Statistics and Probability Letters 84 (2014) 113–120

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Robust errors-in-variables linear regression via Laplace distribution

Jianhong Shi [a], Kun Chen [b], Weixing Song [b,*]

[a] School of Mathematics and Computer Science, Shanxi Normal University, Linfen, 041000, China

[b] Department of Statistics, Kansas State University, Manhattan, KS, 66503, United States

ARTICLE INFO

Article history:
Received 11 July 2013
Received in revised form 27 September 2013
Accepted 28 September 2013
Available online 8 October 2013

MSC:
primary 62F35
secondary 62F10

Keywords:
Least absolute deviation
EM algorithm
Mixture regression model
Normal mixture
Laplace distribution

ABSTRACT

Robust estimation procedures for linear and mixture linear errors-in-variables regression models are proposed based on the relationship between the least absolute deviation criterion and maximum likelihood estimation in a Laplace distribution. The finite sample performance of the proposed procedures is evaluated by simulation studies.

1. Introduction

Developing various robust estimation procedures in linear regression models has been a long-lasting research topic in statistics when data are contaminated with outliers. See Rousseeuw and Leroy (1987) for an extensive introduction to this topic. It is known that the outliers impact more heavily on mixture linear regression than on the classic linear regression models, since the outliers not only affect the estimation of the regression coefficients, but also possibly distort the mixture structure. These problems may become even worse if the covariates in the mixture linear regression models are measured with error.

Compared with the research done for classic linear regression models, there are fewer studies on robust estimation in linear errors-in-variables (EIV) regression models. When the covariate has replications, Carroll and Gallo (1982) proposed an robust estimator for the slope in simple linear EIV models. By extending Huber (1973)'s idea for robust estimation in ordinary regression, Zamar (1989) and Cheng and Van Ness (1990) independently proposed a robust orthogonal regression procedure based on the standardized residuals. Bolfarine and Valle (1994) proposed a robust extension of normal measurement error models by using the t-family of distributions, which is a special case of elliptical distributions, to model the kurtosis of the error distributions, and robustness is achieved by choosing a proper degrees of freedom parameter of the t-distribution. Although there is no increase in the asymptotic variance of the estimation of the regression parameter due to the estimation of the degrees of freedom, the computation is slightly complicated. Other references concerning elliptical measurement error

* Corresponding author.
E-mail address: weixing@ksu.edu (W. Song).

0167-7152/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2013.09.036

models can be found in Valle et al. (1996), Labra et al. (1998), and Lachos et al. (2010). To our best knowledge, there is no discussion on robust inference in mixture linear regression models with measurement errors. This paper will try to fill this void by first extending the least absolute deviation (LAD) procedure to linear EIV regression models and then to mixture linear EIV regression models after the measurement errors are attenuated.

The paper is organized as follows. Based on the calibrated regression model, Section 2 develops an estimation–maximization (EM) algorithm by assuming that the error term follows a Laplace distribution. In Section 3, this methodology is extended to mixture linear EIV regression models. Simulation and comparison studies with other procedures are conducted in Section 4.

2. Robust estimation in EIV linear regression

We start with the following linear EIV regression model: $Y = \alpha+\beta' x+\varepsilon$, $X = x+u$, where Y is a scalar random variable, and x is a latent $d \times 1$ predictor vector which cannot be observed directly; its observed surrogate X is associated to x in an additive way described above, with measurement error u. We shall assume that $E[(\varepsilon, u')] = 0$, $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$, $\mathrm{Var}(u) = \Sigma_u$. For the sake of model identifiability, a commonly used assumption for u is to assume Σ_u to be known. This assumption will be used throughout the paper. In the case of Σ_u being unknown, if replications at each x are available, then a consistent estimator of Σ_u can be obtained. Replacing Σ_u by the consistent estimator in the following algorithm, the procedure still applies after an obvious modification.

If $Z = E(x|X)$ is known, then $E(Y|X) = \alpha+\beta' Z$. This implies a calibrated regression model for Y against X, $Y = \alpha+\beta' Z+\xi$, where ξ has mean $E(\xi|X) = 0$, and therefore it is uncorrelated with X. Then a robust estimator of $\theta = (\alpha, \beta)$ can be constructed based on the following LAD procedure: $\hat\theta = (\hat\alpha, \hat\beta) = \mathrm{argmin}_{\alpha,\beta} \sum_{i=1}^n \left|Y_i - \alpha - \beta' Z_i\right|$. By imposing some conditions as in Koenker (2005) on the model, the consistency of $\hat\theta$ can be proved using a similar argument to that in the classical linear regression model.

