ROBUST MIXTURE MODELING

by

CHUN YU

M.S., Kansas State University, 2008

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

Ordinary least-squares (OLS) estimators for a linear model are very sensitive to unusual values in the design space or outliers among y values. Even one single atypical value may have a large effect on the parameter estimates. In this proposal, we first review and describe some available and popular robust techniques, including some recent developed ones, and compare them in terms of breakdown point and efficiency. In addition, we also use a simulation study and a real data application to compare the performance of existing robust methods under different scenarios. Finite mixture models are widely applied in a variety of random phenomena. However, inference of mixture models is a challenging work when the outliers exist in the data. The traditional maximum likelihood estimator (MLE) is sensitive to outliers. In this proposal, we propose a Robust Mixture via Mean shift penalization (RMM) in mixture models and Robust Mixture Regression via Mean shift penalization (RM^2) in mixture regression, to achieve simultaneous outlier detection and parameter estimation. A mean shift parameter, which is denoted by γ, is added to the mixture models, and penalized by a nonconvex penalty function. With this model setting, we develop an iterative thresholding embedded EM algorithm to maximize the penalized objective function. Comparing with other existing robust methods, the proposed methods show outstanding performance in both identifying outliers and estimating the parameters.

Key words: Robust; Outlier detection; Mixture models; EM algorithm; Penalized likelihood.

ROBUST MIXTURE MODELING

by

CHUN YU

M.S., Kansas State University, 2008

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, PhD

Co-Major Professor
Kun Chen, PhD

Abstract

Ordinary least-squares (OLS) estimators for a linear model are very sensitive to unusual values in the design space or outliers among y values. Even one single atypical value may have a large effect on the parameter estimates. In this proposal, we first review and describe some available and popular robust techniques, including some recent developed ones, and compare them in terms of breakdown point and efficiency. In addition, we also use a simulation study and a real data application to compare the performance of existing robust methods under different scenarios. Finite mixture models are widely applied in a variety of random phenomena. However, inference of mixture models is a challenging work when the outliers exist in the data. The traditional maximum likelihood estimator (MLE) is sensitive to outliers. In this proposal, we propose a Robust Mixture via Mean shift penalization (RMM) in mixture models and Robust Mixture Regression via Mean shift penalization (RM^2) in mixture regression, to achieve simultaneous outlier detection and parameter estimation. A mean shift parameter, which is denoted by γ, is added to the mixture models, and penalized by a nonconvex penalty function. With this model setting, we develop an iterative thresholding embedded EM algorithm to maximize the penalized objective function. Comparing with other existing robust methods, the proposed methods show outstanding performance in both identifying outliers and estimating the parameters.

Key words: Robust; Outlier detection; Mixture models; EM algorithm; Penalized likelihood.

Table of Contents

List of Figures

List of Tables

Acknowledgments

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

I am grateful to Dr. Christopher Pinner for offering generous help and support as chairperson of my final examining committee. I would also like to thank Dr. Haiyan Wang, Dr. Weixing Song and Dr. Jianhan Chen for their willingness to serve on my supervisory committee and for their valuable insight.

My gratefulness extends to everyone who supported me in any respect during the completion of the dissertation.

Chapter 1

Robust Linear Regression: A Review and Comparison

1.1 Introduction

Linear regression has been one of the most important statistical data analysis tools. Given the independent and identically distributed (iid) observations $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, in order to understand how the response y_is are related to the covariates $\boldsymbol{x}_i$s, we traditionally assume the following linear regression model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{1.1}$$

where $\boldsymbol{\beta}$ is an unknown $p \times 1$ vector, and the ε_is are i.i.d. and independent of $\boldsymbol{x}_i$ with $\mathrm{E}(\varepsilon_i \mid \boldsymbol{x}_i) = 0$. The most commonly used estimate for $\boldsymbol{\beta}$ is the ordinary least square (OLS) estimate which minimizes the sum of squared residuals

$$\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2. \tag{1.2}$$

However, it is well known that the OLS estimate is extremely sensitive to the outliers. A single outlier can have large effect on the OLS estimate.

In this paper, we review and describe some available robust methods. In addition, a simulation study and a real data application are used to compare different existing robust methods. The efficiency and breakdown point (Donoho and Huber 1983) are two traditionally used important criteria to compare different robust methods. The efficiency is used to measure the relative efficiency of the robust estimate compared to the OLS estimate when the error distribution is exactly normal and there are no outliers. Breakdown point is to measure the proportion of outliers an estimate can tolerate before it goes to infinity. In this paper, finite sample breakdown point (Donoho and Huber 1983) is used and defined as follows: Let $\mathbf{z}_i = (\boldsymbol{x}_i, y_i)$. Given any sample $\boldsymbol{z} = (\boldsymbol{z}_i, \ldots, \boldsymbol{z}_n)$, denote $T(\boldsymbol{z})$ the estimate of the parameter $\boldsymbol{\beta}$. Let $\boldsymbol{z}'$ be the corrupted sample where any m of the original points of $\boldsymbol{z}$ are replaced by arbitrary bad data. Then the finite sample breakdown point δ^* is defined as

$$\delta^*(\boldsymbol{z}, T) = \min_{1 \le m \le n} \left\{ \frac{m}{n} : \sup_{\boldsymbol{z}'} \|T(\boldsymbol{z}') - T(\boldsymbol{z})\| = \infty \right\}, \tag{1.3}$$

where $\|\cdot\|$ is the Euclidean norm.

Many robust methods have been proposed to achieve high breakdown point or high efficiency or both. M-estimates (Huber, 1981) are solutions of the normal equation with appropriate weight functions. They are resistant to unusual y observations, but sensitive to high leverage points on $\mathbf{x}$. Hence the breakdown point of an M-estimate is $1/n$. R-estimates (Jaeckel 1972) which minimize the sum of scores of the ranked residuals have relatively high efficiency but their breakdown points are as low as those of OLS estimates. Least Median of Squares (LMS) estimates (Siegel 1982) which minimize the median of squared residuals, Least Trimmed Squares (LTS) estimates (Rousseeuw 1983) which minimize the trimmed sum of squared residuals, and S-estimates (Rousseeuw and Yohai 1984) which minimize the variance of the residuals all have high breakdown point but with low efficiency. Generalized S-estimates (GS-estimates) (Croux et al. 1994) maintain high breakdown point

as S-estimates and have slightly higher efficiency. MM-estimates proposed by Yohai (1987) can simultaneously attain high breakdown point and efficiencies. Mallows Generalized M-estimates (Mallows 1975) and Schweppe Generalized M-estimates (Handschin et al. 1975) downweight the high leverage points on $\mathbf{x}$ but cannot distinguish "good" and "bad" leverage points, thus resulting in a loss of efficiencies. In addition, these two estimators have low breakdown points when p, the number of explanatory variables, is large. Schweppe one-step (S1S) Generalized M-estimates (Coakley and Hettmansperger 1993) overcome the problems of Schweppe Generalized M-estimates and are calculated in one step. They both have high breakdown points and high efficiencies. Recently, Gervini and Yohai (2002) proposed a new class of high breakdown point and high efficiency robust estimate called robust and efficient weighted least squares estimator (REWLSE). Lee et al. (2011) and She and Owen (2011) proposed a new class of robust methods based on the regularization of case-specific parameters for each response. They further proved that the M-estimator with Huber's ψ function is a special case of their proposed estimator.

The rest of the paper is organized as follows. In Section 2, we review and describe some of the available robust methods. In Section 3, a simulation study and a real data application are used to compare different robust methods. Some discussions are given in Section 4.

1.2 Robust Regression Methods

1.2.1 M-Estimates

By replacing the least squares criterion (1.2) with a robust criterion, M-estimate (Huber, 1964) of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\hat{\sigma}}\right), \tag{1.4}$$

where $\rho(\cdot)$ is a robust loss function and $\hat{\sigma}$ is an error scale estimate. The derivative of ρ, denoted by $\psi(\cdot) = \rho'(\cdot)$, is called the influence function. In particular, if $\rho(t) = \frac{1}{2}t^2$, then

the solution is the OLS estimate. The OLS estimate is very sensitive to outliers. Rousseeuw and Yohai (1984) indicated that OLS estimates have a breakdown point (BP) of BP $= 1/n$, which tends to zero when the sample size n is getting large. Therefore, one single unusual observation can have large impact on the OLS estimate.

One of the commonly used robust loss functions is Huber's ψ function (Huber 1981), where $\psi_c(t) = \rho'(t) = \max\{-c, \min(c, t)\}$. Huber (1981) recommends using $c = 1.345$ in practice. This choice produces a relative efficiency of approximately 95% when the error density is normal. Another possibility for $\psi(\cdot)$ is Tukey's bisquare function $\psi_c(t) = t\{1 - (t/c)^2\}_+^2$. The use of $c = 4.685$ produces 95% efficiency. If $\rho(t) = |t|$, then *least absolute deviation* (LAD, also called median regression) estimates are achieved by minimizing the sum of the absolute values of the residuals

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left| y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} \right|. \tag{1.5}$$

The LAD is also called L_1 estimate due to the L_1 norm used. Although LAD is more resistent than OLS to unusual y values, it is sensitive to high leverage outliers, and thus has a breakdown point of BP $= 1/n \to 0$ (Rousseeuw and Yohai 1984). Moreover, LAD estimates have a low efficiency of 64% when the errors are normally distributed. Similar to LAD estimates, the general monotone M-estimates, i.e., M-estimates with monotone ψ functions, have a BP $= 1/n \to 0$ due to lack of immunity to high leverage outliers (Maronna, Martin, and Yohai 2006).

1.2.2 LMS Estimates

The LMS estimates (Siegel 1982) are found by minimizing the median of the squared residuals

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \text{Med}\{\left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2\}. \tag{1.6}$$

One good property of the LMS estimate is that it possesses a high breakdown point of near 0.5. However, the LMS estimate has at best an efficiency of 0.37 when the assumption of normal errors is met (see Rousseeuw and Croux 1993). Moreover, LMS estimates do not have a well-defined influence function because of its convergence rate of $n^{-\frac{1}{3}}$ (Rousseeuw 1982). Despite these limitations, the LMS estimate can be used as the initial estimate for some other high breakdown point and high efficiency robust methods.

1.2.3 LTS Estimates

The LTS estimate (Rousseeuw 1983) is defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{q} r_{(i)}\left(\boldsymbol{\beta}\right)^2, \tag{1.7}$$

where $r_{(i)}(\boldsymbol{\beta}) = y_{(i)} - \boldsymbol{x}_{(i)}^T\boldsymbol{\beta}$, $r_{(1)}\left(\boldsymbol{\beta}\right)^2 \leq \cdots \leq r_{(q)}\left(\boldsymbol{\beta}\right)^2$ are ordered squared residuals, $q = [n\left(1-\alpha\right)+1]$, and α is the proportion of trimming. Using q = $\left(\frac{n}{2}\right)$ +1 ensures that the estimator has a breakdown point of BP = 0.5, and the convergence rate of $n^{-\frac{1}{2}}$ (Rousseeuw 1983). Although highly resistent to outliers, LTS suffers badly in terms of very low efficiency, which is about 0.08, relative to OLS estimates (Stromberg, et al. 2000). The reason that LTS estimates call attentions to us is that it is traditionally used as the initial estimate for some other high breakdown point and high efficiency robust methods.

1.2.4 S-Estimates

S-estimates (Rousseeuw and Yohai 1984) are defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \hat{\sigma}\left(r_1\left(\boldsymbol{\beta}\right), \cdots, r_n\left(\boldsymbol{\beta}\right)\right), \tag{1.8}$$

where $r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\hat{\sigma}(r_1(\boldsymbol{\beta}), \cdots, r_n(\boldsymbol{\beta}))$ is the scale M-estimate which is defined as the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) = \delta, \tag{1.9}$$

for any given $\boldsymbol{\beta}$, where δ is taken to be $\mathrm{E}_\Phi[\rho(r)]$. For the biweight scale, S-estimates can attain a high breakdown point of BP = 0.5 and has an asymptotic efficiency of 0.29 under the assumption of normally distributed errors (Maronna, Martin, and Yahai 2006).

1.2.5 Generalized S-Estimates (GS-Estimates)

Croux et al. (1994) proposed generalized S-estimates in an attempt to improve the low efficiency of S-estimators. Generalized S-estimates are defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S_n(\boldsymbol{\beta}), \tag{1.10}$$

where $S_n(\boldsymbol{\beta})$ is defined as

$$S_n(\boldsymbol{\beta}) = \sup\left\{S > 0; \binom{n}{2}^{-1}\sum_{i<j}\rho\left(\frac{r_i - r_j}{S}\right) \geq k_{n,p}\right\}, \tag{1.11}$$

where $r_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$, p is the number of regression parameters, and $k_{n,p}$ is a constant which might depend on n and p. Particularly, if $\rho(x) = I(|x| \geq 1)$ and $k_{n,p} = \left(\binom{n}{2} - \binom{h_p}{2} + 1\right)/\binom{n}{2}$ with $h_p = \frac{n+p+1}{2}$, generalized S-estimator yields a special case, the least quartile difference (LQD) estimator, which is defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q_n(r_1, \ldots, r_n), \tag{1.12}$$

where

$$Q_n = \{|r_i - r_j|; i < j\}_{\binom{h_p}{2}} \tag{1.13}$$

is the $\binom{h_p}{2}$th order statistic among the $\binom{n}{2}$ elements of the set $\{|r_i - r_j| ; i < j\}$. Generalized S-estimates have a breakdown point as high as S-estimates but with a higher efficiency.

1.2.6 MM-Estimates

First proposed by Yohai (1987), MM-estimates have become increasingly popular and are one of the most commonly employed robust regression techniques. The MM-estimates can be found by a three-stage procedure. In the first stage, compute an initial consistent estimate $\hat{\boldsymbol{\beta}}_0$ with high breakdown point but possibly low normal efficiency. In the second stage, compute a robust M-estimate of scale $\hat{\sigma}$ of the residuals based on the initial estimate. In the third stage, find an M-estimate $\hat{\boldsymbol{\beta}}$ starting at $\hat{\boldsymbol{\beta}}_0$.

In practice, LMS or S-estimate with Huber or bisquare functions is typically used as the initial estimate $\hat{\boldsymbol{\beta}}_0$. Let $\rho_0(r) = \rho_1(r/k_0)$, $\rho(r) = \rho_1(r/k_1)$, and assume that each of the ρ-functions is bounded. The scale estimate $\hat{\sigma}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{r_i\left(\hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}}\right) = 0.5. \tag{1.14}$$

If the ρ-function is biweight, then $k_0 = 1.56$ ensures that the estimator has the asymptotic BP $= 0.5$. Note that an M-estimate minimizes

$$L(\beta) = \sum_{i=1}^{n}\rho\left(\frac{r_i\left(\hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}}\right). \tag{1.15}$$

Let ρ satisfy $\rho \leq \rho_0$. Yohai (1987) showed that if $\hat{\boldsymbol{\beta}}$ satisfies $L(\hat{\boldsymbol{\beta}}) \leq (\hat{\boldsymbol{\beta}}_0)$, then $\hat{\boldsymbol{\beta}}$'s BP is not less than that of $\hat{\boldsymbol{\beta}}_0$. Furthermore, the breakdown point of the MM-estimate depends only on k_0 and the asymptotic variance of the MM-estimate depends only on k_1. We can choose k_1 in order to attain the desired normal efficiency without affecting its breakdown point. In order to let $\rho \leq \rho_0$, we must have $k_1 \geq k_0$; the larger the k_1 is, the higher efficiency

the MM-estimate can attain at the normal distribution.

Maronna, Martin, and Yahai (2006) provides the values of k_1 with the corresponding efficiencies of the biweight ρ-function. Please see the following table for more detail.

Efficiency	0.80	0.85	0.90	0.95
k_1	3.14	3.44	3.88	4.68

However, Yohai (1987) indicates that MM-estimates with larger values of k_1 are more sensitive to outliers than the estimates corresponding to smaller values of k_1. In practice, an MM-estimate with bisquare function and efficiency 0.85 ($k_1 = 3.44$) starting from a bisquare S-estimate is recommended.

1.2.7 Generalized M-Estimates (GM-Estimates)

Mallows GM-estimate

In order to make M-estimate resistent to high leverage outliers, Mallows (1975) proposed Mallows GM-estimate that is defined by

$$\sum_{i=1}^{n} w_i \psi \left\{ \frac{r_i\left(\hat{\boldsymbol{\beta}}\right)}{\hat{\sigma}} \right\} \boldsymbol{x}_i = 0, \tag{1.16}$$

where $\psi(e) = \rho'(e)$ and $w_i = \sqrt{1 - h_i}$ with h_i being the leverage of the ith observation. The weight w_i ensures that the observation with high leverage receives less weight than observation with small leverage. However, even "good" leverage points that fall in line with the pattern in the bulk of the data are down-weighted, resulting in a loss of effiency.

Schweppe GM-estimate

Schweppe GM-estimate (Handschin et al. 1975) is defined by the solution of

$$\sum_{i=1}^{n} w_i \psi \left\{ \frac{r_i\left(\hat{\boldsymbol{\beta}}\right)}{w_i \hat{\sigma}} \right\} \boldsymbol{x}_i = 0, \tag{1.17}$$

which adjusts the leverage weights according to the size of the residual r_i. Carroll and Welsh (1988) proved that the Schweppe estimator is not consistent when the errors are asymmetric. Furthermore, the breakdown points for both Mallows and Schweppe GM-estimates are no more than $1/(p+1)$, where p is the number of unknown parameters.

S1S GM-estimate

Coakley and Hettmansperger (1993) proposed Schweppe one-step (S1S) estimate, which extends from the original Schweppe estimator. S1S estimator is defined as

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 + \left[\sum_{i=1}^{n} \psi' \left(\frac{r_i\left(\hat{\boldsymbol{\beta}}_0\right)}{\hat{\sigma} w_i} \right) \boldsymbol{x}_i \boldsymbol{x}_i' \right]^{-1} \times \sum_{i=1}^{n} \hat{\sigma} w_i \psi \left(\frac{r_i\left(\hat{\boldsymbol{\beta}}_0\right)}{\hat{\sigma} w_i} \right) \boldsymbol{x}_i, \tag{1.18}$$

where the weight w_i is defined in the same way as Schweppe's GM-estimate.

The method for S1S estimate is different from the Mallows and Schweppe GM-estimates in that once the initial estimates of the residuals and the scale of the residuals are given, final M-estimates are calculated in one step rather than iteratively. Coakley and Hettmansperger (1993) recommended to use Rousseeuw's LTS for the initial estimates of the residuals and LMS for the initial estimates of the scale and proved that the S1S estimate gives a breakdown point of BP = 0.5 and results in 0.95 efficiency compared to the OLS estimate under the Gauss-Markov assumption.

1.2.8 R-Estimates

The R-estimate (Jaeckel 1972) minimizes the sum of some scores of the ranked residuals

$$\min \sum_{i=1}^{n} a_n\left(R_i\right) r_i, \tag{1.19}$$

where R_i represents the rank of the ith residual r_i, and $a_n\left(\cdot\right)$ is a monotone score function that satisfies

$$\sum_{i=1}^{n} a_n\left(i\right) = 0. \tag{1.20}$$

R-estimates are scale equivalent which is an advantage compared to M-estimates. However, the optimal choice of the score function is unclear. In addition, most of R-estimates have a breakdown point of BP $= 1/n \to 0$. The bounded influence R-estimator proposed by Naranjo and Hettmansperger (1994) has a fairly high efficiency when the errors have normal distribution. However, it is proved that their breakdown point is no more than 0.2.

1.2.9 REWLSE

Gervini and Yohai (2002) proposed a new class of robust regression method called robust and efficient weighted least squares estimator (REWLSE). REWLSE is much more attractive than many other robust estimators due to its simultaneously attaining maximum breakdown point and full efficiency under normal errors. This new estimator is a type of weighted least squares estimator with the weights adaptively calculated from an initial robust estimator.

Consider a pair of initial robust estimates of regression parameters and scale, $\hat{\boldsymbol{\beta}}_0$ and $\hat{\sigma}$ respectively, the standardized residuals are defined as

$$r_i = \frac{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_0}{\hat{\sigma}}.$$

A large value of $|r_i|$ would suggest that $(\boldsymbol{x}_i, y_i)$ is an outlier. Define a measure of proportion

of outliers in the sample

$$d_n = \max_{i>i_0} \left\{ F^+(|r|_{(i)}) - \frac{(i-1)}{n} \right\}^+, \tag{1.21}$$

where $\{\cdot\}^+$ denotes positive part, F^+ denotes the distribution of $|X|$ when $X \sim F$, $|r|_{(1)} \leq \ldots \leq |r|_{(n)}$ are the order statistics of the standardized absolute residuals, and $i_0 = \max\left\{ i : |r|_{(i)} < \eta \right\}$, where η is some large quantile of F^+. Typically $\eta = 2.5$ and the cdf of a normal distribution is chosen for F. Thus those $\lfloor nd_n \rfloor$ observations with largest standardized absolute residuals are eliminated (here $\lfloor a \rfloor$ is the largest integer less than or equal to a).

The adaptive cut-off value is $t_n = |r|_{(i_n)}$ with $i_n = n - \lfloor nd_n \rfloor$. With this adaptive cut-off value, the adaptive weights proposed by Gervini and Yohai (2002) are

$$w_i = \begin{cases} 1, & \text{if } |r_i| < t_n \\ 0, & \text{if } |r_i| \geq t_n. \end{cases} \tag{1.22}$$

Then, the REWLSE is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y}, \tag{1.23}$$

where $\boldsymbol{W} = \text{diag}(w_1, \cdots, w_n)$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, and $\boldsymbol{y} = (y_1, \cdots, y_n)'$.

