Robust Mixture Multivariate Linear Regression by Multivariate Laplace Distribution

Xiongya Li[a], Xiuqin Bai[b], Weixing Song[a,1,*]

[a] *Department of Statistics, Kansas State University, Manhattan, KS 66506*
[b] *Department of Mathematics, Eastern Washington University, Cheney, WA, 99004*

Abstract

Assuming that the error terms follow a multivariate Laplace distribution, we propose a robust estimation procedure for mixture of multivariate linear regression models in this paper. Using the fact that the multivariate Laplace distribution is a scale mixture of the multivariate standard normal distribution, an efficient EM algorithm is designed to implement the proposed robust estimation procedure. The performance of the proposed algorithm is thoroughly evaluated by some simulation and comparison studies.

Keywords: Finite mixtures, Multivariate linear regression, Robust Estimation, Multivariate Laplace Distribution, EM algorithm
2000 MSC: primary 62F35, secondary 62F10

1. Introduction

Finite mixture regression modeling is an efficient tool to investigate the relationship between a response variable and a set of predictors when the underlying population consists of several unknown latent homogeneous groups, and it has been already applied for more than a hundred years since Newcomb (1886). More real examples on finite mixture modelling can be found in Jiang and Tanner (1999), Böhning (2000), McLachlan and Peel (2004), Wedel and Kamakura (2012) and the references therein. Statistical inferences have been discussed extensively for finite mixture modeling when the normality is assumed for the regression error in each cluster. Due to the untractable likelihood function for normal mixture regression models, the unknown regression parameters are often estimated via the expectation and maximization (EM) algorithm. However, the unweighted least squares nature makes the maximum likelihood estimate (MLE) of the regression parameters susceptible of non-robustness to the outliers and the data with heavy tails. Because of its wide application in practice, how to design robust estimation procedures in the finite mixture regression models has attracted much attention from statisticians.

*Corresponding author
Email address: weixing@ksu.edu (Weixing Song)

Extensive research has been done for linear or mixture of linear regression models when the response variable is univariate. For examples, Neykov et al. (2007) proposed a trimmed likelihood estimator (TLE) to robustly estimate the mixtures and the breakdown points of the TLE for the mixture component parameters is also characterized; Replacing the least square criterion in the M step of EM algorithm designed for normal mixtures, Bai et al. (2012) achieved robustness using Tukey's bisquare and Huber's ψ-functions; A class of S-estimators were introduced in Bashir and Carter (2012) and Farcomeni and Greco (2015) which exhibit certain robustness and the parameter estimation is achieved via an expectation-conditional maximization algorithm. Inspired by Pell and McLachlan (2000), Yao et al. (2014) proposed a new robust estimation method for mixture of linear regression by assuming that the mixtures have t-distributions, the EM algorithm is made possible by the fact that t-distribution is a scale mixture of a normal distribution. Due to the selection of degrees of freedom, the procedure in Yao et al. (2014) requires relatively heavy computation although the choice of degrees of freedom provides certain adaptivity to the data. Realizing that the Laplace distribution is also a scale mixture of normal distribution, Song et al. (2014) proposed an alternative robust estimation procedure by assuming the random error has a Laplace distribution, which has a natural connection with the least absolute deviation (LAD) procedure, see Dielman (1984), Li and Arce (2004), and Dielman (2005) for more detail on LAD methodology.

Comparing to the relatively extensive discussion for the univariate response cases, there are fewer work having been done for the multivariate linear regressions. Lin (2010) designed a robust estimation procedure using the multivariate skewed t-distribution, which offers a great deal of flexibility that accommodates asymmetry and heavy tails simultaneously. Xian Wang et al. (2004) proposed a mixture of multivariate t-distribution to fit the multivariate continuous data with a large number of missing values. We haven't seen any work on developing robust estimation procedures for the multivariate linear regression with the multivariate Laplace distribution. We wish there is a multivariate version of Song et al. (2014)'s procedure which should perform equally well in the multivariate linear regression. This is the motivation of the research conducted in the current paper.

The paper is organized as follows. Section 2 introduces the mixture of multivariate linear regression models, and also the definition of the multivariate Laplace distribution, some essential properties of the multivariate Laplace distribution is also discussed. The EM algorithm will be developed in Section 3 for the mixture of multivariate linear regression models. Section 4 includes some simulation and comparison studies to evaluate the performance of the proposed methods.

2. Statistical Model and Multivariate Laplace Distribution

We begin with a brief introduction on the mixture of multivariate linear regression models, and a definition of multivariate Laplace distribution.