Note that the above LAD procedure is equivalent to the maximum likelihood estimation (MLE) procedure based on the Laplace likelihood function $f(\theta; \mathbf{S}) = (\sigma\sqrt{2})^{-n} \exp\left(-\sqrt{2}\sum_{i=1}^n |Y_i - \alpha - \beta' Z_i|/\sigma\right)$, where $\mathbf{S} = (\mathbf{X}, \mathbf{Y})$, $\mathbf{X} = (X_1, \ldots, X_n)$, and $\mathbf{Y} = (Y_1, \ldots, Y_n)$. To find the MLE of α and β, similar to Phillips (2002), an EM algorithm can be developed based on the connection between the Laplace distribution and the normal distribution. This connection was first discussed by Andrews and Mallows (1974).

It is known that, if V is a random variable with density $f_V(v) = v^{-3}\exp(-1/2v^2)$, and, given $V = v$, Y has a normal distribution with mean μ and variance $\sigma^2/2v^2$, then Y marginally has a Laplace distribution with density function $h_Y(y) = \exp(-\sqrt{2}|y-\mu|/\sigma)/(\sqrt{2}\sigma)$. Now, given the latent variable, $V = v$, suppose that Y has a normal distribution with mean $\alpha + \beta' Z$ and variance $\sigma^2/2v^2$. Then the complete log likelihood function of (α, β) based on $\mathbf{P} = (\mathbf{Y}, \mathbf{X}, \mathbf{V})$ is

$$L(\theta; \mathbf{P}) = -\frac{n}{2}\log\pi\sigma^2 - \frac{\sum_{i=1}^n V_i^2(Y_i - \alpha - \beta' Z_i)^2}{\sigma^2} - \sum_{i=1}^n \log V_i^2 - \frac{1}{2}\sum_{i=1}^n \frac{1}{V_i^2},$$

where $\mathbf{V} = (V_1, V_2, \ldots, V_n)$. Following the two steps in the EM algorithm, and assuming that $\theta = (\alpha^{(k)}, \beta^{(k)}, \sigma^{2(k)})$ is the value in the kth iteration, then, in the $(k+1)$th iteration, we have to first calculate the conditional expectation of the complete log likelihood function $L(\theta; \mathbf{P})$ given the observed data set $\mathbf{S}$. In fact,

$$E[L(\theta, \mathbf{P})|\mathbf{S}] = -\frac{n}{2}\log\pi\sigma^2 - \frac{\sum_{i=1}^n \tau_i^{(k)}(Y_i - \alpha - \beta' Z_i)^2}{\sigma^2} + T(S, \theta^{(k)}),$$

where $\tau_i^{(k)} = E[V_i^2|\mathbf{S}, \theta^{(k)}]$. From Phillips (2002), we can show that $\tau_i^{(k)} = \sigma^{(k)}/(\sqrt{2}|Y_i - \alpha^{(k)} - \beta^{(k)'} Z_i|)$ for $i = 1, 2, \ldots, n$, where $T(S, \theta^{(k)})$ does not depend on the unknown parameter θ. Then, maximizing $E[L(\theta, \mathbf{P})|\mathbf{S}]$ with respect to θ is equivalent to maximizing the following:

$$Q(\theta) = -\frac{n}{2}\log\pi\sigma^2 - \frac{\sum_{i=1}^n \tau_i^{(k)}(Y_i - \alpha - \beta' Z_i)^2}{\sigma^2}.$$

Setting $\partial Q(\theta)/\partial\alpha = 0$, $\partial Q(\theta)/\partial\beta = 0$, and $\partial Q(\theta)/\partial\sigma^2 = 0$, and denoting

$$\overline{Y}_\tau = \frac{\sum_{i=1}^n \tau_i^{(k)} Y_i}{\sum_{i=1}^n \tau_i^{(k)}}, \qquad \overline{Z}_\tau = \frac{\sum_{i=1}^n \tau_i^{(k)} Z_i}{\sum_{i=1}^n \tau_i^{(k)}},$$

we can obtain that, in the next iteration, θ should be updated by the following formulae:

$$\beta^{(k+1)} = \left[\sum_{i=1}^{n} \tau_i^{(k)} (Z_i - \overline{Z}_\tau)(Z_i - \overline{Z}_\tau)'\right]^{-1} \left[\sum_{i=1}^{n} \tau_i^{(k)} (Y_i - \overline{Y}_\tau)(Z_i - \overline{Z}_\tau)\right],$$

$$\alpha^{(k+1)} = \overline{Y}_\tau - \beta^{(k+1)'} \overline{Z}_\tau, \qquad \sigma^{2(k+1)} = \frac{2}{n} \sum_{i=1}^{n} \tau_i^{(k)} (Y_i - \alpha^{(k+1)} - \beta^{(k+1)'} Z_i)^2.$$

Then we can start another iteration using $\theta^{(k+1)} = (\alpha^{(k+1)}, \beta^{(k+1)}, \sigma^{2(k+1)})$ as the initial value.

