ROBUST MIXTURES OF REGRESSION MODELS

by

XIUQIN BAI

M.S., Kansas State University, USA, 2010

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2014

Abstract

This proposal contains two projects that are related to robust mixture models. In the first project, we propose a new robust mixture of regression models (Bai et al., 2012). The existing methods for fitting mixture regression models assume a normal distribution for error and then estimate the regression parameters by the maximum likelihood estimate (MLE). In this project, we demonstrate that the MLE, like the least squares estimate, is sensitive to outliers and heavy-tailed error distributions. We propose a robust estimation procedure and an EM-type algorithm to estimate the mixture regression models. Using a Monte Carlo simulation study, we demonstrate that the proposed new estimation method is robust and works much better than the MLE when there are outliers or the error distribution has heavy tails. In addition, the proposed robust method works comparably to the MLE when there are no outliers and the error is normal.

In the second project, we propose a new robust mixture of linear mixed-effects models. The traditional mixture model with multiple linear mixed effects, assuming Gaussian distribution for random and error parts, is sensitive to outliers. We will propose a mixture of multiple linear mixed t distributions to robustify the estimation procedure. An EM algorithm is provided to find the MLE under the assumption of t-distributions for error terms and random mixed effects. Furthermore, we propose to adaptively choose the degrees of freedom for the t-distribution using profile likelihood. In the simulation study, we demonstrate that our proposed model works comparably to the traditional estimation method when there are no outliers and the errors and random mixed effects are normally distributed, but works much better if there are outliers or the distributions of the errors and random mixed effects have heavy tails.

ROBUST MIXTURES OF REGRESSION MODELS

by

XIUQIN BAI

M.S., Kansas State University, USA, 2010

A DISSERTATION

submitted in partial fulfillment of the

requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics

College of Arts and Sciences

KANSAS STATE UNIVERSITY

Manhattan, Kansas

2014

Approved by:

Co-Major Professor

Weixin Yao

Approved by:

Co-Major Professor

Kun Chen

Copyright

Abstract

This proposal contains two projects that are related to robust mixture models. In the first project, we propose a new robust mixture of regression models (Bai et al., 2012). The existing methods for fitting mixture regression models assume a normal distribution for error and then estimate the regression parameters by the maximum likelihood estimate (MLE). In this project, we demonstrate that the MLE, like the least squares estimate, is sensitive to outliers and heavy-tailed error distributions. We propose a robust estimation procedure and an EM-type algorithm to estimate the mixture regression models. Using a Monte Carlo simulation study, we demonstrate that the proposed new estimation method is robust and works much better than the MLE when there are outliers or the error distribution has heavy tails. In addition, the proposed robust method works comparably to the MLE when there are no outliers and the error is normal.

In the second project, we propose a new robust mixture of linear mixed-effects models. The traditional mixture model with multiple linear mixed effects, assuming Gaussian distribution for random and error parts, is sensitive to outliers. We will propose a mixture of multiple linear mixed t distributions to robustify the estimation procedure. An EM algorithm is provided to find the MLE under the assumption of t-distributions for error terms and random mixed effects. Furthermore, we propose to adaptively choose the degrees of freedom for the t-distribution using profile likelihood. In the simulation study, we demonstrate that our proposed model works comparably to the traditional estimation method when there are no outliers and the errors and random mixed effects are normally distributed, but works much better if there are outliers or the distributions of the errors and random mixed effects have heavy tails.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express my appreciation to my major professors, Dr. Weixin Yao and Dr. Kun Chen,for all their encouragement, guidance and suggestions.

My sincere appreciation also goes to Dr Marianne Korten, for her willingness to serve as the chairperson of the committee for my doctoral degree.

I would also like to thank Dr. Paul Nelson, Dr. Juan Du and Dr. Dong Li for their willingness to serve on my committee and for their valuable insight.

My gratefulness extends to all my friends for their help and support during the completion of the report. I would like to thank everyone in the department for their kindness.

Chapter 1

Robust Fitting of Mixture Regression Models

1.1 Introduction

Mixture regression models are widely used to investigate the relationship between variables coming from several unknown latent homogeneous groups. They have applications in many fields, including engineering, genetics, biology, econometrics, and marketing. A typical data set is the tone perception data (Cohen, 1984) which is shown in Figure 1.1. In the tone perception experiment of Cohen (1984), a pure fundamental tone with electronically generated overtones added was played to a trained musician. The overtones were determined by a stretching ratio. The experiment was designed to determine if either of two musical perception theories was reasonable (see Cohen, 1984 for more detail). Based on Figure 1.1, two lines are evident which correspond to the behavior indicated by the two musical perception theories. The two regression lines correspond to correct tuning and tuning to the first overtone, respectively.

The model setting for mixtures of linear regression models can be stated as follows. Let Z be a latent class variable with $P(Z_i = j \mid \mathbf{x}_i) = \pi_j$ for $j = 1, 2, \cdots, m$, where $\mathbf{x}$ is a p-dimensional vector. Given $Z_i = j$, suppose that the response y_i depends on $\mathbf{x}_i$ in a linear way

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_j + \epsilon_{ij}, \tag{1.1.1}$$

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

$$f(y|\mathbf{x}) = \sum_{j=1}^{m} \pi_j \phi(y; \mathbf{x}^T \boldsymbol{\beta}_j, \sigma_j^2), \tag{1.1.2}$$

and the log-likelihood function for observations $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ is

$$\sum_{i=1}^{n} \ln \left[\sum_{j=1}^{m} \pi_j \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_j, \sigma_j^2) \right], \tag{1.1.3}$$

where $\phi(\cdot\,; \mu, \sigma^2)$ is the density function of $N(\mu, \sigma^2)$. See, for example, Jacobs, Jordan, Nowlan, and Hinton (1991), Jiang and Tanner (1999), Wedel and Kamakura (2000), and Skrondal and Rabe-Hesketh (2004), for some applications of model (1.1.2). The unknown parameters in the model (1.1.2) can be estimated by the maximum likelihood estimator (MLE), which maximizes (1.1.3). Note that the maximizer of (1.1.3) does not have an explicit solution and is usually estimated by the EM algorithm (Dempster, Laird, and Rubin, 1977).

Note that different permutations of component parameters will give the same density $f(y \mid \mathbf{x})$ of (1.1.2), which is called label-switching in mixture models. See, for example, Celeux, Hurn, and Robert (2000), Stephens (2000), and Yao and Lindsay (2009) for more detail. Hence, we will say the model (1.1.2) is identifiable up to a permutation of component parameters. To insure the identifiability of the model (1.1.2), we adopt the conditions of Hennig (2000).

Similar to the least squares estimate (LSE) for linear regression, the normality based MLE is sensitive to outliers or heavy-tailed error distributions. For linear regression, the M estimate, which replaces the least squares criterion by a robust criterion, is one of the most commonly used robust estimates for the regression parameters. See, for example, Huber (1973, 1981), Andrews (1974), Rousseeuw and Yohai (1984), Hampel, Ronchetti, Rousseeuw, and Stahel (1986), Yohai (1987), and Rousseeuw and Leroy (1987), for more detail. However, there is little research related to estimating the mixture regression parameters robustly, in part because it is not easy to replace the log-likelihood in (1.1.3) by a robust criterion similar to the M estimate. Neykov, Filzmoser, Dimova, and Neytchev (2007) proposed robust fitting of mixtures using the trimmed likelihood estimator. Markatou (2000) and Shen, Yang, and Wang (2004) proposed using a weight factor for each data to robustify the estimation procedure for mixture regression models. There are also some related robust methods for linear clustering; see, for example, Hennig (2002, 2003), Mueller and Garlipp (2005),

García-Escudero, Gordaliza, San Martín, Van Aelst, and Zamar (2009), and García-Escudero, Gordaliza, Mayo-Iscara, and San Martín (2010).

In this project, we propose a new and simple robust estimation procedure for the mixture regression parameters by modifying the existing EM algorithm rather than focusing on the maximization of the function (1.1.3). Due to the normality assumption, the least squares criterion is used in the M step of EM algorithm for mixture regression models. We propose replacing the least squares criterion in the M step by a robust criterion, such as Tukey's bisquare function. Based on a Monte Carlo study, we demonstrate that the proposed new estimate is robust and much more efficient than the MLE when the data have outliers or the error distribution has heavy tails. Furthermore, the proposed method provides results comparable to the traditional MLE when there are no outliers and the error is exactly normal.

The rest of this chapter is organized as follows. In Section 2, we introduce our new robust estimation procedure for mixture linear regression models. In Section 3, a Monte Carlo simulation study and a real data application are used to illustrate the robustness of the proposed methodology and compare it with the traditional MLE. Some discussions are given in Section 4. Technical conditions and proofs are provided in the Appendix.

1.2 Robust Mixture Regression Models

1.2.1 Introduction to the existing estimate

It is well known that the log-likelihood function (1.1.3) is unbounded and goes to infinity if one observation exactly lies on one component line and the corresponding component variance goes to zero. There has been considerable research dealing with the unbounded likelihood issue. See, for example, Hathaway (1985, 1986), Chen, Tan, and Zhang (2008), and Yao (2010). In this chapter, for simplicity of explanation of our new robust method, we assume equal variance for each component in order to avoid the unboundedness of the mixture likelihood (1.1.3).

The existing EM algorithm to maximize (1.1.3) is as follows.

Algorithm 1. Based on the initial values of $\{\pi_j^{(0)}, \boldsymbol{\beta}_j^{(0)}, \sigma^{(0)}, j = 1, \ldots, m\}$, the EM algorithm iterates between the following E-step and M-step.

E-step: Calculate the classification probabilities

$$p_{ij}^{(k+1)} = \frac{\pi_j^{(k)}\phi(y_i; \mathbf{x}_i^T\boldsymbol{\beta}_j^{(k)}, \sigma^{2(k)})}{\sum_{l=1}^m \pi_l^{(k)}\phi(y_i; \mathbf{x}_i^T\boldsymbol{\beta}_l^{(k)}, \sigma^{2(k)})}, \; i = 1, \ldots, n; j = 1, \ldots, m.$$

M step: Update the parameters

$$\begin{aligned}
\boldsymbol{\beta}_j^{(k+1)} &= \arg\min_{\boldsymbol{\beta}_j} \sum_{i=1}^n p_{ij}^{(k+1)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j)^2 \\
&= (\mathbf{X}^T\mathbf{W}_j^{k+1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_j^{(k+1)}\mathbf{y}, \qquad (1.2.1) \\
\pi_j^{(k+1)} &= \frac{1}{n}\sum_{i=1}^n p_{ij}^{(k+1)}, \\
\sigma^{2(k+1)} &= \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^m p_{ij}^{(k+1)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j^{(k+1)})^2,
\end{aligned}$$

where $j = 1, \ldots, m, \mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T, \mathbf{y} = (y_1, \ldots, y_n)^T$, and $\mathbf{W}_j^{(k+1)}$ is a $n \times n$ diagonal matrix with diagonal elements $\{p_{ij}^{(k+1)}, i = 1, \ldots, n\}$.

It can be seen from (1.2.1) that the MLE based EM algorithm updates $\boldsymbol{\beta}$ by a weighted least squares estimate in the M step, since $\phi(\cdot)$ is a normal density. It is well known that the least squares criterion is sensitive to outliers and heavy-tailed error distributions. In this project, we provide a robust estimation procedure for the mixture regression models.

1.2.2 Robust estimation of a mixture of linear regressions

It is not easy to use the idea of an M estimate to directly replace the objective function (1.1.3) with a robust criteria. In this project, we propose to replace the least squares criterion (1.2.1) in the M step of Algorithm 1 with a robust criterion ρ. Therefore, $\boldsymbol{\beta}_j^{(k+1)}, j = 1, \ldots, m$, is the solution of

$$\sum_{i=1}^n p_{ij}^{(k+1)}\mathbf{x}_i\psi\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j}{\sigma^{(k)}}\right) = 0, \qquad (1.2.2)$$

where $\psi(\cdot) = \rho'(\cdot)$ and $\sigma^{(k)}$ is a robust scale estimate of the error ϵ_{ij}'s. One of the commonly used ρ functions is Huber's ψ-function $\psi_c(t) = \rho'(t) = \max\{-c, \min(c, t)\}$ (Huber, 1981). Huber (1981) recommends using $c = 1.345$ in practice, which produces a relative efficiency of approximately 95% when the error density is normal. Another possibility for $\psi(\cdot)$ is Tukey's bisquare function $\psi_c(t) = t\{1 - (t/c)^2\}_+^2$, which weights the tail contribution of t by a biweight function. In the parametric robustness literature, the use of $c = 4.685$,

which produces 95% efficiency, is recommended. If we use L_1 loss function $\rho(t) = |t|$, we will get the median regression. For more detail, see Huber (1973, 1981), Andrews (1974), Beaton and Tukey (1974), Holland and Welsch (1977), and Hampel, et al. (1986).

Note that

$$\sum_{i=1}^{n} p_{ij}^{(k+1)} \mathbf{x}_i \psi\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j}{\sigma^{(k)}}\right) \approx \sum_{i=1}^{n} p_{ij}^{(k+1)} \mathbf{x}_i W\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j^{(k)}}{\sigma^{(k)}}\right)\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j}{\sigma^{(k)}}\right)$$
$$= \sum_{i=1}^{n} p_{ij}^{*(k+1)} \mathbf{x}_i \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j}{\sigma^{(k)}}\right),$$

where $W(t) = \psi(t)/t$ and

$$p_{ij}^{*(k+1)} = p_{ij}^{(k+1)} W\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j^{(k)}}{\sigma^{(k)}}\right).$$

Based on the above approximation, the solution of (1.2.2) can be approximated by

$$\boldsymbol{\beta}_j^{(k+1)} = \left(\sum_{i=1}^{n} p_{ij}^{*(k+1)} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i=1}^{n} p_{ij}^{*(k+1)} \mathbf{x}_i y_i,$$

which is one step of the iterative reweighting algorithm (Maronna, Martin, and Yohai, 2006, Sec. 4.5.2). Note that $\boldsymbol{\beta}_j^{(k+1)}$ can be considered to be a weighted least squares estimator with the weights $\{p_{ij}^{*(k+1)}, i = 1, \ldots, n\}$.

Based on the above discussions, we propose the following robust estimation procedure for the mixtures of linear regression model (1.1.1).

Algorithm 2. Based on the initial values of $\{\pi_j^{(0)}, \boldsymbol{\beta}_j^{(0)}, \sigma^{(0)}, j = 1, \ldots, m\}$, the proposed robust EM-type algorithm is to iterate the following E-step and M-step.

E-step: Calculate the classification probabilities

$$p_{ij}^{(k+1)} = \frac{\pi_j^{(k)} \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_j^{(k)}, \sigma^{2(k)})}{\sum_{l=1}^{m} \pi_l^{(k)} \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_l^{(k)}.\sigma^{2(k)})}$$

M step: Update the parameters

$$\boldsymbol{\beta}_j^{(k+1)} = \left(\sum_{i=1}^{n} p_{ij}^{*(k+1)} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i=1}^{n} p_{ij}^{*(k+1)} \mathbf{x}_i y_i$$
$$= (\mathbf{X}^T \mathbf{W}_j^{*(k+1)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_j^{*(k+1)} \mathbf{y}, \tag{1.2.3}$$
$$\pi_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(k+1)},$$
$$\sigma^{2(k+1)} = \frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j^{(k+1)})^2 w_{ij}^{(k+1)}, \tag{1.2.4}$$

where $j = 1, \ldots, m$, $\mathbf{W}_j^{*(k+1)}$ is a $n \times n$ diagonal matrix with diagonal elements $\{p_{ij}^{*(k+1)}, i = 1, \ldots, n\}$, and

$$w_{ij}^{(k+1)} = \min \left[1 - \left\{ 1 - \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j^{(k+1)}}{1.56\sigma^{(k)}} \right)^2 \right\}^3, 1 \right] \left(\frac{\sigma^{(k)}}{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j^{(k+1)}} \right)^2.$$

Here, (1.2.4) is our proposed robust scale estimate, which extends the idea of $M - estimate\ of\ scale$ (see Maronna, et al., 2006, section 2.2 for more detail). Note that (1.2.4) is similar to the traditional nonrobust scale estimate for mixtures of regression except for the adjustment factor "2" and the weights $w_{ij}^{(k+1)}$, which are the bisquare weights recommended by Maronna, et al., (2006). One may also apply some other robust scale estimate to get the weights $w_{ij}^{(k+1)}$.

The above proposed method can be easily extended to the unequal variances case. For example, similar to Hathaway (1985, 1986), the above robust EM-type algorithm can be implemented over a constrained parameter space

$$\Omega_C = \{\boldsymbol{\theta} \in \Omega : \sigma_h/\sigma_j \geq C > 0, 1 \leq h \neq j \leq m\}, \tag{1.2.5}$$

where $C \in (0, 1], \boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1^T, \sigma_1, \ldots, \pi_{m-1}, \boldsymbol{\beta}_{m-1}^T, \sigma_{m-1}, \boldsymbol{\beta}_m^T, \sigma_m)^T$, and Ω denotes the unconstrained parameter space.

In (1.1.1), if $\mathbf{x}$ only includes the intercept term 1, the model is the regular normal mixture model. Hence, our proposed robust estimation procedure can be also used to robustly estimate the location parameters in the normal mixture model.

