Two-layer EM algorithm for ALD mixture regression models: A new solution to composite quantile regression

Wang, Shangshan; Xiang, Liming

2017

Wang, S., & Xiang, L. (2017). Two-layer EM algorithm for ALD mixture regression models: A new solution to composite quantile regression. *Computational Statistics & Data Analysis*, 115, 136-154.

<https://hdl.handle.net/10356/83377>

<https://doi.org/10.1016/j.csda.2017.06.002>

© 2017 Elsevier. This is the author created version of a work that has been peer reviewed and accepted for publication by *Computational Statistics & Data Analysis*, Elsevier. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [<http://dx.doi.org/10.1016/j.csda.2017.06.002>].

Downloaded on 26 Apr 2023 19:40:16 SGT

Two-Layer EM Algorithm for ALD Mixture Regression Models: A New Solution to Composite Quantile Regression

Shangshan Wang^{1,2} Liming Xiang^{2*}

¹*School of Economics and Management, Beihang University, Beijing 100191, China*

²*School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
637371, Singapore*

Abstract

We advocate linear regression by modeling the error term through a finite mixture of asymmetric Laplace distributions (ALDs). The model expands the flexibility of linear regression to account for heterogeneity among data and allows us to establish the equivalence between maximum likelihood estimation of the model parameters and the composite quantile regression (CQR) estimation developed by Zou and Yuan (Ann Stat 36:1108-1126, 2008), providing a new likelihood-based solution to CQR. Particularly, we develop a computationally efficient estimation procedure via a two-layer EM algorithm, where the first layer EM algorithm incorporates missing information from the component memberships of the mixture model and nests the second layer EM in its M-step to accommodate latent variables involved in the location-scale mixture representation of the ALD. An appealing feature of the proposed algorithm is that the closed form updates for parameters in each iteration are obtained explicitly, instead of resorting to numerical optimization methods as in the existing work. Computational complexity can be reduced significantly. We evaluate the performance through simulation studies and illustrate its usefulness by analyzing a gene expression dataset.

Key Words: Asymmetric Laplace distribution, Composite quantile regression, EM algorithm, Mixture regression model, Robustness.

1 Introduction

Linear regression analysis is the most widely used statistical technique to explain the relationship between a dependent variable and one or more independent variables. We consider the estimation

* Corresponding author. Tel.: +65 65137451 ; Fax: +65 65158213.

E-mail address: lmxiang@ntu.edu.sg.

of unknown parameter β in a linear regression model

$$Y_i = X_i^T \beta + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ is an n -vector of response, $\mathbf{X} = (X_1, \dots, X_n)^T$ is an $n \times p$ matrix of independent variables with X_i^T being its i th row, $\beta = (\beta_1, \dots, \beta_p)^T$ is an p -vector of unknown parameters and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ is an n -vector of random errors often assumed to be independent and identically distributed with variance σ^2 , independent of $\mathbf{X}$. Popular approaches to estimate β can be generally written in the form of minimizing some prespecified loss function $\rho(\cdot)$, i.e.,

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \rho(Y_i - X_i^T \beta) \right\}. \quad (2)$$

The most common choices for ρ include the quadratic loss $\rho(t) = t^2/2$ corresponding to the least squares (LS) method, which is well-known for its mathematical beauty but inefficient for highly non-normal errors and not resistant to outliers, and the absolute loss $\rho(t) = |t|$ corresponding to least absolute derivation (LAD) regression, which is more robust to outliers than the LS method. There are other types of loss functions used in robust regression, such as Huber's loss, Tukey's bisquare and Hampel's ψ . Different loss functions require different assumptions about the error distribution to ensure desired performance. For example, the quadratic loss function is optimal under the assumption of Gaussian errors, the LAD loss is preferable for heavy-tail distributions and is the most efficient for the double-exponential distribution, and Huber's loss is suitable for contaminated normal distributions.

As another robust alternative to LS regression, quantile regression (QR) with $\rho(t) = t(\tau - I(t < 0))$ for $0 < \tau < 1$ has recently become attractive due to its advantages in avoiding distributional assumptions on the error term and incorporating heterogeneity in the data (Koenker and Bassett Jr, 1978). It includes the LAD regression as a special case where quantile is set to 0.5. Note that the QR estimation can be equivalently obtained using maximum-likelihood estimation (MLE) in linear regression with errors following the asymmetric Laplace distribution (ALD) (Koenker and Machado, 1999; Yu and Moyeed, 2001). Based on the link between both two different estimation methods, Geraci and Bottai (2007, 2014) further studied the likelihood-based approach to the QR for longitudinal data. However, it is known that the relative efficiency of the QR estimator can be arbitrarily small when compared with the LS estimator. To improve the efficiency of the usual quantile regression, Zou and Yuan (2008) proposed the composite quantile regression (CQR) method and derived a universal lower bound of the asymptotic relative efficiency of the CQR with respect to the LS.

In view of the unifying bridge existing between the QR estimation and the MLE under the ALD assumption, we are led to whether the CQR connects to ALD-based likelihood inference. Motivated by the definition of the CQR in which the objective function is a mixture of objective functions from different quantile regression models, we extend linear regression by assuming that the error term follows a finite mixture of the asymmetric Laplace distributions. An estimation procedure is then developed through maximizing a mixture of ALD likelihood functions. This ALD mixture assumption provides great flexibility for linear regression to accommodate the heterogeneity among data caused by heavy-tailed errors or outliers. It also allows us to establish the equivalence between the maximum likelihood estimation of model parameters and the conventional CQR estimation developed by Zou and Yuan (2008).

Modeling unknown error distribution using a mixture distribution has been studied in the literature when a single parametric family is unable to provide a satisfactory model for local variations in the observed data. Bartolucci and Scaccia (2005) and Soffritti and Galimberti (2011) developed regression methods with non-normal errors by assuming a finite mixture of Gaussian components for the error term. Galimberti and Soffritti (2014) extended their methods to a mixture model with each component belonging to a class of t distributions. The EM algorithm is typically employed to accomplish the MLE in such mixture models. Given the mixture structure of our model and the location-scale mixture representation of the ALD, we propose a novel EM-based algorithm, named by a two-layer EM algorithm, to implement our estimation procedure. In the first layer, an EM algorithm is performed to incorporate missing information from the component memberships of the mixture of ALDs. It nests the second layer EM algorithm in its M-step to accommodate latent variables involved in the location-scale mixture representation of the ALD.

One advantage of the proposed algorithm is that the updates for parameters in each iteration are obtained explicitly in closed form, simplifying the optimization tasks involved in the estimation procedure. In particular, the proposed algorithm can be applied directly to estimation of the CQR, providing a new and computationally efficient likelihood-based solution. Instead of resorting to linear programming optimization techniques, which typically have to call high-performance C++ programs as in the existing work, our algorithm is conceptually straightforward, numerically stable and easy to code using R directly. Another advantage of the algorithm is that the update of β obtained at the M-step can be viewed as a weighted composite quantile estimate with data-driven weights adaptively decided by the observed data and the current estimates. This makes more efficient use of data and the efficiency gains of the proposed estimator can be substantial, as

demonstrated later in our numerical results.

The rest of this article is organized as follows. Section 2 briefly reviews the QR, CQR and the representation of ALD as a normal-variance-mean mixture with an exponential mixing distribution. Section 3 presents the finite ALD mixture model and develops the estimation procedure through the two-layer EM algorithm. The application of the proposed method to the CQR is also included in this section. Section 4 provides simulation studies to evaluate the finite sample performance of the proposed method. Section 5 analyzes a gene expression dataset for illustration purpose. Section 6 gives a discussion. Finally, the paper ends with some calculation details in the Appendix.

2 Preliminaries

To motivate the proposed model and estimation method, we briefly review some background results on the asymmetric Laplace distribution, quantile regression and composite quantile regression. Suppose that continuous random variable $W \in \mathbb{R}$ follows an asymmetric Laplace distribution, denoted by $ALD(\mu, \sigma, \tau)$, with location parameter μ , scale parameter $\sigma > 0$ and asymmetric (skewness) parameter $0 < \tau < 1$. Its probability density function is given by

$$f(w; \mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp \left\{ -\rho_\tau \left(\frac{w-\mu}{\sigma} \right) \right\}, \quad (3)$$

where $\rho_\tau(t) = t(\tau - I(t < 0))$ is known as the check function with $I(\cdot)$ denoting the usual indicator function. This three-parameter distribution, initially developed by (Koenker and Machado, 1999), is useful for estimating quantile and quantile regression. See Yu and Zhang (2005) for a summary of its properties and extensions. From (3), It is easy to obtain the cumulative distribution function

$$F(w; \mu, \sigma, \tau) = \begin{cases} \tau \exp \left\{ \rho_\tau \left(\frac{w-\mu}{\sigma} \right) \right\} & w < \mu, \\ 1 - (1-\tau) \exp \left\{ \rho_\tau \left(\frac{w-\mu}{\sigma} \right) \right\} & w \geq \mu. \end{cases} \quad (4)$$

Obviously, the τ th quantile of $ALD(\mu, \sigma, \tau)$ is μ . $ALD(\mu, \sigma, \tau)$ reduces to the standard Laplace distribution or double-exponential distribution when $\tau = 1/2$. In early studies of the ALD random number generator (Kotz et al., 2012; Kozumi and Kobayashi, 2011), a mixture representation was developed for the ALD based on a normal-variance-mean mixture with an exponential mixing distribution. In particular, the ALD (μ, σ, τ) distributed random variable W can be represented hierarchically as follows:

$$W|Z \sim N(\mu + \kappa Z, \nu^2 \sigma Z) \quad \text{and} \quad Z \sim \text{Exp}(\sigma^{-1}), \quad (5)$$

where $\kappa = \frac{1-2\tau}{\tau(1-\tau)}$, $\nu^2 = \frac{2}{\tau(1-\tau)}$, and $\text{Exp}(\sigma^{-1})$ is an exponential distribution with mean σ .

In the typical linear regression model (1), the conditional 100 τ % quantile of $Y_i|X_i$ is $X_i^T\beta + b_\tau$, where b_τ is the 100 τ % quantile of ε_i . For brevity, we shall assume that the density function of ε , denoted by $g(\cdot)$, is nonvanishing everywhere. Thus b_τ is uniquely defined for any $0 < \tau < 1$. Quantile regression estimates β by solving

$$(\hat{b}_\tau, \hat{\beta}^{QR}) = \arg \min_{b, \beta} \sum_{i=1}^n \rho_\tau(Y_i - X_i^T\beta - b) \quad (6)$$

for any $\tau \in (0, 1)$ (Koenker, 2005). Quantile regression estimators are attractive not only because they are robust against non-Gaussian errors, but also because several quantiles convey a more complete picture of the conditional distribution of the dependent variable than the simple mean derived from a traditional least squares approach.

Quantile regression estimator $\hat{\beta}^{QR}$ is the most efficient under the assumption that error term $\varepsilon \sim \text{ALD}(b_\tau, \sigma/\psi, \tau)$ (Koenker and Machado, 1999; Yu and Moyeed, 2001), where the scale ψ makes the variance of ε equal to σ^2 . Unlike the LS estimator, it does not require $\sigma^2 < \infty$ to achieve the root- n consistency and asymptotic normality. However, similar to the LAD estimator, its relative efficiency with respect to the LS estimator can be arbitrarily small for finite variance cases. Zou and Yuan (2008) therefore developed the composite quantile regression (CQR) method to improve the conventional quantile regression. Specifically, denote that $0 < \tau_1 < \dots < \tau_K < 1$ and b_k^* is the 100 τ_k % quantile of ε_i for $1 \leq k \leq K$. The CQR estimator of β is defined by

$$(\hat{b}_1, \dots, \hat{b}_K; \hat{\beta}^{CQR}) = \arg \min_{b_1, \dots, b_K; \beta} \sum_{k=1}^K \left\{ \sum_{i=1}^n \rho_{\tau_k}(Y_i - X_i^T\beta - b_k) \right\}. \quad (7)$$

Typically, the equally spaced quantiles $\tau_k = \frac{k}{K+1}$ are suggested for $k = 1, \dots, K$. Under some regularity conditions, Zou and Yuan (2008) established the asymptotic normality of $\hat{\beta}^{CQR}$ and showed that the asymptotic relative efficiency of the CQR with respect to the LS is always larger than 70%.