It is well known that, under some regularity conditions, if the conditional median of ξ given Z is 0, then the LAD estimates is asymptotically normal,

$$\sqrt{n}(\hat{\alpha} - \alpha_0, (\hat{\beta} - \beta_0)')' \Longrightarrow N(0, C^{-1} D C^{-1}/4),$$

where $C = E f_{\xi|Z}(0)(1, Z')'(1, Z')$, and $D = E(1, Z')'(1, Z')$, and $f_{\xi|Z}$ is the conditional density of ξ given Z. For more details, see Koenker (2005). After obtaining the LAD estimates $\hat{\alpha}$, $\hat{\beta}$ from the above EM algorithm, a consistent estimator for the asymptotic covariance matrix can be obtained by replacing C and D with

$$\hat{C} = \frac{1}{nh} \sum_{i=1}^{n} I(|Y_i - \hat{\alpha} - \hat{\beta}_i' Z_i| \le h/2)(1, Z_i')'(1, Z_i'), \qquad \hat{D} = \frac{1}{n} \sum_{i=1}^{n} (1, Z_i')'(1, Z_i'),$$

respectively, where h is a sequence of positive numbers depending on n such that $h \to 0$ and $n\sqrt{h} \to \infty$. As one referee suggests, another possible way to obtain an estimate of the covariance of the EM estimates is to use the method proposed in Louis (1982), but this suggestion is discouraged by the fact that the likelihood function based on the observed data set is not differentiable with respect to the unknown parameters, although the full likelihood function is; this might invalidate the formula of the information matrix decomposition. The method proposed in Louis (1982) is used in Bolfarine and Rojas (1995), where the likelihood function based on the observed data is normally distributed.

In the classical setup, the latent predictor x and the measurement error u are often assumed to be normally distributed. If $x \sim N(\mu_x, \Sigma_x)$, $u \sim N(0, \Sigma_u)$, and x and u are independent, then we have $E(x|X) = \mu_x + \Sigma_x(\Sigma_x + \Sigma_u)^{-1}(X - \mu_x)$. Σ_u is assumed to be known, and μ_x and Σ_x can be consistently estimated by $\bar{X}$ and $S_X^2 - \Sigma_u$, where $\bar{X}$ is the sample average of the X_i, and S_X^2 is the sample covariance of the X_i.

The parametric assumption on $E(x|X)$ is sometimes not realistic. In such cases, nonparametric estimators of $E(x|X)$ might be considered. A commonly used nonparametric estimator for $E(x|X)$ is the deconvolution kernel estimator. Assume that the density function of u is known. Let K be a symmetric density function on $\mathbb{R}^d$, let ϕ_K denote its characteristic function, and let $h > 0$. Define $L_h(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \exp(-\mathbf{i}t \cdot x)\phi_K(t)/\Phi_u(t/h)dt$, $\mathbf{i} = \sqrt{-1}$. L_h is called the deconvolution kernel function with bandwidth h. Then the deconvolution kernel density estimator of x is defined as

$$\hat{f}_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} L_h\left(\frac{x - X_i}{h}\right). \tag{1}$$

Then $E(x|X)$ can be estimated by $\hat{Z} = \hat{\mu}(X) = \hat{f}_X^{-1}(X) \int x f_u(X - x)\hat{f}_h(x)dx$, where $\hat{f}_X(X)$ is the classical kernel density estimator of X evaluated at X with kernel K. In general, L_h has no explicit form except for some special cases. For example, if $u \sim N(0, \sigma^2)$, and K is so chosen that its characteristic function is $\phi_K(t) = (1 - t^2)_+^3$, then, from Fan and Truong (1993), the deconvolution kernel function L_h is given by

$$L_h(x) = \frac{1}{\pi} \int_0^1 \cos(tx)(1 - t^2)_+^3 \exp\left(\frac{\sigma_u^2 t^2}{2h^2}\right) dt.$$

If u has a double exponential distribution with mean 0 and variance σ_u^2, and the kernel function is chosen to be K, then $L_h(x) = K(x) - \sigma_u^2 K''(x)/(2h^2)$. In particular, if K is further chosen to be Gaussian kernel, then

$$L_h(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\left[1 - \frac{\sigma_u^2}{2h^2}(x^2 - 1)\right]. \tag{2}$$

Thus $\hat{Z}$ can be obtained by numerical integration. For more discussion on the deconvolution technique, see Stefanski and Carroll (1986), Carroll and Hall (1988), Liu and Taylor (1990), and the references therein.

Another possible estimator of the density function of x is the weighted kernel estimator proposed by Hazelton and Turlach (2009). By selecting proper weights, the weighted kernel estimator might have faster convergence rate than the deconvolution kernel density estimator defined in (1). However, a nonparametric regression estimator based on this weighted kernel estimator has not been explored.

3. Robust estimation in mixture EIV linear regression

Many real data examples involving measurement error present chances to consider mixture EIV modeling. For example, Tosteson et al. (1989) described a study about lung function in children: the response was the presence or absence of wheeze in children, which is an indicator of lung dysfunction. The predictor variable is the personal exposure to NO_2. The true personal exposure to NO_2 is not measured; instead, a bivariate variable consisting of observed kitchen and bedroom concentration of NO_2 in the child's home can be observed. The data set consists of two subgroups based on the gender of the children. If these two groups have different data structure, then it would be appropriate to fit the data set using mixture models.

