Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

A spline-assisted semiparametric approach to nonparametric measurement error models

Fei Jiang[a,*], Yanyuan Ma[b], Raymond J. Carroll[c]

[a] *Department of Biostatistics, University of California, San Francisco, United States*
[b] *Department of Statistics, Pennsylvania State University, United States*
[c] *Department of Statistics, Texas A&M University, United States*

ARTICLE INFO

Article history:
Received 15 March 2022
Revised 22 May 2023
Accepted 23 May 2023
Available online xxx

Keywords:
Deconvolution
Errors in covariates
Measurement error
Semiparametrics
Spline

ABSTRACT

A spline-assisted approach is proposed to handle the measurement error problem in treating the pollution and asthma data. It is well known that the minimax rate of convergence in nonparametric regression function estimation of a random variable measured with error is much slower than the rate in the error free case. A different problem is considered. It is shown that if one is willing to impose a relatively mild assumption in requiring that the error-prone variable has a compact support, then standard nonparametric results are obtainable for measurement error models. New and constructive methods to take full advantage of the compact support assumption via spline-assisted semiparametric methods are proposed. It is proven that the new estimator achieves the usual nonparametric rate as if there were no measurement error. Furthermore it is shown that similar spline approach can be used to assess the probability density function on a compact set, using observations with error, and the resulting estimator differs from the true density function only by a constant scale. In addition, it retains the nonparametric density convergence properties of the error free case. The performance of the new methods is demonstrated through simulations and the methods are implemented to analyze the relation between asthma and pollution.

1. Introduction

Conflicting conclusions from different data analyses contribute to controversies in environmental studies. One typical source of the potential problems is measurement error in the risk factors (Biggs et al., 2009), which, if ignored, may induce substantial estimation biases (Carroll et al., 2006). This problem occurs in an asthma study in Beijing where the effect of the pattern matter measure PM2.5 on the mean number of daily asthma emergency room visits (asthma ERV rate) is of interest. We carry out the analyses based on the PM2.5 concentrations obtained from both the Beijing Environmental Protection Bureau (BEPB) and the “Mission China Beijing” web site (Mission-China, 2016) maintained by the U.S. Department of State. Figure 1 shows that the asthma ERV rate decreases with the increase of PM2.5 concentrations based on the BEPB data,

* Corresponding author.:
E-mail addresses: Fei.Jiang@ucsf.edu (F. Jiang), yzm63@psu.edu (Y. Ma), carroll@stat.tamu.edu (R.J. Carroll).

https://doi.org/10.1016/j.ecosta.2023.05.001

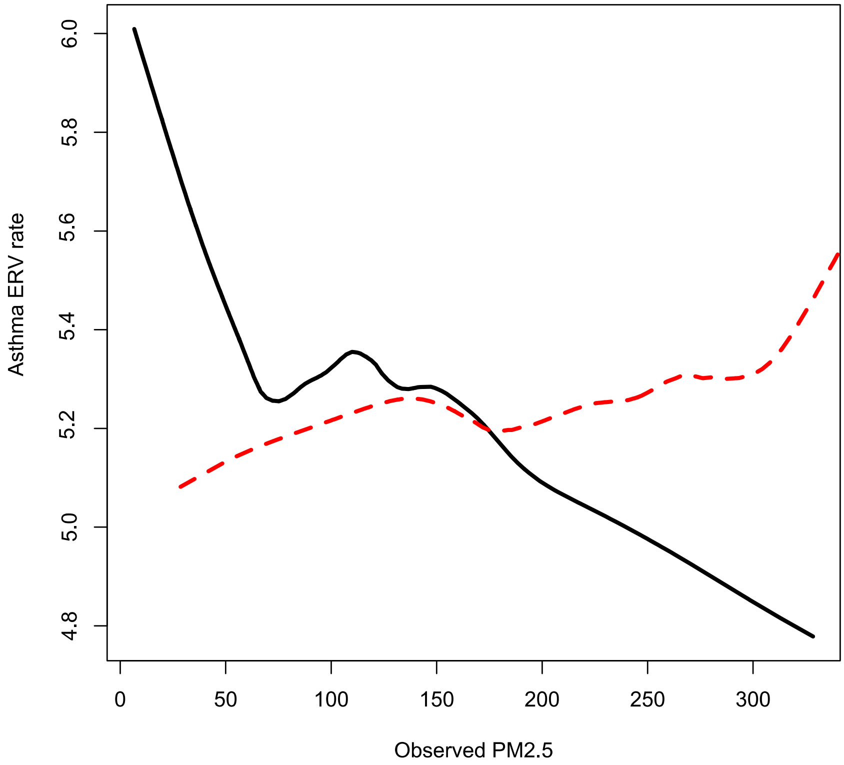

Fig. 1. Local linear regression curves for the asthma ERV rate as a function of the PM2.5 measurements from the Beijing Environmental Protection Bureau (solid line) and the *Mission China Beijing* web site (dashed line)

while it increases based on the "Mission China Beijing" data. It is worth mentioning that the PM2.5 data from BEPB is an average of the reads from seventeen separate monitors. The variation among the PM2.5 reads contributes to measurement errors in statistical terms, if we treat the overall mean concentration as the underlying true PM2.5 level. Therefore, we further consider a nonparametric measurement error model, where estimation is implemented via classical deconvolution methods (Carroll and Hall, 1988; Liu and Taylor, 1989; Stefanski and Carroll, 1990; Zhang, 1990; Fan, 1991). Unfortunately, the deconvolution method did not lead to any conclusive results, see the lower-right panel of Figure 2). We suspect that this is because of the slow convergence rate of the deconvolution method, which is well known in the literature. Hence, in order to extract more information from the data, novel methods with faster estimation convergence rates are desired. Of course, the distribution of the mean PM2.5 levels is important in itself.

These considerations lead us to study two nonparametric estimation problems under errors in covariates. Specifically, let X_i be the error free covariate. Assume that instead of observing X_i, we observe $W_i \equiv X_i + U_i$, where U_i is a mean zero random error independent of X_i and follows a distribution with density function, hereafter pdf, $f_U(u)$. Our main goal is to study the association between the covariate X_i, such as the mean PM2.5 level, and the response Y_i, such as the asthma ERV rate. Consider the nonparametric regression model $Y_i = m(X_i) + \epsilon_i$, where $m(\cdot)$ is an unknown function and ϵ_i is independent of X_i and has mean zero with density $f_\epsilon(\epsilon)$. This is the problem of nonparametric regression with measurement errors. It is well established (Fan and Truong, 1993) that, very slow convergence rates can occur and these rates are minimax. We found that if we assume the support of $f_{X0}(x)$ is contained in a compact set, then we can achieve the same nonparametric convergence rate as in the measurement error free case, which is a new result in the literature. This assumption is certainly true for the PM2.5 values, which are commonly measured on a scale of 0–500. Our second goal is to estimate the density function of X_i when only W_i's are observed. This problem has been studied extensively in the literature (Carroll and Hall, 1988; Liu and Taylor, 1989; Stefanski and Carroll, 1990; Zhang, 1990; Fan, 1991) and it is also well known that the estimator of $f_X(x)$ may converge very slowly. These slow convergence rates are also minimax for general X, and hence there is no hope to improve it without imposing additional constraints on the distribution of X. We show that if we are interested in estimating the density of X on a compact set only, and we ignore a constant scaling, then we can gain much improvement and can achieve the same rate as when there is no measurement error. This is certainly a practically relevant compromise, because most time we are only interested in the density in a finite domain, and a constant multiplier does not change the shape of the density function. This is certainly the case for the PM2.5 levels. The faster convergence rates in both regression and density estimation lead to meaningful results in the pollution-asthma data analysis as shown in Section 5.

The crux of our method is the usage of spline approximation and the semiparametric treatments. Spline representation in measurement error models has been mostly used in the Bayesian framework (Berry et al., 2002; Staudenmayer et al., 2008;