If the initial regression and scale estimates with BP = 0.5 are chosen, the breakdown point of the REWLSE is also 0.5. Furthermore, when the errors are normally distributed, the REWLSE is asymptotically equivalent to the OLS estimates and hence asymptotically efficient.

1.2.10 Robust regression based on regularization of case-specific parameters

She and Owen (2011) and Lee et al. (2011) proposed a new class of robust regression methods using the case-specific indicators in a mean shift model with regularization method. A mean

shift model for the linear regression is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$

where $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, and the mean shift parameter γ_i is nonzero when the ith observation is an outlier and zero, otherwise.

Due to the sparsity of γ_is, She and Owen (2011) and Lee et al. (2011) proposed to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by minimizing the penalized least squares using L_1 penalty:

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\left\{\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\gamma})\right\}^T \left\{\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\gamma})\right\} + \lambda \sum_{i=1}^{n} |\gamma_i|\,, \tag{1.24}$$

where λ are fixed regularization parameters for $\boldsymbol{\gamma}$. Given the estimate $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\beta}}$ is the OLS estimate with $\mathbf{y}$ replaced by $\mathbf{y} - \boldsymbol{\gamma}$. For a fixed $\hat{\boldsymbol{\beta}}$, the minimizer of (1.24) is $\hat{\gamma}_i = sgn(r_i)(|r_i| - \lambda)_+$, that is,

$$\hat{\gamma}_i = \begin{cases} 0, & \text{if } |r_i| \le \lambda \\ y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}, & \text{if } |r_i| > \lambda. \end{cases}$$

Therefore, the solution of (1.24) can be found by iteratively updating the above two steps. She and Owen (2011) and Lee et al. (2011) proved that the above estimate is in fact equivalent to the M-estimate if Huber's ψ function is used. However, their proposed robust estimates are based on different perspective and can be extended to many other likelihood based models.

Note, however, the monotone M-estimate is not resistent to the high leverage outliers. In order to overcome this problem, She and Owen (2011) further proposed to replace the L_1 penalty in (1.24) by a general penalty. The objective function is then defined by

$$L_p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\left\{\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\gamma})\right\}^T \left\{\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\gamma})\right\} + \sum_{i=1}^{n} p_\lambda(|\gamma_i|), \tag{1.25}$$

where $p_\lambda(|\cdot|)$ is any penalty function which depends on the regularization parameter λ. We can find $\hat{\boldsymbol{\gamma}}$ by defining a thresholding function $\Theta(\boldsymbol{\gamma}; \lambda)$ (She and Owen 2009). She and Owen (2009, 2011) proved that for a specific thresholding function, we can always find the corresponding penalty function. For example, the soft, hard, and smoothly clipped absolute deviation (SCAD; Fan and Li, 2001) thresholding solutions of $\boldsymbol{\gamma}$ correspond to L_1, Hard, and SCAD penalty functions, respectively. Minimizing the equation (1.25) yields a sparse $\hat{\boldsymbol{\gamma}}$ for outlier detection and a robust estimate of $\boldsymbol{\beta}$. She and Owen (2011) showed that the proposed estimates of (1.25) with hard or SCAD penalties are equivalent to the M-estimates with certain redescending ψ functions and thus will be resistent to high leverage outliers if a high breakdown point robust estimates are used as the initial values.

1.3 Examples

In this section, we compare different robust methods and report the mean squared errors (MSE) of the parameter estimates for each estimation method. We compare the OLS estimate with seven other commonly used robust regression estimates: the M estimate using Huber's ψ function (M_H), the M estimate using Tukey's bisquare function (M_T), the S estimate, the LTS estimate, the LMS estimate, the MM estimate (using bisquare weights and $k_1 = 4.68$), and the REWLSE. Note that we didn't include the case-specific regularization methods proposed by She and Owen (2011) and Lee et al. (2011) since they are essentially equivalent to M-estimators (She and Owen (2011) did show that their new methods have better performance in detecting outliers in their simulation study).

Example 1. We generate n samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ from the model

$$Y = X + \varepsilon,$$

where $X \sim N(0,1)$. In order to compare the performance of different methods, we consider the following six cases for the error density of ε:

Case I: $\varepsilon \sim N(0,1)$- standard normal distribution.

Case II: $\varepsilon \sim t_3$ - t-distribution with degrees of freedom 3.

Case III: $\varepsilon \sim t_1$ - t-distribution with degrees of freedom 1 (Cauchy distribution).

Case IV: $\varepsilon \sim 0.95N(0,1) + 0.05N(0,10^2)$ - contaminated normal mixture.

Case V: $\varepsilon \sim$ N (0,1) with 10% identical outliers in y direction (where we let the first 10% of $y's$ equal to 30).

Case VI: $\varepsilon \sim$ N (0,1) with 10% identical high leverage outliers (where we let the first 10% of $x's$ equal to 10 and their corresponding $y's$ equal to 50).

Tables 1 and 2 report the mean squared errors (MSE) of the parameter estimates for each estimation method with sample size $n = 20$ and 100, respectively. The number of replicates is 200. From the tables, we can see that MM and REWLSE have the overall best performance throughout most cases and they are consistent for different sample sizes. For Case I, LSE has the smallest MSE which is reasonable since under normal errors LSE is the best estimate; M_H, M_T, MM, and REWLSE have similar MSE to LSE, due to their high efficiency property; LMS, LTS, and S have relative larger MSE due to their low efficiency. For Case II, M_H, M_T, MM, and REWLSE work better than other estimates. For Case III, LSE has much larger MSE than other robust estimators; M_H, M_T, MM, and REWLSE have similar MSE to S. For Case IV, M, MM, and REWLSE have smaller MSE than others. From Case V, we can see that when the data contain outliers in the y-direction, LSE is much worse than any other robust estimates; MM, REWLSE, and M_T are better than other robust estimators. Finally for Case VI, since there are high leverage outliers, similar to LSE, both M_T and M_H perform poorly; MM and REWLSE work better than other robust estimates.

In order to better compare the performance of different methods, Figure 1 shows the plot of their MSE versus each case for the slope (left side) and intercept (right side) parameters for model 1 when sample size $n = 100$. Since the lines for LTS and LMS are above the other lines, S, MM, and REWLSE of the intercept and slopes outperform LTS and LMS estimates throughout all six cases. In addition, the S estimate has similar performance to MM and REWLSE when the error density of ε is Cauchy distribution. However, MM and REWLSE perform better than S-estimates in other five cases. Furthermore, the lines for MM and REWLSE almost overlap for all six cases. It shows that MM and REWLSE are the overall best approaches in robust regression.

Example 2.

$$Y = X_1 + X_2 + X_3 + \varepsilon,$$

where $X_i \sim N(0,1), i = 1,2,3$ and X_i's are independent. We consider the following six cases for the error density of ε:

Case I: $\varepsilon \sim N(0,1)$- standard normal distribution.

Case II: $\varepsilon \sim t_3$ - t-distribution with degrees of freedom 3.

Case III: $\varepsilon \sim t_1$ - t-distribution with degrees of freedom 1 (Cauchy distribution).

Case IV: $\varepsilon \sim 0.95N(0,1) + 0.05N(0,10^2)$ - contaminated normal mixture.

Case V: $\varepsilon \sim N(0,1)$ with 10% identical outliers in y direction (where we let the first 10% of $y's$ equal to 30).

Case VI: $\varepsilon \sim N(0,1)$ with 10% identical high leverage outliers (where we let the first 10% of $x's$ equal to 10 and their corresponding $y's$ equal to 50).

Tables 3 and 4 show the mean squared errors (MSE) of the parameter estimates of each estimation method for sample size $n = 20$ and $n = 100$, respectively. Figure 2 shows the plot of their MSE versus each case for three slopes and the intercept parameters with sample

size $n = 100$. The results in Example 2 tell similar stories to Example 1. In summary, MM and REWLSE have the overall best performance; LSE only works well when there are no outliers since it is very sensitive to outliers; M-estimates (M_H and M_T) work well if the outliers are in y direction but are also sensitive to the high leverage outliers.

Example 3: Next, we use the famous data set found in Freedman et al. (1991) to compare LSE with MM and REWLSE. The data set are shown in Table 5 which contains per capita consumption of cigarettes in various countries in 1930 and the death rates (number of deaths per million people) from lung cancer for 1950. Here, we are interested in how the death rates per million people from lung cancer (dependent variable y) dependent on the consumption of cigarettes per capita (the independent variable x). Figure 1.3 is a scatter plot of the data. From the plot, we can see that USA ($x = 1300, y = 200$) is an outlier with high leverage. We compare different regression parameters estimates by LSE, MM, and REWLSE. Figure 1.3 shows the fitted lines by these three estimates. The LSE line does not fit the bulk of the data, being a compromise between USA observation and the rest of the data, while the fitted lines for the other two estimates almost overlap and give a better representation of the majority of the data.

Table 6 also gives the estimated regression parameters of these three methods for both the complete data and the data without the outlier USA. For LSE, the intercept estimate changes from 67.56 (complete data set) to 9.14 (without outlier) and the slope estimate changes from 0.23 (complete data set) to 0.37 (without outlier). Thus, it is clear that the outlier USA strongly influences LSE. For MM-estimate, after deleting the outlier, the intercept estimate changes slightly but slope estimate remains almost the same. For REWLSE, both intercept and slope estimates remain unchanged after deleting the outlier. In addition, note that REWLSE for the whole data gives almost the same result as LSE without the outlier.

1.4 Discussion

In this article, we describe and compare different available robust methods. Table 7 summarizes the robustness attributes and asymptotic efficiency of most of the estimators we have discussed. Based on Table 7, it can be seen that MM-estimates and REWLSE have both high breakdown point and high efficiency. Our simulation study also demonstrated that MM-estimates and REWLSE have overall best performance among all compared robust methods. In terms of breakdown point and efficiency, GM-estimates (Mallows, Schweppe), Bounded R-estimates, M-estimates, and LAD estimates are less attractive due to their low breakdown points. Although LMS, LTS, S-estimates, and GS-estimates are strongly resistent to outliers, their efficiencies are low. However, these high breakdown point robust estimates such as S-estimates and LTS are traditionally used as the initial estimates for some other high breakdown point and high efficiency robust estimates.

Table 1.1: *MSE of Point Estimates for Example 1 with* $n = 20$

TRUE	OLS	M_H	M_T	LMS	LTS	S	MM	REWLSE
	Case I: $\varepsilon \sim N(0,1)$							
$\beta_0 : 0$	0.0497	0.0532	0.0551	0.2485	0.2342	0.1372	0.0564	0.0645
$\beta_1 : 1$	0.0556	0.0597	0.0606	0.2553	0.2328	0.1679	0.0643	0.0733
	Case II: $\varepsilon \sim t_3$							
$\beta_0 : 0$	0.1692	0.0884	0.0890	0.3289	0.3076	0.1637	0.0856	0.0982
$\beta_1 : 1$	0.1766	0.1041	0.1027	0.4317	0.3905	0.2041	0.1027	0.1189
	Case III: $\varepsilon \sim t_1$							
$\beta_0 : 0$	1003.8	0.2545	0.2146	0.3215	0.2872	0.1447	0.1824	0.1990
$\beta_1 : 1$	1374.1	0.4103	0.3209	0.3659	0.3496	0.1843	0.2996	0.3164
	Case IV: $\varepsilon \sim 0.95N(0,1) + 0.05N(0,10^2)$							
$\beta_0 : 0$	0.3338	0.0610	0.0528	0.2105	0.2135	0.1228	0.0523	0.0538
$\beta_1 : 1$	0.4304	0.0808	0.0644	0.3149	0.2908	0.1519	0.0636	0.0691
	Case V: $\varepsilon \sim N(0,1)$ with outliers in y direction							
$\beta_0 : 0$	9.3051	0.1082	0.0697	0.2752	0.2460	0.1430	0.0671	0.0667
$\beta_1 : 1$	5.5747	0.1083	0.0762	0.2608	0.2029	0.1552	0.0746	0.0801
	Case VI: $\varepsilon \sim N(0,1)$ with high leverage outliers							
$\beta_0 : 0$	0.8045	0.8711	0.8857	0.2161	0.1984	0.1256	0.0581	0.0598
$\beta_1 : 1$	13.426	13.750	13.849	0.3377	0.3019	0.1695	0.0749	0.0749

Table 1.2: *MSE of Point Estimates for Example 1 with* $n = 100$

TRUE	OLS	M_H	M_T	LMS	LTS	S	MM	REWLSE
	Case I: $\varepsilon \sim N(0,1)$							
$\beta_0 : 0$	0.0113	0.0126	0.0125	0.0755	0.0767	0.0347	0.0125	0.0131
$\beta_1 : 1$	0.0096	0.0102	0.0103	0.0693	0.0705	0.0312	0.0103	0.0112
	Case II: $\varepsilon \sim t_3$							
$\beta_0 : 0$	0.0283	0.0154	0.0153	0.0596	0.0659	0.0231	0.0153	0.0170
$\beta_1 : 1$	0.0255	0.0157	0.0164	0.0652	0.0752	0.0356	0.0163	0.0185
	Case III: $\varepsilon \sim t_1$							
$\beta_0 : 0$	40.845	0.0416	0.0310	0.0550	0.0392	0.0201	0.0323	0.0354
$\beta_1 : 1$	39.595	0.0469	0.0387	0.0607	0.0476	0.0274	0.0402	0.0447
	Case IV: $\varepsilon \sim 0.95N(0,1) + 0.05N(0,10^2)$							
$\beta_0 : 0$	0.0650	0.0119	0.0107	0.0732	0.0737	0.0296	0.0107	0.0110
$\beta_1 : 1$	0.0596	0.0126	0.0123	0.0696	0.0775	0.0353	0.0122	0.0134
	Case V: $\varepsilon \sim N(0,1)$ with outliers in y direction							
$\beta_0 : 0$	8.9470	0.0465	0.0107	0.0674	0.0658	0.0283	0.0106	0.0108
$\beta_1 : 1$	0.7643	0.0146	0.0120	0.0611	0.0704	0.0338	0.0119	0.0120
	Case VI: $\varepsilon \sim N(0,1)$ with high leverage outliers							
$\beta_0 : 0$	0.2840	0.2999	0.2983	0.0575	0.0595	0.0234	0.0107	0.0106
$\beta_1 : 1$	13.230	13.591	13.721	0.0624	0.0790	0.0310	0.0127	0.0131

Table 1.3: *MSE of Point Estimates for Example 2 with* $n = 20$

TRUE	OLS	M_H	M_T	LMS	LTS	S	MM	REWLSE
	Case I: $\varepsilon \sim N(0,1)$							
$\beta_0 : 0$	0.0610	0.0659	0.0744	0.3472	0.2424	0.1738	0.0679	0.0800
$\beta_1 : 1$	0.0588	0.0664	0.0752	0.4066	0.3247	0.2299	0.0709	0.1051
$\beta_2 : 1$	0.0620	0.0653	0.0725	0.3557	0.2724	0.2018	0.0716	0.0880
$\beta_3 : 1$	0.0698	0.0719	0.0758	0.3444	0.2657	0.1904	0.0751	0.0999
	Case II: $\varepsilon \sim t_3$							
$\beta_0 : 0$	0.1745	0.1125	0.1168	0.3799	0.3040	0.2326	0.1177	0.1210
$\beta_1 : 1$	0.1998	0.1332	0.1364	0.4402	0.3404	0.2539	0.1311	0.1485
$\beta_2 : 1$	0.1704	0.1203	0.1272	0.4868	0.3831	0.2118	0.1242	0.1461
$\beta_3 : 1$	0.2018	0.1520	0.1732	0.5687	0.4964	0.3145	0.1649	0.2049
	Case III: $\varepsilon \sim t_1$							
$\beta_0 : 0$	248.02	0.3492	0.2579	0.7935	0.4657	0.3615	0.2630	0.2957
$\beta_1 : 1$	209.83	0.4503	0.3713	1.2482	0.9701	0.4355	0.3784	0.4443
$\beta_2 : 1$	93.134	0.4089	0.2936	1.0517	0.6203	0.5086	0.2965	0.3365
$\beta_3 : 1$	374.73	0.4387	0.3206	1.0829	0.7704	0.4717	0.3123	0.4023
	Case IV: $\varepsilon \sim 0.95N(0,1) + 0.05N(0,10^2)$							
$\beta_0 : 0$	0.3245	0.0853	0.0837	0.2820	0.2433	0.1873	0.0785	0.0924
$\beta_1 : 1$	0.3391	0.1026	0.1001	0.4609	0.2875	0.2328	0.0996	0.1047
$\beta_2 : 1$	0.3039	0.0898	0.0938	0.4077	0.3053	0.1887	0.0900	0.1170
$\beta_3 : 1$	0.2618	0.0846	0.0941	0.4560	0.3023	0.2054	0.0900	0.1007
	Case V: $\varepsilon \sim N(0,1)$ with outliers in y direction							
$\beta_0 : 0$	9.9455	0.1442	0.0706	0.3127	0.2334	0.1759	0.0680	0.0713
$\beta_1 : 1$	5.1353	0.1015	0.0636	0.3638	0.2769	0.1508	0.0617	0.0654
$\beta_2 : 1$	5.1578	0.1245	0.0730	0.4647	0.2796	0.1759	0.0690	0.0722
$\beta_3 : 1$	6.0662	0.1273	0.0612	0.3922	0.2733	0.1797	0.0597	0.0654
	Case VI: $\varepsilon \sim N(0,1)$ with high leverage outliers							
$\beta_0 : 0$	1.0096	1.0733	1.1334	0.3339	0.2491	0.1716	0.0821	0.0840
$\beta_1 : 1$	13.663	14.071	14.169	0.4698	0.3126	0.2500	0.1467	0.1031
$\beta_2 : 1$	0.9201	0.9684	1.0108	0.4088	0.2681	0.2064	0.0899	0.1088
$\beta_3 : 1$	0.8538	0.9316	0.9937	0.4411	0.3373	0.2077	0.0709	0.0957

Table 1.4: *MSE of Point Estimates for Example 2 with* $n = 100$

TRUE	OLS	M_H	M_T	LMS	LTS	S	MM	REWLSE
	Case I: $\varepsilon \sim N(0,1)$							
$\beta_0 : 0$	0.0097	0.0108	0.0109	0.0743	0.0690	0.0359	0.0108	0.0119
$\beta_1 : 1$	0.0111	0.0120	0.0121	0.0736	0.0778	0.0399	0.0119	0.0130
$\beta_2 : 1$	0.0100	0.0106	0.0107	0.0713	0.0715	0.0404	0.0107	0.0114
$\beta_3 : 1$	0.0110	0.0116	0.0118	0.0662	0.0712	0.0388	0.0118	0.0121
	Case II: $\varepsilon \sim t_3$							
$\beta_0 : 0$	0.0294	0.0145	0.0159	0.0713	0.0655	0.0330	0.0158	0.0179
$\beta_1 : 1$	0.0464	0.0198	0.0180	0.0651	0.0674	0.0368	0.0181	0.0195
$\beta_2 : 1$	0.0375	0.0183	0.0181	0.0727	0.0733	0.0352	0.0181	0.0195
$\beta_3 : 1$	0.0365	0.0176	0.0167	0.0646	0.0736	0.0344	0.0167	0.0175
	Case III: $\varepsilon \sim t_1$							
$\beta_0 : 0$	36.730	0.0388	0.0287	0.0681	0.0590	0.0317	0.0289	0.0326
$\beta_1 : 1$	31.643	0.0499	0.0351	0.0624	0.0618	0.0262	0.0367	0.0372
$\beta_2 : 1$	41.455	0.0422	0.0337	0.0788	0.0613	0.0321	0.0344	0.0369
$\beta_3 : 1$	29.702	0.0476	0.0317	0.0714	0.0506	0.0320	0.0332	0.0362
	Case IV: $\varepsilon \sim 0.95N(0,1) + 0.05N(0,10^2)$							
$\beta_0 : 0$	0.0591	0.0109	0.0100	0.0656	0.0625	0.0281	0.0100	0.0109
$\beta_1 : 1$	0.0492	0.0122	0.0112	0.0558	0.0643	0.0349	0.0110	0.0115
$\beta_2 : 1$	0.0640	0.0123	0.0110	0.0635	0.0683	0.0337	0.0109	0.0118
$\beta_3 : 1$	0.0696	0.0135	0.0122	0.0573	0.0608	0.0333	0.0122	0.0128
	Case V: $\varepsilon \sim N(0,1)$ with outliers in y direction							
$\beta_0 : 0$	9.1058	0.0560	0.0118	0.0631	0.0579	0.0322	0.0118	0.0120
$\beta_1 : 1$	0.8544	0.0186	0.0137	0.0738	0.0814	0.0377	0.0136	0.0143
$\beta_2 : 1$	0.9538	0.0189	0.0141	0.0672	0.0717	0.0379	0.0140	0.0146
$\beta_3 : 1$	0.8953	0.0193	0.0121	0.0652	0.0696	0.0363	0.0120	0.0123
	Case VI: $\varepsilon \sim N(0,1)$ with high leverage outliers							
$\beta_0 : 0$	0.2673	0.2869	0.2901	0.0632	0.0596	0.0300	0.0114	0.0114
$\beta_1 : 1$	13.259	13.635	13.675	0.0590	0.0658	0.0305	0.0123	0.0127
$\beta_2 : 1$	0.1817	0.1889	0.1922	0.0660	0.0727	0.0344	0.0139	0.0144
$\beta_3 : 1$	0.1546	0.1607	0.1643	0.0668	0.0710	0.0344	0.0107	0.0108

