrm{HZ}}(x)$ computed using the further contaminated data excluding $\mathcal{D}_k$, for $k = 1, \ldots, \delta$, and $w(\cdot)$ is a suitable weight function. Define $\hat{h}_1 = \mathrm{argmin}_{h>0}\mathrm{CV}_1(h)$. In the second round of δ-fold cross validation, another set of further contaminated data is produced according to $W^{**}_{b,j} = W^{*}_{b,j} + U^{**}_{b,j}$, where $\{U^{**}_{b,j}, b = 1, \ldots, B\}_{j=1}^{n}$ are i.i.d. according to $f_U(u)$, for $b = 1, \ldots, B$ and $j = 1, \ldots, n$, also independent of $\{U^{*}_{b,j}, b = 1, \ldots, B\}_{j=1}^{n}$. Similar to the first round, one views $\mathbb{W}^* = \{W^{*}_{b,j}, b = 1, \ldots, B\}_{j=1}^{n}$ as the "unobserved true" covariate values, and considers estimating another target regression function $m^{**}(x) = E(Y|W^* = x)$ using the proposed method based on the "observed" data $\{(Y_j, W^{**}_{b,j})\}_{j=1}^{n}$, for $b = 1, \ldots, B$. Denote this estimator by $\hat{m}^{**}_{\mathrm{HZ}}(x)$. To select a bandwidth for estimating $m^{**}(x)$, one minimizes the following criterion with respect to h,

$$\mathrm{CV}_2(h) = \frac{1}{nB}\sum_{b=1}^{B}\sum_{k=1}^{\delta}\sum_{j\in I_k}\left\{Y_j - \hat{m}^{**(-k)}_{\mathrm{HZ},b}(W^{*}_{b,j})\right\}^2 w(W^{*}_{b,j}),$$

where $\hat{m}^{**(-k)}_{\mathrm{HZ},b}(x)$ is the estimate $\hat{m}^{**}_{\mathrm{HZ}}(x)$ computed using the data $\{(Y_j, W^{**}_{b,j})\}_{j=1}^{n}$ excluding $\mathcal{D}_k$, for $k = 1, \ldots, \delta$. Define $\hat{h}_2 = \mathrm{argmin}_{h>0}\ \mathrm{CV}_2(h)$. Finally, one sets $\hat{h} = \hat{h}_1^2/\hat{h}_2$ as the bandwidth used in $\hat{m}_{\mathrm{HZ}}(x)$ for estimating $m(x)$ based on the original observed data $\{(Y_j, W_j)\}_{j=1}^{n}$.

This bandwidth selection procedure can be computationally cumbersome because, first, in search of $\hat{h}_1$ and $\hat{h}_2$, one evaluates $\mathrm{CV}_1(h)$ and $\mathrm{CV}_2(h)$ on a fine grid of candidate bandwidths; second, as recommended in most SIMEX applications, one needs a B not too small in order to control the Monte Carlo variability when generating further contaminated data and drawing inference based on them repeatedly. To lessen the computational burden, we propose a procedure to refine the search region of h. Take the first round of cross validation described above as an example. Recall that, during this round, $\mathbb{W}$ is viewed as the unobserved true covariate values whereas $\mathbb{W}^*$ is the error-contaminated version of the true covariate values. To narrow down the search region of h when minimizing $\mathrm{CV}_1(h)$, we first find an initial bandwidth, $\tilde{h}_1$. In particular, we obtain $\tilde{h}_1$ by minimizing the following approximated mean integrated squared error (MISE) for the deconvolution kernel density estimator of $f_W(w)$ using $\mathbb{W}^*$ (Stefanski and Carroll 1990),

$$\mathrm{MISE}(h) = \frac{1}{2\pi nh}\int\frac{|\phi_K(t)|^2}{|\phi_U(t/h)|^2}dt + \frac{h^4}{4}\int\left\{f''_W(w)\right\}^2 dw\int x^2K(x)dx, \qquad (29)$$

where $\int\{f''_W(w)\}^2dw$ can be easily estimated using $\mathbb{W}$. After $\tilde{h}_1$ is found, we focus on L grid points within $[0.2\tilde{h}_1, 2\tilde{h}_1]$ when searching for $\hat{h}_1$. This strategy of making use of $\tilde{h}_1$ to zoom in on a search region for $\hat{h}_1$ is motivated by the theoretical finding that the deconvolution kernel regression estimators have the same optimal rates as the deconvolution kernel density estimators. In our extensive trial-and-error simulation experiments under the model settings described in Section 6.3, we considered a wider search region that encompasses $[0.2\tilde{h}_1, 2\tilde{h}_1]$, and we observed all selected h indeed fall in the above refined search region. Similarly, in the second round of cross validation where we search for $\hat{h}_2$ that minimizes $\mathrm{CV}_2(h)$, we search across L grid points within $[0.2\tilde{h}_2, 2\tilde{h}_2]$, where $\tilde{h}_2$ is chosen by minimizing (29), but, different from the first round, now $\int\{f''_W(w)\}^2dw$ there is replaced by $\int\{f''_{W^*}(w)\}^2dw$, which can be easily estimated using $\mathbb{W}^*$.

One may legitimately question our choice of the multiplicative factors, 0.2 and 2, in the

recommended refined search region of h. For a given application, the safe and conservative way to choose h usually involves some trial-and-error. We also recommend that, if the optimal h found within this refined region is too close to one of the boundaries, one may consider pushing that end of the region out slightly and adjusting the search region accordingly. Even with this adjustment and seeking an optimal h in the adjusted region once again, one often still saves some computation time than looking for an optimal h over a wider region formed not based on $\tilde{h}$. Also observed in simulation studies is that, with the so-obtained search region of h at each round of cross validation, one can even use a much smaller B without noticeably compromising the quality of $\hat{m}_{\text{HZ}}(x)$ compared to when a much larger B is used. This refined bandwidth selection procedure and the algorithm for approximating CFT and inverse CFT described in Section 6.1 are implemented in an `R` package called `lpme` created and maintained by the second author, which provides both $\hat{m}_{\text{HZ}}(x)$ and $\hat{m}_{\text{DFC}}(x)$.

