Computational Statistics and Data Analysis 93 (2016) 162–176

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Mixtures of quantile regressions

Qiang Wu [a,*], Weixin Yao [b]

[a] Department of Biostatistics, East Carolina University, Greenville, NC 27834, United States
[b] Department of Statistics, Kansas State University, Manhattan, KS 66506, United States

ARTICLE INFO

*Article history:*
Received 11 June 2013
Received in revised form 9 April 2014
Accepted 17 April 2014
Available online 5 May 2014

*Keywords:*
EM algorithm
Kernel density estimation
Mixture model
Multiple imputation
Quantile regression
Semi-parametric model

ABSTRACT

A semi-parametric mixture of quantile regressions model is proposed to allow regressions of the conditional quantiles, such as the median, on the covariates without any parametric assumption on the error densities. The median as a measure of center is known to be more robust to skewness and outliers than the mean. Modeling the quantiles instead of the mean not only improves the robustness of the model but also reveals a fuller picture of the data by fitting varying quantile functions. The proposed semi-parametric mixture of quantile regressions model is proven to be identifiable under certain weak conditions. A kernel density based EM-type algorithm is developed to estimate the model parameters, while a stochastic version of the EM-type algorithm is constructed for the variance estimation. A couple of simulation studies and several real data applications are conducted to show the effectiveness of the proposed model.

## 1. Introduction

Mixtures of regressions, or clusterwise regressions, have been a longstanding topic in the research of model-based clustering. When the population is heterogeneous and consists of several homogeneous groups, several regression models are simultaneously built to explain the relationships between the response variable and the covariates. The subjects are clustered based on the estimated classification probabilities. Some early results trace back to DeSarbo and Corn (1988), Jones and McLachlan (1992), and Arminger et al. (1999). In a classical mixture of regressions model, the conditional distribution of the response variable $Y$ given the covariates $\mathbf{x}$ can be written as

$$f(y|\mathbf{x},\boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j \phi(y; \mathbf{x}^T\boldsymbol{\beta}_j, \sigma_j^2), \tag{1.1}$$

where $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \sigma_1, \ldots, \pi_m, \boldsymbol{\beta}_m, \sigma_m)$ and $\phi(\cdot; \mu, \sigma^2)$ is the normal probability density function (pdf) with mean $\mu$ and variance $\sigma^2 > 0$. In model (1.1), the unknown parameters include $\pi_j > 0$, $\boldsymbol{\beta}_j = (\beta_{0j}, \ldots, \beta_{pj})^T$, and $\sigma_j^2 > 0$ for $j = 1, \ldots, m$ where the mixing probabilities satisfy $\sum_j \pi_j = 1$. The covariates $\mathbf{x} = (1, \tilde{\mathbf{x}}^T)^T$ usually contain a leading one for fitting intercepts. The parameters can be estimated by the maximum likelihood estimator (MLE) using an EM algorithm. A number of applications of model (1.1) can be found at Wu and Sampson (2009), Skrondal and Rabe-Hesketh (2004), and Wedel and Kamakura (2000).

Much effort has been made recently to improve the robustness of model (1.1). For example, Garcia-Escudero et al. (2010) illustrate a robust clusterwise linear regressions method which trims off a fixed proportion of outlying observations and fits

[*] Corresponding author. Tel.: +1 252 744 6047; fax: +1 252 744 6044.
*E-mail addresses:* wuq@ecu.edu (Q. Wu), wxyao@ksu.edu (W. Yao).

http://dx.doi.org/10.1016/j.csda.2014.04.014
0167-9473/© 2014 Elsevier B.V. All rights reserved.

the rest of the data via a mixture of linear regressions model. This method has improved robustness to noisy data. Following Ingrassia et al. (2012), Ingrassia et al. (2014) develop a family of twelve mixture models each inheriting from a linear $t$-cluster weighted model. Such models allow the group assignments to depend on the covariates and the component distributions to feature heavier than normal tails. Wei (2012) and Yao et al. (2014) review some robust mixture regression models and propose a new one using the $t$-distributions as its components. While being robust to heavy tails of the component distributions, this method also trims the data based on a modified Mahalanobis distance to deal with possible high leverage points. Similarly, Song et al. (2014) introduce a robust mixture model fitting by the Laplace distribution.

Most relevantly to the research in this paper, Hunter and Young (2012) consider a semi-parametric mixture of regressions model

$$f(y|\mathbf{x},\boldsymbol{\theta},\mathbf{G})=\sum_{j=1}^{m}\pi_j g(y-\mathbf{x}^T\boldsymbol{\beta}_j), \tag{1.2}$$

trying to relax the normality assumption to the greatest extent, where $\boldsymbol{\theta}=(\pi_1,\boldsymbol{\beta}_1,\ldots,\pi_m,\boldsymbol{\beta}_m)$ and $g$ is an unknown symmetric pdf with mean equal to zero (its median is also zero since $g$ is symmetric). Hunter and Young (2012) prove that model (1.2) is identifiable for the parameters $\boldsymbol{\theta}$ and the error pdf $g$ up to a permutation on $\boldsymbol{\theta}$ if $\tilde{\boldsymbol{\beta}}_j=(\beta_{1j},\ldots,\beta_{pj})$ for $j=1,\ldots,m$ are distinct vectors in $\mathbb{R}^p$ and the domain of $\tilde{\mathbf{x}}$ contains an open set in $\mathbb{R}^p$. A location-shifted model is a special example of model (1.2). In their method, the parameters and the error pdf are estimated by a kernel density based EM-type algorithm.

When the error pdf is symmetric, the mixture of mean regressions model (1.2) works well in modeling the center location functions. However, there are situations where the error pdf is asymmetric in which case it seems reasonable to consider the median or other quantiles. The median as a measure of center is considered more robust to skewness and outliers than the mean. In this paper, a novel mixture of regressions model is introduced to allow regressions of the conditional quantiles, such as the median, on the covariates. In addition, it allows the component error densities to be different. Denote by $g_j$ the error density of the $j$th component. Under a similar model specification as (1.2), the component pdf $g_j$ is assumed to have its $\tau$th quantile equal to zero. As compared to the traditional mixtures of mean regressions, the mixtures of quantile regressions are more robust to non-normal component distributions and capable of revealing more detailed structure/information of the data by fitting varying conditional quantile functions. A kernel density based EM-type algorithm is developed to estimate the model parameters. In each iteration of the algorithm, the regression parameters are updated using a weighted quantile regression method, and the error pdfs are updated by a constrained kernel density estimation method. Moreover, a stochastic version of the EM-type algorithm based on multiple imputations is constructed for the variance estimation. A couple of simulation studies and several real data applications are conducted to demonstrate the effectiveness of the proposed model.

The rest of this article is organized as follows. In Section 2, we introduce the new mixture of quantile regressions model, prove its identifiability result, and detail the new kernel density based EM-type algorithm. In Section 3, we provide the stochastic EM-type algorithm for the variance estimation. In Sections 4 and 5, we present the simulation studies and the real data applications. Finally, some discussions are given in Section 6.

## 2. Mixtures of quantile regressions

The model setting for a mixture of $\tau$th quantile regressions is as follows. Let $Z$ be a latent class variable with $\Pr(Z=j|\mathbf{x})=\pi_j>0$ for $j=1,\ldots,m$, where $\mathbf{x}=(1,\tilde{\mathbf{x}}^T)^T$ is a $(p+1)$-dimensional vector of covariates with a leading one for fitting intercepts. Given $Z=j$, the response variable $Y$ depends on the covariates $\mathbf{x}$ through

$$Y=\mathbf{x}^T\boldsymbol{\beta}_j(\tau)+\epsilon_j(\tau), \tag{2.1}$$

where $\boldsymbol{\beta}_j(\tau)=(\beta_{0j}(\tau),\ldots,\beta_{pj}(\tau))^T$ are the $\tau$th quantile regression coefficients for the $j$th component. The errors $\epsilon_j(\tau)$ are assumed to be independent of $\mathbf{x}$ and have pdfs $g_j(\cdot)$ whose $\tau$th quantiles are equal to zero. There is no additional constraint on the error pdfs as they are going to be estimated non-parametrically. We assume that the number of components $m>1$ is known in advance. A regular choice for $\tau$ is 0.5 which corresponds to a median regression but it does not have to be. Since the model deals with only one quantile at a time, we suppress its dependency on $\tau$ in the following discussion for the notational ease.

Next, we prove that the mixture of $\tau$th quantile regressions model (2.1) is identifiable for $\boldsymbol{\theta}=(\pi_1,\boldsymbol{\beta}_1,\ldots,\pi_m,\boldsymbol{\beta}_m)$ and $\mathbf{G}=(g_1,\ldots,g_m)$ up to the same permutation on $\boldsymbol{\theta}$ and $\mathbf{G}$ if $\tilde{\boldsymbol{\beta}}_j=(\beta_{1j},\ldots,\beta_{pj})$ for $j=1,\ldots,m$ are distinct vectors in $\mathbb{R}^p$ and the domain of $\tilde{\mathbf{x}}$ contains an open set in $\mathbb{R}^p$. Necessary conditions and the identifiability of the model (2.1) are summarized in Theorem 2.1 whose proof is given in the Appendix. Of course, we have a flexibility to assume an equal error density $g_1=\cdots=g_m$. In this case, a pooled density estimate can be found during the estimation. But the regression parameters $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_m$ must be distinct for the identifiability purpose.

**Theorem 2.1.** *Suppose that in the mixture of quantile regressions model (2.1), the domain of $\tilde{\mathbf{x}}$ contains an open set in $\mathbb{R}^p$, $0<\pi_j<1$, and $\tilde{\boldsymbol{\beta}}_j=(\beta_{1j},\ldots,\beta_{pj})$ are distinct vectors in $\mathbb{R}^p$ for $j=1,\ldots,m$. Then the parameters $\pi_j$, $\boldsymbol{\beta}_j$, and the error pdfs $g_j(\cdot)$ for $j=1,\ldots,m$ are uniquely determined, up to a permutation, by the conditional density $f(y|\mathbf{x})$.*

Hunter and Young (2012) have used similar conditions to prove the identifiability of the model (1.2) under an additional assumption that the component error densities are the same. In a recent but not yet published work, Wang et al. (2012) prove the identifiability of the model (1.2) under more general conditions without requiring the component error densities to be identical or symmetric. In Theorem 2.1, we extend Wang et al. (2012)'s result to the proposed mixture of quantile regressions model.

Since no parametric assumption is made on the component error densities in (2.1), there is no likelihood function to work with in estimating the model parameters. Instead, we propose a kernel density based EM-type algorithm to estimate the parameters $\boldsymbol{\theta}$ and the error pdfs $\mathbf{G}$. Let $\mathcal{X} = \{(Y_i, \mathbf{x}_i),\ i = 1, \ldots, n\}$ be a random sample from the model (2.1).