Initial values: There are many ways to find the initial values for $\{\pi_j^{(0)}, \boldsymbol{\beta}_j^{(0)}, \sigma^{(0)}, j = 1, \ldots, m\}$. One method is to use trimmed likelihood estimates (TLE) (Neykov, et al. 2007). Note that the TLE is robust to both low leverage and high leverage outliers under certain general conditions (Neykov, et al. 2007). Another possible method is that we first randomly partition the data or a subset of the data into m groups. For each group, we use some robust regression method, such as the MM-estimate (Yohai, 1987), to estimate the component regression parameters. Similar partition ideas have been used to find the initial values for finite mixture models (McLachlan and Peel, 2000). In addition, we can also apply the robust linear clustering method to find the initial regression parameter values. See, for example, Hennig (2002, 2003), and García-Escudero, et al. (2009). Note that though, technically, the robust linear clustering methods do not produce consistent regression component estimators. But in many cases, they are close enough to provide good

initial values, since the proposed algorithm doesn't require the initial values to be consistent.

Convergence of Algorithm 2: In the estimating equation (1.2.10), if we replace p_{ij} by z_{ij}, where z_{ij} is the latent component indicator and is equal to 1 if ith observation is from jth component and 0 otherwise, then the corresponding proposed Algorithm 2 can be considered as the *ES algorithm* proposed by Elashoff and Ryan (2004) for estimating equations with missing data. Therefore, the convergence property of the proposed Algorithm 2 can be proved similarly to the ES algorithm of Elashoff and Ryan (2004).

1.2.3 Asymptotic results

In this section, for simplicity of explanation and the proof, we assume that the scale parameter σ used in (1.2.2) is fixed. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_m^T, \pi_1, \ldots, \pi_m)^T$ and $\hat{\boldsymbol{\theta}}_n$ be the estimate found by our proposed robust EM-type Algorithm 2. Note that the $\hat{\boldsymbol{\theta}}_n$ solves the following estimating equations

$$\sum_{i=1}^{n} p_{ij}(\boldsymbol{\theta}) \mathbf{x}_i \psi\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j}{\sigma}\right) = 0, \tag{1.2.6}$$

$$\pi_j = \sum_{i=1}^{n} p_{ij}(\boldsymbol{\theta})/n, \ j = 1, \ldots, m, \tag{1.2.7}$$

where

$$p_{ij}(\boldsymbol{\theta}) = \frac{\pi_j \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_j, \sigma^2)}{\sum_{l=1}^{m} \pi_l \phi(y_i; \mathbf{x}_i^T \boldsymbol{\beta}_l, \sigma^2)}. \tag{1.2.8}$$

Let $\mathbf{z}_i = (\mathbf{x}_i^T, y_i)^T$ and

$$\Psi(\mathbf{z}_i, \boldsymbol{\theta}) = \left\{ p_{i1} \mathbf{x}_i \psi\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_1}{\sigma}\right), \ldots, p_{im} \mathbf{x}_i \psi\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_m}{\sigma}\right), p_{i1} - \pi_1, \ldots, p_{i,m-1} - \pi_{m-1} \right\}^T, \tag{1.2.9}$$

where $p_{ij} = p_{ij}(\boldsymbol{\theta})$ is defined in (1.2.8). Therefore, our proposed estimate $\hat{\boldsymbol{\theta}}_n$ solves the equation

$$S_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \Psi(\mathbf{z}_i, \boldsymbol{\theta}) = 0.$$

Theorem 1.2.1. *Under the regularity conditions (A1)—(A5) in the Appendix, if the error in (1.1.1) is normal, then there exists a sequence* $\{\hat{\boldsymbol{\theta}}_n, n = 1, 2, \ldots, \}$ *such that*

a) $P(\hat{\boldsymbol{\theta}}_n$ *is a solution to* $S_n(\boldsymbol{\theta}) = 0) \to 1$

b) $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$, *where* $\boldsymbol{\theta}_0$ *is the true value of* $\boldsymbol{\theta}$.

Note that the true value of $\boldsymbol{\theta}_0$ is not unique due to the label switching. Therefore, the consistent sequence $\{\hat{\boldsymbol{\theta}}_n, n = 1, 2, \dots, \}$ depend on the specific label of $\boldsymbol{\theta}_0$. The above theorem states that when the error is normal there exists a consistent solution to the equation $S_n(\boldsymbol{\theta}) = 0$. If there is only one root of $S_n(\theta) = 0$, the above theorem tells us that the estimate found by the proposed algorithm must be consistent.

However, like general estimating equations, there may be multiple solutions to the above equation and the selection of a consistent root is usually very difficult. In addition, it is also very difficult to directly prove that the sequence found by our algorithm is consistent. We will provide an empirical way to select the root when multiple roots are found in Section 1.3.

Let

$$A = \mathbf{E}_{\boldsymbol{\theta}_0} \left\{ \frac{\partial \Psi(Z, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right\} \tag{1.2.10}$$

and

$$B = \mathbf{E}_{\boldsymbol{\theta}_0} \{ \Psi(Z, \boldsymbol{\theta}) \Psi(Z, \boldsymbol{\theta})^T \}.$$

Theorem 1.2.2. *Under the regularity conditions (A1)—(A7) in the Appendix, when the error in (1.1.1) is normal, the estimate $\hat{\boldsymbol{\theta}}_n$, given in Theorem 1.2.1, has the following asymptotic distribution*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, V),$$

where $V = A^{-1}BA^{-1}$.

Robustness: Based on our empirical studies, the method based on Tukey's bisquare has greater resistance to high leverage outliers and has overall better performance than the method based on Huber's function. Hennig (2004) treats 1-d mixtures, which is "intercept-only" regression and therefore a special case of what is treated in this project. Hennig (2004) proved that the robust mixture estimates by maximizing some objective functions have low breakdown. It will be interesting to know whether their results can be similarly proved for mixtures of regression models if estimating equations based estimators are used.

Since our proposed estimate solves the equation (2.10), based on the theory of M estimate (Maronna, et al., 2006, section 5.4.2), the influence function of our proposed estimate is

$$\text{If}((\mathbf{x}_0, y_0), \boldsymbol{\theta}_0) = -A^{-1}\Psi((\mathbf{x}_0, y_0), \boldsymbol{\theta}_0),$$

where A is defined in (1.2.10) and Ψ is defined in (1.2.9).

The sample breakdown point is another important measure of the robustness. However, as García-Escudero, et al. (2010) stated, the traditional definition of breakdown point is not the right one to quantify the robustness of clustering regression procedures to outliers, since the robustness of these procedures is not only data dependent but also cluster dependent.

1.3 Simulation Studies and Real Data Application

In this section, we use a Monte Carlo simulation study and the analysis of a real data set to compare our proposed robust estimation procedure with the MLE for mixture regression models. For the proposed robust method, we consider both Tukey's bisquare function with $c = 4.685$ and Huber's ψ function with $c = 1.345$ and denote them by Robust-Bisquare and Robust-Huber, respectively. We run the proposed EM type algorithm until the maximum difference between the updated parameter estimates of two consecutive iterations is less than 10^{-5}. For the MLE, we start the algorithm from 20 random initial values and then choose the converged mode with the largest likelihood. For better comparison, we also include the robust estimates based on the trimmed maximum likelihood estimator (TLE) proposed by Neykov, et al. (2007) with the percentage of trimmed data α set to 0.1. The choice of α plays an important role for the TLE. If α is too large, the TLE will lose much efficiency. If α is too small and the percentage of outliers is more than α then the TLE will fail. In our simulation study, the proportion of outliers is never greater than 0.1.

The TLE is implemented based on the FAST-TLE algorithm (Neykov, et al. 2007 with 20 initial values calculated from 20 randomly chosen sub-samples). For Robust-Bisquare and Robust-Huber, we used 22 initial values that consists of FAST-TLE, robust linear clustering method (García-Escudero, et al. 2009), and 20 initial parameter values used by FAST-TLE. When the proposed algorithm can identify multiple roots, it is important to find the right one. However, finding a consistent root among multiple roots is always a difficult problem for estimating equations. In our simulation study and real data analysis, we used the root, called *modal root*, which most initial values converge to. (One of the motivations of using modal root is that it can be used to approximate the major maximizer of the unknown objective function that defines the estimating equation (1.2.10) if the area associated with major maximizer is larger than the area

associated with any other local minor maximizer/minimizer (Li, Ray, and Lindsay, 2007).) Although it is difficult to give the theoretical support for such choice, our empirical study demonstrates the effectiveness of using such modal root. In addition, our empirical study found that the converged roots starting from FAST-TLE are usually the same as the modal root. Therefore, in practice, to save computation time, one might simply run the proposed algorithm starting from FAST-TLE.

In addition, for mixture models, the label switching issues (Celeux, Hurn, and Robert, 2000; Stephens, 2000; Yao and Lindsay, 2009) also create much trouble when doing comparison using the simulation study. Different labeling strategies might give totally different results and there are no widely accepted labeling methods. In our simulation study, we simply choose the labels by minimizing the distance to the true parameter values. It requires more research to compare different labeling methods.

Example 1. We generate the independent and identically distributed (i.i.d.) data $\{(x_{1i}, x_{2i}, y_i), i = 1, \ldots, n\}$ from the model

$$Y = \begin{cases} 0 + X_1 + X_2 + \epsilon_1, & \text{if } Z = 1; \\ 0 - X_1 - X_2 + \epsilon_2, & \text{if } Z = 2. \end{cases},$$

where Z is a component indicator of Y with $P(Z = 1) = 0.25$, $X_1 \sim N(0, 1), X_2 \sim N(0, 1)$, and ϵ_1 and ϵ_2 have the same distribution as ϵ. Note that the two regression lines will intersect each other when $X_1 = 0$ and $X_2 = 0$. We consider the following five cases:

Case 1: $\epsilon \sim N(0, 1)$ – Standard normal distribution.

Case 2: $\epsilon \sim t_3$ – t-distribution with degrees of freedom 3.

Case 3: $\epsilon \sim t_1$ – t-distribution with degrees of freedom 1 (Cauchy distribution).

Case 4: $\epsilon \sim 0.95N(0, 1) + 0.05N(0, 5^2)$ – Contaminated normal mixture.

Case 5: $\epsilon \sim N(0, 1)$ with 5% of high leverage outliers being $X_1 = 20, X_2 = 20$ and $Y = 100$.

We use Case 1 to test the efficiency of our robust estimation method compared to the traditional MLE when the error is exactly normally distributed and there are no outliers. Case 2 is a heavy-tailed distribution. The t-distributions with degrees of freedom from 3 to 5 are often used to represent the heavy-tailed distributions. Case 3 is an extremely heavy-tailed t distribution with one degree of freedom. Case

4 is a contaminated normal mixture model, which is often used to mimic the outlier situation. The 5% data from $N(0, 5^2)$ are likely to be low leverage outliers. In Case 5, 95% of the observations have the error distribution $N(0, 1)$, but 5% of the observations are replicated high leverage outliers with $X_1 = 20, X_2 = 20$, and $Y = 100$.

Tables 1.1 and 1.2 report the bias and standard errors (Std) of the parameter estimates for each estimate for samples of size $n = 100$ and $n = 400$, respectively. The number of replicates is 1,000. Based on Tables 1.1 and 1.2, we note the following general findings:

1. When there are no outliers and the error is normal (Case I), all methods estimate the parameters well, except that TLE has large bias for some regression parameters. In addition, the MLE works slightly better than the proposed robust methods and Robust-Huber works better than the Robust-Bisquare, especially when sample size is small, such as $n = 100$. (Note that in this case, the traditional MLE, which assumes a normal error, is asymptotically most efficient.)

2. For Cases II to V, all robust estimates work much better than the MLE. In addition, the Robust-Bisquare overall has the best performance. (For Case V, TLE works slightly better than Robust-Bisquare when $n = 400$.)

3. For Case II ($\epsilon \sim t_3$) and IV ($\epsilon \sim 0.95N(0, 1) + 0.05N(0, 5^2)$), the Robust-Huber works better than the TLE. For Case III ($\epsilon \sim t_1$) and V (5% high leverage outliers), the TLE works better than the Robust-Huber, which has a large bias for parameter estimates.

Based on the above findings, we can see that the Robust-Bisquare is robust to both low leverage outliers and high leverage outliers and has the overall best performance. Therefore, in practice, we recommend the use of Robust-Bisquare method.

Table 1.3 reports the average number of found solutions when using 22 initial values for the proposed robust methods. From the table, we can see that in many cases the proposed algorithm can identify multiple solutions and the average number of found roots tends to decrease when sample size increases.

Example 2. We generate the independent and identically distributed (i.i.d.) data $\{(x_i, y_i), i = 1, \ldots, n\}$

from the model

$$Y = \begin{cases} 1 + X + \epsilon_1, & \text{if } Z = 1; \\ 2 + 2X + \epsilon_2, & \text{if } Z = 2; \\ 3 + 5X + \epsilon_3, & \text{if } Z = 3; \end{cases},$$

where Z is a component indicator of Y with $P(Z = 1) = P(Z = 2) = 0.3, P(Z = 3) = 0.4$, $X \sim N(0, 1)$, and ϵ_1, ϵ_2, and ϵ_3 have the same distribution as ϵ. We consider the same five cases for ϵ as in Example 1, except for Case V, in which the 5% high leverage outliers are $X = 20$ and $Y = 200$. Note that in this case all three components have the same sign of the slopes and the first two components are very close.

Tables 1.4 and 1.5 report the bias and standard errors (Std) of the parameter estimates for each estimate for samples of size $n = 100$ and $n = 400$, respectively. The number of replicates is 1,000. Based on Tables 1.4 and 1.5, we can get similar findings to the Example 1, except that TLE also works better than Robust-Huber in Cases II and IV.

Table 1.6 reports the average number of found roots. From the table, we can see that the average number of roots tends to decrease when the sample size increases. In addition, based on Tables 1.3 and 1.6, we can also see that the average number of roots tend to increase when the number of components increases.

Example 3. Next, we use the tone data introduced in Section 1 to illustrate the Robust-Bisquare method and compare it with the MLE. To better see the robustness of our proposed estimate, we have added ten identical high leverage outliers $(0, 4)$ to the original data set (the range of the Actual tone ratio in the original data set is from 1.35 to 3), and refit the data with both the Robust-Bisquare and the MLE. For this data set, Robust-Bisquare found four solutions and 13 out of 22 initial values converged to the modal root. For this data set, both FAST-TLE (Neykov, et al. 2007) and robust linear clustering estimate (García-Escudero, et al. 2009) converge to the modal root. The numbers of initial values converged to the other three minor roots are 4, 3, and 2, respectively.

Figure 1.2 shows the scatter plot with the estimated regression lines generated by MLE (dashed lines) and Robust-Bisquare (solid line) for the data augmented by the outliers (stars). From Figure 1.2, we note that our proposed robust method provides almost the same fit as the one in Figure 1.1 and thus is robust to the added outliers. However, the MLE for one of the components fits the line through the outliers and the MLE for the other component fits the line using the rest of data. In this case, the ten high leverage

outliers have a big impact on the fitted regression lines.

1.4 Discussion

In this project, we propose a new robust estimation procedure for mixture regression models. Instead of modifying the log-likelihood objective function, we propose to modify the existing EM algorithm for mixture regression models by replacing the least squares criterion with a robust criteria in the M step. Our empirical study demonstrates that the proposed method which utilizes the bisquare function works well and is robust and much more efficient than the existing MLE when there are outliers present or the error has heavy tails. In addition, the proposed robust estimation procedure has performance comparable to the MLE when there are no outliers and the error is exactly normal. We believe that similar modifications can be applied to other mixture regression models such as mixtures of generalized linear models. Such extensions will be our future interest.

Although our empirical study demonstrates the effectiveness of the proposed modal root when multiple solutions are found, it requires more research to provide some theoretical guideline for the choice of a consistent root. One method is to find the objective function for the estimating equation (1.2.7) and then choose the root that maximizes the objective function. Similar ideas have been used by McCullagh and Nelder (1989), Li (1993), and Hanfelt and Liang (1995, 1997).

Theorem 1.2.1 and 1.2.2 assume that σ is fixed. The things will be more complicated if σ is estimated. Note that the scale estimator (1.2.4) can be considered as the solution to the estimating equation

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}\rho\left(\frac{y_i - \mathbf{x}_i^T\beta_j}{\sigma}\right) = 0.5, \tag{1.4.1}$$

where $\rho(\cdot)$ corresponds to Tukey's bisquare function. Therefore, if σ is estimated, Theorem 1.2.1 and 1.2.2 can be still proved similarly by adding another estimating equation (1.4.1). However, the asymptotic variance in Theorem 1.2.2 will be different if σ is estimated.

In addition, note that Theorem 1.2.1 only proved the *existence* of a consistent sequence of solutions. The normality results given in Theorem 1.2.2 only applies to that particular consistent sequence found in Theorem 1.2.1. Unfortunately, we are not able to directly prove that the solution found by the proposed algorithm is consistent, which is a very difficult task and requires more research. Therefore, Theorem 1.2.1

and 1.2.2 have very limited practical use. However, one thing that Theorem 1.2.1 can tell us is that the estimate found by the proposed algorithm is consistent if the estimating equations only have one root.

Appendix

The following technical conditions are imposed in this section. They are not the weakest possible conditions, but they are imposed to facilitate the proofs.

Technical Conditions:

A1 $(\mathbf{x}_i, Y_i)$ are independent and identically distributed from some joint density $f(\mathbf{x}, y)$. In addition, the number of distinct $(p-1)$-dimensional hyperplanes which one needs to cover the covariates is no less than m.

A2 The true parameter $\boldsymbol{\theta}_0$ is an interior point of parameter space Ω, i.e., $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j, 1 \leq i \neq j \leq m$, and $\pi_j > 0, j = 1, \ldots, m$.

A3 The $\psi(\cdot)$ function satisfies

$$\int_{-\infty}^{\infty} \psi(t)\phi(t)dt = 0,$$

where $\phi(t)$ is the density for standard normal.

A4 $\psi(t)$ is continuous and $\mathbf{E}_{\boldsymbol{\theta}}\{\Psi(Z, \boldsymbol{\theta})\}$ is differentiable at $\boldsymbol{\theta}_0$ and the derivative matrix is negative (positive) definite.