3 Proposed Methodology

Given the mixture structure of the objective function in (7) and the link between the ALD and QR, it is natural for us to connect the CQR with the ALD-based estimation. To this end, we seek to propose a new mixture regression model based on the asymmetric Laplace distributions, develop its maximum likelihood estimation procedure through an EM-type algorithm and then provide its application to the CQR in this section.

3.1 The Model

We assume that error term ε_i in model (1) follows a K -component mixture of the ALDs with mixing proportions π_k . That is

$$Y_i = X_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \sum_{k=1}^K \pi_k \text{ALD}(b_k, \sigma/\psi_k, \tau_k), \quad (8)$$

where $\pi_k > 0$, $\sum_{k=1}^K \pi_k = 1$, and b_k is the $100\tau_k\%$ quantile of the k th component. The scale ψ_k makes the variance of the k th component equal to σ^2 , $\psi_k^2 = \kappa_k^2 + \nu_k^2$ with $\kappa_k = \frac{1-2\tau_k}{\tau_k(1-\tau_k)}$ and $\nu_k^2 = \frac{2}{\tau_k(1-\tau_k)}$. $0 < \tau_1 < \dots < \tau_K < 1$ are the equally spaced quantiles with $\tau_k = \frac{k}{K+1}$ for $k = 1, \dots, K$ and $K \geq 2$. When $K = 1$, the model reduces to the usual quantile regression.

Under model (8), the pdf of $Y_i|X_i$ can be written as

$$\sum_{k=1}^K \pi_k f(Y_i; \mu_{ik}, \sigma/\psi_k, \tau_k), \quad \mu_{ik} = X_i^T \boldsymbol{\beta} + b_k, \quad (9)$$

where $f(Y_i; \mu_{ik}, \sigma/\psi_k, \tau_k)$ is the pdf of the distribution $\text{ALD}(\mu_{ik}, \sigma/\psi_k, \tau_k)$ evaluated at Y_i . Then the log-likelihood function based on observations $\mathcal{O} = \{X_i, Y_i\}_{i=1}^n$ is given by

$$l_{obs}(\boldsymbol{\theta}, \mathbf{b}; \mathcal{O}) = \sum_{i=1}^n \log \left[\sum_{k=1}^K \frac{\pi_k \psi_k \tau_k (1 - \tau_k)}{\sigma} \exp \left\{ -\rho_{\tau_k} \left(\frac{\psi_k (Y_i - X_i^T \boldsymbol{\beta} - b_k)}{\sigma} \right) \right\} \right], \quad (10)$$

where the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\pi}^T, \boldsymbol{\beta}^T, \sigma)^T$ with mixing probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^T$ and quantile parameters $\mathbf{b} = (b_1, \dots, b_K)^T$.

Note that from the pdf in (9) the proposed model (8) can also be written as a mixture of K linear regression models with the k th component taking the form

$$Y_i = X_i^T \boldsymbol{\beta} + b_k + \tilde{\varepsilon}_{ik}, \quad \tilde{\varepsilon}_{ik} \sim \text{ALD}(0, \sigma/\psi_k, \tau_k). \quad (11)$$

3.2 Estimation Procedure via a Two-layer EM Algorithm

We seek to estimate parameters $\boldsymbol{\theta}$ and $\mathbf{b}$ by maximizing the log-likelihood in (10). The maximum likelihood estimates of parameters in a finite mixture model are typically achieved by using the EM algorithm. In view of the mixture structure of the proposed model and the connection between the QR method and MLE based on the ALD, we propose an EM-type algorithm, called the two-layer EM algorithm. Specifically, the first-layer EM procedure is pursued through reformulating a finite mixture model into an incomplete data problem, as is commonly used in estimation of other finite mixture models. Then it nests the second-layer EM in its M-step based on the normal-variance-mean mixture representation of the ALD for efficiently computing the regression coefficient $\boldsymbol{\beta}$.

3.2.1 First-layer EM Algorithm

To maximize $l_{obs}(\boldsymbol{\theta}, \mathbf{b}; \mathcal{O})$, we first treat component memberships as missing data and introduce latent variables G_{ik} with

$$G_{ik} = \begin{cases} 1, & \text{if } i^{th} \text{ observation is generated from the } k^{th} \text{ component} \\ 0, & \text{otherwise} \end{cases},$$

where $i = 1, \dots, n$, $k = 1, \dots, K$. Let $G = (G_1, \dots, G_n)^T$ with $G_i = (G_{i1}, \dots, G_{iK})^T$ being the component indicator of $\mathcal{O}_i = (X_i, Y_i)$. The complete log-likelihood function of model (8) based on observations $(\mathcal{O}, G)$ can then be written as

$$\begin{aligned} l_c(\boldsymbol{\theta}, \mathbf{b}; \mathcal{O}, G) &= \sum_{i=1}^n \sum_{k=1}^K G_{ik} \log \pi_k + \sum_{i=1}^n \sum_{k=1}^K G_{ik} \log f(Y_i; \mu_{ik}, \sigma / \psi_k, \tau_k) \\ &:= l_{c1}(\boldsymbol{\pi}; G) + l_{c2}(\mathbf{b}, \boldsymbol{\beta}, \sigma; \mathcal{O}, G), \end{aligned}$$

where $\mu_{ik} = X_i^T \boldsymbol{\beta} + b_k$. Plugging the ALD pdf defined by (3) into the term $l_{c2}(\mathbf{b}, \boldsymbol{\beta}, \sigma; \mathcal{O}, G)$, it follows that

$$l_{c2}(\mathbf{b}, \boldsymbol{\beta}, \sigma; \mathcal{O}, G) = l_{c2}^{(1)}(\sigma; G) - l_{c2}^{(2)}(\mathbf{b}, \boldsymbol{\beta}, \sigma; \mathcal{O}, G), \quad (12)$$

where

$$\begin{aligned} l_{c2}^{(1)}(\sigma; G) &= \sum_{i=1}^n \sum_{k=1}^K G_{ik} \log \left(\frac{\psi_k \tau_k (1 - \tau_k)}{\sigma} \right), \\ l_{c2}^{(2)}(\mathbf{b}, \boldsymbol{\beta}, \sigma; \mathcal{O}, G) &= \frac{1}{\sigma} \sum_{i=1}^n \sum_{k=1}^K G_{ik} \psi_k \rho_{\tau_k} (Y_i - X_i^T \boldsymbol{\beta} - b_k). \end{aligned}$$

In the E-step of the EM algorithm, it requires to compute the expectation of the unobservable information G_{ik} given the observed data $\mathcal{O}$ and the current estimates of unknown parameters $\boldsymbol{\theta}^{(m)}$ and $\mathbf{b}^{(m)}$. That is

$$\delta_{ik}^{(m)} = E(G_{ik} | \mathcal{O}; \boldsymbol{\theta}^{(m)}, \mathbf{b}^{(m)}) = \frac{\pi_k^{(m)} f(Y_i; \mu_{ik}^{(m)}, \sigma^{(m)} / \psi_k, \tau_k)}{\sum_{k=1}^K \pi_k^{(m)} f(Y_i; \mu_{ik}^{(m)}, \sigma^{(m)} / \psi_k, \tau_k)}, \quad (13)$$

where $\mu_{ik}^{(m)} = X_i^T \boldsymbol{\beta}^{(m)} + b_k^{(m)}$. With the result in (13), the conditional expectation of the complete log-likelihood function, denoted by $\tilde{l}_c(\boldsymbol{\theta}, \mathbf{b}; \mathcal{O})$, can be derived straightforwardly as

$$\tilde{l}_c(\boldsymbol{\theta}, \mathbf{b}; \mathcal{O}) = \tilde{l}_{c1}(\boldsymbol{\pi}) + \tilde{l}_{c2}(\mathbf{b}, \boldsymbol{\beta}, \sigma; \mathcal{O}),$$

where

$$\tilde{l}_{c1}(\boldsymbol{\pi}) = \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \log \pi_k, \quad (14)$$

$$\tilde{l}_{c_2}(\mathbf{b}, \boldsymbol{\beta}, \sigma; \mathcal{O}) = \tilde{l}_{c_2}^{(1)}(\sigma) - \tilde{l}_{c_2}^{(2)}(\mathbf{b}, \boldsymbol{\beta}, \sigma; \mathcal{O}), \quad (15)$$

$$\tilde{l}_{c_2}^{(1)}(\sigma) = \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \log \left(\frac{\psi_k \tau_k (1 - \tau_k)}{\sigma} \right), \quad (16)$$

$$\tilde{l}_{c_2}^{(2)}(\mathbf{b}, \boldsymbol{\beta}, \sigma; \mathcal{O}) = \frac{1}{\sigma} \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \psi_k \rho_{\tau_k} \left(Y_i - X_i^T \boldsymbol{\beta} - b_k \right). \quad (17)$$

In the M-step, the parameters are re-estimated by maximizing $\tilde{l}_c(\boldsymbol{\theta}, \mathbf{b}; \mathcal{O})$ with respect to $\boldsymbol{\theta}$ and $\mathbf{b}$. First, it is easy to have

$$\pi_k^{(m+1)} = \frac{1}{n} \sum_{i=1}^n \delta_{ik}^{(m)}. \quad (18)$$

The updated estimates $(\mathbf{b}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \sigma^{(m+1)})$ can be found through maximizing $\tilde{l}_{c_2}(\mathbf{b}, \boldsymbol{\beta}, \sigma; \mathcal{O})$ in (15) with respect to $\mathbf{b}, \boldsymbol{\beta}$ and σ , respectively. Note that the term $\tilde{l}_{c_2}^{(1)}(\sigma)$ in (16) doesn't include parameters $\boldsymbol{\beta}$ and $\mathbf{b}$, and thus the updates for $\boldsymbol{\beta}$ and $\mathbf{b}$ are independent of that for σ . Particularly, the updates $(\mathbf{b}^{(m+1)}, \boldsymbol{\beta}^{(m+1)})$ can be obtained via

$$(\mathbf{b}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}) = \arg \min_{b_1, \dots, b_K; \boldsymbol{\beta}} \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \psi_k \rho_{\tau_k} \left(Y_i - X_i^T \boldsymbol{\beta} - b_k \right), \quad (19)$$

and $\sigma^{(m+1)}$ is then obtained by maximizing $\tilde{l}_{c_2}(\mathbf{b}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \sigma; \mathcal{O})$ with respect to σ , which gives

$$\sigma^{(m+1)} = \frac{\sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \psi_k \rho_{\tau_k} \left(Y_i - X_i^T \boldsymbol{\beta}^{(m+1)} - b_k^{(m+1)} \right)}{n}. \quad (20)$$

Substituting $\boldsymbol{\theta}^{(m)}$ and $\mathbf{b}^{(m)}$ by their updated estimates $\boldsymbol{\theta}^{(m+1)}$ and $\mathbf{b}^{(m+1)}$ in (13), $\delta^{(m+1)}$ can be updated. The algorithm repeats the above E-step and M-step until convergence is achieved.