**Algorithm 2.1.** With some initial parameter values $\hat{\boldsymbol{\theta}}^{(b)}$ and error pdfs $\widehat{\mathbf{G}}^{(b)}$ for $b = 0, 1, \ldots,$ the $(b+1)$th iteration of the kernel density based EM-type algorithm updates the parameters and the error pdfs through:

**E Step.** *The E step computes the classification probabilities according to*

$$p_{ij}^{(b+1)} = \Pr(Z_i = j|\mathcal{X}, \hat{\boldsymbol{\theta}}^{(b)}, \widehat{\mathbf{G}}^{(b)}) = \frac{\hat{\pi}_j^{(b)}\hat{g}_j^{(b)}(e_{ij}^{(b)})}{\sum_{l=1}^{m}\hat{\pi}_l^{(b)}\hat{g}_l^{(b)}(e_{il}^{(b)})}, \tag{2.2}$$

*where $e_{ij}^{(b)} = Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_j^{(b)}$. The initial error pdfs can be taken as, for example, the normal densities. But we can also bypass this requirement by assigning initial values directly to the classification probabilities $\{p_{ij}\}$ and proceed to the M step.*

**M Step.** *The M step updates the parameters $\boldsymbol{\theta}$ according to*

$$\hat{\pi}_j^{(b+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(b+1)}, \tag{2.3}$$

*and*

$$\hat{\boldsymbol{\beta}}_j^{(b+1)} = \underset{\boldsymbol{\beta}_j}{\operatorname{argmin}} \sum_{i=1}^{n} p_{ij}^{(b+1)}\rho_\tau(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j), \tag{2.4}$$

*where $\rho_\tau(u) = u(\tau - I(u < 0))$ following Koenker and Bassett (1978). The existence of (2.4) is given by Bai et al. (1992) and a solution to (2.4) is readily available in many statistical packages such as the R package quantreg. To update the error pdfs $\mathbf{G}$, we implement a constrained kernel density estimation*

$$\hat{g}_j^{(b+1)}(t) = \sum_{i=1}^{n}\sum_{l=1}^{2} w_{lj}^{(b+1)}p_{ij}^{(b+1)}K_h(t - e_{ij}^{(b+1)})I_l(e_{ij}^{(b+1)}), \tag{2.5}$$

*where $e_{ij}^{(b+1)} = Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_j^{(b+1)}$, $K_h(t) = h^{-1}K(t/h)$, and $K(t)$ is a kernel function such as the Gaussian kernel. In (2.5), we define $I_1(u) = I(u \le 0)$ and $I_2(u) = I(u > 0)$. The constants $w_{1j}^{(b+1)}$ and $w_{2j}^{(b+1)}$ are found by solving a system of linear equations*

$$\sum_{i=1}^{n}\sum_{l=1}^{2} w_{lj}^{(b+1)}p_{ij}^{(b+1)}I_l(e_{ij}^{(b+1)}) = 1 \tag{2.6}$$

$$\sum_{i=1}^{n}\sum_{l=1}^{2} w_{lj}^{(b+1)}p_{ij}^{(b+1)}v_{ij}^{(b+1)}I_l(e_{ij}^{(b+1)}) = \tau, \tag{2.7}$$

*where $v_{ij}^{(b+1)} = \int_{-\infty}^{0} K_h(t - e_{ij}^{(b+1)})dt$.*

Upon convergence, Algorithm 2.1 finds the parameter estimates $\hat{\boldsymbol{\theta}}$ and the error density estimates $\widehat{\mathbf{G}}$ that best identify the clusters of the quantile regressions (2.4) given that the error densities are in the kernel density forms of (2.5). Algorithm 2.1 is not designed to minimize or maximize a unique objective function, so its convergence criterion is based on the changes in consecutive parameter values, that is, the algorithm is claimed convergence if the sum of the absolute changes in consecutive parameter values does not exceed a pre-specified limit. As one reviewer mentions that because the parameters $\pi_1, \boldsymbol{\beta}_1, \ldots, \pi_m, \boldsymbol{\beta}_m$ are usually not on the same scale, the limit can be also set for the sum of the relative absolute changes.

The constrained kernel density estimate (2.5) and the weights found by solving (2.6) and (2.7) are used to ensure that the density estimates $\hat{g}_j(\cdot)$ always have their $\tau$th quantiles equal to zero. This approach is inspired by the method of Hall and Presnell (1999). In Section 3, we show that the weights found in such a way are asymptotically equivalent to the uniform weights of $1/n$ in the classical kernel density estimation.

A variation can be made to the above algorithm if we employ a homogeneity assumption that all error pdfs are identical, i.e., $g_j = g$ for $j = 1, \ldots, m$. In this case, the E step of the algorithm computes the classification probabilities by

$$p_{ij}^{(b+1)} = \Pr(Z_i = j|\mathcal{X}, \hat{\boldsymbol{\theta}}^{(b)}, \widehat{\mathbf{G}}^{(b)}) = \frac{\hat{\pi}_j^{(b)}\hat{g}^{(b)}(e_{ij}^{(b)})}{\sum_{l=1}^{m}\hat{\pi}_l^{(b)}\hat{g}^{(b)}(e_{il}^{(b)})}.$$

The M step updates the parameters using (2.3) and (2.4) and updates the error pdfs using

$$\hat{g}^{(b+1)}(t) = \sum_{i,j}\sum_{l=1}^{2} w_l^{(b+1)}p_{ij}^{(b+1)}K_h(t - e_{ij}^{(b+1)})I_l(e_{ij}^{(b+1)}), \tag{2.8}$$

where $w_1^{(b+1)}$ and $w_2^{(b+1)}$ are found by solving

$$\sum_{i,j}\sum_{l=1}^{2} w_l^{(b+1)}p_{ij}^{(b+1)}I_l(e_{ij}^{(b+1)}) = 1$$
$$\sum_{i,j}\sum_{l=1}^{2} w_l^{(b+1)}p_{ij}^{(b+1)}v_{ij}^{(b+1)}I_l(e_{ij}^{(b+1)}) = \tau.$$

If the homogeneity assumption is reasonable, then using the pooled estimate $\hat{g}(\cdot)$ may help improve the efficiency of the model estimation. Otherwise, the model estimation and the calculated classification probabilities could be biased. One might start the analysis using Algorithm 2.1 and switch to the equal error density one if there is strong evidence supporting the homogeneity assumption.

## 3. Asymptotics and variance estimation

In this section, we discuss some asymptotic properties of the proposed model estimate from Algorithm 2.1 and introduce a stochastic version of Algorithm 2.1 to estimate the variance–covariance matrix of the parameter estimates.

First, Algorithm 2.1 uses a constrained kernel density estimation method to ensure the estimated densities having their $\tau$th quantiles equal to zero. The idea of this method is inspired by Hall and Presnell (1999). Let $\{X_1, \ldots, X_n\}$ be a random sample from an unknown density $f$ whose $\tau$th quantile equals zero. A constrained kernel density estimate of $f(x)$ can be written as

$$\hat{f}(x|w) = \sum_{i=1}^{n} w_i K_h(x - X_i), \tag{3.1}$$

where $w = (w_1, \ldots, w_n)$ and $\sum_i w_i = 1$. The weights satisfy a constraint $\sum_{i=1}^{n} w_i v_i = \tau$ where $v_i = \int_{-\infty}^{0} K_h(x - X_i)dx$ so that $\hat{f}$ has its $\tau$th quantile equal to zero. Obviously, there are infinitely many weights satisfying the constraint. Hall and Presnell (1999) show, for example, that the ones that minimize the Kullback–Leibler divergence $D(w) = -\sum_i \log(nw_i)/n$ from the uniform weights $(1/n, \ldots, 1/n)$ are $w_i = \{n - (v_i - \tau)c\}^{-1}$ for some constant $c$. They prove that these weights are close to the uniform weights in the sense that $\max_i nw_i$ converges to one in probability as $n \to \infty$. However, finding $c$ involves an iterative Newton–Raphson algorithm. For computational efficiency, we propose to constrain the weights by $w_i = a$ if $X_i \le 0$ or $b$ if $X_i > 0$ for some constants $a$ and $b$. To find $a$ and $b$, it suffices to solve a system of linear equations

$$\sum_{i=1}^{n} aI(X_i \le 0) + \sum_{i=1}^{n} bI(X_i > 0) = 1$$
$$\sum_{i=1}^{n} av_iI(X_i \le 0) + \sum_{i=1}^{n} bv_iI(X_i > 0) = \tau.$$

Using a change of integrals and a Taylor series expansion, we can show that $E[v_iI(X_i \le 0)] = \tau - O(h)$ and $E[v_iI(X_i > 0)] = O(h)$. If we let the bandwidth $h \to 0$ as $n \to \infty$, then our weights are also close to the uniform weights because both $na$ and $nb$ converge to one in probability as $n \to \infty$.

Second, the inference on quantile regressions generally follows three methods: sparsity, rank, and resampling. Let $\{(Y_i, \mathbf{x}_i),\ i = 1, \ldots, n\}$ be a random sample from a quantile regression model whose $\tau$th quantile function is given by $Q_\tau(Y|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$ (we omit the dependence of $\boldsymbol{\beta}$ on $\tau$). The parameter estimates can be found by solving

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}). \tag{3.2}$$

The asymptotic normality of $\hat{\boldsymbol{\beta}}$ is established by Koenker and Bassett (1978) under an assumption that the errors $e_i = Y_i - \mathbf{x}_i^T\boldsymbol{\beta}$ are independent and identically distributed (i.i.d.). The asymptotic result is later extended to heteroscedastic models by Koenker and Zhao (1994). He and Shao (1996) prove a more general asymptotic result when the errors are independent but not necessarily identically distributed. The asymptotic variance–covariance matrix of $\hat{\boldsymbol{\beta}}$ is given generally by

$$V_\tau = \tau(1-\tau)(\mathbf{X}^T F\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T F\mathbf{X})^{-1},$$

where $\mathbf{X}$ is the design matrix and $F = \text{diag}\{f_1(0), \ldots, f_n(0)\}$ is a diagonal matrix of error densities evaluated at zero. For the i.i.d. case, it reduces to

$$V_\tau = \frac{\tau(1-\tau)}{f^2(0)}(\mathbf{X}^T\mathbf{X})^{-1}, \tag{3.3}$$

where $f$ represents the common error density. Koenker (1994) has a discussion on directly estimating $V_\tau$ via approximating the sparsity function $1/f(0)$. He also considers constructing confidence intervals using inversion of rank tests which is later extended to a location-scale model by Koenker and Machado (1999). Bootstrap resampling methods involve repeated applications of quantile regressions to resampled data. Parzen et al. (1994) describe a resampling method using a pivotal quantity. More recently, Kocherginsky et al. (2005) develop a more time efficient Markov chain marginal bootstrap method. The sparsity, rank, and Parzen et al. (1994)'s resampling methods are available in the R package quantreg. Kocherginsky et al. (2005)'s method is available in the R package rqmcmb2. Since in the mixture of quantile regressions model the sparsity function $1/f(0)$ is readily available from the kernel density estimation, we choose to implement the sparsity method (3.3) in estimating the variance–covariance matrix of the parameter estimates.