Table 1.5: *Cigarettes data*

Country	Per capita consumption of cigarette	Deaths rates
Australia	480	180
Canada	500	150
Denmark	380	170
Finland	1100	350
GreatBritain	1100	460
Iceland	230	060
Netherlands	490	240
Norway	250	090
Sweden	300	110
Switzerland	510	250
USA	1300	200

Table 1.6: *Regression estimates for Cigarettes data*

	Complete data		Data without USA	
Estimators	Intercept	Slope	Intercept	Slope
LS	67.5609	0.2284	9.1393	0.3687
MM	7.0639	0.3729	5.9414	0.3753
REWLSE	9.1393	0.3686	9.1393	0.3686

Table 1.7: *Breakdown Points and Asymptotic Efficiencies of Various Regression Estimators*

	Estimator	Breakdown Point	Asymptotic Efficiency
High BP	LMS	0.5	0.37
	LTS	0.5	0.08
	S-estimates	0.5	0.29
	GS-estimates	0.5	0.67
	MM-estimates	0.5	0.85
	REWLSE	0.5	1.00
Low BP	GM-estimates(Mallows,Schweppe)	$1/(p+1)$	0.95
	Bounded R-estimates	< 0.2	0.90-0.95
	Monotone M-estimates	$1/n$	0.95
	LAD	$1/n$	0.64
	OLS	$1/n$	1.00

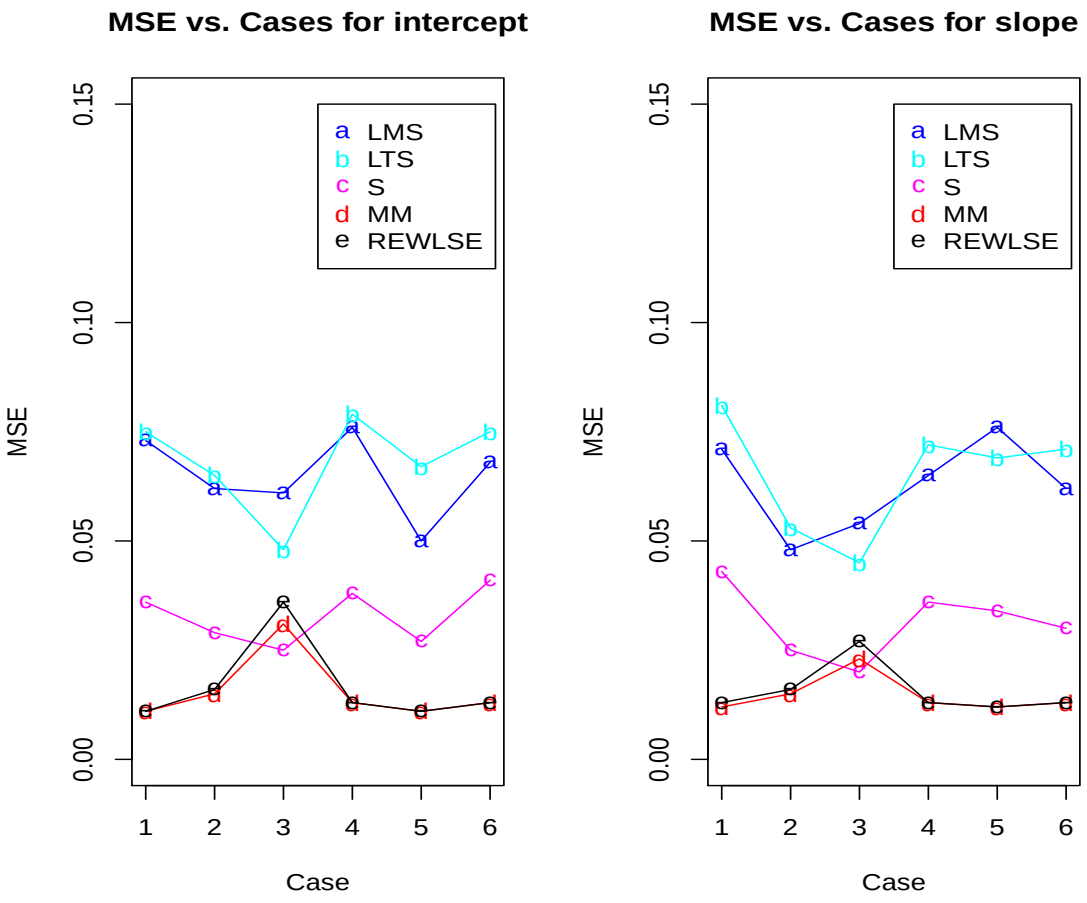

Figure 1.1: *Plot of MSE of intercept (left) and slope (right) estimates vs. different cases for LMS, LTS, S, MM, and REWLSE, for model 1 when* $n = 100$.

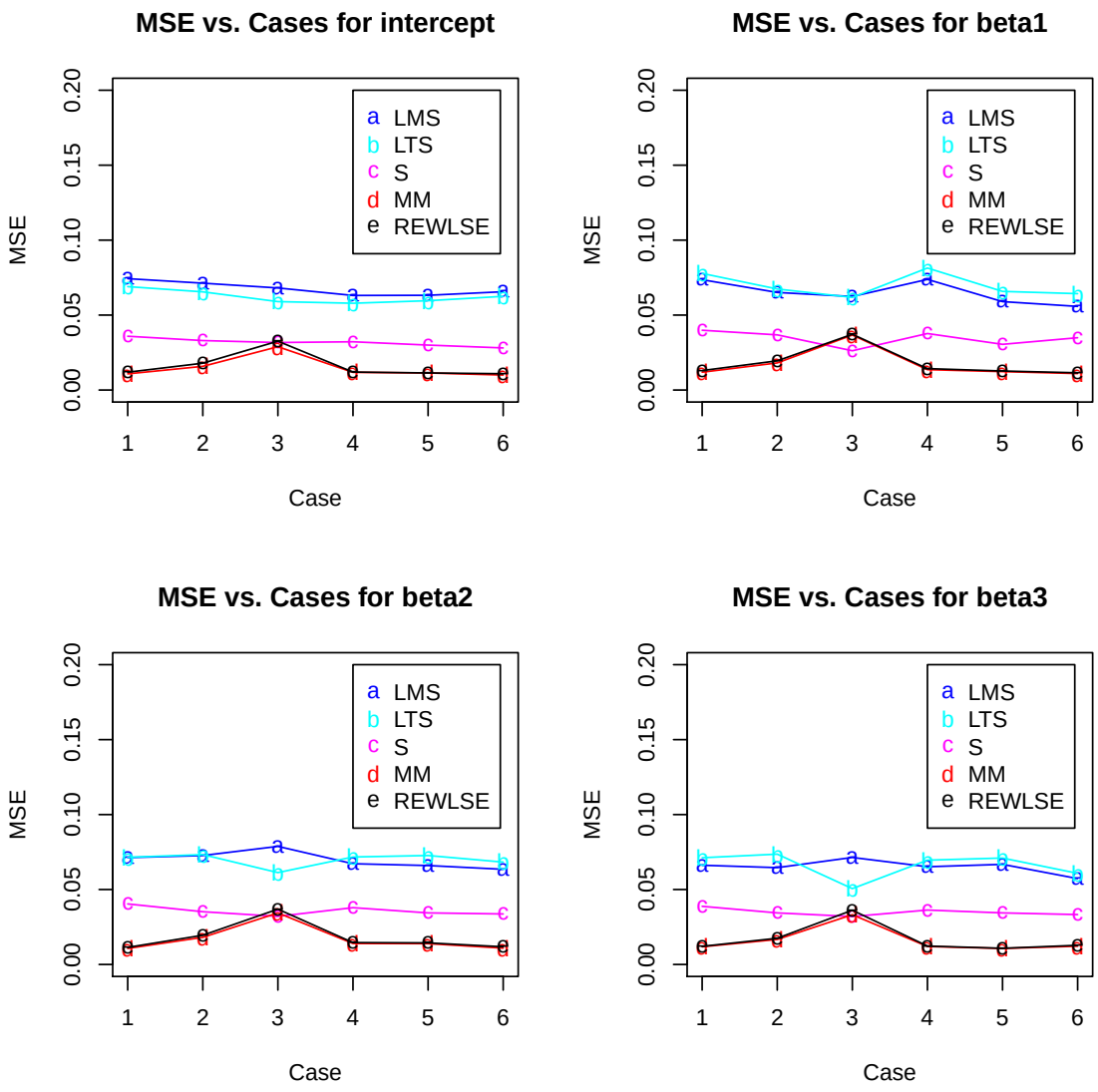

Figure 1.2: *Plot of MSE of different regression parameter estimates vs. different cases for LMS, LTS, S, MM, and REWLSE, for model 2 when* $n = 100$.

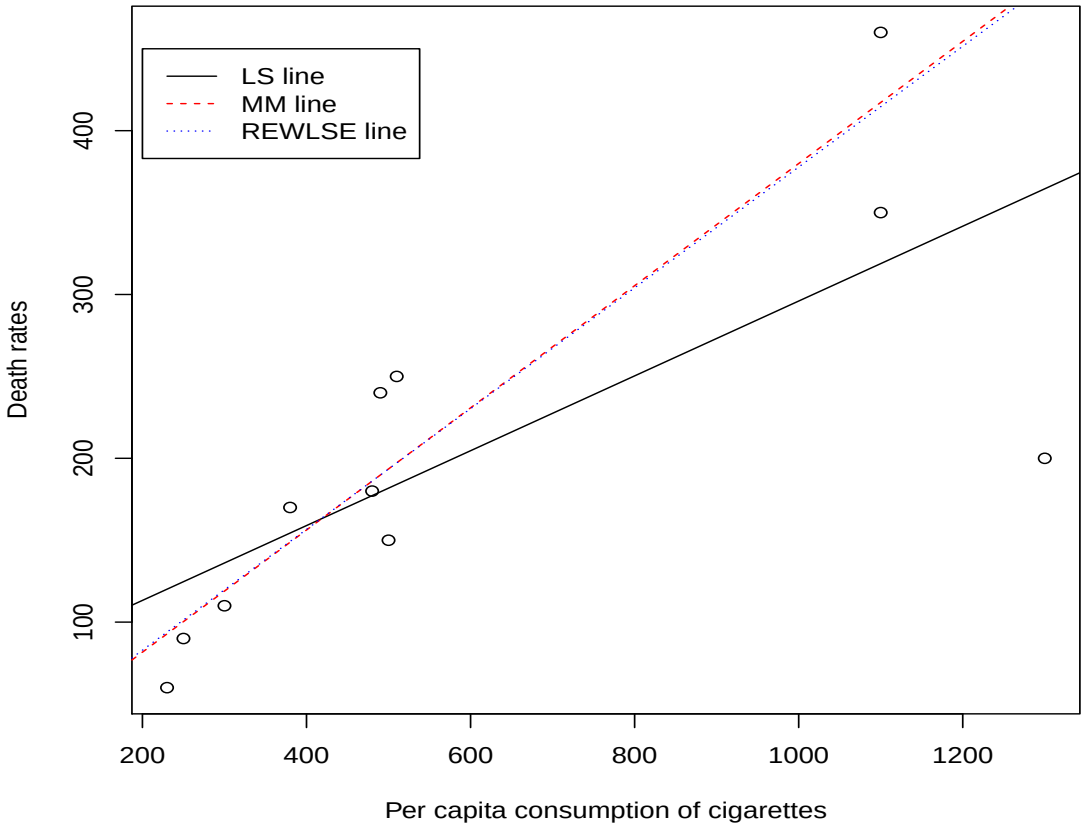

Figure 1.3: *Fitted lines for Cigarettes data*

Chapter 2

Outlier Detection and Robust Mixture Modeling Using Nonconvex Penalized Likelihood

2.1 Introduction

Nowadays finite mixture distributions are increasingly important in modeling a variety of random phenomena (see Everitt and Hand, 1981, Titterington, Smith and Markov, 1985, McLachlan and Basford, 1988, Lindsay, 1995, and Böhning, 1999). The m-component finite normal mixture distribution has probability density

$$f(y;\boldsymbol{\theta}) = \sum_{i=1}^{m} \pi_i \phi(y;\mu_i,\sigma_i^2), \tag{2.1}$$

where $\boldsymbol{\theta} = (\pi_1, \mu_1, \sigma_1; \ldots; \pi_m, \mu_m, \sigma_m)^T$ collects all the unknown parameters, $\phi(\cdot\,;\mu,\sigma^2)$ denotes the density function of $N(\mu,\sigma^2)$, and π_j is the proportion of the jth subpopulation with $\sum_{j=1}^{m} \pi_j = 1$. Given observations $(y_1,\ldots,y_n)$ from model (2.1), the maximum likeli-

hood estimator (MLE) of $\boldsymbol{\theta}$ is given by,

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i; \mu_j, \sigma_j^2) \right\}, \tag{2.2}$$

which does not have an explicit form and is usually calculated by the EM algorithm (Dempster et al. 1977).

The MLE based on the normality assumption possesses many desirable properties such as asymptotic efficiency, however, it is sensitive to the presence of outliers. For the estimation of a single location, many robust methods have been proposed, including the M-estimator (Huber, 1981), the least median of squares (LMS) estimator (Siegel 1982), the least trimmed squares (LTS) estimator (Rousseeuw 1983), the S-estimates (Rousseeuw and Yohai 1984), the MM-estimator (Yohai 1987), and the weighted least squares estimator (REWLSE) (Gervini and Yohai 2002). In contrast, there is much less research on robust estimation of the mixture model, in part because it is not straightforward to replace the log-likelihood in (2.2) by a robust criterion similar to the M-estimation. Peel and McLachlan (2000) proposed a robust mixture modeling using t distribution. Markatou (2000) and Qin and Priebe (2013) proposed using a weighted likelihood for each data point to robustify the estimation procedure for mixture models. Fujisawa and Eguchi (2005) proposed a robust estimation method in normal mixture model using a modified likelihood function. Neykov et al. (2007) proposed robust fitting of mixtures using the trimmed likelihood. Other related robust methods on mixture models include Hennig (2002, 2003), Shen et al. (2004), Bai et al. (2012) and Bashir and Carter (2012).

In this paper, we propose a new robust mixture modelling approach via a mean-shift penalization, which achieves simultaneous outlier detection and robust parameter estimation. A case-specific mean shift parameter vector is added to the mean structure of the mixture model, and it is assumed to be sparse for capturing the rare but possibly severe outlying effects induced by the potential outliers. When the mixture components are assumed to

have equal variances, our method directly extends the robust linear regression approaches proposed by She and Owen (2011) and Lee, MacEachern and Jung (2012). However, even in this case the optimization of the penalized mixture log-likelihood is not trivial, especially for the SCAD penalty (Fan and Li, 2001). For the general case of unequal component variances, the variance heterogeneity of different components complicates the declaration and detection of the outliers, and the naive mean-shift model for the equal variance case is no longer appropriate. We thus propose a scale-free and case-specific mean-shift formulation to achieve the robustness in the general mixture model setup.

2.2 Robust Mixture Model via Mean-Shift Penalization

In this section, we will introduce the proposed robust mixture modelling approach via mean-shift penalization (RMM). To focus on the main idea, we restrict our attention on the normal mixture model. The proposed approach can be readily extended to other mixture models, such as gamma mixture, poisson mixture, and logistic mixture. Due to the inherent difference between the case of equal component variances and the case of unequal component variances, we shall discuss them separately.

2.2.1 RMM for Equal Component Variances

Assume the mixture components have equal variances, i.e., $\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$. The proposed robust mixture model with a mean-shift parameterization is to assume that the observations $(y_1, \ldots, y_n)$ come from the following mixture density

$$f(y_i; \boldsymbol{\theta}, \gamma_i) = \sum_{j=1}^{m} \pi_j \phi(y_i - \gamma_i; \mu_j, \sigma^2), \qquad i = 1, \ldots, n, \tag{2.3}$$

where $\boldsymbol{\theta} = (\pi_1, \mu_1, \ldots, \pi_m, \mu_m, \sigma)^T$ and γ_i is the mean shift parameter for the ith observation, which is nonzero when the ith observation is an outlier and is zero otherwise. Therefore, the sparse estimation of γ_i provides a direct way to identify and accommodate outliers.

Due to the sparsity assumption of γ_i, we propose to maximize the following penalized log-likelihood criterion to conduct model estimation and outlier detection,

$$pl_1(\boldsymbol{\theta}, \boldsymbol{\gamma}) = l_1(\boldsymbol{\theta}, \boldsymbol{\gamma}) - \sum_{i=1}^{n} \frac{1}{w_i} P_\lambda(|\gamma_i|), \tag{2.4}$$

where $l_1(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i - \gamma_i; \mu_j, \sigma^2) \right\}$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$, w_is are the weights to reflect the prior information about how likely it is that the y_is are outliers, $P_\lambda(\cdot)$ is some penalty function used to induce the sparsity in $\boldsymbol{\gamma}$, and λ is a tuning parameter controlling the number of outliers, i.e., the number of nonzero γ_i. To focus on the key idea, we mainly consider $w_1 = w_2 = \ldots = w_n = w$ and discuss the choice of w for different penalty functions.

The commonly used penalty functions include the ℓ_1 norm penalty (Donoho and Johnstone, 1994a; Tibshirani, 1996, 1997) $P_\lambda(\gamma) = \lambda|\gamma|$, the ℓ_0 penalty (Antoniadis, 1997)

$$P_\lambda(\gamma) = \frac{\lambda^2}{2} I(\gamma \neq 0), \tag{2.5}$$

and the SCAD penalty (Fan and Li, 2001)

$$P_\lambda(\gamma) = \begin{cases} \lambda|\gamma|, & \text{if } |\gamma| \leq \lambda, \\ -\left(\frac{\gamma^2 - 2a\lambda|\gamma| + \lambda^2}{2(a-1)} \right), & \text{if } \lambda < |\gamma| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\gamma| > a\lambda, \end{cases} \tag{2.6}$$

where a is a constant usually set to be 3.7. In penalized estimation, each of the above penalty forms corresponds to a thresholding rule, e.g., ℓ_1 penalization corresponds to a soft-threshing rule and ℓ_0 penalization corresponds to a hard-thresholding rule. We mainly focus

on the nonconvex hard penalty and SCAD penalty, due to their superior performance in sparse estimation.

We propose a thresholding embedded EM algorithm to maximize the objective function (2.4). Let

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

and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})$. The complete penalized log-likelihood function based on the complete data $\{(y_i, \boldsymbol{z}_i)), i = 1, 2, \ldots, n\}$ is

$$pl_1^c(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log \left\{ \pi_j \phi(y_i - \gamma_i; \mu_j, \sigma^2) \right\} - \sum_{i=1}^{n} \frac{1}{w} P_\lambda(|\gamma_i|). \tag{2.7}$$

Based on the construction of the EM algorithm, in the E step, given the current estimate $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ at the kth iteration, we need to find the condition expectation of the complete penalized log-likelihood function (2.7), i.e., $\mathrm{E}\{pl_1^c(\boldsymbol{\theta}, \boldsymbol{\gamma}) \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$, which simplifies to the calculation of $\mathrm{E}(z_{ij}|y_i; \boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)})$:

$$p_{ij}^{(k+1)} = \mathrm{E}(z_{ij}|y_i; \boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)}) = \frac{\pi_j^{(k)} \phi(y_i - \gamma_i^{(k)}; \mu_j^{(k)}, \sigma^{2^{(k)}})}{\sum_{j=1}^{m} \pi_j^{(k)} \phi(y_i - \gamma_i^{(k)}; \mu_j^{(k)}, \sigma^{2^{(k)}})}.$$

In the M step, we then update $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ by maximizing $\mathrm{E}\{pl_1^c(\boldsymbol{\theta}, \boldsymbol{\gamma}) \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$. There is no explicit solution, except for the π_js: $\pi_j^{(k+1)} = \sum_{i=1}^{n} p_{ij}^{(k+1)}/n$. We propose to iterate the following two steps until convergence to get $\{\mu_j^{(k+1)}, j = 1, \ldots, m, \sigma^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}\}$:

1. Given μ_js and σ, update $\boldsymbol{\gamma}$ by maximizing

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i; \mu_j, \sigma^2) - \sum_{i=1}^{n} \frac{1}{w} P_\lambda(|\gamma_i|),$$

which is equivalently to minimizing

$$\frac{1}{2}\left\{\gamma_i - \sum_{j=1}^{m} p_{ij}^{(k+1)}(y_i - \mu_j)\right\}^2 + \frac{1}{w}\sigma^2 P_\lambda\left(|\gamma_i|\right), \tag{2.8}$$

separately for each γ_i.

