Electronic Journal of Statistics
Vol. 14 (2020) 970–1023
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1688

Conditional density estimation with covariate measurement error

Xianzheng Huang

Department of Statistics, University of South Carolina, Columbia, SC 29208, USA
e-mail: huang@stat.sc.edu

and

Haiming Zhou

Division of Statistics, Northern Illinois University, DeKalb, IL 60115, USA
e-mail: zhouh@niu.edu

Abstract: We consider estimating the density of a response conditioning on an error-prone covariate. Motivated by two existing kernel density estimators in the absence of covariate measurement error, we propose a method to correct the existing estimators for measurement error. Asymptotic properties of the resultant estimators under different types of measurement error distributions are derived. Moreover, we adjust bandwidths readily available from existing bandwidth selection methods developed for error-free data to obtain bandwidths for the new estimators. Extensive simulation studies are carried out to compare the proposed estimators with naive estimators that ignore measurement error, which also provide empirical evidence for the effectiveness of the proposed bandwidth selection methods. A real-life data example is used to illustrate implementation of these methods under practical scenarios. An R package, `lpme`, is developed for implementing all considered methods, which we demonstrate via an R code example in Appendix B.2.

MSC 2010 subject classifications: Primary 62G08; secondary 62G20.
Keywords and phrases: Bandwidth, bias, cross validation, deconvoluting kernel.

Received June 2019.

1. Introduction

The conditional density of a continuous response Y given a covariate X, denoted by $p(y|x)$, provides a complete picture of the association between Y and X that is valuable for data visualization and exploration. Rosenblatt (1969) is one of the pioneers who considered kernel density estimators for $p(y|x)$. Hyndman et al. (1996) further studied properties of the kernel density estimator based on a random sample, $\{(X_j, Y_j)\}_{j=1}^n$, given by

$$\hat{p}_1(y|x) = \frac{\frac{1}{nh_1h_2}\sum_{j=1}^{n} K_1\left(\frac{X_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right)}{\frac{1}{nh_1}\sum_{j=1}^{n} K_1\left(\frac{X_j - x}{h_1}\right)}, \tag{1.1}$$

where $K_1(t)$ and $K_2(t)$ are kernels, h_1 and h_2 are bandwidths. This estimator originates from two other well studied density estimators. The denominator of (1.1) is the kernel density estimator for the probability density function (pdf) of X, denoted by $f_X(x)$; and the numerator is the kernel density estimator for the joint pdf of (X, Y), denoted by $p(x, y)$. Fan et al. (1996) followed the idea of local polynomial estimation of a mean function (Fan and Gijbels, 1996, Chapter 3) to construct a class of local polynomial estimators for $p(y|x)$. The estimator $\hat{p}_1(y|x)$ in (1.1) belongs to this class, referred to as the local constant estimator. Hyndman and Yao (2002) revised the local polynomial estimators to guarantee non-negativity. Besides $\hat{p}_1(y|x)$, Hyndman et al. (1996) proposed another estimator for $p(y|x)$ with a different estimator for $p(x, y)$ in the numerator, leading to

$$\hat{p}_2(y|x) = \frac{\frac{1}{nh_1h_2}\sum_{j=1}^{n} K_1\left(\frac{X_j - x}{h_1}\right) K_2\left\{\frac{Y_j - \hat{m}(X_j) - y + \hat{m}(x)}{h_2}\right\}}{\frac{1}{nh_1}\sum_{j=1}^{n} K_1\left(\frac{X_j - x}{h_1}\right)}, \tag{1.2}$$

where $\hat{m}(x)$ is an estimator for $m(x) = E(Y|X = x)$, such as a local polynomial estimator. If one replaces $\hat{m}(\cdot)$ with $m(\cdot)$ in (1.2), one obtains the regular kernel density estimator for the density of $e = Y - m(X)$ given X, denoted by $f_{e|X}(e|x)$, which relates to $p(y|x)$ via $p(y|x) = f_{e|X}\{y - m(x)|x\}$. This relationship motivates the construction of $\hat{p}_2(y|x)$ in (1.2). Hyndman et al. (1996) showed that $\hat{p}_2(y|x)$ has a smaller asymptotic mean integrated squared error (MISE) when compared with $\hat{p}_1(y|x)$ under some situations commonly encountered in practice. Hansen (2004) studied $\hat{p}_2(y|x)$ more closely, who referred to $\hat{p}_2(y|x)$ as a two-step estimator to stress the estimation of $m(x)$ that is not needed for $\hat{p}_1(y|x)$, a one-step estimator in contrast.

It is common in practice that a covariate of interest cannot be measured directly or precisely. This motivates our work presented in this article, where we aim to estimate $p(y|x)$ when X is prone to measurement error. Due to error contamination, the observed data are $\{(W_j, Y_j)\}_{j=1}^n$ as opposed to $\{(X_j, Y_j)\}_{j=1}^n$, where W_j is an unbiased surrogate of X_j, for $j = 1, \ldots, n$. We assume in this study a classical additive measurement error model (Carroll et al., 2006, Section 1.2) that relates the observed covariate W and the true covariate X via

$$W_j = X_j + U_j, \tag{1.3}$$

where U_j represents measurement error with mean zero and variance σ_u^2, following a distribution specified by the pdf $f_U(u)$, and is independent of (X_j, Y_j), for $j = 1, \ldots, n$. For reasons related to identifiability issues, we assume $f_U(u)$ known in the majority of the study, and discuss treatments for unknown error distribution in Section 6. Robins et al. (1995) considered estimating unknown parameters in $p(y|x)$ that belongs to a pre-specified parametric family when covariates are missing or measured with error. We are not aware of existing works on estimating $p(y|x)$ nonparametrically in the presence of measurement

error. This article presents solutions to this fundamentally important problem, supplemented with an R package `lpme` (Zhou and Huang, 2017) for easy implementation of the proposed methods.

Using the error contaminated data in $\hat{p}_1(y|x)$ and $\hat{p}_2(y|x)$ leads to two naive estimators for $p(y|x)$ that ignore covariate measurement error,

$$\tilde{p}_1(y|x) = \frac{\frac{1}{nh_1h_2}\sum_{j=1}^{n} K_1\left(\frac{W_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right)}{\frac{1}{nh_1}\sum_{j=1}^{n} K_1\left(\frac{W_j - x}{h_1}\right)}, \tag{1.4}$$

$$\tilde{p}_2(y|x) = \frac{\frac{1}{nh_1h_2}\sum_{j=1}^{n} K_1\left(\frac{W_j - x}{h_1}\right) K_2\left\{\frac{Y_j - \hat{m}^*(W_j) - y + \hat{m}^*(x)}{h_2}\right\}}{\frac{1}{nh_1}\sum_{j=1}^{n} K_1\left(\frac{W_j - x}{h_1}\right)}, \tag{1.5}$$

where $\hat{m}^*(x)$ is an estimator for $m^*(x) = E(Y|W = x)$. These naive estimators are sensible estimators for the conditional density of Y given $W = x$, denoted by $p^*(y|x)$, but are usually inadequate estimators for $p(y|x)$.

In Section 2, we correct the above naive estimators for measurement error, producing two non-naive estimators for $p(y|x)$. Asymptotic properties of the proposed estimators are presented in Section 3. In Section 4 we develop methods for selecting bandwidths involved in these estimators. Finite sample performance of these estimators are demonstrated in comparison with the two naive estimators in simulation studies in Section 5. Practical considerations for implementing the proposed methods are discussed in Section 6, where we entertain a real-life data example. Lastly, in Section 7, we summarize the contribution of our work and discuss future research directions.

2. Proposed estimators

2.1. The rationale

Denote by $p^*(x, y)$ the joint density of (W, Y) evaluated at (x, y). Given the measurement error model in (1.3), one can show that $p^*(x, y)$ is equal to the convolution of $f_U(u)$ and $p(x, y)$ with respect to the first argument, that is,

$$p^*(x, y) = \int p(v, y) f_U(x - v) dv = \{p(\cdot, y) * f_U\}(x). \tag{2.1}$$

where "$*$" in the last expression is the convolution operator. The range of integration in all integrals in this article is the entire real line, unless specified otherwise. Denote by $\phi_g(t)$ the Fourier transform of a function $g(\cdot)$ or the characteristic function of a random variable g. Applying Fourier transform on

both sides of (2.1) yields $\phi_{p^*(\cdot,y)}(t) = \phi_{p(\cdot,y)}(t)\phi_U(t)$, which is equivalent to $\phi_{p(\cdot,y)}(t) = \phi_{p^*(\cdot,y)}(t)/\phi_U(t)$, assuming $\phi_U(t) \neq 0$ for all t. Applying inverse Fourier transform on both sides of the preceding identity gives

$$p(y|x)f_X(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\phi_{p^*(\cdot,y)}(t)}{\phi_U(t)}dt, \tag{2.2}$$

where i is the imaginary unit.

Putting a "hat" on top of each unknown quantity in (2.2) to represent an estimator for this quantity, we obtain a general form of estimators for $p(y|x)$ that account for covariate measurement error,

$$\hat{p}(y|x) = \hat{f}_X^{-1}(x)\cdot\frac{1}{2\pi}\int e^{-itx}\frac{\phi_{\hat{p}^*(\cdot,y)}(t)}{\phi_U(t)}dt. \tag{2.3}$$

With $\hat{p}(x,y) = \hat{p}(y|x)\hat{f}_X(x)$ being an estimator for $p(x,y)$, (2.3) relates $\hat{p}(x,y)$ to $\hat{p}^*(x,y)$, which is a naive estimator for $p(x,y)$ that is suitable for estimating $p^*(x,y)$. The numerators in $\tilde{p}_1(y|x)$ and $\tilde{p}_2(y|x)$ are examples of $\hat{p}^*(x,y)$. Even though, by construction, the integral in (2.2) is real as long as all integrals leading to (2.2) are well defined, the integral in (2.3) can be complex with $p^*(\cdot,y)$ now replaced by $\hat{p}^*(\cdot,y)$. A sensible treatment when the right-hand side of (2.3) returns a complex quantity is to use the real part as an estimator of $p(y|x)$, and argue that the imaginary part is merely a consistent estimator of zero by showing that (2.3) is a consistent estimator of the real-valued $p(y|x)$.

As for the estimator for $f_X(x)$ in (2.3), we adopt the deconvoluting density estimator (Carroll and Hall, 1988; Stefanski and Carroll, 1990),

$$\hat{f}_X(x) = \frac{1}{nh_1}\sum_{j=1}^{n}K_1^*\left(\frac{W_j - x}{h_1}\right), \tag{2.4}$$

where

$$K_1^*(t) = \frac{1}{2\pi}\int e^{-its}\frac{\phi_{K_1}(s)}{\phi_U(-s/h_1)}ds \tag{2.5}$$

is referred to as the deconvoluting kernel. Under conditions (K1) given in Section 3.1, Stefanski and Carroll (1990) showed that

$$E\left\{K_1^*\left(\frac{W-x}{h_1}\right)\middle|\, X\right\} = K_1\left(\frac{X-x}{h_1}\right), \tag{2.6}$$

suggesting that $\hat{f}_X(x)$ has the same bias as the ordinary kernel density estimator for $f_X(x)$ that appears as the common denominator of $\hat{p}_1(y|x)$ and $\hat{p}_2(y|x)$ in (1.1) and (1.2).

2.2. Two estimators accounting for measurement error

Using the numerator of $\tilde{p}_1(y|x)$ as $\hat{p}^*(x,y)$ in (2.3), one can show via straightforward algebra that (2.3) reduces to

$$\hat{p}_3(y|x) = \frac{\frac{1}{nh_1h_2}\sum_{j=1}^{n} K_1^*\left(\frac{W_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right)}{\frac{1}{nh_1}\sum_{j=1}^{n} K_1^*\left(\frac{W_j - x}{h_1}\right)}. \tag{2.7}$$

Looking back at its naive counterpart, $\tilde{p}_1(y|x)$, makes the construction of $\hat{p}_3(y|x)$ in (2.7) transparent. Since the naive estimator $\tilde{p}_1(y|x)$ depends on W_j only via $K_1\{(W_j - x)/h_1\}$, (2.6) suggests that replacing $K_1\{(W_j - x)/h_1\}$ with $K_1^*\{(W_j - x)/h_1\}$, for $j = 1, \ldots, n$, suffices to correct $\tilde{p}_1(y|x)$ for measurement error. This substitution yields $\hat{p}_3(y|x)$.

To correct $\tilde{p}_2(y|x)$ for measurement error is more involved because it depends on $\{W_j\}_{j=1}^n$ in a more complicated way than $\tilde{p}_1(y|x)$ does, and the trick of replacing the regular kernel with a deconvoluting kernel that leads to (2.7) (and also (2.4)) does not work here. Indeed, if one sets $\hat{p}^*(x,y)$ in (2.3) as the numerator of $\tilde{p}_2(y|x)$, an estimator for $p^*(x,y)$ denoted by $\tilde{p}_2(x,y)$, one obtains an estimator for $p(y|x)$ given by

$$\hat{p}_4(y|x) = \hat{f}_X^{-1}(x) \cdot \frac{1}{2\pi} \int e^{-itx} \frac{\phi_{\tilde{p}_2(\cdot,y)}(t)}{\phi_U(t)} dt, \tag{2.8}$$

which cannot be further simplified. For concreteness, in the majority of our study, we use the local linear estimator for $m^*(x)$ as $\hat{m}^*(x)$ in $\tilde{p}_2(x,y)$, with kernel $K_3(t)$ and bandwidth h_3. Considerations of other estimators for $m^*(x)$ are discussed in Sections 5 and 6.

In the absence of measurement error, Hyndman et al. (1996) showed that, under certain conditions (to be presented in Section 3.3), the two-step estimator $\hat{p}_2(y|x)$ often has a lower MISE than the one-step estimator $\hat{p}_1(y|x)$. In the presence of measurement error, we show next that, after correcting $\tilde{p}_2(y|x)$ and $\tilde{p}_1(y|x)$ for measurement error, the comparison between $\hat{p}_4(y|x)$ and $\hat{p}_3(y|x)$ becomes more involved, but $\hat{p}_4(y|x)$ still improves over $\hat{p}_3(y|x)$ under similar conditions.

3. Asymptotic properties

3.1. Preamble

We study properties of $\hat{p}_3(y|x)$ and $\hat{p}_4(y|x)$ under two types of measurement error distributions, namely ordinary smooth distributions and super smooth distributions (Fan, 1991a,b,c). Their definitions are given next.

Definition 3.1 *The distribution of U is ordinary smooth of order b if*

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

for some positive constants b and c.

Definition 3.2 *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 b, b_0, b_1, d_0, d_1, and d_2.

Laplace and gamma distributions are examples of ordinary smooth distributions. Normal and Cauchy distributions are super smooth, for instance. Technical conditions imposed on different functions for the study of asymptotics are listed below. The first set of conditions are solely regarding measurement error U.

Conditions U:

(U1) For all t, $\phi_U(t) \neq 0$.
(U2) $|\phi_U'(t)|_\infty < \infty$.

Condition (U1) is needed to reach (2.2) and for the validity of $K_1^*(t)$ in (2.5). Condition (U2) is imposed to guarantee finite variance for $\hat{f}_X(x)$ when U is ordinary smooth, which is a condition that can be relaxed when U is super smooth due to a stronger condition on $K_1(t)$ given in (K5) included in the following set of conditions.

Conditions K:

(K1) $|\phi_{K_1}(t)/\phi_U(-t/h_1)|_\infty < \infty$, $\int |\phi_{K_1}(t)/\phi_U(-t/h_1)|dt < \infty$.
(K2) $|\phi_{K_1}(t)|_\infty < \infty$ and $|\phi_{K_1}'(t)|_\infty < \infty$.
(K3) $\int |t|^b|\phi_{K_1}(t)|dt < \infty$, $\int |t|^{2b}|\phi_{K_1}(t)|^2dt < \infty$.
(K4) $\int(|t|^b + |t|^{b-1})(|\phi_{K_1}'(t)| + |\phi_{K_1}(t)|)dt < \infty$.
(K5) The support of $\phi_{K_1}(t)$ is $[-1, 1]$.

Condition (K1) is needed to establish (2.6). Conditions (K2)–(K4) are regularity conditions for the variance of $\hat{p}_3(y|x)$ to exist when U is ordinary smooth, whereas only (K5) is needed for this purpose when U is super smooth. Besides conditions on $K_1(t)$ stated above, we choose all three kernels, $K_1(t)$, $K_2(t)$, and $K_3(t)$, to be real and even functions with finite second moments. When $\hat{p}_4(y|x)$ is concerned, once Conditions K are imposed on $K_1(t)$, intuition suggests that having a bounded $K_2(t)$ should suffice to guarantee finite first two moments for $\hat{p}_4(y|x)$ although one should exercise care in formulating more concrete conditions relating to $K_2(t)$. We will come back to this point with more discussions in Section 3.3. For numerical stability and simplicity, we choose both $K_1(t)$ and $K_2(t)$ to be the same kernel in $\hat{p}_4(y|x)$ as done in Masry (1993) for instance. Lastly, it is assumed that $f_X(x)$ does not vanish over the support of X, and it is twice differentiable.

3.2. Properties of the one-step estimator $\hat{p}_3(y|x)$

We derive in Appendix 7 the asymptotic bias and variance of $\hat{p}_3(y|x)$, summarized in the following theorem.

Theorem 3.1 *When U is ordinary smooth of order b, if $nh_1^{1+2b}h_2(h_1^2+h_2^2)^2 \to \infty$ as $n \to \infty$, $h_1, h_2 \to 0$, then*

$$\begin{aligned}&\hat{p}_3(y|x) - p(y|x)\\ =&DB_3(x,y,h_1,h_2) + O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2) + O_p\left(\frac{1}{\sqrt{nh_1^{1+2b}h_2}}\right),\end{aligned} \tag{3.1}$$

where

$$DB_3(x,y,h_1,h_2) = \frac{1}{2f_X(x)}\left[\{p_{xx}(x,y) - p(y|x)f_X''(x)\}\,\mu_{2,1}h_1^2 + p_{yy}(x,y)\mu_{2,2}h_2^2\right] \tag{3.2}$$

is the dominating bias, in which $f_X''(x)$ is the second derivative of $f_X(x)$, $p_{xx}(x,y) = (\partial^2/\partial x^2)p(x,y)$, $p_{yy}(x,y) = (\partial^2/\partial y^2)p(x,y)$, and $\mu_{2,\ell} = \int t^2K_\ell(t)dt$, for $\ell = 1, 2$. When U is super smooth of order b, if $nh_1^{1-2b_2}h_2\exp(-2h_1^{-b}/d_2)(h_1^2+h_2^2)^2 \to \infty$ as $n \to \infty$, $h_1, h_2 \to 0$, then

$$\begin{aligned}&\hat{p}_3(y|x) - p(y|x)\\ =&DB_3(x,y,h_1,h_2) + O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2) + O_p\left\{\frac{\exp(h_1^{-b}/d_2)}{\sqrt{nh_1^{1-2b_2}h_2}}\right\},\end{aligned} \tag{3.3}$$

where $b_2 = b_0I(b_0 < 0.5)$.

In contrast to $\hat{p}_3(y|x)$, Hyndman et al. (1996, Section 3.1) obtained the following result for the error-free counterpart estimator $\hat{p}_1(y|x)$,

$$\begin{aligned}&\hat{p}_1(y|x) - p(y|x)\\ =&\frac{1}{2}\left[\left\{\frac{\partial^2 p(y|x)}{\partial x^2} + 2\frac{\partial p(y|x)}{\partial x}\frac{f_X'(x)}{f_X(x)}\right\}\mu_{2,1}h_1^2 + \frac{\partial^2 p(y|x)}{\partial y^2}\mu_{2,2}h_2^2\right]\\ &+ O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2) + O_p\left(\frac{1}{\sqrt{nh_1h_2}}\right),\end{aligned} \tag{3.4}$$

where $f_X'(x)$ is the first derivative of $f_X(x)$. Straightforward algebra reveal that the dominating bias in (3.4) is equal to the dominating bias of $\hat{p}_3(y|x)$ given in (3.2). Although exhibiting same asymptotic bias, the asymptotic variance of $\hat{p}_3(y|x)$ is inflated due to measurement error when compared to $\hat{p}_1(y|x)$, with more substantial inflation when U is super smooth than when it is ordinary smooth.

3.3. Properties of the two-step estimator $\hat{p}_4(y|x)$

For a generic bivariate function, $g(w,y)$, we define the following double integral transform of $g(w,y)$ via the operator $\mathscr{T}_x(\cdot)$, assuming $|\phi_{g(\cdot,y)}(t)/\phi_U(t)|_\infty < \infty$ and $\int |\phi_{g(\cdot,y)}(t)/\phi_U(t)|dt < \infty$ for each y,

$$\mathscr{T}_x\{g(\cdot,y)\} = \frac{1}{2\pi}\int e^{-itx}\frac{\phi_{g(\cdot,y)}(t)}{\phi_U(t)}dt. \tag{3.5}$$

In Appendix A.2, we establish the following results regarding $\hat{p}_4(y|x)$.