Algorithm 2.1 does not allow an estimation of the variance–covariance matrix, nor does the semi-parametric model setting allow a likelihood analysis. In such cases, one usually turns to resampling methods such as the jackknife or bootstrapping. Particularly for regression parameters, Wu (1986) shows that resampling variances from Case bootstrapping and ordinary jackknife are biased but those from residual bootstrapping are consistent. It is unclear whether parallel results hold true for the mixture of regressions model. But the consistency of the parameter estimates, as well as the variance estimates, in the mixture of regressions model also depends on data patterns. See Section 6 for more discussions on the model consistency. Another competing approach is to bootstrap from the estimated model, i.e., $\hat{f}(y|\mathbf{x}_i, \hat{\boldsymbol{\theta}}, \widehat{\mathbf{G}}) = \sum_j \hat{\pi}_j \hat{g}_j(y - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_j)$. It involves drawing the latent class variables $\mathbf{Z}_i^{(b)} = (Z_{i1}^{(b)}, \ldots, Z_{im}^{(b)})$ from a multinomial distribution $MN(\hat{\pi}_1, \ldots, \hat{\pi}_m, 1)$ and drawing the residuals $e_{ij}^{(b)}$ from the estimated error distributions $\hat{g}_j(t)$ given $\mathbf{Z}_i$. However, when the latent class variables $\mathbf{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ are treated as missing data, multiple imputation methods can be implemented if the variance–covariances of the parameter estimates are available for the augmented data $\{\mathbf{Z}, \mathcal{X}\}$. In the mixture of quantile regressions model, if the error pdfs $\mathbf{G}$ are known, then a Bayesian multiple imputation iterates through an imputation step drawing $\mathbf{Z}^{(b+1)} \sim p(\mathbf{Z}|\mathcal{X}, \tilde{\boldsymbol{\theta}}^{(b)})$ and a posterior step drawing $\tilde{\boldsymbol{\theta}}^{(b+1)} \sim p(\boldsymbol{\theta}|\mathcal{X}, \mathbf{Z}^{(b+1)})$ for $b = 0, 1, \ldots$. Unfortunately, the posterior distribution $p(\boldsymbol{\theta}|\mathcal{X}, \mathbf{Z}^{(b)})$ is not available when the error pdfs $\mathbf{G}$ have unspecified forms. A result from Rubin (1987) shows that imputing $\mathbf{Z}^{(b+1)} \sim p(\mathbf{Z}|\mathcal{X}, \hat{\boldsymbol{\theta}}, \widehat{\mathbf{G}})$ is improper and underestimates the variances. Similarly, imputing $\mathbf{Z}^{(b+1)} \sim p(\mathbf{Z}|\mathcal{X}, \hat{\boldsymbol{\theta}}^{(b)}, \widehat{\mathbf{G}}^{(b)})$ is somewhat improper too. Little and Rubin (2002, pp. 214–217) list some alternative methods for this purpose. For example, one can impute $\mathbf{Z}^{(b+1)} \sim p(\mathbf{Z}|\mathcal{X}, \tilde{\boldsymbol{\theta}}^{(b)}, \widetilde{\mathbf{G}}^{(b)})$ where $\tilde{\boldsymbol{\theta}}^{(b)}$ and $\widetilde{\mathbf{G}}^{(b)}$ are the parameter and error density estimates from a bootstrapped sample $\mathcal{X}_{\text{boot}}^{(b)}$ of $\mathcal{X}$. This method is essentially the Case bootstrapping method.

Finally, we propose a stochastic kernel density based EM-type algorithm to estimate the variance–covariance matrix of the parameter estimates from the mixture of quantile regressions model. It is a multiple imputation method and more time efficient than any of the above mentioned resampling methods. The parameter estimates from Algorithm 2.1 can be treated as initial values for Algorithm 3.1. Following a few burn-in steps, Algorithm 3.1 is carried out for a large number of steps. In practice, a thinning can be used to reduce the autocorrelation among the multiple imputations.

**Algorithm 3.1.** The $(b+1)$th iteration of the stochastic algorithm for $b = 0, 1, \ldots$ starts with some initial parameter values $\tilde{\boldsymbol{\theta}}^{(b)}$ and error pdfs $\widetilde{\mathbf{G}}^{(b)}$.

**E Step.** *The E step draws the random latent class variables from some multinomial distributions as*

$$Z_i^{(b+1)} = (Z_{i1}^{(b+1)}, \ldots, Z_{im}^{(b+1)}) \sim MN(p_{i1}^{(b+1)}, \ldots, p_{im}^{(b+1)}, 1),$$

*for $i = 1, \ldots, n$ where $p_{ij}^{(b+1)}$ are from (2.2) with $\tilde{\boldsymbol{\theta}}^{(b)}$ and $\widetilde{\mathbf{G}}^{(b)}$ substituted for $\hat{\boldsymbol{\theta}}^{(b)}$ and $\widehat{\mathbf{G}}^{(b)}$. Only one of $Z_{i1}^{(b+1)}, \ldots, Z_{im}^{(b+1)}$ can be one and all others are zero. This random draw temporally classifies the ith subject into one of the m clusters.*

**M Step.** *The M step first finds the parameter updates $\hat{\boldsymbol{\theta}}^{(b+1)}$ using (2.3) and (2.4) with all occurrences of $p_{ij}^{(b+1)}$ replaced by $Z_{ij}^{(b+1)}$. It then randomly draws the parameter values as*

$$\tilde{\boldsymbol{\theta}}^{(b+1)} \sim N(\hat{\boldsymbol{\theta}}^{(b+1)}, \text{Var}(\hat{\boldsymbol{\theta}}^{(b+1)})), \tag{3.4}$$

*where* $\mathrm{Var}(\hat{\boldsymbol{\pi}}^{(b+1)}) = \hat{\boldsymbol{\pi}}^{(b+1)}\hat{\boldsymbol{\pi}}^{(b+1)T}/n$, $\mathrm{Var}(\hat{\boldsymbol{\beta}}_j^{(b+1)})$ *are given by* (3.3)*, and the covariances among* $\tilde{\boldsymbol{\pi}}^{(b+1)}, \tilde{\boldsymbol{\beta}}_j^{(b+1)}, \ldots, \tilde{\boldsymbol{\beta}}_m^{(b+1)}$ *are zero. To ensure that* $\tilde{\boldsymbol{\pi}}^{(b+1)}$ *are probabilities that sum to one, we can repeatedly draw its first* $m - 1$ *elements from* (3.4) *until they are all between zero and one and sum to less than one. Random draws* $\widetilde{\mathbf{G}}^{(b+1)}$ *of the error densities can be obtained from a bootstrapped sample of the residuals* $e_{ij}^{(p+1)} = Y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j^{(b+1)}$ *using* (2.5)–(2.7) *with all occurrences of* $p_{ij}^{(b+1)}$ *replaced by* $Z_{ij}^{(b+1)}$*. The residual bootstrapping may help improve the consistency of the variance estimate. An empirical method which does not enjoy the residual bootstrapping property is to compute* $\tilde{\boldsymbol{\theta}}^{(b+1)} = \hat{\boldsymbol{\theta}}_{\mathrm{boot}}^{(b+1)}$ *and* $\widetilde{\mathbf{G}}^{(b+1)} = \widehat{\mathbf{G}}_{\mathrm{boot}}^{(b+1)}$ *from a bootstrapped sample* $(\mathcal{X}_{\mathrm{boot}}^{(b+1)}, \mathbf{Z}_{\mathrm{boot}}^{(b+1)})$ *of* $(\mathcal{X}, \mathbf{Z}^{(b+1)})$.

According to Little and Rubin (2002, pp. 209–212), with the results from $B$ iterations of the stochastic algorithm, the variance–covariance matrix of the parameter estimates $\hat{\boldsymbol{\theta}}$ can be estimated by

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = VW(\hat{\boldsymbol{\theta}}) + (1 + B^{-1})VB(\hat{\boldsymbol{\theta}}),$$

where $\bar{\boldsymbol{\theta}} = \sum_b \hat{\boldsymbol{\theta}}^{(b)}/B$,

$$VB(\hat{\boldsymbol{\theta}}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}^{(b)} - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(b)} - \bar{\boldsymbol{\theta}})^T,$$

$$VW(\hat{\boldsymbol{\pi}}) = \frac{1}{Bn}\sum_{b=1}^{B}\hat{\boldsymbol{\pi}}^{(b)}\hat{\boldsymbol{\pi}}^{(b)T},$$

and

$$VW(\hat{\boldsymbol{\beta}}_j) = \frac{1}{B}\sum_{b=1}^{B}\mathrm{Var}(\hat{\boldsymbol{\beta}}_j^{(b)}).$$

Given $(\mathcal{X}, \mathbf{Z}^{(b)})$, the estimated variance–covariance matrix $\hat{\boldsymbol{\pi}}^{(b)}\hat{\boldsymbol{\pi}}^{(b)}/n$ of the mixing probabilities follows from the multinomial model while $\mathrm{Var}(\hat{\boldsymbol{\beta}}_j^{(b)})$ is computed using (3.3) for each component. Because $\tilde{\boldsymbol{\pi}}^{(b)}, \tilde{\boldsymbol{\beta}}_1^{(b)}, \ldots, \tilde{\boldsymbol{\beta}}_m^{(b)}$ are uncorrelated given $(\mathcal{X}, \mathbf{Z}^{(b)})$, the covariances among $\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\beta}}_j, \ldots, \hat{\boldsymbol{\beta}}_m$ can be estimated simply by the covariances among $\hat{\boldsymbol{\pi}}^{(b)}, \hat{\boldsymbol{\beta}}_j^{(b)}, \ldots, \hat{\boldsymbol{\beta}}_m^{(b)}$ for $b = 1, \ldots, B$. Thus we have an estimated variance–covariance matrix of the parameter estimates $\hat{\boldsymbol{\theta}}$.

The variation in the parameter estimates $\hat{\boldsymbol{\theta}}$ comes from two parts. One is from the uncertainty in group memberships $((1 + B^{-1})VB(\hat{\boldsymbol{\theta}}))$ and the other is from the sampling variations in the data $(VW(\hat{\boldsymbol{\theta}}))$. A ratio of the between-imputation variance to the total variance $|(1 + B^{-1})VB(\hat{\boldsymbol{\theta}})|/|\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})|$ quantifies the fraction of missing information due to the unknown group memberships and so provides an indicator of the separability of the clusters with values closer to zero meaning that the clusters are better separated.

## 4. Simulations

In this section, we conduct a couple of simulation studies to illustrate the effectiveness of the proposed mixture of quantile regressions model and compare the variance estimates from Algorithm 3.1 to those from Case bootstrapping and Model bootstrapping. All simulations use a $\tau = 0.5$, i.e., a mixture of median regressions.