2.1. Mixture of Multivariate Linear Regression

Let G be a latent class variable such that given $G = j$, $j = 1, 2, \ldots, g$, $g \geq 1$, a p-dimensional response Y and a q-dimensional predictor X are in one of the following multivariate linear regression models

$$Y = \beta_j' X + \varepsilon_j, \tag{1}$$

where, for each j, β_j is a $q \times p$ unknown regression coefficient matrix, and ε_j is a p-dimensional random error. Assume ε_j's are independent of X and it is commonly assumed that the density functions f_j of ε_j's are members in a location-scale family with mean 0 and covariance Σ_j. If we further assume that $P(G = j) = \pi_j, j = 1, ..., g$. Then conditioning on X, the density function of Y is given by

$$f(y|x, \theta) = \sum_{j=1}^{g} \pi_j f_\varepsilon(y - \beta_j' x, 0, \Sigma_j), \tag{2}$$

where $\theta = \{\pi_1, \beta_1, \Sigma_1, ..., \pi_g, \beta_g, \Sigma_g\}$. The model (2) is the so called mixture multivariate regression models. The unknown parameters could be estimated by the maximum likelihood estimator (MLE), which maximizes the log-likelihood function (3) based on an independent sample $(X_i, Y_i), i = 1, ..., n$ from (2),

$$L_n(\theta) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{g} \pi_j f_\varepsilon(Y_i, \beta_j' X_i, \Sigma_j) \right]. \tag{3}$$

If $g = 1$, then the mixture linear regression model is simply a multivariate linear regression model.

The traditional maximum likelihood estimation procedure is based on the normality assumption. However, no explicit solution is available due to the untractable expression of (3), and EM algorithm thus developed to obtain its the maximizer. As we mentioned in Section 1, the MLE based on the normality assumption is sensitive to outliers or heavy-tailed error distribution, and we shall develop a robust estimation procedure by assuming that the error distributions are Laplacian.

2.2. Multivariate Laplace Distribution

There are multiple forms of definitions of the multivariate Laplace distribution. For example, the bivariate case was introduced by Ulrich and Chen (1987), and the first form in larger dimensions was discussed in Fang et al. (1990). Later, the multivariate Laplace was introduced as a special case of a multivariate Linnik distribution in Anderson (1992), and the multivariate power exponential distribution in Fernandez et al. (1995) and Ernst (1998). Portilla et al. (2003) presented multivariate Laplace distribution as a Gaussian scale mixture. Kotz et al. (2012) presented the multivariate Laplace distribution formally and thoroughly discussed its probability properties. The multivariate Laplace distribution

is an attractive alternative to the multivariate normal distribution due to its heavier tails. For its application in image and speech recognition, ocean engineering and finance, see Kotz et al. (2012). In this paper, we adopted the following definition from Eltoft et al. (2006).

Definition 1: *A p-dimensional random vector U is called to have a multivariate Laplace distribution, if its density function has the form of*

$$f_U(u) = \frac{2}{(2\pi)^{p/2}|\Sigma|^{1/2}} \left[\frac{Q(u;\mu,\Sigma)}{2}\right]^{\frac{1}{2}(1-\frac{p}{2})} K_{p/2-1}\left(\sqrt{2Q(u;\mu,\Sigma)}\right), \tag{4}$$

where $Q(u;\mu,\Sigma) = (u-\mu)'\Sigma^{-1}(u-\mu)$, $u \in \mathbb{R}^p$, $K_m(x)$ is the modified Bessel function of the second kind with order m. Denote $U \sim ML_p(\mu,\Sigma)$.

The modified Bessel function of the second kind is the solution to the modified Bessel differential equation, and sometimes it is also called the Basset function, the modified Bessel function of the third kind, or the Macdonald function. See Spanier and Oldham (1987), Samko et al. (1987) for more discussion on the Bessel functions. In fact, the multivariate Laplace distribution defined in Definition 1 is also a special case of the symmetric multivariate Bessel distribution defined in Fang et al. (1990). The following lemma provides some important probabilistic properties about the multivariate Laplace distribution.

Lemma 1. *Suppose a random vector U follows the multivariate Laplace distribution with the density function defined in (4), then*

(i) *$EU = \mu$ and $Cov(U) = \Sigma$;*

(ii) *The characteristic function of U is given by $\phi_U(t) = (1 + t'\Sigma t/2)^{-1}\exp(it'\mu)$ for $t \in \mathbb{R}^p$;*

(iii) *Let V be a random variable with density function $f_V(v) = e^{-v}I(v \geqslant 0)$, Z be a p-dimensional standard normal random vector, that is $N_p(0, I)$, V and Z are independent. Then $U = \sqrt{V}\Sigma^{1/2}Z \sim ML_p(0,\Sigma)$;*

(iv) *Assume V and U are defined as above. Then*

$$E\left(\frac{1}{V}\middle| U = u\right) = \sqrt{\frac{2}{u'\Sigma^{-1}u}}\frac{K_{-p/2}(\sqrt{2u'\Sigma^{-1}u})}{K_{1-p/2}(\sqrt{2u'\Sigma^{-1}u})}.$$

Lemma 1 is a summary of some main results in Eltoft et al. (2006). From (i) and the density function of the multivariate Laplace distribution, we can see that the Multivariate Laplace distribution is uniquely determined by its mean vector and covariance matrix. From (ii) we can see that the multivariate Laplace distribution defined by (4) indeed is a natural extension of univariate Laplace distribution. Similar to the univariate Laplace distribution, (iii) indicates that the multivariate Laplace distribution is a scale mixture of multivariate normal distribution, and this property makes it feasible to develop an efficient EM algorithm to implement the proposed robust estimation procedure. The property (iv) is emphasized here since it plays a crucial role in the E step of the developed EM algorithm.