2. Given $\boldsymbol{\gamma}$, the μ_js and σ are updated by

$$\mu_j \leftarrow \frac{\sum_{i=1}^{n} p_{ij}^{(k+1)}(y_i - \gamma_i)}{\sum_{i=1}^{n} p_{ij}^{(k+1)}}, j = 1, \ldots, m,$$
$$\sigma^2 \leftarrow \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij}^{(k+1)}(y_i - \gamma_i - \mu_j)^2}{n}.$$

Note that for the hard penalty, $w^{-1}\sigma^2 P_\lambda\left(|\gamma_i|\right) = \sigma P_{\lambda^*}\left(|\gamma_i|\right)$, where $\lambda^* = \frac{\sigma}{\sqrt{w}}\lambda$. Therefore, if λ is chosen data adaptively, we can simply set $w = 1$ for the hard penalty. However, for the SCAD penalty, such property does not hold and the solution may be affected nonlinearly by the ratio σ^2/w. In order to mimic the unscaled SCAD and use the same a value as suggested by Fan and Li (2001), we need to make sure σ^2/w is close to 1. Therefore, we propose to set $w = \hat{\sigma}^2$ for SCAD penalty, where $\hat{\sigma}^2$ is a robust estimate of σ^2 such as the estimate from the trimmed likelihood estimation (Neykov et al. 2007) or the estimator using the hard penalty assuming $w = 1$.

If the hard penalty is used, (2.8) is minimized by the hard thresholding rule. However, if the SCAD penalty is used, we prove in the following proposition that the minimizer of (2.8) is given by a modified SCAD thresholding rule.

Proposition 1. *Let*

$$\xi_i = \sum_{j=1}^{m} p_{ij}^{(k+1)}(y_i - \mu_j). \tag{2.9}$$

If the penalty function in (2.8) is the hard penalty (2.5), then the thresholding rule to min-

imize (2.8) is

$$\hat{\gamma}_i = \Theta_{hard}(\xi_i; \lambda, \sigma) = \begin{cases} 0, & \text{if } |\xi_i| \leq \sigma\lambda, \\ \xi_i, & \text{if } |\xi_i| > \sigma\lambda. \end{cases}$$

If the penalty function in (2.8) is the SCAD penalty (2.6), then the thresholding rule to minimize (2.8) is

1. *when* $\sigma^2/\hat{\sigma}^2 < a - 1$,

$$\hat{\gamma}_i = \Theta_{SCAD}(\xi_i; \lambda, \sigma) = \begin{cases} sgn(\xi_i)\left(|\xi_i| - \frac{\sigma^2\lambda}{\hat{\sigma}^2}\right)_+, & \text{if } |\xi_i| \leq \lambda + \frac{\sigma^2\lambda}{\hat{\sigma}^2}, \\ \frac{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)\xi_i - sgn(\xi_i)a\lambda}{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)-1}, & \text{if } \lambda + \frac{\sigma^2\lambda}{\hat{\sigma}^2} < |\xi_i| \leq a\lambda, \\ \xi_i, & \text{if } |\xi_i| > a\lambda. \end{cases} \tag{2.10}$$

2. *when* $a - 1 \leq \sigma^2/\hat{\sigma}^2 \leq a + 1$,

$$\hat{\gamma}_i = \Theta_{SCAD}(\xi_i; \lambda, \sigma) = \begin{cases} sgn(\xi_i)\left(|\xi_i| - \frac{\sigma^2\lambda}{\hat{\sigma}^2}\right)_+, & \text{if } |\xi_i| \leq \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda, \\ \xi_i, & \text{if } |\xi_i| > \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda. \end{cases} \tag{2.11}$$

3. *when* $\sigma^2/\hat{\sigma}^2 > a + 1$,

$$\hat{\gamma}_i = \Theta_{SCAD}(\xi_i; \lambda, \sigma) = \begin{cases} 0, & \text{if } |\xi_i| \leq \sqrt{\frac{\sigma^2(a+1)}{\hat{\sigma}^2}}\lambda, \\ \xi_i, & \text{if } |\xi_i| > \sqrt{\frac{\sigma^2(a+1)}{\hat{\sigma}^2}}\lambda. \end{cases} \tag{2.12}$$

The detailed EM algorithm to maximize the penalized log-likelihood (2.4) is summarized in Algorithm 1. The convergence property of the proposed algorithm is summarized in Theorem 2.2.2 below, which follows directly from the property of the EM algorithm, and hence its proof is omitted.

Theorem 2.2.1. *Each iteration of E step and M step of Algorithm 1 monotonically non-decreases the penalized log-likelihood (2.4), i.e.,* $pl_1(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}) \geq pl_1(\boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)})$*, for all* $k \geq 0$.

Algorithm 1 Thresholding Embeded EM Algorithm for Equal Variances Case

Initialize $\boldsymbol{\theta}^{(0)}$ and $\gamma^{(0)}$. Set $k \leftarrow 0$.

repeat

E-Step: Compute the classification probabilities

$$p_{ij}^{(k+1)} = \mathrm{E}(z_{ij}|y_i; \boldsymbol{\theta}^{(k)}) = \frac{\pi_j^{(k)} \phi(y_i - \gamma_i^{(k)}; \mu_j^{(k)}, \sigma^{2^{(k)}})}{\sum_{j=1}^m \pi_j^{(k)} \phi(y_i - \gamma_i^{(k)}; \mu_j^{(k)}, \sigma^{2^{(k)}})}.$$

M-Step: Update $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ by the following two steps:

1.
$$\pi_j^{(k+1)} = \frac{\sum_{i=1}^n p_{ij}^{(k+1)}}{n}, j = 1, \ldots, m.$$

2. Iterating the following steps until convergence to obtain $\{\mu_j^{(k+1)}, j = 1, \ldots, m; \sigma^{2^{(k+1)}}, \boldsymbol{\gamma}^{(k+1)}\}$:

$$(2.a) \qquad \gamma_i \leftarrow \Theta(\xi_i; \lambda, \sigma), i = 1, \ldots, n, \text{ where } \xi_i = \sum_{j=1}^m p_{ij}^{(k+1)}(y_i - \mu_j),$$

$$(2.b) \qquad \mu_j \leftarrow \frac{\sum_{i=1}^n p_{ij}^{(k+1)}(y_i - \gamma_i)}{\sum_{i=1}^n p_{ij}^{(k+1)}}, j = 1, \ldots, m,$$

$$(2.c) \qquad \sigma^2 \leftarrow \frac{\sum_{j=1}^m \sum_{i=1}^n p_{ij}^{(k+1)}(y_i - \gamma_i - \mu_j)^2}{n}.$$

$k \leftarrow k + 1$.

until convergence.

2.2.2 RMM for Unequal Component Variances

When the component variances are unequal, the naive mean shift model (3.3) can not be directly applied, due to the scale difference in the mixture components. To illustrate further,

suppose the standard deviation in the first component is 1 and the standard deviation in the second component is 4. If some weighted residual ξ_i, defined in (2.9), equals to 5, then the ith observation is considered as an outlier if it is from the first component but should not be regarded as an outlier if it belongs to the second component. This suggests that the declaration of outliers in a mixture model shall take into account both the centers and the variabilities of all the components, i.e., an observation is considered as an outlier in the mixture model only if it is far away from all the component centers judged by their own component variabilities.

We propose the following scale-free mean shift model to incorporate the information on component variability,

$$f(y_i;\boldsymbol{\theta},\gamma_i)=\sum_{j=1}^{m}\pi_j\phi(y_i-\gamma_i\sigma_j;\mu_j,\sigma_j^2),\qquad i=1,\ldots,n, \tag{2.13}$$

where with some abuse of notation, $\boldsymbol{\theta}$ is redefined as $\boldsymbol{\theta}=(\pi_1,\mu_1,\sigma_1,\ldots,\pi_m,\mu_m,\sigma_m)^T$. Given observations $(y_1,y_2,\ldots,y_n)$, we estimate the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ by maximizing the following penalized log-likelihood function:

$$pl_2(\boldsymbol{\theta},\boldsymbol{\gamma})=l_2(\boldsymbol{\theta},\boldsymbol{\gamma})-\sum_{i=1}^{n}\frac{1}{w_i}P_\lambda(|\gamma_i|), \tag{2.14}$$

where $l_2(\boldsymbol{\theta},\boldsymbol{\gamma})=\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{m}\pi_j\phi(y_i-\gamma_i\sigma_j;\mu_j,\sigma_j^2)\right\}$. Since the γ_is in (2.14) are scale free, for simplicity we set $w_1=w_2=\ldots=w_n=1$ when no prior information is available.

We again propose a thresholding embedded EM algorithm to maximize (2.14). The complete penalized log-likelihood function constructed based on the complete data $\{(\boldsymbol{z}_i,\boldsymbol{y}_i),i=1,2,\ldots,n\}$, with the same setting of the binary label $\boldsymbol{z}_{ij}$ as the equal component variances case, is

$$pl_2^c(\boldsymbol{\theta},\boldsymbol{\gamma})=\sum_{i=1}^{n}\sum_{j=1}^{m}z_{ij}\log\left\{\pi_j\phi(y_i-\gamma_i\sigma_j;\mu_j,\sigma_j^2)\right\}-\sum_{i=1}^{n}P_\lambda(|\gamma_i|). \tag{2.15}$$

Similar to the arguments in Section 2.2.1, in the E step of the $(k+1)$th iteration, we only need to compute $\mathrm{E}\{pl_2^c(\boldsymbol{\theta}, \boldsymbol{\gamma}) \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$, which simplifies to the calculation of

$$p_{ij}^{(k+1)} = \mathrm{E}(z_{ij}|y_i; \boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)}) = \frac{\pi_j^{(k)} \phi(y_i - \gamma_i^{(k)} \sigma_j^{(k)}; \mu_j^{(k)}, \sigma_j^{2^{(k)}})}{\sum_{j=1}^m \pi_j^{(k)} \phi(y_i - \gamma_i^{(k)} \sigma_j^{(k)}; \mu_j^{(k)}, \sigma_j^{2^{(k)}})}.$$

In the M step, we need to update $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ by maximizing $\mathrm{E}\{pl_2^c(\boldsymbol{\theta}, \boldsymbol{\gamma}) \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)}\}$. Therefore, $\pi_j^{(k+1)} = \sum_{i=1}^n p_{ij}^{(k+1)}/n$, and $\{\mu_j^{(k+1)}, j = 1, \ldots, m, \sigma_j^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}\}$ can be found by iterating the following three steps:

1. Given $\boldsymbol{\gamma}$ and σ_js, μ_js are updated by

$$\mu_j \leftarrow \frac{\sum_{i=1}^n p_{ij}^{(k+1)} (y_i - \gamma_i \sigma_j)}{\sum_{i=1}^n p_{ij}^{(k+1)}}, j = 1, \ldots, m.$$

2. Given $\boldsymbol{\gamma}$ and μ_js, σ_js are updated by

$$\sigma_j^2 \leftarrow \arg\max_{\sigma_j} \sum_{i=1}^n p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i \sigma_j; \mu_j, \sigma_j^2), j = 1, \ldots, m. \tag{2.16}$$

3. Given μ_js and σ_js, update $\boldsymbol{\gamma}$ by minimizing

$$\frac{1}{2} \left[\left\{ \gamma_i - \sum_{j=1}^m \frac{p_{ij}^{(k+1)}}{\sigma_j} (y_i - \mu_j) \right\}^2 \right] + P_\lambda \left(|\gamma_i| \right). \tag{2.17}$$

   separately for each γ_i.

Note that, unlike the equal variances case, the update of σ_j^2 in (2.16) does not have explicit solution and requires some one-dimensional numerical algorithm to sovle, e.g., the Newton-Raphson method. To minimize (2.17), we have the following thresholding solutions for using

the hard and SCAD penalties, respectively:

$$\hat{\gamma}_i = \Theta^*_{hard}(\xi_i;\lambda) = \begin{cases} 0, & \text{if } |\xi_i| \le \lambda, \\ \xi_i, & \text{if } |\xi_i| > \lambda, \end{cases}$$

$$\hat{\gamma}_i = \Theta^*_{SCAD}(\xi_i;\lambda) = \begin{cases} sgn(\xi_i)(|\xi_i| - \lambda)_+, & \text{if } |\xi_i| \le 2\lambda, \\ \frac{(a-1)\xi_i - sgn(\xi_i)a\lambda}{a-2}, & \text{if } 2\lambda < |\xi_i| \le a\lambda, \\ \xi_i, & \text{if } |\xi_i| > a\lambda. \end{cases}$$

where

$$\xi_i = \sum_{j=1}^{m} \frac{p_{ij}^{(k+1)}}{\sigma_j}(y_i - \mu_j).$$

The detailed thresholding embeded EM algorithm to maximize (2.14) can be summarized in Algorithm 2, with its convergence property summarized in Theorem 2.

Theorem 2.2.2. *Each iteration of E step and M step of Algorithm 2 monotonically non-decreases the corresponding objective function, i.e.,* $pl_2(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}) \ge pl_2(\boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)})$*, for all* $k \ge 0$.

2.2.3 Tuning Parameter Selection

In order to apply (2.4) and (2.14) in practice, we need to choose the tuning parameter λ. Here, we provide a data adaptive way to select λ based on the Bayesian information criterion (BIC):

$$BIC(\lambda) = -l_j(\lambda) + \log(n)df(\lambda), \tag{2.18}$$

where $j = 1$ or 2, $l_j(\lambda) = \max_{\boldsymbol{\theta},\boldsymbol{\gamma}} l_j(\boldsymbol{\theta}, \boldsymbol{\gamma})$ is the maximum mixture log-likelihood function for a given tuning parameter λ, and $df(\lambda)$ is the model degrees of freedom which is estimated by the sum of the number of nonzero γ values and the number of mixture component parameters (She and Owen, 2011). The optimal tuning parameter λ is chosen by minimizing BIC(λ)

Algorithm 2 Thresholding Embeded EM Algorithm for Unequal Variances Case

Initialize $\boldsymbol{\theta}^{(0)}$ and $\gamma^{(0)}$. Set $k \leftarrow 0$.

repeat

E-Step: Compute the classification probabilities

$$p_{ij}^{(k+1)} = E(z_{ij}|y_i; \boldsymbol{\theta}^{(k)}) = \frac{\pi_j^{(k)} \phi(y_i - \gamma_i^{(k)} \sigma_j^{(k)}; \mu_j^{(k)}, \sigma_j^{2^{(k)}})}{\sum_{j=1}^m \pi_j^{(k)} \phi(y_i - \gamma_i^{(k)} \sigma_j^{(k)}; \mu_j^{(k)}, \sigma_j^{2^{(k)}})}.$$

M-Step: Update $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ by the following two steps:

1.
$$\pi_j^{(k+1)} = \frac{\sum_{i=1}^n p_{ij}^{(k+1)}}{n}, j = 1, \ldots, m.$$

2. Iterating the following steps until convergence to obtain $\{\mu_j^{(k+1)}, \sigma_j^{2^{(k+1)}}, j = 1, \ldots, m, \boldsymbol{\gamma}^{(k+1)}\}$:

$$(2.a) \qquad \gamma_i \leftarrow \Theta^*(\xi_i; \lambda), \text{ where } \xi_i = \sum_{j=1}^m p_{ij}^{(k+1)} (y_i - \mu_j)/\sigma_j,$$

$$(2.b) \qquad \mu_j \leftarrow \frac{\sum_{i=1}^n p_{ij}^{(k+1)} (y_i - \gamma_i \sigma_j)}{\sum_{i=1}^n p_{ij}^{(k+1)}},$$

$$(2.c) \qquad \sigma_j^2 \leftarrow \arg\max_{\sigma_j} \sum_{i=1}^n p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i \sigma_j; \mu_j, \sigma_j^2).$$

$k \leftarrow k + 1$.

until convergence

over a grid of 100 λ values, equally spaced on the log scale between $\lambda_{\min}$ and $\lambda_{\max}$, where $\lambda_{\max}$ is some large value of λ which corresponds to all zero values of γ_i and $\lambda_{\min}$ is some small value of λ which corresponds to all nonzero values of γ_i.

2.3 Simulation

We conduct several simulation studies to demonstrate the effectiveness of the proposed method and compare it with some of existing estimation methods. We consider two examples: example 1 is equal variance case and example 2 is unequal variance case. For both examples, we set the sample size $n = 400$. For nonzero γ, the absolute value of γ is generated by a uniform distribution either between 5 and 7 or between 11 and 13. We consider two cases of the proportion of outliers: 5% outliers and 10% outliers by adding nonzero γ_is. The number of replicates is 200 for each simulation setting.

Example 1: The samples $(y_1, y_2, \ldots, y_n)$ are generated from model (2.3) with $\pi_1 = 0.3$, $\mu_1 = 0$, $\pi_2 = 0.7$, $\mu_2 = 8$, and $\sigma = 1$. The observations $(y_1, y_2, \ldots, y_{n_1})$ are assigned to the first component (where n_1 is generated by a binomial distribution with $n = 400$ and $p = 0.3$ and n_1 is the sum of 1's) and the rest of observations, $(y_{n_1+1}, \ldots, y_n)$, are assigned to the second component. For 5% outliers case (i.e., 20 nonzero γ_is), the first 5 observations are set to be outliers in the first component and the last 15 observations are set to be outliers in the second component; for 10% outliers case (i.e., 40 nonzero γ_is), the first 10 observations are set to be outliers in the first component and the last 30 observations are set to be outliers in the second component.

Example 2: The samples $(y_1, y_2, \ldots, y_n)$ are generated from model (2.13) with $\sigma_1 = 1$ and $\sigma_2 = 2$. All other model parameters and simulation settings are the same as in Example 1.

2.3.1 Methods and Evaluation Measures

We compare our proposed RMM method using hard and SCAD penalty to one existing robust approach and the traditional MLE. To check the performance of the selection of tuning parameter λ, we also report the "oracle" estimates for both hard and SCAD penalty which are the estimates closest to the true values in the solution path. The seven methods we compared are listed below:

1. traditional MLE assuming the error has normal density (MLE),
2. trimmed likelihood estimator (TLE) proposed by Neykov et al. (2007) with the percentage of trimmed data α set to 0.05 ($\text{TLE}_{0.05}$),
3. TLE with the percentage of trimmed data α set to 0.10 ($\text{TLE}_{0.10}$),
4. the proposed RMM using the hard penalty (Hard),
5. the proposed RMM using the SCAD penalty (SCAD),
6. the oracle estimate using the hard penalty (Hard_{oracle}),
7. the oracle estimate using the SCAD penalty (SCAD_{oracle}).

Note that unlike TLE, the proposed RMM used the data adaptive tuning parameter λ. In addition, unlike our proposed methods, TLE method requires a cutoff value to identify which residuals are outliers. A fixed choice of $\eta = 2.5$ in various situations is applied (Gervini and Yohai, 2002) to identify outliers for TLE method.

To evaluate the performance of different estimators, we report the median squared errors (MeSE) of the parameter estimates. Similar to She and Owen (2011), to evaluate the outlier detection performance, we report (1) the average proportions of masking (M), i.e., the fraction of undetected outliers, (2) the average proportions of swapping (S), i.e., the fraction of good points labeled as outliers, and (3) the joint detection rate (JD), i.e., the proportion of simulations with 0 masking. Ideally, M $\approx$ 0, S $\approx$ 0, and JD $\approx$ 1.

Note, however, for mixture models, there are well known label switching issues (Celeux, et al., 2000; Stephens, 2000; Yao and Lindsay, 2009; Yao, 2012). In our simulation study, the labels are determined by minimizing the distance to true parameter values.

2.3.2 Results

Simulation results of Example 1 are summarized in Table 2.1 – Table 2.4. Tables 2.1 and 2.3 report the three fractions of outlier detection and Tables 2.2 and 2.4 report the median of squared errors (MeSE) of parameter estimates for each estimation method. For equal variance case, both hard and SCAD have similar results to "oracle" estimators. In case I (5% outliers) with either large $|\gamma|$ or small $|\gamma|$, hard, SCAD, $\text{TLE}_{0.05}$, and $\text{TLE}_{0.10}$ gain ideal joint outlier detection rate and fraction of undetected true outliers, and small swamping rate but $\text{TLE}_{0.10}$ has bigger MeSE of parameter estimates with large $|\gamma|$. In case II (10% outliers), hard, SCAD, and $\text{TLE}_{0.10}$ get similar performance in terms of both outlier identification and MeSE. $\text{TLE}_{0.05}$ fails to work with either large or small $|\gamma|$ due to the smaller α setting (less than the proportion of outliers).

Simulation results of Example 2 are summarized in Table 2.5 – Table 2.8. Tables 2.5 and 2.7 report the three fractions of outlier detection and Tables 2.6 and 2.8 report the median of squared errors (MeSE) of parameter estimates for each estimation method. In case I (5% outliers), Hard, SCAD, $\text{TLE}_{0.05}$, and $\text{TLE}_{0.10}$ obtain similar outlier identifying rates. Hard, SCAD, and $\text{TLE}_{0.05}$ have similar MeSE, while $\text{TLE}_{0.10}$ has bigger MeSE for σ. In case II (10% outliers), Hard, SCAD, and $\text{TLE}_{0.10}$ have the similar outlier identifying rates and MeSE for π and μ but $\text{TLE}_{0.10}$ has bigger MeSE for σ with large $|\gamma|$; SCAD fails to work with small $|\gamma|$ but its solution path does include good estimates of the parameters because SCAD_{oracle} has similar results to hard. Therefore, a better method to choose the tuning parameter might be able to improve the performance of SCAD. Like the equal variance case, $\text{TLE}_{0.05}$ performs poorly when there are 10% outliers in the data.

In summary, the proposed Hard has comparable performance to the oracle TLE, that

used the correct trimming proportion α, in the simulation studies in terms of both outlier identifying and MeSE (MSE). The proposed SCAD works well for equal variance case. For unequal variance case, SCAD can still work well when there are 5% outliers or the absolute value of γ is big, but does not work properly when the proportion of outliers in data is 10% and the magnitude of γ is small. A modification on tuning parameter criterion may possibly solve this problem, since the oracle SCAD works well for all cases. The proposed RMM using ℓ_1 norm penalty works with large absolute value of γ when there are 5% outliers but fails to work with small absolute value of γ and more than 5% outliers (The results of soft are omitted here); this agrees with She and Owen (2011). As we expect, the traditional MLE fails to work when there are one or more outliers in the data.