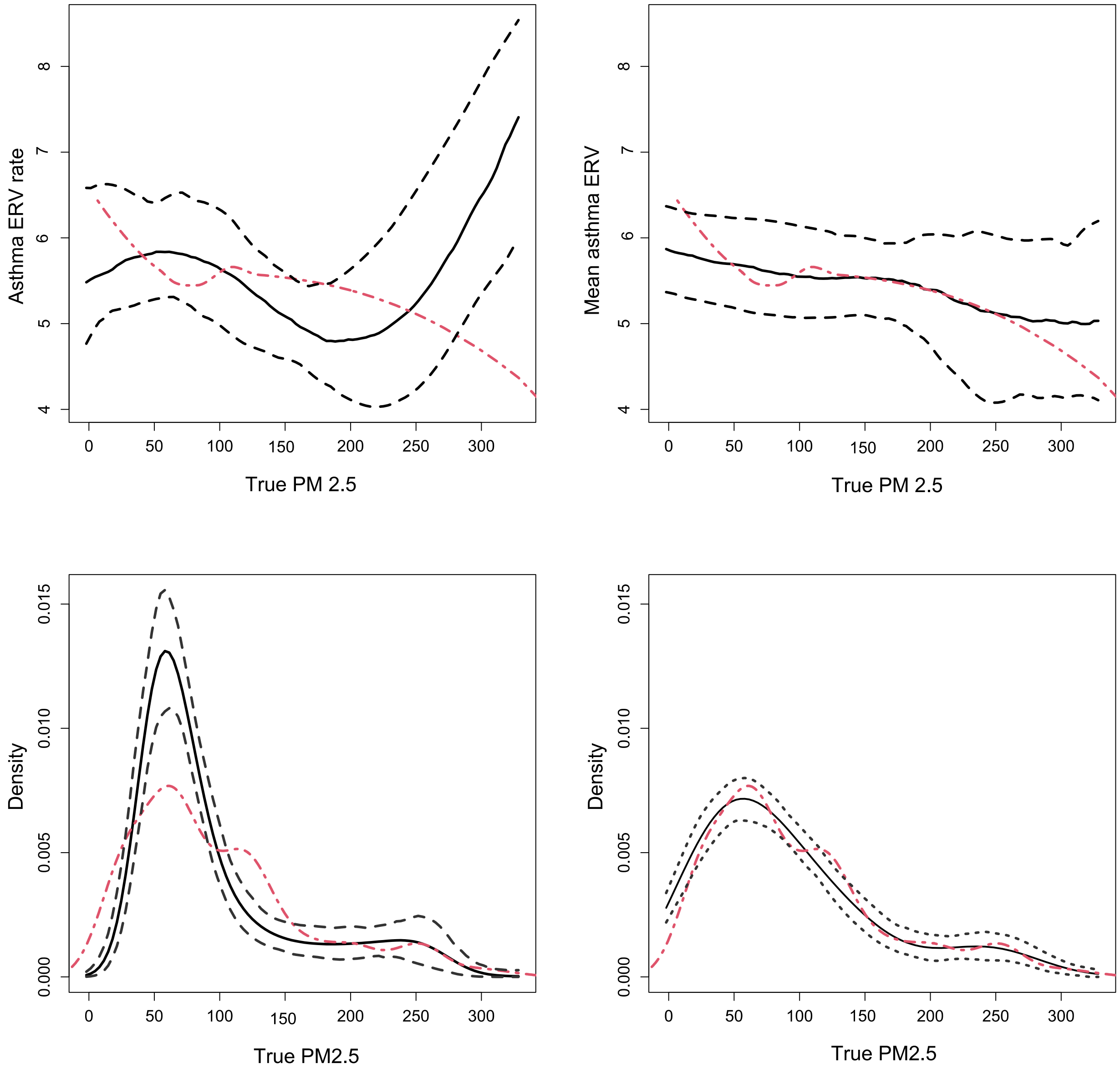

Fig. 2. B-spline semiparametric (upper left) and deconvolution (upper right) regression estimators and their 95% confidence bands. The B-spline MLE (lower left) and deconvolution (lower right) pdf estimators and their 95% confidence bands of PM2.5. The (black) solid lines are the estimators, the dashed lines are the 95% confidence intervals, and the (red) dot-dashed lines are the naive estimator ignoring measurement errors.

Sarkar et al., 2014). To the best of our knowledge, this is the first time it is used in combination with semiparametric treatment, which is indispensable for solving the regression problem in our view. This is the first main contribution of our work. A second contribution is in establishing the nonparametric convergence rate of the resulting estimators and their asymptotic normality, which are not yet available in the literature as far as we know and are important for statistical inference. In achieving these results, we overcome substantial statistical and mathematical difficulties because not only the estimator, but also the estimating equation that generates the estimator, do not have closed forms. The underlying reason that leads to the better convergence rate than the existing literature is the compact support assumption, which enables the usage of B-spline approximation and the subsequent semiparametric treatment.

A key difference between our procedures and the deconvolution procedures is that we restrict our interest in estimating the functions on a compact set while the deconvolution method aims at these functions on the whole real line. In this sense, these two methods are not really comparable. Practically, the possible range of a random variable is often finite, and the relevant information is needed only for functions in a range, so we consider the assumption of the support being compact to be mild. We are very curious if deconvolution methods can achieve the same convergence property if restricted on a finite domain, with possibly some modifications on the existing procedures. To this end, Hall and Qiu (2005) provides some relevant results based on a discrete Fourier transform and its inverse, while we leave this general question as an open problem for researchers who specialize in deconvolution methods.

In the following, we construct the estimation procedures for both the regression mean function and the probability density function in Section 2, and summarize the theoretical properties of our estimators in Section 3. We provide simulation studies to demonstrate the properties of the new estimators in Section 4, and illustrate the methods to study the relation between PM2.5 and asthma ERV rate in Section 5. The paper is concluded with a discussion in Section 6.

2. B-spline-assisted semiparametric estimation procedures

2.1. Regression mean function estimation

To set the notation, we use $f_X(x)$ to denote a generic pdf of the random variable X, and use $f_{X0}(x)$ to denote the true pdf that generates the data. A key observation is that as soon as the nonparametric function $m(x)$ is approximated with the B-spline representation, the regression function itself without measurement error is operationally purely a parametric model (Wang and Yang, 2009), hence we can borrow the treatments designed for parametric measurement error models (Tsiatis and Ma, 2004). Conceptually, we treat the B-spline coefficients as parameters of interest, treat the density function $f_X(x)$ as nuisance parameter, and cast the problem as a semiparametric estimation problem and construct the efficient score function. We now describe the estimation procedure in detail.

First, let $f_X^*(x)$ be a working density function of X. Of course, $f_X^*(x)$ may not be the same as $f_{X0}(x)$, that is, $f_X^*(x)$ is possibly misspecified. We assume that the m is defined on a compact set say, $[0, 1]$, which covers the support of $f_{X0}(x)$. Therefore we only need to consider $m(x)$ on [0,1]. We approximate $m(x)$ using the spline representation $\mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}$. Define

$$\begin{aligned}\mathbf{S}_{\boldsymbol{\beta}}^*(W, Y, \boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \log \int_0^1 f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(W-x) f_X^*(x) d\mu(x) \\ &= -\frac{\int_0^1 f_\epsilon'\{Y - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(W-x) f_X^*(x) \mathbf{B}_r(x) d\mu(x)}{\int_0^1 f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(W-x) f_X^*(x) d\mu(x)},\end{aligned}$$

where $\mu(\cdot)$ is the Lebesgue measure. As the notation suggests, $\mathbf{S}_{\boldsymbol{\beta}}^*(W, Y, \boldsymbol{\beta})$ is the score function with respect to $\boldsymbol{\beta}$ calculated from the joint pdf of (W, Y) under the working model $f_X^*(x)$ and the spline approximation. Due to the possible misspecification of $f_X^*(x)$, the mean of $\mathbf{S}_{\boldsymbol{\beta}}^*(W, Y, \boldsymbol{\beta})$ is not necessarily zero even if the mean function is exactly $\mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}$. Therefore simply solving $\sum_{i=1}^n \mathbf{S}_{\boldsymbol{\beta}}^*(W_i, Y_i, \boldsymbol{\beta}) = \mathbf{0}$ may generate an inconsistent estimator. The idea behind our estimator is to find a function $\mathbf{a}(x, \boldsymbol{\beta})$ so that

$$\begin{aligned}&E\{\mathbf{S}_{\boldsymbol{\beta}}^*(W, Y, \boldsymbol{\beta}) \mid X\} \qquad (1)\\ &= E\left[\frac{\int_0^1 \mathbf{a}(x, \boldsymbol{\beta}) f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(W-x) f_X^*(x) d\mu(x)}{\int_0^1 f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(W-x) f_X^*(x) d\mu(x)} \mid X\right],\end{aligned}$$

and then solve for $\boldsymbol{\beta}$ using the estimating equation

$$\sum_{i=1}^n \left[\mathbf{S}_{\boldsymbol{\beta}}^*(W_i, Y_i, \boldsymbol{\beta}) - \frac{\int_0^1 \mathbf{a}(x, \boldsymbol{\beta}) f_\epsilon\{Y_i - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(W_i-x) f_X^*(x) d\mu(x)}{\int_0^1 f_\epsilon\{Y_i - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(W_i-x) f_X^*(x) d\mu(x)}\right] = \mathbf{0} \qquad (2)$$

ensures the left hand side of (2) has mean zero. The right hand side of (1) is the conditional expectation of $\mathbf{a}(X, \boldsymbol{\beta})$ calculated under the B-spline approximation and the posited model $f_X^*(x)$, hence we alternatively write it as $E^*\{\mathbf{a}(X, \boldsymbol{\beta}) \mid Y_i, W_i, \boldsymbol{\beta}\}$.

To solve for $\mathbf{a}(x, \boldsymbol{\beta})$, we discretize the integral Equation (1) and convert it into a linear system problem. Let $f_X^*(x) = \sum_{j=1}^L c_j I(x = x_j)$, where x_j's are equally spaced discretizing points on [0,1], and $c_j \geq 0, \sum_{j=1}^L c_j = 1$. Then

$$\mathbf{S}_{\boldsymbol{\beta}}^*(W, Y, \boldsymbol{\beta}) = -\frac{\sum_{j=1}^L \mathbf{B}_r(x_j) f_\epsilon'\{Y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(W - x_j) c_j}{\sum_{j=1}^L f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(W - x_j) c_j}.$$

Next, to write out the right hand side of (1) upon discretization, let $\mathbf{A}(\boldsymbol{\beta})$ be an $L \times L$ matrix with its (i, j) entry

$$\begin{aligned}A_{ij}(\boldsymbol{\beta}) = \int & \frac{f_\epsilon\{y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(w - x_j) c_j}{\sum_{j=1}^L f_\epsilon\{y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(w - x_j) c_j} \\ & \times f_\epsilon\{y - \mathbf{B}_r^{\mathrm{T}}(x_i)\boldsymbol{\beta}\} f_U(w - x_i) d\mu(y) d\mu(w).\end{aligned}$$

Let $\mathbf{a}_i = \mathbf{a}(x_i, \boldsymbol{\beta})$, $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_L)$. Further, define $\mathbf{H}_j(\boldsymbol{\beta}) = \{H_{1j}(\boldsymbol{\beta}), \ldots, H_{Lj}(\boldsymbol{\beta})\}$, where

$$H_{ij}(\boldsymbol{\beta}) = \int \frac{-f_\epsilon'\{y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(w - x_j) c_j}{\sum_{j=1}^L c_j f_\epsilon\{y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(w - x_j)} f_\epsilon\{y - \mathbf{B}_r^{\mathrm{T}}(x_i)\boldsymbol{\beta}\} f_U(w - x_i) d\mu(y) d\mu(w),$$

and let $\mathbf{b}(\boldsymbol{\beta})$ be a $p \times L$ matrix, with ith column $\mathbf{b}_i(\boldsymbol{\beta}) = \sum_{j=1}^L H_{ij}(\boldsymbol{\beta}) \mathbf{B}_r(x_j)$. Then $\mathbf{b}(\boldsymbol{\beta}) = \sum_{j=1}^L \mathbf{B}_r(x_j) \mathbf{H}_j(\boldsymbol{\beta})$.