3. Multivariate Laplace-type Estimation Procedure

In this section, we shall develop a robust regression procedure in the mixture of multivariate linear regression models by assuming that the regression errors follow multivariate Laplace distributions with mean vector 0 and possibly different covariance matrices. To be specific, assume that $\varepsilon_j \sim ML(0, \Sigma_j)$ and $\sum_{j=1}^{g} \pi_j = 1$, $g > 1$. Define $G_{ij} = 1$ if the i-th observation (X_i, Y_i) is from the j-th component, and 0 otherwise. Let (X_i, Y_i, G_{ij}), $i = 1, 2, ..., n$, $j = 1, 2, ..., g$ be a sample from the model (1). Once again, recall the notation $Q(u; \mu, \Sigma)$ in Definition 1, and we further denote $Q_{ij} = Q(Y_i; \beta_j' X_i, \Sigma_j)$ for the sake of convenience, the complete likelihood function $L(\theta)$ of $\theta = (\beta_1, ..., \beta_g, \Sigma_1, ..., \Sigma_g, \pi_1, ..., \pi_g)$ can be written as

$$\prod_{i=1}^{n} \prod_{j=1}^{g} \left\{ \left(\frac{2\pi_j}{(2\pi)^{p/2} |\Sigma_j|^{1/2}} \right) \left[\frac{Q_{ij}}{2} \right]^{1/2 - p/4} K_{p/2-1} \left(\sqrt{2Q_{ij}} \right) \right\}^{G_{ij}}.$$

Based on (iii) in Lemma 1, similar to the discussion for the case of $g = 1$, for each (X_i, Y_i), if we can further observe V_i, $i = 1, 2, .., n$, then the complete log-likelihood function of θ, the collection of all unknown parameters, will be given by

$$L(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{g} G_{ij} \log \pi_j - \frac{p}{2} \sum_{i=1}^{n} \sum_{j=1}^{g} G_{ij} \log(2\pi V_i) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{g} G_{ij} \log |\Sigma_j|$$

$$- \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{g} V_i^{-1} G_{ij} Q_{ij} - \sum_{i=1}^{n} \sum_{j=1}^{g} G_{ij} V_i.$$

With the initial values for $\theta^{(0)} = (\pi^{(0)}, \beta^{(0)}, \Sigma^{(0)})$, we have to calculate

$$E[L(\theta)|\theta^{(0)}, \mathbf{D}] = \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_{ij} \log \pi_j - \frac{p}{2} \sum_{i=1}^{n} \sum_{j=1}^{g} E[G_{ij} \log 2\pi V_i | \theta^{(0)}, \mathbf{D}]$$

$$- \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_{ij} \log |\Sigma_j| - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{g} E \left[\frac{G_{ij}}{V_i} Q_{ij} \middle| \theta^{(0)}, \mathbf{D} \right] - \sum_{i=1}^{n} \sum_{j=1}^{g} E(G_{ij} | \theta^{(0)}, \mathbf{D}),$$

where we use $\mathbf{D}$ to denote the complete data set for the sake of brevity. The above conditional expectation can be further written as

$$\sum_{i=1}^{n} \sum_{j=1}^{g} \tau_{ij} \log \pi_j - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_{ij} \log |\Sigma_j| - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{g} \tau_{ij} \delta_{ij} Q_{ij} + \mathbf{R_n}, \tag{5}$$

where $\tau_{ij} = E[G_{ij} | \theta^{(0)}, \mathbf{D}]$, $\delta_{ij} = E\left[V_i^{-1} \middle| \theta^{(0)}, \mathbf{D}, G_{ij} = 1 \right]$, and the reminder term $\mathbf{R_n}$ does not depend on the unknown parameters. Denote $Q_{ij}^{(0)} = Q(Y_i; X_i' \beta_j^{(0)}, \Sigma_j^{(0)})$. From Lemma 1, we know that

$$\delta_{ij} = \sqrt{2} K_{-p/2} \left(\sqrt{2Q_{ij}^{(0)}} \right) \Big/ \sqrt{Q_{ij}^{(0)}} K_{1-p/2} \left(\sqrt{2Q_{ij}^{(0)}} \right).$$

One can further show that, by applying Bayesian formula, $\tau_{ij} = \xi_{ij} / \sum_{l=1}^{g} \xi_{il}$, where

$$\xi_{ij} = \pi_j^{(0)} |\Sigma_j|^{-1/2} [Q_{ij}^{(0)}]^{1/2-p/4} K_{p/2-1}\left(\sqrt{2Q_{ij}^{(0)}}\right).$$