First, we simulate $n = 100, 300$, and $600$ i.i.d. data points $\{(x_i, y_i),\ i = 1, \ldots, n\}$ from the following model

$$Y = \begin{cases} 10 - 10x + \epsilon_1 & \text{if } Z = 1 \\ -10 + 10x + \epsilon_2 & \text{if } Z = 2, \end{cases} \tag{4.1}$$

where $Z = 1$ or $2$ indicate two components with $\Pr(Z = 1) = \Pr(Z = 2) = 0.5$. The covariate $x$ is simulated from the uniform $U(0, 1)$ distribution and the error terms $\epsilon_1$ and $\epsilon_2$ are simulated according to the following model

$$\epsilon_1, \epsilon_2 \sim 0.5N(-1, 1^2) + 0.5N(2, 2^2). \tag{4.2}$$

The error density is chosen so that its median is equal to zero and, consequently, the mixture model uses the median regressions ($\tau = 0.5$) as components. For the kernel density estimation we use a bandwidth of $h = 1.06\sigma n^{-1/5}$ (see Silverman, 1986) where $n$ is the total sample size and $\sigma$ is the standard deviation of the error density if an equal error density is assumed. If unequal error densities are assumed, then $h_j = 1.06\sigma_j n_j^{-1/5}$ for $j = 1, 2$ are calculated separately for each group. We run 500 replicates in order to show properties of the parameter estimates. For each replicate, 500 imputation steps of the stochastic algorithm are recorded. For a comparison, results from 500 Case bootstrapping and Model bootstrapping samples are obtained.

During simulations, we observe that the initial values for Algorithm 2.1 are very important. The nature of the mixture models prohibits using constant initial values like $p_{ij}^{(0)} \equiv p$. The initial values must provide more or less a tendency to separate the groups. Simulating the initial values from the uniform distribution $p_{ij}^{(0)} \sim U(0, 1)$ usually works but it could lead to a saddle-point-like solution. This difficulty can potentially be overcome by starting Algorithm 2.1 from multiple

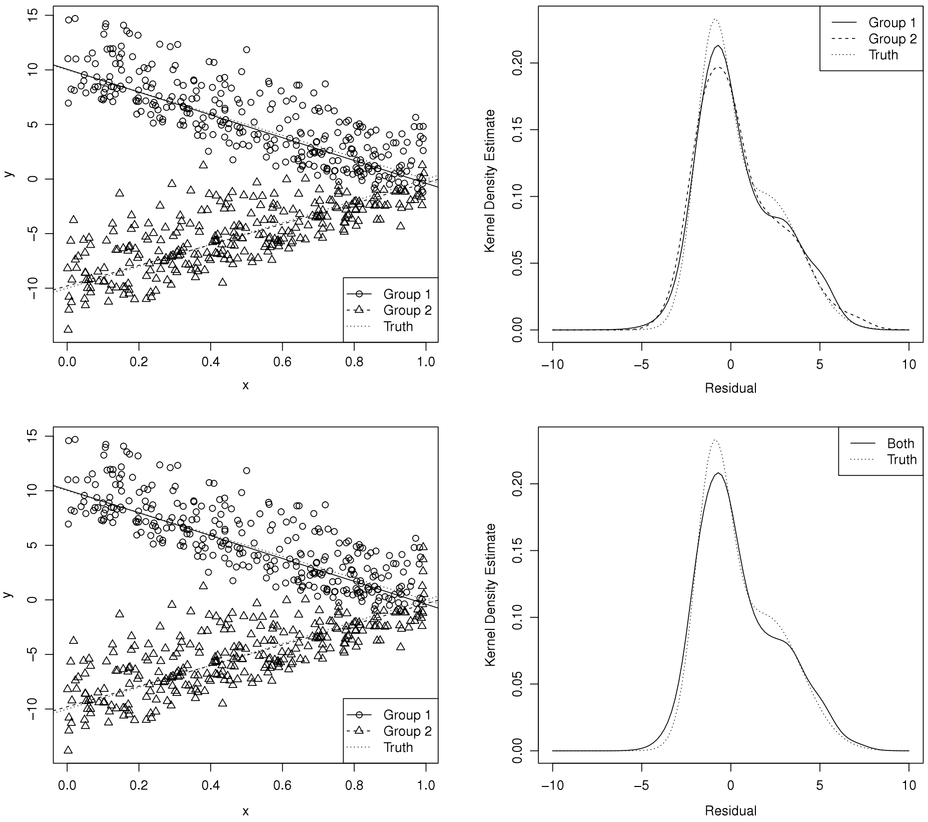

**Fig. 1.** A typical simulation result for $n = 600$. The top two figures illustrate the estimated mixture of quantile regressions model and the two unequal error density estimates superimposed with the true error density in dotted lines. The cases are classified using a cutoff of $p = 0.5$ and shown in different symbols. Similar results are given in the bottom two figures when an equal error density is assumed.

initial values. When starting from multiple initial values, Bai et al. (2012) test successfully a practical solution to choose the root which most of the initial values converge to. For our simulations, we start Algorithm 2.1 from near the true values to maximize the convergence.

Fig. 1 illustrates a typical simulation result for $n = 600$. Both assumptions of an equal error density and unequal error densities are attempted. Algorithm 2.1 successfully identifies the two median regression components and the error densities under both assumptions. The two median regression lines are plotted against the two true median functions. The cases are clustered using $p = 0.5$ as a cutoff and shown in different symbols. The clustering performance is as good as expected. Under both error density assumptions, the algorithm estimates the error densities reasonably well.

Table 1 summarizes the parameter estimates using Algorithm 2.1 on the 500 simulated data sets. Means and variances of the parameter estimates are given under the two error density assumptions and the three sample sizes. The variances in Table 1 can be treated as gold standards against which the three variance estimates in Table 2 are compared. Table 2 summarizes the variance estimates using Case bootstrapping, Model bootstrapping, and Algorithm 3.1 on the 500 simulated data sets. Means and variances of the variance estimates are also tabulated under the two error density assumptions and the three sample sizes. Discrepancies between such variance estimates and the gold standards can be measured in the mean squared error (MSE). The Case bootstrapping variances tend to overestimate the true ones because their means tend to be larger than the gold standards. Variance estimates from Model bootstrapping and Algorithm 3.1 are comparable and generally better than those from Case bootstrapping in terms of the MSE except that Algorithm 3.1 performs slightly better than Model bootstrapping in estimating the variances of the mixing probability $\hat{\pi}_1$. A critical advantage of Algorithm 3.1 is that it is much faster than both Case and Model bootstrapping methods.

Fig. 2 contains pairwise comparisons of variance estimates for $\hat{\pi}_1$, $\hat{\beta}_{10}$ and $\hat{\beta}_{11}$ over 500 replicates for $n = 600$. The figures for $\hat{\beta}_{20}$ and $\hat{\beta}_{21}$ are similar to those of $\hat{\beta}_{10}$ and $\hat{\beta}_{11}$ and so not shown here. These figures show that variance estimates from Algorithm 3.1 are more stable and generally closer to the gold standards than the Case bootstrapping ones. Similar observations hold true between the Model bootstrapping ones and the Case bootstrapping ones.

Another observation from Table 1 is that the estimated group proportion $\hat{\pi}_1$ from Algorithm 2.1 is somewhat biased. For this reason, other parameter estimates are likely to be biased too although the biases are not striking. This is due to the unbalanced nature of the simulated data which we discuss in details in Section 6.

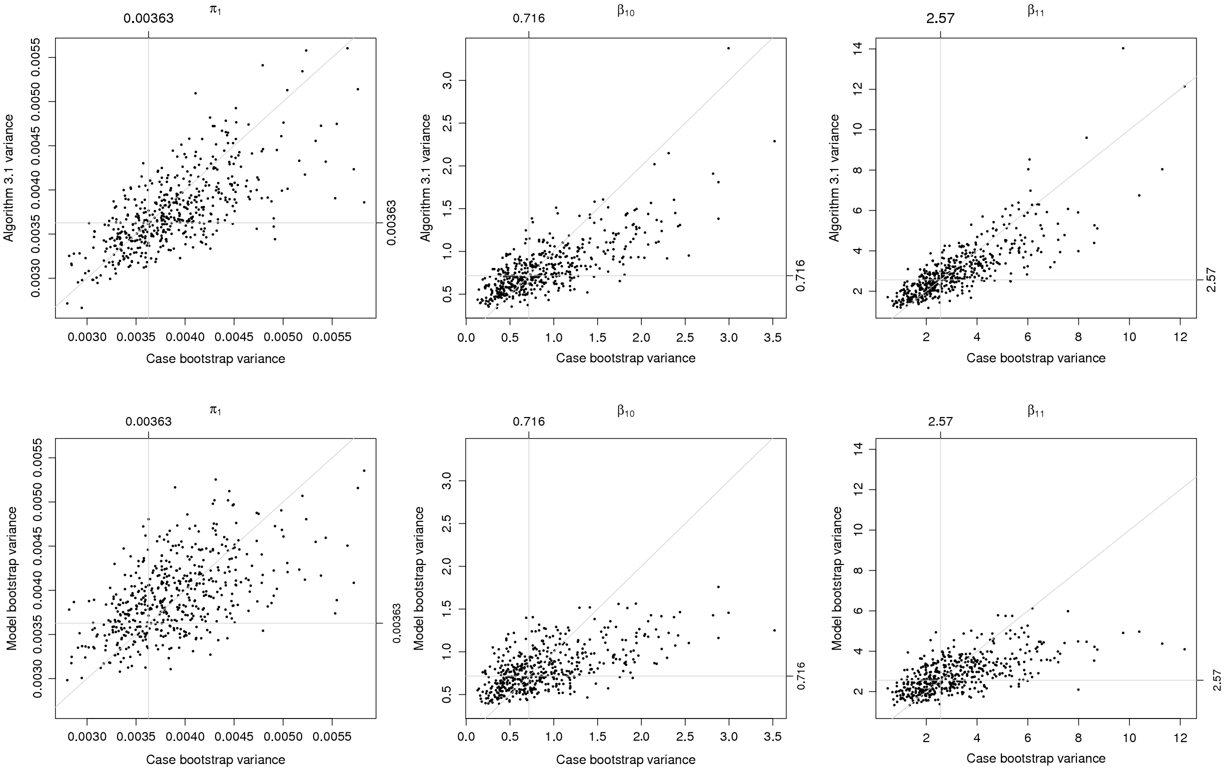

**Fig. 2.** Pairwise comparisons of variance estimates for $\hat{\pi}_1$, $\hat{\beta}_{10}$ and $\hat{\beta}_{11}$ over 500 replicates for $n = 600$. The vertical and horizontal lines represent the gold standards of variance estimates from the 500 simulation replications. The diagonal line is the 45° line of equal variances.