With this notation, the integral Equation (1) becomes $\sum_{j=1}^{L} A_{ij}\mathbf{a}_j = \sum_{j=1}^{L} H_{ij}(\boldsymbol{\beta})\mathbf{B}_r(x_j)$ for $i = 1, \dots, L$, or more concisely, $\mathbf{a}\mathbf{A}^{\mathrm{T}}(\boldsymbol{\beta}) = \mathbf{b}(\boldsymbol{\beta})$. Therefore, $\mathbf{a}(\boldsymbol{\beta}) = \mathbf{b}(\boldsymbol{\beta})\{\mathbf{A}^{-1}(\boldsymbol{\beta})\}^{\mathrm{T}}$, or

$$\mathbf{a}_j(\boldsymbol{\beta}) = \sum_{k=1}^{L} \mathbf{B}_r(x_k)\mathbf{H}_k(\boldsymbol{\beta})\{\mathbf{A}^{-1}(\boldsymbol{\beta})\}^{\mathrm{T}}\mathbf{e}_j, \tag{3}$$

where $\mathbf{e}_j$ is a length L vector with the jth component 1 and all others zero.

Thus, we have obtained $\mathbf{a}(X, \boldsymbol{\beta})$ on the discrete set $x_1, \dots x_L$ and can form

$$\begin{aligned}
&E^*\{\mathbf{a}(X, \boldsymbol{\beta}) \mid Y, W, \boldsymbol{\beta}\}\\
&= \frac{\sum_{j=1}^{L} \mathbf{a}_j(\boldsymbol{\beta}) f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(W - x_j)c_j}{\sum_{j=1}^{L} f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(W - x_j)c_j}\\
&= \sum_{k=1}^{L} \mathbf{B}_r(x_k)\frac{\mathbf{H}_k(\boldsymbol{\beta})\sum_{j=1}^{L}\{\mathbf{A}^{-1}(\boldsymbol{\beta})\}^{\mathrm{T}}\mathbf{e}_j f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(W - x_j)c_j}{\sum_{j=1}^{L} f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(W - x_j)c_j}\\
&= \sum_{k=1}^{L} \mathbf{B}_r(x_k)P_k(W, Y, \boldsymbol{\beta}),
\end{aligned}$$

where

$$P_k(W, Y, \boldsymbol{\beta}) = \frac{\mathbf{H}_k(\boldsymbol{\beta})\sum_{j=1}^{L}\{\mathbf{A}^{-1}(\boldsymbol{\beta})\}^{\mathrm{T}}\mathbf{e}_j f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(W - x_j)c_j}{\sum_{j=1}^{L} f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(W - x_j)c_j}.$$

We then obtain the estimator for $\boldsymbol{\beta}$ by solving the estimating Equation (2) with the corresponding $\mathbf{a}(x, \boldsymbol{\beta})$ plugged in. For simplicity, we use the B-spline with equally spaced knots. In all the functions that are explicitly written to depend on $\boldsymbol{\beta}$, the dependence is always through $\mathbf{B}_r(\cdot)^{\mathrm{T}}\boldsymbol{\beta}$. We summarize the above procedure in the algorithm below.

1. Adopt a B-spline approximation $m(x) = \mathbf{B}_r{}^{\mathrm{T}}(x)\boldsymbol{\beta}$.
2. Posit a working model $f_X^*(x)$.
3. Select equally spaced grid points on the support of X.
4. Approximate all the integrals with respect to x with summations on the grid points.
5. Obtain $\mathbf{a}(x, \boldsymbol{\beta})$ at the grid points from (3).
6. Construct and solve (2) to obtain $\widehat{\boldsymbol{\beta}}$.
7. Obtain $\widehat{m}(x) = \mathbf{B}_r(x)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$.

We show that $\mathbf{B}_r^{\mathrm{T}}(x)\widehat{\boldsymbol{\beta}}$ converges to $m(x)$ whether the working model $f_X^*(x)$ is correct or not. In addition, the convergence is nearly at the nonparametric rate without measurement error, and we derive its estimation variance in Section 3 under mild conditions.

Remark 1. The assumption that $f_{X0}(x)$ has a compact support enables us to represent the mean function with a B-spline representation, hence it is essential. In practice, when the support is unknown, we recommend to represent the mean function on a sufficiently large domain which contains the true support. This ensures consistency of the estimator on the true support. If the domain is excessively large, the matrix $\mathbf{A}(\boldsymbol{\beta})$ will be nearly singular, which can be used as a practical tool to adjust the practical range. Rigorous analysis on estimation of the support under measurement error is challenging despite of some existing work (Kneip et al., 2015) and is beyond the scope of the paper.

2.2. Probability density function estimation

The estimation of $f_{X0}(x)$ is much simpler compared to the regression case. We aim to estimate $f_{X0}(x)$ on a finite set, say [0,1]. We approximate $f_{X0}(x)$ using B-splines (Masri and Redner, 2005). To account that it is a density function, we let the approximation be

$$f_X(x, \boldsymbol{\theta}) \equiv \frac{\exp\{\mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\theta}\}}{\int_0^1 \exp\{\mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\theta}\}dx}, \tag{4}$$

where $\mathbf{B}_r(x)$ is a vector of B-spline basis functions, and $\boldsymbol{\theta}$ is the B-spline coefficient vector. If the true support of $f_{X0}(x)$ goes beyond [0,1], the scaling factor in (4) will be off and the resulting $\widehat{f_X}(x)$ will differ from the true pdf by a constant scaling. For reasons of identifiability, we fix the first component of $\boldsymbol{\theta}$ at zero, i.e., $\theta_1 = 0$, and leave the remaining components $\boldsymbol{\theta}_L$ free. Here, a_L denotes the subvector of a generic vector $\mathbf{a}$ without the first element. Then

$$f_W(w, \boldsymbol{\theta}) \equiv \frac{\int_0^1 \exp\{\mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\theta}\} f_U(w - x)dx}{\int_0^1 \exp\{\mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\theta}\}dx}$$

is an approximation to the pdf of $W = X + U$, a surrogate of X. We then perform simple maximum likelihood estimation (MLE), i.e., we maximize

$$\sum_{i=1}^{n} \log f_W(W_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int_0^1 \exp\{\mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\theta}\} f_U(W_i - x)dx - n\log \int_0^1 \exp\{\mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\theta}\}dx,$$

with respect to $\boldsymbol{\theta}_L$ to obtain $\widehat{\boldsymbol{\theta}}_L$, and then reconstruct $f_X(x, \widehat{\boldsymbol{\theta}})$ and use it as the estimator for $f_{X0}(x)$, i.e., $\widehat{f}_X(x) = f_X(x, \widehat{\boldsymbol{\theta}})$. Here $\widehat{\boldsymbol{\theta}} = (0, \widehat{\boldsymbol{\theta}}_L{}^{\mathrm{T}})^{\mathrm{T}}$.

Having obtained $\widehat{f}_X(x)$, we could use $\widehat{f}_X(x)$ as a way of selecting $f_X^*(x)$ in the regression case considered in Section 2.1. The resulting properties are however more difficult to establish. While the estimation procedure for $f_{X0}(x)$ is extremely simple, it is not as straightforward to establish the large sample properties of the estimator. In Section 3, we will show that $\widehat{f}_X(x)$ converges to $f_{X0}(x)$ at a near-nonparametric rate under mild conditions, up to a scale.

3. Asymptotic results

3.1. Results of regression mean function estimation

To facilitate the description of the theoretical results, we introduce some notation. Define

$$\begin{aligned} P(x, W, Y, \boldsymbol{\beta}) &= \frac{\mathbf{H}(x, \boldsymbol{\beta}) \sum_{j=1}^{L} \{\mathbf{A}^{-1}(\boldsymbol{\beta})\}^{\mathrm{T}} \mathbf{e}_j f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x_j)\boldsymbol{\beta}\} f_U(W - x_j) c_j}{\int_0^1 f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(W - x) f_X^*(x) d\mu(x)}, \\ \mathbf{H}(x, \boldsymbol{\beta}) &= \{H_1(x, \boldsymbol{\beta}), \dots, H_L(x, \boldsymbol{\beta})\}, \\ H_i(x, \boldsymbol{\beta}) &= -\int \frac{f_\epsilon'\{y - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(w - x) f_X^*(x)}{\int_0^1 f_\epsilon\{y - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(w - x_j) f_X^*(x) d\mu(x)} \\ &\qquad \times f_\epsilon\{y - \mathbf{B}_r^{\mathrm{T}}(x_i)\boldsymbol{\beta}\} f_U(w - x_i) d\mu(y) d\mu(w), \end{aligned}$$

and write

$$E^*\{\mathbf{a}(X, \boldsymbol{\beta}) \mid Y, W, \boldsymbol{\beta}\} = \int_0^1 \mathbf{B}_r(x) P(x, W, Y, \boldsymbol{\beta}) d\mu(x);$$

and

$$\mathbf{S}_{\boldsymbol{\beta}}^*(W, Y, \boldsymbol{\beta}) = \int_0^1 \frac{\mathbf{B}_r(x) f_\epsilon'\{Y - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(W - x) f_X^*(x)}{-\int_0^1 f_\epsilon\{Y - \mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\beta}\} f_U(W - x) f_X^*(x)} d\mu(x).$$

We further define $\mathbf{S}_{\boldsymbol{\beta}}^*(W_i, Y_i, m)$, $E^*\{\mathbf{a}(X, m) \mid Y_i, W_i, m\}$, $P(x, W, Y, m)$, $P_k(W, Y, m)$ to be the resulting quantities when we replace all the appearances of $\mathbf{B}_r(\cdot)^{\mathrm{T}}\boldsymbol{\beta}$ in $\mathbf{S}_{\boldsymbol{\beta}}^*(W_i, Y_i, \boldsymbol{\beta})$, $E^*\{\mathbf{a}(X, \boldsymbol{\beta}) \mid Y_i, W_i, \boldsymbol{\beta}\}$, $P(x, W, Y, \boldsymbol{\beta})$, and $P_k(W, Y, m)$ by $m(\cdot)$ respectively. Here $\mathbf{a}(X, m)$ is a function that satisfies

$$E[\mathbf{S}_{\boldsymbol{\beta}}^*(W_i, Y_i, m) - E^*\{\mathbf{a}(X, m) | Y_i, W_i, m\} | X, m] = \mathbf{0},$$

where the last m is used to emphasize that the calculation of the outside expectation depends on m.