A5 In a neighborhood of $\boldsymbol{\theta}_0$, $S_n(\boldsymbol{\theta})$ converges in probability uniformly to $\mathbf{E}_{\boldsymbol{\theta}_0}\{\Psi(Z, \boldsymbol{\theta})\}$, i.e.,

$$\sup_{\boldsymbol{\theta}} \left[\left| n^{-1} \sum_{i=1}^{n} \Psi(Z_i, \boldsymbol{\theta}) - \mathbf{E}_{\boldsymbol{\theta}}\{\Psi(Z, \boldsymbol{\theta})\} \right| : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \leq \delta_n \right] \xrightarrow{p} 0 \text{ if } \delta_n \to 0.$$

A6 $\mathbf{E}_{\boldsymbol{\theta}}\{\Psi(Z, \boldsymbol{\theta})\Psi(Z, \boldsymbol{\theta})^T\}$ and $\mathbf{E}_{\boldsymbol{\theta}}\{\partial\Psi(Z, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\}$ exist and are continuous functions of $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \Omega$ with $\mathbf{E}_{\boldsymbol{\theta}}\{\partial\Psi(Z, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\} \neq 0$ in a neighborhood of $\boldsymbol{\theta}_0$.

A7 $||\partial^2\Psi(Z, \boldsymbol{\theta})/\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j|| \leq M(Z)$ for all $\boldsymbol{\theta}$ and $1 \leq i \leq j \leq 2m-1$, where $M(Z)$ is an integrable function.

The condition A1 is the identifiability conditions for mixtures of liner regression models used by Hennig (2000). The condition $A3$ guarantees $\mathbf{E}\{\Psi(Z, \boldsymbol{\theta})\} = 0$ and thus the existence of a consistent solution to

the estimating functions when the error is normal. If $\psi(\cdot)$ is an odd function, then the Condition A3 is satisfied. The conditional $A5$ is satisfied if $\Psi(Z,\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for every Z and $|\Psi(Z,\boldsymbol{\theta})|$ is dominated by an integrable function, say, $G(Z)$. Here, we put conditions directly on estimating function $\Psi(Z,\boldsymbol{\theta})$ (Godambe, 1991), instead of on $x-$variables. Hennig (2000) pointed out that some limiting conditions on $x-$variables might be needed to get the consistency results. However, we are not able to directly derive the explicit limiting conditions on $x-$variables from Condition A5, which is very cumbersome as stated in Hennig (2000).

Proof of Theorem 1.2.1: From A1 and A3, we have

$$\mathbf{E}\left\{p_{ij}\mathbf{x}_i\psi\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}_j}{\sigma}\right)\Big|\mathbf{x}_i\right\}=\pi_j\mathbf{x}_i\int_{\infty}^{\infty}\phi(t)\psi(t)dt=0. \tag{1.4.2}$$

and

$$\mathbf{E}(p_{ij}\mid\mathbf{x}_i)=\pi_j\int_{-\infty}^{\infty}\phi(y;\mathbf{x}_i^T\boldsymbol{\beta}_j,\sigma^2)dy=\pi_j\int_{-\infty}^{\infty}\phi(t)dt=\pi_j. \tag{1.4.3}$$

Therefore, $\mathbf{E}\{\Psi(x_i,\boldsymbol{\theta}_0)\}=0$.

Let R_n be the collection of all solutions to $S_n(\boldsymbol{\theta})=0$. If $R_n\neq\emptyset$, define $a_n=\inf_{\boldsymbol{\theta}\in R_n}||\boldsymbol{\theta}-\boldsymbol{\theta}_0||$. By definition, there exists a sequence of $\{\hat{\boldsymbol{\theta}}_{n,k}:,k=1,2,\ldots\}$ such that $||\hat{\boldsymbol{\theta}}_{n,k}-\boldsymbol{\theta}_0||\to a_n$ as $k\to\infty$. Noting that the sequence is contained in a bounded set, there exists a subsequence that converges to $\hat{\boldsymbol{\theta}}_{n,0}$, say. Note that $||\hat{\boldsymbol{\theta}}_{n,0}-\boldsymbol{\theta}_0||=a_n$. Since $S_n(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, $S(\hat{\boldsymbol{\theta}}_{n,0})=0$. We define

$$\hat{\boldsymbol{\theta}}_n=\begin{cases}\hat{\boldsymbol{\theta}}_{n,0}, & \text{if } R_n\neq\emptyset;\\ 0, & R_n=\emptyset.\end{cases} \tag{1.4.4}$$

Now we show $\hat{\boldsymbol{\theta}}_n$ satisfies (a) and (b) of Theorem 1.2.1.

Since $\mathbf{E}_{\boldsymbol{\theta}_0}\{S_n(\boldsymbol{\theta})\}=\mathbf{E}_{\boldsymbol{\theta}_0}\{\Psi(Z,\boldsymbol{\theta})\}$ is differentiable at $\boldsymbol{\theta}_0$,

$$\mathbf{E}_{\boldsymbol{\theta}_0}\{S_n(\boldsymbol{\theta})\}-\mathbf{E}_{\boldsymbol{\theta}_0}\{S_n(\boldsymbol{\theta}_0)\}=\frac{\partial}{\partial\boldsymbol{\theta}^T}\mathbf{E}_{\boldsymbol{\theta}_0}\{S_n(\boldsymbol{\theta}_0)\}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+o(||\boldsymbol{\theta}-\boldsymbol{\theta}_0||). \tag{1.4.5}$$

Since $\mathbf{E}_{\boldsymbol{\theta}_0}\{S(\boldsymbol{\theta}_0)\}=0$,

$$(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\mathbf{E}_{\boldsymbol{\theta}_0}\{S_n(\boldsymbol{\theta})\}=(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\frac{\partial}{\partial\boldsymbol{\theta}^T}\mathbf{E}_{\boldsymbol{\theta}_0}\{S_n(\boldsymbol{\theta}_0)\}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^To(||\boldsymbol{\theta}-\boldsymbol{\theta}_0||). \tag{1.4.6}$$

Because $\partial \mathbf{E}_{\boldsymbol{\theta}_0}\{S_n(\boldsymbol{\theta}_0)\}/\partial \boldsymbol{\theta}^T < 0$, we have for sufficiently small $||\boldsymbol{\theta} - \boldsymbol{\theta}_0||$, the above formula (1.4.6) is less than 0. Let $\varepsilon > 0$ be so small such that (1.4.6) is less than 0 on $B(\boldsymbol{\theta}_0, \varepsilon) = \{\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| \leq \varepsilon\}$. Then

$$\sup_{\boldsymbol{\theta} \in \partial B(\boldsymbol{\theta}_0, \varepsilon)} [(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{E}_{\boldsymbol{\theta}_0}\{S_n(\boldsymbol{\theta})\}] < 0,$$

where $\partial B(\boldsymbol{\theta}_0, \varepsilon) = \{\boldsymbol{\theta} : ||\boldsymbol{\theta} - \boldsymbol{\theta}_0|| = \varepsilon\}$.

Based on the uniformly convergence of $S_n(\boldsymbol{\theta})$ to $\mathbf{E}_{\boldsymbol{\theta}_0}\{S_n(\boldsymbol{\theta})\}$ in a neighborhood of $\boldsymbol{\theta}_0$, we have with probability going to 1,

$$\sup_{\boldsymbol{\theta} \in \partial B(\boldsymbol{\theta}_0, \varepsilon)} [(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T S_n(\boldsymbol{\theta})] < 0,$$

Let $A_n = \{\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} : R_n \cap B(\boldsymbol{\theta}_0, \varepsilon) \neq \emptyset\}$. Then on A_n^c, $S_n(\boldsymbol{\theta}) = 0$ has no solution on $B(\boldsymbol{\theta}_0, \varepsilon)$. Define

$$f(\xi) = \frac{S_n(\boldsymbol{\theta}_0 + \varepsilon\xi)}{||S_n(\boldsymbol{\theta}_0 + \varepsilon\xi)||}, ||\xi|| \leq 1.$$

Then $f(\cdot)$ is a continuous function from the closed unit ball to itself. Based on the Brouwer fixed point theorem, we know there exists ξ^* such that $||\xi^*|| \leq 1$ and

$$f(\xi^*) = \xi^* = \frac{S_n(\boldsymbol{\theta}_0 + \varepsilon\xi^*)}{||S_n(\boldsymbol{\theta}_0 + \varepsilon\xi^*)||}.$$

Hence $f(\xi^*)^T\xi^* = \xi^{*T}\xi^*$. Let $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0 + \varepsilon\xi^*$. Then $\boldsymbol{\theta}^* \in B(\boldsymbol{\theta}_0, \varepsilon)$ and

$$\begin{aligned}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)^T S_n(\boldsymbol{\theta}^*) &= \varepsilon\xi^* S_n(\boldsymbol{\theta}_0 + \varepsilon\xi^*) = \varepsilon \frac{S_n(\boldsymbol{\theta}_0 + \varepsilon\xi^*)^T}{||S_n(\boldsymbol{\theta}_0 + \varepsilon\xi^*)||} S_n(\boldsymbol{\theta}_0 + \varepsilon\xi^*) \\ &= \varepsilon ||S_n(\boldsymbol{\theta}_0 + \varepsilon\xi^*)|| > 0.\end{aligned}$$

So, on A_n^c, $(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)^T S_n(\boldsymbol{\theta}^*) > 0$ and

$$C_n \triangleq \{((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) : (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)^T S_n(\boldsymbol{\theta}^*) < 0\} \subset A_n.$$

Note that $P(C_n) \to 1$. Therefore, $P(A_n) \to 1$ and, with probability going to 1, $S_n(\boldsymbol{\theta}) = 0$ has a solution in $B(\boldsymbol{\theta}_0, \epsilon)$ and the defined $\hat{\boldsymbol{\theta}}_n$ must also be in $B(\boldsymbol{\theta}_0, \epsilon)$ satisfying $S(\hat{\boldsymbol{\theta}}_n) = 0$. Therefore, $||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|| < \varepsilon$, and $P(||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|| < \varepsilon) \to 1$.

Proof of Theorem 1.2.2: Based on the Taylor expansion and condition A6, we have

$$0 = S_n(\hat{\boldsymbol{\theta}}) = S_n(\boldsymbol{\theta}_0) + \left\{ \frac{\partial S_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T} + o_p(1) \right\} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

Note that

$$\frac{\partial S_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \Psi(X, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \mathbf{E}_{\boldsymbol{\theta}_0}\left\{\frac{\partial \Psi(Z, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\} + o_p(1) = A + o_p(1).$$

Therefore, $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \{-A + o_p(1)\}^{-1} S_n(\boldsymbol{\theta}_0)$. Based on the central limit theorem, we have $\sqrt{n} S_n(\boldsymbol{\theta}_0) \xrightarrow{d} N(0, B)$, where $B = \mathbf{E}_{\boldsymbol{\theta}}\{\Psi(Z, \boldsymbol{\theta})\Psi(Z, \boldsymbol{\theta})^T\}$. Then by Slutsky's theorem, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = N(0, A^{-1}BA^{-1}).$$

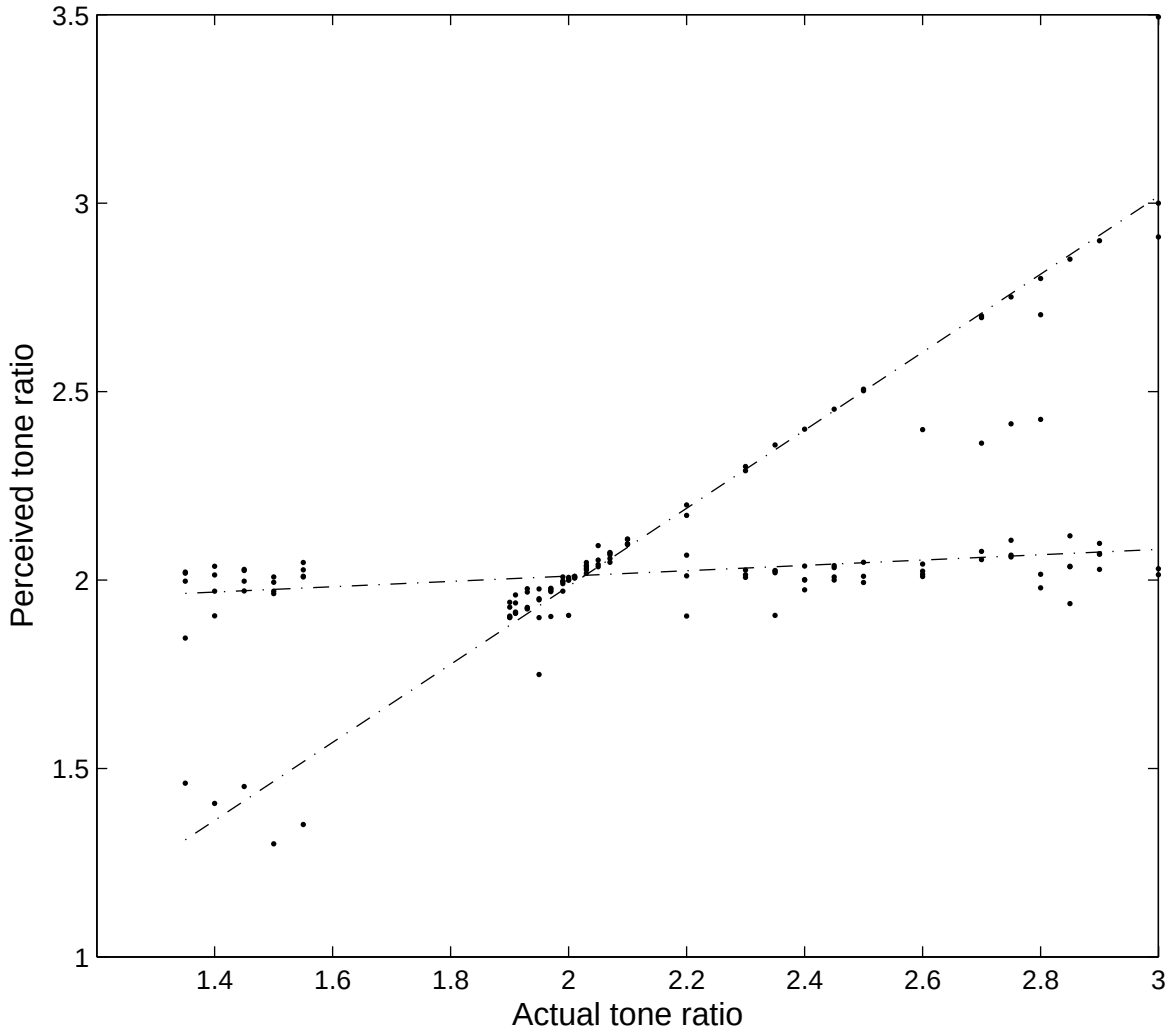

Figure 1.1: *The scatter plot of the tone perception data and the fitted two lines by our proposed method. The predictor is actual tone ratio and the response is the perceived tone ratio by a trained musician.*

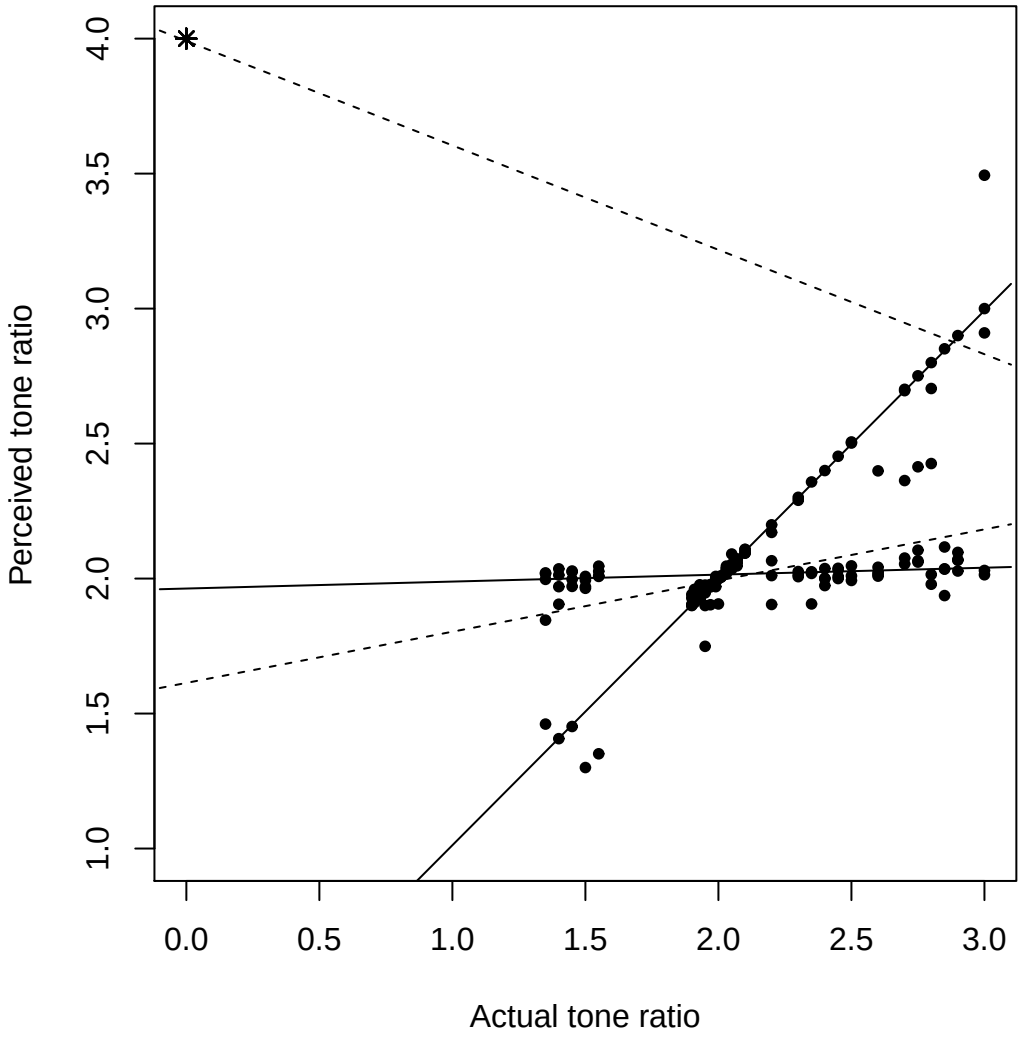

Figure 1.2: *Fitted mixture regression lines with added ten identical outliers* $(0, 4)$ *(denoted by stars at the upper left corner). The solid lines represent the fit by Robust-Bisquare and the dashed lines represent the fit by traditional MLE.*

