Estimation of nonparametric regression models with a mixture of Berkson and classical errors

Zanhua Yin[a], Wei Gao[a,*], Man-Lai Tang[b], Guo-Liang Tian[c]

[a]*School of Mathematics and Statistics, Northeast Normal University, Changchun, P. R. China*
[b]*Hong Kong Baptist University, Hong Kong, P. R. China*
[c]*The University of Hong Kong, Hong Kong, P. R. China*

Abstract

We consider the estimation of nonparametric regression models with the explanatory variable being measured with Berkson errors or with a mixture of Berkson and classical errors. By constructing a compact operator, the regression function is the solution of an ill-posed inverse problem, and we propose an estimation procedure based on Tikhonov regularization. Under mild conditions, the convergence rate of proposed estimator is derived. The finite-sample properties of the estimator are investigated through simulation studies.

Keywords: Berkson error, Classical error, Deconvolution, Ill-posed problem, Kernel methods, Non-parametric regression, Tikhonov regularization.

1. Introduction

In traditional non-parametric regression model analysis, one is interested in the following model

$$Y = g(X) + \epsilon, \tag{1}$$

where $g(\cdot)$ is a smooth function which we wish to estimate and ϵ is a noise variable with $E(\epsilon \mid X) = 0$ and $E(\epsilon^2|X) < \infty$. Here, the explanatory variable X is usually assumed to be directly observable without errors. Both the direct observation and error-free assumptions are however seldom true in many studies. For the violation of the error-free assumption, Armstrong (1998) considered an environmental study which investigated the relation of mean exposure to lead up to age 10 (denoted as X) with intelligence quotient (IQ) among 10-year-old children (denoted as Y) living in the neighborhood of a lead smelter. Each child had one measurement made of blood lead (denoted as W), at a random time during their life. The blood lead measurement (i.e., W) became an approximate measure of mean blood lead over life (X). However, if we were able to make many replicate measurements (at different random time points), the mean would be a good indicator of lifetime exposure. In other words, the measurements of X are subject to errors and W is a perturbation of X. In this case, which is known as the *classical error model*, we observe an i.i.d. sample $(Y_i, W_i), i = 1, \ldots, n$, from

$$Y_i = g(X_i) + \epsilon_i, \qquad W_i = X_i + \varepsilon_i, \tag{2}$$

where $(\epsilon_i, X_i, \varepsilon_i)$, $i = 1, \ldots, n$, are mutually independent and ε represents the classical measurement error variable. The classical error model (2) has attracted considerable attention in the literature, and is by now well understood. See Fan and Truong (1993), Carroll *et al.* (1999), Delaigle and Meister (2007), Delaigle *et al.* (2008) and Delaigle *et al.* (2009). For additional references for non-parametric regression models with classical errors, ones may consult Carroll *et al.* (2006) and the references therein.

In many studies, it is however too costly or impossible to measure X exactly or directly. Instead, a proxy W of X is measured. For the violation of the direct observation assumption, Armstrong (1998) modified the aforementioned

*Corresponding author
Email addresses: yinzh226@nenu.edu.cn (Zanhua Yin), gaow@nenu.edu.cn (Wei Gao)

environmental study in which the children's place of residence at age 10 (assumed known exactly) were classified into three groups by proximity to the smelter - close, medium, far. Random blood lead samples, collected as describe in the aforementioned design, were averaged for each group (denoted as W), and this group mean used as a proxy for lifetime exposure for each child in the group. Here, the same approximate exposure (proxy) is used for all subjects in the same group, and true exposures, although unknown, may be assumed to vary randomly about the proxy. This is the well-known *Berkson error model*. In other words, the explanatory variable X are not directly observable and measurements on its surrogates W are available instead. X is then a perturbation of W. In this case, we observe an i.i.d. sample $(Y_i, W_i), i = 1, \ldots, n$, generated by

$$Y_i = g(X_i) + \epsilon_i, \qquad X_i = W_i + \delta_i, \tag{3}$$

where $(\epsilon_i, W_i, \delta_i), i = 1, 2, \ldots, n$, are mutually independent and δ represents the Berkson measurement error variable. The Berkson error model was first considered by Berkson (1950). Recently, several methods such as least squares estimation method (Huwang and Huang, 2000), minimum distance estimation method (Wang, 2003, 2004) and regression calibration method (Carroll *et al*., 2006) have been studied in the literature. However, all these work mostly focus on specifying some parametric or semiparametric relationship between the explanatory variable and response, and there is little work on nonparametric estimation in the setting of model (3) (e.g. Delaigle *et al*., 2006).

In the Berkson model (3), it is usually assumed that the observable variable W is measured with perfect accuracy. However, this may not be true due to inaccuracy of the measurement process in some situations. In such cases, the measurements of the proxy W are subject to errors, and data can be contaminated by a mixture of Berkson and classical errors. To be specific, we observe a random sample of independent pairs (Y_i, V_i), $i = 1, \ldots, n$, generated by

$$Y_i = g(X_i) + \epsilon_i, \qquad X_i = W_i + \delta_i, \qquad V_i = W_i + \varepsilon_i, \tag{4}$$

where the random variables $W_i \sim f_W$, the berkson errors $\delta_i \sim f_\delta$, the classical errors $\varepsilon_i \sim f_\varepsilon$ and ϵ_i are mutually independent, and the respective error densities f_δ and f_ε are assumed to be known. Obversely, the classical error model (2) and the berkson error model (3) are both included in the mixture model (4). Due to its potentially wide applications, statistical procedures for analyzing model (4) has received more attention recently. For instance, a regression calibration approach was proposed by Reeves *et al*. (1998) and Schafer *et al*. (2002) in a parametric context of random exposure. Mallick *et al*. (2002) considered a Bayesian approach for a semi-parametric regression function. The objective of this paper is to give a nonparametric estimator of the regression function g for the data (Y_i, V_i), $i = 1, \ldots, n$, generated by the mixture model (4).