We further define $S_m(Y, W, m)$ to be a linear operator on L_p whose value at $s(\cdot) \in L_p$ is

$$\begin{aligned} &S_m(Y, W, m)(s) \\ &= \int_0^1 \left[-\frac{f_\epsilon'\{Y - m(x)\} f_U(W - x) f_X^*(x) d\mu(x)}{\int_0^1 f_\epsilon\{Y - m(x)\} f_U(W - x) f_X^*(x) d\mu(x)} - P(x, Y, W, m) \right] s(x) dx. \end{aligned}$$

Theorem 1. *Assume Conditions – and – given in Supplement. Let* $m(x) \in C^q([0, 1])$, $q \geq 1$, *and* $\widehat{m}(x) = \mathbf{B}_r^{\mathrm{T}}(x)\widehat{\boldsymbol{\beta}}$. *Then* $\sup_{x \in [0,1]} |\widehat{m}(x) - m(x)| = O_p\{(nh_b)^{-1/2} + h_b^q\}$. *Specifically,* $\mathrm{bias}\{\widehat{m}(x)\} = E\{\widehat{m}(x)\} - m(x) = O_p(h_b^q) + o_p\{(nh_b)^{-1/2}\}$. *The mean squared error* $\mathrm{MSE}\{\widehat{m}(x)\} \equiv \mathrm{var}\{\widehat{m}(x)\} + \mathrm{bias}\{\widehat{m}(x)\}^2 = O\{h_b^{2q} + (nh_b)^{-1}\}$, *and is minimized at* $N \asymp n^{1/(2q+1)}$ *to be* $O\{n^{2q/(2q+1)}\}$. *Further,*

$$\sqrt{nh_b}[\widehat{m}(x) - m(x) - \mathrm{bias}\{\widehat{m}(x)\}] = n^{-1/2} \sum_{i=1}^{n} Q(W_i, Y_i, x) + o_p(1).$$

If $Nn^{-1/(2q+1)} \to \infty$, *then*

$$\sqrt{nh_b}\{\widehat{m}(x) - m(x)\} = n^{-1/2} \sum_{i=1}^{n} Q(W_i, Y_i, x) + o_p(1).$$

Here

$$
\begin{aligned}
&Q(W_i, Y_i, x)\\
&= \sqrt{h_b}\mathbf{B}_r^{\mathrm{T}}(x)\left[-\left\{E\left(\frac{\partial[\mathbf{S}_{\boldsymbol{\beta}}^*(W_i, Y_i, \boldsymbol{\beta}) - E^*\{\mathbf{a}(X, \boldsymbol{\beta}) \mid Y_i, W_i, \boldsymbol{\beta}\}]}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\mathbf{B}_r(\cdot)^{\mathrm{T}}\boldsymbol{\beta}=m(\cdot)}\right)\right\}^{-1}\right.\\
&\quad \left.\times \mathbf{S}_{\boldsymbol{\beta}}^*(W_i, Y_i, m) - E^*\{\mathbf{a}(X, m) \mid Y_i, W_i, m\}\right].
\end{aligned}
$$

Theorem 1 shows that the B-spline regression mean function estimator has bias of order $O_p(h_b^q) + o_p\{(nh_b)^{-1/2}\}$ and standard error of order $O_p\{(nh_b)^{-1/2}\}$, which is the standard nonparametric regression result without measurement error. The result here requires X to have compact support, so is not comparable with the minimax results established in the literature (Carroll and Hall, 1988; Fan, 1991). Further, to minimize the MSE, we let $h_q \asymp n^{-1/(2q+1)}$, leading to the MSE of order $n^{2q/(2q+1)}$. On the other hand, to suppress the estimation bias asymptotically, we need to under-smooth by setting $h = o\{n^{1/(2q+1)}\}$.

3.2. Results of probability density function estimation

For ease of presentation, we establish the result by assuming that the support of $f_{X0}(x)$ is [0,1]. If not, the results below still hold by replacing $f_{X0}(x)$ with $cf_{X0}(x)$, where c is a constant.

Theorem 2. *Assume Conditions –. Let $f_{X0}(x) \in C^q([0, 1])$, $q \geq 1$ and $\widehat{\boldsymbol{\theta}} = (0, \widehat{\boldsymbol{\theta}}_L{}^{\mathrm{T}})^{\mathrm{T}}$, $\widehat{\boldsymbol{\theta}}_L$ be defined in Proposition and*

$$
\widehat{f}_X(x) = \frac{\exp\{\mathbf{B}_r^{\mathrm{T}}(x)\widehat{\boldsymbol{\theta}}\}}{\int_0^1 \exp\{\mathbf{B}_r^{\mathrm{T}}(v)\widehat{\boldsymbol{\theta}}\}dv}.
$$

Then $\sup_{x\in[0,1]} |\log\{\widehat{f}_X(x)\} - \log\{f_{X0}(x)\}| = O_p\{(nh_b)^{-1/2} + h_b^q\}$. Specifically,

$$
\mathrm{bias}\{\widehat{f}_X(x)\} \equiv E\{\widehat{f}_X(x)\} - f_{X0}(x) = O_p(h_b^q) + o_p\{(nh_b)^{-1/2}\}.
$$

The mean squared error $\mathrm{MSE}\{\widehat{f}_X(x)\} \equiv \mathrm{var}\{\widehat{f}_X(x)\} + \mathrm{bias}\{\widehat{f}_X(x)\}^2 = O\{h_b^{2q} + (nh_b)^{-1}\}$, and is minimized at $N \asymp n^{1/(2q+1)}$ to be $O\{n^{2q/(2q+1)}\}$. Further,

$$
\sqrt{nh_b}\left[\widehat{f}_X(x) - f_{X0}(x) - \mathrm{bias}\{\widehat{f}(x)\}\right] = 1/\sqrt{n}\sum_{i=1}^n Q(W_i, x) + o_p(1).
$$

If $Nn^{-1/(2q+1)} \to \infty$, then

$$
\sqrt{nh_b}\left\{\widehat{f}_X(x) - f_{X0}(x)\right\} = 1/\sqrt{n}\sum_{i=1}^n Q(W_i, x) + o_p(1).
$$

Here

$$
\begin{aligned}
Q(W_i, x) = \sqrt{h_b}\frac{\partial}{\partial \boldsymbol{\theta}_L{}^{\mathrm{T}}}\frac{\exp\{\mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\theta}_0\}}{\int_0^1 \exp\{\mathbf{B}_r^{\mathrm{T}}(x)\boldsymbol{\theta}_0\}dx}&\left(E\left[\frac{\int_0^1 f_{X0}(x)f_U(W_i - x)\mathbf{B}_{rL}(x)\mathbf{B}_{rL}^{\mathrm{T}}(x)dx}{\int_0^1 f_{X0}(x)f_U(W_i - x)dx}\right.\right.\\
&- \frac{\left\{\int_0^1 f_{X0}(x)f_U(W_i - x)\mathbf{B}_{rL}(x)dx\right\}^{\otimes 2}}{\left\{\int_0^1 f_{X0}(x)f_U(W_i - x)dx\right\}^2} - \int_0^1 f_{X0}(x)\mathbf{B}_{rL}(x)\mathbf{B}_{rL}^{\mathrm{T}}(x)dx\\
&\left.\left. + \left\{\int_0^1 f_{X0}(x)\mathbf{B}_{rL}(x)dx\right\}^{\otimes 2}\right]\right)^{-1}\\
&\times \int_0^1 \left\{\frac{f_{X0}(x)f_U(W_i - x)}{\int_0^1 f_{X0}(x)f_U(W_i - x)dx} - f_{X0}(x)\right\}\mathbf{B}_{rL}(x)dx.
\end{aligned}
$$

Theorem 2 shows that the B-spline MLE density estimator has bias of order $O_p(h_b^q) + o_p\{(nh_b)^{-1/2}\}$ and standard error of order $O_p\{(nh_b)^{-1/2}\}$, which is the standard nonparametric density estimation result when no measurement error occurs. In addition, to minimize the MSE, we can use $h_b \asymp n^{-1/(2q+1)}$, leading to the MSE with order $n^{2q/(2q+1)}$. To suppress the estimation bias asymptotically, we need to under-smooth by setting $h = o\{n^{-1/(2q+1)}\}$.

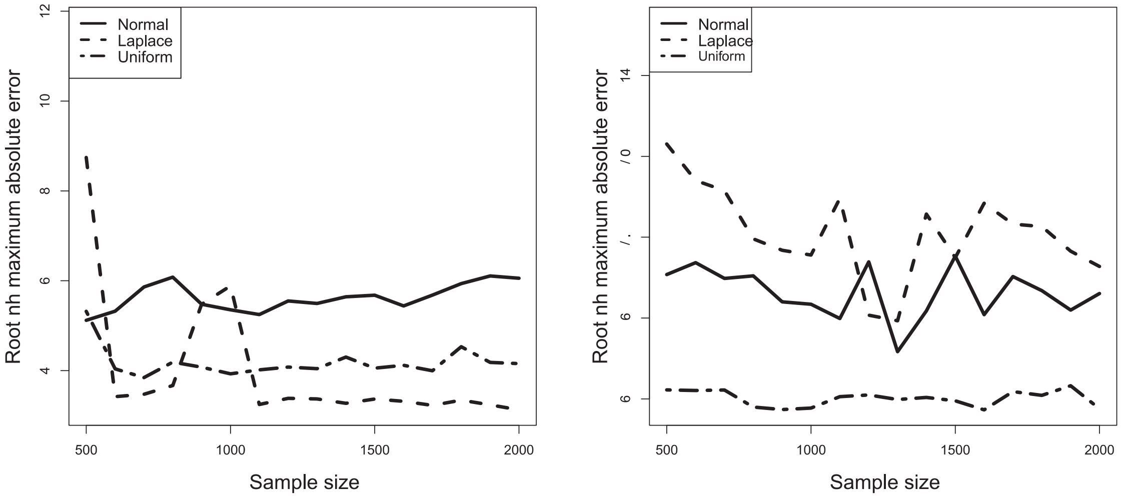

Fig. 3. Performance of the B-spline semiparametric mean estimation (left) and B-spline MLE pdf estimation (right). Results based on 200 simulations. The solid lines are for the Normal measurement error case, the dashed lines are for Laplace error and the dot-dashed lines are for the Uniform error case.