Theorem 3.2 *When U is ordinary smooth of order b, if $nh_1^{1+2b}h_2(h_1^2+h_2^2)^2 \to \infty$ and $h_3 = O(h_2)$ as $n\to\infty$, $h_1,h_2,h_3 \to 0$, then*

$$\begin{aligned}&\hat{p}_4(y|x) - p(y|x)\\ =&DB_4(x,y,h_1,h_2) + O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2) + O_p\left(\frac{1}{\sqrt{nh_1^{1+2b}h_2}}\right),\end{aligned} \tag{3.6}$$

where

$$\begin{aligned}&DB_4(x,y,h_1,h_2)\\ =&\frac{1}{2f_X(x)}\left(\left[p_{xx}(x,y)+\sum_{k=2}^{4}\mathscr{T}_x\{I_k(\cdot,y)\}-p(y|x)f_X''(x)\right]\mu_{2,1}h_1^2+p_{yy}(x,y)\mu_{2,2}h_2^2\right)\end{aligned} \tag{3.7}$$

is the dominating bias, in which

$$\begin{cases} I_2(w,y) = \left\{\dfrac{d^2}{dw^2}m^*(w)\right\}\displaystyle\int p_y(v,y)f_U(w-v)dv,\\ I_3(w,y) = \left\{\dfrac{d}{dw}m^*(w)\right\}^2\displaystyle\int p_{yy}(v,y)f_U(w-v)dv,\\ I_4(w,y) = 2\left\{\dfrac{d}{dw}m^*(w)\right\}\displaystyle\int p_{xy}(v,y)f_U(w-v)dv. \end{cases} \tag{3.8}$$

When U is super smooth of order b, if $nh_1^{1-2b_2}h_2\exp(-2h_1^{-b}/d_2)(h_1^2+h_2^2)^2 \to \infty$ and $h_3 = O(h_2)$ as $n\to\infty$, $h_1,h_2,h_3\to 0$, then

$$\begin{aligned}&\hat{p}_4(y|x) - p(y|x)\\ =&DB_4(x,y,h_1,h_2) + O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2) + O_p\left\{\frac{\exp(h_1^{-b}/d_2)}{\sqrt{nh_1^{1-2b_2}h_2}}\right\}.\end{aligned} \tag{3.9}$$

Similar to the one-step estimator $\hat{p}_3(y|x)$, correcting for measurement error results in a higher variance for the two-step estimator $\hat{p}_4((y|x)$ than its error-

free counterpart $\hat{p}_2(y|x)$. In addition, Theorem 3.2 indicates that, as long as $h_3 = O(h_2)$, the effects of estimating $m^*(x)$ on $\hat{p}_4(y|x)$ are negligible in regard to both bias and variance.

In what follows, we compare the dominating bias of $\hat{p}_4(y|x)$ with those of $\hat{p}_2(y|x)$ and $\hat{p}_3(y|x)$. In the absence of measurement error, Hansen (2004) established the following result regarding the two-step estimator $\hat{p}_2(y|x)$,

$$\begin{aligned}&\hat{p}_2(y|x) - p(y|x)\\ =&\frac{1}{2}\left[\left\{\frac{\partial^2 f_{e|x}(e|x)}{\partial x^2} + 2\frac{\partial f_{e|x}(e|x)}{\partial x}\frac{f_X'(x)}{f_X(x)}\right\}\mu_{2,1}h_1^2 + \frac{\partial^2 f_{e|x}(e|x)}{\partial e^2}\mu_{2,2}h_2^2\right]\\ &+ O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2) + O_p\left(\frac{1}{\sqrt{nh_1h_2}}\right).\end{aligned}$$

Elaborations of derivatives of $f_{e|x}(e|x)$ reveal that Hansen's result suggests the following dominating bias of $\hat{p}_2(y|x)$,

$$\begin{aligned}&\mathrm{DB}_2(x, y, h_1, h_2)\\ =&\frac{1}{2f_X(x)}\left[\left\{f_{X,e,11}^{(2)}(x, e) - p(y|x)f_X''(x)\right\}\mu_{2,1}h_1^2 + p_{yy}(x, y)\mu_{2,2}h_2^2\right],\end{aligned} \tag{3.10}$$

where $f_{X,e}(x, e)$ is the joint density of X and $e = Y - m(X)$, and $f_{X,e,11}^{(2)}(x, e) = (\partial^2/\partial x^2)f_{X,e}(x, e)$. An interesting finding here is that Hansen's dominating bias of the two-step estimator for $p(y|x)$ in the absence of measurement error is generally not equal to the dominating bias of our proposed two-step estimator accounting for measurement error given in (3.7). Starting from $p(x, y) = f_{X,e}\{x, y - m(x)\}$, one can derive $p_{xx}(x, y)$ and show that

$$\begin{aligned}&p_{xx}(x, y)\\ =&f_{X,e,11}^{(2)}(x, e) - m''(x)f_{X,e,2}^{(1)}(x, e) + \{m'(x)\}^2 f_{X,e,22}^{(2)}(x, e) - 2m'(x)f_{X,e,21}^{(2)}(x, e),\end{aligned} \tag{3.11}$$

where $m'(x)$ and $m''(x)$ are the first and second derivatives of $m(x)$, respectively, $f_{X,e,2}^{(1)}(x, e) = (\partial/\partial e)f_{X,e}(x, e)$, $f_{X,e,22}^{(2)}(x, e) = (\partial^2/\partial e^2)f_{X,e}(x, e)$, and $f_{X,e,21}^{(2)}(x, e) = (\partial^2/\partial x\partial e)f_{X,e}(x, e)$. Substituting $p_{xx}(x, y)$ in (3.7) with (3.11), one can see that

$$\begin{aligned}&\mathrm{DB}_4(x, y, h_1, h_2)\\ =&\mathrm{DB}_2(x, y, h_1, h_2) + \frac{\mu_{2,1}h_1^2}{2f_X(x)}\left[\sum_{k=2}^{4}\mathscr{T}_x\{I_k(\cdot, y)\} -\right.\\ &\left. m''(x)f_{X,e,2}^{(1)}(x, e) + \{m'(x)\}^2 f_{X,e,22}^{(2)}(x, e) - 2m'(x)f_{X,e,21}^{(2)}(x, e)\right].\end{aligned} \tag{3.12}$$

Even though there exists an interesting connection between the three functions defined in (3.8) and the last three terms in (3.12), (3.12) does not provide much

insight on how $\hat{p}_4(y|x)$ compares with $\hat{p}_2(y|x)$. We next consider three special cases under which (3.12) can be further simplified in order to gain more insight on the dominating bias associated with different estimators.

The first special case is when $m(x)$ is a constant function, in which case one can show that $m^*(w)$ is also a constant function. Now, by (3.8), all terms in (3.12) following $\mathrm{DB}_2(x, y, h_1, h_2)$ reduce to zero. If fact, by (3.2) and (3.11), $\mathrm{DB}_2(x, y, h_1, h_2) = \mathrm{DB}_3(x, y, h_1, h_2) = \mathrm{DB}_4(x, y, h_1, h_2)$ when $m(x)$ is free of x. The second special case is when there is no measurement error, under which we show in Section B.3 of Appendix A.2 that terms insides the square brackets in (3.12) also reduce to zero, suggesting $\mathrm{DB}_4(x, y, h_1, h_2) = \mathrm{DB}_2(x, y, h_1, h_2)$, as it should be in the absence of measurement error. The third special case results from imposing the conditions stated in Hyndman et al. (1996), under which they concluded that $\hat{p}_2(y|x)$ is superior than $\hat{p}_1(y|x)$. These conditions include that (H1) the covariate is locally uniform near x so that $f'_X(x) \approx 0$ and $f''_X(x) \approx 0$, (H2) $e \perp X$ so that $p(y|x) = f_e\{y - m(x)\}$, and (H3) $m(x)$ is locally linear near x so that $m''(x) \approx 0$. Under Conditions (H1)–(H3), we simplify (3.10), (3.2), and (3.7) in Section B.3 of Appendix A.2 and find that

$$\begin{aligned}
\mathrm{DB}_3(x, y, h_1, h_2) &\approx \mathrm{DB}_2(x, y, h_1, h_2) + 0.5 f''_e(e) \{m'(x)\}^2 \mu_{2,1} h_1^2, \\
\mathrm{DB}_4(x, y, h_1, h_2) &\approx \mathrm{DB}_3(x, y, h_1, h_2) \\
&\quad + 0.5 f''_e(e) \left[\left\{ \frac{d}{dx} m^*(x) \right\}^2 - 2m'(x) \frac{d}{dx} m^*(x) \right] \mu_{2,1} h_1^2.
\end{aligned}$$

It has been observed in many measurement error model settings that $(d/dx)m^*(x)$ attenuates towards zero compared to $m'(x)$, with the exact attenuation factor derived for the case when $m(x)$ is linear, and X and U are normally distributed (Fuller, 2009, Section 1.1). To be more specific, if $m(x) = \beta_0 + \beta_1 x$, where β_0 and β_1 are the intercept and slope parameters, then it has been shown in this case that $m^*(x) = \alpha_0 + \alpha_1 x$, where α_0 and α_1 are the intercept and slope parameters in the naive regression, in which $\alpha_1 = \lambda\beta_1$, with $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ known as the reliability ratio (Carroll et al., 2006, Section 3.2.1), and σ_x^2 being the variance of X. This gives $\mathrm{DB}_3(x, y, h_1, h_2) \approx 0.5 f''_e(e)(\beta_1^2 \mu_{2,1} h_1^2 + \mu_{2,2} h_2^2)$, in contrast to $\mathrm{DB}_4(x, y, h_1, h_2) \approx 0.5 f''_e(e)\{(1-\lambda)^2 \beta_1^2 \mu_{2,1} h_1^2 + \mu_{2,2} h_2^2\}$. In summary, under the third special case, one would usual expect the following trend of comparisons, $|\mathrm{DB}_2(x, y, h_1, h_2)| \leq |\mathrm{DB}_4(x, y, h_1, h_2)| \leq |\mathrm{DB}_3(x, y, h_1, h_2)|$. Therefore, under the same set of conditions considered in Hyndman et al. (1996), the proposed two-step estimator $\hat{p}_4(y|x)$ is still asymptotically superior than the one-step estimator $\hat{p}_3(y|x)$.

We are now in the position to reflect on the findings that the duo of $\hat{p}_2(y|x)$ and $\hat{p}_4(y|x)$ do not share the same dominating bias, whereas the other duo, $\hat{p}_1(y|x)$ and $\hat{p}_3(y|x)$, do. Looking back at the construction of the two proposed estimators accounting for measurement error in Section 2, one can see that they only differ in the estimator of $p^*(x, y)$ used in (2.3) to obtain an estimator for the joint density $p(x, y)$ via the integral transform defined in (3.5). Denote by $\hat{p}_1(x, y)$ and $\hat{p}_2(x, y)$ the numerators of (1.1) and (1.2), respectively, which

are two estimators for $p(x,y)$ in the absence of measurement error. Denote by $\tilde{p}_1(x,y)$ and $\tilde{p}_2(x,y)$ the numerators of (1.4) and (1.5), respectively, which are two estimators for $p^*(x,y)$, viewed as naive estimators for $p(x,y)$ in the presence of measurement error. In the one-step estimator $\hat{p}_3(y|x)$, the estimator for $p(x,y)$ can be expressed as

$$\begin{aligned}\mathscr{T}_x\{\tilde{p}_1(\cdot,y)\} =& \frac{1}{nh_1h_2}\sum_{j=1}^{n}\mathscr{T}_x\left\{K_1\left(\frac{W_j-\cdot}{h_1}\right)\right\}K_2\left(\frac{Y_j-y}{h_2}\right) \\ =& \frac{1}{nh_1h_2}\sum_{j=1}^{n}K_1^*\left(\frac{W_j-x}{h_1}\right)K_2\left(\frac{Y_j-y}{h_2}\right), \text{ by (2.5)}\end{aligned} \tag{3.13}$$

which has the same expectation as that of $\hat{p}_1(x,y)$ according to (2.6). This explains why $\hat{p}_3(y|x)$ and $\hat{p}_1(y|x)$ have the same dominating bias. In contrast, in the two-step estimator $\hat{p}_4(y|x)$, the estimator for $p(x,y)$ is

$$\mathscr{T}_x\{\tilde{p}_2(\cdot,y)\} = \frac{1}{nh_1h_2}\sum_{j=1}^{n}\mathscr{T}_x\left[K_1\left(\frac{W_j-\cdot}{h_1}\right)K_2\left\{\frac{Y_j-\hat{m}^*(W_j)-y+\hat{m}^*(\cdot)}{h_2}\right\}\right], \tag{3.14}$$

of which the expectation is typically not equal to $E\{\hat{p}_2(x,y)\}$. Hence, it is not surprising that, after correcting the naive two-step estimator $\tilde{p}_2(y|x)$ for measurement error, $\hat{p}_4(y|x)$ does not have the same dominating bias as that of $\hat{p}_2(y|x)$.

Contrasting (3.14) with (3.13) also brings awareness that more involved conditions are needed for $\mathscr{T}_x\{\tilde{p}_2(\cdot,y)\}$ to be well-defined. According to (3.13), $\mathscr{T}_x\{\tilde{p}_1(\cdot,y)\}$ is well-defined because $h_1^{-1}\mathscr{T}_x[K_1\{(W-\cdot)/h_1\}]$ is, thanks to Condition (K1). By (3.14), $\mathscr{T}_x\{\tilde{p}_2(\cdot,y)\}$ is well-defined if $(h_1h_2)^{-1}\mathscr{T}_x(K_1\{(W-\cdot)/h_1\}K_2[\{Y-\hat{m}^*(W)-y+\hat{m}^*(\cdot)\}/h_2])$ is, for which sufficient conditions formulated in the same spirit as those in Condition (K1) are that $|\mathrm{CR}(t,Y,W,y)|_\infty < \infty$ and $\int|\mathrm{CR}(t,Y,W,y)|dt<\infty$ with probability one for each y, where

$$\begin{aligned}&\mathrm{CR}(t,Y,W,y)\\ =&\frac{(h_1h_2)^{-1}\int e^{itw}K_1\left(\frac{W-w}{h_1}\right)K_2\left\{\frac{Y-\hat{m}^*(W)-y+\hat{m}^*(w)}{h_2}\right\}dw}{\phi_U(t)}.\end{aligned} \tag{3.15}$$

These sufficient conditions formulated in terms of $\mathrm{CR}(t,Y,W,y)$ essentially imply that the Fourier transform of the product kernel, $K_1(t)K_2\{s(t)\}$, tails off to zero much faster than $\phi_U(t)$ does as $|t|\to\infty$ so that the norm of the complicated ratio in (3.15) is integrable, where $s(\cdot)$ denotes some function of t, introduced here to signify that arguments in $K_1(\cdot)$ and $K_2(\cdot)$ in (3.15) both involve w. Because imposing Condition (K1) already guarantees that the Fourier transform of $K_1(t)$ diminishes fast enough, compared with how fast $\phi_U(t)$ diminishes as $|t|$ diverges, we conjecture that the aforementioned conditions in terms of $\mathrm{CR}(t,Y,W,y)$ are satisfied when $K_2(\cdot)$ is of the same order as $K_1(\cdot)$ so that the

Fourier transform of the product kernel appearing in (3.15) tends to zero no slower than $\phi_{K_1}(t)$ does as $|t| \to \infty$. Indeed, when implementing the proposed two-stage estimation method, we set K_2 the same as K_1 and encounter little numerical complication in obtaining $\hat{p}_4(y|x)$ in the simulation study.

More general analytic comparisons between $\hat{p}_3(y|x)$ and $\hat{p}_4(y|x)$ outside of the aforementioned special cases are unattainable. Empirical evidence from simulation study can shed more light on how they compare with each other and also with the naive estimators. In order to implement the proposed methods, strategies for choosing bandwidths are needed. This is the topic of the next section.

4. Bandwidths selection

4.1. Relevant strategies

The choice of bandwidths in kernel density estimators has a great impact on the estimators. There are two main streams in the literature on bandwidth selection, one relating to the so-called plug-in methods, the other in line with cross validation (CV). Both veins of methodology development start from a criterion that assesses the quality of an estimator, such as the integrated squared error (ISE) of a density estimator, or the MISE. Oftentimes one invokes asymptotic approximations or imposes parametric assumptions, or does both, to simplify a criterion. If the resultant (approximated) criterion can be optimized with respect to a bandwidth explicitly, an asymptotically optimal choice of this bandwidth can be derived. Plug-in methods are based on so-obtained bandwidths, such as the normal reference rule (Silverman, 1986; Scott, 2015). For more complex criteria, a cross validation strategy is often used to estimate the criterion and search for bandwidths that optimize the estimated criterion. Besides plug-in methods and CV methods, Jones et al. (1996) reviewed other bandwidth selection methods for density estimation, including the ones that involve bootstrap estimation of a criterion.

The main challenge bandwidth selection methods attempt to overcome is estimation of the aforementioned criteria. Criteria like ISE, MISE, or asymptotic MISE (AMISE) depend on complicated functionals of unknown densities, and estimating these functionals is often a harder problem than the original problem of density estimation. This challenge is even more formidable in the presence of measurement error. To select bandwidths for marginal density estimation in the presence of measurement error, Delaigle and Gijbels (2004a,b) developed plug-in methods and bootstrap methods based on MISE or AMISE, which require estimation of functionals such as the integrated squared density derivatives using error-prone data. Delaigle and Gijbels (2002) constructed estimators for these functionals, which again involve bandwidths selection.

Later, Delaigle and Hall (2008) combined cross validation with the strategy of simulation extrapolation (SIMEX, Cook and Stefanski, 1994; Stefanski and Cook, 1995) to choose bandwidths in the presence of measurement error. Their

CV-SIMEX method entails estimating a CV criterion and finding a bandwidth twice using error-contaminated data (at two levels of contamination) in the same way one would do when data are error-free. The resulting two bandwidths together lead to a bandwidth accounting for measurement error via an extrapolation step. Compared to methods considered in Delaigle and Gijbels (2004a,b), one novelty of the CV-SIMEX method is that it avoids direct estimation of a CV criterion accounting for measurement error. This is achieved at the price of increased computational burden caused by the combination of CV and SIMEX, each of which is computationally expensive on its own. Moreover, what extrapolant function should be used at the extrapolation step is rarely known (Carroll et al., 2006, Section 5.3.2). Indeed, the extrapolant used in Delaigle and Hall (2008) is only asymptotically justified, i.e., for large sample, under the assumption that error contamination is close to none. For a given application, it is difficult to gauge if the sample size is large enough, relative to the amount of error contamination, for the extrapolation step to yield a bandwidth improving over a naive bandwidth one chooses while ignoring measurement error. A more realistic goal one can achieve by applying the CV-SIMEX method with caution is to somewhat adjust a naive bandwidth in the right direction. This direction is usually upward when measurement errors compromise naive estimation, because, intuitively, a wider bandwidth is needed when measurement errors blur the underlying pattern of association between two variables. Indeed, we observe that a bandwidth used in the proposed estimators that is larger than the naive bandwidth typically yields more satisfactory results in our extensive simulation study.

4.2. Bandwidth selection for $\hat{p}_3(y|x)$

The one-step estimator $\hat{p}_3(y|x)$ depends on two bandwidths in $\mathbf{h} = (h_1, h_2)$. We propose to choose $\mathbf{h}$ by adjusting the naive bandwidths, denoted by $\mathbf{h}_{\text{nv}}^{(1)} = (h_{\text{nv},1}^{(1)}, h_{\text{nv},2}^{(1)})$, obtained via a CV method for estimating $p^*(y|x)$ using $\tilde{p}_1(y|x)$. In particular, we employ the CV method proposed by Fan and Yim (2004) and Hall et al. (2004) to obtain $\mathbf{h}_{\text{nv}}^{(1)}$.

As an estimator for $p^*(y|x)$, the authors considered the ISE of $\tilde{p}_1(y|x)$ given by

$$\begin{aligned}\text{ISE}(\tilde{p}_1) &= \iint \{\tilde{p}_1(y|x) - p^*(y|x)\}^2 f_W(x)\omega(x)dxdy \\ &= \iint \{\tilde{p}_1(y|x)\}^2 f_W(x)\omega(x)dxdy - 2\iint \tilde{p}_1(y|x)p^*(x,y)\omega(x)dxdy \\ &\quad + \int \{p^*(y|x)\}^2 f_W(x)\omega(x)dxdy,\end{aligned} \tag{4.1}$$

where $f_W(x)$ is the pdf of W, and $\omega(x)$ is a nonnegative weight function used to avoid estimating $p^*(y|x)$ at an x around which data are scarce. Observing that the third integral above does not depend on bandwidths, the authors defined a

CV criterion based on the following estimator of the first two integrals in (4.1),

$$\mathrm{CV}(\tilde{p}_1) = \frac{1}{n}\sum_{j=1}^{n}\omega(W_j)\int\{\tilde{p}_{1,-j}(y|W_j)\}^2\,dy - \frac{2}{n}\sum_{j=1}^{n}\omega(W_j)\tilde{p}_{1,-j}(Y_j|W_j), \quad (4.2)$$

where $\tilde{p}_{1,-j}(y|W_j)$ results from computing the estimator $\tilde{p}_1(y|W_j)$ using all observed data except the jth data point, (W_j, Y_j). We set $K_2(t)$ as the Gaussian kernel in $\tilde{p}_1(y|x)$, and thus in $\hat{p}_3(y|x)$ as well. Thanks to this choice of $K_2(t)$, the integral in (4.2) can be derived explicitly, as shown in Appendix B.2, resulting in an elaborated expression of $\mathrm{CV}(\tilde{p}_1)$ provided there. As for the other kernel, $K_1(t)$, in $\tilde{p}_1(y|x)$, and thus also in $\hat{p}_3(y|x)$, we set

$$K_1(t) = \frac{48\cos t}{\pi t^4}\left(1 - \frac{15}{t^2}\right) - \frac{144\sin t}{\pi t^5}\left(2 - \frac{5}{t^2}\right), \quad (4.3)$$

of which the characteristic function is $\phi_{K_1}(s) = (1-s^2)^3 I(-1 \le s \le 1)$, which satisfies Conditions K listed in Section 3.1. Other choices of $K_1(t)$ one may consider that also satisfy Conditions K include the sinc kernel, and the kernel used in Delaigle et al. (2009), of which the characteristic function is $\phi_{K_1}(s) = (1-s^2)^8 I(-1 \le s \le 1)$. As commented in Section 3.1, (K5) in Conditions K can be relaxed when U is ordinary smooth. We keep our choice of $K_1(t)$ to fulfill condition (K5) even when U is ordinary smooth mainly for the numerical stability it renders when computing the deconvoluting kernel $K_1^*(t)$.