When both types of errors are present, nonparametric estimation of g may not be an easy task since, as explained in Section 2, the relation that identifies g is a Fredholm integral equation of the first kind, i.e.,

$$m(w) = \int g(x) f_\delta(x - w)\, \mathrm{d}x, \tag{5}$$

where $m(w) = E[Y \mid W = w]$. The function g is the solution of equation (5), which may lead to an ill-posed problem. Deconvolution is known to be difficult. Carroll *et al.* (2007) proposed a nonparametric estimator using kernel deconvolution techniques, but its calculation is rather complicated since it requires the calculation of a double deconvolution integral and the use of several smoothing parameters. Delaigle and Meister (2011) construct a simple nonparametric estimator based on estimators of the derivatives of $m(\cdot)$, but it need assume the characteristic function of error density f_δ is the inverse of a polynomial (or the error density f_δ is symmetric). In this paper, we propose a new nonparametric estimation approach which consists of two major steps. First, we construct a compact operator T and therefore admits a countable infinite number of eigenvalues and eigenfunctions. Second, using the Tikhonov regularization, we develop an estimator of g based on the operator T and a deconvolution kernel estimator of $m(\cdot)$. Under mild conditions, the convergence rates of the proposed estimator are derived.

This paper is organized as follows. In Section 2, we propose an estimator for the regression function $g(\cdot)$. In Section 3, we derive the convergence rates of our estimator under some regularity conditions. In Section 4, we discuss the computation for the proposed estimator. Section 5 presents some numerical results from simulation studies. A brief discussion is given in Section 6.

2. Methodology

Let $(Y_1, V_1), \ldots, (Y_n, V_n)$ be a random sample from model (4). We define an operator T as follows:

$$(Tg)(w) = \int g(x) f_\delta(x - w)\,\mathrm{d}x,$$

where f_δ is the density of the Berkson error δ. Here and below, unqualified integrals are taken over the whole real line. Let $L^2(\mathbb{R}^2)$ denotes the space of the square-integrable functions with respect to Lebesgue measure on $\mathbb{R}^2$. If the function $f_\delta(x - w) \in L^2(\mathbb{R}^2)$, the operator T is a Hilbert–Schmidt operator. However, this may not be true in some situations (for example, f_δ is a Laplace density or normal density). Hence, the main idea of this paper is to reconstruct T and make it compact.

To be specific, we choose two arbitrary functions $\omega_X(x)$ and $\omega_W(w)$ that satisfy

Condition A:

(A1) Both $\omega_X(x)$ and $\omega_W(w)$ are continuous and bounded density functions, and $\omega_X(x) > 0$; and
(A2) $\int\int f_\delta^2(x - w)\omega_W(w)/\omega_X(x)\,\mathrm{d}x\,\mathrm{d}w < \infty$.

Define

$$L^2(\mathbb{R}, \omega_X) = \left\{\varphi : \mathbb{R} \to \mathbb{R}, \ \text{s.t.} \ \| \varphi \| = \left(\int \varphi^2(x)\omega_X(x)\,\mathrm{d}x\right)^{1/2} < \infty\right\},$$

and

$$L^2(\mathbb{R}, \omega_W) = \left\{\psi : \mathbb{R} \to \mathbb{R}, \ \text{s.t.} \ \| \psi \| = \left(\int \psi^2(w)\omega_W(w)\,\mathrm{d}w\right)^{1/2} < \infty\right\},$$

where $\| \cdot \|$ denotes the norm in these spaces. We further define the operator $T : L^2(\mathbb{R}, \omega_X) \to L^2(\mathbb{R}, \omega_W)$ as

$$(T\varphi)(w) = \int \varphi(x) t(x, w)\omega_X(x)\,\mathrm{d}x,$$

where $t(x, w) = f_\delta(x - w)\omega_W(w)/[\omega_X(x)\omega_W(w)]$ is called the kernel of the operator T. In fact, Condition (A2) implies that $T\varphi \in L^2(\mathbb{R}, \omega_W)$ for any function $\varphi \in L^2(\mathbb{R}, \omega_X)$, and is sufficient condition for T to be a Hilbert–Schmidt operator (see Section 3.2). Hence, it is easy to verify that equation (5) is equivalent to the operator equation

$$(Tg)(w) = m(w). \tag{6}$$

According to equation (6), the function g is the solution of a Fredholm integral equation of the first kind, and this inverse problem is known to be ill-posed and needs a regularization method (see Section 3.2). A variety of regulation schemes are available in the literature (see e.g. Kress 1999), but we focus in this paper on the Tikhonov regularized solution.

We define the adjoint operator T^* of T

$$(T^*\psi)(x) = \int \psi(w) t(x, w)\omega_W(w)\,\mathrm{d}w,$$

where $\psi(w) \in L^2(\mathbb{R}, \omega_W)$. Then, the Tikhonov regularized solution is

$$g^\alpha = (\alpha I + T^*T)^{-1} T^* m. \tag{7}$$

where the penalization term $\alpha\,(\alpha > 0)$ is the regularization parameter.

From (7), we see that to estimate g it only need to estimate $m(\cdot)$. Since $m(w) = E(Y \mid W = w)$ and we observe (Y_i, V_i), where $V_i = W_i + \varepsilon_i$, estimating of $m(\cdot)$ is a classical errors-in-variables problem, and thus we can use the deconvolution kernel estimator of Fan and Truong (1993). Let K denote a kernel function, $h > 0$ a bandwidth and

$$K_\varepsilon(x) = \frac{1}{2\pi} \int \exp(-\mathrm{i}tx) \frac{\phi_K(t)}{\phi_\varepsilon(t/h)}\,\mathrm{d}t,$$

where $\phi_K(\cdot)$ is the Fourier transform of the kernel function $K(\cdot)$, $\phi_\varepsilon(\cdot)$ is the characteristic function of the classical error ε. The deconvolution kernel estimator of $m(\cdot)$, derived by Fan and Truong (1993), is defined by

$$\hat{m}(w) = \sum_{i=1}^{n} K_\varepsilon\Big(\frac{w - V_i}{h}\Big) Y_i \Big/ \sum_{i=1}^{n} K_\varepsilon\Big(\frac{w - V_i}{h}\Big). \tag{8}$$

Based on expression (7), we can now define our estimator of g by

$$\hat{g}^\alpha = (\alpha I + T^*T)^{-1} T^* \hat{m}. \tag{9}$$

where $\hat{m}$ is defined by (8). When the variance of ε in model (4) is equal to 0, which reduces to the Berkson error model (3), we observe (Y_i, W_i) directly, and $\hat{m}$ reduces to the classical Nadaraya–Watson estimator.