Remark 1. For a special case of model (8) in which the mixing probability π_k is fixed to be $1/K$ and $\sigma/\psi_k = 1$, i.e. $\epsilon_i \sim \sum_{k=1}^K \frac{1}{K} \text{ALD}(b_k^*, 1, \tau_k)$, we have the log-likelihood function based on observations

$$l_{\text{obs}}(\boldsymbol{\beta}, \mathbf{b}; \mathcal{O}) = \sum_{i=1}^n \log \left[\sum_{k=1}^K \frac{\tau_k (1 - \tau_k)}{K} \exp \left\{ -\rho_{\tau_k} (Y_i - X_i^T \boldsymbol{\beta} - b_k) \right\} \right]$$

and the complete log-likelihood

$$l_c(\boldsymbol{\beta}, \mathbf{b}; \mathcal{O}, G) = \sum_{i=1}^n \sum_{k=1}^K G_{ik} \left[-\log K + \log(\tau_k (1 - \tau_k)) - \rho_{\tau_k} (Y_i - X_i^T \boldsymbol{\beta} - b_k) \right].$$

Note that G_{ik} is independent of $\mathcal{O}$ and $P(G_{ik} = 1) = 1/K$ in this case. In the E-step of the EM algorithm, after dropping the irrelevant constants, the conditional expectation of the complete log-likelihood becomes

$$\tilde{l}_c(\boldsymbol{\theta}, \mathbf{b}; \mathcal{O}) = - \sum_{i=1}^n \sum_{k=1}^K \rho_{\tau_k} (Y_i - X_i^T \boldsymbol{\beta} - b_k).$$

Therefore, maximizing $\tilde{l}_c(\boldsymbol{\theta}, \mathbf{b}; \mathcal{O})$ over β and $\mathbf{b}$ in the M -step results in the CQR estimator as defined in (7).

Remark 2. The estimator obtained in (19) can be viewed as a weighted composite quantile regression estimator with data-driven weights $\delta_{ik}^{(m)}\psi_k$ adaptively decided by the observed data $\mathcal{O}$ and the current estimates $\theta^{(m)}$ and $b^{(m)}$. The weighted composite quantile regression (WCQR) method has been proposed by Bradic et al. (2011) and Jiang et al. (2012) in the form of

$$(\hat{b}_1, \dots, \hat{b}_K; \hat{\beta}^{CQR}; \mathbf{w}) = \arg \min_{b_1, \dots, b_K; \beta} \sum_{k=1}^K w_k \left\{ \sum_{i=1}^n \rho_{\tau_k}(Y_i - X_i^T \beta - b_k) \right\},$$

where $\mathbf{w} = (w_1, \dots, w_K)^T$ is a vector of weights such that $\|\mathbf{w}\| = 1$ with $\|\cdot\|$ denoting the Euclidean norm, and w_k controls the contribution of the τ_k -th QR. However, it is worth noting that the existing WCQR is essentially different from our data-driven weighted estimation obtained from (19) because the WCQR imposes the same weight w_k for all the observations in the component k and allows w_k to be negative. We conjecture that our proposed estimator is efficient due to its data-driven weighting scheme. It will be verified through numerical studies in Section 4.

3.2.2 Second-layer EM Algorithm

For the minimization problem in (19), the solution cannot be obtained explicitly due to the non-differentiable objective function. Although it can be handled as a large-scale linear programming problem by expanding the covariate space with some new ancillary variables, this would be computationally intensive when a large number of quantiles are considered. Note that the objective function is also a mixture of the objective functions from different quantile regressions. It motivates us to apply the EM algorithm again to this optimization problem. Particularly, we can obtain closed form updated estimates of $(\mathbf{b}, \beta)$ in each iteration step.

In line of the connection between the QR procedure and the MLE based on the ALD error, (19) is equivalent to

$$(\mathbf{b}^{(m+1)}, \beta^{(m+1)}) = \arg \max_{b_1, \dots, b_K; \beta} \sum_{k=1}^K \sum_{i=1}^n \delta_{ik}^{(m)} \psi_k \log f(Y_i - X_i^T \beta - b_k; 0, 1, \tau_k), \quad (21)$$

where $f(Y_i - X_i^T \beta - b_k; 0, 1, \tau_k)$ is the pdf of the distribution $\text{ALD}(0, 1, \tau_k)$ evaluated at $Y_i - X_i^T \beta - b_k$. For fixed $i = 1, \dots, n$ and $k = 1, \dots, K$, the term $\log f(Y_i - X_i^T \beta - b_k; 0, 1, \tau_k)$ can be viewed as the log-likelihood function of the following model

$$Y_i = X_i^T \beta + b_k + \tilde{\varepsilon}_{ik}, \quad \tilde{\varepsilon}_{ik} \sim \text{ALD}(0, 1, \tau_k). \quad (22)$$

Then the computation for (21) is simplified based on the fact that a asymmetric Laplace distributed random variable is a normal-variance-mean mixture with exponential mixing density as given in (5). Let z_i be i.i.d. latent variables from an exponential distribution with mean 1, and u_i be independently distributed as $N(0, z_i)$ given the value of z_i . The model (22) can be rewritten as

$$Y_i = X_i^T \beta + b_k + \kappa_k z_i + \nu_k u_i, \quad (23)$$

where $\kappa_k = \frac{1-2\tau_k}{\tau_k(1-\tau_k)}$ and $\nu_k^2 = \frac{2}{\tau_k(1-\tau_k)}$. For the k th component $\log f(Y_i - X_i^T \beta - b_k; 0, 1, \tau_k)$, the complete log-likelihood function based on observations $(\mathcal{O}_i, z_i)$ is

$$\begin{aligned} l_{ci}(\beta, b_k; \mathcal{O}_i, z_i) &= -\frac{1}{2} \log(2\pi\nu_k^2 z_i) - \frac{(Y_i - X_i^T \beta - b_k - \kappa_k z_i)^2}{2\nu_k^2 z_i} - z_i \\ &= -\frac{1}{2} \log(2\pi\nu_k^2 z_i) - \left\{ \frac{(Y_i - X_i^T \beta - b_k)^2}{2\nu_k^2} z_i^{-1} - \frac{\kappa_k}{\nu_k^2} (Y_i - X_i^T \beta - b_k) + \frac{\kappa_k^2}{2\nu_k^2} z_i \right\} - z_i \\ &\propto -\frac{(Y_i - X_i^T \beta - b_k)^2}{2\nu_k^2} z_i^{-1} + \frac{\kappa_k}{\nu_k^2} (Y_i - X_i^T \beta - b_k). \end{aligned}$$

Given observed data $\mathcal{D} = \{\mathcal{O}, z\}$ with $z = \{z_i, i = 1, \dots, n\}$, the objective function in (21) becomes

$$l_c(\beta, \mathbf{b}; \mathcal{D}) = \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \psi_k \left\{ -\frac{(Y_i - X_i^T \beta - b_k)^2}{2\nu_k^2} z_i^{-1} + \frac{\kappa_k}{\nu_k^2} (Y_i - X_i^T \beta - b_k) \right\}. \quad (24)$$

In the E-step of the second layer EM algorithm, we compute the conditional expectation $E(l_c(\beta, \mathbf{b}; \mathcal{D}) | \mathcal{O}, \beta^{(m)}, \mathbf{b}^{(m)})$, denoted by $\tilde{l}_c(\beta, \mathbf{b}; \mathcal{O})$ for simplicity. By the calculation given in Appendix 1, we have

$$\omega_{ik}^{(m)} = E(z_i^{-1} | \mathcal{O}_i, k; \beta^{(m)}, b_k^{(m)}) = \frac{1}{\tau_k(1-\tau_k) |Y_i - X_i^T \beta^{(m)} - b_k^{(m)}|}. \quad (25)$$

It follows from (24) that

$$\begin{aligned} \tilde{l}_c(\beta, \mathbf{b}; \mathcal{O}) &= \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \psi_k \left[-\omega_{ik}^{(m)} \frac{(Y_i - X_i^T \beta - b_k)^2}{2\nu_k^2} + \frac{\kappa_k}{\nu_k^2} (Y_i - X_i^T \beta - b_k) \right] \\ &= \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \psi_k \left[-\tilde{\omega}_{ik}^{(m)} \frac{(Y_i - X_i^T \beta - b_k)^2}{2} + (1 - 2\tau_k)(Y_i - X_i^T \beta - b_k) \right], \end{aligned} \quad (26)$$

where redefined weights

$$\tilde{\omega}_{ik}^{(m)} = \frac{2\omega_{ik}^{(m)}}{\nu_k^2} = \left| Y_i - X_i^T \beta^{(m)} - b_k^{(m)} \right|^{-1}. \quad (27)$$

In the M-step, we first fix $\mathbf{b}$ at the current estimate $\mathbf{b}^{(m)}$ and then estimate β by maximizing $\tilde{l}_c(\beta, \mathbf{b}^{(m)})$ with respect to β , equivalently, solving equations

$$\frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \psi_k X_i \left\{ \tilde{\omega}_{ik}^{(m)} (Y_i - X_i^T \beta - b_k^{(m)}) - (1 - 2\tau_k) \right\} = 0.$$

A straightforward calculation leads to the following closed-form updated estimate of β .

$$\beta^{(m+1)} = \left(\sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \psi_k \tilde{\omega}_{ik}^{(m)} X_i X_i^T \right)^{-1} \left[\sum_{i=1}^n \sum_{k=1}^K \delta_{ik}^{(m)} \psi_k X_i \left\{ \tilde{\omega}_{ik}^{(m)} (Y_i - b_k^{(m)}) - (1 - 2\tau_k) \right\} \right]. \quad (28)$$

Given $\beta^{(m+1)}$, maximizing $\tilde{l}_c(\beta^{(m+1)}, \mathbf{b}; \mathcal{O})$ with respect to b_k leads to

$$\frac{\psi_k}{2} \sum_{i=1}^n \delta_{ik}^{(m)} \left\{ \tilde{\omega}_{ik}^{(m)} (Y_i - X_i^T \beta^{(m+1)} - b_k) - (1 - 2\tau_k) \right\} = 0, \quad k = 1, 2, \dots, K.$$

Hence, the updated estimate of b_k is obtained as

$$b_k^{(m+1)} = \left(\sum_{i=1}^n \delta_{ik}^{(m)} \tilde{\omega}_{ik}^{(m)} \right)^{-1} \left[\sum_{i=1}^n \delta_{ik}^{(m)} \left\{ \tilde{\omega}_{ik}^{(m)} (Y_i - X_i^T \beta^{(m+1)}) - (1 - 2\tau_k) \right\} \right]. \quad (29)$$

for $k = 1, 2, \dots, K$.