Following the CV method, we search bandwidths that minimize $\mathrm{CV}(\tilde{p}_1)$, resulting in $\mathbf{h}_{\mathrm{nv}}^{(1)} = (h_{\mathrm{nv},1}^{(1)}, h_{\mathrm{nv},2}^{(1)})$. Denote by $\mathbf{h}^{(1)} = (h_1^{(1)}, h_2^{(1)})$ the bandwidths we choose for $\hat{p}_3(y|x)$ to estimate $p(y|x)$. Since Y is observed without error, we set $h_2^{(1)} = h_{\mathrm{nv},2}^{(1)}$; and to account for covariate measurement error, we set

$$h_1^{(1)} = \left(1 + |\rho_{wy}|\sqrt{1-\hat{\lambda}}\right) h_{\mathrm{nv},1}^{(1)}, \quad (4.4)$$

where ρ_{wy} is the sample correlation between W and Y, and $\hat{\lambda} = 1 - \sigma_u^2/s_w^2$ is an estimate of the reliability ratio λ, in which s_w^2 is the sample variance of W. The adjustment of $h_{\mathrm{nv},1}^{(1)}$ given in (4.4) is motivated by the following considerations. When there is no measurement error, certainly no adjustment is needed, which is exactly what (4.4) indicates when $\sigma_u^2 = 0$ (yielding $\hat{\lambda} = 1$). When there exists measurement error but X and Y are independent, W and Y are also independent because U is independent of (X, Y). In this case, since both $p(y|x)$ and $p^*(y|x)$ reduce to the marginal density of Y, accounting for measurement error when estimating $p(y|x)$ is not necessary, and thus neither is adjusting bandwidths for measurement error, which is also what (4.4) suggests with ρ_{wy} consistently estimating the zero correlation. In the presence of measurement error, if X and Y are dependent, it is sensible to inflate the naive bandwidth associated with X to adjust for measurement error, with the adjustment depending on the severity of error contamination and the strength of dependence between X and Y, which can be partially assessed by the correlation between them. In summary,

(4.4) suggests use of the naive bandwidth when no adjustment for measurement error is necessary, and it leads to a different bandwidth by adjusting the naive bandwidth in the right direction otherwise.

4.3. Bandwidth selection for $\hat{p}_4(y|x)$

To select bandwidths in $\mathbf{h} = (h_1, h_2)$ for the two-step estimator $\hat{p}_4(y|x)$, we also begin with some naive bandwidths, denoted by $\mathbf{h}_{\text{nv}}^{(2)} = (h_{\text{nv},1}^{(2)}, h_{\text{nv},2}^{(2)})$, obtained from the CV method for estimating $p^*(y|x)$ using $\tilde{p}_2(y|x)$. Here, the CV criterion is

$$\text{CV}(\tilde{p}_2) = \frac{1}{n}\sum_{j=1}^{n}\omega(W_j)\int \tilde{p}_{2,-j}(y|W_j)^2 dy - \frac{2}{n}\sum_{j=1}^{n}\omega(W_j)\tilde{p}_{2,-j}(Y_j|W_j). \quad (4.5)$$

Even though this criterion is similar to (4.2), there are two complications.

First, when $m^*(y|x)$ is estimated by a local polynomial estimator, as done in the majority of our study, $\tilde{p}_2(y|x)$ involves an additional bandwidth h_3 in $\hat{m}^*(y|x)$. In this case, we use the plug-in method for local polynomial regression (Fan and Gijbels, 1996, Chapter 3) implemented by the R function `locpol` to obtain h_3, with $K_3(t)$ being the Gaussian kernel. For the other two kernels $K_1(t)$ and $K_2(t)$, we set them both as the kernel in (4.3) for ease of numerical implementation as commented in Section 3.1. This choice of $K_2(t)$ causes the second complication, which is that the integral in (4.5) for $\text{CV}(\tilde{p}_2)$ now cannot be derived explicitly. To avoid direct evaluation of this integral, we put the Gaussian kernel back for $K_2(t)$ in (4.5), and proceed with the CV method to choose $\mathbf{h} = (h_1, h_2)$. This produces an elaborated expression of $\text{CV}(\tilde{p}_2)$ provided in equation (C.2) in Appendix B.2 that involves residuals defined by $e_j^* = Y_j - \hat{m}^*(W_j)$. Denote by $\mathbf{h}_{\text{nv}}^{(2)*} = (h_{\text{nv},1}^{(2)*}, h_{\text{nv},2}^{(2)*})$ the bandwidths that minimize (C.2). We then set $\mathbf{h}_{\text{nv}}^{(2)} = (h_{\text{nv},1}^{(2)*}, 0.403h_{\text{nv},2}^{(2)*})$ to acknowledge that the kernel used as $K_2(t)$ in the actual $\tilde{p}_2(y|x)$ is not the Gaussian kernel. The factor $c = 0.403$ used in this adjustment for the bandwidth associated with $K_2(t)$ is deduced as follows. Consider generically estimating the density of a random variable V, $f_V(v)$, via a kernel density estimator with $K(t)$ as the kernel. Silverman (1986, page 45) suggested the following reference rule for choosing bandwidth,

$$h = \left[\frac{8\sqrt{\pi}\int K^2(t)dt}{3\left\{\int t^2K(t)dt\right\}^2}\right]^{1/5} s_v n^{-1/5}, \quad (4.6)$$

where s_v is the sample standard deviation of V. For a given sample of size n, (4.6) provides a relationship between h and $K(t)$. If $K(t)$ is the Gaussian kernel, (4.6) suggests the reference rule of $h = 1.06s_v n^{-1/5}$; and if $K(t)$ is given by (4.3), one has $h = 0.427s_v n^{-1/5}$. The ratio of the latter reference rule over the former gives $c = 0.403$, a sensible scale factor to use when one changes from a Gaussian kernel to the kernel in (4.3).

Lastly, once we have $\mathbf{h}_{\text{nv}}^{(2)}$, we use $\mathbf{h}^{(2)} = (h_1^{(2)}, h_2^{(2)})$ in $\hat{p}_4(y|x)$, where $h_2^{(2)} = h_{\text{nv},2}^{(2)}$ and

$$h_1^{(2)} = \left(1 + |\rho_{we^*}|\sqrt{1-\hat{\lambda}}\right) h_{\text{nv},1}^{(2)}, \tag{4.7}$$

in which ρ_{we^*} is the sample correlation between W and e^*. The adjustment in (4.7) is in the same spirit as (4.4), although we use ρ_{we^*} in place of ρ_{wy}. This replacement is motivated by the fact that $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$ are essentially estimating the conditional density of a mean residual given the corresponding covariate.

5. Simulation study

5.1. Simulation design

We are now in the position to compare finite sample performance of the naive estimators, $\tilde{p}_1(y|x)$ and $\tilde{p}_2(y|x)$, and their non-naive counterparts, $\hat{p}_3(y|x)$ and $\hat{p}_4(y|x)$. In the simulation experiments, we consider the following three models of Y given X:

(C1) $[Y|X = x] \sim N\left(m(x), \sigma^2(x)\right)$, where $m(x) = \sin(\pi x/2)$ and $\sigma(x) = \exp(1 - x/3)/8$;
(C2) $[Y|X = x] \sim 0.5N\left(m(x) - 1, \sigma^2(x)\right) + 0.5N\left(m(x) + 1, \sigma^2(x)\right)$, where $m(x) = \sin(\pi x/2)$ and $\sigma(x) = \exp(1 - x/3)/12$;
(C3) $[Y|X = x] \sim N\left(m(x), \sigma^2(x)\right)$, where $m(x) = x$ and $\sigma(x) = \exp(1 - x/3)/8$.

The three primary (conditional) models are formulated to create two contrasting scenarios under which we compare the four density estimators. One scenario is having a unimodal conditional density (as in (C1) and (C3)) versus a multimodal density (as in (C2)); the other scenario is having a nonlinear conditional mean (as in (C1) and (C2)) versus a linear mean (as in (C3)). The designs of these primary models partly follow the illustrative examples in Sugiyama et al. (2010) with heteroscedastic noise.

In conjunction with each of the three primary models, we vary the true covariate distribution, the measurement error distribution, and the reliability ratio to create four configurations of secondary models: (a) $X \sim N(0,1)$, $U \sim \text{Laplace}(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$; (b) $X \sim N(0,1)$, $U \sim \text{Laplace}(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$; (c) $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$; (d) $X \sim \text{Uniform}(-2, 2)$, $U \sim \text{Laplace}(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$. Contrasting (a) and (b) allows comparison under different severity of error contamination in the covariate. Comparing estimates under (a) and (c) can shed light on effects of different types of measurement error on considered estimators. In particular, the Laplace distribution for U under (a) is an example of ordinary smooth error distributions, whereas the normal distribution for U under (c) provides an example of super smooth

distributions. Finally, the contrast of (a) and (d) provides a testbed for inspecting the performance of estimators when the true covariate has an unbounded support compared to when it has a bounded support.

Putting the three primary models with the four secondary model configurations lead to twelve true model settings, according to each of which we generate 200 Monte Carlo (MC) replicates of size $n = 500$. Given each simulated data set, we carry out two rounds of density estimation. In the first round, to mitigate the confounding effect of data-driven bandwidth selection on the estimation quality, we use the approximated theoretical optimal bandwidths associated with each of the four estimators. Generically denote by $\hat{p}(y|x)$ one of the estimators, the approximated theoretical optimal $\mathbf{h} = (h_1, h_2)$ associated with $\hat{p}(y|x)$ is obtained (through a grid search) by minimizing the empirical integrated squared error (EISE),

$$\text{EISE} = \sum_{j=1}^{\mathcal{M}'} \sum_{k=0}^{\mathcal{M}} \{\hat{p}(y_j|x_k) - p(y_j|x_k)\}^2 f_X(x_k)\Delta\Delta', \tag{5.1}$$

where $\{x_k = x_L + k\Delta\}_{k=0}^{\mathcal{M}}$, Δ is the partition resolution, $\mathcal{M}$ is the largest integer no greater than $(x_U - x_L)/\Delta$, in which $x_U = -2$ and $x_L = 2$; and $\{y_j\}_{j=1}^{\mathcal{M}'}$ is a sequence of grid points equally spaced over the observed sample range of Y, with $y_{j+1} - y_j = \Delta'$. The additional bandwidth, h_3, in $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$ is obtained by minimizing

$$\text{EISE}_m = \sum_{k=0}^{\mathcal{M}} \{\hat{m}^*(x_k) - m^*(x_k)\}^2 f_X(x_k)\Delta. \tag{5.2}$$

In the second round, we use the proposed methods in Sections 4.2 and 4.3 to obtain bandwidths for $\hat{p}_3(y|x)$ and $\hat{p}_4(y|x)$, and apply the CV method in the absence of measurement error to choose bandwidths for $\tilde{p}_1(y|x)$ and $\tilde{p}_2(y|x)$. In the CV criteria used for these methods, we set the weight function $\omega(x) = I(x_L \le x \le x_U)$, where x_U and x_L are the 2.5th and 97.5th percentiles of the observed covariate data, respectively. A similar weight function, as an indicator function over the interval of interest regarding the covariate, is used in Fan and Yim (2004, Section 3.3). Besides the practical consideration in regard to covariate values of interest, one may also choose a weight function to avoid numerical difficulties caused by dividing by numbers close or equal to zero when computing the conditional density estimate as discussed in Hall et al. (2004, Section 2).

As stated in Section 4.1, there exists many different bandwidth selection strategies in the context of density estimation. To have a more focused simulation experiment presented in this article, we avoid going beyond comparing our proposed data-driven bandwidths selection methods with the approximated theoretical optimal approaches, although we did compare the former with their naive counterparts (with results omitted here to save space for other findings) and observe noticeable gain in accuracy of density estimation from adopting

the proposed methods. More comprehensive comparisons between various bandwidth selection methods in conjunction with different density estimators besides the four considered here deserve a manuscript dedicated to reporting simulation study of a larger scale.

5.2. Simulation results

To quantitatively compare different density estimators, we use the EISE defined in (5.1) as the metric to assess the quality of estimates. Figures 1–3 present boxplots of EISE under three primary model configurations when the approximated theoretical optimal bandwidths are used. When comparing a naive estimator with a non-naive one, one can see that adjusting for measurement error clearly leads to estimates of better quality in terms of EISE. On the other hand, $\tilde{p}_2(y|x)$ is less compromised by measurement error than $\tilde{p}_1(y|x)$ is. This can be mostly explained by the findings in Hyndman et al. (1996) and Hansen (2004), which suggest that $\tilde{p}_2(y|x)$ often outperforms $\tilde{p}_1(y|x)$ as estimators for $p^*(y|x)$. Even though estimating $p^*(y|x)$ well typically does not imply reliable estimation of $p(y|x)$, a less satisfactory estimator for the former usually leads to less reliable estimation for the latter. Intuition suggests that correcting a better estimator of $p^*(y|x)$ for measurement error can yield a better non-naive estimator of $p(y|x)$. This intuition is supported by the observations from Figures 1–3 that the most reliable estimator for $p(y|x)$ among the four is $\hat{p}_4(y|x)$ in all considered simulation settings. The benefit of the two-step estimator $\hat{p}_4(y|x)$ compared to $\hat{p}_3(y|x)$ is more evident when the mean function is linear (see panel (d) in Figure 1 in contrast to panel (d) in Figure 3). This can serve as evidence for that adjusting for the mean in the first step then estimating the residual conditional density in the second step leads to better estimates for $p(y|x)$ than a one-step estimator; and this improvement is more noticeable when the dependence of Y on the covariate is mostly explained by the conditional mean that can be well estimated in the first step. As pointed out in Section 3.3, $\hat{p}_4(y|x)$ does not offer any gain asymptotically when compared with $\hat{p}_3(y|x)$ if $m(x)$ is a constant function of x. This is clearly also the case in terms of their finite sample performance. To demonstrate this point, we include in Appendix B.2 boxplots of EISE associated with these two estimators and their naive counterparts when data are generated according to a primary model with a constant $m(x)$. From there, one can see that $\hat{p}_3(y|x)$ behaves very similarly as $\hat{p}_4(x|y)$, and the former is less variable than the latter when the fully data-driven bandwidths are used.

Although $\hat{p}_3(y|x)$ substantially improves over $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$ can perform similarly as $\hat{p}_3(y|x)$ in terms of EISE, especially when error contamination is mild (see, for instance, panel (b) in Figures 1–3 where $\lambda = 0.9$). To compare $\tilde{p}_2(y|x)$ and $\hat{p}_3(y|x)$ more closely in regard to bias and variance, we decompose EISE in (5.1) as follows, where the additional subscript, $\mathrm{MC}(\in \{1, \ldots, 200\})$, is added to signify that, under each simulation setting, there are 200 EISE's recorded for a density estimator, and, for each point (x_k, y_j) at which the density estimate and the true densities are evaluated, there are 200 realizations of a

density estimator,

$$\begin{aligned}\text{EISE}_{\text{MC}} &= \sum_{j=1}^{\mathcal{M}'}\sum_{k=0}^{\mathcal{M}}\left\{\hat{p}_{\text{MC}}(y_j|x_k)-p(y_j|x_k)\right\}^2 f_X(x_k)\Delta\Delta' \\ &= \sum_{j=1}^{\mathcal{M}'}\sum_{k=0}^{\mathcal{M}}\left\{\hat{p}_{\text{MC}}(y_j|x_k)-\bar{p}(y_j|x_k)\right\}^2 f_X(x_k)\Delta\Delta' \qquad (5.3)\\ &+ \sum_{j=1}^{\mathcal{M}'}\sum_{k=0}^{\mathcal{M}}\left\{\bar{p}(y_j|x_k)-p(y_j|x_k)\right\}^2 f_X(x_k)\Delta\Delta' \qquad (5.4)\\ &+ 2\sum_{j=1}^{\mathcal{M}'}\sum_{k=0}^{\mathcal{M}}\left\{\hat{p}_{\text{MC}}(y_j|x_k)-\bar{p}(y_j|x_k)\right\}\left\{\bar{p}(y_j|x_k)-p(y_j|x_k)\right\} f_X(x_k)\Delta\Delta',\end{aligned}$$

where $\bar{p}(y_j|x_k)=\sum_{\text{MC}=1}^{200}\hat{p}_{\text{MC}}(y_j|x_k)/200$ for each point (x_k,y_j). With $\bar{p}(y_j|x_k)$ being the empirical mean of an estimator evaluated at (x_k,y_j), (5.3) can be interpreted as an empirical integrated variance (EIV) associated with a considered estimator, and (5.4) can be viewed as an empirical integrated squared bias (EISB) of the estimator. By construction, the EIV in (5.3) varies across different MC replicates, whereas the EISB in (5.4) does not. Figure 4 shows the ratio of the EISB of $\hat{p}_3(y|x)$ over that of $\tilde{p}_2(y|x)$ under the model setting for panels (a) and (b) in Figure 1. The ratio of EIV of $\hat{p}_3(y|x)$ over that of $\tilde{p}_2(y|x)$, and the ratio of the two EISE's are also depicted in Figure 4. Recall that the true model settings under panels (a) and (b) in each aforementioned figure are the same except for the reliability ratio λ, with $\lambda=0.8$ in (a) and $\lambda=0.9$ in (b). Under both levels of error contamination, one can see in Figure 4 that $\text{EISB}(\hat{p}_3)/\text{EISB}(\tilde{p}_2)<1$ and $\text{EIV}(\hat{p}_3)/\text{EIV}(\tilde{p}_2)>1$, suggesting that $\hat{p}_3(y|x)$ does eliminate some bias in the naive estimator $\tilde{p}_2(y|x)$ at the price of an inflated variance. This price is lower when the error contamination is milder, yielding lower ratios of EIV in (b) compared to those in (a); although milder error contamination also diminishes the amount of bias reduction in $\hat{p}_3(y|x)$ compared to $\tilde{p}_2(y|x)$ since the latter is less compromised in the presence of less measurement error. These comparisons between the two estimators in EISB and EIV explain the resemblance of the estimators in terms of EISE, resulting in $\text{EISE}(\hat{p}_3)/\text{EISE}(\tilde{p}_2)\approx 1$ when $\lambda=0.9$.

To compare $\hat{p}_3(y|x)$ and $\hat{p}_4(y|x)$ in regard to bias and variance separately as in Figure 4, we create Figure 5 to present the ratios of the EISB and EIV of $\hat{p}_4(y|x)$ over those of $\hat{p}_3(y|x)$ under the model setting for panels (a) and (b) in Figure 1. Figure 5 clearly suggests that bias reduction is achieved by $\hat{p}_4(y|x)$ compared to $\hat{p}_3(y|x)$ even outside of the special cases considered in Section 3.3, under which we analytically show the superiority of $\hat{p}_4(y|x)$ over $\hat{p}_3(y|x)$.

Figures 6–8 provide boxplots of EISE associated with density estimates when the fully data-driven bandwidths are used. Table 1 presents medians and interquartile ranges of the EISE depicted in these figures. All patterns described earlier are also observed here, implying great potential of the proposed band-

width selection methods to approximate theoretical optimal bandwidths. It is not surprising to see increased variability across all estimates now, with more uncertainty involved in bandwidth selection, and even more fluctuation when attempts are made to adjust bandwidths for measurement error.

In both rounds of simulation experiments, EISE associated with each of the four considered estimates is higher when the underlying conditional density is bimodal compared to when it is unimodal, or when error contamination is more severe. These are all expected since multimodal densities or noisier data create more unwieldy situations for statistical inference in general. Asymptotic results in Sections 3.2 and 3.3 suggest higher variability for the proposed estimators in the presence of super smooth U than when U is ordinary smooth. The observation that EISE's under panel (c) (with normal U) are higher than those in panel (a) (with Laplace U) in each of Figures 1–8 indicates that the comparison of finite sample variance concurs with the large sample variance comparison. Finally, to demonstrate the effect of sample size, we repeat the simulation study using a much smaller sample size. Figures 9 and 10 show simulation results obtained under the setting with the primary model in (C1) with $n = 200$. Comparing with Figures 1 and 6 where $n = 500$, one can still see similar patterns the estimates exhibit, although more variable EISE are observed for all estimators, especially when the fully data-driven bandwidths are used.