2.4 Real Data Application

We further apply the proposed robust procedure to Acidity dataset (Crawford, 1994; Crawford et al., 1992). The observations are the logarithms of an acidity index measured in a sample 155 lakes in north-central Wisconsin. More details on the data analysis can be found in Crawford (1994), Crawford et al. (1992), and Richardson and Green (1997). Figure 1 shows the histogram of Acidity dataset. Based on the result of Richardson and Green (1997), the posterior for three components was largest. Hence we fit this data set by a three-component normal mixture by the traditional MLE and the proposed RMM using HARD penalty.

Table 2.9 reports the parameter estimates on the Acidity data set. For the original data where there are no outliers, the proposed Hard has similar parameter estimates to that of the traditional MLE. To see the effects of outliers on Hard and MLE, similar to Peel and McLachlan (2000), we add one outlier ($y = 12$) to the original data. Based on Table 9, the proposed Hard is not influenced by the outlier and gives similar parameter estimates to the case of no outliers. However, MLE gives different parameter estimates from the case of no

outliers. In addition, note that MLE provides the same component means for the first and second components. We further add three identical outliers ($y = 12$) to the data. As we expect, Hard still provides similar estimates to the case of no outliers. However, MLE fits a new component to the outliers and gives totally different estimates from the case of no outliers.

2.5 Discussion

The main contribution of this paper is to propose a robust mixture via mean shift penalization model (RMM). In addition, we proposed a thresholding embedded EM algorithm to find the proposed robust estimate. Based on the simulation studies and real data analysis, we can see that RMM with Hard penalty has similar performance to TLE that uses an oracle trimming proportion. Note, however, RMM can adaptively choose the tuning parameter λ based on BIC. In addition, the proposed RMM can naturally detect outliers corresponding to nonzero γ_is.

In this article, we mainly focus on normal mixture model. We think the proposed robust procedure RMM can be also extended to other mixture models, such as mixtures of binomial and mixtures of poisson. In addition, the proposed RMM can be also extended to mixture of linear regression models and mixture of generalized linear models.

Appendix

2.5.1 Proof of Equation (2.8)

The estimate of $\boldsymbol{\gamma}$ is updated by maximizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i; \mu_j, \sigma^2) - \sum_{i=1}^{n} \frac{1}{w} P_\lambda(|\gamma_i|).$$

The problem is separable in each γ_i. Thus each γ_i can be updated by minimizing

$$-\sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i; \mu_j, \sigma^2) + \frac{1}{w} P_\lambda(|\gamma_i|).$$

Note that

$$\begin{aligned}
\log \phi\left(y_i - \gamma_i; \mu_j, \sigma^2\right) &= \log\left[\left(\sigma^2\right)^{-\frac{1}{2}} \cdot \exp\left\{-\frac{\left(y_i - \gamma_i - \mu_j\right)^2}{2\sigma^2}\right\}\right] + \text{const} \\
&= -\frac{1}{2}\log\left(\sigma^2\right) - \frac{\left(y_i - \gamma_i - \mu_j\right)^2}{2\sigma^2} + \text{const.}
\end{aligned}$$

Thus, the solution of γ has the following form,

$$\hat{\gamma}_i = \arg\min_{\gamma_i} \sum_{j=1}^{m} p_{ij} \left\{\frac{1}{2}\log\left(\sigma^2\right) + \frac{\left(y_i - \gamma_i - \mu_j\right)^2}{2\sigma^2}\right\} + \frac{1}{w} P_\lambda\left(|\gamma_i|\right).$$

Since $\frac{1}{2}\sum_{j=1}^{m} p_{ij} \log\left(\sigma^2\right)$ does not depend on γ, we can ignore this term. The second term is

$$\begin{aligned}
\sum_{j=1}^{m} p_{ij} \frac{(y_i - \gamma_i - \mu_j)^2}{2\sigma^2} &= \frac{1}{2\sigma^2} \sum_{j=1}^{m} p_{ij} \left\{\gamma_i^2 - 2\left(y_i - \mu_j\right)\gamma_i + \left(y_i - \mu_j\right)^2\right\} \\
&= \frac{1}{2\sigma^2}\left[\left\{\gamma_i - \frac{\sum_{j=1}^{m} p_{ij}(y_i - \mu_j)}{\sum_{j=1}^{m} p_{ij}}\right\}^2 + \text{const}\right] \\
&= \frac{1}{2\sigma^2}\left[\left\{\gamma_i - \sum_{j=1}^{m} p_{ij}(y_i - \mu_j)\right\}^2 + \text{const}\right],
\end{aligned}$$

where $\sum_{j=1}^{m} p_{ij} = 1$. It follows that

$$\hat{\gamma}_i = \arg\min_{\gamma_i} \frac{1}{2\sigma^2}\left[\left\{\gamma_i - \sum_{j=1}^{m} p_{ij}(y_i - \mu_j)\right\}^2\right] + \frac{1}{w} P_\lambda\left(|\gamma_i|\right).$$

2.5.2 Proof of Equation (2.17)

The parameter $\boldsymbol{\gamma}$ is updated by maximizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i\sigma_j; \mu_j, \sigma_j^2) - \sum_{i=1}^{n} P_\lambda(|\gamma_i|).$$

Again, the problem is separable in each γ_i, and the estimate of each γ_i is obtained by minimizing

$$-\sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i\sigma_j; \mu_j, \sigma_j^2) + P_\lambda(|\gamma_i|).$$

After some algebra, the solution of γ_i has the following form,

$$\hat{\gamma}_i = \arg\min_{\gamma_i} \sum_{j=1}^{m} p_{ij} \left\{ \frac{1}{2}\log\left(\sigma_j^2\right) + \frac{\left(y_i - \gamma_i\sigma_j - \mu_j\right)^2}{2\sigma_j^2} \right\} + P_\lambda\left(|\gamma_i|\right).$$

We have that

$$\begin{aligned}
\sum_{j=1}^{m} p_{ij}\frac{(y_i - \gamma_i\sigma_j - \mu_j)^2}{2\sigma_j^2} &= \sum_{j=1}^{m} \frac{p_{ij}}{2\sigma_j^2}\left\{\gamma_i^2\sigma_j^2 - 2\left(y_i - \mu_j\right)\gamma_i\sigma_j + \left(y_i - \mu_j\right)^2\right\} \\
&= \frac{1}{2}\left[\left\{\gamma_i - \frac{\sum_{j=1}^{m}\frac{p_{ij}}{\sigma_j}(y_i - \mu_j)}{\sum_{j=1}^{m} p_{ij}}\right\}^2 + \text{const}\right] \\
&= \frac{1}{2}\left[\left\{\gamma_i - \sum_{j=1}^{m}\frac{p_{ij}}{\sigma_j}(y_i - \mu_j)\right\}^2 + \text{const}\right],
\end{aligned}$$

where $\sum_{j=1}^{m} p_{ij} = 1$. It follows that

$$\hat{\gamma}_i = \arg\min_{\gamma_i} \frac{1}{2}\left[\left\{\gamma_i - \sum_{j=1}^{m}\frac{p_{ij}}{\sigma_j}(y_i - \mu_j)\right\}^2\right] + P_\lambda\left(|\gamma_i|\right).$$

2.5.3 Proof of SCAD thresholding rule in Proposition 1

The penalized least squares has the following form:

$$\frac{1}{2}(\gamma-\xi)^2+\frac{\sigma^2}{\hat{\sigma}^2}P_\lambda(\gamma) \tag{2.19}$$

where

$$\xi=\frac{\sum_{j=1}^m p_{ij}(y_i-\mu_j)}{\sum_{j=1}^m p_{ij}},$$

Note that for simplicity, we have omitted the subscripts in γ_i and ξ_i.

Consider the first derivative of (2.19) with respect to γ,

$$\frac{\partial\left\{\frac{1}{2}(\gamma-\xi)^2+\frac{\sigma^2}{\hat{\sigma}^2}P_\lambda(\gamma)\right\}}{\partial\gamma}=\gamma-\xi+sgn(\gamma)\frac{\sigma^2}{\hat{\sigma}^2}P'_\lambda(\gamma)$$

where

$$P'_\lambda(\gamma)=\begin{cases}\lambda & \text{if } 0<|\gamma|\leq\lambda,\\ \frac{(a\lambda-|\gamma|)_+}{a-1} & \text{if } \lambda<|\gamma|\leq a\lambda,\\ 0 & \text{if } |\gamma|>a\lambda.\end{cases}$$

We shall check the second derivative of (2.19) in three cases.

Case 1: when $0<|\gamma|\leq\lambda$,

$$\frac{\partial\left\{(\gamma-\xi)+sgn(\gamma)\frac{\sigma^2}{\hat{\sigma}^2}P'_\lambda(\gamma)\right\}}{\partial\gamma}=\frac{\partial\left(\gamma-\xi+sgn(\gamma)\frac{\sigma^2\lambda}{\hat{\sigma}^2}\right)}{\partial\gamma}=1>0.$$

Solving the equation $\gamma-\xi+sgn(\gamma)\frac{\sigma^2\lambda}{\hat{\sigma}^2}=0$, we have $\hat{\gamma}=\xi-\frac{\sigma^2\lambda}{\hat{\sigma}^2}$ and $\hat{\gamma}=-(-\xi-\frac{\sigma^2\lambda}{\hat{\sigma}^2})=\xi+\frac{\sigma^2\lambda}{\hat{\sigma}^2}$.

Case 2: when $\lambda < |\gamma| \leq a\lambda$,

$$\frac{\partial\left\{(\gamma-\xi)+sgn(\gamma)\frac{\sigma^2}{\hat{\sigma}^2}P'_\lambda(\gamma)\right\}}{\partial\gamma} = \frac{\partial\left\{\gamma-\xi+sgn(\gamma)\frac{\sigma^2(a\lambda-|\gamma|)}{\hat{\sigma}^2(a-1)}\right\}}{\partial\gamma} = 1-\frac{\sigma^2}{\hat{\sigma}^2(a-1)}.$$

If $\frac{\sigma^2}{\hat{\sigma}^2} < a-1$, then the second derivative is positive. Solving the equation $\gamma - \xi + sgn(\gamma)\frac{\sigma^2(a\lambda-\gamma)}{\hat{\sigma}^2(a-1)} = 0$, we have $\hat{\gamma} = \frac{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)\xi-a\lambda}{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)-1}$ and $\hat{\gamma} = -\left\{\frac{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)(-\xi)-a\lambda}{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)-1}\right\} = \frac{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)\xi+a\lambda}{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)-1}$.

If $\frac{\sigma^2}{\hat{\sigma}^2} > a-1$, then the second derivative is negative and the solution of the equation $\gamma - \xi + sgn(\gamma)\frac{\sigma^2(a\lambda-\gamma)}{\hat{\sigma}^2(a-1)} = 0$ is not a minimizer of the equation (2.19).

Case 3: when $|\gamma| > a\lambda$,

$$\frac{\partial\left\{(\gamma-\xi)+sgn(\gamma)\frac{\sigma^2}{\hat{\sigma}^2}P'_\lambda(\gamma)\right\}}{\partial\gamma} = \frac{\partial(\gamma-\xi)}{\partial\gamma} = 1 > 0.$$

Solving the equation $\gamma - \xi = 0$, we have $\hat{\gamma} = \xi$.

From the above three cases, we can see that the γ solutions depend on the values of $\frac{\sigma^2}{\hat{\sigma}^2}$ and ξ. Next, we must verify γ solutions in the following scenarios:

When $\sigma^2/\hat{\sigma}^2 < a-1$

Note: For a positive λ, $\frac{\sigma^2}{\hat{\sigma}^2} < a-1$ is equivalent to $\lambda + \frac{\sigma^2\lambda}{\hat{\sigma}^2} < a\lambda$. Since equation (2.19) is symmetric and $\Theta(-\xi;\lambda) = -\Theta(\xi;\lambda)$, we have $\hat{\gamma} = \Theta(-\xi;\lambda) = -\Theta(\xi;\lambda)$. Here we only discuss positive ξ.

1. When $\xi > a\lambda$, γ satisfies Case 3, then we have $\hat{\gamma} = \xi$.
2. When $\lambda + \frac{\sigma^2\lambda}{\hat{\sigma}^2} < \xi \leq a\lambda$, γ satisfies Case 2, then we have $\hat{\gamma} = \frac{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)\xi-a\lambda}{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)-1}$.
3. When $\frac{\sigma^2\lambda}{\hat{\sigma}^2} < \xi \leq \lambda + \frac{\sigma^2\lambda}{\hat{\sigma}^2}$, γ satisfies Case 1, then we have $\hat{\gamma} = \xi - \frac{\sigma^2\lambda}{\hat{\sigma}^2}$.

4. For $0 \leq \xi \leq \frac{\sigma^2\lambda}{\hat{\sigma}^2}$, γ satisfies Case 1. If $\gamma \geq 0$, the first derivative of equation (2.19), $\gamma - \xi + \frac{\sigma^2\lambda}{\hat{\sigma}^2}$, is monotone increasing, so $\hat{\gamma} = 0$; similarly, if $\gamma \leq 0$, the first derivative of equation (2.19) is monotone decreasing, so $\hat{\gamma} = 0$.

In summary, we have

$$\hat{\gamma} = \begin{cases} sgn(\xi)\left(|\xi| - \frac{\sigma^2\lambda}{\hat{\sigma}^2}\right)_+, & \text{if } |\xi| \leq \lambda + \frac{\sigma^2\lambda}{\hat{\sigma}^2} \\ \frac{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)\xi - sgn(\xi)a\lambda}{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)-1}, & \text{if } \lambda + \frac{\sigma^2\lambda}{\hat{\sigma}^2} < |\xi| \leq a\lambda \\ \xi, & \text{if } |\xi| > a\lambda \end{cases}$$

When $a - 1 \leq \sigma^2/\hat{\sigma}^2 \leq a + 1$

Note: For a positive λ, $\frac{\sigma^2}{\hat{\sigma}^2} \geq a - 1$ is equivalent to $\lambda + \frac{\sigma^2\lambda}{\hat{\sigma}^2} \geq a\lambda$. We consider the following subcases:

1. When $|\xi| \leq a\lambda$, based on the result summary when $\sigma^2/\hat{\sigma}^2 < a - 1$,

$$\hat{\gamma} = sgn(\xi)\left(|\xi| - \frac{\sigma^2\lambda}{\hat{\sigma}^2}\right)_+,$$

2. When $a\lambda \leq |\xi| \leq \lambda + \frac{\sigma^2\lambda}{\hat{\sigma}^2}$, for $\hat{\gamma}_1 = sgn(\xi)\left(|\xi| - \frac{\sigma^2\lambda}{\hat{\sigma}^2}\right)_+$, the objective function becomes

$$f_1 = \frac{1}{2}(\hat{\gamma} - \xi)^2 + \frac{\sigma^2}{\hat{\sigma}^2}\lambda|\hat{\gamma}|,$$

and for $\hat{\gamma}_2 = \xi$, the objective function becomes

$$f_2 = \frac{\sigma^2(a+1)\lambda^2}{2\hat{\sigma}^2}.$$

Define $d = f_1 - f_2$. If $d > 0$, then $\hat{\gamma} = \xi$. If $d < 0$, then $\hat{\gamma} = sgn(\xi)\left(|\xi| - \frac{\sigma^2\lambda}{\hat{\sigma}^2}\right)_+$.

(i) When $\xi > \frac{\sigma^2\lambda}{\hat{\sigma}^2}$, $\hat{\gamma}_1 = \xi - \frac{\sigma^2\lambda}{\hat{\sigma}^2}$ and

$$f_1 = \frac{1}{2}\frac{\sigma^{2^2}\lambda^2}{\hat{\sigma}^{2^2}} + \frac{\sigma^2}{\hat{\sigma}^2}\lambda(\xi - \frac{\sigma^2\lambda}{\hat{\sigma}^2}).$$

Then

$$d = f_1 - f_2 = \frac{\sigma^2\lambda^2}{2\hat{\sigma}^2}\left(\frac{2\xi}{\lambda} - a - 1 - \frac{\sigma^2}{\hat{\sigma}^2}\right).$$

When $\xi > \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda$, $d > 0$, so $\hat{\gamma} = \xi$. When $\xi < \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda$, $d < 0$, so $\hat{\gamma} = \xi - \frac{\sigma^2\lambda}{\hat{\sigma}^2}$.
Note that since $\frac{\sigma^2}{\hat{\sigma}^2} > a - 1$, $\frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda > a\lambda$.

Note that in order to result in the soft thresholding rule $\hat{\gamma} = sgn(\xi)\left(|\xi| - \frac{\sigma^2\lambda}{\hat{\sigma}^2}\right)_+$, we need $\frac{\sigma^2\lambda}{\hat{\sigma}^2} \leq \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda$, i.e., $\left[-\frac{\sigma^2\lambda}{\hat{\sigma}^2}, \frac{\sigma^2\lambda}{\hat{\sigma}^2}\right]$ is contained within $\left[-\frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda, \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda\right]$.
Accordingly, $\frac{\sigma^2\lambda}{\hat{\sigma}^2} \leq \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda$ indicates $\frac{\sigma^2}{\hat{\sigma}^2} \leq a + 1$.

(ii) When $0 \leq \xi \leq \frac{\sigma^2\lambda}{\hat{\sigma}^2}$, $\hat{\gamma}_1 = 0$, and $f_1 = \frac{\xi^2}{2}$.

$$\begin{aligned} d = f_1 - f_2 &= \frac{\xi^2}{2} - \frac{\sigma^2(a+1)\lambda^2}{2\hat{\sigma}^2} \\ &< \frac{\sigma^2\lambda^2}{2\hat{\sigma}^2}\left\{\frac{\sigma^2}{\hat{\sigma}^2} - (a+1)\right\}. \end{aligned}$$

Since $\frac{\sigma^2}{\hat{\sigma}^2} \leq a + 1$, $d < 0$, $\hat{\gamma} = 0$.

3. When $|\xi| > \lambda + \frac{\sigma^2\lambda}{\hat{\sigma}^2}$, based on the result summary when $\sigma^2/\hat{\sigma}^2 < a - 1$, $\hat{\gamma} = \xi$.

By summarizing the above three subcases and symmetry property, we have

$$\hat{\gamma} = \begin{cases} sgn(\xi)\left(|\xi| - \frac{\sigma^2\lambda}{\hat{\sigma}^2}\right)_+, & \text{if } |\xi| \leq \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda, \\ \xi, & \text{if } |\xi| > \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda. \end{cases}$$

When $\sigma^2/\hat{\sigma}^2 > a + 1$

For $\frac{\sigma^2}{\hat{\sigma}^2} > a+1$, we have $\frac{\sigma^2\lambda}{\hat{\sigma}^2} > \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda$. Consider the following two subcases:

1. When $\xi > \frac{\sigma^2\lambda}{\hat{\sigma}^2} > \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda$, $\hat{\gamma} = \xi$.

2. When $0 \le \xi \le \frac{\sigma^2\lambda}{\hat{\sigma}^2}$, $\hat{\gamma}_1 = 0$ and

$$d = f_1 - f_2 = \frac{\xi^2}{2} - \frac{\sigma^2(a+1)\lambda^2}{2\hat{\sigma}^2}.$$

If $|\xi| < \sqrt{\frac{\sigma^2(a+1)}{\hat{\sigma}^2}}\lambda$, $d < 0$, then $\hat{\gamma} = 0$; If $|\xi| > \sqrt{\frac{\sigma^2(a+1)}{\hat{\sigma}^2}}\lambda$, $d > 0$, then $\hat{\gamma} = \xi$.