Example 1. We assume the Berkson error δ has a normal distribution with mean zero and variance σ_δ^2, then

$$f_\delta(x - w) = \phi\Big(\frac{x - w}{\sigma_\delta}\Big),$$

where ϕ denotes the p.d.f. of a standard normal distribution. In this case, to ensure Condition A to be valid, a simple choice for ω_W is $\omega_W(w) = \phi(w)$, and ω_X can be computed as

$$\omega_X(x) = \int f_\delta(x - w)\omega_W(w)\,\mathrm{d}w = \phi\Big(x \Big/ \sqrt{1 + \sigma_\delta^2}\Big).$$

Concerning the kernel of T, we have

$$t(x, w)\omega_X(x) = \phi\Big(\frac{x - w}{\sigma_\delta}\Big), \quad \text{and} \quad t(x, w)\omega_W(w) = \phi\Big(\frac{w - \rho x}{\sigma_\delta \sqrt{\rho}}\Big),$$

where $\rho = 1/(1 + \sigma_\delta^2)$.

Example 2. We assume the Berkson error δ has a Laplace distribution with mean zero and variance $2\lambda^2$, then

$$f_\delta(x - w) = \frac{1}{2\lambda} \exp\Big(-\frac{|\, x - w\,|}{\lambda}\Big).$$

Here, we can choose $\omega_W(w) = 0.5\mathbf{1}\{w \in [-1, 1]\}$ and $\omega_X(x) = 0.5\mathbf{1}\{x \in [-1, 1]\}$ to ensure Condition A to be valid.

3. Theoretical properties

3.1. Convergence rate of deconvolution kernel estimator $\hat{m}$

In this section, we focus on the properties of the estimator $\hat{m}(\cdot)$ defined in (8). For this purpose, we present the following regular conditions which are mild and can be found in Fan and Truong (1993).

Condition B:

(B1) The characteristic function of the classical error distribution $\phi_\varepsilon(\cdot)$ does not vanish;
(B2) The density f_W of W is bounded away from 0, and has bounded kth derivative;
(B3) The kernel $K(\cdot)$ is a square integrable k-order bounded symmetric kernel such that $\int |\, x^k K(x)\,|\,\mathrm{d}x < \infty$; and
(B4) The function $m(\cdot)$ has a continuous kth derivative.

The convergence rates of $\hat{m}(\cdot)$ depend on the smoothness of the function $m(\cdot)$ and the regularity conditions on the marginal distribution and the kernel function. They also depend on the tail behaviour of $\phi_\varepsilon(t)$, as Fan and Truong (1993) discussed, which can be classified into the following:

1. Super smooth of order β is

$$d_0|t|^{\beta_0} \exp(-|t|^\beta/\gamma) \leq |\, \phi_\varepsilon(t)\,| \leq d_1|t|^{\beta_1} \exp(-|t|^\beta/\gamma) \qquad as \quad t \to \infty, \tag{10}$$

where d_0, d_1, γ and β are positive constants and β_0 and β_1 are constants.

2. Ordinary smooth of order β is

$$d_0|t|^{-\beta} \leq |\phi_\varepsilon(t)| \leq d_1|t|^{-\beta} \qquad as \quad t \to \infty, \tag{11}$$

for positive constants d_0, d_1 and β.

The following result under super smooth error case is obtained by applying Theorem 2 in Fan and Truong (1993).

Proposition 1. *Suppose that Conditions A and B hold and that the first half inequality of (10) is satisfied. Assume that $\phi_K(t)$ is supported on $[-1, 1]$. Then, for bandwidth $h = d(\log n)^{-1/\beta}$ with $d > (2/\gamma)^{-1/\beta}$, we have*

$$\int [\hat{m}(w) - m(w)]^2 \omega_W(w)\, \mathrm{d}w = O_P[(\log n)^{-2k/\beta}].$$

The next result under ordinary smooth error case is obtained by applying Theorem 4 in Fan and Truong (1993).

Proposition 2. *Suppose that Conditions A and B hold and that the inequality of (11) is satisfied. Assume that $\int |t|^{\beta+1}(|\phi_K(t)| + |\phi_K^{'}(t)|)\, \mathrm{d}t < \infty$ and $\int |t^{\beta+1}\phi_K(t)|^2\, \mathrm{d}t < \infty$. Then, for bandwidth $h = O(n^{-1/(2k+2\beta+1)})$, we have*

$$\int [\hat{m}(w) - m(w)]^2 \omega_W(w)\, \mathrm{d}w = O_P[n^{-2k/(2k+2\beta+1)}].$$

3.2. Convergence rate of $\hat{g}^\alpha$

The main objective of this section is to derive the statistical properties of the estimator $\hat{g}^\alpha(\cdot)$ from the properties of $\hat{m}(\cdot)$, T and T^*. Following Section 2, Condition A2 amounts to assume that T and T^* are Hilbert–Schmidt operators, and is a sufficient condition of compactness of T, T^*, TT^* and T^*T (see Carrasco *et al*., 2007, Theorem 2.34). As a result of compactness, there exists a singular values decomposition, and the singular values of T are the square roots of the eigenvalues of the nonnegative self-adjoint compact operator T^*T. Let λ_j, $j \geq 0$ be the sequence of the nonzero singular values of T and the two orthonormal sequences φ_j of $L^2(\mathbb{R}, \omega_X)$, and ψ_j of $L^2(\mathbb{R}, \omega_W)$ such that (see Kress 1999, Theorem 15.16):

$$T\varphi_j = \lambda_j\psi_j, \;\; T^*\psi_j = \lambda_j\varphi_j; \;\; T^*T\varphi_j = \lambda_j^2\varphi_j, \;\; TT^*\psi_j = \lambda_j^2\psi_j, \;\; \text{for } j \geq 0.$$

Since f_δ, ω_X and ω_W are given, we can consider the eigenvalues and eigenfunctions as known.