As discussed in Hyndman et al. (1996), besides local polynomial estimators, other nonparametric estimators deemed suitable for estimating $m^*(x)$ can be employed in the two-step estimators, such as spline-based estimators. Properties of $\hat{p}_4(y|x)$ in Theorem 3.2, as well as properties of $\tilde{p}_2(y|x)$ established in Hyndman et al. (1996) and Hansen (2004), remain valid provided that the adopted $\hat{m}^*(x)$ converges to $m^*(x)$ faster than the kernel density estimator for the joint density of (W, e^*) converges to the truth. As an example, we use cubic spline estimates for $m^*(x)$ in $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$, and repeat the second round of the simulation experiments. As counterpart plots of Figures 6–8, Appendix B.2 provides these additional boxplots of EISE, which are mostly comparable with Figures 6–8.

6. Application to dietary data

The data set to be analyzed in this section is from the Women's Interview Survey of Health, which contains the food frequency questionnaire (FFQ) intake, measured as percent calories from fat, and six 24-hour food recalls from 271 subjects. It is of interest to estimate the density of the logarithm of FFQ intake (Y) conditioning on one's long-term usual intake (X). The covariate of interest, the long-term usual intake, cannot be observed directly. A common practice in epidemiology studies is to use data from 24-hour food recalls to construct a surrogate (W) of the true covariate. For instance, Liang and Wang (2005) used the average of two 24-hour food recalls from a subject as W and studied the mean of the log-FFQ intake conditioning on X and other error-free covariates; Wang et al. (2012) used the average of six 24-hour food recalls as W and estimated

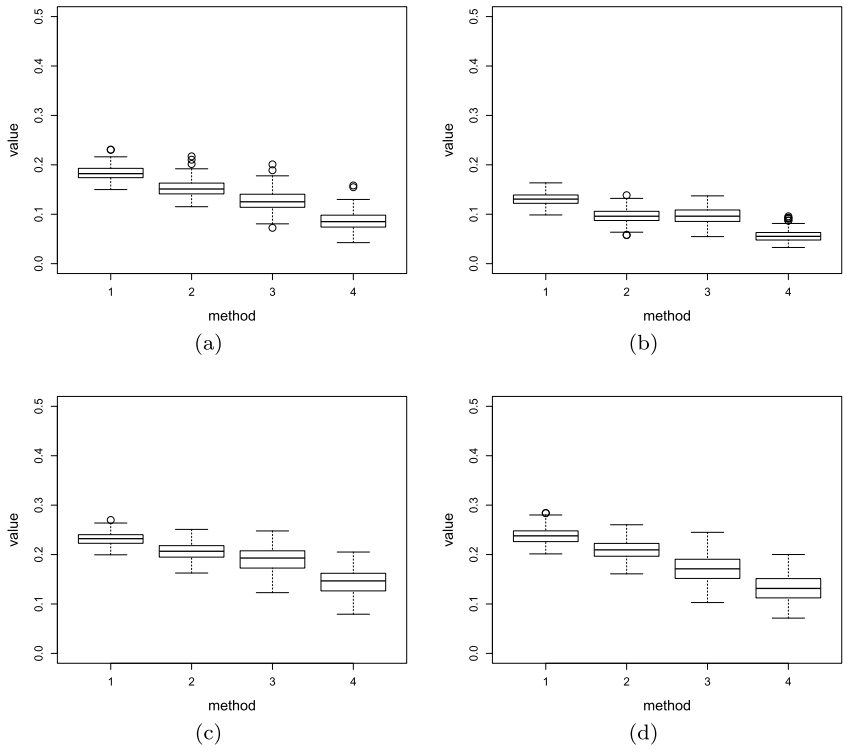

FIG 1. *Boxplots of EISE using the approximated theoretical optimal bandwidths when the primary model is (C1) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

conditional quantiles of the log-FFQ intake. We follow the construction of W in Wang et al. (2012), associated with which the estimated reliability ratio is 0.737. Panel (d) in Figure 11 shows the scatter plot of the log-FFQ versus the so-constructed W from this data set.

For illustration purposes, we estimate the conditional density of the log-FFQ when the long-term usual intake is equal to 6.8, 7.3, and 7.8, respectively. Panels (a)–(c) in Figure 11 depict four estimated density curves, $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, and $\hat{p}_4(y|x)$, at each of the three covariate values. At $x = 6.8$, the two two-step estimates, $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$, are similar but the latter exhibits more distinct peak features, which can be a sign that $\hat{p}_4(y|x)$ corrects $\tilde{p}_2(y|x)$ for measurement error to recover the height around modes of the underlying density. The other non-naive estimate, $\hat{p}_3(y|x)$, resembles $\hat{p}_4(y|x)$ around the highest peak more than the two naive estimates do, and it also differs noticeably from its naive counterpart $\tilde{p}_1(y|x)$ at other regions of the support of Y. At $x = 7.3$, around which data are denser, the difference among the four esti-

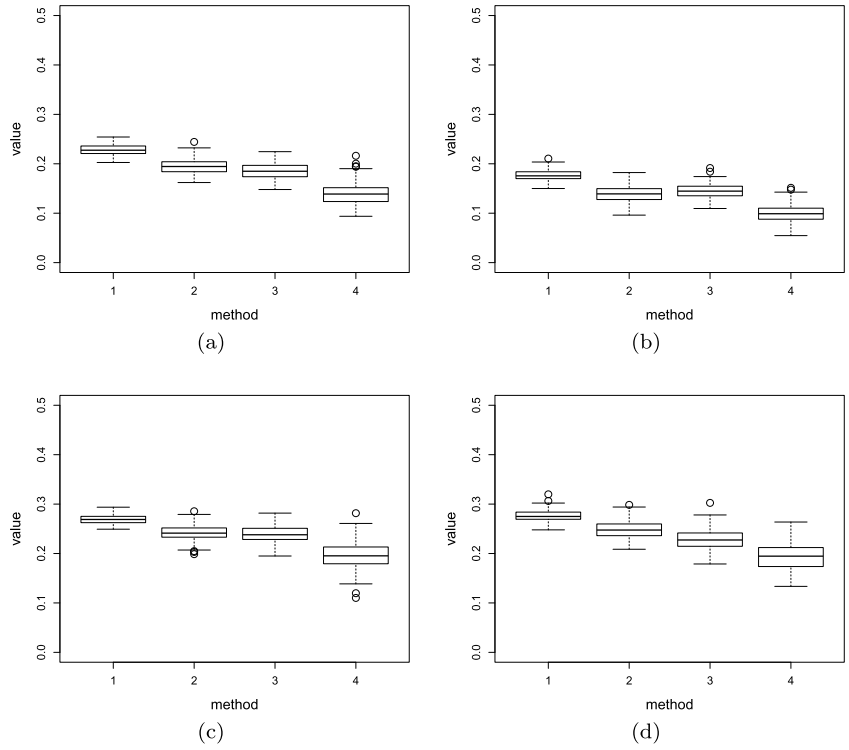

FIG 2. *Boxplots of EISE using the approximated theoretical optimal bandwidths when the primary model is (C2) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

mated density curves appears to be mostly due to whether one uses two-step estimates or one-step estimates. This can be viewed as an example where the effect of measurement error is mild and the two-step estimates lead to improved estimates compared to the one-step estimates. Finally, at $x = 7.8$, around which data become scarce and the association between the response and the covariate may be weaker, the four estimated density curves are less distinguishable. The similarity among the four estimates can be due to low correlation between the response and the true covariate, or that the conditional mean of the response is nearly constant, or lack of sufficient data information for the non-naive estimates effectively correct the naive ones.

We repeat the estimation based on $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$ using the cubic spline estimate for $m^*(\cdot)$ and obtain comparable results in terms of how four estimated densities compare. A figure showing these estimated density curves is given in Appendix B.2. Unlike in simulation studies, here, we actually do not know the measurement error variance σ_u^2 or the distribution family for the measurement

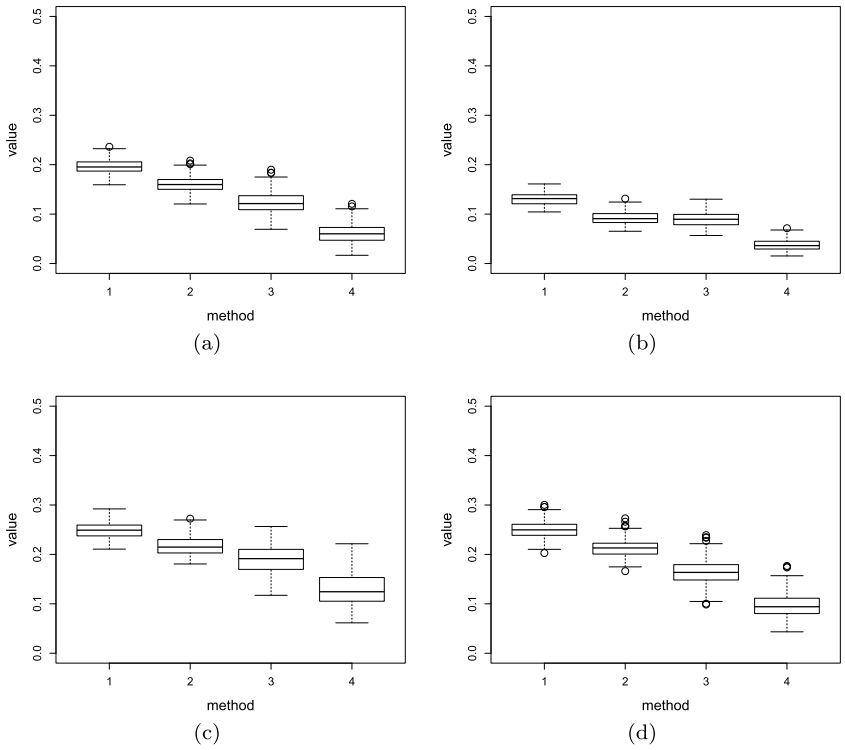

FIG 3. *Boxplots of EISE using the approximated theoretical optimal bandwidths when the primary model is (C3) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

error. We resolved this complication by estimating σ_u^2 via equation (4.3) in Carroll et al. (2006) using repeated measurements (i.e., six 24-hour food recalls from each subject) while assuming Laplace measurement error. Other approaches for estimating σ_u^2 are discussed in Carroll (2014), including that based on correlated repeated measurements (Wang et al., 1996) and those based on validation data or instrumental variables (Buzas et al., 2014). This treatment gives rise to two practical concerns we address next. The first concern relates to misspecification of σ_u^2 in the proposed estimators since an estimated error variance in place of its truth is now used in these estimators. Appendix B.2 presents additional numerical experiments where we repeat part of the simulation studies described in Section 5 but with σ_u^2 set at values different from its truth when obtaining $\hat{p}_3(y|x)$ and $\hat{p}_4(y|x)$. Besides via $\phi_U(t)$, these two estimators also depend on σ_u^2 via bandwidths chosen by the data-driven methods proposed in Section 4. Despite the two sources of dependence on σ_u^2, realizations of $\hat{p}_3(y|x)$ and $\hat{p}_4(y|x)$ from the experiments tend to exhibit smaller EISE than those associated with

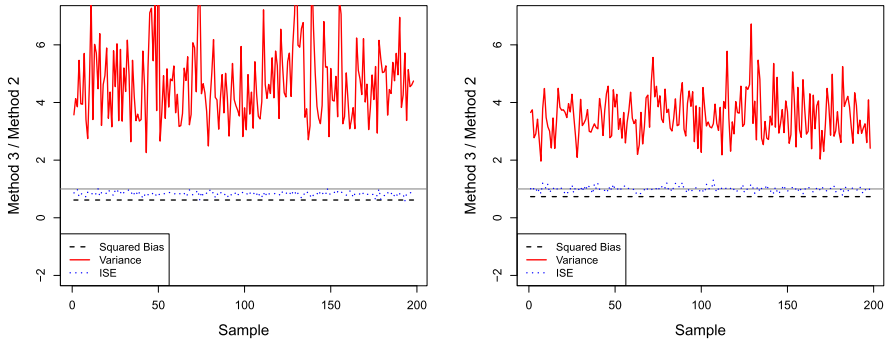

FIG 4. *The ratio of the empirical integrated squared bias (black dashed lines) associated with* $\hat{p}_3(y|x)$ *across 200 Monte Carlo replicates over that associated with* $\tilde{p}_2(y|x)$*, the ratio of the empirical integrated variance (red solid lines) between them, and the ratio of EISE (blue dotted lines) between them, using the approximated theoretical optimal bandwidths when the primary model is (C1) and the secondary models are (a)* $X \sim N(0,1)$*,* $U \sim Laplace(0, \sigma_u/\sqrt{2})$*,* $\lambda = 0.8$*; (b)* $X \sim N(0,1)$*,* $U \sim Laplace(0, \sigma_u/\sqrt{2})$*,* $\lambda = 0.9$*. Method 2 and 3 correspond to* $\tilde{p}_2(y|x)$ *and* $\hat{p}_3(y|x)$*, respectively. The black horizontal solid lines are the reference lines at value one.*

their naive counterparts even when a wrong error variance is used. Hence, although the proposed estimators for $p(y|x)$ are compromised by a misspecified error variance, they remain more superior than the naive estimators provided that the misspecification is not close to ignoring measurement error, e.g., as a result of substantially underestimating σ_u^2. The downside of overestimating σ_u^2 is inflated variability of the non-naive estimators as evidenced in the simulation presented in Appendix B.2. The second concern is in regard to the assumed error distribution. It has been reported in abundant existing studies that nonparametric inference are often fairly robust to distributional assumptions on measurement error (e.g., Delaigle et al., 2009; Zhou and Huang, 2016; Huang and Zhou, 2017). Meister (2004) and Delaigle (2008) provided more theoretical insight on this robustness. If one feels uneasy at assuming a measurement error distribution, one may estimate the characteristic function of U using repeated measurements as proposed by Delaigle et al. (2008), and use this estimate in place of $\phi_U(u)$ in the density estimators. We conjecture that theoretical properties of the resulting density estimators that involve such estimated $\phi_U(u)$ can be derived following similar lines of arguments in Delaigle et al. (2008), which are beyond the scope of the current study.

7. Discussions

In this study we propose two conditional density estimators that account for covariate measurement error by correcting two existing kernel density estimators developed for error-free data. An R code example is provided in Appendix B.2

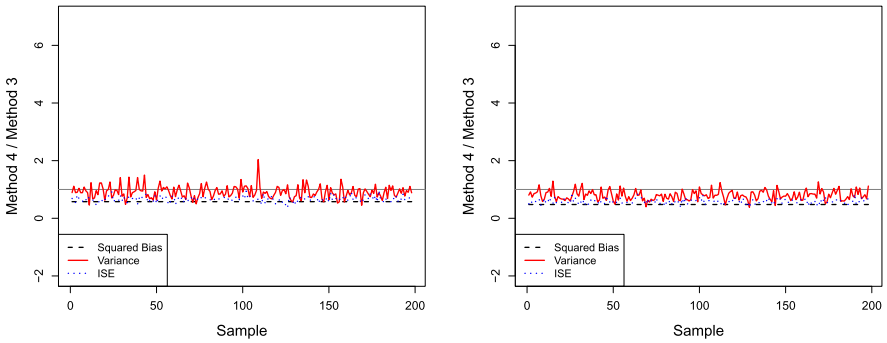

FIG 5. *The ratio of the empirical integrated squared bias (black dashed lines) associated with* $\hat{p}_4(y|x)$ *across 200 Monte Carlo replicates over that associated with* $\hat{p}_3(y|x)$*, the ratio of the empirical integrated variance (red solid lines) between them, and the ratio of EISE (blue dotted lines) between them, using the approximated theoretical optimal bandwidths when the primary model is (C1) and the secondary models are (a)* $X \sim N(0,1)$*,* $U \sim$ *Laplace*$(0, \sigma_u/\sqrt{2})$*,* $\lambda = 0.8$*; (b)* $X \sim N(0,1)$*,* $U \sim$ *Laplace*$(0, \sigma_u/\sqrt{2})$*,* $\lambda = 0.9$*. Method 3 and 4 correspond to* $\hat{p}_3(y|x)$ *and* $\hat{p}_4(y|x)$*, respectively. The black horizontal solid lines are the reference lines at value one.*

to demonstrate use of the R package `lpme` to obtain all four density estimates. When the conditional mean of the response contributes a lot to explaining the dependence of the response on the covariate, the two-step estimators $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$ can substantially benefit from first estimating the mean function. This strategy can even alleviate to some extent the adverse effect of measurement error on naive estimation, even though it can bring in more variability given a finite sample. As one may expect, there will be little return in the effort to account for measurement error when the error contamination is very small. Figure 12 provides comparisons between the four estimators considered in Section 5 in such a scenario, where data for responses are generated according to the primary model in (C1), and $X \sim N(0,1)$, $U \sim \text{Laplace}(0, \sigma_u/\sqrt{2})$, with $\lambda = 0.99$. When the approximated optimal theoretical bandwidths are used, one can see in this figure high resemblance between $\tilde{p}_1(y|x)$ and $\hat{p}_3(y|x)$, as well as between $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$. In addition, Figure 12 indicates that, when one makes the extra effort to select bandwidths using the fully data-driven methods proposed in Section 4, the proposed non-naive estimators exhibit higher variability than their naive counterparts, making the proposed estimators less appealing without gaining noticeable bias reduction.

On the theoretical side, in addition to deriving the asymptotic bias and variance of each proposed estimator, we provide an in-depth comparison between the proposed estimators and their error-free counterparts in regard to the dominating bias. We believe that asymptotic normality of $\hat{p}_3(y|x)$ can be established by proving the Lyapunov's conditions (Billingsley, 2008) under additional regularity conditions following arguments similar to those in Huang and Zhou (2017), although showing the same conditions for $\hat{p}_4(y|x)$ can be much more formidable

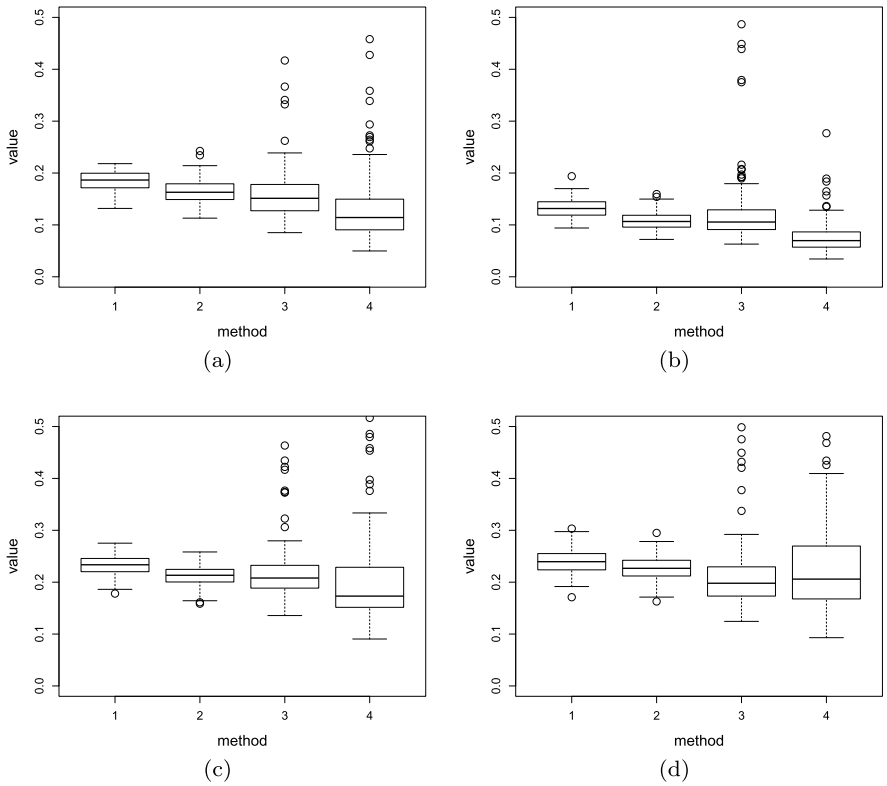

FIG 6. *Boxplots of EISE using the fully data-driven bandwidths when the primary model is (C1) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

due to the higher (than two) order moments of (3.14) arising in the proof. Given the substantial content of this article, we set aside the study of asymptotic normality of proposed estimators and impacts of the smoothness of the covariate and measurement error distributions on their asymptotic distributions for a separate technical note.

The construction of the proposed estimators can be generalized to estimate a multivariate conditional density given a multivariate covariate, some or all elements of which are prone to measurement error. But kernel-based estimators are less well received when there are many variables involved due to the curse of dimensionality (Scott, 2015, Chapter 7) among several other reasons. The use of the integral transform in (3.5) to account for measurement error in some variables, as done in (3.13) and (3.14), only magnifies the challenges in implementing kernel-based density estimation in high dimensional settings. New

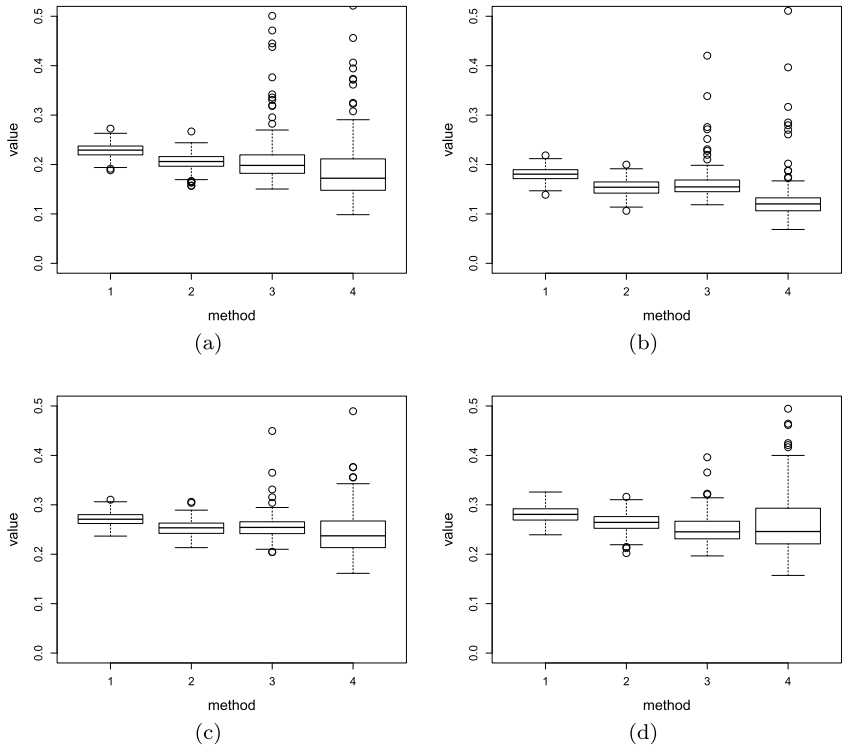

FIG 7. *Boxplots of EISE using the fully data-driven bandwidths when the primary model is (C2) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

strategies for nonparametric conditional density estimation are needed in these settings.