Based on the above discussion, the EM algorithm for estimating θ is as follows:

EM Algorithm:

(1) Choosing initial values for β, Σ, π, say $\beta^{(0)}, \Sigma^{(0)}, \pi^{(0)}$;
then at the $k+1$-th iteration,

(2) E-Step: Calculate $\tau_{ij}^{(k+1)}$, $\delta_{ij}^{(k+1)}$ from above equations with (0) replaced by (k);

(3) M-Step: Update β, Σ, π with

$$\pi_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ij}^{(k+1)}, \quad \beta_j^{(k+1)} = (\mathbf{X}'\mathbf{W}_j\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}_j\mathbf{Y}),$$

$$\Sigma_j^{(k+1)} = \frac{\mathbf{Y}'(\mathbf{W}_j - \mathbf{W}_j\mathbf{X}(\mathbf{X}'\mathbf{W}_j\mathbf{X})^{-1}\mathbf{X}\mathbf{W}_j)\mathbf{Y}}{\sum_{i=1}^{n}\tau_{ij}^{(k+1)}}.$$

where $\mathbf{W}_j = \mathrm{diag}(\tau_{1j}^{(k+1)}\delta_{1j}^{(k+1)}, \tau_{2j}^{(k+1)}\delta_{2j}^{(k+1)}, \ldots, \tau_{nj}^{(k+1)}\delta_{nj}^{(k+1)})$.

(4) Repeat (2) and (3) until certain convergence criterion is met.

The ascent property is a very important characteristic possessed by the EM algorithm in parametric models. It implies that after each iteration, the likelihood at the newly updated estimate is no less than the likelihoods at the previous estimates. The following theorem implies that the proposed EM algorithm dose possess this desired property.

Theorem 1. *Let $\theta^{(k)}$ denote the estimate of θ in the k-th iteration of the EM algorithm, then for any n, $L_n(\theta^{(k+1)}) \geq L_n(\theta^{(k)})$, where $L_n(\theta)$ is defined in (3).*

The main proof of Theorem 1 is similar to the proof of Theorem 3 in Huang et al. (2013). However, a nontrivial modification is needed to accommodate both latent variables G and V.

If we further assume that all Σ_j's are equal, then the common covariance matrix can be estimated by $\Sigma^{(k+1)} = n^{-1}\sum_{j=1}^{g} \mathbf{Y}'(\mathbf{W}_j - \mathbf{W}_j\mathbf{X}(\mathbf{X}'\mathbf{W}_j\mathbf{X})^{-1}\mathbf{X}\mathbf{W}_j)\mathbf{Y}$. One can easily check that if Y is one-dimensional, then all the formulae listed in the M-step of the above EM algorithm are exactly the same as in Song et al. (2014).

As we mentioned in Section 1, the robustness of the EM procedure developed in Song et al. (2014) is well aligned with the close connection between the MLE of the regression coefficients when the error term has a Laplace distribution and the LAD regression, this is also true for the EM procedure we developed above for the multivariate case. Also, from the M-step, we can see that the estimate of the regression coefficients β_j's indeed is a

weighted least squares estimate, and the factor $\delta_{ij}^{(k+1)}$ from the weights $w_{ij}^{(k+1)}$ depends on the $Q_{ij}^{(k)} = (Y_i - (\beta_j^{(k)})'X_i)'(\Sigma_j^{(k)})^{-1}(Y_i - (\beta_j^{(k)})'X_i)$ in a rather complicated way. However, for each i, j, $\delta_{ij}^{(k+1)}$ is indeed a decreasing function of $Q_{ij}^{(k)}$, which indicates that similar to the univariate response case discussed in Song et al. (2014), less weights will be received for those observations with larger residuals in the estimation procedure, which guarantees the robustness of the proposed EM algorithm. A rigorous proof of $\delta_{ij}^{(k+1)}$ being a strictly decreasing function of $Q_{ij}^{(k)}$ can be provided using the Property 12 in the appendix on Bessel functions from Kotz et al. (2012).

Note that when $Q_{ij} = 0$, δ_{ij} will be infinite. This creates some difficulties when we program since very big value of δ_{ij} would make the computation very unstable. This phenomenon is an important aspect, as one reviewer indicated, that characterizes many algorithms for fitting mixture models. For $g = 1$ and the univariate response case, Phillips (2002) noticed that this problem rarely arises, but this does occur often in our case. In fact, same phenomenon is also found in the t-procedure proposed in Yao et al. (2014). Similar to Song et al. (2014), in our simulation study, we adopt a hard threshold rule to control the effect of extremely small Q_{ij} values in each iteration step. Under this rule, $\delta_{ij}^{(k+1)}$ will be assigned a value of 10^6 if the corresponding Q_{ij} equals 0. To see the effects of different choices of the threshold values, we also tried other threshold values, such as $10^8, 10^{10}$, and all these choices produce similar results. Therefore, only the results for 10^6 are reported. Note that numerical instability could also occur if the weights are very small, to deal with this, we use the another hard threshold rule on the value of τ_{ij}, if $\tau_{ij}^{k+1} > 10^{-6}$, then τ_{ij}^{k+1} itself will be used for the next iteration; otherwise, 10^{-6} will be used as the weight for the next iteration. Same technique is used in Yao et al. (2014) and Song et al. (2014). Clearly, searching for more objective thresholding rules or thresholding-free algorithm is still an interesting research topic deserving more effort.