By Picard theorem (see, Kress, 1999), the solution (6) can be represented from the singular value decomposition of T as

$$g = \sum_{j=1}^{\infty} \frac{\langle m, \psi_j\rangle}{\lambda_j}\varphi_j, \;\; \text{with } \langle m, \psi_j\rangle = \int m(w)\psi_j(w)\omega_W(w)\, \mathrm{d}w.$$

Here and blow, we denote by $\langle\cdot,\cdot\rangle$ the scalar product in $L^2(\mathbb{R}, \omega_X)$ or $L^2(\mathbb{R}, \omega_W)$. Above formula clearly demonstrates the ill-posed nature of the equation (6). If we perturb m by $m^\tau = m + \tau\psi_j$, we obtain the solution $g^\tau = g + \tau\varphi_j/\lambda_j$ which can be infinitely far from the true solution g due to the fact that the singular values tend to zero. Looking for one regularized solution is a classical way to overcome this problem. In this paper, we consider the Tikhonov regularized solution g^α at (7).

Note that the regularization bias is

$$\begin{aligned} g - g^\alpha &= [I - (\alpha I + T^*T)^{-1}T^*T]g \\ &= \sum_{j=1}^{\infty} \frac{\alpha}{\alpha + \lambda_j^2}\langle g, \varphi_j\rangle\varphi_j. \end{aligned}$$

In order to control the speed of convergence to zero of the regularization bias $g - g^\alpha$, we introduce the following regularity space Φ_γ for $\gamma > 0$:

$$\Phi_\gamma = \left\{\varphi \in L^2(\mathbb{R}, \omega_X) \;\text{ s.t. }\; \sum_{j=1}^{\infty} \frac{\langle\varphi, \varphi_j\rangle}{\lambda_j^{2\gamma}} < +\infty\right\}.$$

We then obtain the following result by applying Proposition 3.11 in Carrasco *et al*. (2007).

Proposition 3. *Under Condition A, if $g \in \Phi_\gamma$ for $0 < \gamma \leq 2$, we have*

$$\int [g(x) - g^\alpha(x)]^2 \omega_X(x)\,\mathrm{d}x = O(\alpha^\gamma).$$

Therefore, when the regularization parameter α is pushed towards zero, the smoother the function g of interest (i.e. $g \in \Phi_\gamma$ for larger γ) is , the faster the rate of convergence to zero of the regularization bias will be. Now we state the main result of the paper.

Theorem 3.1. *Suppose the conditions of Proposition 1 hold. If $g \in \Phi_\gamma$ for $0 < \gamma \leq 2$, then we have*

$$\int [\hat{g}^\alpha(x) - g(x)]^2 \omega_X(x)\,\mathrm{d}x = O_P\Big[\frac{1}{\alpha^2} \times (\log n)^{-2k/\beta} + \alpha^\gamma\Big].$$

In particular, when $\alpha = O[(\log n)^{-2k/(2\beta+\gamma\beta)}]$, we have

$$\int [\hat{g}^\alpha(x) - g(x)]^2 \omega_X(x)\,\mathrm{d}x = O_P\Big[(\log n)^{-2k\gamma/(2\beta+\gamma\beta)}].$$

Proof: Notice that $g^\alpha = (\alpha I + T^*T)^{-1}T^*m$, then we have

$$\hat{g}^\alpha - g = \hat{g}^\alpha - g^\alpha + g^\alpha - g.$$

By Proposition 3:

$$\int [g(x) - g^\alpha(x)]^2 \omega_X(x)\,\mathrm{d}x = O(\alpha^\gamma).$$

To asses the order of $\hat{g}^\alpha - g^\alpha$, it is worth rewriting it as:

$$\begin{aligned}
\hat{g}^\alpha - g^\alpha &= (\alpha I + T^*T)^{-1}(T^*\hat{m} - T^*m) \\
&= \sum_{j=1}^{\infty} \frac{1}{\alpha + \lambda_j^2} \langle \hat{m} - m, T\varphi_j \rangle \varphi_j \\
&= \sum_{j=1}^{\infty} \frac{\lambda_j}{\alpha + \lambda_j^2} \langle \hat{m} - m, \psi_j \rangle \varphi_j.
\end{aligned}$$

Since $\{\varphi_j\}$ is orthonormal sequence on $L^2(\mathbb{R}, \omega_X)$, we have

$$\int [\hat{g}^\alpha(x) - g^\alpha(x)]^2 \omega_X(x)\,\mathrm{d}x = \sum_{j=1}^{\infty} \frac{\lambda_j^2}{(\alpha + \lambda_j^2)^2} \langle \hat{m} - m, \psi_j \rangle^2.$$

From the properties of scalar product, we have

$$\begin{aligned}
\langle \hat{m} - m, \psi_j \rangle^2 &= \left\{ \int [\hat{m}(w) - m(w)] \psi_j(w) \omega_W(w)\,\mathrm{d}w \right\}^2 \\
&\leq \int [\hat{m}(w) - m(w)]^2 \omega_W(w)\,\mathrm{d}w \int \psi_j^2(w) \omega_W(w)\,\mathrm{d}w.
\end{aligned}$$

Thus, by Proposition 1, we have $\langle \hat{m} - m, \psi_j \rangle^2 = O_P[(\log n)^{-2k/\beta}]$. Since $\sum_{j=1}^{\infty} \lambda_j^2/(\alpha + \lambda_j^2)^2 = O(1/\alpha^2)$ (see e.g. Groetsch 1984), then we have

$$\int [\hat{g}^\alpha(x) - g^\alpha(x)]^2 \omega_X(x)\,\mathrm{d}x = O_P\Big[\frac{1}{\alpha^2} \times (\log n)^{-2k/\beta}\Big].$$

The desired result follows immediately. □

From Theorem 3.1, when the characteristic function of the classical error satisfies inequality (10), we obtain a slower logarithmic rate which corresponds to the usual rate attained in nonparametric deconvolution from super smooth error distribution. Similarly, when the characteristic function of the classical error satisfies inequality (11), we have the following result.