By summarizing the above two subcases and symmetry property, we have

$$\hat{\gamma} = \begin{cases} 0, & \text{if } |\xi| \le \sqrt{\frac{\sigma^2(a+1)}{\hat{\sigma}^2}}\lambda \\ \xi, & \text{if } |\xi| > \sqrt{\frac{\sigma^2(a+1)}{\hat{\sigma}^2}}\lambda \end{cases}$$

Table 2.1: *Outlier Identification Results for Equal Variance Case with Large* $|\gamma|$

	Hard	Hard_{oracle}	SCAD	SCAD_{oracle}	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$
5% outliers						
JD	1.000	1.000	1.000	1.000	1.000	1.000
M	0.000	0.000	0.000	0.000	0.000	0.000
S	0.000	0.000	0.017	0.023	0.012	0.022
10% outliers						
JD	1.000	1.000	1.000	1.000	0.010	1.000
M	0.000	0.000	0.000	0.000	0.732	0.000
S	0.001	0.001	0.042	0.035	0.001	0.013

Table 2.2: *MeSE (MSE) of Point Estimates for Equal Variance Case with Large $|\gamma|$*

	Hard	Hard_{oracle}	SCAD	SCAD_{oracle}	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$	MLE
5% outliers							
π	0.001 (0.001)	0.001 (0.001)	0.001 (0.002)	0.001 (0.003)	0.001 (0.001)	0.001 (0.017)	0.165 (0.220)
μ	0.007 (0.010)	0.007 (0.010)	0.009 (0.017)	0.009 (0.014)	0.007 (0.010)	0.017 (3.159)	38.05 (64.42)
σ	0.001 (0.002)	0.001 (0.002)	0.002 (0.007)	0.001 (0.004)	0.001 (0.001)	0.024 (0.940)	15.89 (628.7)
10% outliers							
π	0.001 (0.001)	0.001 (0.001)	0.001 (0.003)	0.001 (0.003)	0.840 (0.820)	0.001 (0.001)	0.151 (0.236)
μ	0.008 (0.013)	0.008 (0.013)	0.029 (0.039)	0.040 (0.045)	157.0 (153.6)	0.008 (0.013)	40.61 (68.39)
σ	0.003 (0.004)	0.003 (0.004)	0.012 (0.014)	0.001 (0.003)	7.743 (7.729)	0.001 (0.002)	24.73 (8808)

Table 2.3: *Outlier Identification Results for Equal Variance Case with Small $|\gamma|$*

	Hard	Hard_{oracle}	SCAD	SCAD_{oracle}	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$
5% outliers						
JD	0.990	0.990	0.960	1.000	0.99	1.000
M	0.001	0.001	0.030	0.000	0.001	0.000
S	0.004	0.004	0.004	0.081	0.013	0.022
10% outliers						
JD	0.983	0.985	0.925	0.985	0.165	0.96
M	0.004	0.007	0.050	0.001	0.063	0.004
S	0.033	0.031	0.110	0.112	0.001	0.012

Table 2.4: *MeSE (MSE) of Point Estimates for Equal Variance Case with Small $|\gamma|$*

	Hard	Hard_{oracle}	SCAD	SCAD_{oracle}	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$	MLE
5% outliers							
π	0.003 (0.004)	0.003 (0.004)	0.001 (0.001)	0.003 (0.004)	0.001 (0.001)	0.002 (0.017)	0.003 (0.095)
μ	0.023 (0.030)	0.035 (0.031)	0.156 (0.157)	0.041 (0.050)	0.009 (0.014)	0.022 (3.887)	0.251 (13.04)
σ	0.018 (0.027)	0.002 (0.004)	0.404 (0.371)	0.016 (0.020)	0.001 (0.001)	0.022 (0.977)	0.539 (633.4)
10% outliers							
π	0.001 (0.002)	0.001 (0.002)	0.001 (0.002)	0.001 (0.002)	0.001 (0.004)	0.001 (0.001)	0.003 (0.112)
μ	0.017 (0.022)	0.028 (0.033)	0.026 (0.065)	0.026 (0.037)	0.143 (0.723)	0.010 (0.032)	0.806 (13.82)
σ	0.019 (0.020)	0.004 (0.007)	0.034 (0.102)	0.031 (0.038)	0.261 (0.470)	0.001 (0.017)	1.315 (105.4)

Table 2.5: *Outlier Identification Results for Unequal Variance Case with Large* $|\gamma|$

	Hard	Hard_{oracle}	SCAD	SCAD_{oracle}	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$
5% outliers						
JD	1.000	1.000	1.000	1.000	0.967	1.000
M	0.000	0.000	0.000	0.000	0.002	0
S	0.001	0.001	0.003	0.009	0.009	0.031
10% outliers						
JD	1.000	1.000	0.995	0.995	0.000	0.985
M	0.000	0.000	0.005	0.001	0.476	0.000
S	0.001	0.001	0.009	0.013	0.000	0.012

Table 2.6: *MeSE (MSE) of Point Estimates for Unequal Variance Case with Large* $|\gamma|$

	Hard	Hard_{oracle}	SCAD	SCAD_{oracle}	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$	MLE
5% outliers							
π	0.001 (0.001)	0.001 (0.001)	0.001 (0.001)	0.001 (0.001)	0.001 (0.039)	0.001 (0.001)	0.780 (0.767)
μ	0.014 (0.021)	0.014 (0.021)	0.015 (0.022)	0.014 (0.021)	0.022 (17.63)	0.022 (0.031)	92.76 (91.97)
σ	0.023 (0.028)	0.020 (0.024)	0.022 (0.044)	0.016 (0.021)	0.010 (0.551)	0.100 (0.108)	247.5 (243.6)
10% outliers							
π	0.001 (0.001)	0.001 (0.001)	0.001 (0.002)	0.001 (0.001)	0.056 (0.058)	0.001 (0.001)	0.055 (0.058)
μ	0.018 (0.026)	0.018 (0.026)	0.019 (0.030)	0.019 (0.029)	18.13 (18.10)	0.010 (0.013)	11.83 (11.90)
σ	0.036 (0.045)	0.034 (0.042)	0.038 (0.334)	0.036 (0.223)	19.83 (19.92)	1.035 (1.031)	61.33 (61.29)

Table 2.7: *Outlier Identification Results for Unequal Variance Case with Small* $|\gamma|$

	Hard	Hard_{oracle}	SCAD	SCAD_{oracle}	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$
5% outliers						
JD	0.900	0.900	0.805	1.000	0.875	0.990
M	0.015	0.004	0.182	0	0.011	0.0005
S	0.001	0.001	0.008	0.073	0.008	0.031
10% outliers						
JD	0.800	0.800	0.010	1.000	0.000	0.784
M	0.029	0.027	0.920	0.000	0.268	0.016
S	0.001	0.001	0.000	0.103	0.001	0.008

Table 2.8: *MeSE (MSE) of Point Estimates for Unequal Variance Case with Small $|\gamma|$*

	Hard	Hard_{oracle}	SCAD	SCAD_{oracle}	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$	MLE
5% outliers							
π	0.001 (0.001)	0.001 (0.001)	0.001 (0.003)	0.001 (0.001)	0.001 (0.148)	0.001 (0.012)	0.192 (0.174)
μ	0.017 (0.026)	0.017 (0.028)	0.045 (0.071)	0.022 (0.032)	0.025 (17.37)	0.024 (0.600)	21.97 (19.24)
σ	0.009 (0.016)	0.004 (0.008)	0.183 (1.212)	0.004 (0.010)	0.013 (2.031)	0.100 (0.232)	23.76 (20.64)
10% outliers							
π	0.001 (0.001)	0.001 (0.001)	0.027 (0.028)	0.001 (0.001)	0.161 (0.130)	0.001 (0.180)	0.248 (0.257)
μ	0.021 (0.029)	0.022 (0.029)	0.086 (0.105)	0.025 (0.034)	14.45 (10.40)	0.008 (0.012)	30.41 (37.07)
σ	0.016 (0.241)	0.008 (0.015)	12.04 (11.96)	0.010 (0.023)	18.54 (13.74)	1.017 (1.020)	34.52 (30.22)

Table 2.9: *Parameter Estimation on Acidity Data Set*

		π_1	π_2	π_3	μ_1	μ_2	μ_3	σ
MLE	No outliers	0.589	0.138	0.273	4.320	5.682	6.504	0.365
	1 outlier	0.327	0.324	0.349	4.455	4.455	6.448	0.687
	3 outliers	0.503	0.478	0.019	5.105	5.105	12.00	1.028
Hard	No outliers	0.588	0.157	0.255	4.333	5.720	6.545	0.336
	1 outlier	0.591	0.157	0.252	4.333	5.723	6.548	0.334
	3 outliers	0.597	0.157	0.246	4.333	5.729	6.553	0.331

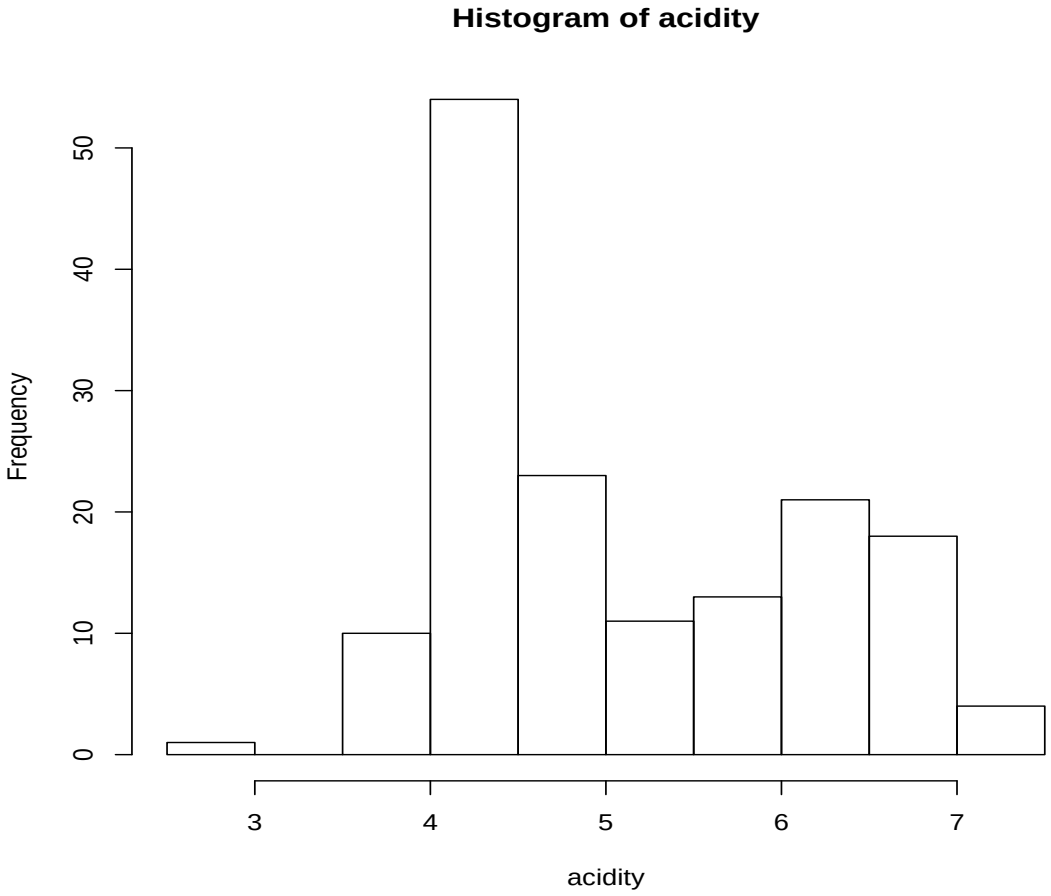

Figure 2.1: *Histogram for Acidity data*

Chapter 3

Outlier Detection and Robust Mixture Regression Using Nonconvex Penalized Likelihood

3.1 Introduction

Given n observations of the response $Y \in \mathbb{R}$ and predictor $X \in \mathbb{R}^p$, multiple linear regression models are commonly used to explore the conditional mean structure of Y given X. However, in many applications, the underlying assumption that the regression relationship is homogeneous across all the observations $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ can be easily violated. Instead, the observations may form several distinct clusters indicating mixed relationships between the response and the predictors. Such heterogeneity can be more appropriately modeled by a *finite mixture regression model*, consisting of, say, m homogeneous groups/components. It is assumed that a linear regression model holds for each of the m components, i.e., when $(y, \boldsymbol{x})$ belongs to the jth component ($j = 1, 2, \ldots, m$), $y = \boldsymbol{x}^T \boldsymbol{\beta}_j + \epsilon_j$, where $\boldsymbol{\beta}_j \in \mathbb{R}^p$ is a fixed and unknown coefficient vector for the jth component, and $\epsilon_j \sim N(0, \sigma_j^2)$. (The intercept term can be included by setting the first element of $\boldsymbol{x}$ as one). Let Z be a latent variable

indicating the class/component membership, such that $P(Z = j) = \pi_j$ for $j = 1, 2, \cdots, m$, where π_js are called mixing proportions. Then the conditional density of $\boldsymbol{y}$ given $\boldsymbol{x}$, without observing Z, is

$$f(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j \phi(y; \boldsymbol{x}^T \boldsymbol{\beta}_j, \sigma_j^2), \tag{3.1}$$

where $\phi(\cdot; \mu, \sigma^2)$ denotes the density function of $N(\mu, \sigma^2)$ and $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \sigma_1; \ldots; \pi_m, \boldsymbol{\beta}_m, \sigma_m)^T$ collects all the unknown parameters of the model.

Since first introduced by Goldfeld and Quandt (1973), the above mixture regression model has been widely used in business, marketing, and social sciences (see Jiang and Tanner, 1999; Böhning, 1999; Wedel and Kamakura, 2000; McLachlan and Peel, 2000; Skrondal and Rabe-Hesketh, 2004; and Frühwirth-Schnatter, 2006). Hennig (2000) proved the identifiability of model (3.1) under some general conditions for the covariates, i.e., model (3.1) is identifiable if m is smaller than the number of distinct $(p-1)$ dimensional hyperplanes needed to cover the covariates of each cluster. Maximum likelihood estimator (MLE) is commonly used to infer the unknown parameter $\boldsymbol{\theta}$ in (3.1), i.e.,

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}_j, \sigma_j^2) \right\}. \tag{3.2}$$

The MLE does not have an explicit form and the problem is usually solved by invoking the EM algorithm (Dempster et al. 1977).

Although the finite mixture models with the maximum likelihood inference have greatly enriched the toolkit of regression analysis, the model is very sensitive to outliers, and failure to accommodate outliers may greatly jeopardize mixture model estimation and inference. Many robust methods have been developed for mixture regression models. Markatou (2000) and Shen et al. (2004) proposed to properly weight each data point to robustify the estimation procedure. Neykov et al. (2007) proposed robust fitting of mixtures using the

trimmed likelihood. Bai et al. (2012) proposed a modified EM algorithm for mixture regression by replacing the least squares criterion in M step with a robust criterion. Bashir and Carter (2012) extended the idea of S-estimator to mixture linear regression. Yao et al. (2014) proposed a robust mixture regression approach using t-distribution. Song et al. (2013) proposed a robust mixture regression model fitting by laplace distribution. There also have been several related robust methods for linear clustering; see, e.g., Hennig (2002, 2003), Mueller and Garlipp (2005), García-Escudero et al. (2009), and García-Escudero et al. (2010).

In this article, we propose a Robust Mixture Regression via Mean shift penalization approach (RMRM or RM^2), to conduct simultaneous outlier detection/accomodation and robust parameter estimation in finite normal mixture regression models. Our method is motivated by She and Owen (2011) and Lee, MacEachern and Jung (2012), in which penalized estimation methods were adopted to induce the sparsity of a case-specific parameter vector for accommodating outliers in linear regression models. Under the general framework of mixture regression, there are several new challenges for adopting the nonconvex penalization methods. For example, the problem of maximizing the likelihood itself becomes a nonconvex problem, which complicates the computation. When the components have unequal variances, the simple mean shift model will not work well since the definition of an outlier may become ambiguous as the scale of the outlying effect of a particular point may vary across different components. We propose to add a component specific mean-shift term for each component and for each observation and these terms are designed to be proportional to the component variances, accounting for the potential heteroscedasticity among different components. We propose an efficient iterative thresholding embeded EM algorithm to solve the nonconvex RM^2 problem, and our proposed estimator is demonstrated to be highly robust against gross outliers and leverage points.

The rest of the article is organized as follows. In Section 3.2, we propose the RM^2 approach. In Section 3.3, we compare the proposed methods to several existing methods

via simulation studies. A real application showcasing the efficacy of the proposed method is presented in Section 3.4, and we conclude the paper in Section 3.5.

3.2 Robust Mixture Regression via Mean-shift Penalization

To illustrate the main idea, we start from the simple case that the mixture components have equal variances, i.e., $\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$. Motivated by the mean-shift linear regression model considered by She and Owen (2011) and Lee, MacEachern, and Jung (2012), it is natural to consider the following mixture model with a mean-shift parameterization, i.e.,

$$f(y_i; \boldsymbol{\theta}, \gamma_i) = \sum_{j=1}^{m} \pi_j \phi(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}_j + \gamma_i, \sigma^2), \qquad i = 1, \ldots, n, \tag{3.3}$$

where $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \ldots, \pi_m, \boldsymbol{\beta}_m, \sigma)^T$. Here, for each observation, a shift parameter, γ_i, is added to its mixture mean structure; we thus refer to the above as a mean shifted mixture model. Without any further constraints on the model parameters, it is obvious that the mean-shift model is over-parameterized and hence the parameters are not fully identifiable. The essence of this formulation lies in the sparsity assumption on γ_i, i.e., we shall assume many γ_is are in fact zero, corresponding to the normal observations, and only a few γ_is are nonzero, corresponding to the outlying observations. Therefore, promoting sparsity of γ_i in model estimation provides a direct way for identifying and accommodating outliers in the mixture model.

Now consider the general case that the mixture components are allowed to have unequal variances, i.e., $\epsilon_j \sim N(0, \sigma_j^2)$. This heteroscedasticity of component variances imposes additional challenges for identifying outliers, as the definition of an "outlier" even becomes ambiguous due to the fact that the mixture components are of different scales. In the general setting, whether an observation is an outlier to a certain component should be judged

based on the scale of that component, and an observation should be declared as an outlier only if it is far away from all the centroids of the mixture components. This motivates us to further extend model (3.3) to take into account the scaling issue. The main idea is to make the case-specific mean shift parameter γ_i be scale invariant, so that the magnitude of γ_i itself represents the standardized distance from the observation to all the cluster centroids. We thus propose the following robust mixture regression model with mean-shift (RM^2),

$$f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \gamma_i) = \sum_{j=1}^{m} \pi_j \phi(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}_j + \gamma_i \sigma_j, \sigma_j^2), \qquad i = 1, \ldots, n, \tag{3.4}$$

where we redefine $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \sigma_1, \ldots, \pi_m, \boldsymbol{\beta}_m, \sigma_m)^T$. The outlying effect is made both case-specific and component-specific, i.e., the outlying effect of the ith observation to the jth component is modeled by $\gamma_i \sigma_j$, depending directly on the scale of the jth component. In this way, γ_i becomes scale free, and can be simply understood as the number of standard deviations shifted from the correct component mean structures.

The efficient and accurate recovery of the sparse vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^T$ holds the key to realize the bearing of the powerful framework of the proposed mean shifted mixture model. In recent years, the penalized estimation approach has undergone exciting developments for sparse learning and variable selection. This motivates us to consider a penalized likelihood approach. Given a random sample $\{(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n\}$ from model (3.4), the log-likelihood function is given by

$$\ell_n(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i - \gamma_i \sigma_j - \boldsymbol{x}_i^T \boldsymbol{\beta}_j; 0, \sigma_j^2) \right\}.$$

We propose a penalized likelihood approach to conduct model estimation and outlier detection,

$$pl_n(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \ell_n(\boldsymbol{\theta}, \boldsymbol{\gamma}) - \sum_{i=1}^{n} P_\lambda(|\gamma_i|), \tag{3.5}$$

where $P_\lambda(\cdot)$ is some penalty function chosen to induce the sparsity in γ, with λ being a

tuning parameter controlling the degrees of penalization (She and Owen, 2011). There are many choices for the penalty function in the above criterion. To list a few, the ℓ_1 norm penalty (Donoho and Johnstone, 1994a; Tibshirani, 1996, 1997) is given by $P_\lambda(\gamma) = \lambda|\gamma|$, the ℓ_0 hard penalty (Antoniadis, 1997) can be written as

$$P_\lambda(\gamma) = \frac{\lambda^2}{2} I(\gamma \neq 0), \tag{3.6}$$

and the SCAD penalty proposed by Fan and Li (2001) is

$$P_\lambda(\gamma) = \begin{cases} \lambda|\gamma|, & \text{if } |\gamma| \leq \lambda, \\ -\left(\frac{\gamma^2 - 2a\lambda|\gamma| + \lambda^2}{2(a-1)}\right), & \text{if } \lambda < |\gamma| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\gamma| > a\lambda, \end{cases} \tag{3.7}$$

where a is a constant usually set to be 3.7. Each of these penalty forms corresponds to certain thresholding rule, thus capable of performing shrinkage and producing exact zero solution, e.g., ℓ_1 penalty corresponds to a soft-threshing rule and ℓ_0 penalty a hard-thresholding rule. The advantages of using nonconvex penalties are well understood. Thus we shall mainly focus on the nonconvex hard penalty and SCAD penalty.