This section will extend the above methodology to mixture linear EIV models. In this setup, we assume that, with probability $\pi_j,\ j = 1, 2, \ldots, g$, (Y, X') comes from one of the following $g \geq 2$ linear EIV regression models:

$$Y = \alpha_j + x'\beta_j + \varepsilon_j, \qquad X = x + u, \quad j = 1, 2, \ldots, g, \tag{3}$$

where $\sum_{j=1}^{g} \pi_j = 1$. The β_j are unknown d-dimensional vectors of regression coefficients, and σ is an unknown positive scalar. The unobservable x, the random errors ε_j, and u are assumed to be independent. We assume that, for each $j = 1, 2, \ldots, g$, $\xi_j = \varepsilon_j + \beta_j'(x - Z)$ has a Laplace distribution with mean 0 and variance σ_j^2. As before, Z is defined as the conditional expectation $E(x|X)$. It can be shown that ξ_j is uncorrelated with Z. Then it is easily seen that, for a sample $\mathbf{S} = \{(X_i', Y_i),\ i = 1, 2, \ldots, n\}$ from model (3), the log-likelihood function of $\theta = (\alpha_1, \beta_1, \sigma_1, \pi_1, \ldots, \alpha_g, \beta_g, \sigma_g^2, \pi_g)$ is

$$L(\theta; \mathbf{S}) = \sum_{i=1}^{n} \log \sum_{j=1}^{g} \frac{\pi_j}{\sqrt{2}\sigma_j} \exp\left(-\frac{\sqrt{2}|Y_i - \alpha_j - Z_i'\beta_j|}{\sigma_j}\right).$$

The maximum likelihood estimate of θ can be obtained by maximizing $L(\theta; \mathbf{S})$ with respect to θ. Usually no explicit solution is available. We will try to develop an EM algorithm to find the maximum likelihood estimate. For this purpose, denote the G_{ij} as latent Bernoulli variables such that $G_{ij} = 1$ if the ith observation (X_i', Y_i) is from jth component, and 0 otherwise. If the complete data set $\mathbf{T} = \{(X_i', Y_i, G_{ij}) : i = 1, \ldots, n,\ j = 1, \ldots, g\}$ is observable, then the complete likelihood function of θ can be written as

$$L(\theta; \mathbf{T}) = \sum_{i=1}^{n} \sum_{j=1}^{g} G_{ij} \log \frac{\pi_j}{\sqrt{2}\sigma_j} \exp\left(-\frac{\sqrt{2}|Y_i - \alpha_j - Z_i'\beta_j|}{\sigma_j}\right).$$

Denote V_i, coupled with (Z_i, Y_i), as the latent scale variables, as we did in Section 2, with $i = 1, 2, \ldots, n$. Then the complete log-likelihood function of θ, based on $\mathbf{D} = \{Z_i, Y_i, V_i, G_{ij} : i = 1, 2 \ldots, n,\ j = 1, 2, \ldots, g\}$ has the form

$$\begin{aligned} L(\theta; \mathbf{D}) = {} & \sum_{i=1}^{n} \sum_{j=1}^{g} G_{ij} \log \pi_j - \sum_{i=1}^{n} \sum_{j=1}^{g} G_{ij} \log \pi \sigma_j^2 - \sum_{i=1}^{n} \sum_{j=1}^{g} \frac{G_{ij} V_i^2 (Y_i - \alpha_j - Z_i'\beta_j)^2}{\sigma_j^2} \\ & + \sum_{i=1}^{n} \sum_{j=1}^{g} G_{ij} \log\left[\frac{1}{V_i^2} \exp\left(-\frac{1}{2V_i^2}\right)\right]. \end{aligned}$$

Based on the EM algorithm, in E-step, we have to calculate the conditional expectation of $L(\theta; \mathbf{D})$ given the observed data set $\mathbf{S} = \{(X_i, Y_i) : i = 1, 2, \ldots, n\}$ and the initial value $\theta^{(0)} = (\pi_1^{(0)}, \alpha_1^{(0)}, \beta_1^{(0)}, \sigma_1^{2(0)}, \ldots, \pi_g^{(0)}, \alpha_g^{(0)}, \beta_g^{(0)}, \sigma_g^{2(0)})$. Since the last term does not depend on the unknown parameters, to maximize $E[L(\theta; \mathbf{D})|\mathbf{S}]$, we can simply drop it from the analysis. But the calculation of this conditional expectation needs the following two terms: $\tau_{ij}^{(1)} = E(G_{ij}|\mathbf{S}, \theta^{(0)})$ and $\delta_{ij}^{(1)} = E[V_i^2|\mathbf{S}, \theta^{(0)}, G_{ij} = 1]$. One can show that