6.3. *Simulation study*

In the simulation experiments, we compare realizations of $\hat{m}_{\text{HZ}}(x)$ and $\hat{m}_{\text{DFC}}(x)$ (with $p = 1$) obtained under the following four model configurations:

(C1) $[Y|X = x] \sim N(m(x), 0.2^2)$, where $m(x) = 2x\exp(-10x^4/81)$, $X = 0.8X_1 + 0.2X_2$, $X_1 \sim f_{X_1}(x) = 0.1875x^2 I_{[-2,2]}(x)$, $X_2 \sim$ uniform$(-1, 1)$, and $U \sim$ Laplace$(0, \sigma_u/\sqrt{2})$.

(C2) $[Y|X = x] \sim N(m(x), 0.5^2)$, where $m(x) = (x + x^2)/4$, $X \sim N(0, 1)$, and $U \sim N(0, \sigma_u^2)$.

(C3) $[Y|X = x] \sim N(m(x), 0.2^2)$, where $m(x) = x^6/30 - 5x^4/6 + 9x^2/2 + x$, $X \sim$ uniform$(-2, 2)$, and $U \sim$ Laplace$(0, \sigma_u/\sqrt{2})$.

(C4) $[Y|X = x] \sim N(m(x), 0.2^2)$, where $m(x) = \cos(x^2) + \sin(x)$, $X \sim$ uniform$(-2, 2)$, and $U \sim$ Laplace$(0, \sigma_u/\sqrt{2})$.

Among these configurations, (C1) is considered in Delaigle et al. (2009); (C2) creates a scenario where the dominating bias of $\hat{m}_{\text{DFC}}(x)$ never vanishes since $m(x)$ is a second-order polynomial; (C3), with $m(x)$ being a higher order polynomial, results in zero dominating bias for $\hat{m}_{\text{DFC}}(x)$ within the support of X at ± 1; and (C4) has $m(x)$ out of the polynomial family yet it can be expanded as a polynomial of infinite order. Besides the model configuration, we also vary the reliability ratio $\lambda = \text{Var}(X)/\{\text{Var}(X) + \sigma_u^2\}$ from 0.7 to 0.95 at increments of 0.05 when generating $\mathbb{W}$. And, under (C2), although the measurement errors are simulated from a normal distribution, we computed the estimates of $m(x)$ assuming a normal U first, and then we repeated the estimation assuming U follows a Laplace distribution. This exercise allows us to observe the effects of a misspecified distribution for U on the estimates. Under each simulation setting, 500 Monte Carol (MC) replicates of sample size $n = 500$ are generated from the true model of (Y, W). For both estimation methods, we used the kernel of which the Fourier transform is given by $\phi_K(t) = (1 - t^2)^8 I_{[-1,1]}(t)$.

Denote by $\hat{m}_{[\cdot]}(x)$ one of the two estimates under comparison generically. For the majority of the simulation experiments, in order to mitigate the confounding effect of a data-driven bandwidth selection method on the quality of $\hat{m}_{[\cdot]}(x)$, we computed $\hat{m}_{[\cdot]}(x)$ using the theoretical optimal bandwidth obtained via minimizing an approximate of the integrated squared error (ISE), ISE $= \int_{x_{\text{L}}}^{x_{\text{U}}} \{\hat{m}_{[\cdot]}(x) - m(x)\}^2 dx$, where $[x_{\text{L}}, x_{\text{U}}]$ is the interval of the true covariate value of interest. This approximated ISE is given by $\sum_{k=0}^{\mathcal{M}} \{\hat{m}_{[\cdot]}(x_k) - m(x_k)\}^2 \Delta$, where Δ is the partition resolution, $\mathcal{M}$ is the largest integer

no greater than $(x_{\mathrm{U}} - x_{\mathrm{L}})/\Delta$, and $x_k = x_{\mathrm{L}} + k\Delta$, for $k = 0, \ldots, \mathcal{M}$. For a small portion of the presented simulation experiments, we used the CV-SIMEX bandwidth selection strategy described in Section 6.2 to select a bandwidth for each of the two estimators. Note that, when choosing a bandwidth for $\hat{m}_{\mathrm{DFC}}(x)$, one should change $\hat{m}^{*}_{\mathrm{HZ}}(x)$ and $\hat{m}^{**}_{\mathrm{HZ}}(x)$ in Section 6.2 to the counterpart estimates $\hat{m}^{*}_{\mathrm{DFC}}(x)$ and $\hat{m}^{**}_{\mathrm{DFC}}(x)$, respectively.

We compare the performance of $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$ with regard to the quality of the entire regression curve estimation over $[x_{\mathrm{L}}, x_{\mathrm{U}}]$, as well as the quality of the estimation of $m(x)$ at individual x's. The quantity used to monitor the overall regression curve estimation is the approximated ISE. And the quantity collected in the simulation to assess the quality of $\hat{m}_{[\cdot]}(x)$ at a particular point $x = x_0$ is the pointwise absolute error (PAE) defined by $\mathrm{PAE}(x_0) = |\hat{m}_{[\cdot]}(x_0) - m(x_0)|$. To better summarize the comparison regarding the accuracy and precision of $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$ at each of the $\mathcal{M}_1$ points of x used to approximate the ISE across 500 MC replicates, we computed the following three summary statistics: first, the pointwise mean absolute error ratio (PmAER) defined by

$$\mathrm{PmAER}(x_k) = \frac{\text{MC average of } |\hat{m}_{\mathrm{HZ}}(x_k) - m(x_k)|}{\text{MC average of } |\hat{m}_{\mathrm{DFC}}(x_k) - m(x_k)|};$$

second, the pointwise standard deviation of absolute error ratio (PsdAER) defined by

$$\mathrm{PsdAER}(x_k) = \frac{\text{MC standard deviation of } |\hat{m}_{\mathrm{HZ}}(x_k) - m(x_k)|}{\text{MC standard deviation of } |\hat{m}_{\mathrm{DFC}}(x_k) - m(x_k)|};$$

and third, the pointwise mean squared error ratio (PMSER) defined by

$$\mathrm{PMSER}(x_k) = \frac{\text{MC average of } |\hat{m}_{\mathrm{HZ}}(x_k) - m(x_k)|^2}{\text{MC average of } |\hat{m}_{\mathrm{DFC}}(x_k) - m(x_k)|^2}.$$

These quantities are presented in Figures 1–4 for (C1)–(C4), respectively. Figure 5 shows the counterpart results of Figure 2 under (C2) when it is (incorrectly) assumed that U follows a Laplace distribution. These five figures depict results obtained when the theoretical optimal h is used. Lastly, Figure 6 is the counterpart of Figure 4 under (C4) with h chosen by the CV-SIMEX bandwidth selection procedure with $B = 10$ and $L = 10$. Very similar performance of the two estimates are observed when larger values of B or L are used in this round of experiment.

