Mixtures of Linear Regression with Measurement Errors

Weixin Yao* and Weixing Song†

Abstract

Existing research on mixtures of regression models are limited to directly observed predictors. The estimation of mixtures of regression for measurement error data imposes challenges for statisticians. For linear regression models with measurement error data, the naive ordinary least squares method, which directly substitutes the observed surrogates for the unobserved error-prone variables, yields an inconsistent estimate for the regression coefficients. The same inconsistency also happens to the naive mixtures of regression estimate, which is based on the traditional maximum likelihood estimator and simply ignores the measurement error. To solve this inconsistency, we propose to use the deconvolution method to estimate the mixture likelihood of the observed surrogates. Then our proposed estimate is found by maximizing the estimated mixture likelihood. In addition, a generalized EM algorithm is also developed to find the estimate. The simulation results demonstrate that the proposed estimation procedures work well and perform much better than the naive estimates.

Key words: EM algorithm; Mixture regression models; Measurement errors; Switching regression models.

*Department of Statistics, Kansas State University. E-mail: wxyao@ksu.edu.
†Department of Statistics, Kansas State University. E-mail:weixing@ksu.edu

1 Introduction

Mixtures of regression models are well known as switching regression models in econometrics literature, which were first introduced by Goldfeld and Quandt (1976). These models are used to investigate the relationship between interested variables coming from several unknown latent components. The model setting for mixtures of regression models can be stated as follows. Let $\mathcal{Z}$ be a latent class variable with $P(\mathcal{Z} = j \mid \mathbf{X} = \mathbf{x}) = \pi_j$ for $j = 1, 2, \cdots, m$, where $\mathbf{x}$ is a p-dimensional vector. Given $\mathcal{Z} = j$, suppose that the response y depends on $\mathbf{x}$ in a linear way

$$y = \mathbf{x}^T \boldsymbol{\beta}_j + \epsilon_j,$$

where $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{pj})^T$ and $\epsilon_j \sim N(0, \sigma_j^2)$. Then the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ can be written as

$$Y|\mathbf{X} = \mathbf{x} \sim \sum_{j=1}^{m} \pi_j N(\mathbf{x}^T \boldsymbol{\beta}_j, \sigma_j^2). \tag{1.1}$$

For more information about the mixtures of regression models (1.1), please see, for example, McLachlan and Peel (2000) and Frühwirth-Schnatter (2006). The unknown parameters in the model (1.1) can be estimated by the maximum likelihood estimator using the EM algorithm (Dempster et al., 1977). Many applications of mixture of regression models can be found in literature, such as in econometrics (Wedel and DeSarbo, 1993; Frühwirth-Schnatter, 2001), and in biology and epidemiology (Wang et al., 1996; Green and Richardson, 2002).

In this article, we will assume that the number of components m is known. When it is unknown, many methods have been proposed to choose the order m. See, for example, the AIC and BIC methods (Leroux, 1992), distance measures based methods (Chen and Kalbfleisch, 1996; James, Priebe, and Marchette, 2001; Charnigo and Sun, 2004; Woo and

Sriram, 2006; Ray and Lindsay, 2008), and hypothesis testing based methods (Chen, Chen, and Kalbfleisch, 2001, 2004). Recently, Chen and Li (2009) and Li and Chen (2010) proposed an EM test approach for testing the order of finite mixtures.

To the best of our knowledge, most existing estimation procedures for mixtures of regression models are limited to directly observed predictors. The estimation of mixtures of regression for measurement error data imposes challenges for statisticians. For linear regression models with measurement error data, it is well known that the naive ordinary least squares method, which directly substitutes the observed surrogates for the unobserved error-prone variables, yields an inconsistent estimate for the regression coefficients. For more information about linear regression with measurement errors, see Fuller (1987). The same inconsistency also happens to the naive mixture of regression estimate, which is based on the traditional maximum likelihood estimator and simply ignores the measurement error. To remove the inconsistency, a deconvolution technique will be used to estimate the mixture likelihood of the observed surrogates, more details will be given later. The proposed estimate is found by maximizing the estimated mixture likelihood of the observed surrogates. A generalized EM algorithm is developed to maximize the estimated mixture likelihood. The ascending property of the proposed algorithm is proved. Using simulation results, we demonstrate that the proposed estimation procedures work well and perform much better than the naive estimates which simply ignore the measurement error.

The rest of this paper is organized as follows. In Section 2, we propose the new estimation procedure to account for the measurement error. A generalized EM algorithm is also proposed to estimate the mixtures of regression with measurement error. In Section 3, we use the simulation study and a real data application to illustrate our proposed estimation procedure. In Section 4, we summarize the proposed method and give a short discussion. The proofs of the ascending property of the proposed algorithm are deferred to Appendix.