Table 1.1: *Bias (Std) of Point Estimates for $n = 100$ in Example 1.*

Case	TRUE	MLE	Robust-Bisquare	Robust-Huber	TLE
I	$\beta_{10} : 0$	0.004(0.309)	-0.018(0.382)	0.015(0.357)	-0.005(0.657)
	$\beta_{20} : 0$	-0.005(0.158)	-0.006(0.220)	-0.005(0.180)	-0.044(0.431)
	$\beta_{11} : 1$	-0.026(0.328)	-0.120(0.492)	-0.080(0.449)	-0.814(0.831)
N(0,1)	$\beta_{21} : -1$	-0.002(0.143)	-0.020(0.207)	0.001(0.149)	0.001(0.238)
	$\beta_{12} : 1$	-0.013(0.318)	-0.119(0.499)	-0.044(0.415)	-0.839(0.867)
	$\beta_{22} : -1$	-0.016(0.138)	-0.008(0.187)	-0.012(0.156)	-0.014(0.205)
	$\pi_1 : 0.25$	0.014(0.071)	0.040(0.129)	0.020(0.074)	0.120(0.107)
II	$\beta_{10} : 0$	0.317(3.144)	-0.001(0.658)	-0.004(0.792)	-0.012(0.775)
	$\beta_{20} : 0$	0.123(2.304)	0.001(0.286)	0.001(0.268)	-0.004(0.319)
	$\beta_{11} : 1$	-0.231(2.519)	-0.181(0.781)	-0.137(0.831)	-0.432(0.761)
t_3	$\beta_{21} : -1$	-0.417(2.173)	-0.062(0.243)	-0.052(0.228)	-0.024(0.236)
	$\beta_{12} : 1$	0.169(2.764)	-0.179(0.765)	-0.048(0.814)	-0.417(0.744)
	$\beta_{22} : -1$	-0.343(2.048)	-0.064(0.275)	-0.066(0.261)	-0.038(0.270)
	$\pi_1 : 0.25$	0.091(0.298)	0.068(0.129)	0.051(0.104)	0.080(0.093)
III	$\beta_{10} : 0$	109.2(1597)	0.117(1.221)	-0.122(7.327)	-0.037(4.070)
	$\beta_{20} : 0$	33.79(412.1)	-0.018(0.837)	0.927(8.547)	-0.257(2.674)
	$\beta_{11} : 1$	131.6(1195)	0.264(1.057)	0.927(5.473)	0.101(3.967)
t_1	$\beta_{21} : -1$	-40.06(233.7)	-0.175(0.901)	-1.082(4.853)	-0.609(3.356)
	$\beta_{12} : 1$	62.25(449.6)	0.180(1.190)	1.751(6.132)	0.018(3.153)
	$\beta_{22} : -1$	-52.49(253.7)	-0.017(0.628)	-1.341(6.329)	-0.393(2.886)
	$\pi_1 : 0.25$	0.238(0.469)	0.133(0.184)	0.124(0.298)	0.120(0.267)
IV	$\beta_{10} : 0$	-0.118(2.307)	0.038(0.565)	0.019(0.514)	0.010(0.683)
	$\beta_{20} : 0$	-0.246(2.218)	-0.052(0.273)	-0.045(0.885)	-0.007(0.309)
$0.95N(0,1)$	$\beta_{11} : 1$	0.044(2.044)	-0.186(0.669)	-0.074(0.613)	-0.564(0.763)
$+0.05N(0,5^2)$	$\beta_{21} : -1$	-0.231(1.668)	0.002(0.187)	0.018(0.349)	0.028(0.215)
	$\beta_{12} : 1$	-0.095(2.240)	-0.102(0.623)	0.016(0.615)	-0.458(0.788)
	$\beta_{22} : -1$	-0.046(1.379)	-0.040(0.185)	-0.073(0.473)	-0.007(0.219)
	$\pi_1 : 0.25$	0.064(0.283)	0.055(0.118)	0.037(0.110)	0.071(0.094)
V	$\beta_{10} : 0$	0.175(2.088)	-0.006(0.870)	0.163(1.569)	0.054(0.722)
	$\beta_{20} : 0$	0.011(0.165)	0.009(0.197)	0.010(0.142)	0.006(0.283)
5% high leverage	$\beta_{11} : 1$	1.501(1.541)	0.185(0.994)	1.608(0.971)	0.240(1.027)
outliers	$\beta_{21} : -1$	0.193(0.192)	0.008(0.151)	0.107(0.156)	-0.009(0.164)
	$\beta_{12} : 1$	1.487(1.543)	0.189(0.865)	1.380(0.975)	-0.172(0.937)
	$\beta_{22} : -1$	-0.216(0.191)	-0.004(0.177)	0.119(0.163)	-0.015(0.176)
	$\pi_1 : 0.25$	-0.095(0.034)	0.003(0.102)	-0.073(0.037)	0.041(0.096)

Table 1.2: *Bias (Std) of Point Estimates for $n = 400$ in Example 1.*

Case	TRUE	MLE	Robust-Bisquare	Robust-Huber	TLE
I	$\beta_{10} : 0$	0.013(0.135)	0.013(0.136)	0.012(0.134)	0.020(0.396)
	$\beta_{20} : 0$	-0.002(0.062)	-0.001(0.065)	-0.001(0.065)	-0.005(0.248)
	$\beta_{11} : 1$	-0.010(0.131)	-0.009(0.139)	-0.008(0.141)	-0.437(0.615)
$N(0,1)$	$\beta_{21} : -1$	0.005(0.063)	0.003(0.061)	0.003(0.061)	0.020(0.075)
	$\beta_{12} : 1$	0.021(0.119)	0.025(0.127)	0.022(0.128)	0.435(0.626)
	$\beta_{22} : -1$	-0.002(0.068)	-0.003(0.070)	-0.002(0.070)	0.017(0.086)
	$\pi_1 : 0.25$	0.007(0.033)	0.009(0.033)	0.009(0.033)	0.035(0.083)
II	$\beta_{10} : 0$	-0.053(3.055)	0.002(0.206)	0.009(0.214)	-0.031(0.230)
	$\beta_{20} : 0$	0.704(3.844)	-0.004(0.085)	-0.004(0.085)	-0.008(0.088)
	$\beta_{11} : 1$	0.279(2.425)	0.005(0.175)	0.038(0.182)	-0.141(0.257)
t_3	$\beta_{21} : -1$	-0.884(3.921)	-0.028(0.080)	-0.048(0.081)	-0.004(0.086)
	$\beta_{12} : 1$	-0.363(1.774)	0.026(0.201)	0.045(0.205)	-0.121(0.216)
	$\beta_{22} : -1$	-0.296(2.487)	-0.014(0.080)	-0.027(0.083)	0.007(0.079)
	$\pi_1 : 0.25$	0.058(0.285)	0.021(0.036)	0.020(0.036)	0.018(0.041)
III	$\beta_{10} : 0$	-100.5(981.6)	-0.097(0.590)	0.655(5.966)	0.066(1.496)
	$\beta_{20} : 0$	4.336(702.2)	0.021(0.156)	-0.282(4.237)	0.168(1.852)
$0.95N(0,1)$	$\beta_{11} : 1$	88.90(342.2)	-0.108(0.632)	1.197(4.321)	-0.100(1.044)
$+0.05N(0,5^2)$	$\beta_{21} : -1$	-111.2(425.4)	-0.105(0.304)	-0.074(1.860)	-0.107(1.025)
	$\beta_{12} : 1$	163.1(888.4)	-0.145(0.578)	0.557(2.669)	-0.130(1.087)
	$\beta_{22} : -1$	-71.85(564.8)	-0.043(0.288)	-0.372(2.191)	-0.044(0.923)
	$\pi_1 : 0.25$	0.210(0.492)	0.096(0.111)	0.037(0.195)	0.059(0.219)
IV	$\beta_{10} : 0$	0.237(2.103)	-0.006(0.162)	-0.004(0.182)	-0.001(0.330)
	$\beta_{20} : 0$	-0.348(2.096)	-0.006(0.069)	-0.007(0.071)	0.009(0.131)
$0.95N(0,1)$	$\beta_{11} : 1$	0.064(1.703)	-0.002(0.166)	0.028(0.161)	-0.213(0.371)
$+0.05N(0,5^2)$	$\beta_{21} : -1$	-0.004(0.503)	-0.002(0.070)	-0.011(0.073)	0.012(0.079)
	$\beta_{12} : 1$	-0.007(1.599)	0.008(0.151)	0.044(0.162)	-0.239(0.402)
	$\beta_{22} : -1$	-0.005(0.893)	0.001(0.065)	-0.011(0.067)	0.015(0.077)
	$\pi_1 : 0.25$	-0.001(0.212)	0.013(0.033)	0.012(0.033)	0.013(0.049)
V	$\beta_{10} : 0$	0.199(1.274)	0.084(0.401)	0.293(1.213)	0.007(0.230)
	$\beta_{20} : 0$	0.006(0.095)	-0.001(0.071)	0.007(0.079)	-0.001(0.082)
	$\beta_{11} : 1$	1.398(0.085)	0.165(0.488)	1.543(0.661)	0.143(0.212)
5% high leverage	$\beta_{21} : -1$	0.242(0.101)	0.006(0.071)	0.113(0.072)	-0.009(0.074)
outliers	$\beta_{12} : 1$	1.587(0.858)	0.183(0.594)	1.438(0.662)	-0.116(0.270)
	$\beta_{22} : -1$	0.254(0.098)	0.012(0.067)	0.014(0.065)	0.001(0.069)
	$\pi_1 : 0.25$	-0.100(0.020)	-0.016(0.038)	-0.074(0.021)	-0.002(0.036)

Table 1.3: *The average number of found solutions for Robust-Bisquare and Robust-Huber based on 22 initial values for Example 1.*

Case	n	Robust-Bisquare	Robust-Huber
I: N(0,1)	100	1.880	1.620
	400	1.330	1.040
II: t_3	100	2.465	2.500
	400	1.610	1.600
III: t_1	100	4.590	4.905
	400	3.920	4.930
IV: $0.95N(0,1) + 0.05N(0,5^2)$	100	2.140	2.035
	400	1.270	1.190
V: 5% high leverage outliers	100	4.440	3.360
	400	3.800	2.770

Table 1.4: *Bias (Std) of Point Estimates for $n = 100$ in Example 2.*

Case	TRUE	MLE	Robust-Bisquare	Robust-Huber	TLE
I	$\beta_{10} : 1$	-0.108(0.406)	-0.068(0.443)	-0.073(0.463)	-0.037(0.465)
	$\beta_{20} : 2$	-0.029(0.559)	0.105(0.567)	0.069(0.569)	0.191(0.604)
	$\beta_{30} : 3$	0.021(0.279)	0.004(0.285)	0.025(0.287)	0.031(0.350)
N(0,1)	$\beta_{11} : 1$	0.022(0.398)	0.068(0.410)	0.078(0.394)	0.346(0.494)
	$\beta_{21} : 2$	0.150(0.785)	0.215(0.756)	0.288(0.844)	0.243(0.919)
	$\beta_{31} : 5$	0.085(0.226)	0.032(0.224)	0.026(0.235)	-0.055(0.303)
	$\pi_1 : 0.3$	-0.003(0.110)	0.007(0.118)	0.008(0.118)	0.026(0.085)
	$\pi_2 : 0.3$	0.024(0.109)	0.011(0.105)	0.011(0.108)	0.021(0.074)
II	$\beta_{10} : 1$	-1.031(2.206)	-0.012(0.577)	-0.157(0.808)	-0.068(0.564)
	$\beta_{20} : 2$	1.032(2.587)	0.141(0.779)	0.178(0.981)	0.152(0.741)
	$\beta_{30} : 3$	0.546(4.015)	0.052(0.379)	0.071(0.426)	0.105(0.452)
t_3	$\beta_{11} : 1$	-0.724(4.654)	-0.005(0.580)	-0.091(0.730)	0.201(0.575)
	$\beta_{21} : 2$	0.361(1.950)	0.424(1.020)	0.258(1.041)	0.429(1.049)
	$\beta_{31} : 5$	1.310(3.588)	0.044(0.320)	0.085(0.360)	-0.113(0.478)
	$\pi_1 : 0.3$	0.026(0.234)	0.041(0.131)	0.016(0.129)	0.031(0.093)
	$\pi_2 : 0.3$	0.067(0.193)	-0.017(0.124)	0.009(0.123)	0.012(0.088)
III	$\beta_{10} : 1$	-18.38(159.7)	-0.014(1.472)	-2.380(11.67)	-0.818(2.663)
	$\beta_{20} : 2$	857.4(9512)	0.472(1.629)	1.926(5.704)	0.717(2.166)
	$\beta_{30} : 3$	13.77(305.1)	0.097(1.478)	1.696(8.679)	0.628(2.326)
t_1	$\beta_{11} : 1$	-40.96(173.9)	-0.011(1.821)	1.561(8.171)	-0.445(2.842)
	$\beta_{21} : 2$	-739.0(8931)	0.361(1.394)	-0.365(4.356)	0.359(1.823)
	$\beta_{31} : 5$	84.69(359.4)	0.205(1.228)	2.121(6.471)	0.393(2.091)
	$\pi_1 : 0.3$	-0.013(0.323)	0.111(0.174)	0.037(0.231)	0.028(0.193)
	$\pi_2 : 0.3$	0.185(0.357)	-0.079(0.166)	0.060(0.196)	0.061(0.177)
IV	$\beta_{10} : 1$	-0.445(5.098)	-0.032(0.516)	-0.258(1.153)	-0.087(0.510)
	$\beta_{20} : 2$	0.845(2.284)	0.109(0.692)	0.091(0.843)	0.161(0.558)
	$\beta_{30} : 3$	0.330(3.579)	0.019(0.278)	0.078(0.492)	0.034(0.357)
$0.95N(0,1)$	$\beta_{11} : 1$	2.226(24.73)	0.066(0.455)	0.001(0.668)	0.288(0.469)
$+0.05N(0,5^2)$	$\beta_{21} : 2$	0.244(2.162)	0.283(0.776)	0.211(0.922)	0.256(0.956)
	$\beta_{31} : 5$	0.944(2.645)	0.016(0.251)	0.066(0.436)	-0.061(0.373)
	$\pi_1 : 0.3$	0.017(0.237)	0.041(0.128)	0.014(0.131)	0.031(0.084)
	$\pi_2 : 0.3$	0.079(0.197)	-0.023(0.132)	0.011(0.127)	0.016(0.081)
V	$\beta_{10} : 1$	0.465(0.209)	0.114(0.454)	0.459(0.235)	-0.064(0.463)
	$\beta_{20} : 2$	0.936(0.233)	0.307(0.600)	0.938(0.256)	0.244(0.723)
	$\beta_{30} : 3$	-2.624(3.700)	-0.224(1.038)	-1.452(2.409)	-0.098(0.844)
5% high leverage	$\beta_{11} : 1$	0.463(0.222)	0.188(0.386)	0.444(0.263)	0.233(0.467)
outliers	$\beta_{21} : 2$	2.922(0.238)	0.569(1.334)	2.918(0.351)	0.275(0.909)
	$\beta_{31} : 5$	4.981(0.185)	0.381(1.331)	4.927(0.121)	0.087(0.779)
	$\pi_1 : 0.3$	0.244(0.065)	0.058(0.131)	0.241(0.071)	0.046(0.099)
	$\pi_2 : 0.3$	0.067(0.063)	-0.005(0.119)	0.068(0.067)	0.007(0.092)