In classical mixture regression problem, the EM algorithm is commonly used to maximize the likelihood, as the component labels are unobservable and can be treated as missing data. Here, we propose an iterative thresholding embeded EM algorithm to maximize the proposed penalized log-likelihood criterion. Let

$$z_{ij} = \begin{cases} 1, & \text{if } i\text{th observation is from } j\text{th component}, \\ 0, & \text{otherwise.} \end{cases}$$

and denote the complete data by $\{(\boldsymbol{x}_i, z_i, y_i), i = 1, 2, \ldots, n\}$, where the component labels $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{im})$ are not observable in practice. The penalized complete log-likelihood

function is

$$pl_n^c(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \ell_n^c(\boldsymbol{\theta}, \boldsymbol{\gamma}) - \sum_{i=1}^{n} P_\lambda(|\gamma_i|) \tag{3.8}$$

where the complete log-likelihood is given by

$$\ell_n^c(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij} \log\left\{\pi_j \phi(y_i - \gamma_i\sigma_j - \boldsymbol{x}_i^T\boldsymbol{\beta}_j; 0, \sigma_j^2)\right\}. \tag{3.9}$$

In the E-step, the conditional expectation of the penalized complete log-likelihood (3.8) is computed, and we then maximize it with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ in the M-step. Specifically, in the M-step, we alternatingly update $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ with the other part held fixed, until convergence is reached. For fixed $\boldsymbol{\gamma}$, both p_{ij} and $\boldsymbol{\beta}$ can be solved explicitly. As each σ_j appears in the mean structure, it no longer has an explicit solution; however, the estimation of each σ_j is separable so that the problem is easily solvable by standard optimization method such as Newton-Raphson. For fixed p_{ij}, $\boldsymbol{\beta}$, and σ_j, $\boldsymbol{\gamma}$ is updated by maximizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(k+1)} \log\phi(y_i - \boldsymbol{\gamma}_i\sigma_j - \boldsymbol{x}_i^T\boldsymbol{\beta}_j; 0, \sigma_j^2) - \sum_{i=1}^{n} P_\lambda(|\gamma_i|).$$

It can be shown that the above problem is separable in each γ_i, for which it suffices to minimize

$$\frac{1}{2}\left[\left\{\gamma_i - \sum_{j=1}^{m}\frac{p_{ij}^{(k+1)}}{\sigma_j}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j)\right\}^2\right] + P_\lambda\left(|\gamma_i|\right). \tag{3.10}$$

The thresholding rules for soft, hard, and SCAD are given, respectively, as follows,

$$\gamma_i = \Theta_{soft}(\xi_i; \lambda) = \begin{cases} 0, & \text{if } |\xi_i| \le \lambda \\ \xi_i - sgn(\xi_i)\lambda, & \text{if } |\xi_i| > \lambda, \end{cases}$$

$$\gamma_i = \Theta_{hard}(\xi_i;\lambda) = \begin{cases} 0, & \text{if } |\xi_i| \le \lambda \\ \xi_i, & \text{if } |\xi_i| > \lambda, \end{cases}$$

and

$$\gamma_i = \Theta_{SCAD}(\xi_i;\lambda) = \begin{cases} sgn(\xi_i)(|\xi_i| - \lambda)_+, & \text{if } |\xi_i| \le 2\lambda \\ \frac{(a-1)\xi_i - sgn(\xi_i)a\lambda}{a-2}, & \text{if } 2\lambda < |\xi_i| \le a\lambda \\ \xi_i, & \text{if } |\xi_i| > a\lambda, \end{cases}$$

where

$$\xi_i = \sum_{j=1}^{m} \frac{p_{ij}}{\sigma_j}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j).$$

The detailed proposed thresholding embeded EM algorithm to maximize the penalized log-likelihood (3.5) is summarized in Algorithm 1. Based on the property of EM algorithm, for any fixed tuning parameter λ, each iteration of the E-step and M-step of Algorithm 1 monotonically non-decreases the penalized log-likelihood function, i.e., $pl_n(\hat{\boldsymbol{\theta}}^{(k+1)}, \hat{\boldsymbol{\gamma}}^{(k+1)}) \ge pl_n(\hat{\boldsymbol{\theta}}^{(k)}, \hat{\boldsymbol{\gamma}}^{(k)})$, for all $k \ge 0$.

The proposed scaled-invariate method is also applicable in the special case that $\epsilon_j \sim N(0, \sigma^2)$ in model (3.1), i.e., the mixture components have equal variance. We use the same procedure as RM^2 for unequal variance case by replacing σ_j with σ. Similar to algorithm 1, the same iterating steps are used except for updating σ^2 with the following formula:

$$(2.b) \qquad \sigma^2 \leftarrow \arg\max_{\sigma^2} \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i\sigma - \boldsymbol{x}_i^T\boldsymbol{\beta}_j; 0, \sigma^2).$$

The proposed EM algorithm is for any fixed tuning parameter λ. In practice, we need to choose an optimal λ and hence an optimal set of parameter estimates. We construct a Bayesian information criterion (BIC) for tuning parameter selection,

$$BIC(\lambda) = -\ell(\lambda) + \log(n) df(\lambda) \tag{3.11}$$

Algorithm 3 Thresholding Embeded EM algorithm for RM^2 with Unequal Variances

Initialize $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$. Set $k \leftarrow 0$.

repeat

(1) E-Step: Compute the conditional expectation:

$$Q(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \left[\log \pi_j + \log \phi(y_i - \gamma_i \sigma_j - \boldsymbol{x}_i^T \boldsymbol{\beta}_j; 0, \sigma_j^2) \right] - \sum_{i=1}^{n} P_\lambda(|\gamma_i|)$$

where

$$p_{ij}^{(k+1)} = E(z_{ij} | y_i; \boldsymbol{\theta}^{(k)}) = \frac{\pi_j^{(k)} \phi(y_i - \gamma_i^{(k)} \sigma_j^{(k)} - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(k)}; 0, \sigma_j^{2^{(k)}})}{\sum_{j=1}^{m} \pi_j^{(k)} \phi(y_i - \gamma_i^{(k)} \sigma_j^{(k)} - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(k)}; 0, \sigma_j^{2^{(k)}})}$$

(2) M-Step: Update $\pi_j^{(k+1)} = \frac{\sum_{i=1}^{n} p_{ij}^{(k+1)}}{n}$ and update other parameters by maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\gamma} | \boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)})$, i.e., start from $(\boldsymbol{\beta}^{(k)}, \sigma_j^{2^{(k)}}, \boldsymbol{\gamma}^{(k)})$ and iterate the following steps until convergence to obtain $(\boldsymbol{\beta}^{(k+1)}, \sigma_j^{2^{(k+1)}}, \boldsymbol{\gamma}^{(k+1)})$:

$$(2.a) \qquad \boldsymbol{\beta}_j \leftarrow \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T p_{ij}^{(k+1)} \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i p_{ij}^{(k+1)} (y_i - \gamma_i \sigma_j) \right),$$

$$(2.b) \qquad \sigma_j^2 \leftarrow \arg\max_{\sigma_j^2} \sum_{i=1}^{n} p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i \sigma_j - \boldsymbol{x}_i^T \boldsymbol{\beta}_j; 0, \sigma_j^2),$$

$$(2.c) \qquad \gamma_i \leftarrow \Theta(\xi_i; \lambda).$$

where Θ denotes a thresholding rule depending on the penalty form adopted, and $\xi_i = \sum_{j=1}^{m} \frac{p_{ij}^{(k+1)}}{\sigma_j} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j)$.

$k \leftarrow k + 1$.

until convergence

where $\ell(\lambda)$ is the mixture log-likelihood function evaluated at the parameter estimates with tuning parameter λ, and $df(\lambda)$ is the degrees of freedom of the resulting model. Following Zou (2006), we estimate the degrees of freedom using the sum of the number of nonzero elements of the vector $\hat{\boldsymbol{\gamma}}(\lambda)$ and the number of component parameters in the mixture model. We fit the model for 100 λ values equally spaced at the log scale in an interval $(\lambda_{\min}, \lambda_{\max})$, where $\lambda_{\min}$ is some λ value for which about 50% of the entries in $\boldsymbol{\gamma}$ are nonzero, and λ_{max} corresponds to some λ value for which $\boldsymbol{\gamma}$ is estimated as a zero vector.

We note that from outlier detection point of view or for practical consideration, there may be other methods to determine the λ value or choose the optimal solution along the solution path. For example, based on prior knowledge, one may decide to discard 5% of the observations as outliers; then a solution with approximately 5% of nonzero $\boldsymbol{\gamma}$ values can be chosen as the final solution. In the scale-invariate model, as $\boldsymbol{\gamma}$ can be interpreted as the number of standard deviations from the mean structure, one may also examine the magnitude of the $\boldsymbol{\gamma}$ estimates to determine the number of possible outliers. Although we mainly use BIC in this paper, we shall see that by formulating the outlier detection problem as a penalized regression method, the many well-studied model selection criteria including C_p, AIC, and GCV are all applicable.

3.3 Simulation

3.3.1 Simulation Setups

We consider two mixture model setups, in which the observations are contaminated with additive outliers, to evaluate the final sample performance of the proposed approach and compare it with several existing methods.

Model 1: For each $i = 1, \ldots, n$, y_i is independently generated by

$$y_i = \begin{cases} 1 - x_{1i} + x_{2i} + \gamma_i \sigma + \epsilon_{i1}, & \text{if } z_{i1} = 1; \\ 1 + 3x_{1i} + x_{2i} + \gamma_i \sigma + \epsilon_{i2}, & \text{if } z_{i1} = 0. \end{cases}$$

where z_{i1} is a component indicator generated from Bernoulli distribution with $P(z_{i1} = 1) = 0.3$; x_{1i} and x_{2i} are independently generated from $N(0, 1)$, and the error terms ϵ_{i1} and ϵ_{i2} are also independently generated from $N(0, \sigma^2)$ with $\sigma^2 = 1$.

Model 2: For each $i = 1, \ldots, n$, y_i is independently generated by

$$y_i = \begin{cases} 1 - x_{1i} + x_{2i} + \gamma_i \sigma_1 + \epsilon_{i1}, & \text{if } z_{i1} = 1; \\ 1 + 3x_{1i} + x_{2i} + \gamma_i \sigma_2 + \epsilon_{i2}, & \text{if } z_{i1} = 0. \end{cases}$$

where z_{i1} is a component indicator generated from Bernoulli distribution with $P(z_{i1} = 1) = 0.3$; x_{1i} and x_{2i} are independently generated from $N(0, 1)$, and the error terms ϵ_{i1} and ϵ_{i2} are independently generated from $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$, respectively, with $\sigma_1^2 = 1$ and $\sigma_2^2 = 4$.

We consider two magnitudes of outliers, i.e., the absolute value of any nonzero mean shift parameter, $\alpha_i = |\gamma_i|$, is generated from uniform distribution either between 5 and 7 or between 11 and 13. We consider two proportions of outliers, either 5% or 10%. In each setting, the sample size is set to be $n = 400$ and we repeat the simulation 200 times. Specifically, in Example 1, we first generate $n = 400$ observations according to Model 1 with all γ_is set to be zero; when there are 5% (10%) outliers, 5 (10) observations from the first component are then replaced by $y_i = 1 - x_{1i} + x_{2i} + \gamma_i + \epsilon_{1i}$ with $x_{1i} = 2$, $x_{2i} = 2$, and $\gamma_i = -\alpha_i$, and 15 (30) observations from the second component are replaced by $y_i = 1 + 3x_{1i} + x_{2i} + \gamma_i + \epsilon_{2i}$ with $x_{1i} = 2$, $x_{2i} = 2$, and $\gamma_i = \alpha_i$. In Example 2, the additive outliers are generated in exactly the same fashion as in Example 1, and the only difference is that the component variances are unequal in the latter example.

3.3.2 Methods and Evaluation Measures

We compare our proposed RM^2 approach with soft, hard, and SCAD penalties to three existing robust approaches and the traditional normal mixture model. To alleviate the inaccuracy in tuning parameter selection and examine the true potential of the proposed approaches, we also report the "oracle" penalized regression estimator for each penalty, which is defined as the solution whose number of selected outliers is the smallest number greater than or equal to the number of true outliers on the solution path. These are the estimators we would have obtained if the true proportion of outliers is known a priori. The eleven methods we compared are listed below:

1. the traditional MLE in mixture linear regression with normally distributed error (MLE);
2. trimmed likelihood estimator (TLE) proposed by Neykov et al. (2007) with the percentage of trimmed data α set to 0.05 (TLE$_{0.05}$),
3. TLE with the percentage of trimmed data α set to 0.10 (TLE$_{0.10}$),
4. the robust estimator based on an modified EM algorithm with bisquare loss (MEM-bisquare) proposed by Bai et al.(2012),
5. the MLE in mixture linear regression assuming t-distributed error (Mixregt),
6. the proposed RM^2 using the hard penalty (Hard),
7. the proposed RM^2 using the SCAD penalty (SCAD),
8. the proposed RM^2 using the soft penalty (Soft),
9. the oracle estimate using the hard penalty (Hard$_{oracle}$),
10. the oracle estimate using the SCAD penalty (SCAD$_{oracle}$),
11. the oracle estimate using the soft penalty (Soft$_{oracle}$).

For fitting mixture models, there are well known label switching issues (Celeux, et al., 2000; Stephens, 2000; Yao and Lindsay, 2009; Yao, 2012). In our simulation study, the labels are determined by minimizing the distance to the true parameter values. To evaluate the estimation performance, we report both the median squared errors (MeSE) and the mean squared errors (MSE) of the parameter estimates. To evaluate the outlier detection performance, similar to She and Owen (2011), we report three measures: the average proportion of masking (M), i.e., the fraction of undetected outliers, the average proportion of swapping (S), i.e., the fraction of good points labeled as outliers, and the joint detection rate (JD), i.e., the proportion of simulations with 0 masking. The simulation results are summarized in Tables 3.1 – 3.8.

3.3.3 Results

The simulation results of Example 1 (equal variance case) are reported in Table 3.1 – Table 3.4. Tables 3.1 and 3.3 report the three fractions of outlier detection and Tables 3.2 and 3.4 report the median of squared errors (MeSE) of parameter estimates for each estimation method. In the case of 5% outliers, all methods gain ideal outlier detection rates and small MeSE of parameter estimates with large $|\gamma|$; all methods except for Soft have high joint outlier detection rate and small MeSE with small $|\gamma|$. In the case of 10% outliers, Hard, SCAD, and $\text{TLE}_{0.10}$ work well in terms of both outlier detection rates and MeSE with large $|\gamma|$, whereas Soft, $\text{TLE}_{0.05}$, MEM-bisquare, and Mixregt have low joint outlier detection rates and big MeSE; with small $|\gamma|$, $\text{TLE}_{0.10}$, and Mixregt work better than other methods in outlier identification but Hard, Hard_{oracle}, and SCAD_{oracle} obtain similar MeSE to those of $\text{TLE}_{0.10}$ and Mixregt.

Table 3.5 – Table 3.8 summarize the simulation results of Example 2 (unequal variance case). Tables 3.5 and 3.7 show the three fractions of outlier detection and Tables 3.6 and 3.8 show the median of squared errors (MeSE) of parameter estimates for each estimation method. All methods except for soft have high joint outlier detection rates when the pro-

portion of outliers is 5% with large $|\gamma|$; SCAD and MEM-bisquare have low joint detection rates with small γ but SCAD_{oracle} has similar performance to Hard. When there are 10% outliers in the data, Hard, SCAD, and $\text{TLE}_{0.10}$ have outstanding performance in terms of outlier identification rates and MeSE with large $|\gamma|$; Hard and SCAD do not work well in terms of joint outlier detection rate with small $|\gamma|$ but MeSE of Hard_{oracle} is comparable to $\text{TLE}_{0.10}$. Mixregt has low joint detection rate with large $|\gamma|$ and high JD rate with small $|\gamma|$ for 10% outliers case.

In summary, $\text{TLE}_{0.10}$ has good results in terms of outliers detection in all cases but has larger MSE for 5% outliers case. $\text{TLE}_{0.05}$ fails to work in the case of 10% outliers due to the small α setting (less than the proportion of outliers). Hard has comparable performance to the oracle TLE and Hard_{oracle} in terms of both outlier detection and MeSE in 5% outliers case with either large or small $|\gamma|$ and in 10% outliers case with large $|\gamma|$. With small $|\gamma|$ and 10% outliers in the data, Hard_{oracle} has better performance than Hard. SCAD performs as well as Hard and SCAD_{oracle} with large $|\gamma|$. But SCAD_{oracle} performs much better than SCAD with small $|\gamma|$. Therefore, a better method to choose the tuning parameter for SCAD and HARD might improve their performance in some cases. Like MLE, Soft is sensitive to high leverage outliers, which has also been noticed by She and Owen (2011).

3.4 Tone Perception Data Analysis

We apply the proposed robust procedure to tone perception data (Cohen, 1984). In the tone perception experiment of Cohen (1984), a pure fundamental tone with electronically generated overtones added was played to a trained musician. The experiment recorded 150 trials from the same musician. The overtones were determined by a stretching ratio, which is the ratio between adjusted tone and the fundamental tone. The purpose of this experiment was to see how this tuning ratio affects the perception of the tone and to determine if either of two musical perception theories was reasonable.

We compare our proposed Hard and traditional MLE after adding ten identical outliers (1.5, 5) into the original data set. Figure 3.1 shows the scatter plot of the data with the estimated regression lines generated by the traditional MLE (dashed lines) and the proposed Hard (solid line) for the data augmented by the outliers (stars). Based on Figure 3.1, the MLE mistakenly assigns the outliers to one component and the rest of the data to another component. In contrast, the proposed method using Hard penalty is not influenced by the added outliers and fits the two regression lines to the two correctly identified components. Using SCAD penalty leads to very similar results.

3.5 Discussion

In this article, we have proposed a robust mixture regression estimation procedure using mean shit model. The new model focuses on outlier detection directly and can also provide a robust model parameter estimate. Based on our simulation results, the proposed RM^2 using the hard penalty (Hard) and data adaptive chosen tuning parameter has overall comparable performance to Hard_{oracle} and the oracle TLE.

In addition, note that Hard_{oracle} and SCAD_{oracle} have better performance than HARD and SCAD in some cases, especially when $|\gamma|$ is small. Therefore, we can further improve the performance of SCAD and HARD if having a better method to choose the tuning parameter. This requires further research.

The traditional definition of breakdown point as a criterion of robustness can not be applied to mixture regression directly. García-Escudero et al. (2010) stated that the traditional definition of breakdown point is not the correct 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. Therefore, construction and investigation of other robustness measures for mixture model setup may be an interesting future research direction.

3.6 Appendix

3.6.1 Proof of Equation (3.10)

The estimate of $\boldsymbol{\gamma}$ is updated based on updated p_{ij}, $\boldsymbol{\beta}$, and σ_j by maximizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - \boldsymbol{\gamma}_i\sigma_j - \boldsymbol{x}_i^T\boldsymbol{\beta}_j; 0, \sigma_j^2) - \sum_{i=1}^{n} P_\lambda(|\gamma_i|).$$

To do this, each γ_i is separately updated by maximizing:

$$\sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i\sigma_j - x_i^T\beta_j; 0, \sigma_j^2) - P_\lambda(|\gamma_i|).$$

Equivalently, the estimate of γ_i is updated by minimizing

$$-\sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - \gamma_i\sigma_j - x_i^T\beta_j; 0, \sigma_j^2) + P_\lambda(|\gamma_i|).$$

Note that

$$\begin{aligned}
&\log \phi\left(y_i - \gamma_i\sigma_j - x_i^T\beta_j; 0, \sigma_j^2\right) \\
&= \log\left[\left(\sigma_j^2\right)^{-\frac{1}{2}} \cdot exp\left\{-\frac{\left(y_i - \gamma_i\sigma_j - x_i^T\beta_j\right)^2}{2\sigma_j^2}\right\}\right] + const \\
&= -\frac{1}{2}\log\left(\sigma_j^2\right) - \frac{\left(y_i - \gamma_i\sigma_j - x_i^T\beta_j\right)^2}{2\sigma_j^2} + const.
\end{aligned}$$

Thus, the solutions of γ have the following form:

$$\gamma_i = argmin\frac{1}{2}\sum_{j=1}^{m} p_{ij}\log\left(\sigma_j^2\right) + \sum_{j=1}^{m} p_{ij}\left\{\frac{\left(y_i - \gamma_i\sigma_j - x_i^T\beta_j\right)^2}{2\sigma_j^2}\right\} + P_\lambda\left(|\gamma_i|\right).$$

Since $\frac{1}{2}\sum_{j=1}^{m} p_{ij}\log\left(\sigma_j^2\right)$ does not depend on γ, we can ignore this term.

The second term:

$$
\begin{aligned}
&\sum_{j=1}^{m} p_{ij} \frac{(y_i - \gamma_i\sigma_j - x_i^T\beta_j)^2}{2\sigma_j^2} \\
&= \sum_{j=1}^{m} \frac{p_{ij}}{2\sigma_j^2} \left\{ \gamma_i^2\sigma_j^2 - 2\left(y_i - x_i^T\beta_j\right)\gamma_i\sigma_j + \left(y_i - x_i^T\beta_j\right)^2 \right\} \\
&= \frac{1}{2}\left[\left\{ \gamma_i - \frac{\sum_{j=1}^{m} \frac{p_{ij}}{\sigma_j}(y_i - x_i^T\beta_j)}{\sum_{j=1}^{m} p_{ij}} \right\}^2 + constant \right] \\
&= \frac{1}{2}\left[\left\{ \gamma_i - \sum_{j=1}^{m} \frac{p_{ij}}{\sigma_j}(y_i - x_i^T\beta_j) \right\}^2 + constant \right],
\end{aligned}
$$

where $\sum_{j=1}^{m} p_{ij} = 1$.