When the theoretical optimal bandwidth is used, as in Figures 1–5, $\hat{m}_{\mathrm{HZ}}(x)$ outperforms $\hat{m}_{\mathrm{DFC}}(x)$ over the majority region of each considered range of x in regard to both accuracy and precision. Even though it is shown in Section 4.2 that the dominating variance of $\hat{m}_{\mathrm{HZ}}(x)$ is higher than that of $\hat{m}_{\mathrm{DFC}}(x)$ when U is ordinary smooth (such as when U follows a Laplace distribution), this large sample trend does not take effect for the majority region of x in these finite sample experiments. The regions where $\hat{m}_{\mathrm{DFC}}(x)$ performs better than $\hat{m}_{\mathrm{HZ}}(x)$ in regard to bias, variance, and MSE are usually neighborhoods of the inflection points of $m(x)$. For instance, under (C3) (see panel (i) in Figure 3), $\hat{m}_{\mathrm{DFC}}(x)$ is less biased than $\hat{m}_{\mathrm{HZ}}(x)$ at the small neighborhoods of ± 1. It is worth pointing out that the gain in accuracy and precision from our estimator compared to the DFC estimator is especially promising at the boundary of x in (C3) and (C4) (see panels (c), (f), and (i) in Figures 3 and 4). In both cases, data points relatively evenly spread over the domain of $m(x)$. Different from (C3) and (C4), in (C1), there are more data points near the boundaries than elsewhere in the domain. Excluding (C2) (since the plotted range of x

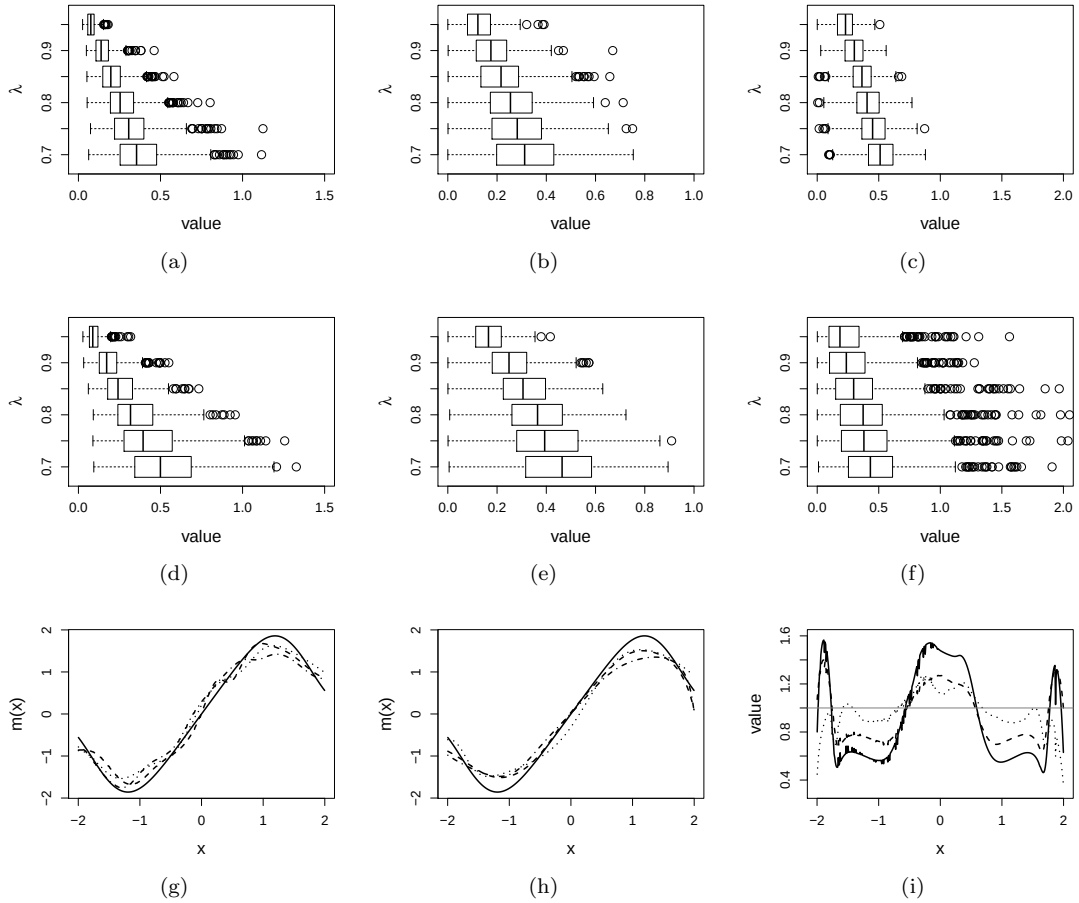

Figure 1. Simulation results under (C1) using the theoretical optimal h. Panels (a) & (d): boxplots of ISEs versus λ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively. Panels (b) & (e): boxplots of PAE(1) versus λ for $\hat{m}_{\mathrm{HZ}}(1)$ and $\hat{m}_{\mathrm{DFC}}(1)$, respectively. Panels (c) & (f): boxplots of PAE(2) versus λ for $\hat{m}_{\mathrm{HZ}}(2)$ and $\hat{m}_{\mathrm{DFC}}(2)$, respectively. Panels (g) & (h): quantile curves when $\lambda = 0.85$ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively, based on ISEs (dashed lines for the first quartile, dotted lines for the second quartile, and dot-dashed lines for the third quartile, solid lines for the truth). Panel (i): PmAER (dashed line), PsdAER (dotted line), and PMSER (solid line) versus x when $\lambda = 0.85$; the horizontal reference line highlights the value 1.

in Figures 2 and 5 is not the entire observed range), (C1) is the only case among all considered cases here that $\hat{m}_{\mathrm{DFC}}(x)$ outperforms $\hat{m}_{\mathrm{HZ}}(x)$ near the boundaries in terms of bias, although $\hat{m}_{\mathrm{HZ}}(x)$ is still substantially less variable there, and its MSE is lower than that of the competing estimator (see panels (c), (f), and (i) of Figure 1). Finally, contrasting Figure 2 and Figure 5, one can see that both estimators are fairly robust to the misspecification of the measurement error distribution.