**Table 1**
A summary of parameter estimates using Algorithm 2.1 on 500 simulations from model (4.1) and (4.2). Results are displayed under the two error density assumptions and the three sample sizes. Variances in this table can be treated as gold standards.

| Error pdf | $n$ | Estimate | $\pi_1 = 0.5$ | $\beta_{10} = 10$ | $\beta_{11} = -10$ | $\beta_{20} = -10$ | $\beta_{21} = 10$ |
|---|---|---|---|---|---|---|---|
| Unequal | 100 | Mean | 0.510 | 10.1 | −10.2 | −10 | 10.2 |
| | | Var | 3.63E−03 | 0.716 | 2.57 | 0.77 | 2.66 |
| | 300 | Mean | 0.508 | 10.1 | −10.2 | −10 | 10.1 |
| | | Var | 1.19E−03 | 0.231 | 0.868 | 0.205 | 0.73 |
| | 600 | Mean | 0.508 | 10 | −10.1 | −10 | 10.1 |
| | | Var | 5.70E−04 | 0.101 | 0.424 | 0.104 | 0.404 |
| Equal | 100 | Mean | 0.506 | 10.1 | −10.1 | −9.98 | 10.2 |
| | | Var | 3.62E−03 | 0.745 | 2.77 | 0.746 | 2.57 |
| | 300 | Mean | 0.505 | 10.1 | −10.2 | −10 | 10.1 |
| | | Var | 1.16E−03 | 0.239 | 0.915 | 0.208 | 0.729 |
| | 600 | Mean | 0.506 | 10 | −10.1 | −10 | 10.1 |
| | | Var | 5.55E−04 | 0.103 | 0.457 | 0.102 | 0.393 |

Second, in order to illustrate the full capacity of Algorithms 2.1 and 3.1. We conduct a second set of 500 simulations of a sample size $n = 300$. The data $\{(x_{1i}, x_{2i}, y_i),\ i = 1, \ldots, n\}$ are simulated from the following model

$$Y = \begin{cases} -20x_1 - 20x_2 + \epsilon_1 & \text{if } Z = 1 \\ \epsilon_2 & \text{if } Z = 2 \\ 20x_1 + 20x_2 + \epsilon_3 & \text{if } Z = 3, \end{cases} \tag{4.3}$$

where $Z = 1, 2,$ or $3$ indicate three components with $\Pr(Z = 1) = \Pr(Z = 2) = \Pr(Z = 3) = 1/3$. The covariates $x_1$ and $x_2$ are again simulated from the uniform $U(0, 1)$ distribution and the error terms $\epsilon_1$, $\epsilon_2$, and $\epsilon_3$ are simulated according to the following models

$$\begin{aligned} &\epsilon_1 \sim \exp(N(1, 1)) - \exp(1) \\ &\epsilon_2 \sim \exp(N(1, 0.5)) - \exp(1) \\ &\epsilon_3 \sim \exp(N(1, 0.25)) - \exp(1). \end{aligned} \tag{4.4}$$

**Table 2**
A summary of variance estimates using Case bootstrapping, Model bootstrapping, and Algorithm 3.1 on 500 simulations from model (4.1) and (4.2). Each variance estimate is based on 500 bootstrap samples or MCMC iterations. Results are displayed under the two error density assumptions and the three sample sizes.

| Error pdf | $n$ | Variance method | $\pi_1 = 0.5$ | | $\beta_{10} = 10$ | | $\beta_{11} = -10$ | | $\beta_{20} = -10$ | | $\beta_{21} = 10$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | Var | Mean | Var | Mean | Var | Mean | Var | Mean | Var |
| Unequal | 100 | Case | 3.88E−03 | 2.66E−07 | 0.897 | 0.290 | 3.19 | 2.91 | 0.925 | 0.421 | 3.41 | 4.63 |
| | | Model | 3.92E−03 | 1.82E−07 | 0.802 | 6.08E−02 | 2.88 | 0.757 | 0.805 | 9.56E−02 | 2.93 | 1.18 |
| | | Algorithm 3.1 | 3.76E−03 | 1.82E−07 | 0.802 | 9.43E−02 | 3.02 | 1.86 | 0.833 | 0.163 | 3.34 | 3.47 |
| | 300 | Case | 1.23E−03 | 1.12E−08 | 0.284 | 2.07E−02 | 1.06 | 0.251 | 0.253 | 1.74E−02 | 0.978 | 0.309 |
| | | Model | 1.26E−03 | 9.20E−09 | 0.250 | 3.84E−03 | 0.931 | 5.66E−02 | 0.237 | 3.54E−03 | 0.891 | 5.35E−02 |
| | | Algorithm 3.1 | 1.22E−03 | 5.91E−09 | 0.252 | 4.51E−03 | 0.915 | 7.75E−02 | 0.240 | 4.52E−03 | 0.920 | 0.113 |
| | 600 | Case | 6.05E−04 | 2.27E−09 | 0.132 | 3.57E−03 | 0.513 | 5.07E−02 | 0.125 | 3.42E−03 | 0.475 | 4.23E−02 |
| | | Model | 6.17E−04 | 1.81E−09 | 0.120 | 5.45E−04 | 0.454 | 8.19E−03 | 0.114 | 5.20E−04 | 0.434 | 8.09E−03 |
| | | Algorithm 3.1 | 6.04E−04 | 8.05E−10 | 0.119 | 6.42E−04 | 0.436 | 1.35E−02 | 0.117 | 6.26E−04 | 0.442 | 1.30E−02 |
| Equal | 100 | Case | 3.82E−03 | 2.91E−07 | 0.895 | 0.286 | 3.30 | 3.23 | 0.912 | 0.397 | 3.34 | 4.41 |
| | | Model | 3.73E−03 | 1.60E−07 | 0.733 | 4.99E−02 | 2.69 | 0.68 | 0.729 | 0.047 | 2.68 | 0.604 |
| | | Algorithm 3.1 | 3.61E−03 | 1.23E−07 | 0.748 | 8.58E−02 | 2.87 | 1.63 | 0.764 | 0.101 | 3.06 | 2.39 |
| | 300 | Case | 1.22E−03 | 1.22E−08 | 0.288 | 2.10E−02 | 1.12 | 0.302 | 0.253 | 1.70E−02 | 0.968 | 0.321 |
| | | Model | 1.20E−03 | 9.45E−09 | 0.234 | 2.61E−03 | 0.889 | 4.14E−02 | 0.228 | 2.18E−03 | 0.860 | 3.65E−02 |
| | | Algorithm 3.1 | 1.18E−03 | 4.40E−09 | 0.237 | 3.43E−03 | 0.869 | 6.84E−02 | 0.232 | 3.15E−03 | 0.890 | 9.87E−02 |
| | 600 | Case | 6.00E−04 | 2.72E−09 | 0.133 | 3.65E−03 | 0.542 | 6.44E−02 | 0.125 | 3.33E−03 | 0.469 | 4.30E−02 |
| | | Model | 5.96E−04 | 1.67E−09 | 0.116 | 3.85E−04 | 0.447 | 6.85E−03 | 0.111 | 3.17E−04 | 0.418 | 5.25E−03 |
| | | Algorithm 3.1 | 5.87E−04 | 6.42E−10 | 0.113 | 5.03E−04 | 0.416 | 1.14E−02 | 0.114 | 4.24E−04 | 0.430 | 9.68E−03 |

**Table 3**
A summary of parameter estimates using Algorithm 2.1 on 500 simulations from model (4.3) and (4.4). All simulations use a sample size $n = 300$ and assume unequal error densities. The variances can be treated as gold standards.

| Estimate | $\pi_1 = 1/3$ | $\pi_2 = 1/3$ | $\pi_3 = 1/3$ | $\beta_{10} = 0$ | $\beta_{11} = -20$ | $\beta_{12} = -20$ |
|---|---|---|---|---|---|---|
| Mean | 0.318 | 0.345 | 0.337 | −0.153 | −19.9 | −19.9 |
| Var | 6.21E−04 | 6.75E−04 | 7.11E−04 | 0.855 | 1.31 | 1.49 |
| Estimate | $\beta_{20} = 0$ | $\beta_{21} = 0$ | $\beta_{22} = 0$ | $\beta_{30} = 0$ | $\beta_{31} = 20$ | $\beta_{32} = 20$ |
| Mean | −2.93E−02 | −2.97E−02 | 5.29E−02 | −5.62E−03 | 20.0 | 20.0 |
| Var | 0.250 | 0.372 | 0.375 | 5.35E−02 | 9.44E−02 | 8.74E−02 |

**Table 4**
A summary of variance estimates using Case bootstrapping, Model bootstrapping, and Algorithm 3.1 on 500 simulations from model (4.3) and (4.4). All simulations use a sample size $n = 300$ and assume unequal error densities. Each variance estimate is based on 500 bootstrap samples or MCMC iterations.

| Variance method | Case | | Model | | Algorithm 3.1 | |
|---|---|---|---|---|---|---|
| | Mean | Var | Mean | Var | Mean | Var |
| $\pi_1 = 1/3$ | 6.63E−04 | 5.87E−09 | 6.00E−04 | 5.20E−09 | 8.00E−04 | 7.22E−09 |
| $\pi_2 = 1/3$ | 6.61E−04 | 4.40E−09 | 6.23E−04 | 5.74E−09 | 8.13E−04 | 7.16E−09 |
| $\pi_3 = 1/3$ | 7.18E−04 | 3.36E−09 | 7.17E−04 | 3.63E−09 | 7.52E−04 | 9.30E−10 |
| $\beta_{10} = 0$ | 1.18 | 0.504 | 1.39 | 1.40 | 1.43 | 1.03 |
| $\beta_{11} = -20$ | 1.82 | 0.830 | 2.17 | 3.53 | 2.33 | 2.89 |
| $\beta_{12} = -20$ | 1.81 | 0.775 | 2.14 | 3.18 | 2.29 | 2.61 |
| $\beta_{20} = 0$ | 0.309 | 2.80E−02 | 0.375 | 3.00E−02 | 0.298 | 1.52E−02 |
| $\beta_{21} = 0$ | 0.491 | 5.21E−02 | 0.578 | 6.35E−02 | 0.485 | 3.77E−02 |
| $\beta_{22} = 0$ | 0.490 | 4.78E−02 | 0.577 | 6.35E−02 | 0.483 | 3.75E−02 |
| $\beta_{30} = 0$ | 7.07E−02 | 1.30E−03 | 7.55E−02 | 2.99E−03 | 6.30E−02 | 4.09E−04 |
| $\beta_{31} = 20$ | 0.113 | 2.30E−03 | 0.122 | 6.49E−03 | 0.105 | 9.66E−04 |
| $\beta_{32} = 20$ | 0.119 | 2.63E−03 | 0.123 | 7.72E−03 | 0.105 | 1.03E−03 |

The three error distributions in (4.4) are different but all have medians equal to zero. The parameters are estimated using Algorithm 2.1. Bootstrap and Algorithm 3.1 variances are obtained based on 500 bootstrap samples or MCMC iterations. The parameter estimates from the 500 simulated data sets are summarized in Table 3 and the variance estimates are summarized in Table 4. Means and variances are given for both the parameter estimates using Algorithm 2.1 and the variance estimates using all three methods. The variances in Table 3 can be treated as gold standards against which the variance estimates in Table 4 are compared. As can be seen, the parameters are successfully estimated by Algorithm 2.1 and all three variance estimates are reasonably close to the gold standards.