Bandwidth selection has been a hurdle for which no unified solution seems to exist that is numerically convenient and effective for most kernel-based estimation problems. We develop strategies for our proposed estimators aiming to, first, take advantage of existing well accepted bandwidth selection methods in the absence of measurement error, and second, adjust the bandwidths for measurement error in the right direction. Achieving the first goal frees one from estimating a CV criterion using error-prone data. We reach the second goal by a simple adjustment of naive bandwidths that depends on the severity of measurement error and the correlation between the covariate and the response or a mean residual. A more refined adjustment demands systematic investigation on relationships between naive bandwidths and theoretically optimal bandwidths accounting for measurement error.

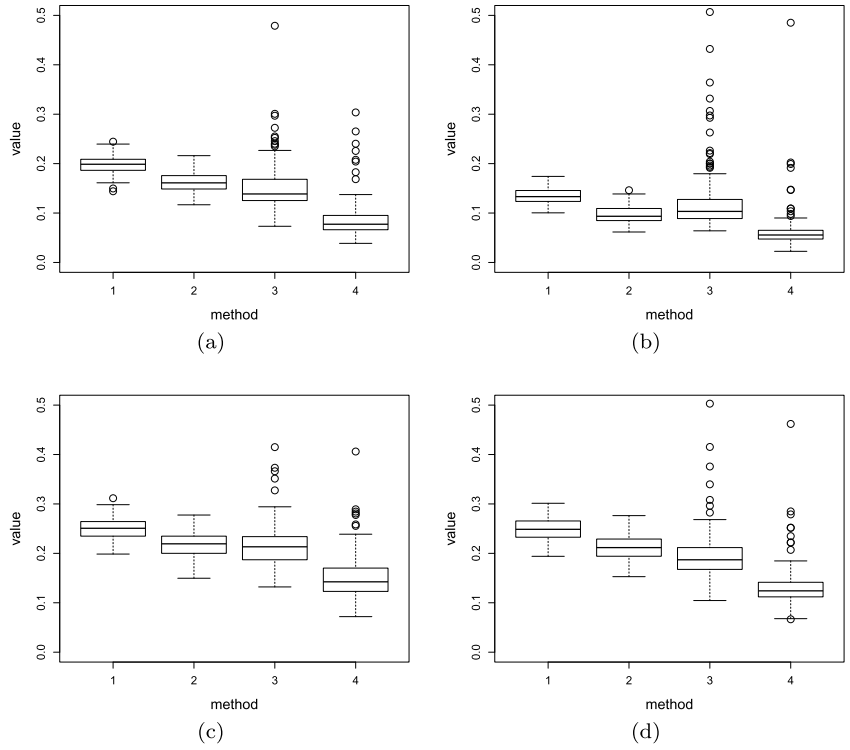

FIG 8. *Boxplots of EISE using the fully data-driven bandwidths when the primary model is (C3) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

Appendix A: Proof of Theorem 3.1

The construction of $\hat{p}_3(y|x)$ can be viewed as $\hat{p}_3(y|x) = \hat{f}_X^{-1}(x)\hat{p}_3(x,y)$, where $\hat{f}_X(x)$ is the deconvoluting kernel estimator of $f_X(x)$ in (2.4), and

$$\hat{p}_3(x,y) = \frac{1}{nh_1h_2}\sum_{j=1}^{n} K_1^*\left(\frac{W_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right) \tag{A.1}$$

is an estimator of $p(x,y)$.

We next approximate $\hat{f}_X(x)$ and $\hat{p}_3(x,y)$ via the decomposition, $A = E(A) + O_p\{\sqrt{\mathrm{Var}(A)}\}$, for a random variable A under regularity conditions.

A.1. Approximation of $\hat{f}_X(x)$

Because the mean of $\hat{f}_X(x)$ is the same as the mean of the regular kernel density estimator of $f_X(x)$ in the absence of measurement error, which is well established

TABLE 1

Medians and interquartile ranges (in parenthesis) of EISE associated with four estimators across 200 Monte Carlo replicates when the fully data-driven bandwidths are used. Data are generated according to primary models (C1)–(C3) along with secondary models (a)–(d) formulated in Section 5.1. Method 1, 2, 3, 4 correspond to $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, and $\hat{p}_4(y|x)$, respectively

Model	Method	(a)	(b)	(c)	(d)
(C1)	1	0.186 (0.028)	0.134 (0.021)	0.234 (0.023)	0.239 (0.028)
	2	0.163 (0.030)	0.108 (0.022)	0.215 (0.027)	0.225 (0.029)
	3	0.151 (0.049)	0.112 (0.044)	0.205 (0.033)	0.194 (0.057)
	4	0.114 (0.059)	0.075 (0.029)	0.181 (0.051)	0.195 (0.154)
(C2)	1	0.230 (0.020)	0.179 (0.015)	0.270 (0.017)	0.276 (0.021)
	2	0.206 (0.023)	0.150 (0.020)	0.254 (0.020)	0.263 (0.022)
	3	0.196 (0.036)	0.153 (0.021)	0.254 (0.025)	0.245 (0.038)
	4	0.167 (0.059)	0.115 (0.030)	0.240 (0.062)	0.251 (0.086)
(C3)	1	0.200 (0.026)	0.135 (0.022)	0.250 (0.025)	0.251 (0.031)
	2	0.164 (0.029)	0.096 (0.023)	0.220 (0.027)	0.213 (0.037)
	3	0.145 (0.042)	0.105 (0.036)	0.209 (0.045)	0.182 (0.049)
	4	0.079 (0.025)	0.056 (0.018)	0.143 (0.044)	0.126 (0.042)

(Scott, 2015, equation (6.16)), one has

$$E\left\{\hat{f}_X(x)\right\} = f_X(x) + 0.5 f_X''(x)\mu_{2,1}h_1^2 + O(h_1^4), \tag{A.2}$$

where $f_X''(x)$ is the second derivative of $f_X(x)$, and $\mu_{2,\ell} = \int t^2 K_\ell(t)dt$, for $\ell = 1, 2$.

Also similar to the variance result for the ordinary kernel density estimator of $f_X(x)$ (Scott, 2015, equation (6.17)), one can show that

$$\mathrm{Var}\left\{\hat{f}_X(x)\right\} = \frac{f_W(x)R(K_1^*)}{nh_1} + O(n^{-1}), \tag{A.3}$$

where $f_W(\cdot)$ is the density of W, and $R(K_1^*) = \int\{K_1^*(t)\}^2 dt$. In the sequel, we use $R(g)$ to denote $\int g^2(t)dt$ for a square integrable function $g(t)$. Note that $R(K_1^*)$ depends on h_1 since $K_1^*(t)$ depends on h_1, and by Lemmas B.4 and B.9 in Delaigle et al. (2009), under Conditions U and Conditions K in the main article,

$$R(K_1^*) = \begin{cases} O(h_1^{-2b}), & \text{if } U \text{ is ordinary smooth,} \\ O\left\{h_1^{2b_2}\exp(2h_1^{-b}/d_2)\right\}, & \text{if } U \text{ is super smooth,} \end{cases} \tag{A.4}$$

where $b_2 = b_0 I(b_0 < 0.5)$.

By (A.2)–(A.4), one has, when U is ordinary smooth,

$$\hat{f}_X(x) = f_X(x) + 0.5 f_X''(x)\mu_{2,1}h_1^2 + O(h_1^4) + O_p\left(\frac{1}{\sqrt{nh_1^{1+2b}}}\right), \tag{A.5}$$

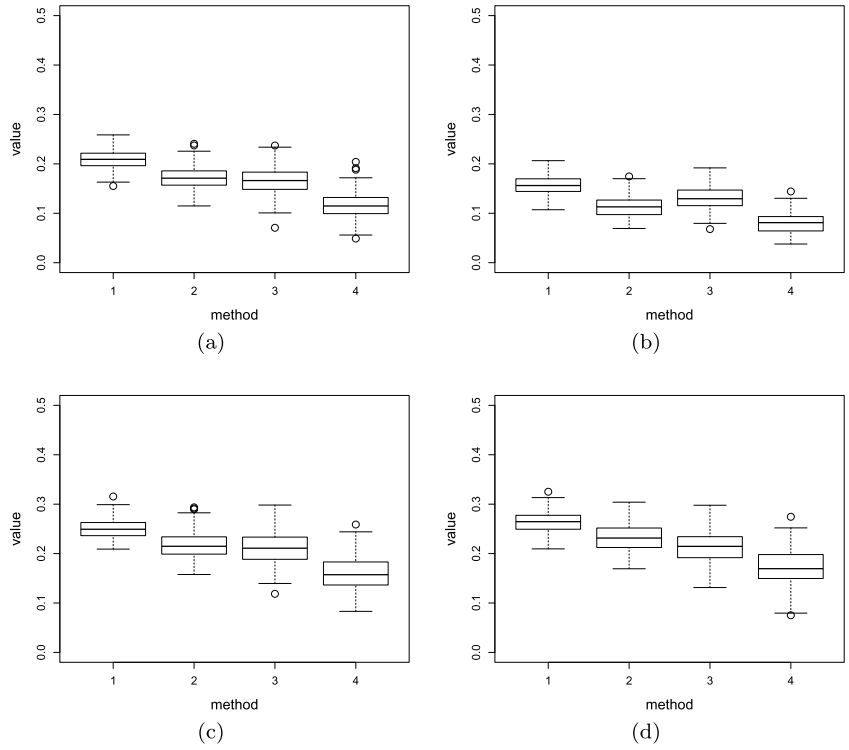

FIG 9. *Boxplots of EISE using the approximated theoretical optimal bandwidths when the primary model is (C1) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively. The sample size is* $n = 200$.

and, when U is super smooth,

$$\hat{f}_X(x) = f_X(x) + 0.5 f_X''(x)\mu_{2,1}h_1^2 + O(h_1^4) + O_p\left\{\frac{\exp(h_1^{-b}/d_2)}{\sqrt{nh_1^{1-2b_2}}}\right\}. \tag{A.6}$$

Following (A.5) and (A.6), one has, for ordinary smooth U,

$$\hat{f}_X^{-1}(x) = f_X^{-1}(x) - 0.5 f_X^{-2}(x) f_X''(x)\mu_{2,1}h_1^2 + O(h_1^4) + O_p\left(\frac{1}{\sqrt{nh_1^{1+2b}}}\right), \tag{A.7}$$

and, for super smooth U,

$$\hat{f}_X^{-1}(x) = f_X^{-1}(x) - 0.5 f_X^{-2}(x) f_X''(x)\mu_{2,1}h_1^2 + O(h_1^4) + O_p\left\{\frac{\exp(h_1^{-b}/d_2)}{\sqrt{nh_1^{1-2b_2}}}\right\}. \tag{A.8}$$

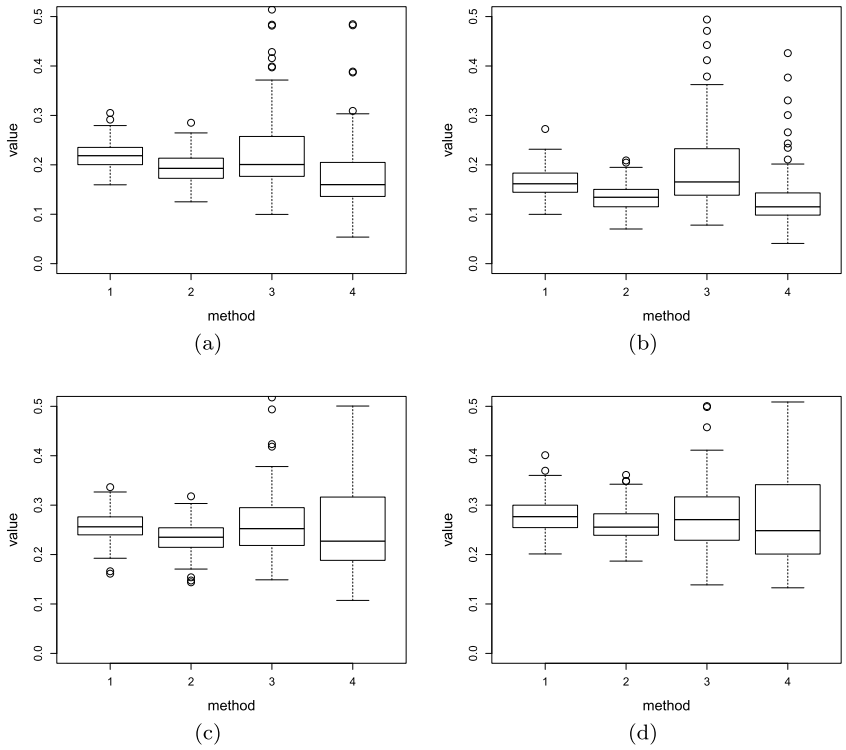

FIG 10. *Boxplots of EISE using the fully data-driven bandwidths when the primary model is (C1) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively. The sample size is* $n = 200$.

A.2. Approximation of $\hat{p}_3(x, y)$

Because the mean of $\hat{p}_3(x, y)$ is the same as the mean of the regular bivariate kernel density estimator of $p(x, y)$ in the absence of measurement error, which has been established (Scott, 2015, equation (6.40)), one has

$$\begin{aligned} &E\{\hat{p}_3(x,y)\} \\ =&p(x,y) + 0.5\{p_{xx}(x,y)\mu_{2,1}h_1^2 + p_{yy}(x,y)\mu_{2,2}h_2^2\} + O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2), \end{aligned} \tag{A.9}$$

where $p_{xx}(x,y) = (\partial^2/\partial x^2)p(x,y)$ and $p_{yy}(x,y) = (\partial^2/\partial y^2)p(x,y)$.

Following similar derivations leading to the asymptotic variance of a regular bivariate kernel density estimator in the absence of measurement error (Scott, 2015, equation (6.41)), one can show that

$$\text{Var}\{\hat{p}_3(x,y)\} = \frac{p^*(x,y)R(K_1^*)R(K_2)}{nh_1h_2} + O(n^{-1}). \tag{A.10}$$

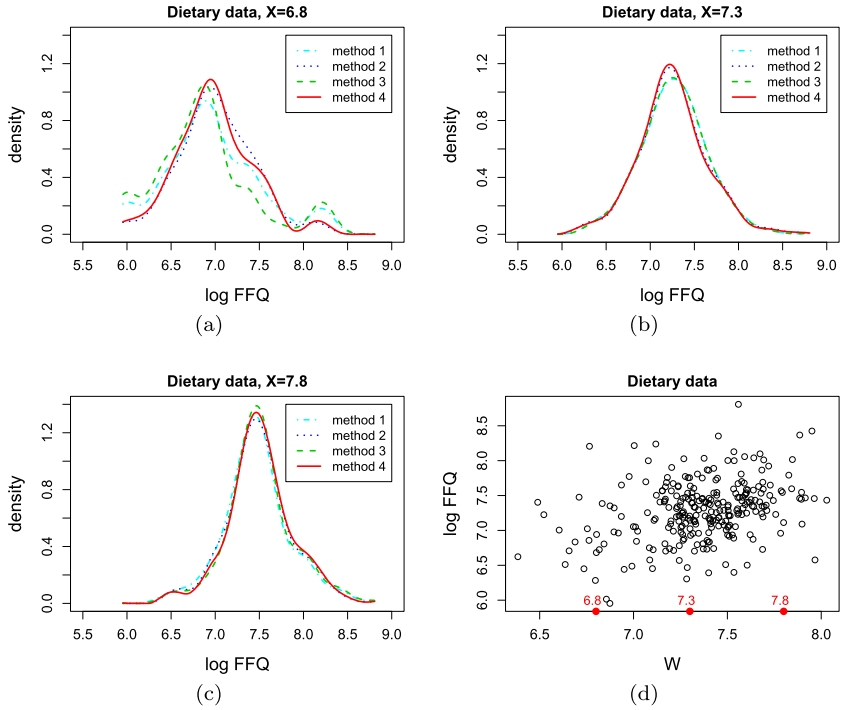

FIG 11. *Naive estimates of the conditional density of the logarithm of FFQ intake corresponding to* $\tilde{p}_1(y|x)$ *(cyan dash-dotted lines) and* $\tilde{p}_2(y|x)$ *(blue dotted lines), and two non-naive density estimates,* $\hat{p}_3(y|x)$ *(green dashed lines) and* $\hat{p}_4(y|x)$ *(red solid lines) when* $x = 6.8$ *(in panel (a)), 7.3 (in panel (b)), and 7.8 (in panel (c)), respectively. In each panel of (a)–(c), method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$*,* $\tilde{p}_2(y|x)$*,* $\hat{p}_3(y|x)$*, and* $\hat{p}_4(y|x)$*, respectively. The scatter plot of the observed response versus the observed covariate values is shown in panel (d), where the three values of* x *at which* $p(y|x)$ *is estimated are highlighted in red dots on the horizontal axis.*

By (A.9), (A.10), and (A.4), one has, for ordinary smooth U,

$$\hat{p}_3(x,y) = p(x,y) + 0.5\{p_{xx}(x,y)\mu_{2,1}h_1^2 + p_{yy}(x,y)\mu_{2,2}h_2^2\} + O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2) + O_p\left(\frac{1}{\sqrt{nh_1^{1+2b}h_2}}\right), \tag{A.11}$$

and, for super smooth U,

$$\hat{p}_3(x,y) = p(x,y) + 0.5\{p_{xx}(x,y)\mu_{2,1}h_1^2 + p_{yy}(x,y)\mu_{2,2}h_2^2\} + O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2) + O_p\left\{\frac{\exp(h_1^{-b}/d_2)}{\sqrt{nh_1^{1-2b_2}h_2}}\right\}. \tag{A.12}$$

The result in Theorem 3.1 regarding $\hat{p}_3(y|x) - p(y|x)$ is obtained from (A.7)

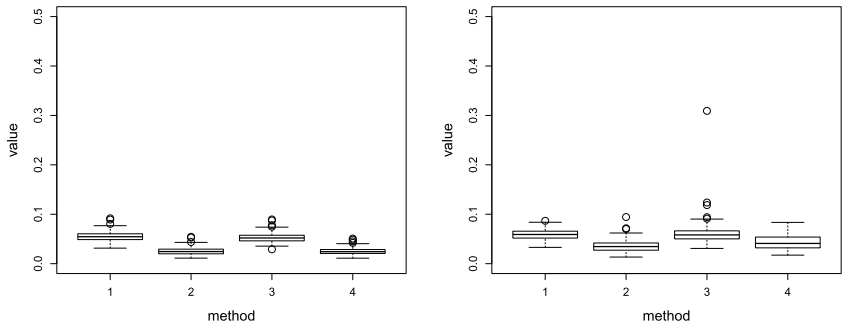

FIG 12. *Boxplots of EISE using the approximated theoretical optimal bandwidths (in panel (a)) and boxplots of EISE using the fully data-driven bandwidths (in panel (b)) when the primary model is (C1) and the secondary model is* $X \sim N(0,1)$, $U \sim$ *Laplace*$(0, \sigma_u/\sqrt{2})$, $\lambda = 0.99$. *Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively. The sample size is* $n = 500$.

multiplying (A.11) for ordinary smooth U, and (A.8) multiplying (A.12) for super smooth U.

Appendix B: Proof of Theorem 3.2

Because the integral transform $\mathscr{T}_x(\cdot)$ defined in (3.5) is a linear operator by construction, it can commute with another linear operator, such as expectation. In addition, $\mathscr{T}_x\{g(\cdot, y)\} = g(x, y)$ if $\phi_U(t) = 1$ for all t.

The construction of $\hat{p}_4(y|x)$ originates from $\hat{p}_4(y|x) = \hat{f}_X^{-1}(x)\hat{p}_4(x, y)$, where $\hat{p}_4(x, y)$ is an estimator of $p(x, y)$ obtained via the aforementioned integral transform of the following estimator for $p^*(x, y)$,

$$\tilde{p}_2(x,y) = \frac{1}{nh_1h_2}\sum_{j=1}^{n} K_1\left(\frac{W_j - x}{h_1}\right) K_2\left\{\frac{Y_j - \hat{m}^*(W_j) - y + \hat{m}^*(x)}{h_2}\right\}. \quad \text{(B.1)}$$

More specifically,

$$\begin{aligned}\hat{p}_4(x,y) &= \frac{1}{2\pi}\int e^{-itx}\frac{\phi_{\tilde{p}_2(\cdot,\,y)}(t)}{\phi_U(t)}dt \\ &= \frac{1}{2\pi}\int e^{-itx}\frac{\int e^{itw}\tilde{p}_2(w,y)dw}{\phi_U(t)}dt \\ &= \mathscr{T}_x\{\tilde{p}_2(\cdot,y)\}. \end{aligned} \quad \text{(B.2)}$$

We next use the mean and variance results for $\tilde{p}_2(x, y)$ to obtain those for $\hat{p}_4(x, y)$.