To conclude this section, we would like to point out that the proposed EM algorithm based on the multivariate Laplace distribution is robust against outliers along the y-direction, but not in the x-direction. Therefore, certain modification is needed to equip the proposed method with some robustness against the outliers in x-direction. Here we recommend a pre-screening method. That is, exclude the observations which is deemed to be an outlier in x-direction before applying the proposed EM algorithm. For this purpose, we first calculate the leverage value for each observation using the formula $h_{jj} = n^{-1} + (n-1)^{-1}MD_j$, where $MD_j = (X_j - \bar{X})'S^{-1}(X_j - \bar{X})$, $\bar{X}$, S are the sample mean and sample covariance matrix of X_j's, respectively. The j-th observation will be identified as a high leverage point if $h_{jj} > 2p/n$, where p is the dimension of X. Some robust estimation of the population mean and covariance matrix of X can be used instead of the sample mean and sample covariance. For example, the minimum covariance determinant (MCD) estimators developed in Rousseeuw (1999), and the Stahel-Donoho (SD) estimator from Stahel (1981) and Donoho (1982). More discussion on this matter can be found in Yao et al. (2014) and Song et al. (2014).

4. Simulation Studies

To evaluate the performance of the proposed robust estimation procedure, we conduct some simulation studies in this section. In the first simulation, a comparison study is made between the proposed method and the MCD-based robust multivariate regression procedure discussed in Rousseeuw et al. (2004). Note that this study is done only for the non-mixture case, due to the MCD-based robust estimation procedure does not have a clearly workable extension to the mixture cases. In the second simulation, a case of $g > 1$ will be considered. We shall compare the proposed method with other two methods, the traditional MLE assuming the error has a multivariate normal density and the robust mixture regression model based on the multivariate t distribution.

Simulation 1: Among many robust multivariate regression procedures, the one based on the MCD has been enjoyed great popularity since its introduction by Rousseeuw et al. (2004). To be specific, let $Z_i = (Y_i', X_i')'$, and $\mathbf{Z}_n = \{Z_1, \ldots, Z_n\}$. The MCD regression first looks for the subset $\{Z_{i_1}, \ldots, Z_{i_h}\}$ of size h of $\mathbf{Z}_n$ whose covariance matrix has the smallest determinant, then the usual least squared estimation procedure (mcdLSE) is applied to the selected subset to obtain the estimates for the regression coefficients. Common choices of h are $h \approx n/2$ or $3n/4$. To increase the efficiency, Rousseeuw et al. (2004) proposed three reweighted least squared estimation procedures by reweighting the location (mcdLoc), the regression (mcdReg), both the location and the regression (mcdLR). In the simulation, the data are generated from the multivariate regression models $Y = \beta' X + \varepsilon$, where β is a $q \times p = 4 \times 10$ matrix with entries randomly generated from a uniform distribution on $[0, 10]$, X follows a 4-dimensional multivariate normal distribution with mean 0 and identity covariance matrix. The regression errors ε are chosen from 6 different distributions: (a) the multivariate standard normal; (b) the multivariate Laplace distribution with identity covariance matrix; (c) the multivariate t distribution with degrees of freedom 1; (d) the multivariate t-distribution with degrees of freedom 3; (e) the normal mixture $0.95N(0, I) + 0.05N(0, 50I)$, and (f) a multivariate normal with 5% x-direction high leverage outliers, all x-values being 10 and all y-values 2.

Case (a) is often used to evaluate the efficiency of different estimation methods compared to the traditional MLE when error is exactly multivariate normally distributed and there are no outliers. Under case (b), the proposed estimation procedure will provide the MLE of unknown parameters, which, as in the first case, would serve as a baseline to evaluate the performance of other estimation procedures. Both case (c) and (d) are heavy tailed distributions and are often used in the literature to mimic the outlier situations. Case (e) would produce 5% low leverage outliers, and in case (f), 5% of the observations are replicated serving as the high leverage outliers, which will be used to check the robustness of estimation procedures against the high leverage outliers.

The sample size of $n = 200$ is used in the simulation study. For each case, the simulation is repeated 200 times. The average L_2-norm of the differences of the estimated β values from their true values are used as the criterion to evaluate the performance of

the proposed (L-EM) and the MCD based estimation procedures. Table 1 is a summary of the simulation results.