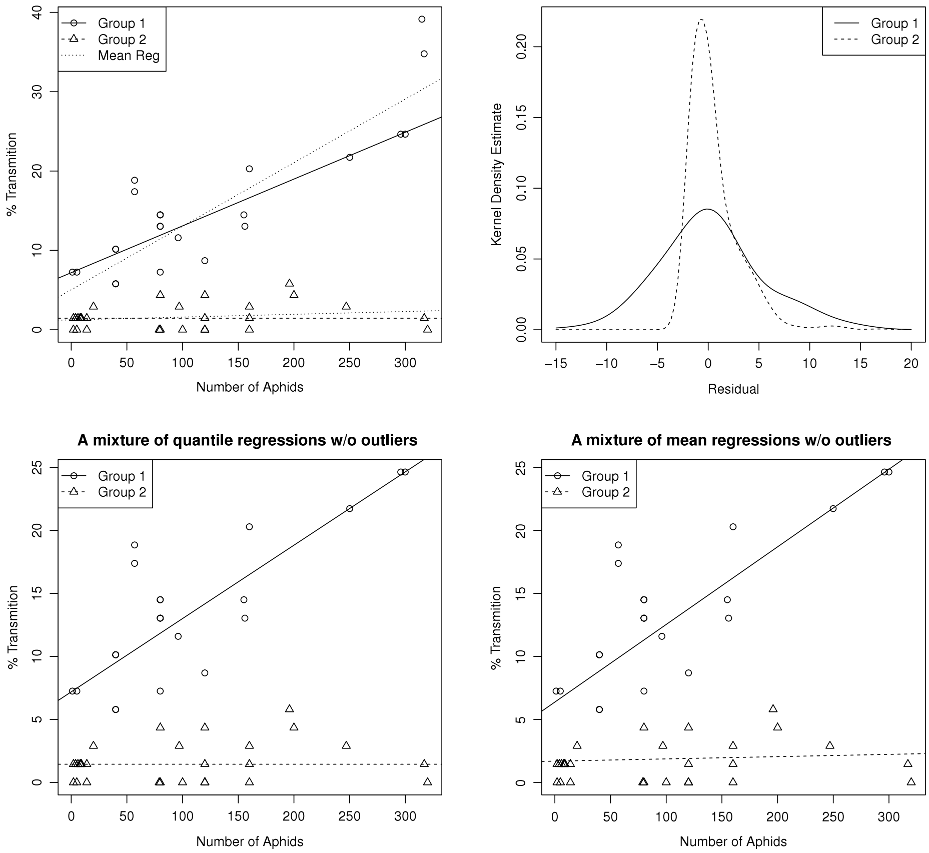

**Fig. 3.** A mixture of median regressions of percent virus transmission on the number of aphids. A mixture of mean regressions is included for a comparison. The kernel density estimates of the two error densities are given. The mixture of median regressions model and the mixture of mean regressions model are refitted when the two outlying observations with high virus transmission rates are excluded.

## 5. Illustrative examples

In this section, the newly proposed mixture of quantile regressions method is applied to three illustrative examples: the aphids data example from Boiteau et al. (1998), the tone data from Cohen (1984), and the engine data from Brinkman (1981). Advantages of the mixture of quantile regressions model over the mixture of mean regressions model is demonstrated. Algorithm 2.1 starting from uniformly generated initial values is used to find the parameter estimates and Algorithm 3.1 is used to estimate the variances.

First, Boiteau et al. (1998) study the effect of aphids on virus spreading from infected to healthy tobacco plants. A controlled experiment was conducted to measure the number of newly infected plants under differing numbers of aphids introduced into the experimental environment. The data from 51 trials of the study are plotted in Fig. 3 which shows a dichotomous pattern of two regression lines. For this reason, Boiteau et al. (1998) fit a mixture of regressions model with two mean regression components of percent virus transmission on the number of aphids. The estimated regression lines are (1) $\hat{y} = 5.0342 + 0.0801x$ and (2) $\hat{y} = 1.2447 + 0.0035x$ with an estimated group proportion of $\hat{\pi}_1 = 0.5017$. When applying the mixture of quantile regressions model with $\tau = 0.5$, the estimated intercepts and slopes of the two median regression lines are (1) $\hat{\beta}_{10} = 7.1874$ (6.3654) and $\hat{\beta}_{11} = 0.0590$ (2.1025E−4) and (2) $\hat{\beta}_{20} = 1.4493$ (0.5527) and $\hat{\beta}_{21} = 0.0000$ (2.8636E−5) with an estimated group proportion $\hat{\pi}_1 = 0.4315$ (8.4783E−3). The values in the parentheses are the variance estimates from Algorithm 3.1. The results are depicted in Fig. 3 where the cases are classified using a cutoff of $p = 0.5$ to the classification probabilities and shown in different symbols. The mean regression lines are included in the plot for a comparison. As can be seen, there is a big difference between the mean and the median regression lines in group 1 which is likely due to the two outlying observations at the top right corner. The kernel density estimates of the error densities show moderate skewness to the right in both groups. There is also evidence that the two error densities are unequal which shows the superiority of the mixture of median regressions model to the mixture of mean regressions model (1.2).

To further demonstrate the robustness of the mixture of quantile regressions model to possible outliers, the two outlying observations with their percents virus transmission beyond 30% are removed and both the median and the mean regression models are refitted to the rest of the data. Fig. 3 shows that the refitted median regressions do not change much but the

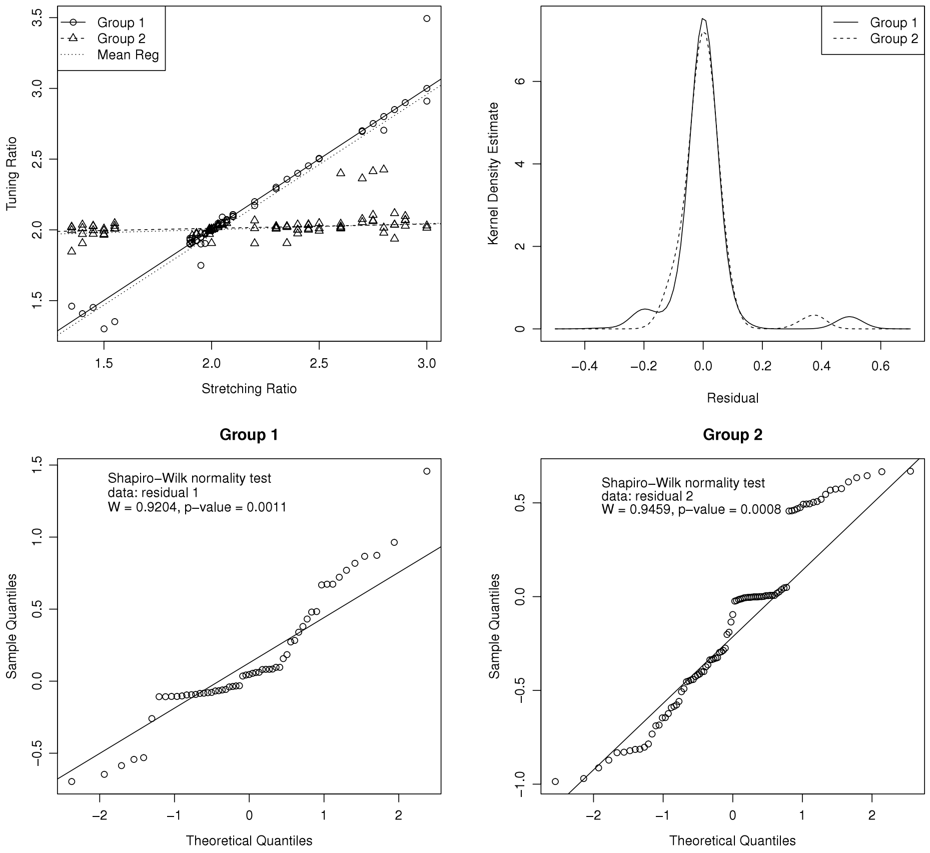

**Fig. 4.** A mixture of two median regressions of the tuning ratio on the stretching ratio. A mixture of mean regressions is included for a comparison. The kernel density estimates of the two error densities are given. QQ-plots of the residuals within each group are provided and the corresponding Shapiro–Wilk's normality tests are conducted.

refitted mean regressions depart largely from the original ones. Therefore, the two outlying observations have a big impact on the mean regression estimates.

Second, Cohen (1984) describes a tone perception experiment, in which a pure fundamental tone with electronically generated overtones added was played to a trained musician. The overtones were determined by a stretching ratio. The response variable is the tuning ratio, which is the ratio between the adjusted tone and the fundamental tone, and the predictor is the stretching ratio. There are 150 trials recorded from the same musician. The purpose of this experiment was to see how this tuning ratio affects the perception of the tone and determine if either of two musical perception theories was reasonable (see Cohen, 1984, for more details). The scatterplots in Fig. 4 show that two homogeneous groups are evident which correspond to correct tuning and tuning to the first overtone, respectively. The estimated mean regression lines are (1) $\hat{y} = -0.0193 + 0.9923x$ and (2) $\hat{y} = 1.9164 + 0.0426x$ with an estimated group proportion $\hat{\pi}_1 = 0.3022$. Applying the mixture of median regressions model, we estimate the two median regression lines to have coefficients equal to (1) $\hat{\beta}_{10} = 3.22\text{E} - 3$ (4.76E−3) and $\hat{\beta}_{11} = 0.999$ (9.41E−4) and (2) $\hat{\beta}_{20} = 1.95$ (6.28E−4) and $\hat{\beta}_{21} = 3.04\text{E} - 2$ (1.29E−4) with $\hat{\pi}_1 = 0.373$ (2.89E−3). The results are depicted in Fig. 4 where the cases are classified using a cutoff of $p = 0.5$ to the classification probabilities and shown in different symbols. The plot also includes the mean regression lines which can be found close to the median regression lines. This is not surprising because we have a pretty big sample size after all. The estimated error densities show some bi- or triple-modal patterns. This is because of some suspicious outliers outside the patterns of the two regression models.

The mixture of median regressions model is superior to the normal mixture model (1.1) by dropping the normality assumption. Therefore, in practice, if there is no prior information about the error density, the proposed mixture of quantile regressions model can potentially be used as an exploratory tool to check any parametric assumption of the error density. However, normality tests in the mixture model setting are not easy because of the uncertainty in the group memberships. It is unclear whether we can directly use the residuals from the mixture models for normality testing or we should truncate the residuals based on the classification probabilities. If we do truncate, shall we weight the truncated residuals by the classification probabilities? Another approach is probably to resample the residuals by the classification probabilities and treat the resampled residuals as from a single population. More research is needed to address all these questions. For the tone data, we show the normal Q–Q plots of the residuals within each group while the cases are classified using the hard boundary

**Table 5**
Three mixtures of quantile regressions of the equivalence ratio on the concentration of nitrous oxide using $\tau = 0.25, 0.5$, and 0.75. Parameter estimates are found using Algorithm 2.1 and variances are estimated using Algorithm 3.1.

| Quantile (%) | Parameter | $\beta_{10}$ | $\beta_{11}$ | $\beta_{20}$ | $\beta_{21}$ |
|---|---|---|---|---|---|
| 25 | Estimate | 1.223 | −0.07982 | 0.5358 | 0.07892 |
| | Variance | 7.45E−05 | 1.73E−05 | 4.56E−04 | 7.67E−05 |
| 50 | Estimate | 1.240 | −8.17E−02 | 5.57E−01 | 9.09E−02 |
| | Variance | 1.45E−04 | 3.75E−05 | 5.37E−04 | 9.44E−05 |
| 75 | Estimate | 1.263 | −0.08301 | 0.6146 | 0.08153 |
| | Variance | 1.55E−04 | 3.63E−05 | 3.12E−04 | 5.25E−05 |