Theorem 3.2. *Suppose the conditions of Proposition 2 hold. If $g \in \Phi_\gamma$ for $0 < \gamma \leq 2$, then we have*

$$\int [\hat{g}^\alpha(x) - g(x)]^2 \omega_X(x)\,\mathrm{d}x = O_P\Big[\frac{1}{\alpha^2} n^{-2k/(2k+2\beta+1)} + \alpha^\gamma\Big].$$

In particular, for $\alpha = O[n^{-2k/[(2k+2\beta+1)(\gamma+2)]}]$, we have

$$\int [\hat{g}^\alpha(x) - g(x)]^2 \omega_X(x)\,\mathrm{d}x = O_P\Big[n^{-2k\gamma/[(2k+2\beta+1)(\gamma+2)]}\Big].$$

By some modifications of the proof of Theorem 3.1, the proof of Theorem 3.2 is straightforward and is omitted.

4. Computation

In this section, we discuss the computation of the estimator $\hat{g}^\alpha$. This estimator is a solution to the equation:

$$(\alpha I + T^*T)g = T^*\hat{m},$$

or equivalently

$$\hat{g}^\alpha(x) = \sum_{j=1}^{\infty} \frac{1}{\alpha + \lambda_j^2} \langle T^*\hat{m}, \varphi_j\rangle \varphi_j(x). \tag{12}$$

Unfortunately, computing the estimator (12) requires specifying expression of the eigenvalues and eigenfunctions. Blow, we explain the estimate approach of the eigenvalues and eigenfunctions which was adopted by Carrasco and Florens (2009).

Using the importance sampling, T can be estimated by

$$(\hat{T}\varphi)(w) = \frac{1}{B}\sum_{k=1}^{B} \varphi(x_k)\frac{f_\delta(x_k - w)}{\omega_X(x_k)},$$

where $\{x_k\}, k = 1, \ldots, B$, is an i.i.d. sample drawn from ω_X. Similarly, T^* can be approached by

$$(\hat{T}^*\psi)(x) = \frac{1}{B}\sum_{k=1}^{B} \psi(w_k)\frac{f_\delta(x - w_k)}{\omega_X(x)},$$

where $\{w_k\}, k = 1, \ldots, B$, is an i.i.d. sample drawn from ω_W.

Therefore $(T^*T\varphi)(x)$ can be approximated by

$$\frac{1}{B}\sum_{k=1}^{B}\Big[\frac{1}{B}\sum_{l=1}^{B} \varphi(x_l)\frac{f_\delta(x_l - w_k)}{\omega_X(x_l)}\Big]\frac{f_\delta(x - w_k)}{\omega_X(x)}.$$

Note that $T^*T\varphi_j = \lambda_j^2\varphi_j$. We can calculate the eigenvalues and eigenfunctions of above operator by solving

$$\frac{1}{B}\sum_{k=1}^{B}\Big[\frac{1}{B}\sum_{l=1}^{B} \varphi_j(x_l)\frac{f_\delta(x_l - w_k)}{\omega_X(x_l)}\Big]\frac{f_\delta(x - w_k)}{\omega_X(x)} = \lambda_j^2\varphi_j(x). \tag{13}$$

Hence $\varphi_j(x)$ is necessarily of the form: $\varphi_j(x) = \sum_{k=1}^{B} \theta_k^j f_\delta(x - w_k)/\omega_X(x)$. Replacing in (13), we see that solving (13) is equivalent to finding the B nonzero eigenvalues $\hat{\lambda}_1^2, \ldots, \hat{\lambda}_B^2$ and eigenvectors $\hat{\theta}^1, \ldots, \hat{\theta}^B$ of a $B \times B$-matrix M with principle element

$$M_{k,l} = \frac{1}{B^2}\sum_{s=1}^{B} \frac{f_\delta(x_s - w_k)f_\delta(x_s - w_l)}{\omega_X^2(x_s)}.$$