When the bandwidth chosen by the refined CV-SIMEX method is used, as in Figure 6, both estimates become more variable, with our estimates better than the DFC estimates over most of the 500 MC replicates. As mentioned earlier, increasing B to a larger value does not substantially change our estimate. More importantly, using a B smaller than ten affects our estimator far less than it affects the DFC estimator, which suffers tremendous numerical instability when B is smaller.

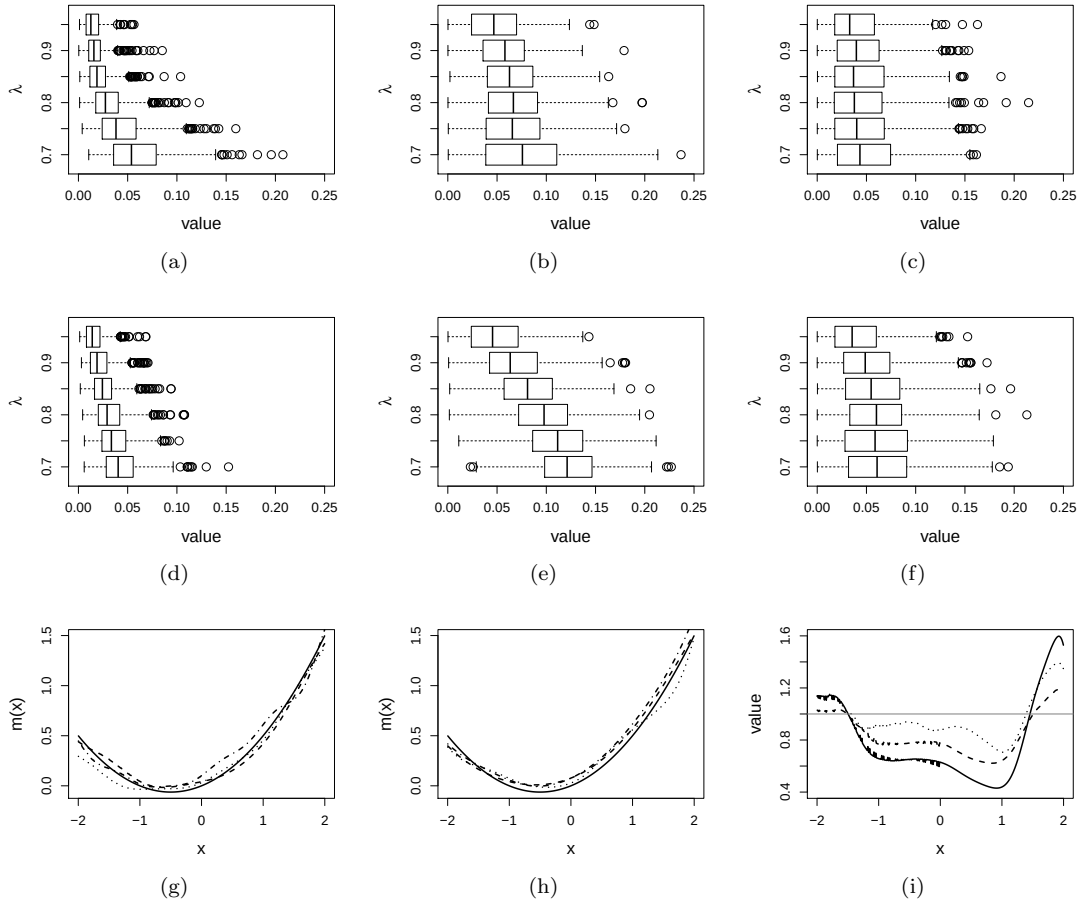

Figure 2. Simulation results under (C2) using the theoretical optimal h. Panels (a) & (d): boxplots of ISEs versus λ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively. Panels (b) & (e): boxplots of PAE(0) versus λ for $\hat{m}_{\mathrm{HZ}}(0)$ and $\hat{m}_{\mathrm{DFC}}(0)$, respectively. Panels (c) & (f): boxplots of PAE(-1) versus λ for $\hat{m}_{\mathrm{HZ}}(-1)$ and $\hat{m}_{\mathrm{DFC}}(-1)$, respectively. Panels (g) & (h): quantile curves when $\lambda = 0.85$ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively, based on ISEs (dashed lines for the first quartile, dotted lines for the second quartile, and dot-dashed lines for the third quartile, solid lines for the truth). Panel (i): PmAER (dashed line), PsdAER (dotted line), and PMSER (solid line) versus x when $\lambda = 0.85$; the horizontal reference line highlights the value 1.

6.4. *Motorcycle data*

We now apply the two estimation methods to error-contaminated data sets created based on the motorcycle crash data from a simulated motorcycle crash designed to test crash helmets (available under `R` library `MASS`). The original data set consists of 133 measurements of head acceleration measured in standard gravity acceleration (gs) at various times in milliseconds after impact. It is of interest to estimate the underlying head acceleration, Y, as a function of time after impact, X. Having the error-free data in this example allows us to have a reference estimate of the regression function with which the estimates based on error-prone data compare.