2 Mixtures of regression with measurement error

2.1 Introduction to the new method

In this section, we consider the mixtures of regression when the $\mathbf{X}$ or part of the $\mathbf{X}$ in (1.1) can not be observed directly and instead the surrogate, denoted by $\mathbf{W}$, of $\mathbf{X}$ is observed. The mixtures of regression with measurement error model assumes that

$$\begin{aligned} P(\mathcal{Z} = j \mid \mathbf{W}, \mathbf{X}) &= \pi_j \\ Y|\mathbf{X} = \mathbf{x}, \mathcal{Z} = j, \mathbf{W} &\sim N(\mathbf{x}^T\boldsymbol{\beta}_j, \sigma_j^2) \\ \mathbf{W} &= \mathbf{X} + \mathbf{U} \end{aligned} \tag{2.1}$$

where $\mathbf{W}$ is an observed surrogate of $\mathbf{X}$ and $\mathbf{U}$ is the measurement error and independent of $(\mathbf{X}, \mathbf{Y}, \mathcal{Z})$. Denote by $f_U(\mathbf{u})$ the density of $\mathbf{U}$ (some elements of $\mathbf{U}$ might have degenerate distributions if the corresponding elements of X are measured without errors). We first consider the situation in which the distribution of U, $f_U(\mathbf{u})$, is known, we will study the case when it is unknown later on.

The naive estimation method for the model (2.1) will simply ignores the measurement error $\mathbf{U}$ and estimate $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \sigma_1, \pi_1, \ldots, \boldsymbol{\beta}_m, \sigma_m, \pi_m)$ by maximizing the log-likelihood

$$\sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \frac{\pi_j}{\sigma_j} \phi \left\{ (y_i - \mathbf{w}_i^T \boldsymbol{\beta}_j)/\sigma_j \right\} \right\}, \tag{2.2}$$

where $\phi(\cdot)$ is the normal density for standard normal $N(0, 1)$. Similar to the least squares method for linear regression with measurement error, the naive estimate by maximizing (2.2) is not consistent, since the wrong model and likelihood function are used. We will also demonstrate this inconsistency using our simulation studies in Section 3.

If σ_js are unequal, it is well known that the log-likelihood function (2.2) is unbounded and goes to infinity if one observation exactly lies on one component line and the corresponding component variance goes to zero. When the likelihood is unbounded, we define the MLE as the maximum interior/local mode. Hathaway (1985) provided some theoretical support of using the maximum interior/local mode. There has been considerable research dealing with the unbounded mixture likelihood issue. See, for example, Hathaway (1985, 1986), Chen, Tan, and Zhang (2008), and Yao (2010).

In order to account for the measurement error in the mixture of regression model, we need to find the conditional density of Y given W. Given $Z = j$, the conditional density of Y given $\mathbf{W} = \mathbf{w}$ is

$$f_j(y \mid \mathbf{w}, \boldsymbol{\theta}_j) = \int f(y \mid \mathbf{x}, \boldsymbol{\theta}_j) f(\mathbf{x} \mid \mathbf{w}) d\mathbf{x} = \frac{1}{\sigma_j} \int \phi\left\{(y - \mathbf{x}^T \boldsymbol{\beta}_j)/\sigma_j\right\} f(\mathbf{x} \mid \mathbf{w}) d\mathbf{x} \qquad (2.3)$$

where $\boldsymbol{\theta}_j = (\beta_{1j}, \ldots, \beta_{pj}, \sigma_j)^T$. For simplicity of notation, here, we use $f(\cdot)$ to denote the generic density. Therefore $Y \mid \mathbf{W} = \mathbf{w} \sim \sum_{j=1}^m \pi_j f_j(y \mid \mathbf{w}, \boldsymbol{\theta}_j)$, and the log-likelihood for $\boldsymbol{\theta}$ is

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^n \log\left\{\sum_{j=1}^m \pi_j f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j)\right\}, \qquad (2.4)$$

where $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\theta}_1, \ldots, \pi_m, \boldsymbol{\theta}_m)^T$. Then our proposed new estimate of $\boldsymbol{\theta}$ is the maximizer of (2.4). A generalized EM algorithm to maximizer (2.4) will be provided in Section 2.2.

In many cases, $f(\mathbf{x} \mid \mathbf{w})$ might be unknown. Denote by $\hat{f}(\mathbf{x} \mid \mathbf{w})$ the estimated conditional distribution of $\mathbf{x}$ given $\mathbf{w}$. Then we propose to estimate $\boldsymbol{\theta}$ by maximizing the estimated log-likelihood

$$\log \hat{L}(\boldsymbol{\theta}) = \sum_{i=1}^n \log\left\{\sum_{j=1}^m \pi_j \hat{f}_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j)\right\}, \qquad (2.5)$$

where

$$\hat{f}_j(y \mid \mathbf{w}, \boldsymbol{\theta}_j) = \frac{1}{\sigma_j} \int \phi\left\{(y - \mathbf{x}^T\boldsymbol{\beta}_j)/\sigma_j\right\} \hat{f}(\mathbf{x} \mid \mathbf{w}) d\mathbf{x}.$$

We will provide the method of estimating $f(\mathbf{x} \mid \mathbf{w})$ in Section 2.3. Denote by $\hat{\boldsymbol{\theta}}$ the maximizer of (2.5).

