UC Riverside
UC Riverside Previously Published Works

Title
Outlier detection and robust mixture modeling using nonconvex penalized likelihood

Permalink
https://escholarship.org/uc/item/8vw5m7q9

Journal
Journal of Statistical Planning and Inference, 164(1)

ISSN
0378-3758

Authors
Yu, Chun
Chen, Kun
Yao, Weixin

Publication Date
2015-09-01

DOI
10.1016/j.jspi.2015.03.003

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Outlier Detection and Robust Mixture Modeling Using Nonconvex Penalized Likelihood

Chun Yu, [*] Kun Chen,[†] Weixin Yao, [‡]

Abstract

Finite mixture models are widely used in a variety of statistical applications. However, the classical normal mixture model with maximum likelihood estimation is prone to the presence of only a few severe outliers. We propose a robust mixture modeling approach using a mean-shift formulation coupled with nonconvex sparsity-inducing penalization, to conduct simultaneous outlier detection and robust parameter estimation. An efficient iterative thresholding-embedded EM algorithm is developed to maximize the penalized log-likelihood. The efficacy of our proposed approach is demonstrated via simulation studies and a real application on Acidity data analysis.

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

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,

[*]Chun Yu is Instructor, School of Statistics, Jiangxi University of Finance and Economics, Nanchang 330013, China. Email: chuckyu0106@126.com.

[†]Kun Chen is Assistant Professor, Department of Statistics, University of Connecticut, Storrs, CT, 06269. Email: kun.chen@uconn.edu.

[‡]Weixin Yao is corresponding author, Associate Professor, Department of Statistics, University of California, Riverside, CA 92521. Email: weixin.yao@ucr.edu.

McLachlan and Basford, 1988, Lindsay, 1995, and Böhning, 1999). The m-component finite normal mixture distribution has probability density

$$f(y;\boldsymbol{\theta}) = \sum_{i=1}^{m} \pi_i \phi(y;\mu_i,\sigma_i^2), \tag{1.1}$$

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 jth subpopulation with $\sum_{j=1}^{m}\pi_j = 1$. Given observations $(y_1,\ldots,y_n)$ from model (1.1), the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is given by,

$$\hat{\boldsymbol{\theta}}_{\text{mle}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{m} \pi_j \phi(y_i;\mu_j,\sigma_j^2)\right\}, \tag{1.2}$$

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, the method is not robust and both parameter estimation and inference can fail miserably in the presence of outliers. Our focus in this paper is hence on robust estimation and accurate outlier detection in mixture model. For the simpler problem of estimating 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 (1.2) by a robust criterion similar to 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), Bashir and Carter (2012), Yao et al. (2014), and Song et al. (2014)

We propose a new robust mixture modelling approach based on a mean-shift model formulation coupled with 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 caused 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 we thus propose a general scale-free and case-specific mean-shift formulation to solve the general problem.

2 Robust Mixture Model via Mean-Shift Penalization

In this section, we 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 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 two cases separately.

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

$$f(y_i; \boldsymbol{\theta}, \gamma_i) = \sum_{j=1}^{m} \pi_j \phi(y_i - \gamma_i; \mu_j, \sigma^2), \qquad i = 1, \ldots, n, \tag{2.1}$$

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. Apparently, without any constraints, the addition of the mean-shift parameters results in an overly parameterized model. The key here is to assume that the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$ is sparse, i.e., γ_i is zero when the ith data point is a normal observation with any of the m mixture components, and only when the ith observation is an outlier, γ_i becomes nonzero to capture the outlying effect. Therefore, the sparse estimation of $\boldsymbol{\gamma}$ provides a direct way to accommodate and identify outliers.

Due to the sparsity assumption of $\boldsymbol{\gamma}$, we propose to maximize the following penalized log-likelihood criterion to conduct model estimation and outlier detection,

$$pl_1(\boldsymbol{\theta}, \boldsymbol{\gamma}) = l_1(\boldsymbol{\theta}, \boldsymbol{\gamma}) - \sum_{i=1}^{n} \frac{1}{w_i} P_\lambda(|\gamma_i|) \tag{2.2}$$

where $l_1(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i - \gamma_i; \mu_j, \sigma^2) \right\}$, w_is are some prespecified weights, $P_\lambda(\cdot)$ 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. In practice, w_is can be chosen to reflect any available prior information about how likely that y_is are outliers; to focus on the main idea, here we mainly consider $w_1 = w_2 = \ldots = w_n = w$, and discuss the choice of w for different penalty functions.