Following the same treatment of this data set in Silverman (1985), we assume independent model errors of the model of Y given X. Based on the original data, we first obtain the local linear estimate of $m(x)$ using the `R` function `locpol` in the `locpol` package, with the bandwidth chosen by cross validation (Wang and Jones 1995) implemented by function `regCVBwSelC` in the same `R` package. Compared to the fitted curves using error-prone data, this $\hat{m}(x)$ can be viewed as the "ideal" estimate in the sense that, intuitively,

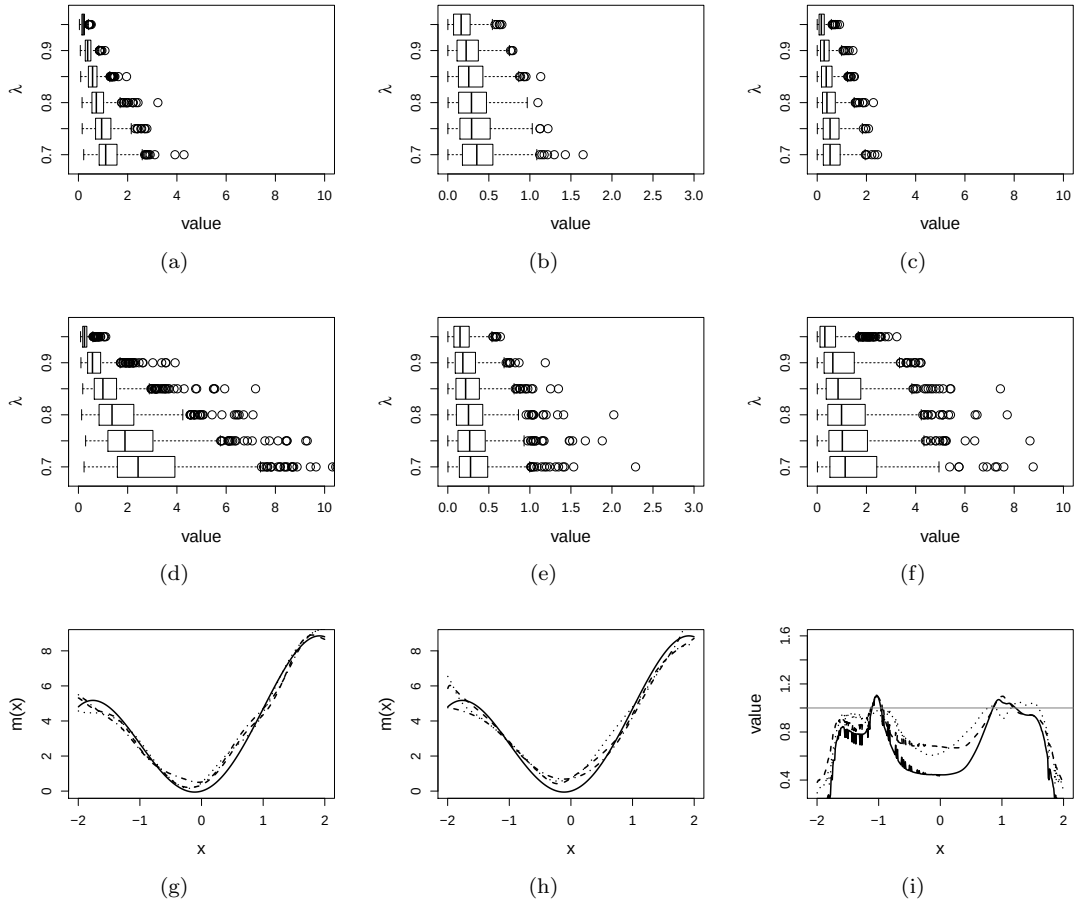

Figure 3. Simulation results under (C3) using the theoretical optimal h. Panels (a) & (d): boxplots of ISEs versus λ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively. Panels (b) & (e): boxplots of PAE(1) versus λ for $\hat{m}_{\mathrm{HZ}}(1)$ and $\hat{m}_{\mathrm{DFC}}(1)$, respectively. Panels (c) & (f): boxplots of PAE(2) versus λ for $\hat{m}_{\mathrm{HZ}}(2)$ and $\hat{m}_{\mathrm{DFC}}(2)$, respectively. Panels (g) & (h): quantile curves when $\lambda = 0.8$ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively, based on ISEs (dashed lines for the first quartile, dotted lines for the second quartile, dot-dashed lines for the third quartile, and solid lines for the truth). Panel (i): PmAER (dashed line), PsdAER (dotted line), and PMSER (solid line) versus x when $\lambda = 0.8$; the horizontal reference line highlights the value 1.

one cannot do better than this with error-contaminated data. We use this ideal curve as the reference curve in our follow-up experiments, where we contaminate X with simulated independent Laplace measurement errors to achieve different levels of estimated reliability ratio λ. At each level of λ, we use the error-contaminated data to estimate the acceleration curve using the two estimation methods, both assuming Laplace U. This experiment of curve estimation following data contamination is repeated 500 times at each level of λ. We obtained very similar results when we contaminated X with simulated normal U while assuming Laplace U when estimating the curves using both methods.

Figure 7 depicts results from this experiment, including boxplox of ISE at each λ level across 500 estimates associated with each estimator viewing the ideal curve as the "truth" of $m(x)$, the fitted curves when $\lambda = 0.95$ selected according to quantiles of ISE when the approximated theoretical optimal h is used for each method, and the counterpart fitted curves when the refined CV-SIMEX method is used to select h with $B = 10$ and $L = 10$. Using the ideal estimate as the benchmark, our estimate again appears to be less biased and less variable at all considered levels of error contamination than the DFC estimate.