Table 1.5: *Bias (Std) of Point Estimates for $n = 400$ in Example 2.*

Case	TRUE	MLE	Robust-Bisquare	Robust-Huber	TLE
I	$\beta_{10} : 1$	-0.053(0.204)	0.064(0.217)	0.064(0.214)	0.108(0.254)
	$\beta_{20} : 2$	0.045(0.196)	0.040(0.208)	0.067(0.211)	0.240(0.242)
	$\beta_{30} : 3$	0.006(0.098)	0.007(0.103)	0.007(0.103)	0.027(0.207)
N(0,1)	$\beta_{11} : 1$	0.010(0.187)	0.007(0.187)	0.014(0.187)	0.304(0.268)
	$\beta_{21} : 2$	0.004(0.176)	0.011(0.181)	0.032(0.184)	-0.138(0.483)
	$\beta_{31} : 5$	0.019(0.085)	0.015(0.091)	0.015(0.090)	-0.053(0.150)
	$\pi_1 : 0.3$	-0.003(0.059)	-0.002(0.059)	-0.004(0.059)	0.020(0.050)
	$\pi_2 : 0.3$	0.004(0.063)	0.003(0.063)	0.004(0.062)	0.012(0.050)
II	$\beta_{10} : 1$	-0.949(4.354)	-0.129(0.452)	-0.243(0.429)	-0.214(0.324)
	$\beta_{20} : 2$	1.604(4.427)	0.131(0.453)	0.165(0.573)	0.218(0.317)
	$\beta_{30} : 3$	0.506(7.373)	0.018(0.122)	0.030(0.137)	0.009(0.164)
t_3	$\beta_{11} : 1$	-0.698(4.114)	0.082(0.298)	0.009(0.645)	0.242(0.280)
	$\beta_{21} : 2$	-0.058(3.883)	0.064(0.356)	0.028(0.545)	-0.058(0.378)
	$\beta_{31} : 5$	2.161(6.046)	0.027(0.123)	0.056(0.122)	-0.034(0.134)
	$\pi_1 : 0.3$	0.024(0.275)	0.025(0.094)	0.008(0.094)	0.014(0.057)
	$\pi_2 : 0.3$	0.095(0.215)	-0.022(0.088)	-0.001(0.090)	0.009(0.056)
III	$\beta_{10} : 1$	105.6(1066)	0.078(1.117)	-7.375(11.74)	1.804(2.506)
	$\beta_{20} : 2$	185.3(1106)	0.135(0.818)	1.749(7.543)	0.378(1.658)
	$\beta_{30} : 3$	460.8(2960)	-0.010(1.013)	2.829(8.789)	0.436(1.717)
t_1	$\beta_{11} : 1$	-375.4(1443)	0.307(0.743)	-0.611(0.654)	0.545(1.529)
	$\beta_{21} : 2$	-130.0(796.0)	0.302(1.081)	-0.772(6.175)	0.381(1.617)
	$\beta_{31} : 5$	705.9(2646)	0.057(0.471)	0.524(3.727)	0.091(0.888)
	$\pi_1 : 0.3$	-0.026(0.295)	0.154(0.130)	-0.066(0.243)	-0.011(0.230)
	$\pi_2 : 0.3$	0.181(0.301)	-0.148(0.133)	0.138(0.160)	0.084(0.179)
IV	$\beta_{10} : 1$	-2.045(4.149)	-0.020(0.255)	-0.204(0.955)	-0.084(0.292)
	$\beta_{20} : 2$	0.787(2.473)	0.063(0.245)	0.143(0.511)	0.220(0.292)
	$\beta_{30} : 3$	0.739(3.728)	0.010(0.121)	0.019(0.123)	-0.001(0.151)
$0.95N(0,1)$	$\beta_{11} : 1$	-0.339(3.860)	0.032(0.205)	0.035(0.328)	0.293(0.263)
$+0.05N(0,5^2)$	$\beta_{21} : 2$	0.273(2.249)	0.053(0.242)	-0.063(0.434)	-0.050(0.389)
	$\beta_{31} : 5$	1.055(3.095)	-0.007(0.098)	0.013(0.096)	-0.035(0.132)
	$\pi_1 : 0.3$	-0.034(0.279)	0.019(0.077)	0.001(0.083)	0.023(0.055)
	$\pi_2 : 0.3$	0.148(0.186)	-0.020(0.082)	0.001(0.087)	0.001(0.062)
V	$\beta_{10} : 1$	0.459(0.093)	0.092(0.212)	0.459(0.107)	-0.102(0.256)
	$\beta_{20} : 2$	0.966(0.104)	0.069(0.232)	0.968(0.106)	0.171(0.299)
	$\beta_{30} : 3$	-2.945(2.395)	0.092(0.113)	-1.724(1.856)	-0.008(0.124)
5% high leverage	$\beta_{11} : 1$	0.482(0.108)	0.042(0.244)	0.468(0.126)	0.204(0.261)
outliers	$\beta_{21} : 2$	2.916(0.099)	0.126(0.829)	2.936(0.097)	-0.104(0.237)
	$\beta_{31} : 5$	4.996(0.119)	0.021(0.477)	4.936(0.092)	-0.040(0.118)
	$\pi_1 : 0.3$	0.235(0.031)	0.021(0.081)	0.235(0.030)	0.011(0.056)
	$\pi_2 : 0.3$	0.083(0.031)	0.007(0.083)	0.083(0.030)	-0.006(0.059)

Table 1.6: *The average number of the found solutions for Robust-Bisquare and Robust-Huber based on 22 initial values for Example 2.*

Case	n	Robust-Bisquare	Robust-Huber
I: N(0,1)	100	3.370	3.400
	400	2.380	2.290
II: t_3	100	3.690	4.055
	400	2.920	3.460
III: t_1	100	5.635	5.465
	400	5.620	5.930
IV: $0.95N(0,1) + 0.05N(0,5^2)$	100	3.540	3.665
	400	2.690	3.180
V: 5% high leverage outliers	100	5.600	3.740
	400	5.200	3.400

Chapter 2

Robust Mixture of Linear Mixed Models Using Multivariate t Distribution

2.1 Introduction

Linear mixed models have been widely applied in many disciplines, including agriculture, genetics, marketing, and industrial statistics, where multiple correlated measurements are made on each unit of interest. Random-effects models were first introduced in Fisher, R. A., (1918). The best linear unbiased estimates (BLUE) of fixed effects and the best linear unbiased predictions (BLUP) of random effects have been extensively studied (Henderson, et al. 1959; Hartley, 1967; Robinson, 1991; Mclean, et al. 1991).Subsequently, mixed modeling has become a major area of statistical research.

The classical linear mixed model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{b} + \boldsymbol{\epsilon} \tag{2.1.1}$$

where $\mathbf{y}$ is the $N \times 1$ response vector, $\mathbf{X}$ is the $N \times p$ design matrix for the fixed effects, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed-effect coefficients, $\mathbf{U}$ is the $N \times q$ design matrix for the random effects, $\mathbf{b} \sim N_q(0, \boldsymbol{\Psi})$ is the $q \times 1$ vector of random effect coefficients, and $\boldsymbol{\epsilon}$ is the $N \times 1$ vector of errors for observations and assumed to have multivariate normal distribution with mean 0 and variance $\boldsymbol{\Lambda}$. Based on the above model setup,

$\mathbf{X}\boldsymbol{\beta}$ models the fixed effects and $\mathbf{Ub}$ models the random effects. It follows that $\mathbf{y}$ has a multivariate normal distribution with mean $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\mathbf{V} = \text{cov}(\mathbf{y}) = \mathbf{U}\boldsymbol{\Psi}\mathbf{U}^T + \boldsymbol{\Lambda}$. The main goal here is still to model the relationship between some response variable and some predictor variables. Linear mixed models are thus considered important extensions of the conventional linear regression models for handling dependent data, which arise in various problems, e.g., when the observations are taken on groups of related individuals, or when repeated measurements are made over time on the same set of individuals. For clarity and without loss of generality, in the sequel we shall mainly refer to the repeated measurement setup when presenting our proposed methodology, similar to Celeux et al. (2005).

In many applications, however, the underlying assumption that the regression relationship is homogeneous across all the subjects could be violated. Of particular interest is the situation that the subjects may form several distinct clusters, indicating mixed regression relationships. Such heterogeneity can be modeled by a finite mixture regression model, consisting of, say, m homogeneous groups/components. Suppose there are I subjects under study, and n_i repeated measurements are gathered on the ith subjects, for $i = 1, \dots I$. We consider a mixture linear mixed model setup as follows. For each $i = 1, \dots I$, let Z_i be a latent variable with $P(Z_i = j) = \pi_j$, $j = 1, \dots m$. Given $Z_i = j$, we assume that the response $\mathbf{y}_i \in \mathbb{R}^{n_i}$ follows a linear mixed model, i.e.,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_j + \mathbf{U}_i\mathbf{b}_{ij} + \mathbf{e}_{ij} \tag{2.1.2}$$

where $\mathbf{X}_i \in \mathbb{R}^{n_i \times p}$ is the fixed-effect covariate matrix, $\boldsymbol{\beta}_j \in \mathbb{R}^p$ a fixed-effect coefficient vector, $\mathbf{U}_i \in \mathbb{R}^{n_i \times q}$ the random-effect covariate matrix, $\mathbf{b}_{ij}$ the random-effect coefficient vector which is thought as random, and $\mathbf{e}_{ij}$ the random error vector. Following the conventional formulations of the normal mixture model and the mixed model, it is natural to assume that

$$\mathbf{b}_{ij} \sim N_q(0, \boldsymbol{\Psi}_j), \qquad \mathbf{e}_{ij} \sim N_{n_i}(0, \boldsymbol{\Lambda}_{ij}),$$

and all $\mathbf{b}_{ij}$s, $\mathbf{e}_{ij}$s, for $i = 1, \dots, I$ and $j = 1, \dots, m$ are independent. Usually each error covariance matrix $\boldsymbol{\Lambda}_{ij}$ is assumed to be dependent on i only through its dimension, e.g., an AR(1) correlation structure with some correlation parameter ρ so that $\boldsymbol{\Lambda}_{ij} = \boldsymbol{\Lambda}(\rho, i)$. The correlation structure among each n_i observations on subject i is induced and modeled by the random component $\mathbf{U}_i\mathbf{b}_{ij}$. Conditional on $Z_i = j$, the joint

distribution of $(\mathbf{y}_i, \mathbf{b}_{ij})$ is

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{b}_{ij} \end{pmatrix} \Big| Z_i = j \sim N_{n_i+q}\left(\begin{pmatrix} \mathbf{X}_i\boldsymbol{\beta}_j \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{U}_i\boldsymbol{\Psi}_j\mathbf{U}_i^T + \boldsymbol{\Lambda}_{ij} & \mathbf{U}_i\boldsymbol{\Psi}_j \\ \boldsymbol{\Psi}_j\mathbf{U}_i^T & \boldsymbol{\Psi}_j \end{pmatrix} \right), \tag{2.1.3}$$

and the mixture distribution of $\mathbf{y}_i$ itself, without observing Z_i, is

$$\mathbf{y}_i \sim \sum_{j=1}^{m} \pi_j N_{n_i}(\mathbf{X}_i\boldsymbol{\beta}_j, \mathbf{U}_i\boldsymbol{\Psi}_j\mathbf{U}_i^T + \boldsymbol{\Lambda}_{ij}). \tag{2.1.4}$$

Although the above normal mixture linear mixed model is quite appealing in modeling the regression relationship with the aforementioned hierarchically clustered data, one potential drawback of the model is that it can be very sensitive to outliers in the observations, an undesirable property inherited from the normal mixture model. Motivated by Lange et al. (1989), Welsh and Richardson (1997) and Pinheiro et al. (2001),we propose a new mixture linear mixed model by replacing the normal distribution with multivariate t distribution. For each mixture component, we assume the response and the random effects jointly follow a multivariate t distribution, in a similar fashion as (2.1.3), to conveniently robustify the estimation procedure. An efficient generalized EM algorithm is developed for conducting maximum likelihood estimation. The degrees of freedom parameters of the t distributions are chosen data adaptively for achieving flexible tradeoff between estimation robustness and efficiency. We demonstrate via simulation study that the proposed approach is indeed robust and can be much more efficient than the traditional normal mixture model when outliers are present in the data, and in the absence of outliers the proposed approach leads to comparable performance to that of the normal mixture model. An application on lung growth of children further showcases the efficacy of the proposed approach.

2.2 Robust t-Mixture Linear Mixed Models

2.2.1 The t-mixture of linear mixed models

In practice, outliers and anomalies are bounded to occur, and failure to accommodate outliers may put both the model estimation and inference in jeopardy. This motivates us to propose a robust t-mixture of linear mixed models. Given $Z_i = j$, we start by assuming that the joint distribution of $(\mathbf{y}_i, \mathbf{b}_{ij})$ is

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{b}_{ij} \end{pmatrix} \mid Z_i = j \sim t_{n_i+q}\left(\begin{pmatrix} \mathbf{X}_i\boldsymbol{\beta}_j \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{U}_i\boldsymbol{\Psi}_j\mathbf{U}_i^T + \boldsymbol{\Lambda}_{ij} & \mathbf{U}_i\boldsymbol{\Psi}_j \\ \boldsymbol{\Psi}_j\mathbf{U}_i^T & \boldsymbol{\Psi}_j \end{pmatrix}, \nu_j \right), \tag{2.2.1}$$

where we use $t_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ to denotes a n-dimensional multivariate t distribution with mean vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$ and degrees of freedom ν; in the sequel we use $t_n(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ to denote its probability density function. Throughout, the error covariance matrices are assumed to take the form $\boldsymbol{\Lambda}_{ij} = \sigma_j^2 \mathbf{R}_i$, for $i = 1, \ldots, I$, $j = 1, \ldots, m$, where $\mathbf{R}_i$ are known matrices taken to be the identity matrix, unless otherwisely noted.

The proposed approach essentially assumes that $\mathbf{y}_i$ follows a mixture distribution,

$$\mathbf{y}_i \sim \sum_{j=1}^{m} \pi_j t_{n_i}(\mathbf{X}_i \boldsymbol{\beta}_j, \mathbf{U}_i \boldsymbol{\Psi}_j \mathbf{U}_i^T + \boldsymbol{\Lambda}_{ij}, \nu_j), \tag{2.2.2}$$

and given the data for $i = 1, \ldots, I$, the log-likelihood function is

$$\sum_{i=1}^{I} \ln \left\{ \sum_{j=1}^{m} \pi_j t_{n_i}(\mathbf{y}_i; \mathbf{X}_i \boldsymbol{\beta}_j, \mathbf{U}_i \boldsymbol{\Psi}_j \mathbf{U}_i^T + \boldsymbol{\Lambda}_{ij}, \nu_j) \right\}. \tag{2.2.3}$$

Comparing to model (2.1.4), we have used the multivariate t distribution to replace the multivariate normal distribution, following similar idea in Lange et al. (1989). This extension allows us to carry out the mixture mixed-effect model analysis for the data involving errors with longer-than-normal tails. The degrees of freedom parameter of the t distribution is allowed to be chosen data adaptively, and it provides a convenient way for achieving flexible tradeoff between robustness and efficiency, i.e., in the special case $\nu = 1$, the distribution becomes a multivariate Cauchy distribution, and as $\nu \to \infty$, the distribution rolls back to the multivariate normal. Also note that in the above model we have directly specified the distribution of $\mathbf{y}_i$ as the multivariate t, instead of separately specifying the distributions of the random effects and the error terms, as the latter is unnecessary and may lead to untractable or inconvenient marginal distribution of $\mathbf{y}_i$.

To understand better about model (2.2.2), we shall discuss several of its alternative representations, which may ultimately facilitate the model estimation using maximum likelihood method, to be elaborated in the next section. It is known that the multivariate t distribution can be written as a normal scale mixture distribution, i.e., its probability density function $t(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ can be expressed as

$$t(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \int_0^\infty f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}/u) g(u; \frac{\nu}{2}, \frac{\nu}{2}) du,$$

where f denotes the normal density and g the Gamma density. In light of the above characterization, it is

convenient to express model (2.2.2) as a hierarchical model,

$$\begin{aligned}\mathbf{y}_i \mid b_{ij}, \tau_{ij}, j = 1, \ldots, m &\sim \sum_{j=1}^{m} \pi_j N(\mathbf{X}_i\boldsymbol{\beta}_j + \mathbf{U}_i\mathbf{b}_{ij}, \frac{1}{\tau_{ij}}\boldsymbol{\Lambda}_{ij}),\\ b_{ij} \mid \tau_{ij} &\sim N(\mathbf{0}, \frac{1}{\tau_{ij}}\boldsymbol{\Psi}_j), \text{ for } j = 1, \ldots, m,\\ \tau_{ij} &\sim \text{Gamma}(\frac{\nu_j}{2}, \frac{\nu_j}{2}), \text{ for } j = 1, \ldots, m.\end{aligned}$$

Model (2.2.2) could also be written in a conventional mixed-model form. Given $Z_i = j$,

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_j + \mathbf{U}_i\mathbf{b}_{ij} + \mathbf{e}_{ij}, \qquad i = 1, \ldots, I,$$

where $\mathbf{b}_{ij} \sim t_q(\mathbf{0}, \boldsymbol{\Psi}_j, \nu_j)$, and $\mathbf{e}_{ij} \sim t_{n_i}(\mathbf{0}, \boldsymbol{\Lambda}_{ij}, \nu_j)$. Conditional on τ_{ij}, $\mathbf{b}_{ij}$ is independent of $\mathbf{e}_{ij}$, which means that in general $\mathbf{b}_{ij}$ and $\mathbf{e}_{ij}$ are uncorrelated but not independent, for any $\nu_j < \infty$. It is now clear that in our proposed method, both $\mathbf{b}_{ij}$ and $\mathbf{e}_{ij}$ follow multivariate t distribution, and thus the method is robust against potential outliers in both the random effects or the within-subject random errors.

By integrating out $\mathbf{b}_{ij}$, the hierarchical model can be equivalently expressed as

$$\begin{aligned}\mathbf{y}_i \mid \tau_{ij}, j = 1, \ldots, m &\sim \sum_{j=1}^{m} \pi_j N(\mathbf{X}_i\boldsymbol{\beta}_j, \frac{1}{\tau_{ij}}(\mathbf{U}_i\boldsymbol{\Psi}_j\mathbf{U}_i^T + \boldsymbol{\Lambda}_{ij})),\\ \tau_{ij} &\sim \text{Gamma}(\frac{\nu_j}{2}, \frac{\nu_j}{2}), \text{ for } j = 1, \ldots, m.\end{aligned}$$

The conditional distribution of τ_{ij} can then be readily derived,

$$\tau_{ij} \mid \mathbf{y}_i, Z_i = j \sim \text{Gamma}\left(\frac{\nu_j + n_i}{2}, \frac{\nu_j + \delta_{ij}^2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}_j, \sigma_j^2)}{2}\right),$$

where

$$\delta_{ij}^2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}_j, \sigma_j^2) = (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j)^T(\mathbf{U}_i\boldsymbol{\Psi}_j\mathbf{U}_i^T + \boldsymbol{\Lambda}_{ij})^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j). \tag{2.2.4}$$

Therefore,

$$E(\tau_{ij} \mid \mathbf{y}_i, Z_i = j) = \frac{\nu_j + n_i}{\nu_j + \delta_{ij}^2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}_j, \sigma_j^2)}. \tag{2.2.5}$$

The above results will be useful in the proposed generalized EM algorithm in next section.