3.2.3 Implementation

To simplify the expressions of above updated estimates, we denote by $\mathbf{1}_n$ a $n \times 1$ vector with all elements equal to 1, $\mathbf{C} = \mathbf{1} - 2\boldsymbol{\tau}$ with $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)^T$, $\boldsymbol{\psi}^2 = \boldsymbol{\kappa}^2 + \boldsymbol{\nu}^2$ with $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_K)^T$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K)^T$. Furthermore,

$$\begin{aligned} \Delta^{(m)} &= (\delta_{ik}^{(m)})_{n \times K}, \quad \tilde{\Omega}^{(m)} = (\tilde{\omega}_{ik}^{(m)})_{n \times K}, \\ W^{(m)} &= \text{diag} \left\{ \sum_{k=1}^K \delta_{1k}^{(m)} \tilde{\omega}_{1k}^{(m)} \psi_k, \dots, \sum_{k=1}^K \delta_{nk}^{(m)} \tilde{\omega}_{nk}^{(m)} \psi_k \right\} = \text{diag} \left\{ \left(\Delta^{(m)} \circ \tilde{\Omega}^{(m)} \right) \boldsymbol{\psi} \right\}, \\ \Phi^{(m)} &= \text{diag} \left\{ \sum_{i=1}^n \delta_{i1}^{(m)} \tilde{\omega}_{i1}^{(m)}, \dots, \sum_{i=1}^n \delta_{iK}^{(m)} \tilde{\omega}_{iK}^{(m)} \right\} = \text{diag} \left\{ \left(\Delta^{(m)} \circ \tilde{\Omega}^{(m)} \right)^T \mathbf{1}_n \right\}, \\ \Upsilon^{(m)} &= \text{diag} \left\{ \sum_{i=1}^n \delta_{i1}^{(m)}, \dots, \sum_{i=1}^n \delta_{iK}^{(m)} \right\} = \text{diag} \{ \Delta^{(m)T} \mathbf{1}_n \}, \\ \Xi^{(m)} &= (\Xi_{ik}^{(m)})_{n \times K}, \quad \Xi_{ik}^{(m)} = \rho_{\tau_k} \left(Y_i - X_i^T \beta^{(m+1)} - b_k^{(m+1)} \right), \end{aligned}$$

where the notation $(\Delta^{(m)} \circ \tilde{\Omega}^{(m)})$ denotes the entry-wise multiplication of matrices of $\delta_{ik}^{(m)}$ and $\tilde{\omega}_{ik}^{(m)}$. So it returns a $n \times K$ matrix with the (i, k) th entry being the product of $\delta_{ik}^{(m)}$ and $\tilde{\omega}_{ik}^{(m)}$. Similarly, we denote $(\mathbf{C} \circ \boldsymbol{\psi})$ and $(\mathbf{b}^{(m)} \circ \boldsymbol{\psi})$ as the entry-wise multiplication of vectors $\mathbf{C}$ and $\boldsymbol{\psi}$ with the k th entry being $C_k \psi_k$ and $b_k^{(m)} \psi_k$, respectively. Using these notations, we can rewrite the

updating formulas for π , β , b and σ in (18), (20), (28) and (29), respectively as follows:

$$\pi^{(m+1)} = \frac{\Delta^{(m)T} \mathbf{1}_n}{n}, \quad (30)$$

$$\beta^{(m+1)} = \left(\mathbf{X}^T W^{(m)} \mathbf{X} \right)^{-1} \mathbf{X}^T \left\{ W^{(m)} \mathbf{Y} - (\Delta^{(m)} \circ \tilde{\Omega}^{(m)}) (\mathbf{b}^{(m)} \circ \psi) - \Delta^{(m)} (\mathbf{C} \circ \psi) \right\}, \quad (31)$$

$$\mathbf{b}^{(m+1)} = \{ \Phi^{(m)} \}^{-1} \left\{ \left(\Delta^{(m)} \circ \tilde{\Omega}^{(m)} \right)^T (\mathbf{Y} - \mathbf{X}^T \beta^{(m+1)}) - \Upsilon^{(m)} \mathbf{C} \right\}, \quad (32)$$

$$\sigma^{(m+1)} = \frac{\mathbf{1}_n^T \left(\Delta^{(m)} \circ \Xi^{(m)} \right) \psi}{n}. \quad (33)$$

In summary, the proposed two-layer EM algorithm can be described as follows:

Algorithm 1 ALD mixture regression

- 1: Choose initial values $\theta^{(0)} = (\pi^{(0)T}, \beta^{(0)T}, \sigma^{(0)})^T$ and $\mathbf{b}^{(0)}$; Let $m = 0$.
 - 2: Given current estimates $\theta^{(m)}$ and $\mathbf{b}^{(m)}$, compute $\delta_{ik}^{(m)}$ and $\tilde{\omega}_{ik}^{(m)}$ through (13) and (27).
 - 3: Perform formulas (30)-(33) to obtain updated estimates $\pi^{(m+1)}, \beta^{(m+1)}, \sigma^{(m+1)}$ and $\mathbf{b}^{(m+1)}$; Let $m = m + 1$.
 - 4: Iterate between steps 2 and 3 until the algorithm converges numerically.
-

Remark 3. It is easy to see that the resulting estimator of β is robust to outliers from the updating formula (31), which is actually an iteratively re-weighted least squares (IRLS) procedure. Note that $\tilde{\omega}_{ik}^{(m)}$ in (27) is reciprocally related to the term $|Y_i - X_i^T \beta^{(m)} - b_k^{(m)}|$, implying that the larger residuals give smaller values of $\tilde{\omega}_{ik}^{(m)}$ and hence impose less weights of the corresponding observations on the estimates. Besides robustness, due to the equivalence between the proposed estimation procedure and the CQR method, our resulting estimator of β naturally inherits other nice properties from the CQR estimator, such as the asymptotic normality and relative efficiency. These properties of the proposed estimator are demonstrated through simulation in the next section.

To implement the proposed algorithm, extra attention should be paid to the E-steps of both layers. In the first layer, when assuming perfect fit, the quantile regression estimator satisfies that $|Y_i - X_i^T \beta^{(m)} - b_k^{(m)}| = 0$ for a subset of observations Koenker (2005). In this case, $\tilde{\omega}_{ik}^{(m)}$ in (27) will be considerably large, leading to numerical instability of the algorithm. To overcome this problem, some modified weighting strategy can be incorporated similar to the work of Phillips (2002). That is, choose a small $\epsilon > 0$ and set $\tilde{\omega}_{ik}^{(m)}$ as in (27) if $|Y_i - X_i^T \beta^{(m)} - b_k^{(m)}| \geq \epsilon$ for all observations; otherwise, set $\tilde{\omega}_{ik}^{(m)} = 1$ for cases with $|Y_i - X_i^T \beta^{(m)} - b_k^{(m)}| < \epsilon$, and $\tilde{\omega}_{ik}^{(m)} = \frac{\epsilon}{|Y_i - X_i^T \beta^{(m)} - b_k^{(m)}|}$ for all other cases. These adjusted weights are consistent with the original ones in the sense that cases with smaller residuals should get heavier weights. Alternatively, we adopt another simplified weighting

strategy, i.e., set $\tilde{\omega}_{ik}^{(m)} = \{|Y_i - X_i^T \boldsymbol{\beta}^{(m)} - b_k^{(m)}| + \epsilon\}^{-1}$ and $\epsilon = 10^{-6}$ in our numerical analysis. On the other hand, numerical instability could also occur in the second layer of the algorithm if the weights $\delta_{ik}^{(m)}$ are very small. A common way to deal with it is to impose a hard threshold on $\delta_{ik}^{(m)}$ in (13). In particular, for a pre-assigned small $\tilde{\epsilon} > 0$, we set $\tilde{\delta}_{ik}^{(m)} = \tilde{\epsilon}$ if $\delta_{ik}^{(m)} < \tilde{\epsilon}$, and $\tilde{\delta}_{ik}^{(m)} = \delta_{ik}^{(m)}$ otherwise. Then we replace $\delta_{ik}^{(m)}$ with $\tilde{\delta}_{ik}^{(m)}$ for the iteration as did in Song et al. (2014). Again, we use $\tilde{\epsilon} = 10^{-6}$ in our numerical studies.

We start the proposed algorithm near the ordinary LS or a LAD solution for $\boldsymbol{\beta}$ with equal weights $\omega_k = 1/K$ and sample $100\tau_k\%$ quantile for b_k , $k = 1, \dots, K$ unless otherwise specified. This initial value of $\boldsymbol{\beta}$ appears to work well in our numerical studies. We note that the algorithm proposed maintains the framework of the EM algorithm and thus the monotonic increasing feature of the log-likelihood (10), i.e., $l_{obs}(\boldsymbol{\theta}^{(m+1)}, \mathbf{b}^{(m+1)}; \mathcal{O}) \geq l_{obs}(\boldsymbol{\theta}^{(m)}, \mathbf{b}^{(m)}; \mathcal{O})$. However, the EM algorithm is not guaranteed to converge to the MLE. When multiple local maxima are suspected, one can start the algorithm from several different randomly chosen initial values in an attempt to find the global maximum. Finally, the proposed algorithm terminates when the new iterate fails to improve the objective function by a specified tolerance, say 10^{-6} , which is a good trade-off between precision and computation time of the solution.

3.3 Application to Composite Quantile Regression

We now apply the algorithm established in Section 3.3 to solve the minimization problem in (7) directly. The CQR method developed by Zou and Yuan (2008) aims to estimate the coefficient $\boldsymbol{\beta}$ only, while the estimation of σ is not considered. As addressed in Remark 1, letting $\delta_{ik} = 1/K$ and $\sigma/\psi_k = 1$ in the estimation procedure in Section 3.3 results in the CQR estimator as given by (7), providing a new likelihood-based approach for the composite quantile regression estimation. For brevity's sake, we summarize main steps of the corresponding algorithm below, while more details of the algorithm construction are provided in Appendix 2.

The main advantage of the algorithm above is that the closed form updates for parameters in each iteration step are obtained explicitly, instead of resorting to linear programming optimization techniques as in the existing work. Computational complexity can be reduced significantly.

3.4 Selection of the number of components K

To determine the number of components K is an important and challenging issue for mixture regression models, including the ALD mixture regression proposed. Tian et al. (2016) studied the

Algorithm 2 Composite quantile regression

- 1: Choose initial values $\beta^{(0)}$ and $\mathbf{b}^{(0)}$; Let $m = 0$.
- 2: Calculate $\tilde{\omega}_{ik}^{(m)} = |Y_i - X_i^T \beta^{(m)} - b_k^{(m)}|^{-1}$.
- 3: Perform the following formulas to get the updated estimates of β and b ,

$$\begin{aligned}\beta^{(m+1)} &= \left(\mathbf{X}^T \tilde{W}^{(m)} \mathbf{X}\right)^{-1} \mathbf{X}^T \left\{ \tilde{W}^{(m)} \mathbf{Y} - \tilde{\Omega}^{(m)} \mathbf{b}^{(m)} - (\mathbf{1}_K^T \mathbf{C}) \mathbf{1}_n \right\}, \\ \mathbf{b}^{(m+1)} &= \{\tilde{\Phi}^{(m)}\}^{-1} \left\{ \tilde{\Omega}^{(m)T} (\mathbf{Y} - \mathbf{X}^T \beta^{(m+1)}) - n \mathbf{C} \right\},\end{aligned}$$

where

$$\begin{aligned}\tilde{W}^{(m)} &= \text{diag} \left\{ \sum_{k=1}^K \tilde{\omega}_{1k}^{(m)}, \dots, \sum_{k=1}^K \tilde{\omega}_{nk}^{(m)} \right\} = \text{diag} \{ \tilde{\Omega}^{(m)} \mathbf{1}_K \}, \\ \tilde{\Phi}^{(m)} &= \text{diag} \left\{ \sum_{i=1}^n \tilde{\omega}_{i1}^{(m)}, \dots, \sum_{i=1}^n \tilde{\omega}_{iK}^{(m)} \right\} = \text{diag} \{ \tilde{\Omega}^{(m)T} \mathbf{1}_n \}.\end{aligned}$$

Let $m = m + 1$.

- 4: Iterate between steps 2 and 3 until the algorithm converges numerically.
-

“AIC-like-criterion” and “BIC-like-criterion” to select K for the CQR only. Their criteria imposed a too heavy penalty on K and hence resulted in the selection of smaller values for K as they addressed.