4. Simulation studies

We conducted two simulation studies to illustrate the finite sample performance of our regression mean and density function estimators. In all our simulations, X is generated from a beta distribution with both shape parameters equal to 4. We then generated the measurement error U from three models:

- I(a): a normal distribution with mean 0, variance 0.25, denoted by $N(0, 0.25)$,
- I(b): a Laplace distribution with mean 0 and scale $0.5/\sqrt{2}$, denoted by $\mathrm{Lap}(0, 0.5/\sqrt{2})$,
- I(c): a uniform distribution on $[-\sqrt{3/4}, \sqrt{3/4}]$, denoted by $\mathrm{Unif}(-\sqrt{3/4}, \sqrt{3/4})$.

We also consider three models with error variabilities

- II(a): $N(0, 0.0025)$, II(b): $\mathrm{Lap}(0, 0.05/\sqrt{2})$, II(c): $\mathrm{Unif}(-0.125, 0.125)$.

These settings cover the scenarios where data have low and high measurement error radiability, defined by $\mathrm{var}(X)/\mathrm{var}(W)$. The measurement error reliabilities are 0.11, 0.11, 0.14 for Models I (a), (b) (c), and are 0.92, 0.92, 0.85 for Model II (a), (b) and (c).

4.1. Performance of the B-spline-assisted semiparametric mean estimator

We now evaluated the finite sample performance of the B-spline semiparametric mean regression method described in Section 2.1. We used sample sizes n from 500 to 2000, used cubic B-splines, with the number of knots equal to the smallest integer larger than $1.3n^{1/5}$. Throughout the numerical implementation, we select $f^*(x)$ to be a discrete uniform distribution with positive values at 10 evenly spaced points on the support. In this case we generated X from a beta distribution with both shape parameters equal to 2. The true regression mean function is $m(x) = \sin(2\pi x)$ and we generated the regression model errors ϵ from a normal distribution with mean zero and variance 0.25. We generated the measurement errors U from the three different distributions described in Models I(a)–I(c).

In the right panel of Figure 3, we plotted the averaged root-(nh_b) times the maximum absolute error (MAE) calculated via $\sqrt{nh_b}\sup_x|\widehat{m}(x) - m(x)|$ as a function of the sample size n. The curves stabilize as sample size increases, and is largely flat after $n = 1000$, indicating that $\sup_x|\widehat{m}(x) - m(x)|$ has order $(nh_b)^{-1/2}$.

We further compared the B-spline semiparametric method with the deconvolution method (Fan and Truong, 1993) in the nonparametric mean regression model, where we use the two-stage plug-in bandwidth selection method (Delaigle and Gijbels, 2002) in implementing the deconvolution method. In Figure 4, we plotted the average $\sup_x|\widehat{m}(x) - m(x)|$ over 200 simulations for both methods. With moderate to significant amount of noise, the B-spline semiparametric method greatly outperforms the deconvolution method with smaller average error.

We further reduced the measurement error variability and generated the errors from Models II (a)–(c) to investigate the regression estimator. We plotted the mean function estimates and the 90% confidence bands for the B-spline semiparametric and deconvolution methods in Figure 5 for sample size 500, and also provided the same results in Figures, and in the Supplement. for sample sizes 1000, 2000, as well as 200. These results indicate that the B-spline semiparametric estimator indeed outperforms the deconvolution method. Their performance difference in terms of the average

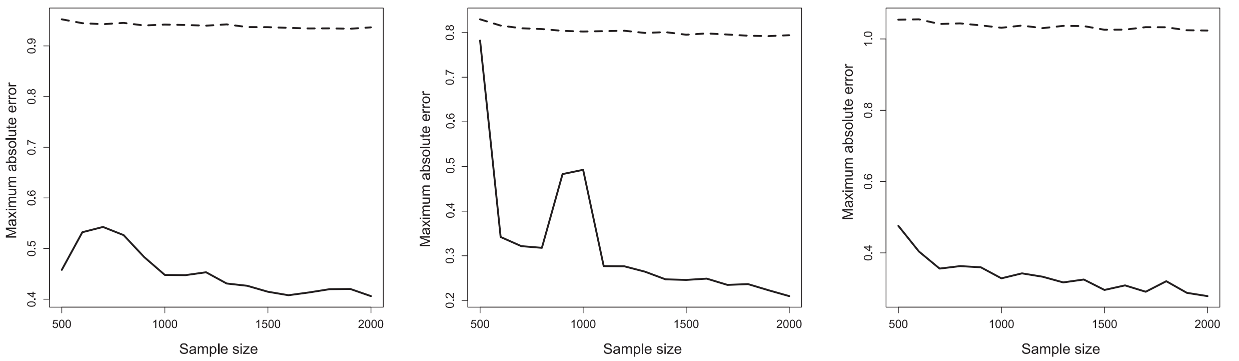

Fig. 4. Comparison of mean estimators based on the B-spline semiparametric method (solid) and the deconvolution (dashed) method, when measurement errors are Normal (left), Laplace (middle) and Uniform (right) respectively. Average MAE $\sup_x |\widehat{m}_X(x) - m_X(x)|$ is computed based on 200 simulations at sample sizes from 500 to 2000.

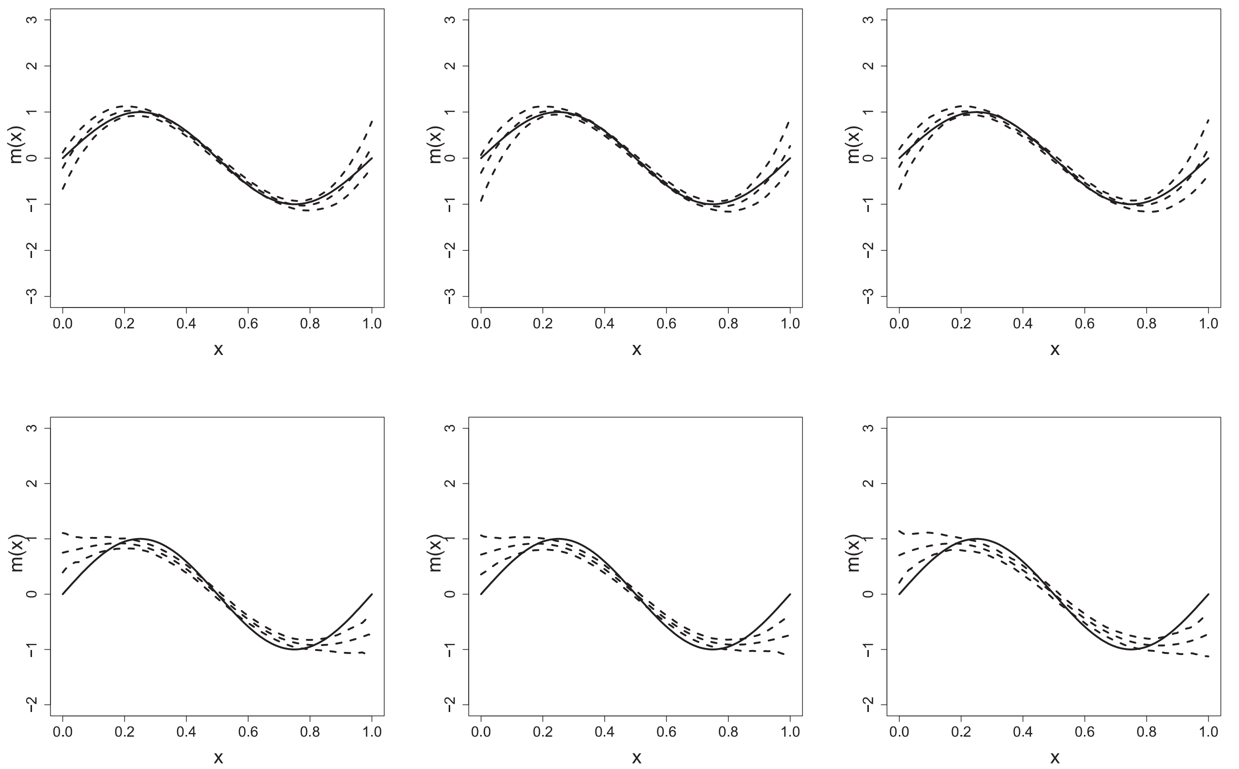

Fig. 5. B-spline semiparametric regression estimation (top) and deconvolution estimation (bottom) from 200 simulations: The solid lines represent the true functions and the dashed lines represent the estimated functions and their 90% confidence bands. The rows are the results for Models II (a)–(c), respectively. Sample size is 500.

L_2 norm of the difference between the estimated and true mean curves defined by $E[\int\{\widehat{m}(x) - m(x)\}^2 dx]$, and the average MAE defined by $E\{\sup_x |\widehat{m}(x) - m(x)|\}$ between the estimated and true mean curves are further provided in Table 1. The deconvolution density estimation and nonparametric regression are implemented using the code from the web site https://researchers.ms.unimelb.edu.au/~aurored/links.html.

4.2. Performance of the B-spline-assisted density estimator

To evaluate the B-spline method for estimating the density functions, we generated data sets of sample sizes from $n = 500$ to $n = 2000$. We used cubic B-splines with the number of knots equal to the smallest integer larger than $1.3n^{1/5}$.

In the left panel of Figure 3, we plotted the averaged root-(nh_b) MAE, calculated as $\sqrt{nh_b}\sup_x |\widehat{f}_X(x) - f_{X0}(x)|$, versus the sample sizes n. Following Theorem 2, the root-(nh_b) MAE has a constant order. This translates to the curves in the plots that are nearly bounded, which is what we observe, especially when the sample size grows larger than 700.