2.2.2 An efficient generalized EM algorithm for maximum likelihood estimation

We propose to conduct maximum likelihood estimation and inference of the proposed robust t-mixture linear mixed model. Direct maximization of the log-likelihood function (2.2.3) constructed from mixture

multivariate t distributions is quite difficult. In this section, we derive an efficient generalized EM algorithm to solve the problem, extending the work by Pinheiro et al. (2001) in the context of linear mixed model. The EM algorithm is commonly applied in problems with missing or incomplete data, which is particularly suitable here, in view of the alternative hierarchical model representation of the t-mixture model discussed in the previous section.

Let $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_I\}, \mathbf{b} = \{\mathbf{b}_{ij}; i = 1, \ldots, I, j = 1, \ldots, m\}$, and $\boldsymbol{\tau} = \{\tau_{ij}; i = 1, \ldots, I, j = 1, \ldots, m\}$. Let

$$Z_{ij} = \begin{cases} 1 & \text{if the } i\text{th subject is from the } j\text{th component,} \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathbf{Z} = \{Z_{ij}; i = 1, \ldots, I, j = 1, \ldots, m\}$. Similarly, let $\boldsymbol{\pi} = \{\pi_j; j = 1, \ldots, m\}$, $\boldsymbol{\beta} = \{\boldsymbol{\beta}_j; j = 1, \ldots, m\}$, $\boldsymbol{\Psi} = \{\boldsymbol{\Psi}_j; j = 1, \ldots, m\}$, $\boldsymbol{\sigma}^2 = \{\sigma_j^2; j = 1, \ldots, m\}$, and $\boldsymbol{\nu} = \{\nu_j; j = 1, \ldots, m\}$.

In our problem, $\mathbf{y}$ consists of the observed data and $(\mathbf{b}, \boldsymbol{\tau}, \mathbf{Z})$ can be treated as the missing data. Based on the hierarchical model formulation, the likelihood of the complete data $(\mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z})$ given the covariates is,

$$\prod_{i=1}^{I}\prod_{j=1}^{m}\left\{\pi_j f(\mathbf{y}_i; \mathbf{X}_i\boldsymbol{\beta}_j + \mathbf{U}_i\mathbf{b}_{ij}, \frac{1}{\tau_{ij}}\boldsymbol{\Lambda}_{ij}) f(\mathbf{b}_{ij}; \mathbf{0}, \frac{1}{\tau_{ij}}\boldsymbol{\Psi}_j) g(\tau_{ij}; \frac{\nu_j}{2}, \frac{\nu_j}{2})\right\}^{Z_{ij}}.$$

It follows that the complete log-likelihood function is

$$\begin{aligned}
&\ell(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\sigma}^2, \boldsymbol{\nu} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \\
=&\sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij}\ln(\pi_j) \\
&+ \sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij}\left\{-\frac{1}{2}\ln|\frac{1}{\tau_{ij}}\sigma_j^2\mathbf{R}_i| - \frac{1}{2}\mathbf{E}_{ij}^T\left(\frac{1}{\tau_{ij}}\sigma_j^2\mathbf{R}_i\right)^{-1}\mathbf{E}_{ij} + \text{const}\right\} \\
&+ \sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij}\left\{-\frac{1}{2}\ln|\frac{1}{\tau_{ij}}\boldsymbol{\Psi}_j| - \frac{1}{2}(\mathbf{b}_{ij})^T\left[\frac{1}{\tau_{ij}}\boldsymbol{\Psi}_j\right]^{-1}\mathbf{b}_{ij} + \text{const}\right\} \\
&+ \sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij}\left\{\left(\frac{\nu_j}{2} - 1\right)\ln(\tau_{ij}) - \frac{\tau_{ij}}{2}\nu_j - \ln\left(\Gamma\left(\frac{\nu_j}{2}\right)\right) + \frac{\nu_j}{2}\ln\left(\frac{\nu_j}{2}\right)\right\},
\end{aligned}$$

where $\mathbf{E}_{ij} = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j - \mathbf{U}_i\mathbf{b}_{ij}$, and we have used the setting that $\boldsymbol{\Lambda}_{ij} = \sigma_j^2\mathbf{R}_i$. We shall separate the above log-likelihood function into four parts, based on the parameters involved, i.e., let

$$\begin{aligned}
\ell(\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\sigma}^2, \boldsymbol{\nu} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) =& \ell_0(\boldsymbol{\pi} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) + \ell_1(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \\
&+ \ell_2(\boldsymbol{\Psi} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) + \ell_3(\boldsymbol{\nu} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}),
\end{aligned}$$

where

$$
\begin{aligned}
\ell_0(\boldsymbol{\pi} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) =& \sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \ln(\pi_j), \\
\ell_1(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) =& \sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \left(\left\{ -\frac{n_i}{2} \ln \sigma_j^2 - \frac{\tau_{ij}}{2\sigma_j^2} \mathbf{E}_{ij}^T \boldsymbol{R_i}^{-1} \mathbf{E}_{ij} \right\} \right) \\
=& -\sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \frac{n_i}{2} \ln \sigma_j^2 \\
& -\sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \left[\frac{\tau_{ij}}{2\sigma_j^2} \mathrm{tr} \left\{ \mathbf{R}_i^{-1} (\mathbf{y}_i - \mathbf{U}_i \mathbf{b}_{ij})(\mathbf{y}_i - \mathbf{U}_i \mathbf{b}_{ij})^T \right\} \right] \\
& +\sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \left\{ \frac{\tau_{ij}}{\sigma_j^2} \boldsymbol{\beta}_j^T \mathbf{X}_i^T \mathbf{R}_i^{-1} (\mathbf{y}_i - \mathbf{U}_i \mathbf{b}_{ij}) \right\} \\
& -\sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \left(\frac{\tau_{ij}}{2\sigma_j^2} \boldsymbol{\beta}_j^T \mathbf{X}_i^T \mathbf{R}_i^{-1} \mathbf{X}_i \boldsymbol{\beta}_j \right), \\
\ell_2(\boldsymbol{\Psi} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) =& \sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \left(-\frac{1}{2} \ln |\boldsymbol{\Psi}_j| \right) - \frac{1}{2} \sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \left(\tau_{ij} \mathbf{b}_{ij}^T \boldsymbol{\Psi}_j^{-1} \mathbf{b}_{ij} \right) \\
=& -\frac{1}{2} \sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \ln |\boldsymbol{\Psi}_j| - \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Psi}_j^{-1} \sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \tau_{ij} \mathbf{b}_{ij} \mathbf{b}_{ij}^T \right),
\end{aligned}
$$

and

$$
\ell_3(\boldsymbol{\nu} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) = \sum_{i=1}^{I}\sum_{j=1}^{m} Z_{ij} \left[\left\{ \frac{\nu_j}{2} \left(\ln(\frac{\nu_j}{2}) + \ln(\tau_{ij}) - \tau_{ij} \right) - \ln(\tau_{ij}) - \ln \left(\Gamma(\frac{\nu_j}{2}) \right) \right\} \right].
$$

Let $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\sigma}^2, \boldsymbol{\nu})$, collecting all the unknown parameters. Given $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, we now derive the expected complete data log-likelihood, $E\{\ell(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \mid \mathbf{y}, \widehat{\boldsymbol{\theta}}\}$, with respect to the missing data $(\mathbf{b}, \boldsymbol{\tau}, \mathbf{Z})$ conditional on the observed data $\mathbf{y}$, which simplifies to the calculations of the following quantities,

$$
\begin{aligned}
p_{ij} &= E(Z_{ij} = 1 \mid \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}), \\
\widehat{\tau}_{ij} &= E(\tau_{ij} \mid \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}, Z_{ij} = 1), \\
\widehat{\mathbf{b}}_{ij} &= E(\mathbf{b}_{ij} \mid \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}, Z_{ij} = 1, \tau_{ij}), \\
\widehat{\boldsymbol{\Omega}}_{ij} &= \tau_{ij} \mathrm{cov}(\mathbf{b}_{ij} \mid \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \mathbf{y}, Z_{ij} = 1, \tau_{ij}).
\end{aligned}
$$

From (2.2.2), it is easy to show that

$$
p_{ij} = \frac{\pi_j t_{n_i}(\mathbf{y}_i; \mathbf{X}_i \widehat{\boldsymbol{\beta}}_j, \mathbf{U}_i \widehat{\boldsymbol{\Psi}}_j \mathbf{U}_i^T, \widehat{\nu}_j)}{\sum_{j=1}^{m} \pi_j t_{n_i}(\mathbf{y}_i; \mathbf{X}_i \widehat{\boldsymbol{\beta}}_j, \mathbf{U}_i \widehat{\boldsymbol{\Psi}}_j \mathbf{U}_i^T, \widehat{\nu}_j)}. \tag{2.2.6}
$$

By (2.2.5), we have

$$\widehat{\tau}_{ij} = \frac{\widehat{\nu}_j + n_i}{\widehat{\nu}_j + \delta_{ij}^2(\boldsymbol{\beta}_j, \widehat{\boldsymbol{\Psi}}_j, \widehat{\sigma}_j^2)}, \tag{2.2.7}$$

where $\delta_{ij}^2(\boldsymbol{\beta}_j, \widehat{\boldsymbol{\Psi}}_j, \widehat{\sigma}_j^2)$ is defined as in (2.2.4). Next, based on the assumed multivariate t model (2.2.1) and its normal scale mixture representation,

$$\mathbf{b}_{ij} \mid \mathbf{y}_i, Z_{ij} = 1, \tau_{ij} \sim N_q\bigg(\mathbf{A}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_j), \frac{1}{\tau_{ij}}(\boldsymbol{\Psi}_j - \mathbf{A}\mathbf{U}_i\boldsymbol{\Psi}_j)\bigg),$$

where $\mathbf{A} = \boldsymbol{\Psi}_j\mathbf{U}_i^T(\mathbf{U}_i\boldsymbol{\Psi}_j\mathbf{U}_i^T + \sigma_j^2\mathbf{R}_i)^{-1}$. It follows that

$$\begin{aligned}\widehat{\mathbf{b}}_{ij} &= \widehat{\boldsymbol{\Psi}}_j\mathbf{U}_i^T\bigg(\mathbf{U}_i\widehat{\boldsymbol{\Psi}}_j\mathbf{U}_i^T + \widehat{\sigma}_j^2\mathbf{R}_i\bigg)^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_j) \\ &= \bigg(\widehat{\boldsymbol{\Psi}}_j^{-1} + \frac{1}{\widehat{\sigma}_j^2}\mathbf{U}_i^T\mathbf{R}_i^{-1}\mathbf{U}_i\bigg)^{-1}\frac{1}{\widehat{\sigma}_j^2}\mathbf{U}_i^T\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_j),\end{aligned} \tag{2.2.8}$$

and

$$\widehat{\boldsymbol{\Omega}}_{ij} = \widehat{\boldsymbol{\Psi}}_j - \widehat{\boldsymbol{\Psi}}_j\mathbf{U}_i^T(\mathbf{U}_i\widehat{\boldsymbol{\Psi}}_j\mathbf{U}_i^T + \widehat{\sigma}_j^2\mathbf{R}_i)^{-1}\mathbf{U}_i\widehat{\boldsymbol{\Psi}}_j = \bigg(\widehat{\boldsymbol{\Psi}}_j^{-1} + \frac{1}{\widehat{\sigma}_j^2}\mathbf{U}_i^T\mathbf{R}_i^{-1}\mathbf{U}_i\bigg)^{-1}. \tag{2.2.9}$$

Based on the above results, we have

$$E\left(\ell_0(\boldsymbol{\pi} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \mid \mathbf{y}, \widehat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{I}\sum_{j=1}^{m} p_{ij}\ln(\pi_j), \tag{2.2.10}$$

$$\begin{aligned}&E\left(\ell_1(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \mid \mathbf{y}, \widehat{\boldsymbol{\theta}}\right) \\ &= -\sum_{i=1}^{I}\sum_{j=1}^{m} p_{ij}\frac{n_i}{2}\ln\sigma_j^2 - \sum_{i=1}^{I}\sum_{j=1}^{m} p_{ij}\frac{1}{2\sigma_j^2}\mathrm{tr}\bigg[\mathbf{R}_i^{-1}\left\{\mathbf{U}_i\widehat{\boldsymbol{\Omega}}_{ij}\mathbf{U}_i^T + \widehat{\tau}_{ij}(\mathbf{y}_i - \mathbf{U}_i\widehat{\mathbf{b}}_{ij})(\mathbf{y}_i - \mathbf{U}_i\widehat{\mathbf{b}}_{ij})^T\right\}\bigg] \\ &+ \sum_{i=1}^{I}\sum_{j=1}^{m} p_{ij}\frac{1}{\sigma_j^2}\widehat{\tau}_{ij}\boldsymbol{\beta}_j^T\mathbf{X}_i^T\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{U}_i\widehat{\mathbf{b}}_{ij}) - \sum_{i=1}^{I}\sum_{j=1}^{m} p_{ij}\frac{1}{2\sigma_j^2}\widehat{\tau}_{ij}\boldsymbol{\beta}_j^T\mathbf{X}_i^T\mathbf{R}_i^{-1}\mathbf{X}_i\boldsymbol{\beta}_j,\end{aligned} \tag{2.2.11}$$

$$\begin{aligned}&E\left(\ell_2(\boldsymbol{\Psi} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \mid \mathbf{y}, \widehat{\boldsymbol{\theta}}\right) \\ &= -\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{m} p_{ij}\ln|\boldsymbol{\Psi}_j| - \frac{1}{2}\mathrm{tr}\bigg\{\boldsymbol{\Psi}_j^{-1}\sum_{i=1}^{I}\sum_{j=1}^{m} p_{ij}(\widehat{\tau}_{ij}\widehat{\mathbf{b}}_{ij}\widehat{\mathbf{b}}_{ij}^T + \widehat{\boldsymbol{\Omega}}_{ij})\bigg\},\end{aligned} \tag{2.2.12}$$

and

$$\begin{aligned}E\left(\ell_3(\boldsymbol{\nu} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \mid \mathbf{y}, \widehat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{I}\sum_{j=1}^{m} p_{ij}\bigg[&\frac{\nu_j}{2}\bigg\{\ln(\frac{\nu_j}{2}) + E[\ln(\tau_{ij}) \mid \mathbf{y}, \widehat{\boldsymbol{\theta}}, Z_{ij} = 1] - \widehat{\tau}_{ij}\bigg\} \\ &- E[\ln(\tau_{ij}) \mid \mathbf{y}, \widehat{\boldsymbol{\theta}}, Z_{ij} = 1] - \ln\{\Gamma(\frac{\nu_j}{2})\}\bigg].\end{aligned} \tag{2.2.13}$$

Following Khodabina and Alireza (2010) and based on properties of generalized Gamma distribution,

$$E\left(\ln(\tau_{ij}) \mid \mathbf{y}, \widehat{\boldsymbol{\theta}}, Z_{ij}=1\right) = \ln \widehat{\tau}_{ij} + \left\{\psi\left(\frac{\nu_j + n_i}{2}\right) - \ln\left(\frac{\nu_j + n_i}{2}\right)\right\},$$

where

$$\psi\left(\frac{\nu_j + n_i}{2}\right) = \frac{\partial\Gamma(\frac{\nu_j+n_i}{2})}{\partial(\frac{\nu_j+n_i}{2})}/\Gamma(\frac{\nu_j + n_i}{2}).$$

Now we are ready to fully describe our proposed generalized EM algorithm for conducting maximum likelihood estimation.

Initialization: Set $k=0$; obtain some initial estimates of the parameters $\boldsymbol{\theta}^{(0)}$, including $\pi_j^{(0)}$, $\boldsymbol{\beta}_j^{(0)}$, $\boldsymbol{\Psi}_j^{(0)}$, $\nu_j^{(0)}$, and $\sigma_j^{2(0)}$, for $j=1,\ldots,m$.

E-step: At $(k+1)^{th}$ iteration, given $\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}$, compute $p_{ij}^{(k+1)}, \mathbf{b}_{ij}^{(k+1)}, \tau_{ij}^{(k+1)}$ and $\boldsymbol{\Omega}_{ij}^{(k+1)}$ based on (2.2.6), (2.2.8), (2.2.7) and (2.2.9), respectively, for $i=1,\ldots,I$ and $j=1,\ldots,m$. Subsequently, the four components of the expected complete log-likelihood can be constructed from (2.2.10), (2.2.11), (2.2.12), and (2.2.13), respectively.