In order to make a trade-off between the goodness of fit and model complexity, we propose a heuristic information criterion called ABIC, which is an analogy to the conventional BIC, for determining the number of components in the ALD mixture model (8). The ABIC is defined as

$$\text{ABIC}(K) = \frac{-2l_{\text{obs}}(\hat{\boldsymbol{\theta}}_K, \hat{\mathbf{b}}_K; \mathcal{O})/K + (2K + p) \log n}{n}, \quad (34)$$

where l_{obs} is the log-likelihood function given in (10); $\hat{\boldsymbol{\theta}}_K = (\hat{\boldsymbol{\pi}}_K^T, \hat{\boldsymbol{\beta}}_K^T, \hat{\sigma}_K)^T$, $\hat{\mathbf{b}}_K$ and $\hat{\sigma}_K$ are proposed estimates; $2K + p$ is the number of independent parameters. The value of K is then determined by minimizing $\text{ABIC}(K)$ with respect to K . Particularly letting $\pi_k = \frac{1}{K}$ for $k = 1, \dots, K$ and ignoring σ_K in the model (8), the value of K in the CQR can be obtained by

$$\hat{K}_{\text{CQR}} = \arg \min_K \frac{2}{K} \sum_{k=1}^K \sum_{i=1}^n \rho_{\tau_k}(Y_i - X_i^T \hat{\beta}^{(K)} - \hat{b}_k^{(K)}) + (K + p) \log n. \quad (35)$$

Note that this heuristic approach for determining K has not been rigorously justified. Whether it leads to an optimal value of K consistently for the proposed model is still an open question that needs further investigation, but is beyond the scope of the current work. Our simulation results in the next section will demonstrate its effectiveness.

4 Simulation Study

We carry out simulations to examine the performance of the proposed estimation procedures in Section 3.2 and Section 3.3 (labeled ALD-Mix and CQR, respectively) in comparison with the least squares (LS) and least absolute derivation (LAD) methods. We generate 500 data sets with $n = 100$ and 200 each according to the model

$$Y = X^T \beta + \varepsilon, \quad \text{where } X \sim N_p(0, \Sigma_x), p = 3, \Sigma_x = (0.5^{|i-j|})_{p \times p} \text{ and } \beta = (3, 1.5, 2). \quad (36)$$

This regression model is the oracle model considered in Zou and Yuan (2008), Tibshirani (1996) and Fan and Li (2001). The following six distributions are considered for the error term.

CASE 1. $\varepsilon \sim N(0, 3)$;

CASE 2. ε follows a mixture of normal distributions, $0.95N(0, 1) + 0.05N(0, 41)$, and the variance of ε is 3.

CASE 3. ε follows a Student t -distribution with 3 degrees of freedom.

CASE 4. ε follows the Cauchy distribution

CASE 5. $\varepsilon = \sigma \varepsilon^*$, where ε^* follows the double exponential distribution, and let $\sigma = \sqrt{6}/2$ to make the variance of ε be 3.

CASE 6. $\varepsilon = \sigma \varepsilon^*$, where ε^* follows the logistic distribution, and let $\sigma = \frac{3}{\pi}$ to make the variance of ε be 3.

Case 1 is used to show the efficiency of the proposed estimation procedures compared with the LS when the error is exactly normally distributed. In this case, the LS estimator reduces to the traditional maximum likelihood estimator (MLE). The contaminated normal mixture model in Case 2 mimics the situation with outliers, where 20% of data generated from $N(0, 41)$ are likely to be low leverage outliers. Case 3 refers to a heavy-tailed distribution. Case 4 specifies the error having an extreme heavy-tailed distribution. Case 5 aims to evaluate the efficiency of the proposed estimation procedures relative to the LAD when the error is double exponentially distributed. In this case the LAD estimator reduces to the traditional MLE. Case 6 is employed to assess the efficiency of the proposed estimators when the error has the logistic distribution. In this case, the CQR estimator achieves the information bound asymptotically (Zou and Yuan, 2008).

Table 1: Simulation results of average biases of estimated regression coefficients and average model errors (ME) obtained by different estimation methods based on 500 replications, where numbers in parentheses are corresponding sampling standard deviations.

Model	β	n=100				n=200			
		ALD-Mix	CQR	LS	LAD	ALD-Mix	CQR	LS	LAD
Case 1	β_1	0.0014 (0.2006)	0.0032 (0.2073)	0.0073 (0.1965)	0.0021 (0.2621)	0.0079 (0.1454)	0.0042 (0.1441)	0.0051 (0.1429)	0.0065 (0.1766)
	β_2	-0.0020 (0.2232)	-0.0040 (0.2305)	-0.0030 (0.2253)	-0.0119 (0.2851)	-0.0093 (0.1659)	-0.0104 (0.1660)	-0.0092 (0.1621)	-0.0192 (0.2048)
	β_3	0.0039 (0.2070)	0.0041 (0.2109)	-0.0007 (0.2046)	0.0064 (0.2525)	0.0048 (0.1413)	0.0068 (0.1439)	0.0044 (0.1394)	0.0113 (0.1759)
	ME	0.0945 (0.0848)	0.0987 (0.0808)	0.0925 (0.0783)	0.1506 (0.1228)	0.0473 (0.0440)	0.0481 (0.0390)	0.0461 (0.0381)	0.0718 (0.0600)
Case 2	β_1	0.0025 (0.1325)	0.0050 (0.1312)	0.0009 (0.2115)	0.0035 (0.1484)	-0.0010 (0.0951)	-0.0022 (0.0904)	-0.0011 (0.1341)	-0.0019 (0.1082)
	β_2	0.0050 (0.1520)	0.0006 (0.1431)	0.0075 (0.2247)	0.0036 (0.1716)	0.0008 (0.1049)	0.0014 (0.0996)	0.0029 (0.1409)	-0.0017 (0.1226)
	β_3	-0.0014 (0.1347)	-0.0012 (0.1316)	-0.0063 (0.2085)	0.0026 (0.1545)	-0.0059 (0.0926)	-0.0055 (0.0882)	-0.0007 (0.1298)	-0.0045 (0.1103)
	ME	0.0419 (0.0479)	0.0392 (0.0349)	0.0982 (0.1035)	0.0542 (0.0495)	0.0197 (0.0196)	0.0181 (0.0146)	0.0393 (0.0371)	0.0267 (0.0206)
Case 3	β_1	-0.0014 (0.1526)	0.0001 (0.1483)	0.0064 (0.1975)	0.0007 (0.1647)	0.0104 (0.1086)	0.0121 (0.1041)	0.0110 (0.1382)	0.0127 (0.1117)
	β_2	-0.0035 (0.1773)	0.0007 (0.1670)	-0.0042 (0.2351)	-0.0047 (0.1792)	0.0012 (0.1105)	-0.0028 (0.1083)	-0.0013 (0.1472)	-0.0059 (0.1171)
	β_3	0.0096 (0.1540)	0.0062 (0.1452)	0.0026 (0.1905)	0.0090 (0.1644)	-0.0056 (0.1071)	-0.0053 (0.1023)	-0.0033 (0.1383)	-0.0045 (0.1090)
	ME	0.0546 (0.0587)	0.0493 (0.0478)	0.0917 (0.0996)	0.0602 (0.0576)	0.0262 (0.0252)	0.0242 (0.0193)	0.0437 (0.0456)	0.0274 (0.0214)
Case 4	β_1	0.0079 (0.2042)	0.0158 (0.2299)	1.1577 (18.0526)	0.0123 (0.2163)	0.0057 (0.1382)	0.0032 (0.1505)	-0.5183 (13.9114)	0.0055 (0.1343)
	β_2	0.0072 (0.2352)	0.0043 (0.2782)	-0.3905 (16.6704)	0.0065 (0.2526)	-0.0047 (0.1569)	-0.0074 (0.1693)	1.4175 (31.1121)	-0.0074 (0.1521)
	β_3	-0.0052 (0.2033)	0.0028 (0.2444)	-0.4160 (27.2095)	0.0001 (0.2170)	0.0069 (0.1426)	0.0037 (0.1558)	-0.4188 (17.5218)	0.0038 (0.1388)
	ME	0.0922 (0.0863)	0.1268 (0.1126)	923.8960 (11924.2571)	0.1049 (0.0833)	0.0432 (0.0763)	0.0525 (0.0493)	664.5983 (577.1068)	0.0412 (0.0389)
Case 5	β_1	0.0006 (0.1810)	-0.0022 (0.1692)	-0.0020 (0.1976)	0.0037 (0.1773)	-0.0059 (0.1251)	-0.0069 (0.1191)	-0.0050 (0.1466)	-0.0084 (0.1144)
	β_2	0.0012 (0.2017)	0.0077 (0.1964)	0.0135 (0.2287)	0.0000 (0.2044)	0.0078 (0.1349)	0.0076 (0.1307)	0.0060 (0.1556)	0.0067 (0.1295)
	β_3	0.0006 (0.1768)	-0.0042 (0.1792)	-0.0089 (0.2065)	0.0004 (0.1740)	0.0034 (0.1277)	0.0052 (0.1205)	0.0036 (0.1432)	0.0046 (0.1151)
	ME	0.0720 (0.0781)	0.0677 (0.0616)	0.0913 (0.0867)	0.0697 (0.0607)	0.0362 (0.0341)	0.0326 (0.0266)	0.0472 (0.0372)	0.0304 (0.0271)
Case 6	β_1	-0.0019 (0.2024)	-0.0076 (0.1984)	-0.0052 (0.2025)	-0.0174 (0.2374)	-0.0018 (0.1389)	-0.0028 (0.1367)	0.0020 (0.1437)	-0.0069 (0.1542)
	β_2	0.0012 (0.2310)	0.0039 (0.2266)	0.0010 (0.2346)	0.0122 (0.2581)	-0.0032 (0.1581)	-0.0019 (0.1552)	-0.0051 (0.1586)	0.0030 (0.1760)
	β_3	0.0020 (0.2028)	0.0038 (0.2015)	0.0038 (0.2057)	-0.0035 (0.2256)	0.0053 (0.1384)	0.0007 (0.1343)	0.0012 (0.1411)	0.0041 (0.1530)
	ME	0.0941 (0.0853)	0.0907 (0.0813)	0.0952 (0.0929)	0.1203 (0.1019)	0.0440 (0.0372)	0.0416 (0.0341)	0.0452 (0.0393)	0.0549 (0.0440)

We consider quantiles with $\tau_k = \frac{k}{20}$, $k = 1, 2, \dots, 19$ as recommended in Zou and Yuan (2008). The performance of each estimation method is first evaluated based on the average bias, sample standard deviation (SD) and average model error (ME) of the resulting estimates across 500 replications. The model error, defined by $E[(\hat{\beta} - \beta)^T \Sigma_x (\hat{\beta} - \beta)]$, is an effective way of evaluating model adequacy versus model complexity. As an estimate of the model error, the average model error is computed by $ME = \frac{1}{500} \sum_{j=1}^{500} (\hat{\beta}_j - \beta)^T \Sigma_x (\hat{\beta}_j - \beta)$, where $\hat{\beta}_j$ is the estimate of β from the j^{th} simulation run for $j = 1, \dots, 500$.