2.2 Estimation Algorithm

To maximize (2.4) (or (2.5)) is not trivial. Here, we propose a generalized EM algorithm to maximize (2.4). Define a vector of component indicator $\mathbf{z}_i = (z_{i1}, \ldots, z_{im})^T$, where

$$z_{ij} = \begin{cases} 1, & \text{if } (\mathbf{w}_i, y_i) \text{ is from the } j\text{-th component;} \\ 0, & \text{otherwise.} \end{cases}$$

Then the complete log-likelihood function for the complete data $\{(\mathbf{w}_i, y_i, \mathbf{z}_i),$ $i = 1, \ldots, n\}$, by omitting some irrelevant constants, is

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} \left\{\log \pi_j + \log f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j)\right\}.$$

Based on the properties of EM algorithm, in the $(k+1)$th E step, we need to calculate $\mathrm{E}\left\{l_c(\boldsymbol{\theta}) \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}\right\}$, where $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{\theta}^{(k)}$ is the estimate of $\boldsymbol{\theta}$ at kth step. Since $l_c(\boldsymbol{\theta})$ is a linear function of z_{ij}'s, in the E step, we only need to calculate

$$p_{ij}^{(k+1)} = \mathrm{E}\left\{Z_{ij} \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}\right\} = \frac{\pi_j^{(k)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})}{\sum_{l=1}^{m} \pi_l^{(k)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})}, i = 1, \ldots, n,\ j = 1, \ldots, m. \qquad (2.6)$$

In the M step, we need to find $\boldsymbol{\theta}$ by maximizing

$$Q(\boldsymbol{\theta}) = \mathrm{E}\left\{ l_c(\boldsymbol{\theta}) \mid \boldsymbol{\theta}^{(k)}, \mathbf{y} \right\} = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \{ \log \pi_j + \log f_j(y_i \mid w_i, \boldsymbol{\theta}_j) \}. \tag{2.7}$$

Hence

$$\pi_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(k+1)}, \tag{2.8}$$

and $\boldsymbol{\theta}_j$ is the maximizer of

$$\sum_{i=1}^{n} p_{ij}^{(k+1)} \log f_j(y_i \mid w_i, \boldsymbol{\theta}_j). \tag{2.9}$$

Therefore, $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{pj})^T$ is the solution of

$$\begin{aligned}
0 =& \frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{n} p_{ij}^{(k+1)} \frac{\partial \log f_j(y_i \mid w_i, \boldsymbol{\theta}_j)}{\partial \boldsymbol{\beta}_j} \\
=& \sum_{i=1}^{n} p_{ij}^{(k+1)} \frac{\int \phi\{(y_i - \mathbf{x}^T \boldsymbol{\beta}_j)/\sigma_j\} (y_i - \mathbf{x}^T \boldsymbol{\beta}_j) \mathbf{x} f(\mathbf{x} \mid \mathbf{w}_i) d\mathbf{x}}{f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j) \sigma_j^3} \\
\approx& \sigma_j^{-2} \left[\sum_{i=1}^{n} p_{ij}^{(k+1)} y_i \int \tau_{ij}^{(k+1)}(\mathbf{x}) \mathbf{x} d\mathbf{x} - \left\{ \sum_{i=1}^{n} p_{ij}^{(k+1)} \int \tau_{ij}^{(k+1)}(\mathbf{x}) \mathbf{x}\mathbf{x}^T d\mathbf{x} \right\} \boldsymbol{\beta}_j \right],
\end{aligned}$$

where

$$\tau_{ij}^{(k+1)}(\mathbf{x}) = f(\mathbf{x} \mid \boldsymbol{\theta}_j^{(k)}, y_i, \mathbf{w}_i) = \frac{\phi\{(y_i - \mathbf{x}^T \boldsymbol{\beta}_j^{(k)})/\sigma_j^{(k)}\} f(\mathbf{x} \mid \mathbf{w}_i)}{f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)}) \sigma_j^{(k)}} \tag{2.10}$$

is the conditional density of $\mathbf{x}$ given the $\mathbf{w}_i$, y_i and the current estimate $\boldsymbol{\theta}_j^{(k)}$. Hence, based on the above approximation, we can update $\boldsymbol{\beta}_j$ by

$$\boldsymbol{\beta}_j^{(k+1)} = \left\{ \sum_{i=1}^{n} p_{ij}^{(k+1)} \int \tau_{ij}^{(k+1)}(\mathbf{x}) \mathbf{x}\mathbf{x}^T d\mathbf{x} \right\}^{-1} \left\{ \sum_{i=1}^{n} p_{ij}^{(k+1)} y_i \int \tau_{ij}^{(k+1)}(\mathbf{x}) \mathbf{x} d\mathbf{x} \right\}. \tag{2.11}$$

The σ_j^2 is the solutions of

$$\begin{aligned} 0 =& \frac{\partial Q(\boldsymbol{\theta})}{\partial \sigma_j^2} = \sum_{i=1}^{n} p_{ij}^{(k+1)} \left[\frac{\int \phi\{(y_i - \mathbf{x}^T\boldsymbol{\beta}_j)/\sigma_j\}(y_i - \mathbf{x}^T\boldsymbol{\beta}_j)^2 f(\mathbf{x} \mid \mathbf{w}) dx}{2\sigma_j^5 f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j)} - \frac{1}{2\sigma_j^2} \right] \\ \approx& (2\sigma_j^4)^{-1} \sum_{i=1}^{n} p_{ij}^{(k+1)} \left[\int \tau_{ij}^{(k+1)}(\mathbf{x})(y_i - \mathbf{x}^T\boldsymbol{\beta}_j^{(k+1)})^2 d\mathbf{x} - \sigma_j^2 \right]. \end{aligned}$$