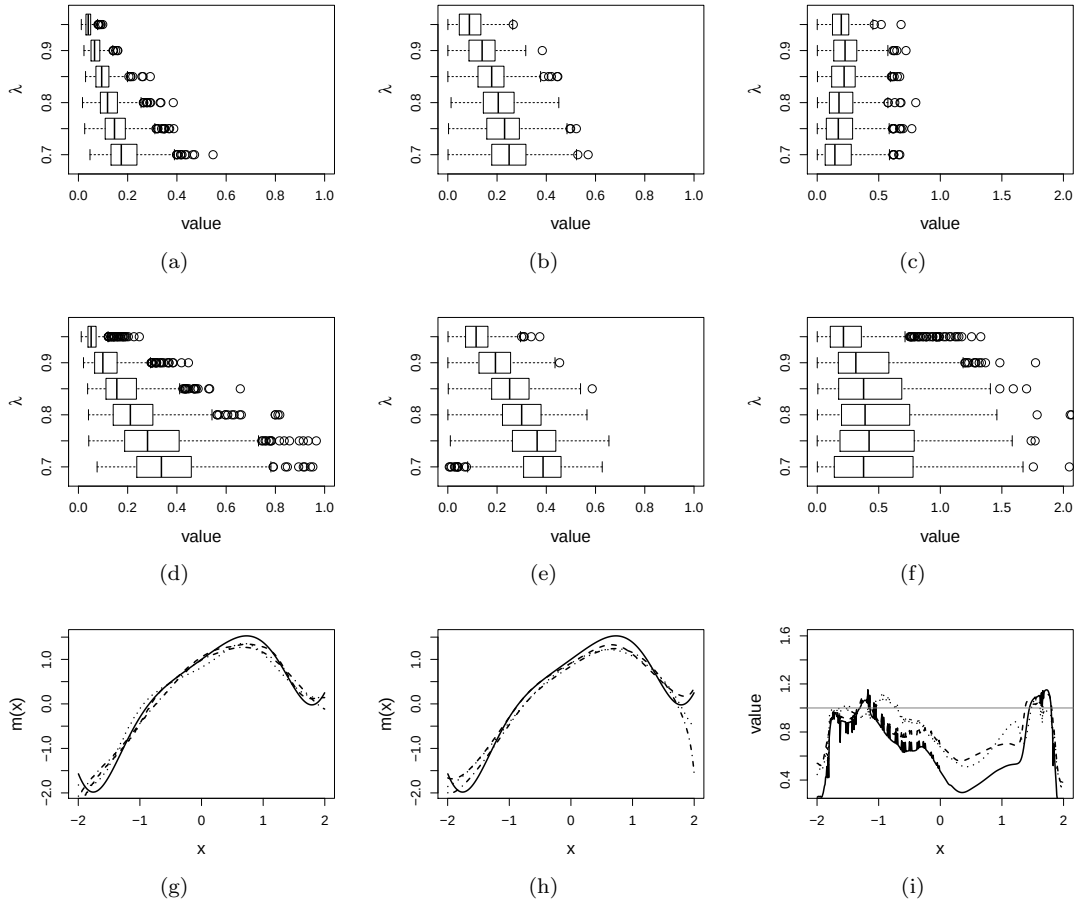

Figure 4. Simulation results under (C4) using the theoretical optimal h. Panels (a) & (d): boxplots of ISEs versus λ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively. Panels (b) & (e): boxplots of PAE(1) versus λ for $\hat{m}_{\mathrm{HZ}}(1)$ and $\hat{m}_{\mathrm{DFC}}(1)$, respectively. Panels (c) & (f): boxplots of PAE(2) versus λ for $\hat{m}_{\mathrm{HZ}}(2)$ and $\hat{m}_{\mathrm{DFC}}(2)$, respectively. Panels (g) & (h): quantile curves when $\lambda = 0.8$ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively, based on ISEs (dashed lines for the first quartile, dotted lines for the second quartile, dot-dashed lines for the third quartile, and solid lines for the truth). Panel (i): PmAER(dashed line), PsdAER (dotted line), and PMSER (solid line) versus x when $\lambda = 0.8$; the horizontal reference line highlights the value 1.

And when the refined CV-SIMEX method is used to select h, our estimator suffers less numerical instability compared to the competing method.

7. Discussion

In this study, we proposed a local polynomial estimator of the regression function when the covariate is measured with error. The proposed estimator makes direct use of the naive inference for the regression function and the covariate density as a whole, which leads to relatively more transparent connections between the properties of the proposed estimator and those of the inference from error-free data. Using these connections, we rigorously derived the asymptotic properties of the proposed estimator in comparison with the estimator proposed by Delaigle et al. (2009). Under very similar regularity conditions, besides the asymptotic normality that both estimators possess, we compare the asymptotic bias and variance of these estimators. These large sample comparisons suggest that the new estimator can be less biased than the competing estimator. Results