Some commonly used sparsity-inducing penalty functions include the ℓ_1 penalty (Donoho and Johnstone, 1994a; Tibshirani, 1996, 1997)

$$P_\lambda(\gamma) = \lambda |\gamma|, \tag{2.3}$$

the ℓ_0 penalty (Antoniadis, 1997)

$$P_\lambda(\gamma) = \frac{\lambda^2}{2} I(\gamma \neq 0), \tag{2.4}$$

and the SCAD penalty (Fan and Li, 2001)

$$P_\lambda(\gamma) = \begin{cases} \lambda|\gamma| & \text{if } |\gamma| \leq \lambda, \\ -\left(\frac{\gamma^2 - 2a\lambda|\gamma| + \lambda^2}{2(a-1)}\right) & \text{if } \lambda < |\gamma| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\gamma| > a\lambda, \end{cases} \tag{2.5}$$

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-thresholding rule, and ℓ_0 penalization corresponds to a hard-thresholding rule. It is also known that the convex ℓ_1 penalization often leads to over-selection and substantial bias in estimation. Indeed, as shown by She and Owen (2011) in the context of linear regression, ℓ_1 penalization in mean-shift model has connections with M-estimation using Huber's loss and usually leads to poor performance in outlier detection. Similar phenomenon is also observed in our extensive numerical studies. However, if there are no high leverage outliers, the ℓ_1 penalty works well and succeeds to detect the outliers, see for examples, Dalalyan and Keriven (2012); Dalalyan and Chen (2012); Nguyen and Tran (2013).

We propose a thresholding embedded EM algorithm to maximize the objective function (2.2). 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 $\mathbf{z}_i = (z_{i1}, \ldots, z_{im})$. The complete penalized log-likelihood function based on the complete data $\{(y_i, \mathbf{z}_i), i = 1, 2, \ldots, n\}$ is

$$pl_1^c(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \sum_{i=1}^n \sum_{j=1}^m z_{ij} \log\left\{\pi_j \phi(y_i - \gamma_i; \mu_j, \sigma^2)\right\} - \sum_{i=1}^n \frac{1}{w} P_\lambda(|\gamma_i|). \tag{2.6}$$

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 conditional expectation of the complete penalized log-likelihood function (2.6), i.e., $\mathrm{E}\{pl_1^c(\boldsymbol{\theta},\boldsymbol{\gamma}) \mid \boldsymbol{\theta}^{(k)},\boldsymbol{\gamma}^{(k)}\}$. The problem simplifies to the calculation of $\mathrm{E}(z_{ij}|y_i;\boldsymbol{\theta}^{(k)},\boldsymbol{\gamma}^{(k)})$,

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

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,

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

We propose to alternatingly update $\{\sigma,\mu_j, j = 1,\ldots,m\}$ and $\boldsymbol{\gamma}$ until convergence to get $\{\mu_j^{(k+1)}, j = 1,\ldots,m;\sigma^{(k+1)},\boldsymbol{\gamma}^{(k+1)}\}$. Given $\boldsymbol{\gamma}$, μ_js and σ are updated by

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

Given μ_js and σ, $\boldsymbol{\gamma}$ is updated by maximizing

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

which is equivalently to minimizing

$$\frac{1}{2}\left\{\gamma_i - \sum_{j=1}^m p_{ij}^{(k+1)}(y_i-\mu_j)\right\}^2 + \frac{1}{w}\sigma^2 P_\lambda\left(|\gamma_i|\right), \tag{2.7}$$

separately for each γ_i, $i = 1,\ldots,n$. A detailed derivation is presented in the Appendix. For either the ℓ_1 or the ℓ_0 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 these penalties. 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 that 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 ℓ_0 penalty assuming $w = 1$.