Based on the above approximation, we can update σ_j by

$$\sigma_j{}^{(k+1)} = \left[\left\{ \sum_{i=1}^{n} p_{ij}^{(k+1)} \right\}^{-1} \sum_{i=1}^{n} p_{ij}^{(k+1)} \int \tau_{ij}^{(k+1)}(\mathbf{x}) \left\{ y_i - \mathbf{x}^T\boldsymbol{\beta}_j^{(k+1)} \right\}^2 d\mathbf{x} \right]^{1/2}. \tag{2.12}$$

If we assume σ_j's are equal, i.e., $\sigma_1 = \sigma_2 = \cdots = \sigma_m = \sigma$, then we can update σ by

$$\sigma^{(k+1)} = \left[n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \int \tau_{ij}^{(k+1)}(\mathbf{x}) \left\{ y_i - \mathbf{x}^T\boldsymbol{\beta}_j^{(k+1)} \right\}^2 d\mathbf{x} \right]^{1/2}. \tag{2.13}$$

We will prove in Theorem 2 that the iterations from (2.10) to (2.12) can be also considered as an EM algorithm for the the objective function (2.9) with $\mathbf{x}_i$'s as missing latent variables (See the proof of Theorem 2 in the Appendix for more detail). Therefore, one may run the full iteration of (2.10) — (2.12) to get the update value $\boldsymbol{\theta}^{(k+1)}$. However, based on the properties of EM algorithm, each iteration from (2.10) to (2.12) increases the objective function (2.9) and thus suffice for the monotone increasing of (2.4) for the whole algorithm from (2.6) to (2.12).

Based on the above descriptions, we propose the following generalized EM algorithm (GEM; Dempster, Laird, and Rubin 1977) to estimate $\boldsymbol{\theta}$.

Algorithm 1. *Starting with* $\boldsymbol{\theta}^{(0)}$*, in* $(k+1)$*th step*

E-Step*: Calculate the classification probabilities* $p_{ij}^{(k+1)}$*'s using (2.6).*

M-Step*: Update π_j's, $\boldsymbol{\beta}_j$'s and σ_j's based on (2.8), (2.11), and (2.12).*

Theorem 1. *Each iteration of the E and M steps in Algorithm 1 will monotonically increase the log-likelihood (2.4), i.e.,*

$$\log L(\boldsymbol{\theta}^{(k+1)}) \geq \log L(\boldsymbol{\theta}^{(k)}),$$

for all k, where $\log L(\boldsymbol{\theta})$ is defined in (2.4).

2.3 Estimation of $f(\mathbf{x} \mid \mathbf{w})$

Notice that

$$f(\mathbf{x} \mid \mathbf{w}) = \frac{f_X(\mathbf{x}) f(\mathbf{w} \mid \mathbf{x})}{f_W(\mathbf{w})},$$

where $f(\mathbf{w} \mid \mathbf{x}) = f_U(\mathbf{w} - \mathbf{x})$ is assumed to be known and $f_W(\mathbf{w})$ can be estimated by kernel density estimator. In fact, the proposed estimation procedure in Algorithm 1 for $\boldsymbol{\theta}$ does not depend on $f_W(\mathbf{w})$, since it does not involve the unknown parameters. Therefore, we only need to estimate $f_X(\mathbf{x})$ in order to estimate $f(\mathbf{x} \mid \mathbf{w})$. Estimating $f_X(\mathbf{x})$ when f_U is given has been a long standing research problem for measurement error model. In this article, we use the nonparametric deconvolution method to estimate $f_X(\mathbf{x})$.

For any p-dimensional density function L, let ϕ_L denote its characteristic function and define

$$K_h(\mathbf{x}) = \frac{1}{(2\pi)^p} \int_{R^p} \exp(-\mathbf{i}\mathbf{t}'\mathbf{x}) \frac{\phi_L(\mathbf{t})}{\phi_U(\mathbf{t}/h)} dt, \quad \mathbf{i} = \sqrt{-1},$$

where h is a positive number. Then the deconvolution kernel estimate of $f(\mathbf{x})$ with the bandwidth h is defined as

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} K_h \left(\frac{\mathbf{x} - \mathbf{w}_i}{h} \right). \tag{2.14}$$

The asymptotic properties of this deconvolution kernel estimate of $f(\mathbf{x})$ were thoroughly

discussed in literature. See Stefanski and Carroll (1986, 1990), Fan (1991a,1991b) and the references therein for more details. Very often the deconvolution kernel function K_h is not tractable, but in some particular cases, K_h does have explicit forms. For example, Fan and Truong (1993) showed that if $f_U(u)$ has double exponential distribution

$$f_U(u) = (\sqrt{2}\sigma)^{-1} \exp(-\sqrt{2}|u|/\sigma),$$

then

$$K_h(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \left[1 - \frac{\sigma^2}{2h^2}(x^2 - 1)\right],$$

and if $f_U(u)$ has normal distribution $N(0, \sigma^2)$, then

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

If $f(\mathbf{x})$ has a parametric form, then one can certainly construct more efficient estimates. In our examples, in order to reduce the dependence of our method on the parametric assumption of $f_X(\mathbf{x})$ and enhance the generality of our method, we will use the nonparametric deconvolution method to estimate $f_X(\mathbf{x})$. Based on our empirical study, the proposed estimate based on nonparametric deconvolution method is not very sensitive to the distribution assumption of the measurement error.