Simulation results reported in Table 1 are very encouraging. In general, all the average biases are very small, which verifies that all four types of estimators are unbiased. When sample size n increases, the standard error for the average bias and the average model error decrease. Note that the LS and LAD estimators are equivalent to the MLE in cases 1 and 5, respectively. Therefore, each of them is efficient in the corresponding case when the sample size tends to infinity. As expected, the biases, SDs and MEs of the LS estimates are all smaller than those of the corresponding LAD estimates in Case 1, while the LAD estimates achieve smaller SDs and MEs in comparison with the LS in Case 5. It is worth noting that the proposed estimators perform very close to the efficient one in both cases in terms of standard error and ME. This finding demonstrates that the ALM-Mix and CQR estimators are nearly as efficient as the MLE in Cases 1 and 5. Our simulation results verify that the CQR estimator is asymptotically efficient in Case 6, where the CQR method yields the smallest SDs and MEs among all four methods, and the ALD-Mix procedure performs the second best in terms of SD and ME. In the presence of outliers and heavy tails in Cases 2 and 3, the performance of the proposed methods are also better than that of the others in terms of SD and ME. Particularly, the LS method fails in Case 4 in which the error distribution has infinite variance, while the ALD-Mix and LAD methods still work very well in this case.

In practice, the SEs of the ALD-Mix and CQR estimators can be estimated through the bootstrap. To demonstrate the behavior of the Bootstrap SE in comparison with the sample SD of the estimated β , Table 2 reports SEs and SDs of the ALD-Mix and CQR estimates based on 500 simulation replications with 500 bootstrap resamples each and sample size $n = 200$. It can be seen that the bootstrap SEs are very close to the corresponding empirical SDs for all six cases.

Next we document the asymptotic normality of the proposed estimators. For the sake of brevity, we consider the ALD-Mix estimator only. Based on 500 replications, Figure 1 presents normal qqplots of the standardized estimates $(\hat{\beta}_r - \beta_r)/sd(\hat{\beta}_r)$, $r = 1, 2, 3$, for Case 1 with $n = 100$ and $n = 200$ in the top and bottom panels, respectively. It can be seen that there is satisfactory

Table 2: Results of the ALD-Mix and CQR estimates and comparison between their estimated standard errors (SE) via the Bootstrap method and empirical standard derivations (SD) across 500 simulated samples for all 6 cases, where sample size $n = 200$ and the true values of $(\beta_1, \beta_2, \beta_3)$ are $(3, 1.5, 2)$.

Model	β_1			β_2			β_3		
	Estimate	SE	SD	Estimate	SE	SD	Estimate	SE	SD
ALD-Mix									
Case 1	3.0079	0.1362	0.1454	1.4907	0.1561	0.1659	2.0048	0.1381	0.1413
Case 2	2.9990	0.0925	0.0951	1.5008	0.1097	0.1049	1.9941	0.0950	0.0926
Case 3	3.0104	0.1072	0.1086	1.5012	0.1156	0.1105	1.9944	0.1007	0.1071
Case 4	3.0057	0.1305	0.1382	1.4953	0.1581	0.1569	2.0069	0.1417	0.1426
Case 5	2.9941	0.1279	0.1251	1.5078	0.1308	0.1349	2.0034	0.1293	0.1277
Case 6	2.9982	0.1363	0.1389	1.4968	0.1634	0.1581	2.0053	0.1294	0.1384
CQR									
Case 1	3.0042	0.1311	0.1441	1.4896	0.1607	0.1660	2.0068	0.1471	0.1439
Case 2	2.9978	0.0923	0.0904	1.5014	0.1062	0.0996	1.9945	0.0847	0.0882
Case 3	3.0121	0.1067	0.1041	1.4972	0.1084	0.1083	1.9947	0.1072	0.1023
Case 4	3.0032	0.1538	0.1505	1.4926	0.1637	0.1693	2.0037	0.1472	0.1558
Case 5	2.9931	0.1193	0.1191	1.5076	0.1301	0.1307	2.0052	0.1267	0.1205
Case 6	2.9972	0.1392	0.1367	1.4981	0.1516	0.1552	2.0007	0.1364	0.1343

agreement between the empirical distribution of the standardized ALD-Mix estimates and the distribution $N(0, 1)$, and the difference between these two distributions becomes negligible as sample size n increases to 200. Similar findings can be obtained for the other cases. To save space, we present only the QQ-plots with $n = 200$ for Cases 2-6 in Figures 2

To further illustrate the nice performance of the proposed mixture of ALDs in fitting the data, we compare the estimated density under the fitted model with the nonparametric kernel density estimate of the error term. Based on a randomly generated sample $\{y_i, x_i\}_{i=1}^n$ from each case, we estimate the density of the error term under the fitted model by $g(\omega) = \sum_{k=1}^K \hat{\pi}_k f(\omega; \hat{b}_k, \hat{\sigma}/\psi_k, \tau_k)$, and compute the corresponding kernel density estimate from observed errors $y_i - x_i^T \beta_0$ using R function *density()*, where $\{\hat{\pi}_k\}_{k=1}^K$, $\{\hat{b}_k\}_{k=1}^K$ and $\hat{\sigma}$ are the ALD-Mix estimates of parameters and β_0 is the true value of parameter β . Figures 3 and 4 depict both estimated density functions for all six cases when sample size $n = 100$ and $n = 200$, respectively. The estimated density curves under the fitted mixture of ALDs closely matches the kernel density curves in all cases, and a significant

Figure 1: Normal QQ-plots of the standardized ALD-Mix estimates (corresponding to $(\hat{\beta}_1 - \beta_1)/\text{sd}(\hat{\beta}_1)$, $(\hat{\beta}_2 - \beta_2)/\text{sd}(\hat{\beta}_2)$ and $(\hat{\beta}_3 - \beta_3)/\text{sd}(\hat{\beta}_3)$ in columns 1-3 from left to right, respectively) for Case 1 when $n = 100$ (Top row) and $n = 200$ (Bottom row).

improvement is obtained as the sample size increases.

In addition, we report the computational times of the proposed estimation procedures for the mixture of ALDs and the CQR. To quantify the speed gains afforded by the proposed algorithms, Table 3 summarizes the average and range of computing times in seconds over 500 replications in all 6 cases, where the sample size $n = 200$. All computations were performed on a laptop with a 3.40GHz Intel Core i7-6700 CPU processor. The computational times for both estimation procedures are reasonably short, especially for the CQR with solutions delivered in a few seconds for each case.

Finally, we turn our attention to the performance of the ABIC approach for selecting the value of K . To evaluate the behaviors of the ABIC in selecting K , we conduct simulations, in which data are generated from the same working model (36) but the error term ϵ_i follows the K -component mixture of the ALDs with equal proportions $\pi_k = 1/K$, i.e. $\epsilon_i \sim \sum_{k=1}^K \frac{1}{K} \text{ALD}(0, 1/\psi_k, \tau_k)$, where $\psi_k^2 = \kappa_k^2 + \nu_k^2$ with $\kappa_k = \frac{1-2\tau_k}{\tau_k(1-\tau_k)}$, $\nu_k^2 = \frac{2}{\tau_k(1-\tau_k)}$ and the asymmetric parameter $\tau_k = k/(K+1)$. First, nine scenarios are considered with the true value of K as $2, \dots, 10$, respectively. We conduct 500 simulation runs for each scenario with the sample size $n = 200$ and implement Algorithm 1

Figure 2: Normal QQ-plots of the standardized ALD-Mix estimates (corresponding to $(\hat{\beta}_1 - \beta_1)/\text{sd}(\hat{\beta}_1)$, $(\hat{\beta}_2 - \beta_2)/\text{sd}(\hat{\beta}_2)$ and $(\hat{\beta}_3 - \beta_3)/\text{sd}(\hat{\beta}_3)$ in columns 1-3 from left to right, respectively) for Cases 2-6 (corresponding to rows 1-5 from the bottom to the top, respectively) when $n = 200$.

Figure 3: Comparison between the kernel estimate of the error distribution (Kernel) and the estimated mixture of asymmetric Laplace distributions (Estimated ALD) based on a random sample of size $n = 100$ over six different cases.

Figure 4: Comparison between the kernel estimate of the error distribution (Kernel) and the estimated mixture of asymmetric Laplace distributions (Estimated ALD) based on a random sample of size $n = 200$ over six different cases.

Table 3: Average and range of computing times in seconds based on 500 replications for the proposed ALD-Mix and CQR estimation procedures in Cases 1-6, using a laptop with an Intel Core i7-6700 3.40GHz CPU processor.

Method		Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
ALD-Mix	Mean	20.713	20.201	12.037	3.282	11.003	17.015
	Min	18.113	17.801	3.957	0.008	4.627	4.857
	Max	23.519	22.192	20.203	20.554	20.122	20.733
CQR	Mean	4.304	3.044	3.141	3.744	4.206	2.817
	Min	2.620	1.894	0.003	0.003	0.230	0.670
	Max	7.006	5.066	7.228	7.538	8.798	6.455

for parameter estimation and the ABIC for determining K in each simulation run. We use four measures of the performance: the average of selected K , median of selected K and correct selection and false discovery rates (CSR and FDR), which are given by

$$\text{CSR} = \frac{\sum_{j=1}^{500} I(\hat{K}_j = K)}{500} \quad \text{and} \quad \text{FDR} = \frac{\sum_{j=1}^{500} |\hat{K}_j - K|}{\sum_{j=1}^{500} \hat{K}_j},$$

respectively, where $\hat{K}_j$ is the value of K selected in the j^{th} simulation run, $j = 1, \dots, 500$. Results presented in Table 4 show that the ABIC approach performs well in selecting the correct number of components especially when $K > 4$. When $K \leq 4$, the ABIC incurs some loss in CSR but controls FDR effectively.

Table 4: Simulation results of the ABIC based on 500 replications in terms of average selected K ($\text{Ave}\hat{K}$), standard derivation of selected K (SD), median selected K ($\text{Med}\hat{K}$), median absolute deviation of selected K (MAD), correct selection rate (CSR) and false discovery rate (FDR).

	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$	$K = 9$	$K = 10$
$\text{Ave}\hat{K}$	1.9840	2.9700	4.1260	4.9660	6.0060	7.0360	8.0080	8.9440	9.8680
(SD)	(0.7880)	(0.8477)	(0.8739)	(0.4704)	(0.4543)	(0.3937)	(0.4652)	(0.7759)	(0.6867)
$\text{Med}\hat{K}$	2.0000	3.0000	4.0000	5.0000	6.0000	7.0000	8.0000	9.0000	10.0000
(MAD)	(0.7413)	(1.4826)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
CSR	0.5000	0.4860	0.5280	0.9160	0.9200	0.9320	0.8920	0.9060	0.8480
FDR	0.2700	0.1940	0.1435	0.0252	0.0203	0.0137	0.0180	0.0249	0.0264

For two scenarios with the true value of $K = 5$ and 7, respectively, Figure 5 shows a comparison between the ABIC and conventional BIC curves obtained from a randomly selected simulation run.

The ABIC approach can effectively identify the value of K that minimizes the curve, while the conventional BIC fails. Based on our numerical experience, the BIC is nearly linear with K in almost all simulation runs.

Figure 5: The proposed ABIC (Left panel) and the conventional BIC (Right panel) curves based on one simulated sample when the error terms follow a 7-component (Top panel) and 5-component (Bottom panel) mixture of ALDs, respectively.

5 Real Data Analysis

To demonstrate the practical application of the proposed method, we revisit a Bardet-Biedl syndrome gene expression data set described by Scheetz et al. (2006) in a study of rat eye disease. The data set consists of gene expression levels of 18,975 genes obtained from $n = 120$ rat samples after the two preprocessing steps. The main objective of the analysis is to find relevant genes that are correlated with gene *TRIM32* (probe ID 1289163_at), a gene known to cause the eye disease

Bardet-Biedl syndrome. Wang et al. (2012) has verified the significant heterogeneity in these data and suggested regression analysis for conditional quantiles.