As shown in Proposition 1 below, when the ℓ_1 penalty is used, (2.7) is minimized by a soft thresholding rule, and when the ℓ_0 penalty is used, (2.7) is minimized by a hard thresholding rule. When the SCAD penalty is used, however, the problem is solved by a modified SCAD thresholding rule, which is shown in Lemma 1.

Proposition 1. *Let $\xi_i = \sum_{j=1}^m p_{ij}^{(k+1)}(y_i - \mu_j)$. Let $w = 1$ in (2.7). When the penalty function in (2.7) is the ℓ_1 penalty (2.8), the minimizer of (2.7) is given by*

$$\hat{\gamma}_i = \Theta_{soft}(\xi_i; \lambda, \sigma) = sgn(\xi_i)\left(|\xi_i| - \sigma\lambda\right)_+, \tag{2.8}$$

where $a_+ = \max(a, 0)$. When the penalty function in (2.7) is the ℓ_0 penalty (2.9), the minimizer of (2.7) is given by

$$\hat{\gamma}_i = \Theta_{hard}(\xi_i; \lambda, \sigma) = \xi_i I(|\xi_i| > \sigma\lambda), \tag{2.9}$$

where $I(\cdot)$ denotes the indicator function.

Lemma 1. *Let $\xi_i = \sum_{j=1}^m p_{ij}^{(k+1)}(y_i - \mu_j)$. Let $w = \hat{\sigma}^2$ in (2.7), a robust estimator of σ^2. When the penalty function in (2.7) is the SCAD penalty (2.5), the minimizer of (2.7) is given by*

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

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

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

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

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

$$\hat{\gamma}_i = \Theta_{scad}(\xi_i;\lambda,\sigma) = \xi_i I(|\xi_i| > \sqrt{\frac{\sigma^2(a+1)}{\hat{\sigma}^2}}\lambda). \tag{2.12}$$

The detailed EM algorithm is summarized in Algorithm 1. For simplicity, we have used $\Theta(\xi_i;\lambda,\sigma)$ to denote a general thresholding rule determined by the adopted penalty function, e.g., the modified SCAD thresholding rule $\Theta_{\text{scad}}(\xi_i;\lambda,\sigma)$ defined in Lemma 1. The convergence property of the proposed algorithm is summarized in Theorem 2.1 below, which follows directly from the property of the EM algorithm, and hence its proof is omitted.

Theorem 2.1. *Each iteration of E step and M step of Algorithm 1 monotonically non-decreases the penalized log-likelihood (2.2), i.e.,* $pl_1(\boldsymbol{\theta}^{(k+1)},\boldsymbol{\gamma}^{(k+1)}) \ge pl_1(\boldsymbol{\theta}^{(k)},\boldsymbol{\gamma}^{(k)})$*, for all* $k \ge 0$.

2.2 RMM for Unequal Component Variances

When the component variances are unequal, the naive mean-shift model (2.1) 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 Proposition 1, 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 a general scale-free mean-shift model to incorporate the information on com-

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

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.

ponent 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, we can 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). As the construction is similar to the case of equal variances, we omit the details of its derivation. The proposed EM algorithm is presented in Algorithm 2, and here we shall briefly remark the main changes. Unlike in the case of equal variances, the update of σ_j^2 in (2.17), with other parameters held fixed, does not have explicit solution in general and requires some numerical algorithm to solve, e.g., the Newton-Raphson method; as the problem is one dimensional, the computation remains very fast. In the case of unequal variances, the problem of updating $\boldsymbol{\gamma}$, with other parameters held fixed, is still separable in each γ_i, i.e., at the $(k+1)$th iteration,

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

It can be readily shown that the solution is given by simple threhsolding rules. In particular, using the ℓ_1 penalty leads to $\hat{\gamma}_i=\Theta_{\text{soft}}(\xi_i;\lambda,1)$ and using the ℓ_0 penalty leads to $\hat{\gamma}_i=\Theta_{\text{hard}}(\xi_i;\lambda,1)$, where Θ_{soft} and Θ_{hard} are defined in Proposition 1, and here in the case of unequal variance, ξ_i becomes