Remark: If $f_U(\mathbf{u})$ is only unknown due to the covariance matrix $\mathbf{\Sigma}_U = \text{Cov}(\mathbf{U})$, for example if $\mathbf{U}$ is normal with mean zero with unknown covariance matrix, we can estimate $\mathbf{\Sigma}_U$ based on the partially replicated observations, $\mathbf{W}_{ij} = \mathbf{X}_i + \mathbf{U}_{ij}$ for $j = 1, \ldots, J_i$ (Carroll, et al. 2006, chap 3). Let $\bar{\mathbf{W}}_i = J_i^{-1} \sum_{i=1}^{J_i} \mathbf{W}_{ij}$ and $\bar{\mathbf{U}}_i = J_i^{-1} \sum_{i=1}^{J_i} \mathbf{U}_{ij}$. Then a consistent

estimate of $\boldsymbol{\Sigma}_U$ is

$$\hat{\boldsymbol{\Sigma}}_U = \sum_{i=1}^{n}\sum_{j=1}^{J_i}(\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)(\mathbf{W}_{ij} - \bar{\mathbf{W}}_i)^T / \sum_{i=1}^{n}(J_i - 1).$$

Note that $\text{Cov}(\bar{\mathbf{U}}_i) = J_i^{-1}\boldsymbol{\Sigma}_U$. By mimicking the idea of linear regression with measurement error, we can also use a bias corrected estimation equation weighted by the classification probabilities to update $\boldsymbol{\beta}_j$'s in the M step of Algorithm 1

$$\boldsymbol{\beta}_j^{(k+1)} = \arg\min_{\boldsymbol{\beta}_j} \sum_{i=1}^{n} p_{ij}^{(k+1)} \left\{ (y_i - \bar{\mathbf{W}}_i^T\boldsymbol{\beta}_j)^2 - J_i^{-1}\boldsymbol{\beta}_j^T\hat{\boldsymbol{\Sigma}}_U\boldsymbol{\beta}_j \right\}. \tag{2.15}$$

2.4 Bandwidth Selection

When $f(\mathbf{x})$ is assumed to be unknown, we need to estimate it first based on the deconvolution method proposed in Section 2.3. Therefore, a choice of a bandwidth h for (2.14) is needed. In practice, data driven methods can be used for bandwidth selection, such as cross-validation (CV). Denote by $\mathcal{D}$ as the full data set. We then partition $\mathcal{D}$ into a training set $\mathcal{R}_l$ and test set $\mathcal{T}_l$, $\mathcal{D} = \mathcal{T}_l \cup \mathcal{R}_l$ $l = 1, \cdots, J$. We use the training set $\mathcal{R}_l$ to obtain the estimates $\hat{\boldsymbol{\theta}}$. We consider a likelihood version CV, which is given by

$$CV = \sum_{l=1}^{J}\sum_{q\in\mathcal{T}_l} \log\left\{ \sum_{j=1}^{m} \hat{\pi}_j \hat{f}_j(y_q \mid \mathbf{w}_q, \hat{\boldsymbol{\theta}}_j) \right\}. \tag{2.16}$$

The optimal bandwidth is selected when CV is maximized. Based on our empirical experience, the cross-validation tends to provide a smaller bandwidth than the optimal one.

3 Examples

In this section, the sampling behavior of the proposed mixture of regression estimate with measurement error is examined by a Monte Carlo simulation study.

Example 1: We generate the independent and identically distributed (i.i.d.) data $\{(x_i, y_i, w_i), i = 1, \ldots, n\}$ from the model

$$Y = \begin{cases} -12 + 4X + \epsilon_1, & \text{if } \mathcal{Z} = 1; \\ 12 - 4X + \epsilon_2, & \text{if } \mathcal{Z} = 2. \end{cases}$$
$$W = X + U\ ,$$

where $\mathcal{Z}$ is the latent component indicator of Y with $P(\mathcal{Z} = 1) = 0.4$, $X \sim Unif(2, 4), \epsilon_1 \sim N(0, 1)$, and $\epsilon_2 \sim N(0, 1)$. Note that the above two lines intersect each other at $X = 3$, which is the center of $Unif(2, 4)$. Therefore, the two components have some overlap around $X = 3$. To study the effect of measurement error distribution of U on the proposed estimator, we consider the following two cases:

Case I: U has a normal distribution with mean zero.

Case II: U has a double exponential distribution with mean zero.

The variance of U is chosen so that the reliability ratio (Fuller, 1987):

$$r = \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \mathrm{Var}(U)} = 0.70. \tag{3.1}$$

For each simulated data set, we estimate the mixture of regression parameters by three methods:

(a) the naive method which ignores the measurement error and maximizes (2.2) directly,

(b) the proposed new method assuming a normal measurement error (New-Norm).

(c) the proposed new method assuming a double exponential measurement error (New-Double).

As will be demonstrated in this simulation study, the proposed estimate is very robust to the distribution assumption of the measurement error.

We compare the performance of different methods based on the mean squared errors (MSE). For example, for π_1,

$$MSE(\hat{\pi}_1) = \frac{1}{500}\sum_{t=1}^{500}(\hat{\pi}_{1t} - \pi_1)^2$$

where $\hat{\pi}_{1t}$ is the estimate of π_1 based on t^{th} replication and π_1 is the true value, which is 0.4 in this example.