Regression analysis fails to be applied directly to such ultra-high dimensional data due to various challenges of computational expediency, statistical accuracy and algorithmic stability. Fan and Lv (2008) proposed to first use a fast and rough method of reducing the dimensionality to a moderate size (usually below the sample size), and then a more sophisticated technique, such as a penalized likelihood method based on the LASSO or SCAD penalty, can be applied to perform variable selection and parameter estimation simultaneously in the regression context. Following this idea, in the screening step we first screen down from 18,975 genes to 3,000 genes, which have the largest variances in gene expression levels. Then we compute the marginal correlation coefficients between each of these 3,000 genes and the gene *TRIM32*, and select the top 200 genes with the largest absolute correlation coefficients. In the step of variable selection, we utilize the regularization path of the adaptive Lasso developed by Huang et al. (2008) and select 14 important genes (variables) as a result. The important variables here refer to those variables whose regression coefficients differ from zero, while most regression coefficients are often assumed to be zeros in sparse modeling. The aforementioned procedure for reducing the number of candidate predictors has also been used by Kim et al. (2008) for regression models and Wang et al. (2012) for quantile regression. The selected 14 genes are given in the first column of Table 5.

Now we perform regression analysis of *TRIM32* (Y) on these selected 14 genes using the proposed method. The number of components K is determined by the ABIC approach given in Section 3.4. We present a plot of the ABIC values versus different values of K in Figure 6, which shows that the lowest value of the ABIC are attained at $K = 6$. Results obtained from the ALD-Mix and CQR estimation procedures with $K = 6$ are reported in Table 5 in comparison with those of the LS and LAD methods. The numbers in the parentheses coupled with each estimate in Table 5 are the standard errors estimated using the bootstrap method based on 500 bootstrap samples. To assess the quality of each estimation method, we adopt “leave-one-out” cross-validation. To be more specific, we conduct the estimation procedure based on 119 samples with i th sample dropped, and calculate the prediction errors $(Y_i - \hat{Y}_i)^2$ for $i = 1, \dots, 120$. The last row in Table 5 presents the average prediction error for all 120 samples.

Our analysis based on all the ALD-Mix, CQR, LS and LAD methods reveals that probes 1383996_at and 1386683_at are significant at 0.05 level. This implies that both probes have stronger effects on gene *TRIM32* than the other selected probes. Note that the probe 1383996_at was

confirmed to be important across different quantiles by Wang et al. (2012) using penalized quantile regression, while the probe 1386683_at was not detected in their work. In addition, the probe 1383110_at is found to be significant by the LAD method only, consistent with Wang et al's finding that it might be important only for a certain quantile. In terms of prediction error, the ALD-Mix estimator outperforms the other three estimators, followed by the CQR estimator which shows a slight improvement over the LS and LAD. Although the gains made in PE by the proposed method seem not "substantial", it is not surprising as far as the robust estimation is concerned in the study. As a result of an efficient and robust estimation procedure, our ALD-Mix estimator has the smallest PE but sometimes additional biases which may be a relatively good price paid in return for a greater reduction in SE.

Table 5: Analysis of the gene expression data: estimated coefficients, prediction errors and their corresponding estimated standard errors (in parentheses) for regression of TRIM32 on the selected 14 genes, where the number of components $K = 6$ in the ALD-Mix and CQR.

Probe ID	ALD-Mix	CQR	LS	LAD
1369353_at	-0.0161 (0.0200)	-0.0138 (0.0195)	-0.0125 (0.0211)	-0.0205 (0.0207)
1370429_at	0.0023 (0.0201)	0.0009 (0.0211)	-0.0057 (0.0223)	-0.0043 (0.0244)
1374106_at	0.0116 (0.0208)	0.0069 (0.0214)	0.0047 (0.0239)	-0.0008 (0.0244)
1384204_at	0.0061 (0.0123)	0.0066 (0.0128)	0.0020 (0.0145)	-0.0051 (0.0153)
1389584_at	0.0130 (0.0167)	0.0140 (0.0180)	0.0017 (0.0212)	0.0240 (0.0242)
1379495_at	0.0124 (0.0124)	0.0149 (0.0127)	0.0177 (0.0164)	0.0138 (0.0157)
1379971_at	0.0213 (0.0201)	0.0219 (0.0200)	0.0153 (0.0217)	0.0229 (0.0229)
1381787_at	-0.0155 (0.0115)	-0.0154 (0.0119)	-0.0115 (0.0142)	-0.0101 (0.0146)
1382835_at	0.0226 (0.0213)	0.0250 (0.0212)	0.0366 (0.0232)	0.0275 (0.0224)
1383110_at	0.0082 (0.0122)	0.0157 (0.0134)	0.0211 (0.0150)	0.0352 (0.0161)
1383673_at	0.0113 (0.0153)	0.0099 (0.0168)	0.0131 (0.0185)	0.0115 (0.0222)
1383749_at	-0.0260 (0.0148)	-0.0213 (0.0149)	-0.0230 (0.0164)	-0.0084 (0.0182)
1383996_at	0.0315 (0.0091)	0.0288 (0.0095)	0.0324 (0.0113)	0.0274 (0.0127)
1386683_at	-0.0213 (0.0100)	-0.0205 (0.0104)	-0.0332 (0.0121)	-0.0245 (0.0129)
Prediction errors	0.0074	0.0077	0.0081	0.0078

Figure 6: The plot of $ABIC$ values versus different values of K in the analysis of the gene expression data.

6 Discussion

Based on the asymmetric Laplace distribution, we have proposed a new mixture regression model to account for heterogeneity among data and thus provide a robust and efficient modeling approach, which includes the CQR model as a special case. We have developed an estimation procedure implemented through a two-layer EM algorithm, which includes two nested EM algorithms owing to the facts of a mixture of ALDs assumed for the error term and a normal-variance-mean mixture formulation of the ALD. Notably, the estimates obtained from the iteration of the proposed algorithm can be viewed as weighted composite quantile estimates with data-driven weights adaptively decided by the observed data and the current estimates, leading to more efficient estimates. In this viewpoint, the proposed method may provide a new perspective for the weighted composite quantile regression in Jiang et al. (2012). The two-layer EM algorithm can be directly applied to the composite quantile regression, providing an alternative efficient algorithm. In comparison with the existing linear programming based implementation given by Zou and Yuan (2008) for the CQR, the proposed method is conceptually straightforward and numerically stable. Based on our experience in numerical studies and the proposed algorithm converges rapidly and easy-to-program in R as it produces explicit estimation formulas in each iteration.

In a recent work of Tian et al. (2016), an EM algorithm is also developed for the CQR estimation. The major difference between our work and theirs resides in the scope of study. Our ALD mixture

regression model provides a more general and flexible linear regression framework to account for heterogeneity among data and includes the CQR as a special case, while their work focuses on the CQR only. Although both works performing the CQR estimation via the EM algorithm, they used the composite ALD with auxiliary scale vector $\sigma = (\sigma_1, \dots, \sigma_K)^T$, while the proposed algorithm 2 is based on the location parameter. The heuristic information criterion, ABIC, proposed in this paper provides a more effective approach for the selection of the number of components K in comparison with Tian et al. (2016)'s AIC or BIC -type criterion, which tends to choose small values of K due to a too heavy penalty imposed. However, more detailed properties of the proposed ABIC deserve further investigation from both theoretical and practical perspectives.

In this paper, the standard errors of the proposed estimators are computed via the bootstrap. There are some existing methods for estimating the standard error of an estimator obtained from the EM algorithm. For example, Louis (1982) developed a formula for computing the observed information matrix in terms of the complete and missing information matrices, while the calculation of the missing information involves the conditional expectation of the outer product of the complete-data score vector; Jamshidian and Jennrich (2000) proposed to use numerical differentiation to yield the Hessian matrix. Note that our proposed algorithm involves two EM algorithms with one nested the other, which complicates the problem so that these existing methods cannot be readily extended and some further study is expected in the future.

Acknowledgement

This research is supported by the Singapore MOE AcRF (MOE2013-T2-2-118, Tier 1 RG30/12) and the National Natural Science Foundation (Grant No. 71420107025, 11501586). The authors would like to thank the Associate Editor and two anonymous reviewers for their valuable comments and suggestions on the earlier version of this paper.

Appendix 1: Detailed Calculation for $\omega_{ik}^{(m)}$

In this section, we show the calculation details of $\omega_{ik}^{(m)}$ in (25) of subsection 3.2. In fact, it is easy to calculate from (23) that the conditional distribution of z_i is proportional to

$$z_i^{-1/2} \exp \left[-\frac{1}{2} \left\{ \varsigma_{ik}^2 z_i^{-1} + \gamma_k^2 z_i \right\} \right], \quad (37)$$

where $\varsigma_{ik}^2 = (Y_i - X_i^T \beta - b_k)^2 / \nu_k^2$ and $\gamma_k^2 = 2 + \kappa_k^2 / \nu_k^2$. Note that (37) is the kernel of a generalized inverse Gaussian (GIG) distribution, thus

$$[z_i | \mathcal{O}_i, k, \beta, b_k] \sim \text{GIG}\left(\frac{1}{2}, \varsigma_{ik}, \gamma_k\right).$$

For the general $\text{GIG}(u; v, \varsigma, \gamma)$ with $u > 0$, $-\infty < v < \infty$, $\varsigma > 0$, and $\gamma > 0$, Karlis (2002) showed that the moments around the original of the $\text{GIG}(u; v, \varsigma, \gamma)$ distribution are given by

$$E(z^r) = \left(\frac{\varsigma}{\gamma}\right)^r \frac{K_{v+r}(\varsigma\gamma)}{K_v(\varsigma\gamma)},$$

where $K_v(\cdot)$ is a modified Bessel function of the third kind (Karlis, 2002; Kozumi and Kobayashi, 2011). For $v = 1/2$ in our setting, it holds that

$$E(z^{-1}) = \left(\frac{\varsigma}{\gamma}\right)^{-1} \frac{K_{-1/2}(\varsigma\gamma)}{K_{1/2}(\varsigma\gamma)} = \frac{\gamma}{\varsigma}, \quad (38)$$

where the last equality holds because $K_v(\cdot) = K_{-v}(\cdot)$ (Abramowitz and Stegun, 1972). Substitute $\kappa_k = \frac{1-2\tau_k}{\tau_k(1-\tau_k)}$, $\nu_k^2 = \frac{2}{\tau_k(1-\tau_k)}$, $\varsigma_{ik} = |Y_i - X_i^T \beta - b_k| / \nu_k$ and $\gamma_k = (2 + \kappa_k^2 / \nu_k^2)^{1/2}$ into (38), it follows that

$$E\left(z_i^{-1} | \mathcal{O}_i, k; \beta^{(m)}, b_k^{(m)}\right) = \left\{ \tau_k(1 - \tau_k) |Y_i - X_i^T \beta^{(m)} - b_k^{(m)}| \right\}^{-1}$$

as given in (25).