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

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 $\{\mu_j^{(k+1)},\sigma_j^{2^{(k+1)}}, j = 1,\ldots,m,\boldsymbol{\gamma}^{(k+1)}\}$:

$$(2.a) \qquad \gamma_i \leftarrow \Theta(\xi_i;\lambda,1), \text{ where } \xi_i = \sum_{j=1}^m p_{ij}^{(k+1)}(y_i-\mu_j)/\sigma_j, \tag{2.15}$$

$$(2.b) \qquad \mu_j \leftarrow \frac{\sum_{i=1}^n p_{ij}^{(k+1)}(y_i - \gamma_i\sigma_j)}{\sum_{i=1}^n p_{ij}^{(k+1)}}, \tag{2.16}$$

$$(2.c) \qquad \sigma_j^2 \leftarrow \arg\max_{\sigma_j} \sum_{i=1}^n p_{ij}^{(k+1)} \log\phi(y_i - \gamma_i\sigma_j;\mu_j,\sigma_j^2). \tag{2.17}$$

$k \leftarrow k+1$.

until convergence

As the γ_is become scale free, the thresholding rule for solving SCAD becomes much simpler, and it is given by (2.10) when setting the quantity $\sigma^2/\hat{\sigma}^2 = 1$, i.e.,

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

Similar to Theorem 2.1, it is easy to check that the monotone non-decreasing property remains hold for Algorithm 2. We note that in both algorithms, we have used an iterative algorithm aiming to fully maximize the expected complete log-likelihood under penalization. It can be seen that in this blockwise coordinate descent algorithm, each loop of $(2.a) - (2.c)$ monotonically non-decreases the objective function. Therefore, an alternative strategy is to run $(2.a) - (2.c)$ only a few times or even just once in each M-step; the resulting generalized EM algorithm continues to possess the desirable convergence property. Based on our limited experience, however, this method generally does not lead to significant saving in computation, because the iterations in the M-step only involve simple operations and partially solving M-step may slow down the overall convergence. Nevertheless, it is worthwhile to point out this strategy, as it can be potentially useful when more complicated penalization methods are required.

2.3 Tuning Parameter Selection

When using robust estimation or outlier detection methods, it is usually required to choose a "threshold" value, e.g., the percentage of observations to eliminate, or the cutoff to declare extreme residuals. In our method, selecting "threshold" becomes the tuning parameter selection problem in penalized regression (2.2) and (2.14). As such, many well-developed methodologies including cross validation and information criterion based approaches are all applicable, and the turning parameter λ can be selected in an objective way, based on predictive power of the model or the balance between model goodness of fit and complexity. Here, we provide a data adaptive way to select λ based on a Bayesian information criterion (BIC), due to its computation efficiency and proven superior performance on variable selection,

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

where $j = 1$ or 2, $l_j^*(\lambda) = l_j(\hat{\boldsymbol{\theta}}(\lambda), \hat{\boldsymbol{\gamma}}(\lambda))$ is the mixture log-likelihood evaluated at the estimator $(\hat{\boldsymbol{\theta}}(\lambda), \hat{\boldsymbol{\gamma}}(\lambda))$ obtained by maximizing the penalized likelihood criterion (2.2) or (2.14) with λ being the tuning parameter, and $\mathrm{df}(\lambda)$ 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. In practice, the optimal tuning parameter λ is chosen by minimizing $\mathrm{BIC}(\lambda)$ 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 resulting in all zero values in $\hat{\boldsymbol{\gamma}}$, corresponding to the case of no outlier, and $\lambda_{\min}$ is some small value of λ resulting in roughly 40% nonzero values in $\hat{\boldsymbol{\gamma}}$, since in reality the proportion of outliers is usually quite small. The models with various λ values are fitted sequentially. The previous solution is used as the initial value for fitting the next model to speed up the computation. As such, our proposed method is able to search conveniently over a whole spectrum of possible models.

In mixture model, it is a foremost task to determine the number of mixture component m. The problem may be resolved based on prior knowledge of the underlying data generation process. In many applications where no prior information is available, we suggest to conduct the penalized mixture model analysis with a few plausible m values, and use the proposed BIC criterion to select both the number of component m and the amount of penalization λ.