$$\tau_{ij}^{(1)} = \frac{\pi_j^{(0)} \sigma_j^{-(0)} \exp\left(-\sqrt{2}|Y_i - \alpha_j^{(0)} - Z_i'\beta_j^{(0)}|/\sigma_j^{(0)}\right)}{\sum_{m=1}^{g} \pi_m^{(0)} \sigma_m^{-(0)} \exp\left(-\sqrt{2}|Y_i - \alpha_m^{(0)} - Z_i'\beta_m^{(0)}|/\sigma_m^{(0)}\right)}, \tag{4}$$

$$\delta_{ij}^{(1)} = \frac{\sigma_j^{(0)}}{\sqrt{2}|Y_i - \alpha_j^{(0)} - Z_i'\beta_j^{(0)}|}. \tag{5}$$

Then we have to maximize the following:

$$Q(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_{ij}^{(1)} \log \pi_j - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_{ij}^{(1)} \log \sigma_j^2 - \sum_{i=1}^{n} \sum_{j=1}^{g} \frac{\tau_{ij}^{(1)} \delta_{ij}^{(1)} (Y_i - \alpha_j - Z_i'\beta_j)^2}{\sigma_j^2}.$$

By setting $\partial Q(\theta)/\partial\pi_j = 0$, we see that π_j should be updated by

$$\pi_j^{(1)} = \sum_{i=1}^{n} \tau_{ij}^{(1)}/n, \tag{6}$$

and by setting $\partial Q(\theta)/\partial\alpha_j = 0$, $\partial Q(\theta)/\partial\beta_j = 0$, $\partial Q(\theta)/\partial\sigma_j^2 = 0$, and denoting $w_{ij}^{(1)} = \tau_{ij}^{(1)}\delta_{ij}^{(1)}$,

$$\overline{Y}_j^{(1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(1)} Y_i}{\sum_{i=1}^{n} w_{ij}^{(1)}}, \qquad \overline{Z}_j^{(1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(1)} Z_i}{\sum_{i=1}^{n} w_{ij}^{(1)}}, \tag{7}$$

we can obtain that, in the next iteration, θ should be updated by the following formulae:

$$\beta_j^{(1)} = \left[\sum_{i=1}^{n} w_{ij}^{(1)}(Z_i - \overline{Z}_j^{(1)})(Z_i - \overline{Z}_j^{(1)})'\right]^{-1} \left[\sum_{i=1}^{n} w_{ij}^{(1)}(Y_i - \overline{Y}_j^{(1)})(Z_i - \overline{Z}_j^{(1)})\right],$$

$$\alpha_j^{(1)} = \overline{Y}_j^{(1)} - \beta_j^{(1)'}\overline{Z}_j^{(1)}, \qquad \sigma_j^{2(1)} = \frac{2\sum_{i=1}^{n} w_{ij}^{(1)}(Y_i - \alpha_j^{(1)} - \beta_j^{(1)'}Z_i)^2}{\sum_{i=1}^{n} \tau_{ij}^{(1)}}.$$

Thus we propose the following EM algorithm to maximize $L(\theta; \mathbf{D})$ with respect to θ.

EM algorithm

1. Choose an initial value for $\theta = (\pi_1, \alpha_1, \beta_1, \sigma_1^2, \ldots, \pi_g, \alpha_g, \beta_g, \sigma_g^2)$.
2. E-step: at the kth iteration, calculate $\tau_{ij}^{(k)}$ and $\delta_{ij}^{(k)}$ from Eqs. (4) and (5) with (0) replaced by $(k-1)$.
3. M-step: at the kth iteration, update π_j, $\overline{Y}_j$, and $\overline{Z}_j$ using (6) and (7) with (1) replaced by (k), and calculate the updated values for α_j, β_j, and σ_j^2 as follows:

$$\beta_j^{(k)} = \left[\sum_{i=1}^{n} w_{ij}^{(k)}(Z_i - \overline{Z}_j^{(k)})(Z_i - \overline{Z}_j^{(k)})'\right]^{-1} \left[\sum_{i=1}^{n} w_{ij}^{(k)}(Y_i - \overline{Y}_j^{(k)})(Z_i - \overline{Z}_j^{(k)})\right],$$

$$\alpha_j^{(k)} = \overline{Y}_j^{(k)} - \beta_j^{(k)'}\overline{Z}_j^{(k)}, \qquad \sigma_j^{2(k)} = \frac{2\sum_{i=1}^{n} w_{ij}^{(k)}(Y_i - \alpha_j^{(k)} - \beta_j^{(k)'}Z_i)^2}{\sum_{i=1}^{n} \tau_{ij}^{(k)}}.$$

4. Repeat steps 2 and 3 until convergence is obtained.

If the σ_j^2 are all the same, then, in the above EM algorithm, a common initial value should be used for these variances, and σ^2 can be updated in the M-step by

$$\sigma^{2(k)} = \frac{2\sum_{i=1}^{n}\sum_{j=1}^{g} w_{ij}^{(k)}(Y_i - \alpha_j^{(k)} - \beta_j^{(k)'}Z_i)^2}{n}.$$

The robustness of the proposed EM procedure indeed is a consequence of the adoption of the LAD regression. It is also noted that, in the formulae of the updated β_j in the kth iteration, the factor $\delta_{ij}^{(k)}$ is reciprocally related to $|Y_i - \alpha_j^{(k-1)} - Z_i'\beta_j^{(k-1)}|$, implying that larger fitted residuals give smaller values of δ_{ij}, and hence downweight the corresponding observations when calculating the estimates.