Table 1: Simulation results for $g = 1$

Error	Normal	mcdLSE	mcdLoc	mcdReg	mcdLR	L-EM
(a)	0.2060	0.2990	0.2424	0.5472	0.3102	0.2517
(b)	0.4097	0.4487	0.3920	0.6791	0.4457	0.3358
(c)	40868.81	0.8786	0.8227	1.0766	0.8419	0.8012
(d)	40869.43	1.2239	1.1637	1.4826	1.2001	1.1172
(e)	40870.17	1.5287	1.4184	2.0247	1.5314	1.4045
(f)	41888.27	1.8193	1.6604	2.5675	1.8476	123.5556

Clearly, one can see the proposed estimation procedure performs better than the MCD-based estimation procedures for all chosen scenarios (a)-(e), and the MLE based on the normal distribution is not resistant to the outliers at all. However, the worse performance in (f) indicates that the proposed estimate is not robust to the outliers in x-directions.

Simulation 2: For convenience, we will denote N-EM the EM algorithm based on Normal distribution, t-EM the EM algorithm based on t distribution and L-EM the EM algorithm based on Laplace distribution. To implement the t-EM procedure, the profile likelihood method discussed in Yao et al. (2014) is adopted to determine the proper degrees of freedom. The data are generated from the mixture of multivariate linear regression models with $g = 2$: $Y = \beta_1' X + \varepsilon_1$ if $G = 1$ and $Y = \beta_2' X + \varepsilon_2$ if $G = 2$, where G is a component indicator of Y with $P(G = 1) = 0.25$. The true regression coefficients are chosen to be

$$\beta_1 = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{14} & \beta_{15} & \beta_{16} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 3 \end{pmatrix}, \quad \beta_2 = -\beta_1.$$

The covariates $X \in R^2$ are generated from $N_2(0, I_{2\times 2})$, the random errors ε_1 and ε_2 have the same distribution as ε. We will consider the following six error distributions: (a) $\varepsilon \sim N(0, I_{3\times 3})$; (b) $\varepsilon \sim$ 3-dimensional Laplace distribution with mean 0 and identity covariance matrix; (c) $\varepsilon \sim t_1$, the 3-dimensional t distribution with 1 degrees of freedom, denoted as $MT_3(1)$; (d) $\varepsilon \sim t_3$, the 3-dimensional t distribution with 3 degrees of freedom, denoted as $MT_3(3)$; (e) $\varepsilon \sim 0.95N(0, I_{3\times 3}) + 0.05N(0, 50I_{3\times 3})$; (f) $\varepsilon \sim N(0, I_{3\times 3})$ with 5% high leverage outliers in both x- and y-directions ($X_1 = X_2 = 5, Y = 100$); and (g) $\varepsilon \sim N(0, I_{3\times 3})$ with 5% high leverage outliers only in x-direction ($X_1 = X_2 = 5, Y = 0$).

The sample size of $n = 100$ is used in the simulation study. For each case, the simulation is repeated 200 times. Same criterion as in simulation 1 are used as the criterion to evaluate the performance of various estimation procedures, except that this is done separately for π, $(\beta_{11}, \beta_{12}, \beta_{13})$ and $(\beta_{21}, \beta_{22}, \beta_{23})$. The simulation results are summarized in Table 2.

From the simulation results, we can see that if the true distribution of ε is normal, the MSEs of traditional MLE procedure are slightly smaller than two robust estimation procedures, which indicates the proposed estimation procedure and the procedure based on the multivariate t distribution are as efficient as the traditional MLE. For other cases when the distribution of ε has heavier tail or there are high leverage outliers in the data set, traditional MLE fails to provide reasonable estimates. The robust estimation via multivariate t distribution performs well, except when high leverage outliers are present in the data set. The computation of robust multivariate t distribution is intensive due to the estimation of degrees of freedom parameters. The simulation results clearly show that the proposed method in the paper outperforms or is at least comparable to any other methods except for some scenarios, for example, when ε has a lighter tail, the MSEs of proposed method are slightly larger than the traditional MLE method. However, when the ε has a heavier tail, the MSEs of proposed method are comparable to robust multivariate t distribution, and when the high leverage outliers are present in both directions, the proposed method outperforms any other methods. It is also clear that the proposed method and the t-procedures is not very robust when the outliers appear in the x-direction.