3 Simulation

3.1 Setups

We conduct simulation studies to investigate the effectiveness of the proposed approach and compare it with several existing methods. We consider both the case of equal variances and the case of unequal variances. In each setup to be elaborated below, we first generate independent observations from a normal mixture distribution; a few outliers are then created by adding random mean-shift to some of the observations. The sample size is set to $n = 200$, and we consider two proportions of outliers, i.e., $p_{\mathcal{O}} = 5\%$ and $p_{\mathcal{O}} = 10\%$. The number of replicates is 200 for each simulation setting.

Example 1: The samples $(y_1, y_2, \dots, y_n)$ are generated from model (2.1) with $\pi_1 = 0.3$, $\mu_1 = 0$, $\pi_2 = 0.7$, $\mu_2 = 8$, and $\sigma = 1$. That is, the size of the first component n_1 is generated from a binomial distribution with $n_1 \sim \mathrm{Bin}(n, p = 0.3)$, and consequently the size of the second component is given by $n_2 = n - n_1$. To create $100p_{\mathcal{O}}\%$ outliers, we randomly choose $3np_{\mathcal{O}}/10$ many observations from component 1, and each of them is added a random mean shift $\gamma \sim \mathrm{Unif}([-5, -7])$. Similarly $7np_{\mathcal{O}}/10$ outliers are created by adding random mean shift $\gamma \sim \mathrm{Unif}([5, 7])$ to observations from component 2.

Example 2: The samples $(y_1, y_2, \dots, y_n)$ are generated from model (2.13) with $\pi_1 = 0.3$, $\mu_1 = 0$, $\sigma_1 = 1$, $\pi_2 = 0.7$, $\mu_2 = 8$, and $\sigma_2 = 2$. All other settings are the same as in Example 1, except that when generating outliers, we add an amount $\mathrm{Unif}([-5\sigma_1, -7\sigma_1])$ to observations from component 1 and $\mathrm{Unif}([5\sigma_2, 7\sigma_2])$ to observations from component 2.

In the above simulation examples, the majority of data points form two well-separated clusters. There are very few extreme observations (5% or 10%), which are far away from both the cluster centers. As such, it is appropriate to model these anomaly observations as outliers in a two-component mixture model.

3.2 Methods and Evaluation Metrics

We use our proposed RMM approaches with several different penalty forms including ℓ_0, ℓ_1 and SCAD penalties, denoted as Soft, Hard and SCAD, respectively. For each penalty, our approach efficiently produces a solution path with varying numbers of outliers. The optimal solution is selected by the BIC criterion. To investigate the performance of BIC and to better understand the true potential of each penalization method, we also report an "oracle" estimator, which is defined as the solution having the best outlier detection performance along the fitted solution path. When there are multiple such solutions on the path, we choose the one gives the best parameter estimates. These oracle estimators are denoted as $\text{Soft}_{\mathcal{O}}$, $\text{Hard}_{\mathcal{O}}$ and $\text{SCAD}_{\mathcal{O}}$. We note that the oracle estimators rely on the knowledge of the true parameter values, and thus they are not feasible to compute in practice. Nevertheless, as we shall see below, they provide interesting information about the behaviors of different penalty forms. We also compare our RMM approach to the nonrobust maximum likelihood estimation method (MLE) and the robust trimmed likelihood estimation method (TLE) proposed by Neykov et al. (2007), with the percentage of trimmed data α set to either 0.05 ($\text{TLE}_{0.05}$) or 0.10 ($\text{TLE}_{0.1}$). TLE methods require a cutoff value η to identify extreme residuals; following Gervini and Yohai (2002), we set $\eta = 2.5$.

To evaluate the outlier detection performance, we report (1) the proportion of masking (M%), i.e., the fraction of undetected outliers, (2) the proportion 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% $\approx$ 100%. To evaluate the performance of parameter estimation, we report both the mean squared errors (MSE) and the robust median squared errors (MeSE) of the parameter estimates.