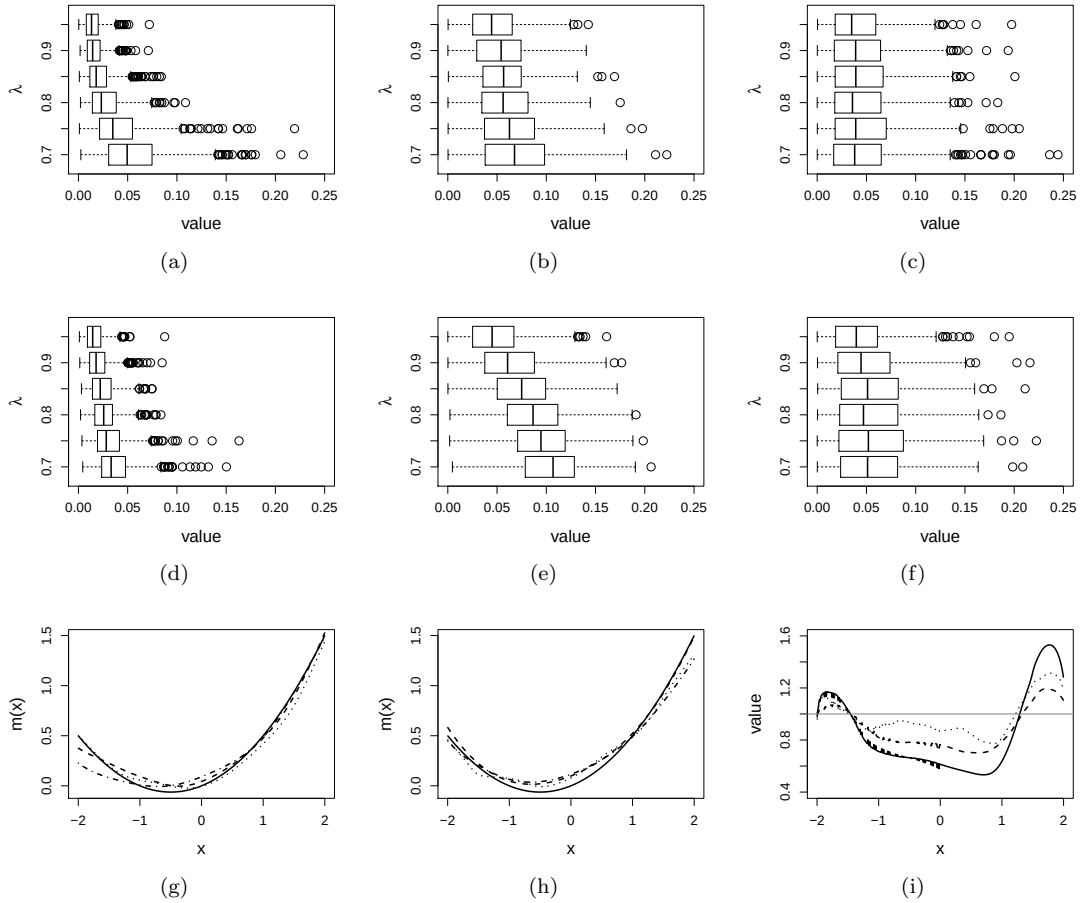

Figure 5. Simulation results under (C2) using the theoretical optimal h, with U-distribution misspecified as Laplace. Panels (a) & (d): boxplots of ISEs versus λ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively. Panels (b) & (e): boxplots of PAE(0) versus λ for $\hat{m}_{\mathrm{HZ}}(0)$ and $\hat{m}_{\mathrm{DFC}}(0)$, respectively. Panels (c) & (f): boxplots of PAE(-1) versus λ for $\hat{m}_{\mathrm{HZ}}(-1)$ and $\hat{m}_{\mathrm{DFC}}(-1)$, respectively. Panels (g) & (h): quantile curves when $\lambda = 0.85$ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively, based on ISEs (dashed lines for the first quartile, dotted lines for the second quartile, and dot-dashed lines for the third quartile, solid lines for the truth). Panel (i): PmAER (dashed line), PsdAER (dotted line), and PMSER (solid line) versus x when $\lambda = 0.85$; the horizontal reference line highlights the value 1.

from extensive simulation study also support this finding.

To implement the proposed method, we thoughtfully refined the CV-SIMEX bandwidth selection method proposed by Delaigle and Hall (2008) to narrow the search region of h, which in turn allows us to use a much smaller B in the SIMEX implementation without compromising noticeably the estimate. This refinement greatly reduces the computational burden for the otherwise intrinsically cumbersome bandwidth selection procedure due to the marriage of cross validation and SIMEX.

The comparison between the proposed estimator and the DFC estimator at the boundary of the support of X in our simulation study appears to depend on the distribution of X. Even though the proposed estimator appears to suffer less numerical instability when the refined CV-SIMEX method is used to select h compared to the competing method, it can still be rather challenging to estimate the curve near the boundary. The properties of our estimator near the boundary deserve further investigation, which can lead to ways to improve its behavior near the boundary.

Another practically interesting follow-up research topic is to consider a similar strat-

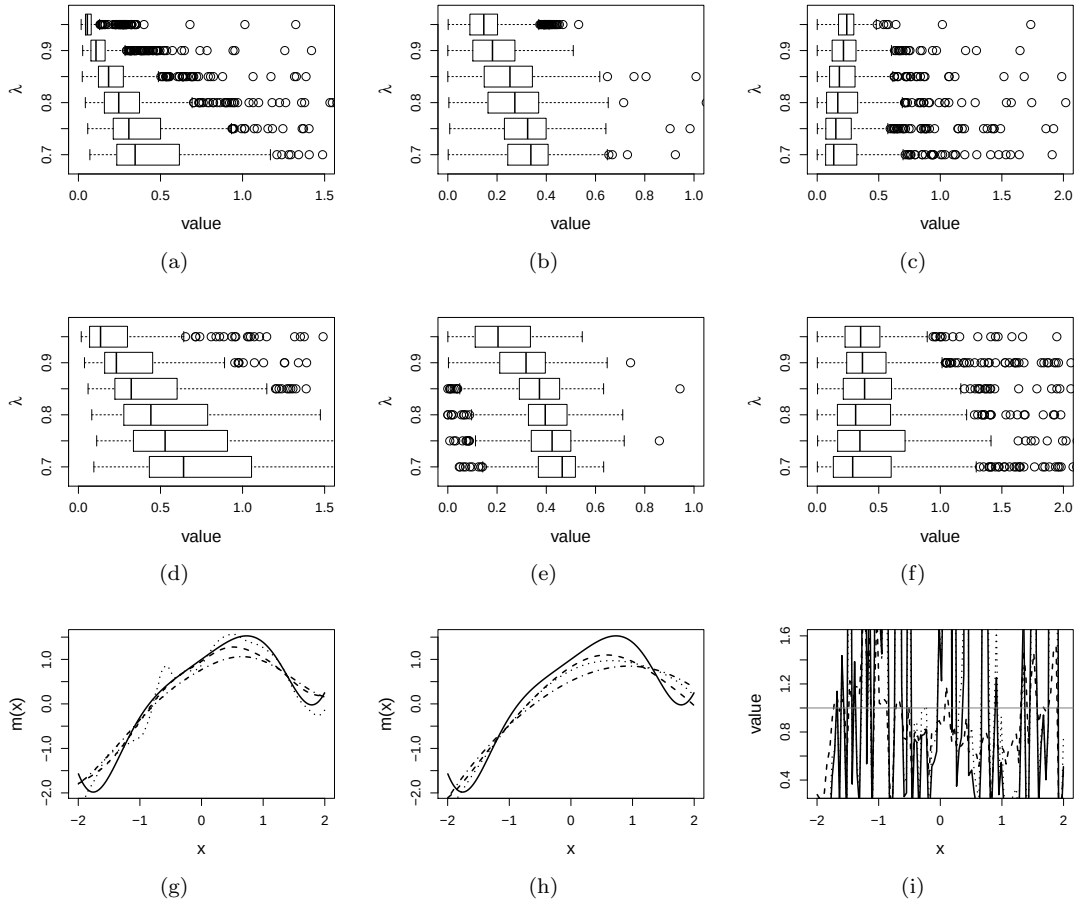

Figure 6. Simulation results under (C4) using CV-SIMEX bandwidth selection. Panels (a) & (d): boxplots of ISEs versus λ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively. Panels (b) & (e): boxplots of PAE(1) versus λ for $\hat{m}_{\mathrm{HZ}}(1)$ and $\hat{m}_{\mathrm{DFC}}(1)$, respectively. Panels (c) & (f): boxplots of PAE(2) versus λ for $\hat{m}_{\mathrm{HZ}}(2)$ and $\hat{m}_{\mathrm{DFC}}(2)$, respectively. Panels (g) & (h): quantile curves when $\lambda = 0.8$ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively, based on ISEs (dashed lines for the first quartile, dotted lines for the second quartile, and dot-dashed lines for the third quartile, solid lines for the truth). Panel (i): PmAER (dashed line), PsdAER (dotted line), and PMSER (solid line) versus x when $\lambda = 0.8$; the horizontal reference line highlights the value 1.

egy of estimating the regression function when the measurement error distribution is unknown and replicate measures of the true covariate are available.