Table 1
Comparison between the B-spline semiparametric estimator/MLE and deconvolution method. Average the average L_2 norm of the difference between the estimated and true curves and average MAE over 200 simulations are reported.

regression estimation: $E(\int[\widehat{m}(x)-m(x)]^2dx)$						
	B-spline semiparametric			Deconvolution		
	$n=500$	$n=1000$	$n=2000$	$n=500$	$n=1000$	$n=2000$
Model II(a)	0.030	0.025	0.018	0.407	0.404	0.402
Model II(b)	0.052	0.022	0.007	0.287	0.279	0.276
Model II(c)	0.025	0.013	0.007	0.502	0.490	0.495
regression estimation: $E\{\sup_x \lvert\widehat{m}(x)-m(x)\rvert\}$						
	B-spline semiparametric			Deconvolution		
	$n=500$	$n=1000$	$n=2000$	$n=500$	$n=1000$	$n=2000$
Model II(a)	0.370	0.263	0.175	0.908	0.796	0.762
Model II(b)	0.425	0.264	0.163	0.857	0.828	0.779
Model II(c)	0.414	0.291	0.219	0.880	0.832	0.801
pdf estimation: $E(\int[\widehat{f}(x)-f_{X0}(x)]^2dx)$						
	B-spline MLE			Deconvolution		
	$n=500$	$n=1000$	$n=2000$	$n=500$	$n=1000$	$n=2000$
Model II(a)	0.008	0.004	0.002	0.011	0.007	0.005
Model II(b)	0.009	0.004	0.002	0.012	0.007	0.004
Model II(c)	0.010	0.006	0.003	0.018	0.017	0.014
pdf estimation: $E\{\sup_x \lvert\widehat{f}_X(x)-f_{X0}(x)\rvert\}$						
	B-spline MLE			Deconvolution		
	$n=500$	$n=1000$	$n=2000$	$n=500$	$n=1000$	$n=2000$
Model II(a)	0.203	0.156	0.107	0.238	0.212	0.160
Model II(b)	0.213	0.155	0.104	0.230	0.197	0.158
Model II(c)	0.230	0.181	0.131	0.315	0.242	0.230

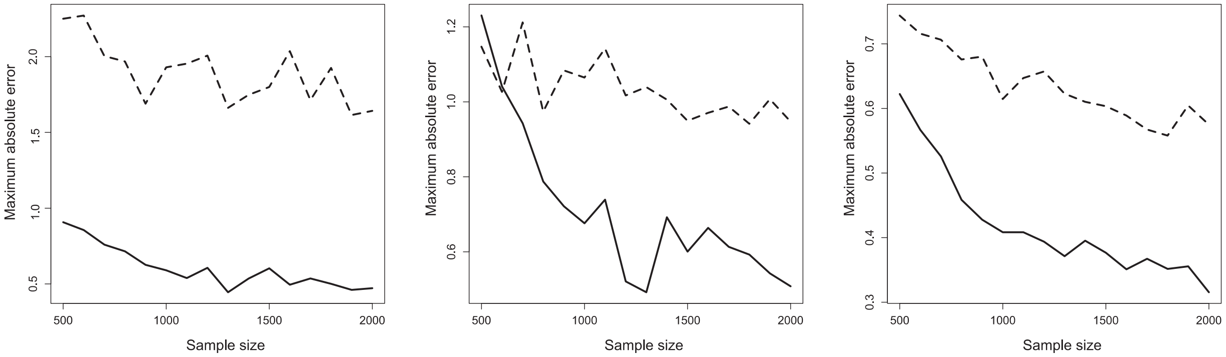

Fig. 6. Comparison of pdf estimators based the B-spline MLE (solid) and the deconvolution (dashed) method, when measurement errors are Normal (left), Laplace (middle) and Uniform (right) respectively. Average MAE $\sup_x \lvert\widehat{f}_X(x)-f_X(x)\rvert$ is computed based on 200 simulations at sample sizes from 500 to 2000.

We also compared the B-spline MLE method with the widely used deconvolution method (Stefanski and Carroll, 1990) for density estimation. In Figure 6, we plotted the average values of $\sup_x \lvert\widehat{f}_X(x)-f_{X0}(x)\rvert$ based on 200 simulations for both methods at different sample sizes. We adopted the two-stage plug-in bandwidth selection method proposed in Delaigle and Gijbels (2002) in implementing the deconvolution method. Unsurprisingly, results in Figure 6 indicate that the B-spline MLE outperformed the deconvolution method with a rather significant gain in this case. We suspect that such a gain is a direct result of the slow convergence rate of the deconvolution method. Although a large measurement error causes difficulties for both methods, the deconvolution method deteriorates much faster than our method.

To further examine the performance of individual estimated pdf curves from both methods, we also plotted the estimated mean density curves for sample sizes 500, 1000 and 2000. Because the deconvolution method performs poorly when the measurement errors are large, see Figure 6, we reduced the error variances, and generated the three error distributions from Models I(a), I(b) and I(c). With reduced error variability, we plotted the resulting density estimates and their 90%

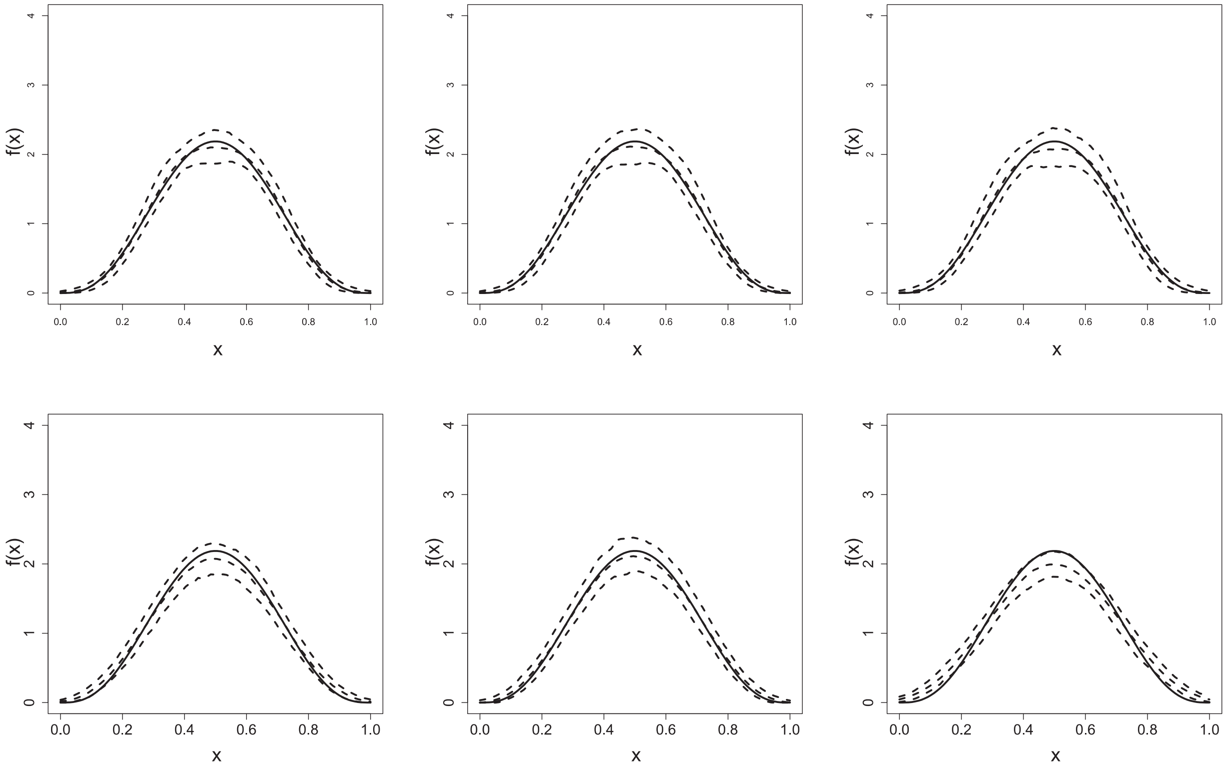

Fig. 7. B-spline MLE density estimation (top) and deconvolution estimation (bottom) from 200 simulations: The solid lines represent the true functions and the dashed lines represent the estimated functions and their 90% confidence bands. The first row to the third row are the results for Models II (a)–(c), respectively. The sample size is 500.

confidence bands. Figure 7 contains the results of the B-spline MLE and the deconvolution estimator using the two-stage plug-in bandwidth (Delaigle and Gijbels, 2002), at the sample size 500. Although not as dramatic as the large error case, the B-spline MLE is still closer to the true pdf with narrower confidence band, hence is more precise than the deconvolution method.

We also provide similar comparisons at sample sizes 1000 and 2000 in Figures and in the Supplement. For a quantitative comparison, we also computed the average L_2 norm of the difference between the true pdf curve and the estimated pdf curve and the average MAE in Table 1. These results show that the B-spline MLE performs consistently better than deconvolution. The performance of the two methods largely follows the same pattern when sample sizes are smaller, although the improvement of the B-spline method over the deconvolution method is of course not as dramatic. We provide the results for $n = 200$ in Figure in the Supplement.

4.3. Robustness of the B-spline estimators

We examine the performance of the proposed method when the support of X is misspecified. We first generate U from Models II(a)–II(c), and simulate X from a beta distribution with shape parameters 4, and generate Y following the same procedure as in Section 4.1. We specify the support of X to be $[0, 2]$ in the estimation, while note that the true support of X is [0,1]. We plotted the mean of the estimates and the 90% confidence bands for the B-spline semiparametric and density estimator in Figure 8, where the sample size is $n = 500$. The results show that the proposed method provides accurate estimation of $m(x)$ and $f_{X0}(x)$ on the support of X. For completeness, we also provided the estimation results on [1,2], which is beyond the true support. We can see that the estimated density function correctly detects that there is no mass in this region, while the estimated regression function seems to extrapolate the relation near the boundary.