A very important usage of mixture model is for clustering. From the fitted mixture model, the Bayes classification rule assigns the ith observation to cluster j such that $j = \arg\max_k p_{ik}$, where p_{ik}, $k = 1, \ldots, m$, are the set of cluster probabilities for the ith observation directly produced from the EM algorithm. We thus compute the average misclassification rate (Mis%) to evaluate the clustering performance of each method. We note that for mixture models, there are well-known label switching issues (Celeux, et al., 2000; Stephens, 2000; Yao and Lindsay, 2009; Yao, 2012a, 2012b). Roughly speaking, the mixture likelihood function is invariant to the permutation of the component labels, so that the component parameters are not identifiable marginally since they are exchangeable. As a consequence, the estimation results from different simulation runs are not directly comparable, as the mixture components in each simulation run can be labeled arbitrarily. In our examples, the component labels in each simulation are aligned to the reference label of the true parameter values, i.e., the labels are chosen by minimizing the distance from the resulting parameter estimates to the true parameter values.

3.3 Results

The simulation results are summarized in Tables 1 to 4. Not surprisingly, MLE fails in all the cases. This demonstrates that robust mixture modeling is indeed needed in the presence of rare but severe outliers.

In case of equal variances, both Hard and SCAD perform very well, and their performance on outlier detection is very similar to their oracle counterparts. While the Soft method performs well in outlier detection when $p_{\mathcal{O}} = 5\%$, its performance becomes much worse when $p_{\mathcal{O}} = 10\%$ mainly due to masking. The observed nonrobustness of Soft is consistent with the results in She and Owen (2011). In terms of parameter estimation, Hard and Hard$_{\mathcal{O}}$ perform the best among the RMM methods. On the other hard, SCAD$_{\mathcal{O}}$ performs better than Soft$_{\mathcal{O}}$ and they are slightly outperformed by SCAD and Soft, respectively. This interesting phenomenon reveals some important behaviors of the penalty functions. When using the ℓ_0 penalty, the effect of an outlier is completely captured by its estimated mean-shift parameter whose magnitude is not penalized, so once an observation is detected as an outlier, i.e., its mean-shift parameter is estimated to be nonzero, it does not affect parameter estimation any more. However, when using ℓ_1 type penalty, due to its inherit shrinkage effects on the mean-shift parameters, the model tries to accommodate the effects of severe outliers in estimation. Even if an observation is detected as an outlier with a nonzero mean-shift, it may still partially affects parameter estimation as the magnitude of the mean-shift parameter is shrunk towards zero. As a consequence, the oracle estimator which has the best outlier detection performance does not necessarily leads to the best estimation. Since the SCAD penalty can be regarded as a hybrid between ℓ_0 and ℓ_1, it exhibits behaviors that are characteristics of both of ℓ_0 and ℓ_1. Further examination of the simulation results reveals that Soft$_{\mathcal{O}}$ (SCAD$_{\mathcal{O}}$) tends to require a stronger penalty than the Soft (SCAD) estimator in order to reduce false positives, which induces heavier shrinkage of $\boldsymbol{\gamma}$, and consequently the former is distorted more by the outliers than the latter. The TLE method leads to satisfactory results when the trimming proportion is correctly specified. It loses efficiency when the trimming proportion is too large and fails to be robust when the trimming proportion is too small. Our RMM methods can achieve comparable performance to the oracle TLE that assumes the correct trimming proportion.

In case of unequal variances, the behaviors of the RMM estimators and their oracle counterparts are similar to those in the case of equal variances. Hard still performs the best among all feasible estimators in both outlier detection and parameter estimation. SCAD and Soft work satisfactorily when $p_{\mathcal{O}} = 5\%$. However, when $p_{\mathcal{O}} = 10\%$, the two methods may fail to detect outliers and their average masking rates become 18.72% and 55.67%, respectively. Again, this can be explained by the shrinkage effects on the mean-shift parameters induced by the penalty forms. Nevertheless, SCAD is affected much less and thus performs much better in parameter estimation then Soft.