Supplemental materials

The supplement to this article contains Appendices A–C referenced in Sections 3, 4, and 5.

References

Bailey, D. and Swarztrauber, P. (1994), 'A fast method for the numerical evaluation of continuous Fourier and Laplace transforms', *SIAM Journal on Scientific Computing*, 15, 1105–1110.

Carroll, R. and Hall, P. (1988), 'Optimal rates of convergence for deconvoluting a density', *Journal of the American Statistical Association*, 83, 1184–1186.

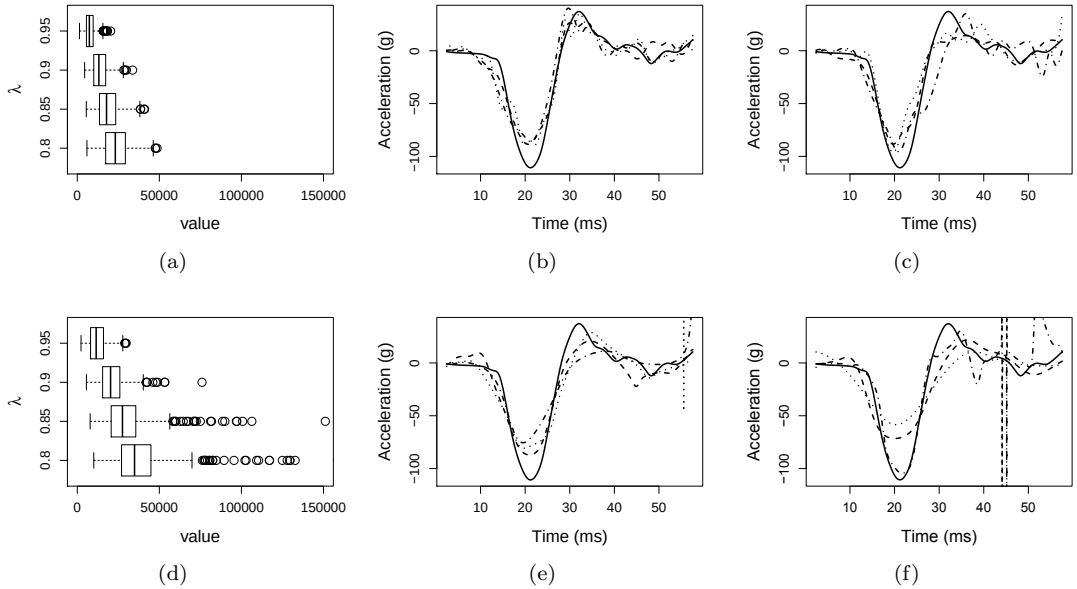

Figure 7. Results for motorcycle data. Panels (a) & (d): boxplots of ISEs versus λ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively. Panels (b) & (e): quantile curves when $\lambda = 0.95$ for $\hat{m}_{\mathrm{HZ}}(x)$ and $\hat{m}_{\mathrm{DFC}}(x)$, respectively, based on ISEs (dashed lines for the first quartile, dotted lines for the second quartile, and dot-dashed lines for the third quartile, solid lines for the "truth") when the approximated theoretical optimal h is used. Panels (c) & (f): counterpart quantile curves of those in panels (b) & (e) when h is chosen by the refined CV-SIMEX procedure.

Carroll, R., Ruppert, D., Stefanski, L. A., Crainiceanu, C. M. (2006), *Measurement error in nonlinear models: A model perspective* (2nd ed.), Chapman & Hall/CRC. Boca Raton, FL.

Davis, P. J. and Rabinowitz, P. (1984), *Methods of Numerical Integration*. Academic Press.

Delaigle, A. (2014), 'Nonparametric kernel methods with errors-in-variables: constructing estimators, computing them, and avoiding common mistakes', *Australian & New Zealand Journal of Statistics*, 56, 105–124.

Delaigle, A., Fan, J., and Carroll, R. (2009), 'A design-adaptive local polynomial estimator for the error-in-variables problem', *Journal of the American Statistical Association*, 104, 348–359.

Delaigle, A. and Hall, P. (2008), 'Using SIMEX for smoothing-parameter choice in errors-in-variables problems', *Journal of the American Statistical Association*, 103, 280-287.

Fan, J. (1991), 'Asymptotic normality for deconvolution kernel density estimators', *Sankhya, Series A*, 53, 97–110.

Fan, J. (1991), 'Global behavior of deconvolution kernel estimates', *Statistica Sinica*, 1, 541–551.

Fan, J. (1991), 'On the optimal rates of convergence for nonparametric deconvolution problems', *The Annals of Statistics*, 19, 1257–1272.

Fan, J. and Gijbels, I. (1996), *Local polynomial modelling and its applications*, Chapman and Hall/CRC, Boca Raton.

Fan, J. and Truong, Y. K. (1993), 'Nonparametric regression with errors in variables', *The Annals of Statistics*, 21, 1900–1925.

Nadaraya, E. (1964), 'On estimating regression', *Theory of probability and its applications*, 9, 141–142.

Ruppert, D. and Wand, M. P. (1994), 'Multivariate locally weighted least squares regression', *The Annals of Statistics*, 22, 1346–1370.

Schennach, S. M. (2004a), 'Estimation of nonlinear models with measurement error', *Econometrica*, 72, 33–75.

Schennach, S. M. (2004b), 'Nonparametric regression in the presence of measurement error', *Econometric Theory*, 20, 1046–1093.

Silverman, B.W. (1985), 'Some aspects of the spline smoothing approach to non-parametric re-

gression curve fitting', *Journal of Royal Statistical Society, Series B*, 47, 1–52.
Stefanski, L. A. and Boos, D. D. (2002), 'The calculus of M-estimation', *The American Statisticians*, 56, 29–38.
Stefanski, L. A. and Carroll, R. J. (1990), 'Deconvoluting kernel density estimators', *Statistics*, 21, 169–184.
Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*, Chapman and Hall, London.
Watson, G. (1964), 'Smooth regression analysis', *Sankhya, Series A*, 26, 359–372.
Zwanzig, S. (2006), 'On local linear estimation in nonparametric error-in-variables models', *Theory of Stochastic Processes*, 28, 316–327.