Extra attention should be paid when programming the proposed EM algorithm. In the case of $g = 1$, Schlossmacher (1973) warned that, if a perfect LAD fit occurs, i.e., $|Y_i - \alpha_j^{(k-1)} - Z_i'\beta_j^{(k-1)}| = 0$ for some $i, j, k \geq 1$ and $\alpha_j^{(k-1)}$, $\beta_j^{(k-1)}$, then the algorithm will eventually give $|Y_i - \alpha_j^{(k-1)} - Z_i'\beta_j^{(k-1)}|$ being close to 0 during the iteration process. As a result, $\delta_{ij}^{(k)}$ will be very large, and numerical instability would follow. Although Phillips (2002) noticed that this problem rarely arises in the case of $g = 1$, it does occur often in the mixture case, due to the fact that more than one regression model provide more chance for a perfect LAD fitting. It is not quite reasonable to adopt Schlossmacher (1973)'s weight scheme by setting $\delta_{ij}^{(k)} = 0$ whenever $|Y_i - \alpha_j^{(k-1)} - Z_i'\beta_j^{(k-1)}| < e$ for a pre-assigned $e > 0$. In our simulation study, we simply adopt a hard threshold rule to control the extremely small LAD residuals in each iteration step. Under this rule, $\delta_{ij}^{(k)}$ will be assigned a value of 10^6

for any perfect LAD fit. Other threshold values, such as 10^8 and 10^{10}, produce almost identical results. For the sake of brevity, only the case of 10^6 is reported.

Numerical instability could also occur if the weights are very small. A common way to deal with this issue is to impose a hard threshold on $\tau_{ij}^{(k)}$ obtained in the kth iteration. Namely, for a pre-specified value e, say, if $\tau_{ij}^{(k)} > e$, then $\tau_{ij}^{(k)}$ itself will be used for the next iteration; otherwise, e will be used as the weight for the next iteration. In our simulation study, $e = 10^{-6}$ is adopted.

4. Simulation studies

To assess the finite sample performance of the proposed robust estimation procedure, two simulation studies are conducted in this section. To resolve the model non-identifiability issue resulting from the label switching in mixture modeling, we simply choose the labels by minimizing the distance between the estimates and the true parameter values. In fact, there are no widely accepted labeling standards, and the effects of labeling schemes on comparison different estimation procedures deserve independent research in the future.

To see the effect of different distributions of ε on various estimation methods, we consider the following five cases: (1), $\varepsilon \sim N(0, 1)$; (2), $\varepsilon \sim$ Laplace distribution with mean 0 and variance1; (3). $\varepsilon \sim t_1$, t-distribution with one degree of freedom or the Cauchy distribution; (4), $\varepsilon \sim t_3$, t-distribution with three degrees of freedom; (5), $\varepsilon \sim 0.95N(0, 1) + 0.05N(0, 25)$, a mixture of two normal distributions.

Case 1 is often used to evaluate the efficiency of different estimation methods compared to the traditional MLE when the error is exactly normally distributed and there are no outliers. For Case 2, the estimation methods proposed in this paper will provide the MLE of unknown parameters, which, as in the first case, would serve as a reference line to evaluate the performance of other estimation procedures. Both Cases 3 and 4 are heavy-tailed distributions, and they are often used in the literature to mimic the outlier situations. Case 5 would produce 5% data likely to be low leverage outliers. In all simulation studies, the iteration is terminated whenever the difference of the likelihood functions in the current step and the previous step is less than 10^{-6}.

Four estimation methods were compared in the simulation study: (1), the maximum likelihood method based on the normality assumption (MLE); (2), the trimmed likelihood estimator (TLE) proposed by Neykov et al. (2007); (3), the robust modified EM algorithm based on bisquare (Bisquare) proposed by Bai et al. (2012); (4), the proposed robust EM mixture regression method based on a Laplace distribution (MixregL). In order to see the effect of the measurement error on all procedures, two different values of σ_u^2, 0.1 and 0.3, are considered. In both simulations, sample sizes $n = 100$ and 200 are considered, each scenario will be repeated 200 times, and the total mean square errors (MSEs) from the estimates of α, β, and π will be used to evaluate the relative performance for these four estimation procedures.