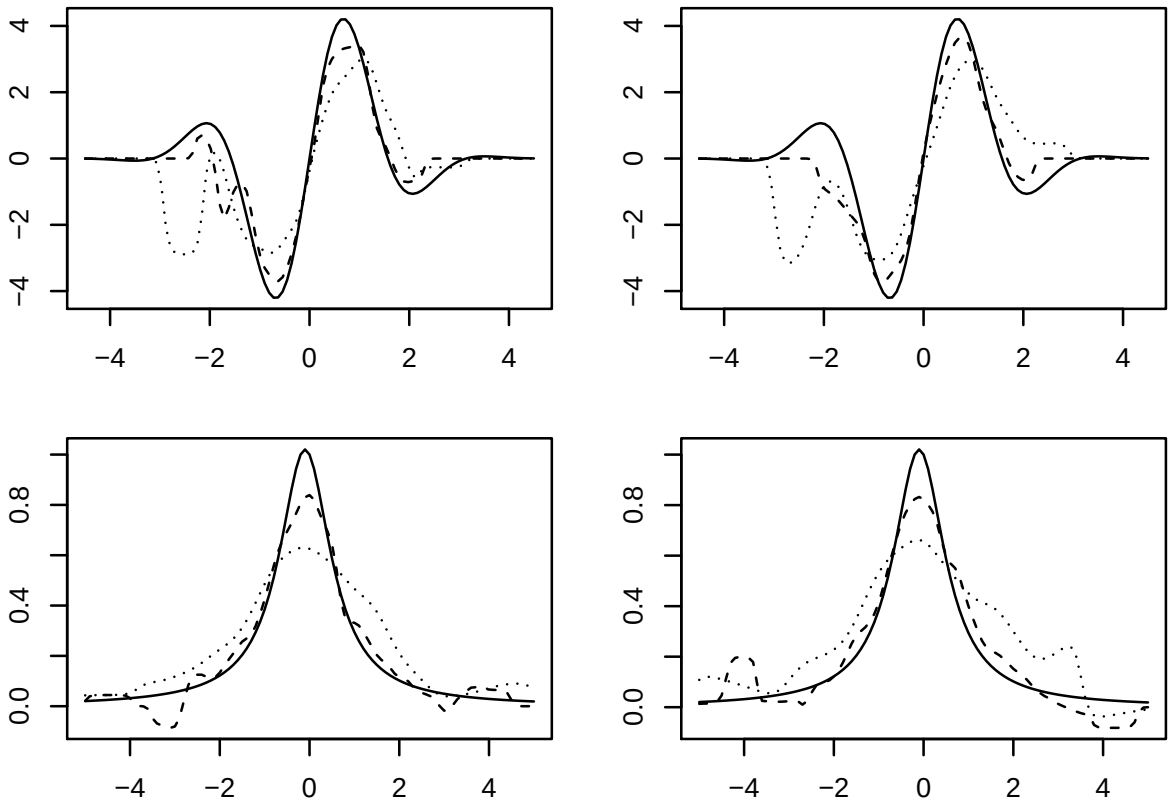

Figure 1: Median curves of 500 estimators of regression functions (a) (top) and (b) (bottom) for samples of size $n = 250$ in the mixture model and $\delta \sim$ Laplace and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ (left) or $\delta \sim N(0, \sigma_\varepsilon^2)$ and $\varepsilon \sim$ Laplace (right) with $(\sigma_\delta^2/\sigma_W^2, \sigma_\varepsilon^2/\sigma_W^2) = (0.1, 0.2)$. The solid, dashed, and dotted curves represent $g(x)$, $\hat{g}^\alpha(x)$ and $\hat{g}_I(x)$ respectively.

Then the estimators of eigenfunctions are

$$\hat{\varphi}_j(x) = \sum_{k=1}^{B} \hat{\theta}_k^j f_\delta(x - w_k)/\omega_X(x), \quad j = 1, \ldots, B,$$

associated with $\hat{\theta}^1, \cdots, \hat{\theta}^B$. The $\hat{\varphi}_j$ need to be orthonormalized.

In addition, the term

$$\langle T^* \hat{m}, \varphi_j \rangle = \int (T^* \hat{m})(x) \varphi_j(x) \omega_X(x) \, \mathrm{d}x$$

can be estimated by

$$\widehat{\langle T^* \hat{m}, \varphi_j \rangle} = \frac{1}{B} \sum_{k=1}^{B} (\hat{T}^* \hat{m})(x_k) \hat{\varphi}_j(x_k).$$

Hence, by expression (12), we obtain $\hat{g}^\alpha$:

$$\hat{g}^\alpha(x) = \sum_{j=1}^{B} \frac{1}{\alpha + \hat{\lambda}_j^2} \widehat{\langle T^* \hat{m}, \varphi_j \rangle} \hat{\varphi}_j(x).$$

5. Simulation studies

In this section, we conduct several simulations to numerically evaluate performances of the proposed estimator. To implement our method (9), the smoothing parameter h and the regularization parameter α should be chosen. In our simulation studies, we use the following two-dimensional cross-validation (CV) approach, selecting (h, α) as

$$(\hat{h}, \hat{\alpha}) = \arg \min_{(h,\alpha)} \sum_{i=1}^{n} \frac{Y_i - \hat{g}^{\alpha}(V_i)}{1 - n^{-1} \sum_{i=1}^{n} \hat{g}^{\alpha}(V_i)}. \tag{14}$$

To compare the proposed estimator with existing estimators, we consider the naive kernel estimator (denoted as $\hat{g}_I$), which is the standard Nadaraya–Watson estimator based on direct data from (Y_i, V_i), $i = 1, \ldots, n$. It should be pointed out that $\hat{g}_I$ can serve as a gold standard in the simulation study, even though it is practically unachievable due to measurement errors. The performance of estimator g^{est} is assessed by using the square root of average square errors (RASE)

$$\text{RASE} = \left\{ \frac{1}{M} \sum_{s=1}^{M} [g^{est}(u_s) - g(u_s)]^2 \right\}^{1/2},$$

where u_s, $s = 1, \ldots, M$, are grid points at which $g^{est}(u_s)$ is evaluated.

We applied our estimator to data from models (3) and (4), where the regression functions g taken from the examples of Carroll *et al.* (2007):

(a) $g(x) = 5 \sin(2x) \exp(-16x^2/50)$, $\epsilon \sim N(0, 0.15)$, $W \sim N(0, 0.5)$ (sinusoidal), and,
(b) $g(x) = (2x^2 + 0.4x + 1)^{-1}$, $\epsilon \sim N(0, 0.01)$, $W \sim N(0, 2)$ (sharp unimodal).

We took the errors δ and ε to be either normal or Laplace distribution with zero mean. Specially, if $\delta \sim Normal$ (or $\delta \sim Laplace$), we chose ω_X and ω_W as in Example 1 (or Example 2). For pure Berkson error model (3), we observed (Y_i, W_i) directly (i.e., $V_i = W_i$), and used the Nadaraya–Watson estimator with a standard normal kernel to calculate $m(\cdot)$. For mixture model (4), we calculated $m(\cdot)$ via (8), and adopt the kernel $K(\cdot)$ corresponding to $\phi_K(t) = (1 - t^2)^8 \mathbf{1}\{t \in [-1, 1]\}$, which is commonly used in deconvolution problems.

In our simulations we consider sample sizes $n = 50$ or 250, and in each case 500 simulated data sets were generated from model (3) or model (4). We calculated the corresponding 500 estimators of the curve g, using our method or using the naive kernel estimator, and reported the corresponding 500 calculated RASEs. To calculate $\hat{g}^{\alpha}$, we selected the parameters (h, α) as in (14). For $\hat{g}_I$, we used the standard normal kernel, and the bandwidth was selected by generalized cross-validation (GCV).