We further consider a setting where the $f_U(\cdot)$ and $f_\epsilon(\cdot)$ are misspecified. In the simulation, we generate X from a beta distribution with shape parameters 4, U and ϵ from a $\text{Lap}(0, 0.5/\sqrt{2})$ distribution. In the estimation, we misspecify U and ϵ to be $N(0, 0.25)$ distributed, which have the same first two moments as the true distributions. We present the B-spline semiparametric regression estimator for $m(x)$ in the left panel of Figure 9. The results suggest that the estimators is close to the true values even though we misspecify both f_U and f_ϵ in the estimation. In addition, we consider a setting where U and X are correlated. In the first simulation, we generate X from a beta distribution with shape parameters 4, ϵ from a $N(0, 0.25)$ distribution. Furthermore, we generate a random variable G from a gamma distribution with shape parameter X and rate parameter $b = \sqrt{X/0.25}$, and let $U = G - X/b$. This specification yields $E(U|X) = 0$ and $\text{var}(U|X) = 0.25$. In the estimation, we

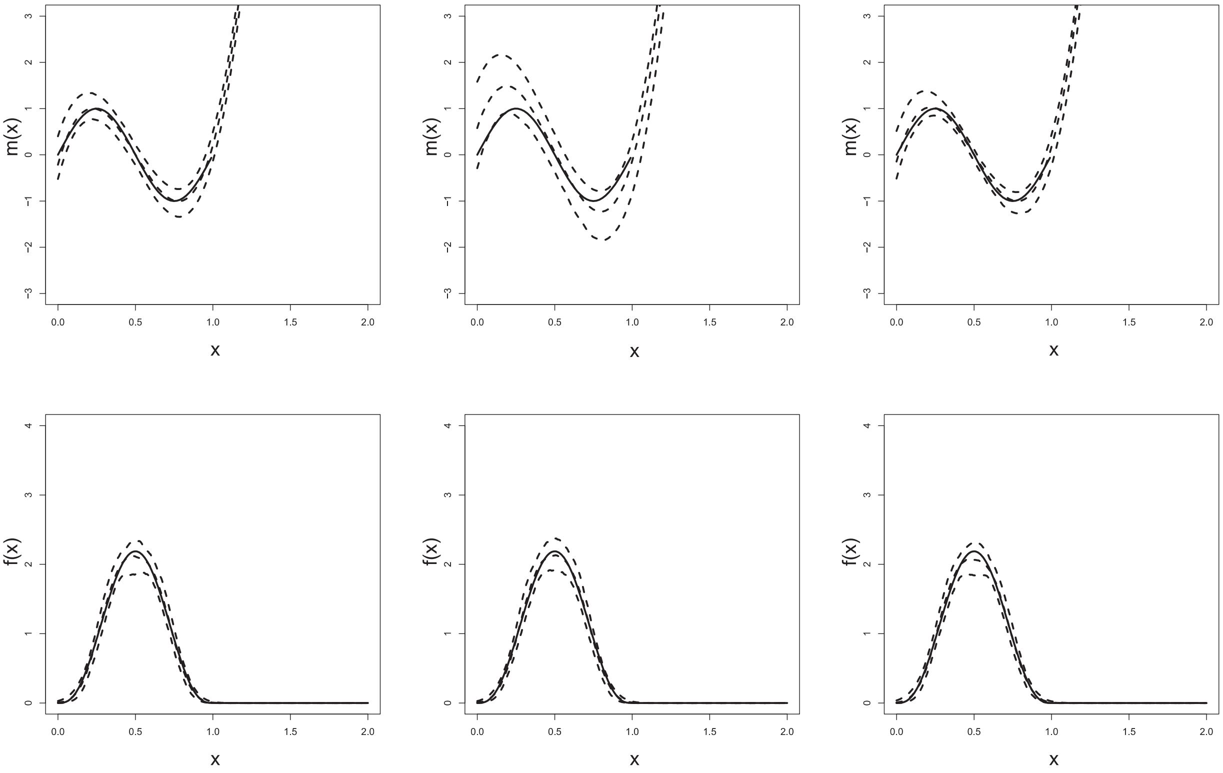

Fig. 8. B-spline semiparametric regression estimation (top) and B-spline MLE density estimation (bottom) and from 200 simulations when the support of X is misspecified: The solid lines represent the true functions and the dashed lines represent the estimated functions and their 90% confidence bands. The first column to third column are the results for Models II (a)–(c), respectively. Sample size is 500.

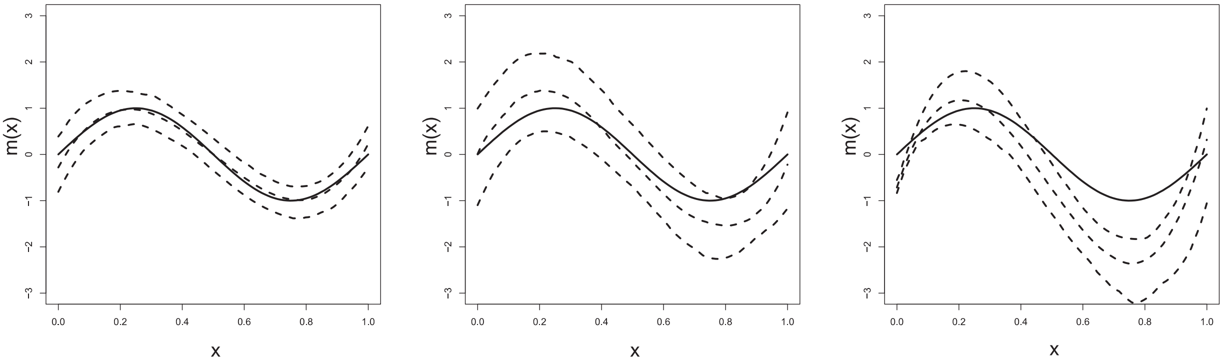

Fig. 9. The B-spline semiparametric regression estimation from 200 simulations when the distributions of U and ϵ are misspecified (left), when U and X are correlated and the first two conditional moment of U is correctly specified (middle), and when U and X are correlated while the first two conditional moments of U are misspecified (right): The solid lines represent the true functions and the dashed lines represent the estimated functions and their 90% confidence bands. Sample size is 500.

misspecify U to be $N(0, 0.25)$ distributed, which matches the first two moments of the conditional distribution of U given X. We present the B-spline semiparametric regression estimator for $m(x)$ in the middle panel of Figure 9. The results suggest that when X and U are correlated, but the first two conditional moments of U are specified correctly, the proposed method provides reasonable estimators with 90% confidence interval covers the true values. In the second simulation, we generate X from a shifted beta distribution with shape parameters 4 and recentered to 0, ϵ from a $N(0, 0.25)$ distribution, U from a normal distribution with mean X and variance 0.223. In the estimation, we misspecify U to be $N(0, 0.25)$ distributed, which only matches the first two moments of the marginal distribution of U. We present the B-spline semiparametric regression estimator for $m(x)$ in the right panel of Figure 9. The results show that when X and U are correlated, but the first two conditional moments of U are misspecified, the estimators deviate from the true value. Three observations agree with our

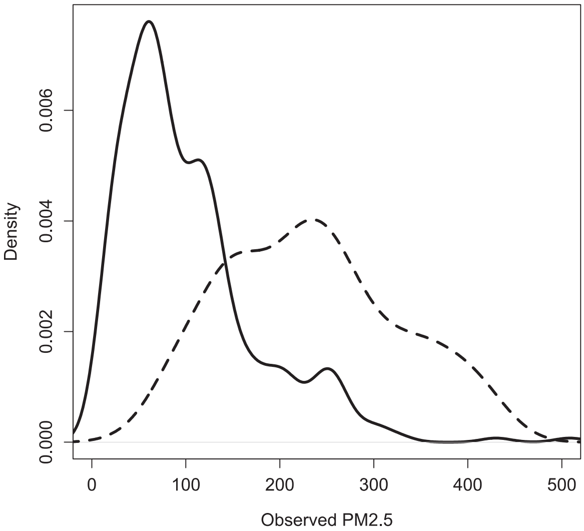

Fig. 10. The estimated pdf of PM2.5 without considering measurement error, based on data from the Beijing Environmental Protection Bureau (solid line) and the "Mission China" web site (dashed-line).

experience, in that often the first two moments of $f_{U|X}(u,x)$ play a very important role. When $E(U \mid X)$ and $\mathrm{var}(U \mid X)$ are correctly specified, the estimator often shows some robustness performance. Of course, more rigorous investigations are needed on parameter estimation when X and U are correlated, which is beyond the scope of the current work.

5. PM2.5 and Daily Asthma ERV rate

Heavy PM2.5 air pollution has become a serious problem in China and its possible effect on respiratory diseases has been a concern in public health. Starting from 2012, the Beijing Environmental Protection Bureau (BEPB) has been recording the daily PM2.5 levels in Beijing. Based on these data, Xu et al. (2016) studied the effect of PM2.5 on daily asthma ERV rate in 2013. Specifically, they explored the PM2.5 effect on the number of daily asthma emergency room visits (ERV) in ten hospitals in Beijing, but found no significant effect. In fact, the mean number of daily asthma ERVs even shows a decreasing trend along the increase of measured PM2.5. This contradicts the general conclusion that PM2.5 has short term adverse effects on asthma (Fan et al., 2016).

A potential reason of this inconsistency is the errors in the PM2.5 measurements which were not taken into account in the above analysis. Indeed, there is much debate on the accuracy of the PM2.5 reports, especially in the early years such as 2013. For example, we compared the daily average PM2.5 reports in 2013 from 17 ambient air quality monitoring stations and those reported by the "Mission China Beijing" web site (Mission-China, 2016) maintained by the U.S. Department of State, and show the two estimated pdfs of PM2.5 in Figure 10. It is clear that the estimated pdfs of PM2.5 from the two sources are very different, where the PM2.5 concentrations obtained from the BEPB yields a pdf estimate with the mode to the left of that obtained from the "Mission China Beijing", indicating a generally less severe air pollution problem. This motivates us to consider the measurement error issue in studying the effect of PM2.5 on the daily asthma ERV.