Therefore,

$$
\gamma_i = argmin \frac{1}{2}\left[\left\{ \gamma_i - \sum_{j=1}^{m} \frac{p_{ij}}{\sigma_j}(y_i - x_i^T\beta_j) \right\}^2 \right] + P_\lambda\left(|\gamma_i|\right).
$$

Table 3.1: *Outlier Identification Results for Equal Variance Case with Large* $|\gamma|$

	5% outliers			10% outliers		
	M	S	JD	M	S	JD
Hard	0.000	0.001	1.000	0.000	0.002	1.000
Hard_{oracle}	0.000	0.001	1.000	0.000	0.000	1.000
SCAD	0.005	0.014	0.995	0.001	0.003	0.994
SCAD_{oracle}	0.000	0.031	1.000	0.000	0.002	1.000
Soft	0.066	0.017	0.920	0.840	0.005	0.000
Soft_{oracle}	0.000	0.033	1.000	0.179	0.024	0.375
$\text{TLE}_{0.05}$	0.000	0.007	1.000	0.749	0.050	0.000
$\text{TLE}_{0.10}$	0.000	0.003	1.000	0.000	0.007	1.000
MEM-bisquare	0.000	0.005	1.000	0.639	0.061	0.145
Mixregt	0.000	0.078	1.000	0.313	0.096	0.555

Table 3.2: *MeSE (MSE) of Point Estimates for Equal Variance Case with Large* $|\gamma|$

5% outliers						
	π		β		σ	
	MSE	MeSE	MSE	MeSE	MSE	MeSE
Hard	0.002	0.001	0.058	0.042	0.020	0.005
Hard_{oracle}	0.002	0.001	0.050	0.024	0.024	0.004
SCAD	0.003	0.001	0.053	0.042	0.018	0.005
SCAD_{oracle}	0.002	0.001	0.049	0.041	0.004	0.002
Soft	0.010	0.006	0.771	0.193	1.957	0.045
Soft_{oracle}	0.007	0.005	0.126	0.119	0.462	0.459
$\text{TLE}_{0.05}$	0.002	0.001	0.047	0.037	0.002	0.001
$\text{TLE}_{0.10}$	0.002	0.001	0.085	0.067	0.025	0.023
MEM-bisquare	0.002	0.001	0.050	0.041	0.007	0.004
Mixregt	0.003	0.002	0.090	0.080	0.123	0.121
MLE	0.470	0.680	17.20	20.33	2.912	2.920
10% outliers						
	π		β		σ	
	MSE	MeSE	MSE	MeSE	MSE	MeSE
Hard	0.002	0.001	0.059	0.047	0.015	0.006
Hard_{oracle}	0.001	0.001	0.071	0.044	0.002	0.002
SCAD	0.002	0.001	0.088	0.047	0.037	0.005
SCAD_{oracle}	0.001	0.001	0.012	0.045	0.002	0.002
Soft	0.293	0.409	17.85	19.64	3.308	3.026
Soft_{oracle}	0.065	0.017	11.96	6.176	2.195	2.518
$\text{TLE}_{0.05}$	0.274	0.046	50.94	49.08	0.298	0.275
$\text{TLE}_{0.10}$	0.002	0.001	0.057	0.046	0.002	0.001
MEM-bisquare	0.279	0.043	39.81	45.74	0.143	0.120
Mixregt	0.212	0.005	18.05	0.174	0.058	0.056
MLE	0.075	0.014	11.55	10.09	4.462	4.459

Table 3.3: *Outlier Identification Results for Equal Variance Case with Small* $|\gamma|$

	5% outliers			10% outliers		
	M	S	JD	M	S	JD
Hard	0.002	0.001	0.965	0.038	0.001	0.615
Hard_{oracle}	0.000	0.060	1.000	0.005	0.001	0.790
SCAD	0.001	0.001	0.950	0.957	0.001	0.000
SCAD_{oracle}	0.001	0.054	0.985	0.119	0.059	0.575
Soft	0.906	0.000	0.000	0.959	0.001	0.000
Soft_{oracle}	0.002	0.054	0.955	0.263	0.031	0.000
$\text{TLE}_{0.05}$	0.002	0.008	0.965	0.490	0.015	0.000
$\text{TLE}_{0.10}$	0.000	0.026	0.995	0.002	0.007	0.945
MEM-bisquare	0.038	0.008	0.865	0.471	0.019	0.200
Mixregt	0.000	0.074	0.990	0.007	0.050	0.930

Table 3.4: *MeSE (MSE) of Point Estimates for Equal Variance Case with Small* $|\gamma|$

5% outliers						
	π		β		σ	
	MSE	MeSE	MSE	MeSE	MSE	MeSE
Hard	0.002	0.001	0.067	0.048	0.013	0.006
Hard_{oracle}	0.002	0.001	0.057	0.048	0.005	0.002
SCAD	0.003	0.001	0.116	0.087	0.016	0.008
SCAD_{oracle}	0.002	0.001	0.080	0.068	0.006	0.003
Soft	0.003	0.001	1.056	1.031	0.281	0.259
Soft_{oracle}	0.002	0.001	0.306	0.286	0.058	0.054
$\text{TLE}_{0.05}$	0.002	0.001	0.060	0.054	0.002	0.001
$\text{TLE}_{0.10}$	0.002	0.001	0.093	0.086	0.027	0.026
MEM-bisquare	0.003	0.001	1.237	0.058	0.009	0.004
Mixregt	0.002	0.001	0.102	0.089	0.120	0.121
MLE	0.003	0.001	1.091	1.078	0.315	0.308
10% outliers						
	π		β		σ	
	MSE	MeSE	MSE	MeSE	MSE	MeSE
Hard	0.002	0.001	0.134	0.046	0.015	0.006
Hard_{oracle}	0.001	0.001	0.057	0.049	0.008	0.007
SCAD	0.002	0.001	2.310	2.273	0.563	0.565
SCAD_{oracle}	0.002	0.001	0.784	0.129	0.131	0.008
Soft	0.003	0.001	2.357	2.322	0.538	0.539
Soft_{oracle}	0.003	0.001	1.734	1.685	0.344	0.340
$\text{TLE}_{0.05}$	0.015	0.003	7.472	0.937	0.142	0.141
$\text{TLE}_{0.10}$	0.002	0.001	0.058	0.050	0.002	0.001
MEM-bisquare	0.029	0.004	7.397	1.314	0.134	0.103
Mixregt	0.005	0.001	0.247	0.111	0.074	0.074
MLE	0.003	0.001	2.386	2.347	0.576	0.567

Table 3.5: *Outlier Identification Results for Unequal Variance Case with Large* $|\gamma|$

	5% outliers			10% outliers		
	M	S	JD	M	S	JD
Hard	0.000	0.001	1.000	0.000	0.001	1.000
Hard_{oracle}	0.000	0.001	1.000	0.000	0.000	1.000
SCAD	0.002	0.004	0.995	0.003	0.005	0.990
SCAD_{oracle}	0.000	0.010	1.000	0.000	0.005	1.000
Soft	0.894	0.005	0.050	0.960	0.001	0.000
Soft_{oracle}	0.005	0.225	0.995	0.728	0.084	0.010
$\text{TLE}_{0.05}$	0.004	0.008	0.915	0.656	0.018	0.000
$\text{TLE}_{0.10}$	0.008	0.032	0.845	0.003	0.008	0.900
MEM-bisquare	0.062	0.006	0.915	0.722	0.012	0.010
Mixregt	0.000	0.078	1.000	0.461	0.097	0.200

Table 3.6: *MeSE (MSE) of Point Estimates for Unequal Variance Case with Large* $|\gamma|$

5% outliers						
	π		β		σ	
	MSE	MeSE	MSE	MeSE	MSE	MeSE
Hard	0.003	0.001	0.104	0.091	0.036	0.028
Hard_{oracle}	0.003	0.001	0.102	0.091	0.029	0.025
SCAD	0.004	0.001	0.129	0.096	0.114	0.025
SCAD_{oracle}	0.003	0.001	0.112	0.095	0.029	0.016
Soft	0.689	0.757	36.76	37.74	160.0	182.2
Soft_{oracle}	0.011	0.004	0.405	0.196	1.574	0.556
$\text{TLE}_{0.05}$	0.077	0.002	9.160	0.096	0.502	0.023
$\text{TLE}_{0.10}$	0.259	0.007	1.528	0.219	1.756	0.655
MEM-bisquare	0.087	0.004	9.835	0.115	0.637	0.102
Mixregt	0.008	0.003	0.421	0.182	0.683	0.655
MLE	0.761	0.763	43.20	41.84	186.2	186.5
10% outliers						
	π		β		σ	
	MSE	MeSE	MSE	MeSE	MSE	MeSE
Hard	0.003	0.001	0.122	0.100	0.060	0.052
Hard_{oracle}	0.003	0.001	0.122	0.100	0.056	0.044
SCAD	0.006	0.002	0.319	0.115	1.837	0.044
SCAD_{oracle}	0.003	0.002	0.205	0.108	0.094	0.046
Soft	0.587	0.590	39.49	38.87	193.5	194.2
Soft_{oracle}	0.570	0.589	46.68	45.69	110.7	112.9
$\text{TLE}_{0.05}$	0.654	0.679	98.20	90.68	1.960	1.970
$\text{TLE}_{0.10}$	0.063	0.002	10.37	0.125	0.403	0.018
MEM-bisquare	0.622	0.652	94.93	86.53	2.397	2.291
Mixregt	0.516	0.638	70.46	81.49	0.968	0.998
MLE	0.593	0.593	40.89	38.98	188.1	195.2

Table 3.7: *Outlier Identification Results for Unequal Variance Case with Small* $|\gamma|$

	5% outliers			10% outliers		
	M	S	JD	M	S	JD
Hard	0.003	0.001	0.955	0.649	0.000	0.125
Hard_{oracle}	0.001	0.001	0.995	0.051	0.006	0.725
SCAD	0.828	0.001	0.055	0.951	0.001	0.000
SCAD_{oracle}	0.001	0.070	0.980	0.300	0.061	0.215
Soft	0.889	0.001	0.000	0.952	0.001	0.000
Soft_{oracle}	0.000	0.233	1.000	0.423	0.050	0.000
$\text{TLE}_{0.05}$	0.004	0.008	0.945	0.672	0.017	0.000
$\text{TLE}_{0.10}$	0.001	0.029	0.980	0.005	0.008	0.885
MEM-bisquare	0.234	0.007	0.590	0.734	0.008	0.000
Mixregt	0.001	0.085	0.990	0.092	0.060	0.820

Table 3.8: *MeSE (MSE) of Point Estimates for Unequal Variance Case with Small* $|\gamma|$

5% outliers						
	π		β		σ	
	MSE	MeSE	MSE	MeSE	MSE	MeSE
Hard	0.003	0.001	0.146	0.104	0.038	0.020
Hard_{oracle}	0.003	0.001	0.125	0.112	0.022	0.011
SCAD	0.114	0.032	5.877	3.617	1.797	1.726
SCAD_{oracle}	0.003	0.001	0.167	0.136	0.022	0.013
Soft	0.123	0.037	6.296	3.819	1.954	1.814
Soft_{oracle}	0.003	0.002	0.451	0.425	0.023	0.015
$\text{TLE}_{0.05}$	0.004	0.002	0.129	0.111	0.031	0.020
$\text{TLE}_{0.10}$	0.017	0.003	0.863	0.145	0.237	0.176
MEM-bisquare	0.183	0.005	9.725	0.193	0.443	0.123
Mixregt	0.007	0.003	0.210	0.178	0.700	0.711
MLE	0.443	0.583	16.67	18.66	5.714	2.926
10% outliers						
	π		β		σ	
	MSE	MeSE	MSE	MeSE	MSE	MeSE
Hard	0.086	0.019	7.360	6.213	1.743	1.764
Hard_{oracle}	0.005	0.003	0.300	0.112	0.265	0.043
SCAD	0.150	0.077	10.98	8.265	3.037	3.005
SCAD_{oracle}	0.007	0.002	3.412	4.103	1.090	1.208
Soft	0.150	0.077	10.97	8.264	3.043	3.011
Soft_{oracle}	0.062	0.025	7.040	6.000	1.891	1.767
$\text{TLE}_{0.05}$	0.437	0.487	25.00	23.95	1.555	1.552
$\text{TLE}_{0.10}$	0.004	0.002	0.145	0.111	0.034	0.027
MEM-bisquare	0.429	0.477	22.85	22.22	2.362	2.290
Mixregt	0.074	0.004	4.298	0.303	0.513	0.444
MLE	0.310	0.361	16.44	17.62	3.480	3.613

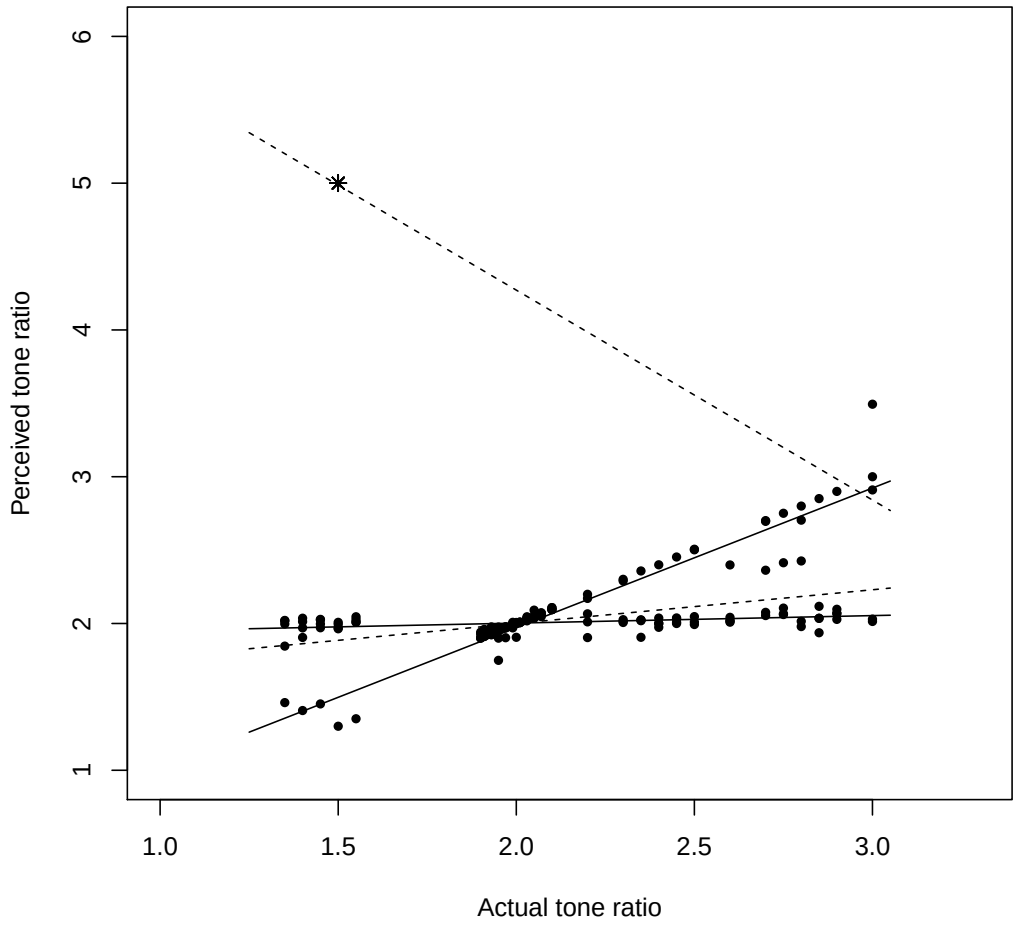

Figure 3.1: *The scatter plot of the tone perception data and the fitted mixture regression lines with added ten identical outliers* $(1.5, 5)$ *(denoted by stars at the upper left corner). The predictor is actual tone ratio and the response is the perceived tone ratio by a trained musician. The solid lines represent the fit by the proposed Hard and the dashed lines represent the fit by the traditional MLE.*

Bibliography

[1] Antoniadis, A. (1997), Wavelets in Statistics: A Review (with discussion). *Journal of the Italian Statistical Association*, 6, 97-144.

[2] Bai, X., Yao, W., and Boyer, J. E. (2012), Robust fitting of mixture regression models. *Computational Statistics and Data Analysis*, 56, 2347-2359.

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

[4] Böhning, D. (1999), *Computer-Assisted Analysis of Mixtures and Applications*. Boca Raton, FL: Chapman and Hall/CRC.

[5] Carroll, R. J. and Welsch, A. H. (1988), A Note on Asymmetry and Robustness in Linear Regression. *Journal of American Statisitcal Association*, 4, 285-287.

[6] 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.

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

[8] Coakley, C. W. and Hettmansperger, T. P. (1993), A Bounded Influence, High Breakdown, Efficient Regression Estimator. *Journal of American Statistical Association*, 88, 872-880.

[9] Crawford, S. L., Degroot, M. H., Kadane, J. B., and Small, M. J. (1992), Modeling lake-chemistry distributions-approximate Bayesian methods for estimating a finite-mixture model. *Technometrics*, 34, 441-453.

[10] Crawford, S. L. (1994), An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89, 259-267.

[11] Croux, C., Rousseeuw, P. J., and Hössjer O. (1994), Generalized S-estimators. *Journal of American Statistical Association*, 89, 1271-1281.

[12] 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, Ser. B*, 39, 1-38.

[13] Donoho, D. L. and Huber, P. J. (1983), The Notation of Break-down Point, *in A Festschrift for E. L. Lehmann, Wadsworth*

[14] Donoho, D. L. and Johnstone, I. M. (1994a), Ideal Spatial Adaptation by Wavelet shrinkage, *Biometrika*, 81, 425-455.

[15] Everitt, B. S. and Hand D. J. (1981), Finite Mixture Distributions. Chapman and Hall, London

[16] Freedman, W. L., Wilson, C. D., and Madore, B. F. (1991), New Cepheid Distances to Nearby Galaxies Based on BVRI CCD Photometry. *Astrophysical Journal*, 372, 455-470.

[17] Fan, J. and Li, R. (2001), Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96, 1348-1360.

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

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

[20] 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 Series*, B71, 301-318.

[21] Gervini, D. and Yohai, V. J. (2002), A Class of Robust and Fully Efficient Regression Estimators. *The Annals of Statistics*, 30, 583-616.

[22] Goldfeld, S. M. and Quandt, R. E. (1973), A Markov model for switching regression. *Journal of Econometrics*, 1, 3-15.

[23] Handschin, E., Kohlas, J., Fiechter, A., and Schweppe, F. (1975), Bad Data Analysis for Power System State Estimation. *IEEE Transactions on Power Apparatus and Systems*, 2, 329-337.

[28] Huber, P.J. (1981), Robust Statistics. *New York: John Wiley and Sons.*

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

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

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

[28] Huber, P.J. (1981), Robust Statistics. *New York: John Wiley and Sons.*

[29] Jackel, L.A. (1972), Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals. *Annals of Mathematical Statistics*, 5, 1449-1458.

[30] 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.

[31] Lee, Y., MacEachern, S. N., and Jung, Y. (2011), Regularization of Case-Specific Parameters for Robustness and Efficiency. *Submitted to the Statistical Science*.

[32] Lindsay, B. G. (1995), Mixture Models: Theory, Geometry and Applications, *NSF-CBMS Regional Conference Series in Probability and Statistics*, Vol. 5. Institute of mathematical Statistics and the American Statisitcal Association, Alexandra, VA.

[33] Mallows, C.L. (1975), On Some Topics in Robustness. unpublished memorandum, Bell Tel. Laboratories, Murray Hill.

[34] Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006), Robust Statistics. *John Wiley*.

[35] Markatou, M. (2000), Mixture models, robustness, and the weighted likelihood methodology. *Biometrics*, 56, 483-486.

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

[37] McLachlan, G. J. and Basford, K. E. (1988), Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

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

[39] 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.

[40] Naranjo, J.D., Hettmansperger, T. P. (1994), Bounded Influence Rank Regression. *Journal of the Royal Statistical Society B*, 56, 209-220.

[41] 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.

[42] Richardson, S. and Green, P. J. (1997), On Bayesian analysis of mixtures with an unknown number of components (with discussion).*Journal of The Royal Statistical Society Series*, B , 59, 731-792.

[43] Rousseeuw, P.J.(1982), Least Median of Squares regression. Resaerch Report No. 178, Centre for Statistics and Operations research, VUB Brussels.

[44] Rousseeuw, P.J.(1983), Multivariate Estimation with High Breakdown Point. Resaerch Report No. 192, Center for Statistics and Operations research, VUB Brussels.

[45] Rousseeuw, P.J. and Croux, C.(1993), Alternatives to the Median Absolute Deviation. *Journal of American Statistical Association*, 94, 388-402.

[46] Rousseeuw, P.J. and Yohai, V. J. (1984), Robust Regression by Means of S-estimators. *Robust and Nonlinear Time series*, J. Franke, W. Härdle and R. D. Martin (eds.),Lectures Notes in Statistics 26, 256-272, New York: Springer.

[47] She, Y. (2009), Thresholding-Based Iterative Selection Procedures for Model Selection and Shrinkage. *Electronic Journal of Statistics*, 3, 384C415.

[48] She, Y. and Owen, A. (2011), Outlier Detection Using Nonconvex Penalized Regression. *Journal of the American Statistical Association*, 106, 626-639.

[49] 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*, 2004, Vol. 3192/2004, 208-215.

[50] Siegel, A.F. (1982), Robust Regression Using Repeated Medians. *Biometrika*, 69, 242-244.

[51] Song, W., Yao, W., and Xing Y. (2013), Robust mixture regression model fitting by laplace distribution. To appear at *Computational Statistics and Data Analysis*

[52] Stromberg, A. J., Hawkins, D. M., and Hössjer, O. (2000), The Least Trimmed Differences Regression Estimator and Alternatives. *Journal of American Statistical Association*, 95, 853-864.

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

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

[55] Tibshirani, R. J. (1996), Regression Shrinkage and Selection via the LASSO. *Journal of The Royal Statistical Society Series*, B58, 267-288.

[56] Tibshirani, R. J. (1996), The LASSO Method for Variable Selection in the Cox Model. *Statistics in Medicine*, 16, 385-395.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), Statistical Analysis of Finite Mixture Distribution. Wiley, New York.

[57] 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.

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

[59] Yao, W. (2012a), A Simple Solution to Bayesian Mixture Labeling. *Statistics and Computing.*

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

[61] Yao, W., Wei, Y., and Yu, C. (2014), Robust mixture regression using the t-distribution. To appear at *Computational Statistics and Data Analysis*, 71, 116-127.

[62] Yohai, V. J. (1987), High Breakdown-point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*, 15, 642-656.

[63] Zou, H. (2006), The Adaptive Lasso and Its Oracle Properties. *Journal of American Statistical Association*, 101, 1418-1429.

Bibliography