Table 1: The RASE comparison for the estimators $\hat{g}^{\alpha}(x)$ and $\hat{g}_I(x)$. Let $\kappa = (\sigma_{\delta}^2/\sigma_W^2, \sigma_{\varepsilon}^2/\sigma_W^2)$, and simply denote $\delta \sim$ Normal and $\varepsilon \sim$ Laplace by (N, L), and other similar.

			$\hat{g}^{\alpha}(x)$				$\hat{g}_I(x)$			
curve	n	κ	(L, L)	(L, N)	(N, L)	(N, N)	(L, L)	(L, N)	(N, L)	(N, N)
(a)	50	(0.1,0.1)	0.9306	0.9941	1.1415	1.0315	1.1197	1.1228	1.1162	1.2631
		(0.1,0.2)	0.9587	1.1225	1.1913	1.2179	1.4099	1.4111	1.4145	1.5519
		(0.1,0.3)	1.0272	1.1913	1.5425	1.4690	1.7224	1.6978	1.6255	1.7025
	250	(0.1,0.1)	0.8159	0.8205	0.8463	0.9596	0.9108	1.0534	1.0590	0.9896
		(0.1,0.2)	0.8368	0.9988	1.0096	1.1406	1.1072	1.3749	1.3320	1.3256
		(0.1,0.3)	0.8765	1.0699	1.2499	1.2632	1.4324	1.6040	1.5587	1.5086
(b)	50	(0.1,0.1)	0.1020	0.1153	0.1192	0.1300	0.1162	0.1187	0.1400	0.1676
		(0.1,0.2)	0.1183	0.1269	0.1285	0.1637	0.1453	0.1463	0.1697	0.1718
		(0.1,0.3)	0.1261	0.1356	0.1470	0.1726	0.1643	0.1817	0.1884	0.2216
	250	(0.1,0.1)	0.1010	0.1030	0.1175	0.1100	0.1105	0.1140	0.1187	0.1179
		(0.1,0.2)	0.1153	0.1237	0.1183	0.1378	0.1375	0.1386	0.1323	0.1428
		(0.1,0.3)	0.1217	0.1319	0.1364	0.1411	0.1600	0.1738	0.1622	0.1661

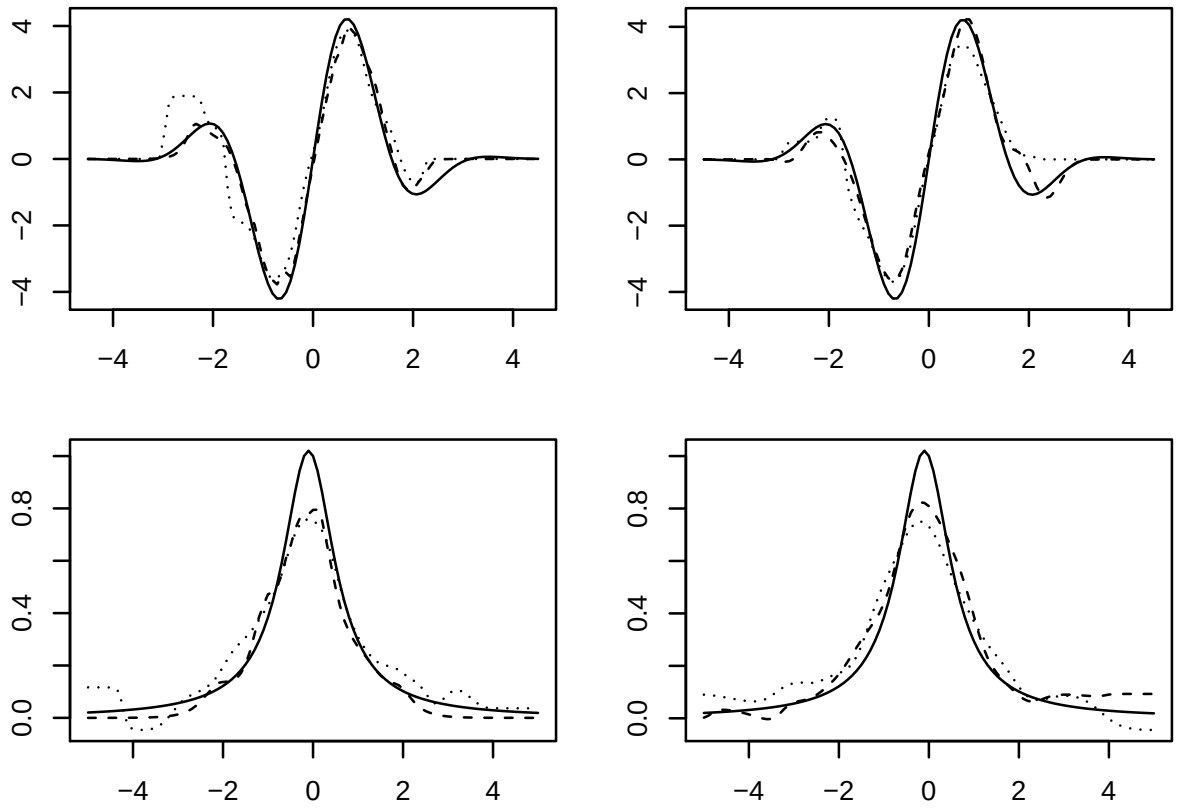

Figure 2: Median curves of 500 estimators of regression functions (a) (top) and (b) (bottom) for samples of size $n = 250$ in the pure Berkson model and $\delta \sim$ Laplace (left) or $\delta \sim N(0, \sigma_\varepsilon^2)$ (right) with $\sigma_\delta^2 = 0.1 \times \sigma_W^2$. The solid, dashed, and dotted curves represent $g(x)$, $\hat{g}^\alpha(x)$ and $\hat{g}_I(x)$ respectively.