Table 2: Simulation results for $g = 2$

Error	N-EM	L-EM	t-EM
(a)	(0.002, 0.027, 0.015)	(0.003, 0.056, 0.022)	(0.003, 0.042, 0.022)
(b)	(0.002, 0.082, 0.030)	(0.002, 0.020, 0.022)	(0.002, 0.032, 0.014)
(c)	(0.034, 5.054, 2.604)	(0.004, 0.190, 0.039)	(0.004, 0.062, 0.026)
(d)	(0.004, 0.162, 0.077)	(0.003, 0.030, 0.042)	(0.003, 0.033, 0.019)
(e)	(0.003, 0.336, 0.176)	(0.002, 0.043, 0.045)	(0.002, 0.032, 0.029)
(f)	(0.031, 49.362, 3.645)	(0.004, 0.073, 0.196)	(0.017, 11.409, 0.260)
(g)	(0.009, 12.870, 0.075)	(0.006, 2.478, 0.992)	(0.004, 7.409, 0.016)

In summary, the simulation results indicate that the performance of the proposed robust estimation procedure is, in most of cases, comparable to the t-procedure. However, the extra step for finding a proper degrees of freedom makes the t-procedure is more computationally extensive than the proposed estimation procedure. Also, when $p = 1$, the natural connection between the LAD (least absolute deviation) estimate and the MLE based on Laplace distributions appears more attractive. That said, we do not intend to say the proposed robust estimation procedure is better than the t-procedure in all aspects. In fact, the extra degrees of freedom might provide t-procedure an extra adaptivity to the data. Except for proposing a computationally efficient robust estimation procedure for the mixtures of multivariate linear regression, the significance of this paper is to provide another alternative to robustly estimate the regression parameters in such models. In real application, collectively using all the available robust estimation methods might provide us more accurate information on the data structures.

Appendix: Proof of The Ascent Property

Proof. Suppose the latent class variable G has probability mass function $P(G = j) = \pi_j, j = 1, 2, \ldots, g$. Let $\psi(v)$ be the density function of the exponential variable V. $\psi(v|y, x; \theta)$ as the conditional density function of V given $Y = y, X = x$, and $w(j, v|y, x, \theta)$ as the conditional joint mass-density function of G and V given $Y = y, X = x$, and θ. Then we have $w(j, v|y, x, \theta) = \pi_j \phi(y|\beta_j' x, v\Sigma_j)\psi(v)/f(y|x, \theta)$, where $f(y|x, \theta)$ is defined in (2). Note that for given $\theta^{(k)}$, we have $\sum_{j=1}^{g} \int w(j, v|Y_i, X_i, \theta^{(k)})dv = 1$, therefore,

$$
\begin{aligned}
L_n(\theta) &= \sum_{i=1}^{n} \log \left[\sum_{j=1}^{g} \pi_j f_\varepsilon(Y_i - \beta_j' X_i, 0, \Sigma_j) \right] = \sum_{i=1}^{n} \left[\sum_{j=1}^{g} \int \log f(Y_i|X_i, \theta) w(j, v|Y_i, X_i, \theta^{(k)})dv \right] \\
&= \sum_{i=1}^{n} \left[\sum_{j=1}^{g} \int \log \left[\pi_j \phi(Y_i|\beta_j' X_i, v\Sigma_j)\psi(v) \right] w(j, v|Y_i, X_i, \theta^{(k)})dv \right] \\
&\quad - \sum_{i=1}^{n} \left[\sum_{j=1}^{g} \int \log w(j, v|Y_i, X_i, \theta) w(j, v|Y_i, X_i, \theta^{(k)})dv \right] \hat{=} L_{n1}(\theta) - L_{n2}(\theta).
\end{aligned}
$$

Recall the definition $\tau_{ij}^{(k+1)}$ in the E-step, and note that for any i, j, k,

$$
\phi(Y_i|\beta_j^{(k)'} X_i, v\Sigma_j^{(k)})\psi(v) = \psi(v|Y_i, X_i, \theta_j^{(k)}) f_\varepsilon(Y_i - \beta_j^{(k)'} X_i, 0, \Sigma_j^{(k)}),
$$

we have

$$
\frac{\pi_j^{(k)} \phi(Y_i|\beta_j^{(k)'} X_i, v\Sigma_j^{(k)})\psi(v)}{f(Y_i|X_i, \theta^{(k)})} = \tau_{ij}^{(k+1)} \psi(v|Y_i, X_i; \theta_j^{(k)}).
$$

Then by the definition of $w(j, v|y, x, \theta)$, we can show that $L_{n1}(\theta)$ is exactly the expression in (5). Therefore, the M-step in the proposed EM algorithm implies that $n^{-1}L_{n1}(\theta^{(k)}) \geq n^{-1}L_{n1}(\theta^{(k)})$.