Hansen (2004) showed that, despite the extra estimation of $m^*(\cdot)$, the asymptotic variance of the two-step estimator for $p^*(x, y)$, namely $\tilde{p}_2(x, y)$, is of the

same order as that of the one-step estimator given by

$$\tilde{p}_1(x,y) = \frac{1}{nh_1h_2}\sum_{j=1}^{n} K_1\left(\frac{W_j - x}{h_1}\right) K_2\left(\frac{Y_j - y}{h_2}\right). \tag{B.3}$$

Since $\hat{p}_3(x,y) = \mathscr{T}_x\{\tilde{p}_1(\cdot,y)\}$, which is the same as how $\hat{p}_4(x,y)$ relates to $\tilde{p}_2(x,y)$ in (B.2), the asymptotic variance of $\hat{p}_4(x,y)$ is also of the same order as that of $\hat{p}_3(x,y)$, which is provided in Section A.2.

In our study, we set $\hat{m}^*(\cdot)$ as the local linear estimator of $m^*(\cdot)$ with kernel $K_3(t)$ and bandwidth h_3. Following the proof in Hansen (2004, Section 10), one can show that, if $h_3 = O(h_2)$ as h_2 and h_3 tend to zero,

$$\begin{aligned} E\{\tilde{p}_2(x,y)\} =& p^*(x,y) + 0.5\left\{g_1(x,y)\mu_{2,1}h_1^2 + g_2(x,y)\mu_{2,2}h_2^2\right\} + \\ & O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2), \end{aligned} \tag{B.4}$$

where

$$g_1(x,y) = f^{(2)}_{W,e^*,11}(x,e^*) = \left[\frac{\partial^2}{\partial w^2} f_{W,e^*}(w,e^*)\right]\Big|_{w=x,e^*=y-m^*(x)}, \tag{B.5}$$

$$g_2(x,y) = f^{(2)}_{W,e^*,22}(x,e^*) = \left[\frac{\partial^2}{\partial e^{*2}} f_{W,e^*}(w,e^*)\right]\Big|_{w=x,e^*=y-m^*(x)}, \tag{B.6}$$

in which $f_{W,e^*}(w,e^*)$ is the joint density of W and $e^* = Y - m^*(W)$. It follows that, by commuting the operations of expectation and $\mathscr{T}_x$,

$$\begin{aligned} & E\{\hat{p}_4(x,y)\} \\ =& \mathscr{T}_x[E\{\tilde{p}_2(\cdot,y)\}] \\ =& \mathscr{T}_x\{p^*(\cdot,y)\} + 0.5\left[\mathscr{T}_x\{g_1(\cdot,y)\}\mu_{2,1}h_1^2 + \mathscr{T}_x\{g_2(\cdot,y)\}\mu_{2,2}h_2^2\right] \\ & + O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2), \end{aligned}$$

where

$$\mathscr{T}_x\{p^*(\cdot,y)\} = p(x,y), \tag{B.7}$$

$$\mathscr{T}_x\{g_1(\cdot,y)\} = p_{xx}(x,y) + \sum_{k=2}^{4} \mathscr{T}_x\{I_k(\cdot,y)\}, \tag{B.8}$$

$$\mathscr{T}_x\{g_2(\cdot,y)\} = p_{yy}(x,y), \tag{B.9}$$

in which $I_k(w,y)$, for $k = 2,3,4$, are defined in (3.8). Among (B.7)–(B.9), (B.7) can be proved using (2.1) in the main article, (B.8) and (B.9) are proved in Section B.1. In conclusion, we have

$$\begin{aligned} & E\{\hat{p}_4(x,y)\} \\ =& p(x,y) + 0.5\left(\left[p_{xx}(x,y) + \sum_{k=2}^{4}\mathscr{T}_x\{I_k(\cdot,y)\}\right]\mu_{2,1}h_1^2 + p_{yy}(x,y)\mu_{2,2}h_2^2\right) + \\ & O(h_1^4) + O(h_2^4) + O(h_1^2h_2^2). \end{aligned} \tag{B.10}$$

Combining (B.10) with the variance rate of $\hat{p}_4(x,y)$, we have, for ordinary smooth U,

$$\begin{aligned}&\hat{p}_4(x,y)\\ =&p(x,y)+0.5\left(\left[p_{xx}(x,y)+\sum_{k=2}^{4}\mathscr{T}_x\{I_k(\cdot,y)\}\right]\mu_{2,1}h_1^2+p_{yy}(x,y)\mu_{2,2}h_2^2\right)\\ &+O(h_1^4)+O(h_2^4)+O(h_1^2h_2^2)+O_p\left(\frac{1}{\sqrt{nh_1^{1+2b}h_2}}\right),\end{aligned} \tag{B.11}$$

and, for super smooth U,

$$\begin{aligned}&\hat{p}_4(x,y)\\ =&p(x,y)+0.5\left(\left[p_{xx}(x,y)+\sum_{k=2}^{4}\mathscr{T}_x\{I_k(\cdot,y)\}\right]\mu_{2,1}h_1^2+p_{yy}(x,y)\mu_{2,2}h_2^2\right)\\ &+O(h_1^4)+O(h_2^4)+O(h_1^2h_2^2)+O_p\left\{\frac{\exp(h_1^{-b}/d_2)}{\sqrt{nh_1^{1-2b_2}h_2}}\right\}.\end{aligned} \tag{B.12}$$

The result in Theorem 3.2 regarding $\hat{p}_4(y|x)-p(y|x)$ is obtained from (A.7) multiplying (B.11) for ordinary smooth U, and (A.8) multiplying (B.12) for super smooth U.

B.1. Proof of (B.8) and (B.9)

In order to derive the two transforms on the left-hand side of (B.8) and (B.9), we need to elaborate the two functions, $g_1(x,y)$ and $g_2(x,y)$, defined in (B.5) and (B.6). For notational clarity, we first derive partial derivatives of $f_{W,e^*}(w,e^*)$, viewing it as a function of w and e^*, before evaluating the partial derivatives at $w=x$ and $e^*=y-m^*(x)$ to obtain $g_1(x,y)$ and $g_2(x,y)$.

Because

$$f_{W,e^*}(w,e^*)=p^*(w,y)=\int p(v,y)f_U(w-v)dv, \tag{B.13}$$

where $y=m^*(w)+e^*$, one has

$$\begin{aligned}&\frac{\partial}{\partial w}f_{W,e^*}(w,e^*)\\ =&\left\{\frac{d}{dw}m^*(w)\right\}\int p_y(v,y)f_U(w-v)dv+\int p(v,y)f_U'(w-v)dv\\ =&\left\{\frac{d}{dw}m^*(w)\right\}\int p_y(v,y)f_U(w-v)dv+\int p_x(v,y)f_U(w-v)dv,\end{aligned}$$

where integration-by-part is used to obtain the last integral, $p_x(v,y)$ is equal to $(\partial/\partial x)p(x,y)$ evaluated at $(x,y)=(v,y)$, and $p_y(v,y)$ is equal to $(\partial/\partial y)p(x,y)$ evaluated at $(x,y)=(v,y)$. It follows that

$$\begin{aligned}
&\frac{\partial^2}{\partial w^2} f_{W,e^*}(w,e^*)\\
&= \left\{\frac{d^2}{dw^2}m^*(w)\right\}\int p_y(v,y)f_U(w-v)dv + \left\{\frac{d}{dw}m^*(w)\right\}^2\int p_{yy}(v,y)f_U(w-v)dv+\\
&\quad \left\{\frac{d}{dw}m^*(w)\right\}\int p_y(v,y)f_U'(w-v)dv + \left\{\frac{d}{dw}m^*(w)\right\}\int p_{xy}(v,y)f_U(w-v)dv+\\
&\quad \int p_x(v,y)f_U'(w-v)dv\\
&= \left\{\frac{d^2}{dw^2}m^*(w)\right\}\int p_y(v,y)f_U(w-v)dv + \left\{\frac{d}{dw}m^*(w)\right\}^2\int p_{yy}(v,y)f_U(w-v)dv+\\
&\quad \left\{\frac{d}{dw}m^*(w)\right\}\int p_{xy}(v,y)f_U(w-v)dv + \left\{\frac{d}{dw}m^*(w)\right\}\int p_{xy}(v,y)f_U(w-v)dv+\\
&\quad \int p_{xx}(v,y)f_U(w-v)dv\\
&= \int p_{xx}(v,y)f_U(w-v)dv + \left\{\frac{d^2}{dw^2}m^*(w)\right\}\int p_y(v,y)f_U(w-v)dv+\\
&\quad \left\{\frac{d}{dw}m^*(w)\right\}^2\int p_{yy}(v,y)f_U(w-v)dv + 2\left\{\frac{d}{dw}m^*(w)\right\}\int p_{xy}(v,y)f_U(w-v)dv.
\end{aligned}$$

Evaluating the last expression at $w=x$ and $e^*=y-m^*(x)$ gives $g_1(x,y)=\sum_{k=1}^4 I_k(x,y)$, where $I_1(x,y)$ is equal to $\int p_{xx}(v,y)f_U(w-v)dv$ evaluated at $(w,y)=(x,y)$, and $I_2(x,y)$, $I_3(x,y)$, $I_4(x,y)$ are the three functions defined in (3.8) evaluated at $(w,y)=(x,y)$, respectively. It is worth pointing out that, in the absence of measurement error, (B.13) can be simply viewed as $f_{W,e^*}(w,e^*)=p^*(w,y)=p(x,y)$, which is symbolically equivalent to viewing $\int p(v,y)f_U(w-v)dv$ as $p(x,y)$. Following this viewpoint, one has the following definitions of $I_k(x,y)$ in the absence of measurement error, for $k=2,3,4$,

$$\begin{cases} I_2(x,y) = m''(x)p_y(x,y),\\ I_3(x,y) = \{m'(x)\}^2 p_{yy}(x,y),\\ I_4(x,y) = 2m'(x)p_{xy}(x,y). \end{cases} \tag{B.14}$$

To this end, one has $\mathscr{T}_x\{g_1(\cdot,y)\}=\sum_{k=1}^4\mathscr{T}_x\{I_k(\cdot,y)\}$, where

$$\begin{aligned}
\mathscr{T}_x\{I_1(\cdot,y)\} &= \frac{1}{2\pi}\int e^{-itx}\frac{1}{\phi_U(t)}\int e^{itw}\int p_{xx}(v,y)f_U(w-v)dvdwdt\\
&= \frac{1}{2\pi}\int e^{-itx}\frac{1}{\phi_U(t)}\int e^{itv}p_{xx}(v,y)\int e^{it(w-v)}f_U(w-v)dwdvdt\\
&= \frac{1}{2\pi}\int e^{-itx}\frac{1}{\phi_U(t)}\int e^{itv}p_{xx}(v,y)\phi_U(t)dvdt\\
&= p_{xx}(x,y).
\end{aligned}$$

This proves (B.8), where the latter three transforms, $\mathscr{T}_x\{I_k(\cdot,y)\}$, for $k=2,3,4$, cannot be further simplified in the presence of measurement error without additional assumptions, such as those on $m^*(w)$.

To show (B.9), we first derive $g_2(x,y)$ defined in (B.6). By (B.13), it is easy to see that $f^{(2)}_{W,e^*,22}(w,e^*)=\int p_{yy}(v,y)f_U(w-v)dv$. Thus $g_2(w,y)=\int p_{yy}(v,y)f_U(w-v)dv$. It follows that

$$\begin{aligned}\mathscr{T}_x\{g_2(\cdot,y)\}&=\frac{1}{2\pi}\int e^{-itx}\frac{1}{\phi_U(t)}\int e^{itw}\int p_{yy}(v,y)f_U(w-v)dvdwdt\\&=\frac{1}{2\pi}\int e^{-itx}\frac{1}{\phi_U(t)}\int e^{itv}p_{yy}(v,y)\int e^{it(w-v)}f_U(w-v)dwdvdt\\&=\frac{1}{2\pi}\int e^{-itx}\frac{1}{\phi_U(t)}\int e^{itv}p_{yy}(v,y)\phi_U(t)dvdt\\&=p_{yy}(x,y).\end{aligned}$$

This completes the proof of (B.9).

B.2. Consideration of two special cases in Section 3

We state in Section 3 in the main article that, in the absence of measurement error, (3.12) reduces to $\mathrm{DB}_4(x,y,h_1,h_2)=\mathrm{DB}_2(x,y,h_1,h_2)$. We first prove this statement in this subsection.

Because $p(x,y)=f_{X,e}\{x,y-m(x)\}$, one has

$$\begin{aligned}p_y(x,y)&=(\partial/\partial y)f_{X,e}\{x,y-m(x)\}=f^{(1)}_{X,e,2}(x,e),\\p_{yy}(x,y)&=(\partial^2/\partial y^2)f_{X,e}\{x,y-m(x)\}=f^{(2)}_{X,e,22}(x,e),\\p_{xy}(x,y)&=(\partial/\partial x)f^{(1)}_{X,e,2}\{x,y-m(x)\}=f^{(2)}_{X,e,21}(x,e)-m'(x)f^{(2)}_{X,e,22}(x,e).\end{aligned}$$

Using these three results in (B.14), one has that, in the absence of measurement error,

$$\begin{aligned}&\sum_{k=2}^{4}\mathscr{T}_x\{I_k(\cdot,y)\}\\=&\sum_{k=2}^{4}I_k(x,y)\\=&m''(x)f^{(1)}_{X,e,2}(x,e)+\{m'(x)\}^2f^{(2)}_{X,e,22}(x,e)\\&+2m'(x)\left\{f^{(2)}_{X,e,21}(x,e)-m'(x)f^{(2)}_{X,e,22}(x,e)\right\}\\=&m''(x)f^{(1)}_{X,e,2}(x,e)-\{m'(x)\}^2f^{(2)}_{X,e,22}(x,e)+2m'(x)f^{(2)}_{X,e,21}(x,e),\end{aligned}$$

which cancel with the last three terms in (3.12) in the main article. This proves that $\mathrm{DB}_4(x,y,h_1,h_2)=\mathrm{DB}_2(x,y,h_1,h_2)$ in the absence of measurement error.

In another special case considered in Section 3, we impose the following the conditions stated in Hyndman et al. (1996): (H1) the covariate is locally uniform near x so that $f'_X(x) \approx 0$ and $f''_X(x) \approx 0$, (H2) $e \perp X$ so that $p(y|x) = f_e\{y - m(x)\}$, and (H3) $m(x)$ is locally linear near x so that $m''(x) \approx 0$. Next, we simplify the following dominating bias associated with $\hat{p}_2(y|x)$, $\hat{p}_3(y|x)$, and $\hat{p}_4(y|x)$, respectively,

$$\mathrm{DB}_2(x,y,h_1,h_2) = \frac{1}{2f_X(x)}\left[\left\{f^{(2)}_{X,e,11}(x,e) - p(y|x)f''_X(x)\right\}\mu_{2,1}h_1^2 + p_{yy}(x,y)\mu_{2,2}h_2^2\right], \tag{B.15}$$

$$\mathrm{DB}_3(x,y,h_1,h_2) = \frac{1}{2f_X(x)}\left[\{p_{xx}(x,y) - p(y|x)f''_X(x)\}\,\mu_{2,1}h_1^2 + p_{yy}(x,y)\mu_{2,2}h_2^2\right] \tag{B.16}$$

$$\mathrm{DB}_4(x,y,h_1,h_2) = \frac{1}{2f_X(x)}\left(\left[p_{xx}(x,y) + \sum_{k=2}^{4}\mathscr{T}_x\{I_k(\cdot,y)\} - p(y|x)f''_X(x)\right]\mu_{2,1}h_1^2 + p_{yy}(x,y)\mu_{2,2}h_2^2\right). \tag{B.17}$$

First, by (H2),

$$p_{yy}(x,y) = \frac{\partial^2}{\partial y^2}\{f_X(x)p(y|x)\} = \frac{\partial^2}{\partial y^2}\left[f_X(x)f_e\{y - m(x)\}\right] = f_X(x)f''_e(e). \tag{B.18}$$

Second, by (H1) and (H2), $f^{(2)}_{X,e,11}(x,e)$ and $f''_X(x)$ are approximately zero. Hence, (B.15) reduces to

$$\mathrm{DB}_2(x,y,h_1,h_2) \approx 0.5f''_e(e)\mu_{2,2}h_2^2.$$

Third,

$$\begin{aligned} p_{xx}(x,y) &= \frac{\partial^2}{\partial x^2}\{f_X(x)p(y|x)\} \\ &= \frac{\partial}{\partial x}\left[f'_X(x)f_e(e) - f_X(x)f'_e(e)m'(x)\right], \text{ by (H2)}, \\ &\approx -\frac{\partial}{\partial x}\left[f_X(x)f'_e(e)m'(x)\right], \text{ by (H1)}, \\ &= -f'_X(x)f'_e(e)m'(x) + f_X(x)f''_e(e)\{m'(x)\}^2 - f_X(x)f'_e(e)m''(x) \\ &\approx f_X(x)f''_e(e)\{m'(x)\}^2, \text{ by (H1) and (H3)}. \end{aligned} \tag{B.19}$$

Hence, (B.16) simplifies to

$$\mathrm{DB}_3(x,y,h_1,h_2) \approx 0.5f''_e(e)\left[\{m'(x)\}^2\mu_{2,1}h_1^2 + \mu_{2,2}h_2^2\right].$$

Lastly, to simplify $\mathrm{DB}_4(x,y,h_1,h_2)$, we need to look into the transform $\mathscr{T}_x\{I_k(\cdot,y)\}$, for $k = 2,3,4$. The only reason that it is more difficult to obtain closed-form expressions of these three transforms than the same transform

of $I_1(w,y) = \int p_{xx}(v,y) f_U(w-v)dv$ is the additional function of w (as derivatives of $m^*(w)$) outside of the integrals in (3.8). If $m^*(w)$ is approximately linear so that $(d/dw)m^*(w)$ is approximately a constant, then this only obstacle disappears. This additional assumption can be satisfied for some measurement error models given Condition (H3). With this assumption on $m^*(w)$, one immediately has $I_2(w,y) \approx 0$ according to (3.8) and thus $\mathscr{T}_x\{I_2(\cdot,y)\} \approx 0$. Following the same idea behind the derivation for $\mathscr{T}_x\{I_1(\cdot,y)\}$ and the proof for (B.9) in Section B.1, one can show that,

$$\begin{aligned}\mathscr{T}_x\{I_3(\cdot,y)\} &\approx \left\{\frac{d}{dx}m^*(x)\right\}^2 p_{yy}(x,y)\\ &= \left\{\frac{d}{dx}m^*(x)\right\}^2 f_X(x) f_e''(e), \text{ by (B.18)},\end{aligned}$$

and

$$\begin{aligned}\mathscr{T}_x\{I_4(\cdot,y)\} &\approx 2\left\{\frac{d}{dx}m^*(x)\right\} p_{xy}(x,y)\\ &= 2\left\{\frac{d}{dx}m^*(x)\right\}\frac{\partial}{\partial y}\left\{\frac{\partial}{\partial x}f_X(x)f_e(e)\right\}, \text{ by (H2)},\\ &= 2\left\{\frac{d}{dx}m^*(x)\right\}\frac{\partial}{\partial y}\left\{-f_X(x)f_e'(e)m'(x)\right\}, \text{ by (H1)},\\ &= -2\left\{\frac{d}{dx}m^*(x)\right\} f_X(x) f_e''(e) m'(x).\end{aligned}$$

Putting these results of $\mathscr{T}_x\{I_k(\cdot,y)\}$, for $k = 2,3,4$, along with (B.19) and (B.18), back in (B.17), one has

$$\mathrm{DB}_4(x,y,h_1,h_2) \approx 0.5 f_e''(e)\left[\left\{m'(x) - \frac{d}{dx}m^*(x)\right\}^2 \mu_{2,1}h_1^2 + \mu_{2,2}h_2^2\right].$$

In summary, under the aforementioned special case, we have

$$\begin{aligned}&\mathrm{DB}_2(x,y,h_1,h_2) \approx 0.5 f_e''(e)\mu_{2,2}h_2^2,\\ &\mathrm{DB}_3(x,y,h_1,h_2) \approx 0.5 f_e''(e)\left[\{m'(x)\}^2 \mu_{2,1}h_1^2 + \mu_{2,2}h_2^2\right],\\ &\mathrm{DB}_4(x,y,h_1,h_2) \approx 0.5 f_e''(e)\left[\left\{m'(x) - \frac{d}{dx}m^*(x)\right\}^2 \mu_{2,1}h_1^2 + \mu_{2,2}h_2^2\right].\end{aligned}$$

These three approximations imply the comparison between $\mathrm{DB}_3(x,y,h_1,h_2)$ and $\mathrm{DB}_2(x,y,h_1,h_2)$, and that between $\mathrm{DB}_4(x,y,h_1,h_2)$ and $\mathrm{DB}_3(x,y,h_1,h_2)$ summarized in Section 3.3.