Simulation 1. In this simulation, the sample data $(X_{i1}, X_{i2}, Y_i)_{i=1}^n$ are generated from the following two-component mixture errors-in-variables linear regression model with $d = 2$ and $g = 2$:

$$Y = \begin{cases} 0 + x_1 + x_2 + \varepsilon_1, & \text{if } G = 1, \\ 0 - x_1 - x_2 + \varepsilon_2, & \text{if } G = 2, \end{cases} \qquad \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

where G is the latent component indicator. The true regression coefficients for the first component are $\alpha_1 = 0$ and $\beta_1 = (1, 1)'$, and for the second component they are $\alpha_2 = 0$ and $\beta_2 = (-1, -1)'$. The predictor vector (x_1, x_2) will be assumed to be independent bivariate normal with mean $(0, 0)$ and equal variances 1, and the measurement error vector (u_1, u_2) has an independent normal distribution with mean $(0, 0)$ and equal variances σ_u^2. If we fit $E(x|X)$ using all the distributional assumption, then $Z = E(x|X) = \mu_x + \Sigma_{xx}\Sigma_{XX}^{-1}(X - \mu_x)$, and μ_x can be estimated by the sample mean $\bar{X} = (\bar{X}_1, \bar{X}_2)$, Σ_{XX} by the sample covariance matrix S_X^2 of X, and Σ_{xx} by $S_X^2 - \Sigma_{uu}$. Since the deconvolution kernel regression estimate is computationally extensive in two-dimensional cases, we will not discuss this nonparametric alternative here.

The simulation results are presented in Table 1. We can see that, if the true distribution of ε is normal, the MSEs of the MLE procedure are larger than those from our proposed method, but the superiority of MLE over all other methods becomes clear when the sample size gets larger. For other cases when the distribution of ε has a heavier tail, or the data are contaminated by some outliers, MLE fails to provide reasonable estimates.

The overall performance of TLE and Bisquare is better than that of MLE, but is still less desirable if ε has a heavier tail. See the simulation results for Case III, where ε has a t-distribution with one degree of freedom. The simulation results clearly show that the proposed method in the paper outperforms or is at least comparable to any other method when $n = 100$.

It is also noted that, when σ_u^2 gets larger, all the methods get worse. Except for Case III, the proposed method is slightly inferior than the TLE and Bisquare methods even for larger sample sizes, implying that a larger measurement error has a larger adverse effect on the proposed method.

Simulation 2. In this simulation study, we consider the following linear errors-in-variables mixture regression model with x being one dimensional, and $g = 2$.

$$Y = \begin{cases} 0 + x + \varepsilon_1, & \text{if } G = 1, \\ 0 - x + \varepsilon_2, & \text{if } G = 2, \end{cases} \qquad X = x + u.$$

Table 1
MSEs from Simulation 1.

n	Case	MLE		TLE		Bisquare		MixregL	
		$\sigma_u = 0.1$	$\sigma_u = 0.3$	$\sigma_u = 0.1$	$\sigma_u = 0.3$	$\sigma_u = 0.1$	$\sigma_u = 0.3$	$\sigma_u = 0.1$	$\sigma_u = 0.3$
100	I	0.430	0.555	0.601	0.772	0.467	0.577	0.202	0.469
	II	0.477	0.696	0.276	0.454	0.337	0.510	0.121	0.329
	III	1534.465	1962.242	11.987	11.503	7.925	8.596	0.967	1.420
	IV	10.993	8.587	0.743	0.868	1.169	1.599	0.273	0.536
	V	8.080	15.451	0.614	0.861	0.713	0.990	0.229	0.518
200	I	0.173	0.229	0.453	0.502	0.174	0.238	0.121	0.374
	II	0.154	0.202	0.137	0.207	0.128	0.170	0.074	0.267
	III	453.333	480.913	6.811	6.661	5.447	6.000	0.582	1.008
	IV	2.749	4.597	0.326	0.453	0.325	0.447	0.149	0.437
	V	4.266	4.537	0.328	0.376	0.215	0.285	0.131	0.391

Table 2
MSEs from Simulation 2.

n	Case	MLE		TLE		Bisquare		MixregL	
		$\sigma_u = 0.1$	$\sigma_u = 0.3$	$\sigma_u = 0.1$	$\sigma_u = 0.3$	$\sigma_u = 0.1$	$\sigma_u = 0.3$	$\sigma_u = 0.1$	$\sigma_u = 0.3$
100	I	0.544	0.645	2.347	2.352	0.577	0.682	0.275	0.383
	II	1.251	2.048	1.562	1.441	0.404	0.569	0.103	0.194
	III	763.911	740.434	12.809	13.619	5.068	5.110	0.824	0.819
	IV	8.785	9.880	2.062	2.207	1.561	1.971	0.380	0.434
	V	11.382	15.799	2.226	2.165	0.908	1.014	0.279	0.381
200	I	0.161	0.209	1.766	1.767	0.161	0.206	0.204	0.233
	II	0.311	0.328	1.161	1.086	0.178	0.206	0.092	0.121
	III	636.634	640.183	8.974	9.656	3.347	3.955	0.691	0.700
	IV	9.269	10.964	1.406	1.436	0.713	0.773	0.196	0.222
	V	9.960	11.361	1.461	1.689	0.263	0.299	0.186	0.202