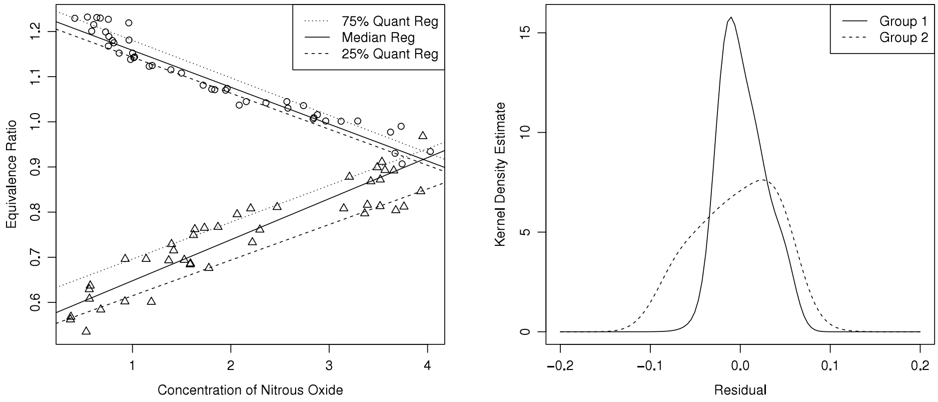

**Fig. 5.** There mixtures of quantile regressions of the equivalence ratio on the concentration of nitrous oxide for $\tau = 0.25, 0.5$, and 0.75. The kernel density estimates of the two error densities are given under the mixture of median regressions model.

of $p = 0.5$. Results in Fig. 4 show that the residuals are by no means from some normal distributions. The Shapiro–Wilk's tests of normality have $p$-values less than 0.001 for both groups. The Q–Q plot and the Shapiro–Wilk's test are only effective when the components are well-separated and questionable when the regression lines overlap.

Finally, Brinkman (1981) introduces an engine data set containing two related variables. The response variable is the equivalence ratio, which measures the richness of the air–ethanol mix for burning ethanol in a single-cylinder automobile test and the predictor is the concentration of nitrous oxide in engine exhaust, normalized by engine work. The observations are collected on 87 different engines. From Fig. 5, one can see that there are two homogeneous groups. Therefore, a single linear regression will not fit the data very well and a two component mixture of regressions should be used instead. The estimated mean regression lines are (1) $\hat{y} = 1.2470 - 0.0829x$ and (2) $\hat{y} = 0.5674 + 0.0846x$ with an estimated group proportion $\hat{\pi}_1 = 0.5164$. However, it could be very interesting to fit several mixtures of quantile regressions to the same data and show the conditional quantile regions of groups. Fig. 5 illustrates three mixtures of quantile regressions fitted to the engine data using $\tau = 0.25, 0.5$, and 0.75, respectively. For both groups, the 25% and 75% quantile regression lines define the conditional inter-quartile regions. Table 5 summarizes the fitted models. However, there are a couple of difficulties when fitting multiple mixtures of quantile regressions to the same data. First, the quantile regression lines within each group might cross each other so that the conditional quantiles at some $x$ values are not in a proper order. To ensure that all conditional quantiles are in a proper order, further restrictions on the model fitting are necessary but they are out of the scope of this paper. Second, although we do not expect the estimated classification probabilities to be much different while using different quantile levels, it could happen in real data applications, especially when the sample size is small. This can potentially create an ambiguity in terms of the classification. For this application on the engine data, all cases are classified using the mixture of median regressions with a cutoff of $p = 0.5$ to the classification probabilities and shown in different symbols in Fig. 5. Also from the mixture of median regressions, the error density estimate for group 1 is skewed to the right and that for group 2 is skewed to the left.

## 6. Considerations

For the mixture of regressions model in general, if the error pdfs $\mathbf{G}$ are completely known or known up to a finite number of parameters, then a regular EM algorithm can be used to find the MLE $\hat{\boldsymbol{\theta}}$ of the parameters $\boldsymbol{\theta}$. For example, $\mathbf{G}$ can be the normal densities $\phi(\cdot/\sigma_j)/\sigma_j$ with unknown scale parameters $\sigma_j$ for $j = 1, \ldots, m$. When $\mathbf{G}$ are known, an equivalent estimating equation approach solves

$$\sum_{i=1}^{n} \mathbf{x}_i p_{ij}(\boldsymbol{\theta}) \xi(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j) = 0, \tag{6.1}$$

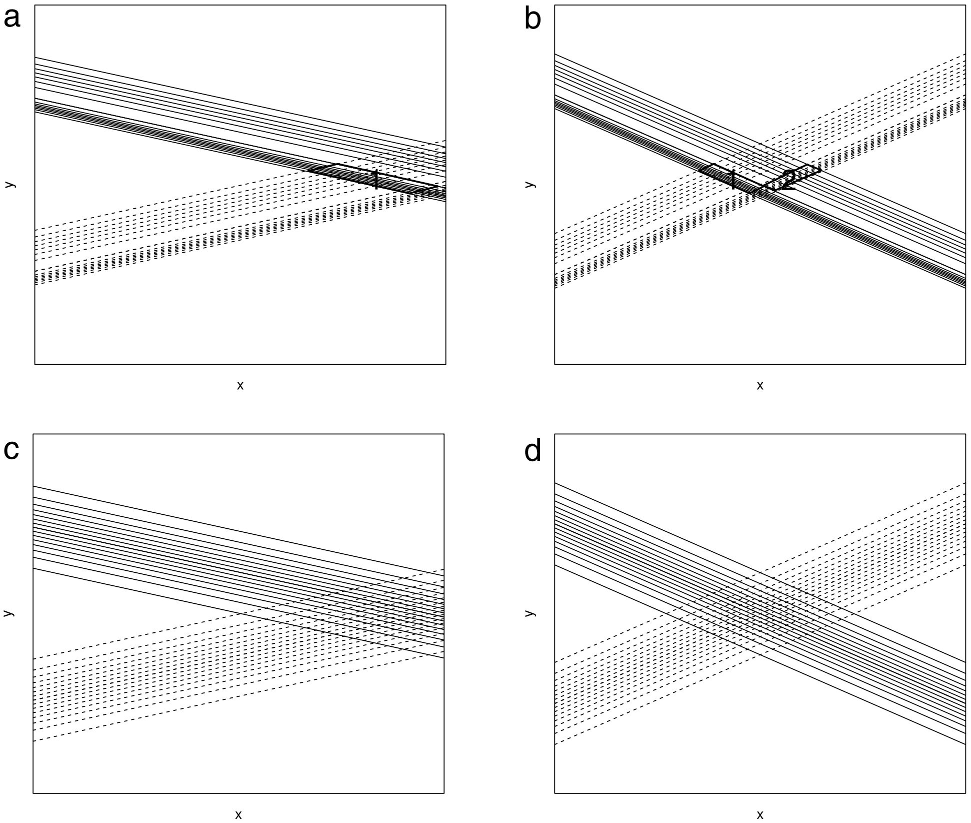

**Fig. 6.** Prototypes of mixtures of two regression models. The solid lines and the dashed lines represent the two groups. The line density represents the density of regression errors. Figures (a) and (b) contain two clusters with skewed error distributions while figures (c) and (d) have symmetric error distributions.

and

$$\pi_j = \frac{1}{n}\sum_{i=1}^{n} p_{ij}(\boldsymbol{\theta}), \tag{6.2}$$

for $j = 1, \dots, m$ where

$$p_{ij}(\boldsymbol{\theta}) = \frac{\pi_j g_j(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j)}{\sum\limits_{l=1}^{m} \pi_l g_l(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}_l)}.$$

For mixtures of mean regressions $\xi(a) = a$ is an identity function, while for mixtures of $\tau$th quantile regressions $\xi(a) = I(a \le 0) - \tau$. When **G** are known up to a finite number of parameters, similar estimating equations can be constructed with additional equations to solve for the parameters in **G**. From either the MLE theory or the estimating equation theory, the estimators $\hat{\boldsymbol{\theta}}$ enjoy nice asymptotic properties such as being consistent and asymptotically normal.

However, such asymptotic results are not available for semi-parametric models (1.2) or (2.1). The EM-type Algorithm 2.1 is designed to solve (6.1) and (6.2) together with a kernel density estimation of the unknown error pdfs **G**. But when **G** are modeled non-parametrically, (6.1) and (6.2) are not estimating equations any more, at least not in the sense of Tsiatis (2010). This is because $p_j(\boldsymbol{\theta})\xi(Y - \mathbf{x}^T\boldsymbol{\beta}_j)$ for $j = 1, \dots, m$ are generally not orthogonal to the nuisance tangent space of the model (see Tsiatis, 2010, pp. 73–87 for more details). For this reason, potential bias in $\hat{\boldsymbol{\theta}}$ can arise which we find true in our simulation studies.

In general, the bias can occur when the clusters have imbalanced intersections. This usually happens when the error pdfs are asymmetric. In Fig. 6(a), the section labeled by "1" is an imbalanced intersection of high and low densities of the two groups. The subjects in this section are more likely to be clustered into the top group which leads to a bias in the semi-parametric estimation. In Fig. 6(b), the two imbalanced intersections "1" and "2" are counter-balanced by each other, so the bias can be reduced or eliminated. In Fig. 6(c) and (d), all intersections are balanced because the error densities are symmetric, so the estimates should be consistent. At this stage, Algorithms 2.1 and 3.1 represent some practical methods that work reasonably well in clustering the subjects and model estimation. Further research is required for possible improvements on the asymptotics.

## Acknowledgments

The authors would like to thank two anonymous referees for their insightful comments that led to an improved manuscript and Dr. Paul Vos in Department of Biostatistics at East Carolina University for reviewing the revision.

## Appendix

**Lemma A.1.** *For any nonzero real number* $\lambda$*,*

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}e^{is\lambda}ds=0.$$

**Proof of Theorem 2.1.** The proof is adapted from Wang et al. (2012). Note that the conditional density $f(y|\mathbf{x})$ is

$$f(y|\mathbf{x})=\sum_{j=1}^{m}\pi_j\tilde{g}_j(y-\tilde{\mathbf{x}}^T\tilde{\boldsymbol{\beta}}_j),$$

where $\tilde{g}_j(t)=g_j(t-\beta_{0j})$ and thus $\tilde{g}_j$ has its $\tau$th quantile equal to $\beta_{0j}$. The characteristic function for the conditional distribution of $y$ given $\mathbf{x}$ is

$$\phi_{y|\mathbf{x}}(t)=\int_{-\infty}^{\infty}e^{iyt}\sum_{j=1}^{m}\pi_j\tilde{g}_j(y-\tilde{\mathbf{x}}^T\tilde{\boldsymbol{\beta}}_j)dy=\sum_{j=1}^{m}\pi_je^{i\tilde{\mathbf{x}}^T\tilde{\beta}_jt}\phi_{\tilde{g}_j}(t),$$

where $i$ is the imaginary unit, $\phi_{\tilde{g}_j}(t)$ is the characteristic function of $\tilde{g}_j(y)$.