Appendix C: Derivations of the CV criteria associated with $\tilde{p}_1(y|x)$ and $\tilde{p}_2(y|x)$

The cross validation (CV) criterion proposed by Fan and Yim (2004) and Hall et al. (2004) for choosing bandwidths in the estimator of $p^*(y|x)$, $\tilde{p}_1(y|x)$, is

given by (4.2), the integral in which can be derived explicitly as follows when $K_2(t)$ is the Gaussian kernel.

By the definition of $\tilde{p}_{1,-j}(y|W_j)$, one has

$$\begin{aligned}
&\int \{\tilde{p}_{1,-j}(y|W_j)\}^2\,dy \\
=&\int \left\{ \frac{\frac{1}{(n-1)h_1h_2}\sum_{j'\neq j} K_1\left(\frac{W_{j'}-W_j}{h_1}\right) K_2\left(\frac{Y_{j'}-y}{h_2}\right)}{\frac{1}{(n-1)h_1}\sum_{j'\neq j} K_1\left(\frac{W_{j'}-W_j}{h_1}\right)} \right\}^2 dy \\
=&\left\{\sum_{j'\neq j} K_1\left(\frac{W_{j'}-W_j}{h_1}\right)\right\}^{-2} \int \frac{1}{h_2^2}\sum_{j_1\neq j}\sum_{j_2\neq j}\left\{K_1\left(\frac{W_{j_1}-W_j}{h_1}\right)\times\right. \\
&\left. K_1\left(\frac{W_{j_2}-W_j}{h_1}\right) K_2\left(\frac{Y_{j_1}-y}{h_2}\right) K_2\left(\frac{Y_{j_2}-y}{h_2}\right)\right\} dy \\
=&\left\{\sum_{j'\neq j} K_1\left(\frac{W_{j'}-W_j}{h_1}\right)\right\}^{-2} \frac{1}{h_2^2}\sum_{j_1\neq j}\sum_{j_2\neq j}\left\{K_1\left(\frac{W_{j_1}-W_j}{h_1}\right)\times\right. \\
&\left. K_1\left(\frac{W_{j_2}-W_j}{h_1}\right) \int K_2\left(\frac{Y_{j_1}-y}{h_2}\right) K_2\left(\frac{Y_{j_2}-y}{h_2}\right) dy\right\},
\end{aligned}$$

in which

$$\begin{aligned}
&\int K_2\left(\frac{Y_{j_1}-y}{h_2}\right) K_2\left(\frac{Y_{j_2}-y}{h_2}\right) dy \\
=&h_2\int K_2(t) K_2\left(t-\frac{Y_{j_1}-Y_{j_2}}{h_2}\right) dt \\
=&h_2\int \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\cdot\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(t-\frac{Y_{j_1}-Y_{j_2}}{h_2}\right)^2\right\} dt \\
=&\frac{h_2}{\sqrt{4\pi}}\exp\left\{-\left(\frac{Y_{j_1}-Y_{j_2}}{2h_2}\right)^2\right\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\int \{\tilde{p}_{1,-j}(y|W_j)\}^2\,dy \\
=&\frac{\frac{1}{\sqrt{4\pi}h_2}\sum_{j_1\neq j}\sum_{j_2\neq j} K_1\left(\frac{W_{j_1}-W_j}{h_1}\right) K_1\left(\frac{W_{j_2}-W_j}{h_1}\right)\exp\left\{-\left(\frac{Y_{j_1}-Y_{j_2}}{2h_2}\right)^2\right\}}{\left\{\sum_{j'\neq j} K_1\left(\frac{W_{j'}-W_j}{h_1}\right)\right\}^2}.
\end{aligned}$$

Using this result in the first half of (4.2), and using the definition of $\tilde{p}_{1,-j}(y|W_j)$ in the second half of (4.2) leads to the following elaboration of (4.2),

$$\begin{aligned}
&\mathrm{CV}(\tilde{p}_1)\\
=&\frac{1}{nh_2}\sum_{j=1}^{n}\omega(W_j)\times\\
&\left[\frac{\frac{1}{\sqrt{4\pi}}\sum_{j_1\neq j}\sum_{j_2\neq j}K_1\left(\frac{W_{j_1}-W_j}{h_1}\right)K_1\left(\frac{W_{j_2}-W_j}{h_1}\right)\exp\left\{-\left(\frac{Y_{j_1}-Y_{j_2}}{2h_2}\right)^2\right\}}{\left\{\sum_{j'\neq j}K_1\left(\frac{W_{j'}-W_j}{h_1}\right)\right\}^2}\right.\\
&\left.-2\frac{\sum_{j'\neq j}^{n}K_1\left(\frac{W_{j'}-W_j}{h_1}\right)K_2\left(\frac{Y_{j'}-Y_j}{h_2}\right)}{\sum_{j'\neq j}K_1\left(\frac{W_{j'}-W_j}{h_1}\right)}\right].
\end{aligned}\tag{C.1}$$

Following similar derivations leading to (C.1), one can show that, with $K_2(t)$ being the Gaussian kernel, (4.5) becomes

$$\begin{aligned}
&\mathrm{CV}(\tilde{p}_2)\\
=&\frac{1}{nh_2}\sum_{j=1}^{n}\omega(W_j)\times\\
&\left[\frac{\frac{1}{\sqrt{4\pi}}\sum_{j_1\neq j}\sum_{j_2\neq j}K_1\left(\frac{W_{j_1}-W_j}{h_1}\right)K_1\left(\frac{W_{j_2}-W_j}{h_1}\right)\exp\left\{-\left(\frac{e^*_{j_1}-e^*_{j_2}}{2h_2}\right)^2\right\}}{\left\{\sum_{j'\neq j}K_1\left(\frac{W_{j'}-W_j}{h_1}\right)\right\}^2}\right.\\
&\left.-2\frac{\sum_{j'\neq j}^{n}K_1\left(\frac{W_{j'}-W_j}{h_1}\right)K_2\left(\frac{e^*_{j'}-e^*_j}{h_2}\right)}{\sum_{j'\neq j}K_1\left(\frac{W_{j'}-W_j}{h_1}\right)}\right],
\end{aligned}\tag{C.2}$$

where $e^*_j = Y_j - \hat{m}^*(W_j)$.

Appendix D: Boxplots of EISE associated with four density estimators when $m(x) \equiv 1$

We simplify the primary model setting (C1) in Section 5.1 in the main article to create the following primary model setting,

(C4) $[Y|X = x] \sim N\big(m(x), \sigma^2(x)\big)$, where $m(x) \equiv 1$ and $\sigma(x) = e^{1-x/3}/8$.

Along with the secondary model settings (a)–(d) stated in Section 5.1 in the main article, we now have four data generating processes, according to each of which data of the form $\{(W_j, Y_j)\}_{j=1}^{500}$ are generated independently 200 times. Figure D.1 provides boxplots of EISE associated with $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, and $\hat{p}_4(y|x)$ when the approximated theoretical optimal bandwidths are used, which suggest that all four estimators perform similarly. Figure D.2 shows the same boxplots when the fully data-driven bandwidths are used. From there one can see that the two non-naive estimators are more variable than their naive counterparts, but are otherwise comparable. Between the two non-naive estimators, $\hat{p}_3(y|x)$ is slightly less variable than $\hat{p}_4(y|x)$.

Appendix E: Boxplots of EISE under the simulation settings in Section 5 when cubic spline estimates of the mean function are used

Figures E.1–E.3 are counterpart plots of Figures 6–8 in the main article, where the cubic spline is used in $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$ to estimate $m^*(\cdot)$.

Appendix F: Estimated density curves using dietary data when cubic spline estimates of the mean function are used

Figure F.1 is a counterpart plot of Figure 11 in the main article, where dietary data are used to estimate $p(y|x)$, but with cubic spline estimate for the mean function in $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$.

Appendix G: Boxplots of EISE associated with four density estimators when σ_u^2 is correctly specified and when it is misspecified

In this experiment, we generate data following the primary model specified in (C1) and the secondary model configuration (a) described in Section 5.1, where the true measurement error variance is $\sigma_u^2 = 0.25$, corresponding to $\lambda = 0.8$. Based on each of 200 simulated data sets, each of size $n = 500$, besides computing $\tilde{p}_1(y|x)$ and $\tilde{p}_2(y|x)$, we compute $\hat{p}_3(y|x)$ and $\hat{p}_4(y|x)$ while assuming σ_u^2 at its truth and three other misspecified values corresponding to $\lambda = 0.7, 0.9, 0.99$. All bandwidths are chosen using the data-driven methods described in Section 4.

Figure G.1 contains boxplots of EISE across 200 Monte Carlo replicates at each assumed level of σ_u^2, i.e., at each assumed level of $\lambda(= 0.8, 0.7, 0.9, 0.99)$. When comparing with the case without misspecifying the value of σ_u^2 (in panel

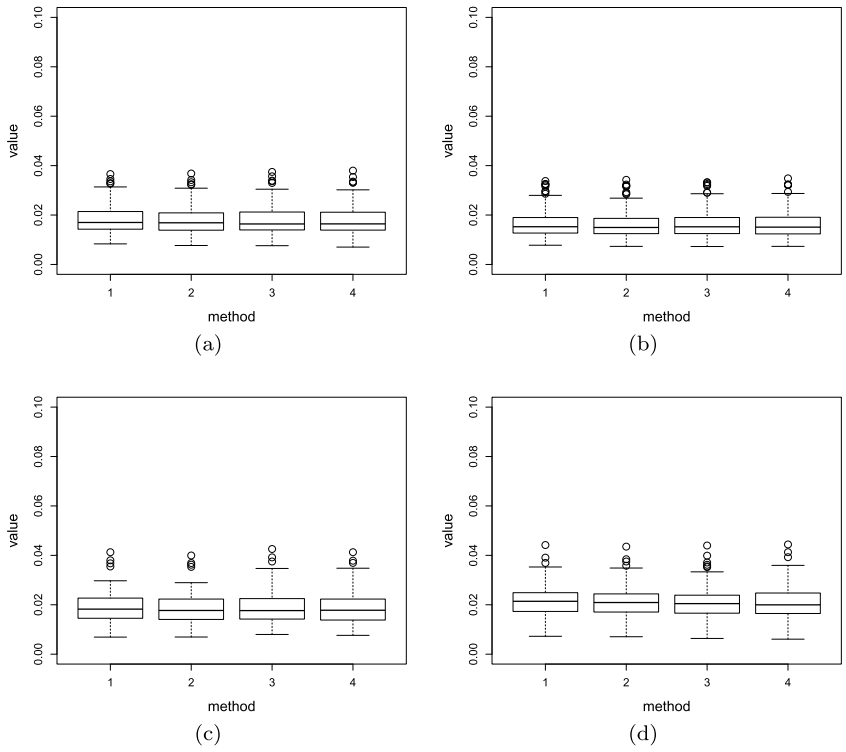

FIG D.1. *Boxplots of EISE using the approximated theoretical optimal bandwidths when the primary model is (C4) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim$ *Laplace*$(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim$ *Laplace*$(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim$ *Uniform*$(-2,2)$, $U \sim$ *Laplace*$(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$. *Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

(a)), one can see that even when one sets σ_u^2 at a higher level than the truth (in panel (b)) or at a lower level (in panel (c)), each proposed non-naive estimator, $\hat{p}_3(y|x)$ or $\hat{p}_4(y|x)$, still outperforms its naive counterpart, that is, $\tilde{p}_1(y|x)$ or $\tilde{p}_2(y|x)$, in the sense that the median EISE associated with $\hat{p}_3(y|x)$ or $\hat{p}_4(y|x)$ is still smaller than that of $\tilde{p}_1(y|x)$ or $\tilde{p}_2(y|x)$. The variabilities of the two proposed estimators with an assumed σ_u^2 value much larger than the truth, as the case in panel (b), are higher compared to when one uses the correct σ_u^2 or sets it at a smaller value than the truth. This can be caused by, besides involving a wrong $\phi_U(t)$ in the proposed estimators, one uses a larger bandwidth h_1 by setting $\hat{\lambda}$ in (4.4) and (4.7) at some smaller value than what one would use had one used the true value of σ_u^2.

Certainly, if one assumes a low enough value for σ_u^2 such that it is close to assuming no measurement error, as in panel (d) of Figure G.1, then all four estimates behave similarly. Table G.1 presents medians and IQRs corresponding to the EISEs depicted in Figure G.1.

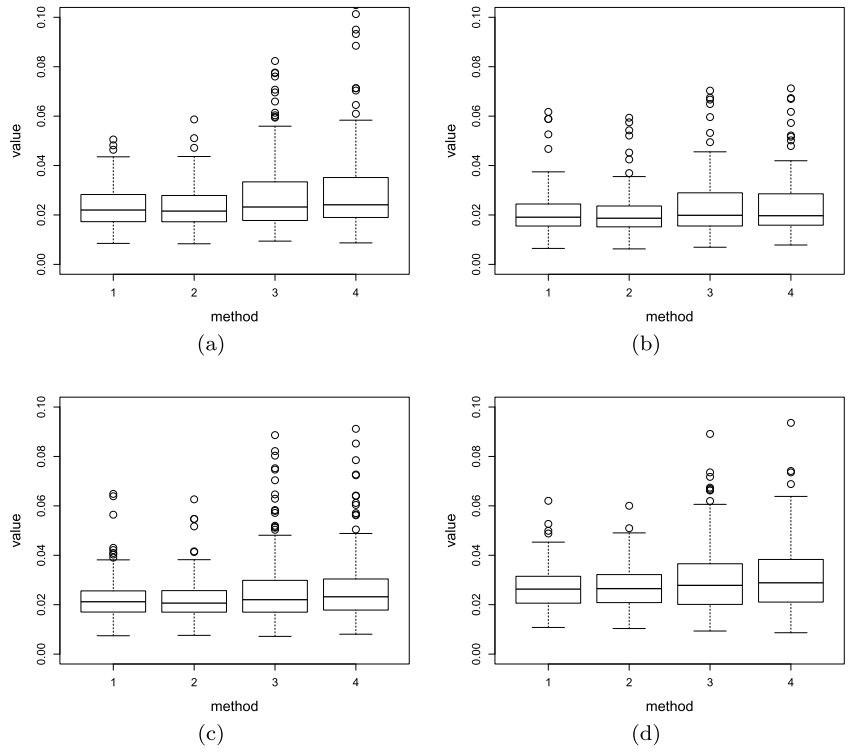

FIG D.2. *Boxplots of EISE using the fully data-driven bandwidths when the primary model is (C4) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim$ *Laplace*$(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim$ *Laplace*$(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim$ *Uniform*$(-2,2)$, $U \sim$ *Laplace*$(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

Appendix H: An example R code for estimating conditional densities using the R package `lpme`

For illustration purposes, we generate a data set of size $n = 1000$ under the primary model configuration (C3) specified in the main article, with $X \sim N(0,1)$, $U \sim \text{Laplace}(0, \sigma_u/\sqrt{2})$, and $\lambda = 0.8$. The following code is used to generate data.

```
## X - True covariates
## W - Observed covariates
## Y - Individual response
rm(list=ls())
library(lpme)
## Generate Laplace random numbers
rlap = function (use.n, location = 0, scale = 1)
{
location <- rep(location, length.out = use.n)
scale <- rep(scale, length.out = use.n)
```

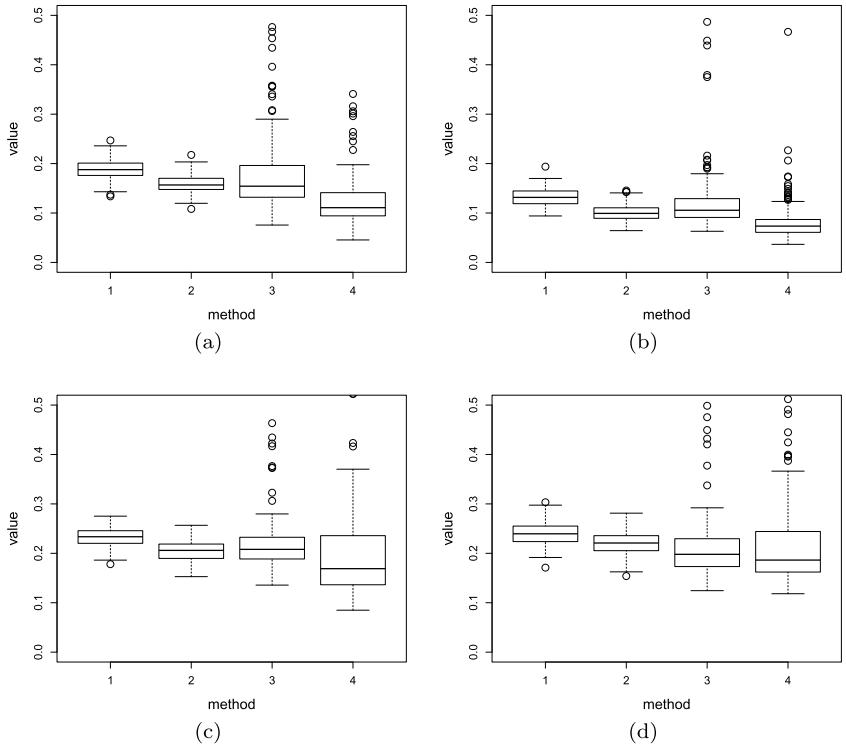

FIG E.1. *Boxplots of EISE using the fully data-driven bandwidths with the cubic spline mean estimate when the primary model is (C1) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$. *Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

```
rrrr <- runif(use.n)
location - sign(rrrr - 0.5) * scale *
(log(2) + ifelse(rrrr < 0.5, log(rrrr), log1p(-rrrr)))
}
## Function f(y|x) to be estimated
mofx = function(x){ x }
sofx = function(x){ exp(1-x/3)/8 }
wide = 0.04; ymin=-4; ymax=3;
x = seq(-2, 2, wide);
y = seq(-4, 3, wide);
nx = length(x)
ny = length(y);
## True density function
fy_x=function(y,x) dnorm(y, mofx(x), sofx(x));

####################  Generate data ################
set.seed(2017)
n = 1000 ## sample size:
sigma_x = 1; X = rnorm(n, 0, sigma_x);
```

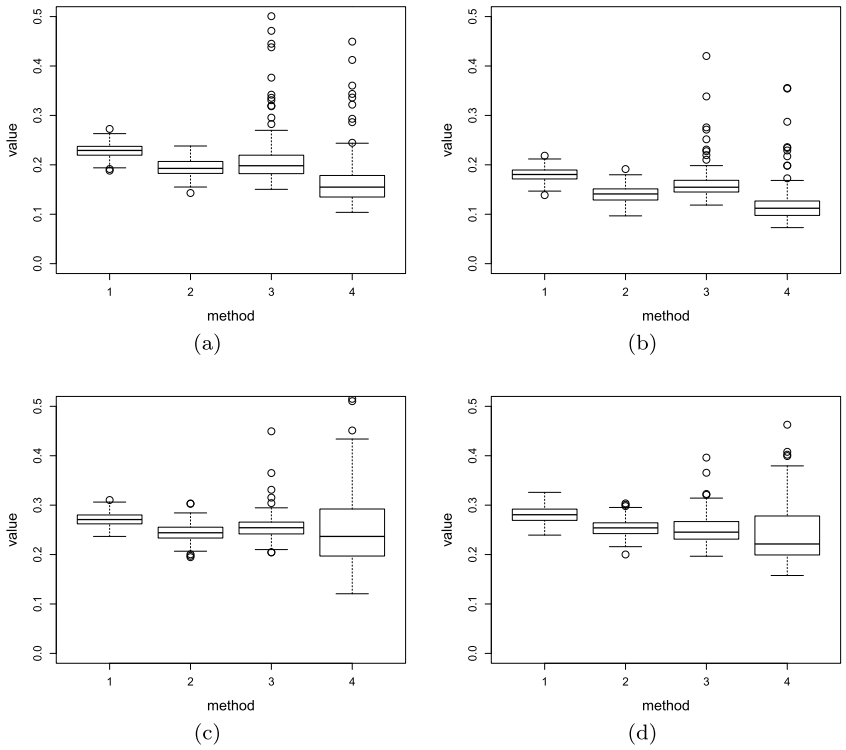

FIG E.2. *Boxplots of EISE using the fully data-driven bandwidths with the cubic spline mean estimate when the primary model is (C2) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

```
Y = rep(0, n);
for(i in 1:n){
Y[i] = mofx(X[i]) + rnorm(1, 0, sofx(X[i]));
}
## reliability ratio
lambda = 0.8;
sigma_u  = sqrt(1/lambda-1)*sigma_x;
W = X + rlap(n, 0, sigma_u/sqrt(2));
```

Panel (d) in Figure H.1 shows the scatter plot of the response Y versus the covariate X, with the corresponding realizations of W imposed. The following code is used to obtain the conditional density estimates, $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$ and $\hat{p}_4(y|x)$, at grid points `x` and `y` defined in above code. Panels (a)–(c) in Figure H.1 depict these four density estimates when $x = -1.5, 0, 1.5$, respectively.

```
##----- Method 1: naive estimate without mean adjustment  ----
## kernel functions
K1 = "Gauss"; const1 = 1.06;
```