Similar to Bordes, Chauveau, and Vandekerkhove (2007), we use the true initial values for $\boldsymbol{\theta}$ in our GEM algorithm, in order to avoid the possible bias introduced by different starting values among replications or label switching issues (Diebolt and Robert, 1994; Stephens, 2000; Yao and Lindsay, 2009; Yao, 2012a, 2012b).

In Table 1 and 2, we report the relative efficiency between the naive method and our proposed new methods based on the ratio of the MSE of the naive method to that of the proposed estimators. From the Tables, we can see that the new methods, which incorporate the measurement error, work much better than the naive method and the gain can be substantial even for small sample size. In addition, it can be seen that the new methods are very robust to the distribution assumption of the measurement error.

Example 2 (Tone perception data): In the tone perception experiment of Cohen (1984), a pure fundamental tone was played to a trained musician with electronically generated

Table 1: Relative efficiency, Proposed vs Naive (normal measurement error).

Sample size		β_{10}	β_{11}	β_{20}	β_{21}	σ_1	σ_2	π_1
n=100	New-Norm	5.615	7.142	4.802	6.386	1.354	2.580	2.048
	New-Double	4.896	6.204	4.447	5.858	1.712	3.179	1.772
n=200	New-Norm	15.585	19.671	14.942	19.515	3.042	5.230	2.521
	New-Double	15.015	19.088	14.200	18.286	3.863	5.829	2.486
n=400	New-Norm	35.187	42.905	35.998	45.643	8.460	14.843	4.341
	New-Double	33.628	41.445	29.529	37.398	7.891	14.763	4.336

Table 2: Relative efficiency, Proposed vs Naive (double exponential measurement error).

Sample size		β_{10}	β_{11}	β_{20}	β_{21}	σ_1	σ_2	π_1
n=100	New-Norm	2.644	2.704	2.688	2.834	1.661	1.997	1.947
	New-Double	2.922	3.066	2.743	2.896	2.593	3.372	1.852
n=200	New-Norm	5.654	5.752	6.348	6.626	3.470	5.084	1.835
	New-Double	6.467	6.549	7.621	7.978	5.044	7.558	1.909
n=400	New-Norm	9.048	9.283	13.231	13.457	5.344	7.059	1.860
	New-Double	11.204	11.499	17.435	17.824	7.079	9.566	1.945

overtones added, which were determined by a stretching ratio of x. $x = 2$ corresponds to the harmonic pattern usually heard in traditional definite pitched instruments. The musician was instructed to tune an adjustable tone to the octave above the fundamental tone. y gives the ratio of the adjusted tone to the fundamental, i.e., $y = 2.0$ would be the correct tuning for all x-values. The tuning ratio, which is the ratio between adjusted tone and the fundamental tone, was recorded for 150 trials from the same musician. The purpose of this experiment was to see how this tuning ratio affects the perception of the tone. Furthermore, the experiment was designed to determine if either of two musical perception theories was reasonable (see Cohen, 1980 for more detail). A scatter plot of these data can be found in

Figure 1 and two lines are evident which correspond to the behavior indicated by the two musical perception theories.

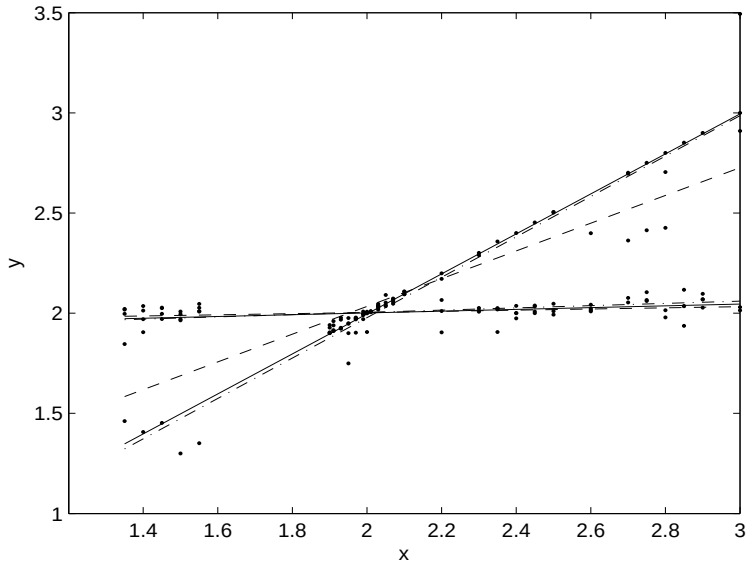

Figure 1: *The scatter plot of the original tone perception data and the fitted regression lines by different methods when the measurement error is added. The predictor is actual tone ratio and the response is the perceived tone ratio by a trained musician. The solid lines are based on the new method, the dash-dash lines are based on the naive method, and the dash-dot lines are based on the oracle method.*

To see the impact of measurement error, under the constraint (3.1), we add a measurement error $N(0, 0.3^2)$ to the predictor x. Denote by W the surrogate of X. We fit the data (W, Y) using both naive method, which ignores the measurement error, and the proposed new method assuming double exponential error. For comparison, we also add an oracle method which uses the (X, Y) directly. We plot these fits in Figure 1. From Figure 1, we can see that the regression lines estimated by the new method and the oracle method are almost overlap. However, the naive estimate has some bias for one of the component lines.