Therefore, it suffices to show that in probability, $L_{n2}(\theta^{(k+1)}) \leq L_{n2}(\theta^{(k)})$. Note that the difference $L_{n2}(\theta^{(k+1)}) - L_{n2}(\theta^{(k)})$ is equivalent to

$$
\begin{aligned}
&\sum_{i=1}^{n} \left[\sum_{j=1}^{g} \int \log \frac{w(j, v|Y_i, X_i, \theta^{(k+1)})}{w(j, v|Y_i, X_i, \theta^{(k)})} w(j, v|Y_i, X_i, \theta^{(k)})dv \right] \\
&= \sum_{i=1}^{n} \left[\sum_{j=1}^{g} \pi_j \int \log \frac{w(v|Y_i, G = j, X_i, \theta^{(k+1)})}{w(v|Y_i, G = j, X_i, \theta^{(k)})} w(v|Y_i, G = j, X_i, \theta^{(k)})dv \right]
\end{aligned}
$$

which is less than 0 by the Kullback-Leibler information inequality applied to the conditional density function $w(v|Y_i, G = j, X_i, \theta^{(k)})$ for each i, j, k. □

References

Anderson, D. N., 1992. A multivariate linnik distribution. Statistics & probability letters 14 (4), 333–336.

Bai, X., Yao, W., Boyer, J. E., 2012. Robust fitting of mixture regression models. Computational Statistics & Data Analysis 56 (7), 2347–2359.

Bashir, S., Carter, E., 2012. Robust mixture of linear regression models. Communications in Statistics-Theory and Methods 41 (18), 3371–3388.

Böhning, D., 2000. Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others. Vol. 81. CRC press.

Dielman, T., 1984. Least absolute value estimation in regression models: An annotated bibliography. Communication in Statistics - Theory and Methods 4, 513–541.

Dielman, T., 2005. Least absolute value regression: recent contributions. Journal of Statistical Computation and Simulation 75, 263–286.

Donoho, D., 1982. Breakdown properties of multivariate location estimators. Qualifying paper, Harvard University, Boston.

Eltoft, T., Kim, T., Lee, T.-W., 2006. On the multivariate laplace distribution. IEEE Signal Processing Letters 13 (5), 300–303.

Ernst, M. D., 1998. A multivariate generalized laplace distribution. Computational Statistics 13 (2), 227–232.

Fang, K.-T., Kotz, S., Ng, K. W., 1990. Symmetric multivariate and related distributions. Chapman and Hall.

Farcomeni, A., Greco, L., 2015. Robust methods for data reduction. CRC Press, Taylor & Francis Group, LLC.

Fernandez, C., Osiewalski, J., Steel, M. F., 1995. Modeling and inference with v-spherical distributions. Journal of the American Statistical Association 90 (432), 1331–1340.

Huang, M., Li, R., Wang, S., 2013. Nonparametric mixture of regression models. Journal of the American Statitical Association 108 (503), 929–941.

Jiang, W., Tanner, M. A., 1999. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. Annals of Statistics, 987–1011.

Kotz, S., Kozubowski, T., Podgorski, K., 2012. The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance. Springer Science & Business Media.

Li, Y., Arce, G., 2004. A maximum likelihood approach to least absolute deviation regression. Journal of applied signal processing 12, 1762–1769.

Lin, T.-I., 2010. Robust mixture modeling using multivariate skew t distributions. Statistics and Computing 20 (3), 343–356.

McLachlan, G., Peel, D., 2004. Finite mixture models. John Wiley & Sons.

Newcomb, S., 1886. A generalized theory of the combination of observations so as to obtain the best result. American journal of Mathematics, 343–366.

Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P., 2007. Robust fitting of mixtures using the trimmed likelihood estimator. Computational Statistics & Data Analysis 52 (1), 299–308.

Pell, D., McLachlan, 2000. Robust mixture modelling using the t-distribution. Statistics and Computing 10, 339–348.

Portilla, J., Strela, V., Wainwright, M. J., Simoncelli, E. P., 2003. Image denoising using scale mixtures of gaussians in the wavelet domain. IEEE Transactions on Image processing 12 (11), 1338–1351.

Rousseeuw, P., Van Aelst, S., Van Driessen, K., Agullø, J., 2004. Robust multivariate regression. Technometrics 46 (3), 293–305.

Rousseeuw, P.J., V. D. K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Samko, S. G., Kilbas, A. A., Marichev, O. I., 1987. Fractional integrals and derivatives. 1993. Gordon & Breach Science Publishers.

Song, W., Yao, W., Xing, Y., 2014. Robust mixture regression model fitting by laplace distribution. Computational Statistics & Data Analysis 71, 128–137.

Spanier, J., Oldham, K. B., 1987. An atlas of functions. Taylor & Francis/Hemisphere.

Stahel, W. A., 1981. Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. thesis, ETH Zürich.

Ulrich, G., Chen, C., 1987. A bivariate double exponential distribution and its generalization. In: ASA Proceedings on Statistical Computing. pp. 127–129.

Wedel, M., Kamakura, W. A., 2012. Market segmentation: Conceptual and methodological foundations. Vol. 8. Springer Science & Business Media.

Xian Wang, H., bing Zhang, Q., Luo, B., Wei, S., 2004. Robust mixture modelling using multivariate t-distribution with missing information. Pattern Recognition Letters 25 (6), 701–710.

Yao, W., Wei, Y., Yu, C., 2014. Robust mixture regression using the t-distribution. Computational Statistics & Data Analysis 71, 116–127.