Figure 1 shows the regression function curve $g(x)$, the curves of the medians of 500 estimates $\hat{g}^\alpha(x)$ and $\hat{g}_I(x)$ under different settings of δ and ε for sample size $n = 250$, in the two examples (a) and (b) respectively. From this figure, we see clearly that the proposed estimator $\hat{g}^\alpha(x)$ and smoothing parameter selection method appeared to perform very well for the test functions considered in this study. In comparison, $\hat{g}_I(x)$ clearly targeted the wrong curve as we expected.

Table 1 summarizes the results shown in Figure 1 numerically. The estimated RASE which were evaluated at 101 grid points of x are presented. Our results show that the estimator $\hat{g}^\alpha(x)$ worked better than the naive estimator $\hat{g}_I(x)$ in all cases. Also, as the sample size increases, the quality of the estimator has a significant improvement (i.e., the corresponding RASEs decrease). For any nonparametric method in measurement error regression problem, the quality of the estimator also depends on the discrepancy of the observed sample. That is, the performance of the estimator depends on the variances of measurement error. Here, we compared the results for different variance ratios $(\sigma_\delta^2/\sigma_W^2, \sigma_\varepsilon^2/\sigma_W^2)$. It is noteworthy that the effect of the variances on the estimator performance was obvious.

Figure 2 shows the median curves of 500 estimators of regression functions (a) and (b) for samples of size $n = 250$ in the pure Berkson model, when $\delta \sim$ Laplace or $\delta \sim N(0, \sigma_\varepsilon^2)$ with $\sigma_\delta^2 = 0.1 \times \sigma_W^2$. As expected, our proposed estimator substantially outperformed the estimator that completely ignores any measurement errors. Our results show that our proposed estimator also works well in the pure Berkson model.

6. Discussion

In this paper, we propose a new method for estimating non-parametric regression models with the explanatory variable being measured with pure Berkson errors or with a mixture of Berkson and classical errors. We start by

deriving the conditional expectation of unknown objective regression function given the proxy variable that help us obtaining a Fredholm integral equation of the first kind. So the regression function is the solution of an ill-posed problem and we propose an estimator based on Tikhonov regularization. The difficulty with our approach comes from the fact that how to choose two available density functions $\omega_X(x)$ and $\omega_W(w)$ which are able to construct a compact operator T. This is of future research interest.

References

[1] Armstrong, B.G. (1998) Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup. Environ. Med.,* **55,** 651–656.

[2] Berkson, J. (1950) Are there two regression problems? *J. Am. Statist. Ass.,* **45,** 164–180.

[3] Carrasco, M. and Florens, J.P. (2009) Spectral method for deconvolving a density. *forthcoming Econometric Theory*.

[4] Carrasco, M., Florens, J.P. and Renault, E. (2007) *Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regulation*, Handbook of Econometrics, Elsevier, North Holland, 5633–5751.

[5] Carroll, R.J., Delaigle, A. and Hall, P. (2007) Nonparametric regression estimation from data contaminated by a mixture of Berkson and classical errors. *J. R. Statist. Soc. B,* **69,** 859–878.

[6] Carroll, R.J., Maca, J.D. and Ruppert, D. (1999) Nonparametric regression in the presence of measurement error. *Biometrika,* **86,** 541–554.

[7] Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006) *Measurement Error in Nonlinear Models*, second edition. Chapman and Hall CRC Press, Boca Raton.

[8] Delaigle, A., Fan, J. and Carroll, R.J. (2009) A Design-adaptive Local Polynomial Estimator for the Errors-in-Variables Problem. *J. Am. Statist. Ass.,* **104,** 348–359.

[9] Delaigle, A., Hall, P. and Meister, A. (2008) On deconvolution with repeated measurements. *Ann. Statist.,* **36,** 665–685.

[10] Delaigle, A., Hall, P. and Qiu, P. (2006) Nonparametric methods for solving the Berkson errors-in-variables problem. *J. R. Statist. Soc. B ,* **68,** 201–220.

[11] Delaigle, A., and Meister, A. (2007) Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *J. Am. Statist. Ass.,* **102,** 1416–1426.

[12] Delaigle, A., and Meister, A. (2011) Rate-optimal nonparametric estimation in classical and Berkson errors-in-variables problems. *J. Statist. Pla. Inf.,* **141,** 102–114.

[13] Fan, J. and Truong, Y.K. (1993) Nonparametric regression with errors in variables. *Ann. Statist.,* **21,** 1900–1925.

[14] Gössl, C. and Küchenhoff, H. (2001) Bayesian analysis of logistic regression with an unknown change point and covariate measurement error. *Statist. Med.,* **20,** 3109–3121.

[15] Groetsch, C. (1984) *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, London.

[16] Huwang, L. and Huang, H.Y.S. (2000) On errors-in-variables in polynomial regression - Berkson case. *Statist. Sin,* **10,** 923–936.

[17] Kress, R. (1999) *Linear Integral Equations*, Springer.

[18] Küchenhoff, H. and Carroll, R.J. (1997) Segmented regression with errors in predictors: semi-parametric and parametric methods. *Statist. Med.,* **16,** 169–188.

[19] Mallick, B., Hoffman, F.O. and Carroll, R.J. (2002) Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada test site. *Biometrics,* **58,** 13–20.

[20] Reeves, G.K., Cox, D.R., Darby, S.C. and Whitley, E. (1998) Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statist. Med.,* **17,** 2157–2177.

[21] Schafer, M., Mullhaupt, B. and Clavien, P.A. (2002) Evidence-based pancreatic head resection for pancreatic cancer and chronic pancreatitis. *Ann. Surg.,* **236,** 137–148.

[22] Ulm, K. (1991) A statistical method for assessing a threshold in epidemiological studies. *Statist. Med.,* **10,** 341–349.

[23] Wang, L. (2003) Estimation of nonlinear Berkson-type measurement error models. *Statist. Sin,* **13,** 1201–1210.

[24] Wang, L. (2004) Estimation of nonlinear models with Berkson measurement errors. *Ann. Statist.,* **32,** 2559–2579.