Table 3 reports the mixtures of regression parameter estimates. For comparison, we also include the new method assuming the normal measurement error, which is the true one.

Table 3: Regression parameter estimates for the tone perception data with measurement error

	Oracle	Naive method	New-Double	New-Norm
β_{10}	1.892	1.943	1.908	1.909
β_{11}	0.055	0.029	0.046	0.046
β_{20}	-0.038	0.596	-0.057	-0.043
β_{21}	1.007	0.725	1.041	1.015
σ_1	0.063	0.049	0.048	0.048
σ_2	0.114	0.281	0.201	0.203
π_1	0.674	0.747	0.752	0.751

From the table, it can be seen that both new methods have closer results to the oracle one than the naive method. The naive estimate has larger bias for β_{20} and β_{21}. In addition, both new methods, assuming different measurement errors, provide similar results. Therefore, our new method is not every sensitive to the distribution assumption of the measurement error.

4 Concluding Remarks

In this article, we proposed a method to estimate the mixture of linear regression with measurement errors by maximizing the "corrected" log-likelihood (2.4). In addition, we also proposed a generalized EM algorithm to compute the MLE. The simulation results demonstrate that the proposed estimation procedures work well and perform much better than the naive MLE which simply ignores the measurement error. Note that the generic identifiability of finite mixtures of regression models does not follow from the generic identifiability of Gaussian mixtures. It will be interesting to know whether we can use the similar identifiability conditions of Hennig (2000) for regular mixtures of regression along with the assumption on $f_U(\mathbf{u})$ and $f_X(\mathbf{x})$ to insure the identifiability of the model (2.1), when the measurement

error exists. This requires more research. In addition, it will be also interesting to investigate the asymptotic properties of proposed estimates. However, we think the proof won't be easy since it involves both the measurement error and the nonparametric estimated density $f(\mathbf{x} \mid \mathbf{w})$.

APPENDIX: PROOFS

Proof of Theorem 1:

$$\begin{aligned}
\log\hat{L}(\boldsymbol{\theta}^{(k+1)}) - \log\hat{L}(\boldsymbol{\theta}^{(k)}) &= \sum_{i=1}^{n} \log\left\{ \frac{\sum_{j=1}^{m} \pi_j^{(k+1)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k+1)})}{\sum_{l=1}^{m} \pi_j^{(k)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})} \right\} \\
&= \sum_{i=1}^{n} \log\left\{ \sum_{j=1}^{m} \frac{\pi_j^{(k)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})}{\sum_{l=1}^{m} \pi_j^{(k)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})} \frac{\pi_j^{(k+1)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k+1)})}{\pi_j^{(k)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})} \right\} \\
&= \sum_{i=1}^{n} \log\left\{ \sum_{j=1}^{m} p_{ij}^{(k+1)} \frac{\pi_j^{(k+1)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k+1)})}{\pi_j^{(k)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})} \right\} \\
&\geq \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \log\left\{ \frac{\pi_j^{(k+1)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k+1)})}{\pi_j^{(k)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})} \right\}
\end{aligned}$$

Therefore,

$$\log\hat{L}(\boldsymbol{\theta}^{(k+1)}) - \log\hat{L}(\boldsymbol{\theta}^{(k)}) \geq 0$$

if we can prove

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \log\left\{ \frac{\pi_j^{(k+1)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k+1)})}{\pi_j^{(k)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})} \right\} \geq 0.$$

Let $e_i^{(k)} = y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j^{(k)}$. Then,

$$
\begin{aligned}
&\sum_{i=1}^{n} \log\left\{ \frac{f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k+1)})}{f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})} \right\} \\
=&\sum_{i=1}^{n} \log\left\{ \frac{(\sigma_j^{(k+1)})^{-1} \int \phi\left\{(y - \mathbf{x}^T \boldsymbol{\beta}_j^{(k+1)})/\sigma_j^{(k)}\right\} f(\mathbf{x} \mid \mathbf{w}_i) d\mathbf{x}}{(\sigma_j^{(k)})^{-1} \int \phi\left\{(y - \mathbf{x}^T \boldsymbol{\beta}_j^{(k)})/\sigma_j^{(k)}\right\} f(\mathbf{x} \mid \mathbf{w}_i) d\mathbf{x}} \right\} \\
=&\sum_{i=1}^{n} \log\left\{ \int \frac{\phi\left\{e_i^{(k)}/\sigma_j^{(k)}\right\} f(\mathbf{x} \mid \mathbf{w}_i)}{\int \phi\left\{e_i^{(k)}/\sigma_j^{(k)}\right\} f(\mathbf{x} \mid \mathbf{w}_i) d\mathbf{x}} \frac{(\sigma_j^{(k+1)})^{-1} \phi\left\{e_i^{(k+1)}/\sigma_j^{(k+1)}\right\} f(\mathbf{x} \mid \mathbf{w}_i)}{(\sigma_j^{(k)})^{-1} \phi\left\{e_i^{(k)}/\sigma_j^{(k)}\right\} f(\mathbf{x} \mid \mathbf{w}_i)} d\mathbf{x} \right\} \\
=&\sum_{i=1}^{n} \log\left\{ \int \tau_{ij}^{(k+1)}(\mathbf{x}) \frac{(\sigma_j^{(k+1)})^{-1} \phi\left\{e_i^{(k+1)}/\sigma_j^{(k+1)}\right\} f(\mathbf{x} \mid \mathbf{w}_i)}{(\sigma_j^{(k)})^{-1} \phi\left\{e_i^{(k)}/\sigma_j^{(k)}\right\} f(\mathbf{x} \mid \mathbf{w}_i)} d\mathbf{x} \right\} \\
\geq&\sum_{i=1}^{n} \left\{ \int \tau_{ij}^{(k+1)}(\mathbf{x}) \log \frac{(\sigma_j^{(k+1)})^{-1} \phi\left\{e_i^{(k+1)}/\sigma_j^{(k+1)}\right\}}{(\sigma_j^{(k)})^{-1} \phi\left\{e_i^{(k)}/\sigma_j^{(k)}\right\}} d\mathbf{x} \right\} \\
\geq&0
\end{aligned}
$$