We do not have the distributional assumption on x, and the deconvolution estimator of $E(x|X)$ will be used. Assume that u has a double exponential distribution with mean 0 and variance σ_u^2. Then, by choosing K to be the standard normal kernel, the product deconvolution kernel function will be $L_h(x)$, with $L_h(\cdot)$ defined by (2). The corresponding estimate for $E(x|X_i)$ will be

$$Z_i = \hat{E}(x|X_i) = \frac{\sum_{j=1}^{n} \int x f_u(X_i - X_j) L_h((x - X_j)/h) dx}{\sum_{j=1}^{n} K((X_i - X_j)/h)}.$$

The above integrals do not have explicit analytic expressions, and the Riemann sums over $[-6, 6]$ with subintervals of length 0.01 are used to evaluate them.

The simulation results are reported in Table 2. Similar to Simulation 1, when σ_u^2 gets larger, the MSEs for all methods get larger. But, different from the previous simulation, the proposed method outperforms or is comparable to all other estimation procedures in all scenarios.

Acknowledgment

Jianhong Shi's research is supported by the Natural Science Foundation of Shanxi Province, China (2013011002-1).

References

Andrews, D.F., Mallows, C.L., 1974. Scale mixtures of normal distributions. Journal of the Royal Statistical Society. Series B (Methodological) 36 (1), 99–102.
Bai, X., Yao, W., Boyer, J.E., 2012. Robust fitting of mixture regression models. Computational Statistics and Data Analysis 56, 2347–2359.
Bolfarine, H., Rojas, M., 1995. Structural comparative calibration using the em-algorithm. Journal of Applied Statistics 22, 277–292.
Bolfarine, H., Valle, R., 1994. Robust modelling in measurement error models using the t distribution. Revista Brasileira de Probabilidade e Estatistica 1, 67–84.
Carroll, R., Gallo, P., 1982. Some aspects of robustness in functional errors-in-variables regression models. Communications in Statistics Series A 11, 2573–2585.
Carroll, R., Hall, P., 1988. Optimal rates of convergence for deconvolving a density. Journal of the American Statistical Association 83, 1184–1186.
Cheng, C., Van Ness, J.W., 1990. Bounded influence errors-in-variables regression. Contemporary Mathematics 112, 227–241.
Fan, J.Q., Truong, Y.K., 1993. Nonparametric regression with errors-in-variables. Annals of Statistics 21 (4), 1900–1925.
Hazelton, M.L., Turlach, B.A., 2009. Nonparametric density deconvolution by weighted kernel estimators. Statistics and Computing 19 (3), 217–228.
Huber, P., 1973. Robust regression: asymptotic, conjectures, and Monte Carlo. The Annals of Statistics 1, 799–821.

Koenker, R., 2005. Quantile Regression. Cambridge University Press.
Labra, F., Bolfarine, H., Valle, R., 1998. Elliptical functional model. Journal of Multivariate Analysis 65, 36–57.
Lachos, V., Labra, F., Bolfarine, H., Gosh, P., 2010. Multivariate measurement error models based on scale mixtures of the skew-normal distribution. Statistics 44, 541–556.
Liu, M., Taylor, R., 1990. Simulation and computation of a nonparametric density estimator for the deconvolution problem. Journal of Statistical Computation and Simulation 35, 145–167.
Louis, T., 1982. Finding the observed information matrix when using the em algorithm. Journal of the Royal Statistical Society: Series B 44, 226–233.
Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P., 2007. Robust fitting of mixtures using the trimmed likelihood estimator. Computational Statistics and Data Analysis 52, 299–308.
Phillips, R., 2002. Least absolute deviations estimation via the em algorithm. Statistics and Computing 12, 281–285.
Rousseeuw, P., Leroy, A., 1987. Robust Regression and Outlier Detection. John Wiley & Sons, Inc.
Stefanski, L., Carroll, R., 1986. Deconvoluting kernel density estimators. Statistics 21, 169–184.
Schlossmacher, E.J., 1973. An iterative technique for absolute deviations curve fitting. Journal of the American Statistical Association 68, 857–859.
Tosteson, T., Stefanski, L., Schafer, D., 1989. A measurement-error model for binary and ordinal regression. Statistics in Medicine 8, 1139–1147.
Valle, R.A., Bolfarine, H., Labra, F., 1996. Ultrastructural elliptical models. Canadian Journal of Statistics 24, 207–216.
Zamar, R., 1989. Robust estimation in the errors-in-variables model. Biometrika 76, 149–160.