M-step:

M-0: Obtain $\pi_j^{(k+1)}$, $j=1,\ldots,m$, by maximizing $E\left(\ell_0(\boldsymbol{\pi} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}\right)$, with respect to $\boldsymbol{\pi}$,

$$\pi_j^{(k+1)} = \frac{1}{I}\sum_{i=1}^{I} p_{ij}^{(k+1)}.$$

M-1: Given $\sigma_j^2 = \sigma_j^{2(k)}$, $j=1,\ldots,m$, obtain $\boldsymbol{\beta}_j^{(k+1)}$, $j=1,\ldots,m$, by maximizing

$$E\left(\ell_1(\boldsymbol{\beta}, \boldsymbol{\sigma}^{2(k)} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}\right),$$

with respect to $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}_j^{(k+1)} = \left\{\sum_{i=1}^{I} p_{ij}^{(k+1)} \frac{\tau_{ij}^{(k+1)}}{\sigma_j^{2(k)}} \mathbf{X}_i^T \mathbf{R}_i^{-1} \mathbf{X}_i\right\}^{-1} \left\{\sum_{i=1}^{I} p_{ij}^{(k+1)} \frac{\tau_{ij}^{(k+1)}}{\sigma_j^{2(k)}} \mathbf{X}_i^T \mathbf{R}_i^{-1} (\mathbf{y}_i - \mathbf{U}_i \mathbf{b}_{ij}^{(k+1)})\right\}.$$

M-2: Given $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{(k+1)}$, $j=1,\ldots,m$, obtain $\sigma_j^{2(k+1)}$, $j=1,\ldots,m$ by maximizing

$$E\left(\ell_1(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\sigma}^2 \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}\right),$$

with respect to $\boldsymbol{\sigma}^2$,

$$\sigma_j^{2(k+1)} = \frac{\sum_{i=1}^{I} p_{ij}^{(k+1)} \left\{\tau_{ij}^{(k+1)} \mathbf{E}_{ij}^T \mathbf{R}_i^{-1} \mathbf{E}_{ij} + \mathrm{tr}(\boldsymbol{\Omega}_{ij}^{(k+1)} \mathbf{U}_i^T \mathbf{R}_i^{-1} \mathbf{U}_i)\right\}}{\sum_{i=1}^{I} p_{ij}^{(k+1)} n_i},$$

M-3: Obtain $\boldsymbol{\Psi}_j^{(k+1)}$, $j = 1, \ldots, m$, by maximizing $E\left(\ell_2(\boldsymbol{\Psi} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}\right)$ with respect to $\boldsymbol{\Psi}$,

$$\boldsymbol{\Psi}_j^{(k+1)} = \frac{\sum_{i=1}^{I} p_{ij}^{(k+1)} (\tau_{ij}^{(k+1)} \mathbf{b}_{ij}^{(k+1)} (\mathbf{b}_{ij}^{(k+1)})^T + \boldsymbol{\Omega}_{ij}^{(k+1)})}{\sum_{i=1}^{I} p_{ij}^{(k+1)}}.$$

M-4: Obtain $\nu_j^{(k+1)}$, $j = 1, \ldots, m$, by maximizing $E\left(\ell_3(\boldsymbol{\nu} \mid \mathbf{y}, \mathbf{b}, \boldsymbol{\tau}, \mathbf{Z}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}\right)$ with respect to $\boldsymbol{\nu}$.

$$\nu_j^{(k+1)} = \arg\max_{\nu_j} \left[\sum_{i=1}^{I} p_{ij}^{(k+1)} \frac{\nu_j}{2} \left\{ \ln(\frac{\nu_j}{2}) + E[\ln(\tau_{ij}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}, Z_{ij} = 1] - \tau_{ij}^{(k+1)} \right\} - \ln\{\Gamma(\frac{\nu_j}{2})\} \right].$$

The problem is separable in each ν_j. Although these one-dimensional problems do not admit explicit solutions, they can be solved by numerical optimization methods, e.g., the Newton-Raphson algorithm or the secant method. However, we find that the above approach may not be always stable, partly due to the high nonlinearity of the objective function. Alternatively, we can replace M-4 by carrying out constrained estimation of the actual log-likelihood (2.2.3) with respect to the unknown degrees of freedom parameters, with all the other parameters held fixed at their currently updated values Pinheiro et al. (2001).

M-4*: Obtain $\nu_j^{(k+1)}$, $j = 1, \ldots, m$, by maximizing (2.2.3) with respect to $\boldsymbol{\nu}$, with $\boldsymbol{\pi} = \boldsymbol{\pi}^{(k+1)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}$, $\boldsymbol{\Psi} = \boldsymbol{\Psi}^{(k+1)}$, and $\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^{2(k+1)}$.

In the case that $\nu_j = \nu$, $j = 1, \ldots, m$, it is convenient to use a profile likelihood approach to avoid either M-4 or M-4* step entirely in the EM algorithm, i.e., conduct maximum likelihood estimation with ν held fixed, for a grid of ν values, say, $\nu = 1, \ldots, 20$, and then the final estimate of the degrees of freedom is selected as the one that gives the largest log-likelihood.

In the M-step, we do not aim to fully maximize the expected log-likelihood, as it requires iteratively solving M-1 and M-2, which may be computationally inefficient. Nevertheless, solving each of the five subproblems once in the M-step monotonically increases the expected log-likelihood, so that the stable monotone convergence property of the EM algorithm is preserved. The E-step and M-step are carried out alternatingly, until convergence is reached, i.e., the log-likelihood function (2.2.3) stops increasing up to a small tolerance value. Based on our limited experience, the proposed algorithm works well in terms of both computational stability and efficiency.

2.3 Simulation Study

We generate the data from the model

$$\mathbf{y}_i = \begin{cases} \mathbf{X}_i\boldsymbol{\beta}_1 + \mathbf{U}_i\mathbf{b}_{i1} + \mathbf{e}_{i1}, & \text{if } Z_i = 1; \\ \mathbf{X}_i\boldsymbol{\beta}_2 + \mathbf{U}_i\mathbf{b}_{i2} + \mathbf{e}_{i2}, & \text{if } Z_i = 2. \end{cases}$$

where $i = 1, \ldots, I$, $\boldsymbol{\beta}_1 = (1, 1, 0, 0)^T$, $\boldsymbol{\beta}_2 = (0, 0, 1, 1)^T$, and $\pi_1 = P(Z_i = 1) = 0.4$. The rows of the covariates $\mathbf{X}_i \in \mathbb{R}^{n_i \times 4}$ are independently generated from $N_4(\mathbf{0}, \mathbf{I})$. The rows of $\mathbf{U}_i \in \mathbb{R}^{n_i \times 2}$ are independently generated from $N_2(\mathbf{0}, \mathbf{I})$.

We consider the following three types of the random effects and the error distributions.

1. t distribution: $\mathbf{e}_{ij} \sim t_{n_i}(\mathbf{0}, \boldsymbol{\Lambda}_{ij}, \nu)$, $\mathbf{b}_{ij} \sim t_q(\mathbf{0}, \boldsymbol{\Psi}_j, \nu)$, and given τ_{ij}, $\mathbf{b}_{ij}$ and $\mathbf{e}_{ij}$ are conditionally independent. That is, $\mathbf{b}_{ij} \mid \tau_{ij} \sim N(0, \frac{1}{\tau_{ij}}\boldsymbol{\Psi}_j)$ and $\mathbf{e}_{ij} \mid \tau_{ij} \sim N(0, \frac{1}{\tau_{ij}}\boldsymbol{\Lambda}_{ij})$. We set $\boldsymbol{\Lambda}_{ij}$ as identity matrix and $\boldsymbol{\Psi}_j$ as diagonal matrix with diagonal elements 1 and off-diagonal elements 0.5. We consider three degrees of freedom values, i.e., $\nu \in \{1, 3, 5\}$.

2. Normal distribution: $\mathbf{e}_{ij} \sim N_{n_i}(\mathbf{0}, \boldsymbol{\Lambda}_{ij})$ and $\mathbf{b}_{ij} \sim N_q(\mathbf{0}, \boldsymbol{\Psi}_j)$, where we set $\boldsymbol{\Lambda}_{ij}$ as identity matrix and $\boldsymbol{\Psi}_j$ as diagonal matrix with diagonal elements 1 and off-diagonal elements 0.5.

3. Contaminated normal distribution: $\mathbf{e}_{ij} \sim 0.95N_{n_i}(\mathbf{0}, \mathbf{I}) + 0.05N_{n_i}(\mathbf{0}, 25\mathbf{I})$ and $\mathbf{b}_{ij} \sim 0.95N_q(\mathbf{0}, \mathbf{I}) + 0.05N_q(\mathbf{0}, 25\mathbf{I})$.

We have experimented with various sample sizes and numbers of replicates. In particular, the following four cases are considered,

Case 1: $n_i = 8$, $I = 100$.

Case 2: $n_i = 8$, $I = 200$.

Case 3: $n_i = 4$, $I = 200$.

Case 4: $n_i = 4$, $I = 400$.

The simulation is replicated 500 times under each setting.

We compare normal mixture mixed-effect approach to our proposed robust t-mixture method. Our implemented EM algorithm can be readily simplified to fit the normal mixture model; an even simpler and

more straightforward approach is to use the algorithm for the t-mixture model with the degrees of freedom held fixed at a very large number, say, 1000, so that the t distribution essentially becomes very close to normal. Similar to Bordes et al. (2007)and Hunter and Young (2012), we use the true parameter values as the initial values to start the EM algorithm, in order to avoid the possible bias introduced by different starting values among replications or label switching issues (Celeux, et al., 2000; Stephens, 2000; Yao and Lindsay, 2009), so as to compare the "best-case" results of the various estimation methods. The degrees of freedom estimates in the t-mixture model is determined based on the aforementioned profile likelihood approach.

In Table 2.1, we report the average degrees of freedom estimates using the t-mixture model under the aforementioned five mixed effect and error structures. As the tail of the assumed mixed effect and error distribution becomes heavier, the estimated degrees of freedom becomes smaller on average as expected. Therefore, the proposed approach captures the tail behavior of the mixed effect and error distributions quite well.

In Tables 2.2–2.5, we report the median squared errors (MedSE) for parameter estimates and the relative efficiencies of our proposed t-mixture method as compared to the conventional normal mixture model. In Figures 2.1–2.2, we also show the MedSE for some of parameter estimates for cases 1 and 2. Our t-mixture approach works very well and consistently outperforms the normal mixture model when the random effects and error distributions are of heavy tail or are contaminated by outliers. Even when the random effects and the error terms follow normal distribution, the performance of the t-mixture model is comparable to that of the normal mixture model. This is essentially because the latter method can be treated as a special case of our proposed robust t-mixture model, and thus the efficiency loss is minimal when no outlier presents in the data. When the true model has t-distributed random effects and errors, the relative efficiency estimates may be very high. This is because the normal mixture model may fail miserably when applied to heavy-tailed Cauchy or close-to-Cauchy distributions.

2.4 Lung Growth Data Analysis

We consider a dataset on lung growth of girls, from a study of air pollution and health in six cities across the U.S.; see see Dockery et al. (1983) for the details of the study. Here we focus on the records gathered from Topeka, Kansas. The lung growth status of 300 girls in Topeka were tracked. Most of them were enrolled in the first or second grade and between the ages of six and seven, and measurements of participants were obtained annually until graduation from high school or loss to follow-up Dockery et al. (1983). We have omitted the subjects with only one record, and now the number of observations gathered on each of the remaining 252 subjects over time ranges from 2 to 12.

We use the logarithmic forced expiratory volume in one second (fev1) as the response variable. Specially, this variable measures the volume of air that can be forcibly exhaled from the lungs in the first second of a forced expiratory maneuver, and it is critically important in the diagnosis of obstructive and restrictive diseases and is a commonly-used measure of lung function from the pulmonary function tests. We are interested in modeling the lung growth pattern over time, and thus the age variable is used as both the fixed-effect covariate and the random-effect covariate. It is also of great interest to investigate whether the subjects form several distinct clusters or groups that exhibit different behaviors on lung growth. We thus fit the data based on the traditional normal mixture of linear mixed models and the proposed robust t-mixture of linear mixed models. Following Heinzl et al. (2013), a three-component mixture model is used.

Table 2.6 shows the estimated parameters. Based on the profile likelihood approach, the degrees of freedom of the t-mixture model is estimated to be $\widehat{\nu} = 28$, which is quite large. This result suggests that the random effects and the errors may be approximately normally distributed in this application. To test our robust estimation approach, however, we add some artificial outliers for some arbitrarily selected subjects in the dataset and refit the t-mixture model. Using contaminated datasets with outliers in *one* subject, the estimated degrees of freedom is $\nu = 9$, and using the contaminated datasets with outliers in *two* subjects, the estimate becomes $\nu = 6$. The decrease in the estimated degrees of freedom as the number of outliers increases clearly demonstrates the robustness of the proposed approach. In addition, compared to the estimates of the traditional normal mixture of linear mixed models, the parameter estimates for the new method does not change much when the outliers are added into the data set.

Our analysis reveals some interesting cluster structure. In Figure 2.3, three distinct groups can be clearly distinguished by the intercept and slope estimates bases on the mixed effects. Girls assigned to different clusters are marked with different colors and symbols. It appears that cluster 1 (blue, triangle) consists of the girls who had initial low-level lung function and then experienced relatively fast lung growth to their adulthood. In contrast, cluster 2 (red, circle) consists of the girls who had relatively high level of initial lung development and then experienced relatively slow lung growth to their adulthood. Cluster 3 (black, cross) is the smallest cluster of the three, which appears to consist of the girls who had relatively low level of initial lung development and also experienced relatively slow lung growth over time.

2.5 Discussion

We have proposed a robust mixture linear mixed model, using multivariate t distribution to robustify the model estimation and inference. It is interesting to extend our model to other distributions that possessing certain robustness properties, e.g., mixture of Laplace distributed mixed effects and random errors. It is also worthwhile to apply the trimmed-likelihood idea to the mixture linear mixed model setups. The recently developed penalized estimation approaches may also be adopted to directly capture and accommodate potential outliers.

Figure 2.1: *The median squared errors of Case 1. Solid line is for the t-mixture method and dashed line is for the normal mixture method. The five conditions refer to five scenarios of the random effects and error distributions, i.e., t_1, t_3, t_5, normal, and contaminated normal, respectively.*

#replicates	#subjects	Estimated degrees of freedom				
		t_1	t_3	t_5	Normal	Contaminated Normal
$n_i = 8$	$I = 100$	1.605	5.665	6.716	12.46	4.839
	$I = 200$	1.605	3.832	9.868	12.46	4.133
$n_i = 4$	$I = 200$	2.575	6.149	7.199	10.65	5.665
	$I = 400$	1.488	5.252	7.766	12.46	3.015

Table 2.1: *Degrees of freedom estimation results, based on 500 simulation runs.*

Estimator		Random Effects and Error Distribution				
		t_1	t_3	t_5	Normal	Contaminated N.
$\widehat{\pi}_1$	MedSE(NMM)	0.196	0.014	0.014	0.009	0.014
	MedSE(tMM)	0.064	0.010	0.014	0.009	0.009
	Efficiency	3.063	1.400	1.000	1.000	1.556
$\widehat{\beta}_{11}$	MedSE(NMM)	0.265	0.106	0.005	0.004	0.010
	MedSE(tMM)	0.197	0.006	0.003	0.004	0.004
	Efficiency	1.345	17.667	1.667	1.000	2.500
$\widehat{\beta}_{21}$	MedSE(NMM)	0.279	0.110	0.005	0.004	0.012
	MedSE(tMM)	0.216	0.007	0.004	0.004	0.004
	Efficiency	1.292	15.714	1.250	1.000	3.000
$\widehat{\beta}_{31}$	MedSE(NMM)	0.276	0.094	0.006	0.003	0.010
	Median(tMM)	0.237	0.008	0.005	0.003	0.004
	Efficiency	1.165	11.750	1.200	1.000	2.500
$\widehat{\beta}_{41}$	MedSE(NMM)	0.265	0.118	0.005	0.004	0.012
	Median(tMM)	0.192	0.008	0.004	0.004	0.004
	Efficiency	1.380	14.750	1.250	1.000	3.000
$\widehat{\beta}_{12}$	MedSE(NMM)	7.871	0.014	0.005	0.003	0.011
	MedSE(tMM)	0.085	0.005	0.004	0.003	0.004
	Efficiency	92.600	1.280	1.250	1.000	2.750
$\widehat{\beta}_{22}$	MedSE(NMM)	7.516	0.015	0.006	0.003	0.012
	MedSE(tMM)	0.078	0.004	0.004	0.003	0.004
	Efficiency	92.600	3.750	1.500	1.000	3.000
$\widehat{\beta}_{32}$	MedSE(NMM)	7.235	0.017	0.008	0.003	0.012
	MedSE(tMM)	0.081	0.005	0.004	0.003	0.004
	Efficiency	89.321	3.400	2.000	1.000	3.000
$\widehat{\beta}_{42}$	MedSE(NMM)	5.869	0.017	0.006	0.003	0.012
	MedSE(tMM)	0.081	0.005	0.005	0.003	0.004
	Efficiency	72.457	3.400	1.200	1.000	3.000

Table 2.2: *Simulation results for Case 1:* $n_i = 8$, $I = 100$.

Estimator		Random Effects and Error Distribution				
		t_1	t_3	t_5	Normal	Contaminated N.
$\widehat{\pi}_1$	MedSE(NMM)	0.211	0.010	0.011	0.012	0.014
	MedSE(tMM)	0.054	0.010	0.011	0.012	0.011
	Efficiency	3.907	1.000	1.000	1.000	1.273
$\widehat{\beta}_{11}$	MedSE(NMM)	0.289	0.041	0.004	0.001	0.004
	MedSE(tMM)	0.151	0.002	0.002	0.001	0.002
	Efficiency	1.914	20.500	2.000	1.000	2.000
$\widehat{\beta}_{21}$	MedSE(NMM)	0.270	0.040	0.003	0.001	0.005
	MedSE(tMM)	0.139	0.002	0.002	0.002	0.002
	Efficiency	1.942	20	1.500	0.500	2.500
$\widehat{\beta}_{31}$	MedSE(NMM)	0.266	0.034	0.003	0.001	0.005
	MedSE(tMM)	0.161	0.003	0.002	0.001	0.002
	Efficiency	1.652	11.333	1.500	1.000	2.500
$\widehat{\beta}_{41}$	MedSE(NMM)	0.271	0.040	0.004	0.002	0.004
	MedSE(tMM)	0.155	0.002	0.002	0.002	0.002
	Efficiency	1.748	20.000	2.000	1.000	2.000
$\widehat{\beta}_{12}$	MedSE(NMM)	7.753	0.008	0.004	0.002	0.007
	MedSE(tMM)	0.031	0.002	0.002	0.002	0.002
	Efficiency	250.097	4.000	2.00	1.000	3.500
$\widehat{\beta}_{22}$	MedSE(NMM)	5.797	0.008	0.004	0.001	0.008
	MedSE(tMM)	0.028	0.002	0.002	0.001	0.002
	Efficiency	207.036	4.000	2.000	1.000	4.000
$\widehat{\beta}_{32}$	MedSE(NMM)	6.116	0.008	0.004	0.001	0.009
	MedSE(tMM)	0.035	0.002	0.002	0.002	0.002
	Efficiency	175.029	4.000	2.000	0.500	4.500
$\widehat{\beta}_{42}$	MedSE(NMM)	6.783	0.009	0.003	0.002	0.008
	MedSE(tMM)	0.033	0.002	0.002	0.002	0.002
	Efficiency	204.545	4.500	1.500	1.000	4.000

Table 2.3: *Simulation results for Case 2:* $n_i = 8$, $I = 200$.