by noting that $\boldsymbol{\beta}_j^{(k+1)}$ in (2.11) and $\sigma_j^{(k+1)}$ maximizes

$$
\sum_{i=1}^{n} \left\{ \int \tau_{ij}^{(k+1)}(\mathbf{x}) \log\left[\sigma_j^{-1} \phi\left\{(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j)/\sigma_j\right\}\right] d\mathbf{x} \right\}.
$$

Since

$$
\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \log \pi_j^{(k+1)} - \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \log \pi_j^{(k)} \geq 0,
$$

we have

$$
\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \log\left\{ \frac{\pi_j^{(k+1)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k+1)})}{\pi_j^{(k)} f_j(y_i \mid \mathbf{w}_i, \boldsymbol{\theta}_j^{(k)})} \right\} \geq 0.
$$

References

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*. Chapman & Hall/CRC.

Charnigo, R. and Sun, J. (2004). Testing homogeneity in a mixture distribution via the L2-distance between competing models. *Journal of the American Statistical Association*, 99, 488-498.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society*, Ser. B, 63, 19-29.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society*, Ser. B, 66, 95-115.

Chen, J. and Kalbfleisch, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *The Canadian Journal of Statistics*, 24, 167-175.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37, 2523-2542.

Chen, R. C. H. and Liu, W. B. (2001). The consistency of estimators in finite mixture models. *Scandinavian Journal of Statistics*, 28, 603-616.

Chen, J., Tan, X., and Zhang, R. (2008). Inference for normal mixture in mean and variance. *Statistica Sincia*, 18, 443-465.

Cohen, E. (1984). Some effects of inharmonic partials on interval perception. *Music Perception*, 1, 323-349.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society*, Ser. B, 56, 363-375.

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

Fan, J. (1991b). Asymptotic normality for deconvolving kernel density estimators. *Sankhyā, Series A*, 53, 97-110.

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

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of American and Statisical Association*. 96, 194-209.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer, 2006.

Fuller. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

Goldfeld, S. M. and Quandt, R. E. (1973). A markov model for switching regression. *Journal of Econometrics*. 1, 3-16

Green, P. J. and Richardson, S. (2002). Hidden markov models and disease mapping. *Journal of American and Statisical Association*. 97, 1055-1070.

Hathaway, R.J. (1985). A constrained formulation of maximum-likelihood estimation for nomral mixture distributions. *Annals of Statistics*, 13, 795-800.

Hathaway, R.J. (1986). A constrained EM algorithm for univariate mixutres. *Journal of Statistical Computation and Simulation*, 23, 211-230.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17, 273-296.

James, L. F., Priebe, C. E., and Marchette, D. J. (2001). Consistent estimation of mixture complexity. *The Annals of Statistics*, 29, 1281-1296.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20, 1350-1360.

Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *the Journal of American Statistical Association*, 105, 1084-1092.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Ray, S. and Lindsay, B. G. (2008). Model selection in high-dimensions: A quadratic-risk based approach. *Journal of the Royal Statistical Society*, Ser. B, 70, 95-118.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society*, Ser. B, 62, 795-809.

Stefanski, L. A. and Carroll,R.J. (1986). Deconvolution kernel density estimators. Technical report.

Stefanski,L. A. and Carroll, R. J. (1990). Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Society*, Ser. B, 52, 345-359.

Wang, P., Puterman, M. L., Cockburn, I., and Le, N. (1996). Mixed poisson regression models with covariate dependent rates. *Biometrics*. 52, 381-400.

Wedel, M. and DeSarbo, W. S. (1993). A latent class binomial logit methodology for the analysis of paired comparison data. *Decision Sciences*. 24, 1157-1170.

Woo, M. and Sriram, T. N. (2006). Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101, 1475-1485.

Yao, W. (2010). A profile likelihood method for normal mixture with unequal variance. *Journal of Statistical Planning and Inference*, 140, 2089-2098.

Yao, W. (2012a). Bayesian mixture labeling and clustering. *Communications in Statistics - Theory and Methods*, 41, 403-421.

Yao, W. (2012b). Model based labeling for mixture models. *Statistics and Computing*, 22, 337-347.

Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *Journal of American Statistical Association*, 104, 758-767.