We restrict our analysis of PM2.5 to the range from 0 to 400, since the smallest and largest recorded PM2.5 values are 6.65 and 328.41, respectively. Instead of taking a log transform (Eckert et al., 1997), we work with the original data form. For computational convenience, we first rescaled the observed PM2.5 by dividing 328.41 − 6.65, the range of the observed PM2.5 data, for each observation to get scaled PM2.5. After the rescaling, the mean and variance of W_i are 0.290 and 0.045, respectively. The data set we analyzed contains 337 observations. In the data available to us, the ith observation contains the number of daily asthma ERVs, which we treat as response Y_i, the average scaled PM2.5 level over 17 ambient air quality monitoring stations from BEPB, which we denote as W_i, and the PM2.5 level from "Mission China Beijing", which we write as W_{0i}. To carry out the analysis, we let the true scaled PM2.5 value be X_i, and let W_{ki} and U_{ki} be the observed PM2.5 value and its corresponding measurement error at the kth monitoring station, $k = 1, \dots, 17$. We assume U_{ki}'s are independent of each other and of X_i. Then $W_i = \sum_{k=1}^{17} W_{ki}/17$ and we use the average W_i as our surrogate measurement of X_i. To obtain the measurement error variance associated with W_i, we write $\overline{U}_i = \sum_{k=1}^{17} U_{ki}/17$, and get $W_i = X_i + \overline{U}_i$. Because our preliminary analysis result in Figure 10 suggests a possible discrepancy between the measurements in W_i and in W_{0i}, we allow a potential bias term b and model $W_{0i} = b + X_i + U_{0i}$, where U_{0i} is the measurement error of W_{0i}. We assume all the U_{ki}, $k = 0, \dots, 17$ to have the same distribution with mean zero, and to be independent of each other, and we estimate b by $\widehat{b} = n^{-1}(\sum_{i=1}^{n} W_{0i} - \sum_{i=1}^{n} W_i)$, which yields the value $\widehat{b} = 0.41$. We further estimated the variance of $\overline{U}_i$ based on $\mathrm{var}(\overline{U}) = \{\mathrm{var}(W_{0i} - W_i)\}/18$. This yields $\widehat{\mathrm{var}}(\overline{U}) = 0.008$. Further, because $\overline{U}_i$ is the average of 17 U_{ki}'s, it is sensible to assume that $\overline{U}_i$ has a normal distribution. Note that we do not assume normality on U_{0i}. We consider the support of X to be $[0, 1 + 1.65\sqrt{0.008}] = [0, 1.1]$

as a conservative approximation, where 1.65 is the 95% quantile for the standard normal distribution and we factor in that X is non-negative. The data from “Mission China Beijing”, which is considered more reliable, has the maximum value 0.99 after the same scaling. Thus, taking both into account, we treat the support of X to be [0,1] as well.

Based on the preliminary analysis above, we proceed to estimate the mean regression function of asthma ERVs conditional on PM2.5, i.e., $E(Y_i \mid X_i = x)$, and the pdf of PM2.5, i.e., $f_{X0}(x)$, using the B-spline-assisted semiparametric/MLE methods in Sections 2.1 and 2.2. Taking into account that the PM2.5 levels are potentially temporally correlated, instead of using the results in Sections 3, we implemented 100 block bootstraps (Hall, 1985) to estimate the asymptotic variances of the resulting estimators, which is a standard method to estimate asymptotic variances for correlated data. We also compared the B-spline semiparametric regression and density estimators with the deconvolution regression and density estimators. In implementing the B-spline approximation, we used two and three equally spaced knots respectively, and in implementing the deconvolution methods, we used bandwidth 0.05. The number of knots is chosen based on the simulation studies in Section 4.

As a side note, the bandwidth 0.05 is the least we need in order to achieve stable result for the deconvolution estimator for this data set. In fact, in selecting the bandwidth, we implemented both the crossvalidation (Stefanski and Carroll, 1990) and the plug-in (Delaigle and Gijbels, 2002) methods. Both procedures led to very small bandwidths that induce large numerical errors. For this reason, we increased the bandwidth to 0.05.

The upper panel in Figure 2 provides the estimated mean of Y as a function of X. The B-spline semiparametric estimator shows a fluctuating pattern in the range from 0 to 200. In the range of PM2.5 concentration larger than 200 (about 11.3% of the observations), it shows clearly an increasing trend, which agrees with the conclusion in Fan et al. (2016) that the exposure to high PM2.5 has an adverse effect on asthma ERV rate. In contrast, the relation from the deconvolution estimator is similar to that of the local linear regression estimator ignoring the measurement errors, and it is unable to detect the increasing trend of the asthma ERVs as the PM2.5 level increases.

The lower panel in Figure 2 shows the estimated pdfs based on the B-spline and deconvolution methods. Compared with the kernel estimator in the same plots which ignores the measurement errors, the B-spline method shows more difference than the deconvolution estimator. In fact, the noise-to-signal ratio is more than $\text{var}(\bar{U}_i)/\text{var}(X_i) = 0.25$ (the measurement reliability is 0.8), hence measurement error issue is likely not ignorable.

6. Discussion

Motivated by the asthma study in Beijing, we have developed a B-spline based semiparametric estimator for nonparametric regression mean function estimation, and a B-spline based method for nonparametric pdf estimation, when the covariates are measured with error. The performance of both procedures are superior to the widely used deconvolution methods, in terms of both their faster convergence rate and smaller estimation errors. Viewing its satisfactory performance in analyzing the asthma data, the B-spline estimation was helpful in resolving the disputes in the environmental studies attributing to the measurement errors.

For convenience, we have assumed a known distribution of U throughout the text. This assumption is usually imposed in the literature for identifiability reason. In practice, the distribution of U is often estimated based on additional information and preprocessing, and then the estimated $f_U(\cdot)$ is used in the main analysis. Typical procedures include using multiple measurements, validation data or instrumental variables. In fact, provided the problem is identifiable, we can allow unknown parameters in $f_U(\cdot)$ as well, and we estimate these parameters together with the B-spline coefficients by concatenating them. We also assumed the distribution of ϵ to be known for convenience. The same procedure will work if we allow unknown parameters in $f_\epsilon(\cdot)$, by concatenating the B-spline coefficients to the parameters in $f_\epsilon(\cdot)$. Furthermore, we can also handle the situation where $f_\epsilon(\cdot)$ is only assumed to be a distribution with mean zero, without any specific parametric form. To this end, technically we only need to replace the method of Tsiatis and Ma (2004) by the method of Garcia and Ma (2017), after the spline approximation. Finally, the independence between U and X is also not necessary for our method, although it is critically important for the deconvolution method. To accommodate the dependence between X and U in our procedure, all we need to do is to replace $f_U(w-x)$ with $f_{W|X}(w,x)$ and the whole procedure goes through as long as the model is still identifiable. Considering that the assumptions of known $f_U(\cdot)$, $f_\epsilon(\cdot)$ and independence between X and U are often violated in practice, the flexibility of the our method provides robustness against these assumptions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work is partially supported by grants from the National Science Foundation and the National Institutes of Health including K training award 1K25AG071840-01A1.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecosta.2023.05.001

References

Berry, S.M., Carroll, R.J., Ruppert, D., 2002. Bayesian smoothing and regression splines for measurement error problems. Journal of the American Statistical Association 97, 160–169.

Biggs, R., Carpenter, S.R., Brock, W.A., 2009. Spurious certainty: how ignoring measurement error and environmental heterogeneity may contribute to environmental controversies. BioScience 59, 65–76.

Carroll, R.J., Hall, P., 1988. Optimal rates of convergence for deconvolving a density. Journal of the American Statistical Association 83, 1184–1186.

Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. Measurement error in nonlinear models: a modern perspective. CRC press.

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

Eckert, R., Carroll, R., Wang, N., 1997. Transformations to additivity in measurement error models. Biometrics 53, 262–272.

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

Fan, J., Li, S., Fan, C., Bai, Z., Yang, K., 2016. The impact of pm2.5 on asthma emergency department visits: a systematic review and meta-analysis. Environmental Science and Pollution Research 23, 843–850.

Fan, J., Truong, Y.K., 1993. Nonparametric regression with errors in variables. Annals of Statistics 21, 1900–1925.

Garcia, T., Ma, Y., 2017. Simultaneous treatment of unspecified heteroskedastic model error distribution and mismeasured covariates for restricted moment models. Journal of Econometrics 200, 194–206.

Hall, P., 1985. Resampling a coverage pattern. Stochastic processes and their applications 20 (2), 231–246.

Hall, P., Qiu, P., 2005. Discrete-transform approach to deconvolution problems. Biometrika 92, 135–148.

Kneip, A., Simar, L., Van Keilegom, I., 2015. Frontier estimation in the presence of measurement error with unknown variance. Journal of Econometrics 184 (2), 379–393.

Liu, M.C., Taylor, R.L., 1989. A consistent nonparametric density estimator for the deconvolution problem. Canadian Journal of Statistics 17, 427–438.

Masri, R., Redner, R.A., 2005. Convergence rates for uniform b-spline density estimators on bounded and semi-infinite domains. Nonparametric Statistics 17, 555–582.

Mission-China, 2016. Accessed:09-30. http://www.stateair.net/web/historical/1/1.html.

Sarkar, A., Mallick, B.K., Carroll, R.J., 2014. Bayesian semiparametric regression in the presence of conditionally heteroscedastic measurement and regression errors. Biometrics 70 (4), 823–834.

Staudenmayer, J., Ruppert, D., Buonaccorsi, J.P., 2008. Density estimation in the presence of heteroscedastic measurement error. Journal of the American Statistical Association 103, 726–736.

Stefanski, L.A., Carroll, R.J., 1990. Deconvoluting kernel density estimators. Statistics 21, 169–184.

Tsiatis, A.A., Ma, Y., 2004. Locally efficient semiparametric estimators for functional measurement error models. Biometrika 91, 835–848.

Wang, L., Yang, L., 2009. Spline estimation of single-index models. Statistica Sinica 19 (2), 765–783.

Xu, Q., Li, X., Wang, S., Wang, C., Huang, F., Gao, Q., Wu, L., Tao, L., Guo, J., Wang, W., et al., 2016. Fine particulate air pollution and hospital emergency room visits for respiratory disease in urban areas in beijing, china, in 2013. PloS One 11, e0153099.

Zhang, C.H., 1990. Fourier methods for estimating mixing densities and distributions. Annals of Statistics 18, 806–830.