Estimator		Random Effects and Error Distribution				
		t_1	t_3	t_5	Normal	Contaminated N.
$\widehat{\pi}_1$	MedSE(NMM)	0.208	0.039	0.012	0.011	0.068
	MedSE(tMM)	0.079	0.013	0.012	0.011	0.008
	Efficiency	2.633	3.000	1.000	1.000	8.500
$\widehat{\beta}_{11}$	MedSE(NMM)	0.253	0.181	0.007	0.004	0.095
	MedSE(tMM)	0.253	0.008	0.006	0.002	0.005
	Efficiency	1.000	2.130	1.167	2.000	19.000
$\widehat{\beta}_{21}$	MedSE(NMM)	0.228	0.201	0.006	0.003	0.104
	MedSE(tMM)	0.246	0.010	0.005	0.002	0.006
	Efficiency	0.927	20.100	1.200	1.500	5.567
$\widehat{\beta}_{31}$	MedSE(NMM)	0.251	0.194	0.007	0.003	0.106
	MedSE(tMM)	0.250	0.009	0.005	0.002	0.005
	Efficiency	1.008	21.556	1.400	1.500	21.200
$\widehat{\beta}_{41}$	MedSE(NMM)	0.249	0.203	0.009	0.004	0.113
	MedSE(tMM)	0.241	0.008	0.006	0.002	0.006
	Efficiency	1.029	25.375	1.500	2.000	18.833
$\widehat{\beta}_{12}$	MedSE(NMM)	11.846	0.335	0.007	0.004	0.043
	MedSE(tMM)	0.405	0.011	0.005	0.002	0.004
	Efficiency	29.249	30.455	1.400	2.000	10.750
$\widehat{\beta}_{22}$	MedSE(NMM)	16.726	0.209	0.007	0.004	0.048
	MedSE(tMM)	0.443	0.009	0.006	0.002	0.004
	Efficiency	37.756	23.222	1.167	2.000	12.000
$\widehat{\beta}_{32}$	MedSE(NMM)	15.735	0.270	0.007	0.004	0.045
	MedSE(tMM)	0.337	0.009	0.005	0.002	0.004
	Efficiency	46.691	30.000	1.400	2.000	11.250
$\widehat{\beta}_{42}$	MedSE(NMM)	15.323	0.275	0.008	0.003	0.035
	MedSE(tMM)	0.379	0.009	0.006	0.002	0.004
	Efficiency	40.456	30.556	1.333	1.500	8.750

Table 2.4: *Simulation results for Case 3:* $n_i = 4$, $I = 200$.

Estimator		Random Effects and Error Distribution				
		t_1	t_3	t_5	Normal	Contaminated N.
$\widehat{\pi}_1$	MedSE(NMM)	0.222	0.211	0.012	0.009	0.224
	MedSE(tMM)	0.044	0.054	0.012	0.009	0.010
	Efficiency	5.045	3.907	1.000	1.000	22.400
$\widehat{\beta}_{11}$	MedSE(NMM)	0.275	0.289	0.007	0.002	0.177
	MedSE(tMM)	0.083	0.151	0.006	0.002	0.002
	Efficiency	3.313	1.914	1.167	1.000	88.500
$\widehat{\beta}_{21}$	MedSE(NMM)	0.280	0.270	0.006	0.002	0.174
	MedSE(tMM)	0.084	0.139	0.005	0.002	0.002
	Efficiency	3.333	1.942	1.200	1.000	87.000
$\widehat{\beta}_{31}$	MedSE(NMM)	0.279	0.266	0.007	0.002	0.181
	MedSE(tMM)	0.079	0.161	0.005	0.002	0.002
	Efficiency	3.532	1.652	1.400	1.000	90.500
$\widehat{\beta}_{41}$	MedSE(NMM)	0.276	0.271	0.009	0.002	0.180
	MedSE(tMM)	0.075	0.155	0.006	0.002	0.002
	Efficiency	3.680	1.748	1.600	1.000	90.000
$\widehat{\beta}_{12}$	MedSE(NMM)	14.856	7.753	0.007	0.002	0.042
	MedSE(tMM)	0.024	0.031	0.005	0.002	0.002
	Efficiency	619.000	250.097	1.400	1.000	21.000
$\widehat{\beta}_{22}$	MedSE(NMM)	17.778	5.797	0.007	0.002	0.059
	MedSE(tMM)	0.025	0.028	0.006	0.002	0.002
	Efficiency	711.120	207.036	1.167	1.000	29.500
$\widehat{\beta}_{32}$	MedSE(NMM)	12.837	6.116	0.007	0.002	0.043
	MedSE(tMM)	0.030	0.035	0.005	0.002	0.002
	Efficiency	427.900	175.029	1.400	1.000	21.500
$\widehat{\beta}_{42}$	MedSE(NMM)	18.654	6.783	0.008	0.002	0.041
	MedSE(tMM)	0.030	0.033	0.006	0.001	0.002
	Efficiency	621.8	205.545	1.333	2.000	20.500

Table 2.5: *Simulation results for Case 4:* $n_i = 4$, $I = 400$.

	Original		With 1 outlier		With 2 outliers	
	t_{28}	Normal	t_9	Normal	t_6	Normal
$\widehat{\pi}_1$	0.248	0.235	0.281	0.569	0.297	0.196
$\widehat{\pi}_2$	0.688	0.704	0.652	0.402	0.630	0.765
$\widehat{\beta}_{01}$	-0.010	-0.010	-0.041	-0.126	-0.056	-0.175
$\widehat{\beta}_{11}$	0.074	0.074	0.074	0.083	0.075	0.083
$\widehat{\beta}_{02}$	-0.350	-0.341	-0.361	-0.336	-0.368	-0.274
$\widehat{\beta}_{12}$	0.092	0.091	0.093	0.090	0.093	0.087
$\widehat{\beta}_{03}$	-0.307	-0.296	-0.293	-0.418	-0.279	-0.335
$\widehat{\beta}_{13}$	0.074	0.073	0.074	0.090	0.075	0.088

Table 2.6: *Estimation results for the Topeka girls lung function data analysis.*

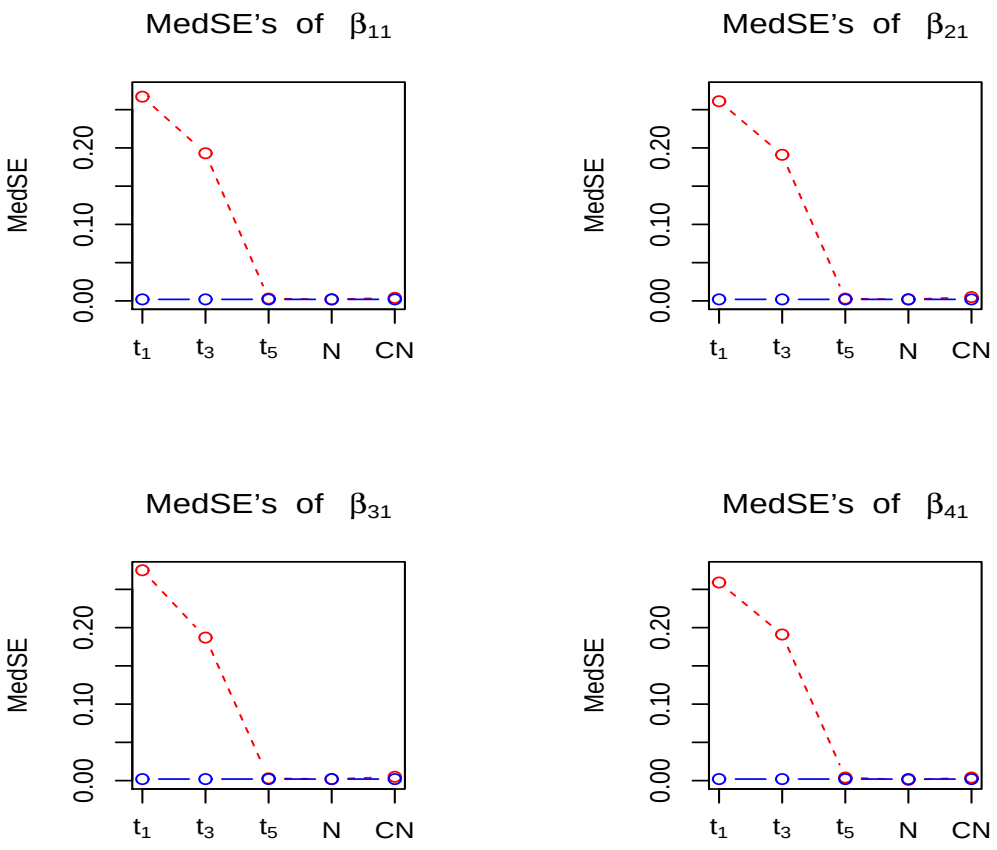

Figure 2.2: *The median squared errors of Case 2. All the settings are the same as in Figure 2.1.*

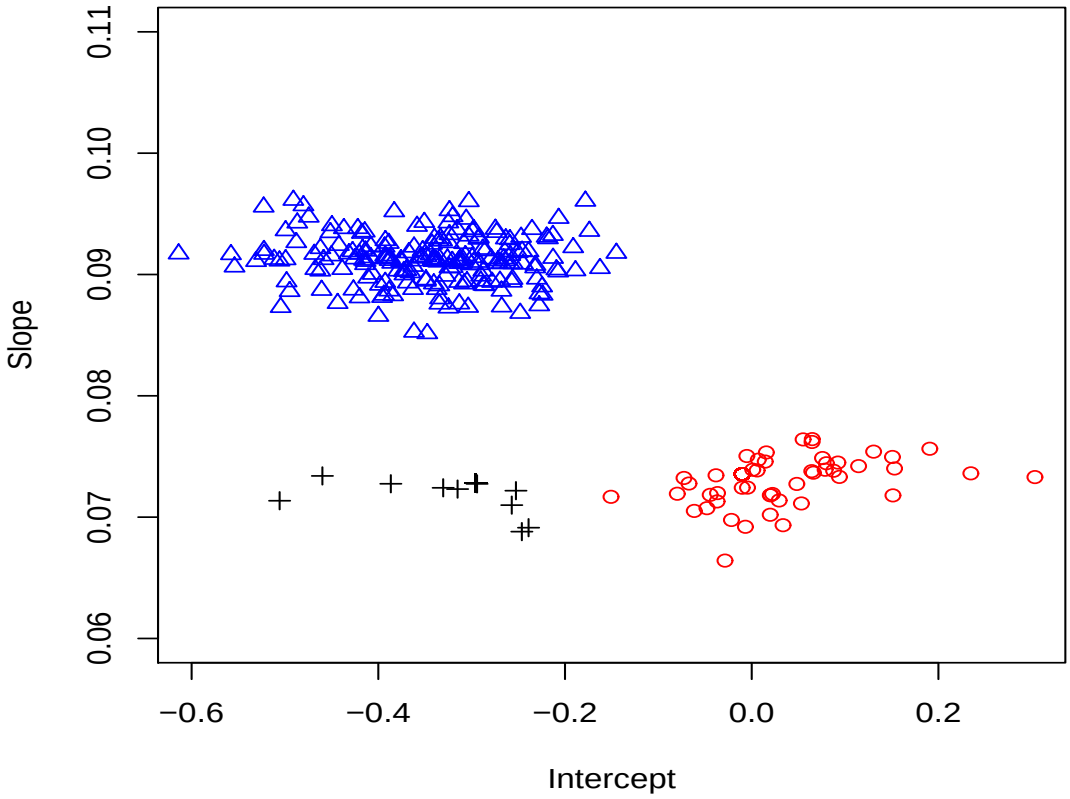

Figure 2.3: *Cluster patten revealed by the t-mixture model based on the estimated intercept and slope parameters of the mixed effects.*

Bibliography

[1] Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics*, 16, 523-531.

[2] Bai, X., Yao, W., Boyer, J, E. (2012). Robust fitting of mixture regression models. *Computational Statistics Data Analysis*, 2347-2359.

[3] Beaton, A. E. and Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16, 147-185.

[4] Bordes, L., Chauveau, D., and Vandekerkhove, P. (2007). A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis*, 51, 5429-5443

[5] Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95, 957-970.

[6] Celeux, G., Martin, O., and Lavergne, Ch. (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modeling*, 5, 1-25.

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

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

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

[10] Dockery, D. W., Berkery, C. S., Ware, J. H., Speizer, F. E., and Ferris, B. G. (1983). Distribution of fvc and fev1 in children 6 to 11 years old. *American Review of Respiratory Disease*, 128, 405-12.

[11] Elashoff, M. and Ryan, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, 13, 48-65.

[12] Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2): 399-433.

[13] García-Escudero, L. A., Gordaliza, A., Mayo-Iscara, A., and San Martín, R. (2010). Robust clusterwise linear regression through trimming. *Computational Statistics and Data Analysis*, 54, 3057-3069.

[14] García-Escudero, L. A., Gordaliza, A., San Martín, R., Van Aelst, S., and Zamar, R. (2009). Robust linear clustering. *Journal of the Royal Statistical Society*, B, 71, 301-318.

[15] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

[16] Hanfelt, J. J. and Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating equations. *Biometrika*, 82, 461-477.

[17] Hanfelt, J.J. and Liang, K.-Y. (1997). Approximate likelihood for generalized linear errors-in-variables models. *Journal of the Royal Statistical Society*, B59, 627-637.

[18] Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93108.

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

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

[21] Heinzl, F., and Tutz, G. (2013). Clustering in linear mixed models with approximate Dirichlet process mixtures using EM algorithm. *Statistical Modeling*, 13, 41-67.

[22] Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. M. (1959). The estimation of

environmental and genetic trends from records subject to culling. *International Biometric Society*, 15 (2): 192-218.

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

[24] Hennig, C. (2002). Fixed point clusters for linear regression: computation and comparison. *Journal of Classification*, 19, 249-276.

[25] Hennig, C. (2003). Clusters, outliers, and regression: fixed point clusters. *Journal of Multivariate Analysis*, 86, 183-212.

[26] Hennig, C. (2004). Breakdown points for maximum likelihood-estimators of location-scale mixtures. *Annals of Statistics*, 32, 1313-1340.

[27] Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least squares. *Computations in Statistics*, A6, 813-827.

[28] Huber, P. J. (1973). Robust regression: asymptotics, conjectures, and monte carlo. *Annals of Statistics*, 1, 799-821.

[29] Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

[30] Hunter, D. R., and Young, D. S. (2012). Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics*, 24 (1): 19-38

[31] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1): 79-87.

[32] Jiang, W. and Tanner, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *The Annals of Statistics*, 27, 987-1011.

[33] Khodabina, M., and Alireza, A. (2010). Some properties of generalized gamma distribution. *Mathematical Sciences*, 4, 9-28.

[34] Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statitical modeling using the t distribution. *Journal of the American Statistical Association*, 84, 881-896.

[35] Li, B. (1993). A deviance function for the quasi-likelihood method. *Biometrika*, 80, 741-753.

[36] Li, J., Ray, S., and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(8), 1687-1723.

[37] Markatou, M. (2000). Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, 56, 483486.

[38] Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

[39] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

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

[41] Mclean, R. A., Sanders, W. L., Stroup, W. W. (1991). A unified approach to mixed linear models. *The American Statistician*, 45(1): 54-64.

[42] Mueller, C. H. and Garlipp, T. (2005). Simple consistent cluster methods based on redescending M-estimators with an application to edge identification in images. *Journal of Multivariate Analysis*, 92, 359-385.

[43] Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. (2007). Robust fitting of mixtures using the Trimmed Likelihood Estimator. *Computational Statistics and Data Analysis*, 52, 299-308.

[44] Pinheiro, J. C., Liu, CH. H., Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t-distribution. *Journal of Computational and Graphical Statistics*, 10 (2), 249-276.

[45] Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6, 15-32.

[46] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outier Detection*. Wiley, New York.

[47] Rousseeuw, P. J and Yohai, V. J. (1984). Robust regression by means of s-estimators. *Robust and Nonlinear Time Series Analysis.* Franke, J. , Hdle W. and Martin R. D. (eds.), *Lectures Notes in Statistics* 26, 256-272, New York: Springer.

[48] Shen, H., Yang, J., and Wang, S. (2004). Outlier detecting in fuzzy switching regression models. *Artificial Intelligence: Methodology, Systems, and Applications Lecture Notes in Computer Science*, 3192, 208-215.

[49] Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models.* Boca Raton. Chapman and Hall/CRC.

[50] Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of Royal Statistical Society*, B62, 795-809.

[51] Wedel, M. and Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations.* 2nd edition, Norwell, MA: Kluwer Academic Publishers. *Journal of Classification.* Springer, New York.

[52] Welsh, A. H. and Richardson, A. M. (1997). Approaches to the robust estimation of mixed models. *Handbook of Statistics*, 15 of Maddala, G. S., and Rao, C. R. (1997), chapter 13, 343-384.

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

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

[55] Yohai, V. J. (1987). High breakdown point and high efficiency estimates for regression. *The Annals of Statistics*, 15, 642-65.