Appendix 2: Detailed Derivation of Algorithm 2

Recall that the pdf of $\text{AID}(\mu, \sigma, \tau)$ is

$$f(t; \mu, \sigma, \tau) = \frac{\tau(1 - \tau)}{\sigma} \exp \left\{ -\rho_\tau \left(\frac{t - \mu}{\sigma} \right) \right\}.$$

The CQR method defined in (7) can be rewritten as

$$(\hat{b}_1, \dots, \hat{b}_K; \hat{\beta}^{\text{CQR}}) = \arg \max_{b_1, \dots, b_K; \beta} \sum_{k=1}^K \sum_{i=1}^n \log f(Y_i - X_i^T \beta - b_k; 0, 1, \tau_k),$$

where $\log f(Y_i - X_i^T \beta - b_k; 0, 1, \tau_k)$ can be viewed as the log-likelihood function of the following model

$$Y_i = X_i^T \beta + b_k + \tilde{\varepsilon}_{ik}, \quad \tilde{\varepsilon}_{ik} \sim \text{ALD}(0, 1, \tau_k),$$

for $i = 1, \dots, n$ and $k = 1, \dots, K$. Similar to Section 3.2.2, let z_i be the latent variable independently and identically following an exponential distribution with mean 1, and u_i be independently distributed as $N(0, z_i)$ given z_i . Then the model above is equivalent to

$$Y_i = X_i^T \beta + b_k + \kappa_k z_i + \nu_k u_i,$$

where $\kappa_k = \frac{1-2\tau_k}{\tau_k(1-\tau_k)}$ and $\nu_k^2 = \frac{2}{\tau_k(1-\tau_k)}$, and the complete log-likelihood function based on observations $\mathcal{O}_i = (X_i, Y_i)$ and z_i is obtained by

$$l_{ci}(\boldsymbol{\beta}, b_k; \mathcal{O}_i, z_i) \propto -\frac{(Y_i - X_i^T \boldsymbol{\beta} - b_k)^2}{2\nu_k^2} z_i^{-1} + \frac{\kappa_k}{\nu_k^2} (Y_i - X_i^T \boldsymbol{\beta} - b_k).$$

Given data $\mathcal{D} = \{\mathcal{O}, z\}$ with $z = (z_i, i = 1, \dots, n)$, we have

$$l_c(\boldsymbol{\beta}, \mathbf{b}; \mathcal{D}) = \sum_{i=1}^n \sum_{k=1}^K -\frac{(Y_i - X_i^T \boldsymbol{\beta} - b_k)^2}{2\nu_k^2} z_i^{-1} + \frac{\kappa_k}{\nu_k^2} (Y_i - X_i^T \boldsymbol{\beta} - b_k).$$

When using the EM algorithm, the E-step calculates the conditional expectation $E(l_c(\boldsymbol{\beta}, \mathbf{b}; \mathcal{D}) | \mathcal{O}, \boldsymbol{\beta}^{(m)}, \mathbf{b}^{(m)})$.

Note that

$$\omega_{ik}^{(m)} = E(z_i^{-1} | \mathcal{O}_i, k; \boldsymbol{\beta}^{(m)}, b_k^{(m)}) = \frac{1}{\tau_k(1-\tau_k) |Y_i - X_i^T \boldsymbol{\beta}^{(m)} - b_k^{(m)}|}.$$

It follows that

$$\begin{aligned} \tilde{l}_c(\boldsymbol{\beta}, \mathbf{b}) &= E(l_c(\boldsymbol{\beta}, \mathbf{b}; \mathcal{D}) | \mathcal{O}, \boldsymbol{\beta}^{(m)}, \mathbf{b}^{(m)}) = \sum_{i=1}^n \sum_{k=1}^K -\omega_{ik}^{(m)} \frac{(Y_i - X_i^T \boldsymbol{\beta} - b_k)^2}{2\nu_k^2} + \frac{\kappa_k}{\nu_k^2} (Y_i - X_i^T \boldsymbol{\beta} - b_k) \\ &= \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K \left[-\tilde{\omega}_{ik}^{(m)} \frac{(Y_i - X_i^T \boldsymbol{\beta} - b_k)^2}{2} + (1 - 2\tau_k)(Y_i - X_i^T \boldsymbol{\beta} - b_k) \right], \end{aligned}$$

where the redefined weights $\tilde{\omega}_{ik}^{(m)}$ is the same as that in (27). In the M-step, given the current estimate of $\mathbf{b}$, $\mathbf{b}^{(m)}$ differentiating $\tilde{l}_c(\boldsymbol{\beta}, \mathbf{b}^{(m)})$ with respect to $\boldsymbol{\beta}$ leads to

$$\sum_{i=1}^n \sum_{k=1}^K X_i \left\{ \tilde{\omega}_{ik}^{(m)} (Y_i - X_i^T \boldsymbol{\beta} - b_k^{(m)}) - (1 - 2\tau_k) \right\} = 0,$$

and consequently the updating formulae for $\boldsymbol{\beta}$ is obtained by

$$\boldsymbol{\beta}^{(m+1)} = \left(\sum_{i=1}^n \sum_{k=1}^K \tilde{\omega}_{ik}^{(m)} X_i X_i^T \right)^{-1} \left[\sum_{i=1}^n \sum_{k=1}^K X_i \left\{ \tilde{\omega}_{ik}^{(m)} (Y_i - b_k^{(m)}) - (1 - 2\tau_k) \right\} \right].$$

Let $\mathbf{1}_n$ be a $n \times 1$ vector with all elements equal to 1, $\mathbf{C} = \mathbf{1} - 2\boldsymbol{\tau}$ with $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)^T$, $\boldsymbol{\psi}^2 = \boldsymbol{\kappa}^2 + \boldsymbol{\nu}^2$ with $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_K)^T$ and $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K)^T$. Furthermore, $\tilde{\Omega}^{(m)} = (\tilde{\omega}_{ik}^{(m)})_{n \times K}$, $\tilde{W}^{(m)} = \text{diag}\{\sum_{k=1}^K \tilde{\omega}_{1k}^{(m)}, \dots, \sum_{k=1}^K \tilde{\omega}_{nk}^{(m)}\} = \text{diag}\{\tilde{\Omega}^{(m)} \mathbf{1}_K\}$, we have

$$\boldsymbol{\beta}^{(m+1)} = \left(\mathbf{X}^T \tilde{W}^{(m)} \mathbf{X} \right)^{-1} \mathbf{X}^T \left\{ \tilde{W}^{(m)} \mathbf{Y} - \tilde{\Omega}^{(m)} \mathbf{b}^{(m)} - (\mathbf{1}_K^T \mathbf{C}) \mathbf{1}_n \right\}.$$

Now we update $\mathbf{b}$ with $\boldsymbol{\beta}$ fixed at $\boldsymbol{\beta}^{(m+1)}$. By differentiating $\tilde{l}_c(\boldsymbol{\beta}^{(m+1)}, \mathbf{b})$ with respect to b_k , we have

$$\sum_{i=1}^n \left\{ \tilde{\omega}_{ik}^{(m)} (Y_i - X_i^T \boldsymbol{\beta}^{(m+1)} - b_k) - (1 - \tau_k) \right\} = 0, \quad k = 1, 2, \dots, K.$$

Thus, the updating formulae for b_k can be obtained by

$$b_k^{(m+1)} = \left(\sum_{i=1}^n \tilde{\omega}_{ik}^{(m)} \right)^{-1} \sum_{i=1}^n \left[\tilde{\omega}_{ik}^{(m)} (Y_i - X_i^T \beta^{(m+1)}) - (1 - \tau_k) \right], \quad k = 1, 2, \dots, K.$$

Let $\tilde{\Phi}^{(m)} = \text{diag}\{\sum_{i=1}^n \tilde{\omega}_{i1}^{(m)}, \dots, \sum_{i=1}^n \tilde{\omega}_{iK}^{(m)}\} = \text{diag}\{\tilde{\Omega}^{(m)T} \mathbf{1}_n\}$. The updating formulae for $\mathbf{b}$ becomes

$$\mathbf{b}^{(m+1)} = \{\tilde{\Phi}^{(m)}\}^{-1} \left\{ \tilde{\Omega}^{(m)T} (\mathbf{Y} - \mathbf{X}^T \beta^{(m+1)}) - n\mathbf{C} \right\}.$$

We repeat the E-step and M-step are repeated alternately until convergence. Hence Algorithm 2 is obtained.

References

- M. Abramowitz and I. A. Stegun. Handbook of mathematical functions: with formulas, graphs, and mathematical tables. Number 55. Courier Dover Publications, 1972.
- F. Bartolucci and L. Scaccia. The use of mixtures for dealing with non-normal regression errors. Computational Statistics & Data analysis, 48(4):821–834, 2005.
- J. Bradic, J. Fan, and W. Wang. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3):325–349, 2011.
- J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360, 2001.
- J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5):849–911, 2008.
- G. Galimberti and G. Soffritti. A multivariate linear regression analysis using finite mixtures of t distributions. Computational Statistics & Data Analysis, 71:138–150, 2014.
- M. Geraci and M. Bottai. Quantile regression for longitudinal data using the asymmetric Laplace distribution. Biostatistics, 8(1):140–154, 2007.
- M. Geraci and M. Bottai. Linear quantile mixed models. Statistics and Computing, 24(3):461–479, 2014.

- J. Huang, S. Ma, and C. H. Zhang. Adaptive Lasso for sparse high-dimensional regression models. Statistica Sinica, 18(4):1603, 2008.
- M. Jamshidian and R. I. Jennrich. Standard errors for em estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62(2):257–270, 2000.
- X. Jiang, J. Jiang, and X. Song. Oracle model selection for nonlinear models based on weighted composite quantile regression. Statistica Sinica, 22(4):1479–1506, 2012.
- D. Karlis. An EM type algorithm for maximum likelihood estimation of the normal-inverse Gaussian distribution. Statistics & Probability letters, 57(1):43–52, 2002.
- Y. Kim, H. Choi, and H.-S. Oh. Smoothly clipped absolute deviation on high dimensions. Journal of the American Statistical Association, 103(484):1665–1673, 2008.
- R. Koenker. Quantile regression. Number 38. Cambridge university press, 2005.
- R. Koenker and G. Bassett Jr. Regression quantiles. Econometrica: Journal of the Econometric Society, pages 33–50, 1978.
- R. Koenker and J. A. Machado. Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association, 94(448):1296–1310, 1999.
- S. Kotz, T. Kozubowski, and K. Podgorski. The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance. Springer Science & Business Media, 2012.
- H. Kozumi and G. Kobayashi. Gibbs sampling methods for bayesian quantile regression. Journal of Statistical Computation and Simulation, 81(11):1565–1578, 2011.
- T. A. Louis. Finding the observed information matrix when using the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), pages 226–233, 1982.
- R. F. Phillips. Least absolute deviations estimation via the EM algorithm. Statistics and Computing, 12(3):281–285, 2002.
- T. E. Scheetz, K. Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, T. L. Casavant, V. C. Sheffield, and E. M. Stone. Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proceedings of the National Academy of Sciences, 103(39):14429–14434, 2006.

- G. Soffritti and G. Galimberti. Multivariate linear regression with non-normal errors: a solution based on mixture models. Statistics and Computing, 21(4):523–536, 2011.
- W. Song, W. Yao, and Y. Xing. Robust mixture regression model fitting by Laplace distribution. Computational Statistics & Data Analysis, 71:128–137, 2014.
- Y. Tian, Q. Zhu, and M. Tian. Estimation of linear composite quantile regression using EM algorithm. Statistics & Probability Letters, 2016.
- R. Tibshirani. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.
- L. Wang, Y. Wu, and R. Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. Journal of the American Statistical Association, 107(497):214–222, 2012.
- K. Yu and R. A. Moyeed. Bayesian quantile regression. Statistics & Probability Letters, 54(4):437–447, 2001.
- K. Yu and J. Zhang. A three-parameter asymmetric Laplace distribution and its extension. Communications in Statistics-Theory and Methods, 34(9-10):1867–1879, 2005.
- H. Zou and M. Yuan. Composite quantile regression and the oracle model selection theory. The Annals of Statistics, pages 1108–1126, 2008.