Suppose that the proposed model has another representation

$$f(y|\mathbf{x})=\sum_{k=1}^{l}\tau_kh_k(y-\tilde{\mathbf{x}}^T\boldsymbol{\gamma}_k),$$

where $\gamma_k$ are distinct, $0\le\tau_k\le 1$, and $\sum_{k=1}^{l}\tau_k=1$. Then

$$\phi_{y|\mathbf{x}}(t)=\sum_{k=1}^{l}\tau_ke^{i\tilde{\mathbf{x}}^T\gamma_kt}\phi_{h_k}(t),$$

and therefore,

$$\sum_{j=1}^{m}\pi_je^{i\tilde{\mathbf{x}}^T\tilde{\beta}_jt}\phi_{\tilde{g}_j}(t)=\sum_{k=1}^{l}\tau_ke^{i\tilde{\mathbf{x}}^T\gamma_kt}\phi_{h_k}(t). \tag{A.1}$$

For any fixed $1\le q\le m$, multiplying both sides of Eq. (A.1) by $e^{-i\tilde{\mathbf{x}}^T\tilde{\beta}_qt}$ gives

$$\pi_q\phi_{\tilde{g}_q}(t)+\sum_{1\le j\ne q\le m,}\pi_je^{i\tilde{\mathbf{x}}^T(\tilde{\beta}_j-\tilde{\beta}_q)t}\phi_{\tilde{g}_j}(t)=\sum_{k=1}^{l}\tau_ke^{i\tilde{\mathbf{x}}^T(\gamma_k-\tilde{\beta}_q)t}\phi_{h_k}(t).$$

Since $\tilde{\boldsymbol{\beta}}_j$ are distinct vectors in $\mathbb{R}^p$, $\tilde{\boldsymbol{\beta}}_j-\tilde{\boldsymbol{\beta}}_q$ are nonzero vectors in $\mathbb{R}^p$ for $j\ne q$, where $\mathbb{R}$ is the set of all real numbers.

Next, we will prove that one of $\boldsymbol{\gamma}_k$'s is equal to $\tilde{\boldsymbol{\beta}}_q$. Let us first assume that $\boldsymbol{\gamma}_k-\tilde{\boldsymbol{\beta}}_q$ are all nonzero vectors in $\mathbb{R}^p$. Let

$$F_1=\cup_{1\le j\ne q\le m}\{\tilde{\mathbf{x}}\in\mathcal{R}^p:\tilde{\mathbf{x}}^T(\tilde{\boldsymbol{\beta}}_j-\tilde{\boldsymbol{\beta}}_q)=0\}$$

and

$$F_2=\cup_{k=1}^{l}\{\tilde{\mathbf{x}}\in\mathcal{R}^p:\tilde{\mathbf{x}}^T(\boldsymbol{\gamma}_k-\tilde{\boldsymbol{\beta}}_q)=0\}.$$

Note that both $F_1$ and $F_2$ consist of the union of several $(p-1)$-dimensional hyper-planes of $\mathbb{R}^p$ and thus both have zero measure. Let $D$ be an open subset inside the domain of $\tilde{\mathbf{x}}$. Then, there is a vector $\boldsymbol{\nu}\in D$, such that $\boldsymbol{\nu}^T(\tilde{\boldsymbol{\beta}}_j-\tilde{\boldsymbol{\beta}}_q)\ne 0$, $1\le j\le m$, $j\ne q$, and that $\boldsymbol{\nu}^T(\boldsymbol{\gamma}_k-\tilde{\boldsymbol{\beta}}_q)\ne 0$, $k=1,\dots,l$. Let $\varepsilon>0$ be such that $a\boldsymbol{\nu}\in D$ for all $a\in(1-\varepsilon,1+\varepsilon)$. Then

$$\pi_q\phi_{\tilde{g}_q}(t)+\sum_{1\le j\ne q\le m,}\pi_je^{ia\boldsymbol{\nu}^T(\tilde{\beta}_j-\tilde{\beta}_q)t}\phi_{\tilde{g}_j}(t)=\sum_{k=1}^{l}\tau_ke^{ia\boldsymbol{\nu}^T(\gamma_k-\tilde{\beta}_q)t}\phi_{h_k}(t).$$

For any fixed $t \neq 0$, denote $\eta_j = \boldsymbol{\nu}^T(\tilde{\boldsymbol{\beta}}_j - \tilde{\boldsymbol{\beta}}_q)t,\ j = 1, \dots, m, \xi_k = \boldsymbol{\nu}(\boldsymbol{\gamma}_k - \tilde{\boldsymbol{\beta}}_q)t,\ k = 1, \dots, l$. Then $\eta_j \neq 0$ for $j \neq q$ and $\xi_k \neq 0$. Therefore,

$$\pi_q \phi_{\tilde{g}_q}(t) + \sum_{1\le j\ne q\le m,} \pi_j e^{ia\eta_j} \phi_{\tilde{g}_j}(t) = \sum_{k=1}^{l} \tau_k e^{ia\xi_k} \phi_{h_k}(t), \quad 1-\varepsilon < a < 1+\varepsilon.$$

By extension theorem (Whitney, 1934), the above equality holds for all complex number $a \in \mathbb{R}$.

Then

$$\frac{1}{2T}\int_{-T}^{T}\left\{\pi_q \phi_{\tilde{g}_q}(t) + \sum_{1\le j\ne q\le m,} \pi_j e^{ia\eta_j} \phi_{\tilde{g}_j}(t)\right\} da = \frac{1}{2T}\int_{-T}^{T}\left\{\sum_{k=1}^{l} \tau_k e^{ia\xi_k} \phi_{h_k}(t)\right\} da.$$

Applying Lemma A.1, we obtain that $\pi_q \phi_{\tilde{g}_q}(t) = 0$. Letting $t \to 0$, we have $\pi_q = 0$, which contradicts the assumption that $\pi_q > 0$. Therefore, there must be a $1 \le \tilde{q} \le l$ such that $\boldsymbol{\gamma}_{\tilde{q}} = \tilde{\boldsymbol{\beta}}_q$. Based on the similar arguments above, we can have $\pi_q \phi_{\tilde{g}_q}(t) = \tau_{\tilde{q}} \phi_{h_{\tilde{q}}}(t)$. Letting $t \to 0$, we have $\tau_{\tilde{q}} = \pi_q$, and therefore, $\phi_{h_{\tilde{q}}}(t) = \phi_{g_q}(t)$, which in turns implies that $h_{\tilde{q}}(y) = g_q(y)$.

Note that the above results should hold for any $1 \le q \le m$. In addition, since the $\tilde{g}_j$'s are identifiable, the $\beta_{0j}$'s, which are $\tau$th quantile of $\tilde{g}_j$'s, are also identifiable.

## References

Arminger, G., Stein, P., Wittenberg, J., 1999. Mixtures of conditional mean- and covariance-structure models. Psychometrika 64, 475–494.

Bai, Z.D., Rao, C.R., Wu, Y., 1992. *M*-estimation of multivariate linear regression parameters under a convex discrepancy function. Statist. Sinica 2, 237–254.

Bai, X., Yao, W., Boyer, J.E., 2012. Robust fitting of mixture regression models. Comput. Statist. Data Anal. 56, 2347–2359.

Boiteau, G., Singh, M., Singh, R.P., Tai, G.C.C., Turner, T.R., 1998. Rate of spread of PVY-n by alate Myzus persicae (sulzer) from infected to healthy plants under laboratory conditions. Potato Res. 41, 335–344.

Brinkman, N., 1981. Ethanol fuel-a single-cylinder engine study of efficiency and exhaust emissions. SAE Trans. 90, 1410–1427.

Cohen, E., 1984. Some effects of inharmonic partials on interval perception. Music Percept. 1, 323–349.

DeSarbo, W.S., Corn, L.W., 1988. A maximum likelihood methodology for clusterwise linear regression. J. Classification 5, 249–282.

Garcia-Escudero, L., Gordaliza, A., Mayo-Iscar, A., Martin, R.S., 2010. Robust clusterwise linear regression through trimming. Comput. Statist. Data Anal. 54, 3057–3069.

Hall, P., Presnell, B., 1999. Density estimation under constraints. J. Comput. Graph. Statist. 8, 259–277.

He, X., Shao, Q.M., 1996. A general Bahadur representation of *M*-estimators and its application to linear regression with nonstochastic designs. Ann. Statist. 24, 2608–2630.

Hunter, D.R., Young, D.S., 2012. Semiparametric mixtures of regressions. J. Nonparametr. Stat. 24, 19–38.

Ingrassia, S., Minotti, S.C., Punzo, A., 2014. Model-based clustering via linear cluster-weighted models. Comput. Statist. Data Anal. 71, 159–182.

Ingrassia, S., Minotti, S.C., Vittadini, G., 2012. Local statistical modeling via a cluster-weighted approach with elliptical distributions. J. Classification 29, 363–401.

Jones, P.N., McLachlan, G.J., 1992. Fitting finite mixture models in a regression context. Austral. J. Statist. 34, 233–240.

Kocherginsky, M., He, X., Mu, Y., 2005. Practical confidence intervals for regression quantiles. J. Comput. Graph. Statist. 14, 41–55.

Koenker, R., 1994. Confidence intervals for regression quantiles. In: Mandl, P., Hušková, M. (Eds.), Asymptotic Statistics. In: Contributions to Statistics, Physica-Verlag HD, pp. 349–359. http://dx.doi.org/10.1007/978-3-642-57984-4_29.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33–50.

Koenker, R., Machado, J.A., 1999. Goodness of fit and related inference processes for quantile regression. J. Amer. Statist. Assoc. 94, 1296–1310.

Koenker, R., Zhao, Q., 1994. *L*-estimation for linear heteroscedastic models. J. Nonparametr. Stat. 3, 223–235.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, second ed. John Wiley and Sons, Inc., Hoboken, New Jersey.

Parzen, M.I., Wei, L.J., Ying, Z., 1994. A resampling method based on pivotal estimating functions. Biometrika 81, 341–350.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman and Hall/CRC, Boca Raton, FL.

Song, W., Yao, W., Xing, Y., 2014. Robust mixture regression model fitting by Laplace distribution. Comput. Statist. Data Anal. 71, 128–137.

Tsiatis, A.A., 2010. Semiparametric Theory and Missing Data. Springer, New York.

Wang, S., Yao, W., Hunter, D., 2012. Mixture of linear regression models with unknown error density. http://www-personal.ksu.edu/~wxyao/material/submitted/mixlinnonerr.pdf. Unpublished manuscript.

Wedel, M., Kamakura, W.A., 2000. Market Segmentation: Conceptual and Methodological Foundations. Kluwer, Dordrecht.

Wei, Y., 2012. Robust mixture regression models using *t*-distribution. Technical Report. Department of Statistics, Kansas State University.

Whitney, H., 1934. Analytic extension of differential functions defined in closed sets. Trans. Amer. Math. Soc. 36, 63–89.

Wu, C.F.J., 1986. Jackknife, bootstrap and other resampling methods in regression analysis. Ann. Statist. 14, 1261–1295.

Wu, Q., Sampson, A.R., 2009. Mixture modeling with applications in schizophrenia research. Comput. Statist. Data Anal. 53, 2563–2572.

Yao, W., Wei, Y., Yu, C., 2014. Robust mixture regression using *t*-distribution. Comput. Statist. Data Anal. 71, 116–127.