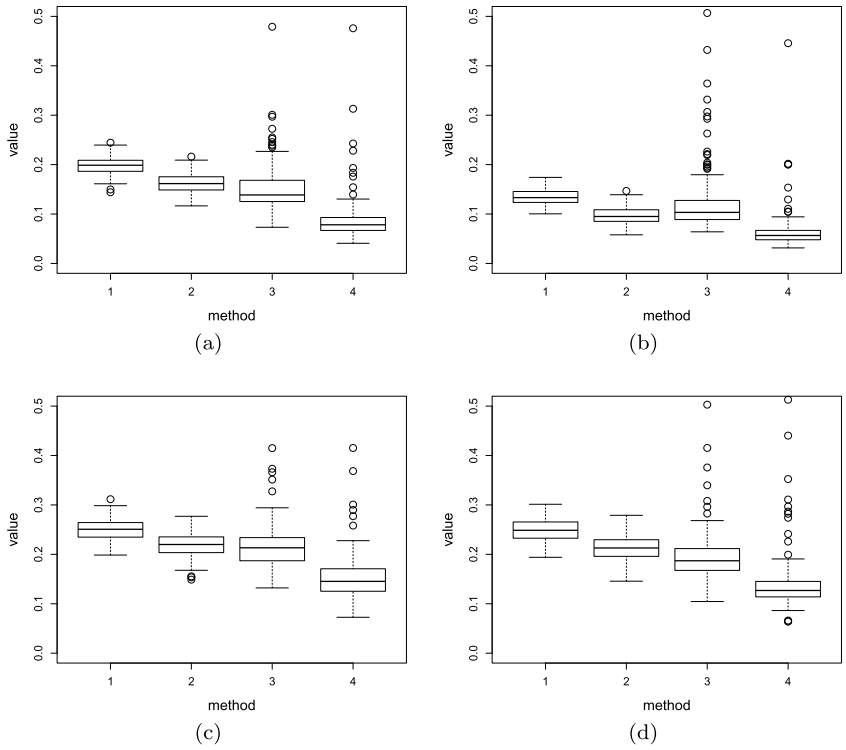

FIG E.3. *Boxplots of EISE using the fully data-driven bandwidths with the cubic spline mean estimate when the primary model is (C3) and the secondary models are (a)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$*; (b)* $X \sim N(0,1)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.9$*; (c)* $X \sim N(0,1)$, $U \sim N(0, \sigma_u^2)$, $\lambda = 0.8$*; (d)* $X \sim Uniform(-2,2)$, $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$. *Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively.*

```
K2 = "Gauss"; const2 = 1.06;
## initial reference rule
hxyhat = c(sd(W)*const1, sd(Y)*const2)*n^(-1/5);
## grid points for searching bandwidths
h1 = hxyhat[1]*seq(0.2, 1.5, length.out = 10 )
h2 = hxyhat[2]*seq(0.2, 1.5, length.out = 10 )
ptm<-proc.time()
fitbw1 = densityregbw(Y, W, xinterval = c(min(x), max(x)), h1 = h1, h2 = h2,
                      K1 = K1, K2 = K2)
systime1=proc.time()-ptm; systime1;
ptm<-proc.time()
fhat1 = densityreg(Y, W, bw = fitbw1$bw, xgrid = x, ygrid = y,
                   K1 = K1, K2 = K2);
systime11=proc.time()-ptm; systime11;
##----- Method 2: naive estimate with mean adjustment  ----
## kernel functions
K1 = "Gauss"; const1 = 1.06;
K2 = "Gauss"; const2 = 1.06;
## initial reference rule
```

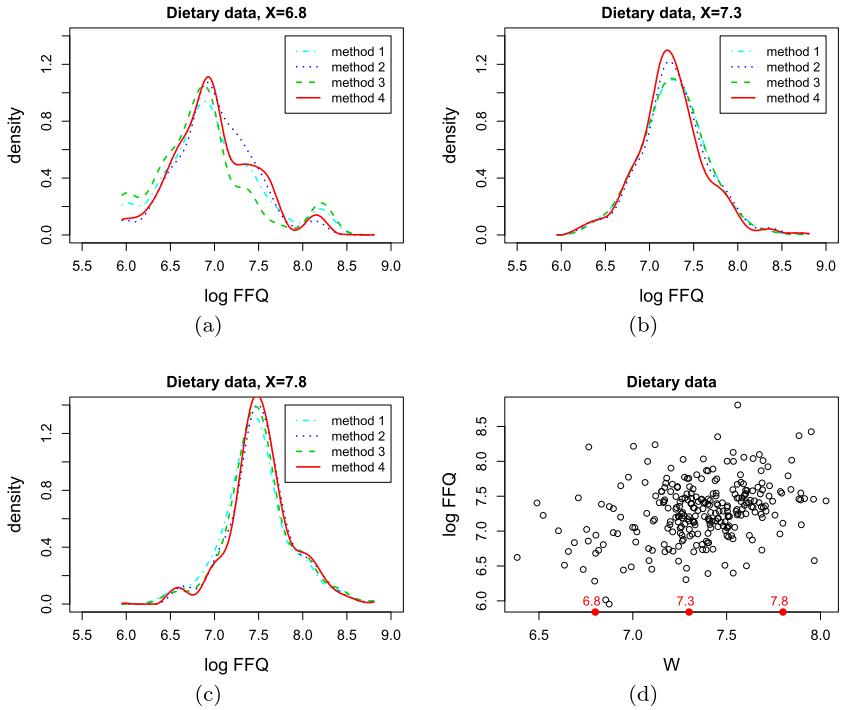

FIG F.1. *Naive estimates of conditional density of the logarithm of FFQ intake corresponding to $\tilde{p}_1(y|x)$ (cyan dash-dotted lines) and $\tilde{p}_2(y|x)$ (blue dotted lines), and two non-naive density estimates, $\hat{p}_3(y|x)$ (green dashed lines) and $\hat{p}_4(y|x)$ (red solid lines) when $x = 6.8$ (in panel (a)), 7.3 (in panel (b)), and 7.8 (in panel (c)), respectively. In each panel of (a)–(c), method 1, 2, 3, 4 correspond to $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, and $\hat{p}_4(y|x)$, respectively. The cubic spline estimate of the mean function is used in $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$. The scatter plot of the observed response versus the observed covariate values is shown in panel (d), where the three values of x at which $p(y|x)$ is estimated are highlighted in red dots on the horizontal axis.*

```
hxyhat = c(sd(W)*const1, sd(Y)*const2)*n^(-1/5);
## grid points for searching bandwidths
h1 = hxyhat[1]*seq(0.5, 3, length.out = 10 )
h2 = hxyhat[2]*seq(0.2, 1.5, length.out = 10 )
ptm<-proc.time()
fitbw2 = densityregbw(Y, W, xinterval = c(min(x), max(x)), h1 = h1, h2 = h2,
                      K1 = K1, K2 = K2, mean.estimate = "kernel")
systime2=proc.time()-ptm; systime2;
ptm<-proc.time()
fhat2 = densityreg(Y, W, bw = fitbw2$bw, xgrid = x, ygrid = y,
                   K1 = K1, K2 = K2, mean.estimate = "kernel");
systime22=proc.time()-ptm; systime22;

##----- Method 3: proposed method without mean adjustment  ----
## kernel functions
K1 = "SecOrder"; const1 = 0.427398;
K2 = "Gauss"; const2 = 1.06;
```

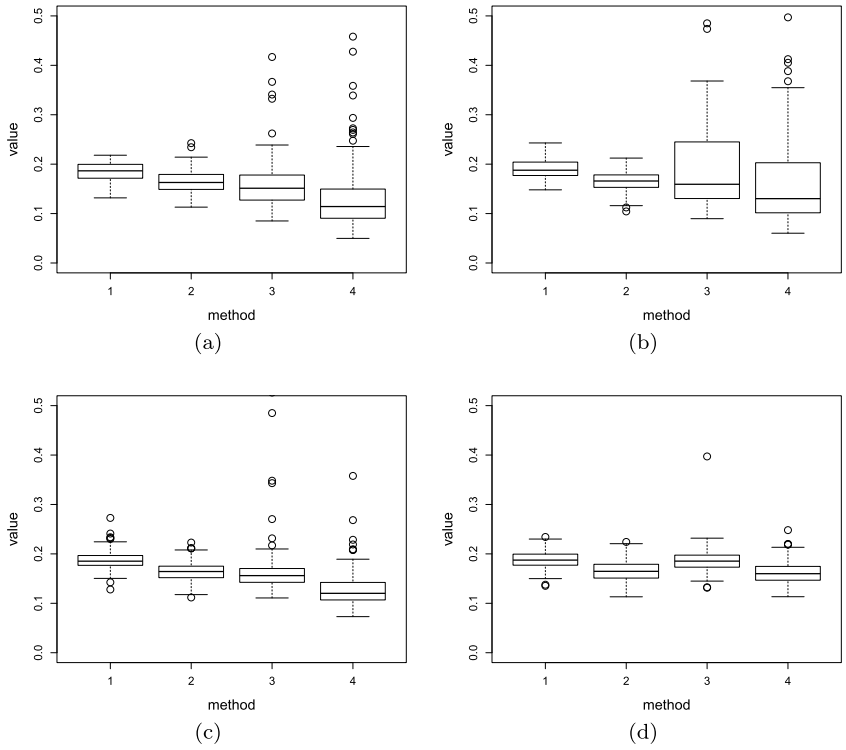

FIG G.1. *Boxplots of EISE using the fully data-driven bandwidths when the primary model is (C1), the secondary model is* $X \sim N(0,1)$ *and* $U \sim Laplace(0, \sigma_u/\sqrt{2})$, $\lambda = 0.8$. *Panel (a) presents the results when true* σ_u^2 *is used. Panels (b), (c) and (d) present the results when one misspecifies* σ_u^2 *such that the reliability* λ *is assumed to be 0.7, 0.9 and 0.99, respectively. Method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$, $\tilde{p}_2(y|x)$, $\hat{p}_3(y|x)$, *and* $\hat{p}_4(y|x)$, *respectively. The sample size is* $n = 500$.

```
## initial reference rule
hxyhat = c(sd(W)*const1, sd(Y)*const2)*n^(-1/5);
## grid points for searching bandwidths
h1 = hxyhat[1]*seq(0.2, 1.5, length.out = 10 )
h2 = hxyhat[2]*seq(0.2, 1.5, length.out = 10 )
ptm<-proc.time()
fitbw3 = densityregbw(Y, W, xinterval = c(min(x), max(x)), sig = sigma_u,
                      h1 = h1, h2 = h2, K1 = K1, K2 = K2)
systime3=proc.time()-ptm; systime3;
ptm<-proc.time()
fhat3 = densityreg(Y, W, bw = fitbw3$bw, xgrid = x, ygrid = y, sig = sigma_u,
                   K1 = K1, K2 = K2);
systime33=proc.time()-ptm; systime33;

##----- Method 4: proposed method wit mean adjustment  ----
## kernel functions
K1 = "SecOrder"; const1 = 0.427398;
K2 = "SecOrder"; const2 = 0.427398;
```

TABLE G.1
Median and IQR (in parenthesis) of the EISE associated with each of the four considered estimators using the fully data-driven bandwidths under (C1) with σ_u^2 correctly specified (corresponding to panel (a) in Figure G.1) and misspecified (corresponding to panels (b)–(d) in Figure G.1)

Method	(a)	(b)	(c)	(d)
1	0.186 (0.028)	0.188 (0.027)	0.185 (0.020)	0.187 (0.023)
2	0.163 (0.030)	0.167 (0.026)	0.164 (0.023)	0.165 (0.028)
3	0.151 (0.049)	0.159 (0.114)	0.156 (0.028)	0.185 (0.023)
4	0.114 (0.059)	0.130 (0.100)	0.120 (0.035)	0.160 (0.028)

```
## initial reference rule
hxyhat = c(sd(W)*const1, sd(Y)*const2)*n^(-1/5);
## grid points for searching bandwidths
h1 = hxyhat[1]*seq(0.5, 3, length.out = 10 )
h2 = hxyhat[2]*seq(0.2, 1.5, length.out = 10 )
ptm<-proc.time()
fitbw4 = densityregbw(Y, W, xinterval = c(min(x), max(x)), sig = sigma_u,
                      h1 = h1, h2 = h2, K1 = K1, K2 = K2, mean.estimate = "kernel")
systime4=proc.time()-ptm; systime4;
ptm<-proc.time()
fhat4 = densityreg(Y, W, bw = fitbw4$bw, xgrid = x, ygrid = y, sig = sigma_u,
                   K1 = K1, K2 = K2, mean.estimate = "kernel");
systime44=proc.time()-ptm; systime44;
```

The function `densityregbw` in the R package `lpme` (Zhou and Huang, 2017) is used for bandwidths selection. We explain five arguments in this function next.

(i) The argument `sig` allows one to specify the standard deviation of the measurement error. Its default value is `NULL`, suggesting that one assumes no measurement error. In the above code, letting `sig = NULL` or leaving it unspecified leads to the naive estimates, $\tilde{p}_1(y|x)$ and $\tilde{p}_2(y|x)$; and we set `sig = sigma_u` with a pre-defined valeue for `sigma_u` to obtain the non-naive estimates, $\hat{p}_3(y|x)$ and $\hat{p}_4(y|x)$.

(ii) The argument `mean.estimate` is where one specifies the type of estimates for the mean function $m^*(\cdot)$. If left unspecified, it takes the default value of `NULL`, corresponding to the density estimation methods that do not require estimating the mean function. This is value for this argument when computing $\tilde{p}_1(y|x)$ and $\hat{p}_3(y|x)$ in the above code. To compute $\tilde{p}_2(y|x)$ and $\hat{p}_4(y|x)$ in the example code, we set `mean.estimate = ''kernel''` to use the local linear estimate for the mean function. For these two density estimates, one may set `mean.estimate = ''spline''` to estimate the mean function using spline-based estimates, and use the argument `spline.df` to specify the order of the spline. The default value of `spline.df` is 5.

(iii) The arguments `K1` and `K2` correspond to the kernel functions $K_1(t)$ and $K_2(t)$ used in the main article. Choices for each one include the Gaussian kernel, `''Gauss''`, and the second order kernel, `''SecOrder''`, defined in (4.3) in the main article. In the current version, not all the combinations

are supported, and one will receive an error message if one chooses a combination of `K1` and `K2` that is not supported.

(iv) The arguments `h1` and `h2` are used to specify the searching grid points for bandwidths h_1 and h_2. When unspecified, bandwidths selected based on reference rules are used.

(v) The argument `xinterval` is used to specify the values x_L and x_U in the main article.

The function `densityregbw` returns an object with three variables, `bw` (selected bandwidths), `h1` (searched grid points for h_1), and `h2` (searched grid points for h_2).

The function `densityreg` is used for density estimation. Some arguments in this function are the same as those used in `densityregbw`. Two additional arguments in this function are `xgrid` and `ygrid`, which are used to specify the grid points for x and y in estimating $p(y|x)$, respectively. The function `densityreg` returns an object with three variables, `fitxy` (a matrix of fitted values with rows corresponding to x values), `xgrid` (grid points for x), and `ygrid` (grid points for y).

Acknowledgments

We are grateful to the Associate Editor and referee for their constructive comments and suggestions on an earlier version of the manuscript. The first author would also like to thank Professor David W. Scott at Rice University, for insightful discussions with her during the early stage of this research project.

References

Billingsley, P. (2008). *Probability and measure*. John Wiley & Sons. MR0534323

Buzas, J. S., Stefanski, L. A., and Tosteson, T. D. (2014). Measurement error. *Handbook of epidemiology*, pages 1241–1282.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement error in nonlinear models: a modern perspective*, volume 105. Chapman & Hall/CRC. MR2243417

Carroll, R. J. (2014). Measurement error in epidemiologic studies. *Wiley StatsRef: Statistics Reference Online*.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186. MR0997599

Cook, J. and Stefanski, L. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328. MR1379467

Delaigle, A. (2008). An alternative view of the deconvolution problem. *Statistica Sinica*, pages 1025–1045. MR2440402

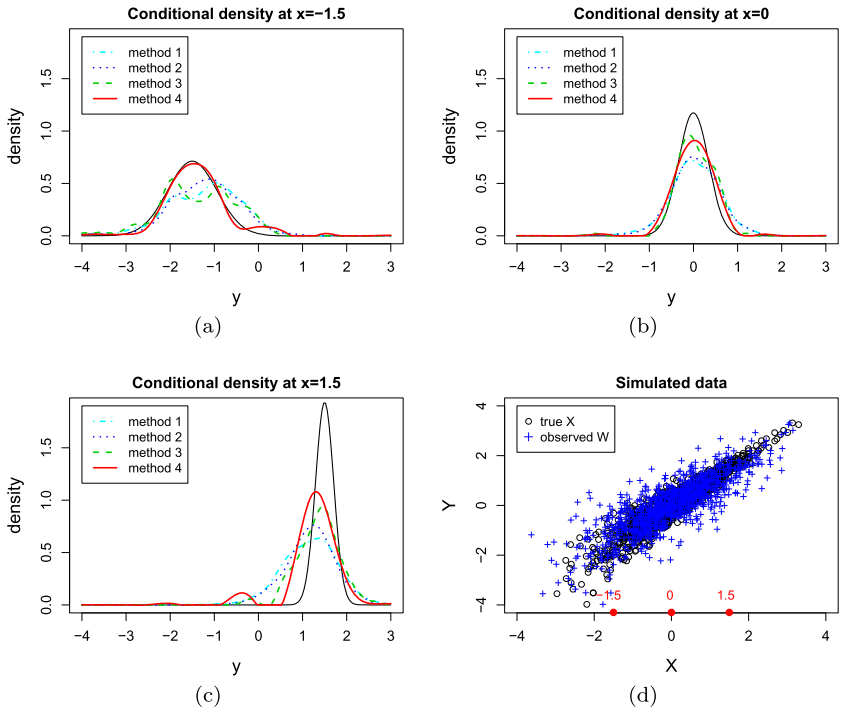

FIG H.1. *Estimated conditional density curves in panels (a)–(c) obtained from the example R code, with simulated data shown in panel (d). In each panel of (a)–(c), method 1, 2, 3, 4 correspond to* $\tilde{p}_1(y|x)$ *(cyan dash-dotted lines),* $\tilde{p}_2(y|x)$ *(blue dotted lines),* $\hat{p}_3(y|x)$ *(green dashed lines), and* $\hat{p}_4(y|x)$ *(red solid lines), respectively.*

Delaigle, A., Fan, J., and Carroll, R. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association*, 104(485):348–359. MR2504382

Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):869–886. MR1979392

Delaigle, A. and Gijbels, I. (2004a). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Annals of the Institute of Statistical Mathematics*, 56(1):19–47. MR2053727

Delaigle, A. and Gijbels, I. (2004b). Practical bandwidth selection in deconvolution kernel density estimation. *Computational statistics & data analysis*, 45(2):249–267. MR2045631

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *Journal of the American Statistical Association*, 103(481):280–287. MR2394636

Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *The Annals of Statistics*, pages 665–685. MR2396811

Fan, J. (1991a). Asymptotic normality for deconvolution kernel density estimators. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 97–110. MR1177770

Fan, J. (1991b). Global behavior of deconvolution kernel estimates. *Statistica Sinica*, pages 541–551. MR1130132

Fan, J. (1991c). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272. MR1126324

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, volume 66. Chapman & Hall/CRC. MR1383587

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206. MR1399164

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834. MR2126035

Fuller, W. A. (2009). *Measurement error models*, volume 305. John Wiley & Sons. MR2301581

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026. MR2109491

Hansen, B. E. (2004). Nonparametric conditional density estimation. https://www.ssc.wisc.edu/~bhansen/papers/ncde.pdf.

Huang, X. and Zhou, H. (2017). An alternative local polynomial estimator for the error-in-variables problem. *Journal of Nonparametric Statistics*, 29(2):301–325. MR3635015

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336. MR1422114

Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Journal of nonparametric statistics*, 14(3):259–278. MR1905751

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407. MR1394097

Liang, H. and Wang, N. (2005). Partially linear single-index measurement error models. *Statistica Sinica*, pages 99–116. MR2125722

Masry, E. (1993). Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stochastic processes and their applications*, 47(1):53–74. MR1232852

Meister, A. (2004). On the effect of misspecifying the error density in a deconvolution problem. *Canadian Journal of Statistics*, 32(4):439–449. MR2125855

Robins, J. M., Hsieh, F., and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 409–424. MR1323347

Rosenblatt, M. (1969). Conditional probability density and regression estima-

tors. *Multivariate analysis II*, 25:31. MR0254987

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons. MR3329609

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall. MR0848134

Stefanski, L. A. and Carroll, R. J. (1990). Deconvolving kernel density estimators. *Statistics: A Journal of Theoretical and Applied Statistics*, 21(2):169–184. MR1054861

Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90(432):1247–1256. MR1379467

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., and Okanohara, D. (2010). Conditional density estimation via least-squares density ratio estimation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 781–788. MR2895762

Wang, H. J., Stefanski, L. A., and Zhu, Z. (2012). Corrected-loss estimation for quantile regression with covariate measurement errors. *Biometrika*, 99(2):405–421. MR2931262

Wang, N., Carroll, R., and Liang, K.-Y. (1996). Quasilikelihood estimation in measurement error models with correlated replicates. *Biometrics*, pages 401–411.

Zhou, H. and Huang, X. (2016). Nonparametric modal regression in the presence of measurement error. *Electronic Journal of Statistics*, 10(2):3579–3620. MR3575565

Zhou, H. and Huang, X. (2017). *lpme: Nonparametric Estimation of Measurement Error Models*. Version 1.1.1.