We have investigated the problem of selecting the number of mixture components using the proposed BIC criterion. In Example 2 with unequal variances and $p_{\mathcal{O}} = 5\%$, we use the RMM method to fit models with 2, 3, and 4 mixture components. The two-component model is selected 100%, 98% and 63% of the time when using Hard, SCAD and Soft, respectively, based on 200 simulated datasets. The results are similar using Example 1 and/or $p_{\mathcal{O}} = 10\%$. These results again suggest that RMM works well with nonconvex penalty forms. In Table 5, we compare the average computation times. As expected, RMM tends to be slightly slower than TLE and MLE, mainly because the M-step has to be solved by an iterative procedure. In general, the computation time of RMM increases slightly as the proportion of outliers increases, and the case of unequal variances needs slightly longer time to compute than the case of equal variances. Nevertheless, the proposed RMM method remains to be very computationally efficient and the speed can be further improved with more careful implementation. (A user-friendly R package for RMM will be made available to the public).

In summary, our RMM approach using nonconvex penalization, together with the proposed BIC criterion, achieves the dual goal of accuracy outlier detection and robust parameter estimation. In practice, the proportion of extreme outliers is usually very small in mixture model setup, and we suggest to use either the ℓ_0 or the SCAD penalty. Other nonconvex penalty forms such as the minimax concave penalty (MCP) (Zhang, 2010) can also be used.

4 Acidity Data Analysis

We 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 of 155 lakes in north-central Wisconsin. More details on the data can be found in Crawford (1994), Crawford et al. (1992), and Richardson and Green (1997). Figure 1 shows the histogram of the observed acidity indices.

Following Richardson and Green (1997), we fit the data by a three-component normal mixture model with equal variances, using both the raditional MLE method and the proposed RMM approach with ℓ_0 penalty. The tuning parameter in RMM is selected by BIC. Table 6 shows the parameter estimates. In the original data, there does not appear to be outliers, and the proposed RMM approach results in very similar parameter estimates to that of the traditional MLE. This shows that RMM does not lead to efficiency loss when there is no outlier, and its performance is as good as that of MLE.

Following Peel and McLachlan (2000), to examine the effects of outliers, we add one outlier ($y = 12$) to the original data. While RMM is not influenced by the outlier and gives similar parameter estimates to the case of no outliers, MLE leads to very different parameter estimates. Note the first and second components are estimated to have the same mean based on MLE, thus the model essentially has only two components. We then add three identical outliers ($y = 12$) to the data. As expected, RMM still provides similar estimates as before. However, MLE fits a new component to the outliers and gives drastically different estimates comparing to the case of no outliers. In fact, in both cases, RMM successfully detects the added extreme observations as outliers, so that the parameter estimation remains unaffected. This example shows that our proposed RMM method provides a stable and robust way for fitting mixture models, especially in the presence of severe outliers.

5 Discussion

We have developed a robust mixture modelling approach under the penalized estimation framework. Our robust method with nonconvex penalizaton is capable of conducting simultaneous outlier detection and robust parameter estimation. The method has comparable performance to TLE that uses an oracle trimming proportion. However, our method can efficiently produce a solution path consisting of solutions with varying number of outliers, so that the proportion of outliers and the accommodation of them in estimation can both be efficiently determined data adaptively.

There are many directions for future research. It is pressing to investigate the theoretical properties of the proposed RMM approach, e.g., the selection consistency of outlier detection. As RMM is formulated as a penalized estimation problem, the many established results on penalized variable selection may shed light on this problem; see. e.g., Khalili (2007) and Stadler (2010). Our proposed general scaled-dependent outlier detection model shares similar idea with the reparameterized model proposed by Stadler (2010), and our model can be written as a penalized mixture regression problem. However, their approach for establishing the oracle properties of the penalized estimator is not directly applicable to our problem, as in our case the design matrix associated with the mean-shift parameters becomes a fixed identity matrix of dimension n. We have mainly focused on normal mixture model in this paper, but the method can be readily extended to other mixture models, such as mixtures of binomial and mixtures of Poisson. It would also be interesting to extend the method to multivariate mixture models and mixture regression models.

Acknowledgements

We thank the two referees and the Associate Editor, whose comments and suggestions have helped us to improve the paper significantly. Yao's research is supported by NSF grant DMS-1461677.

Appendix

Derivation of Equation (2.7)

The estimate of $\boldsymbol{\gamma}$ is obtained 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, and thus each γ_i can be updated by minimizing

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

Using the from of the normal density, the solution has the following form,

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

Note that $\sum_{j=1}^{m} p_{ij} \log\left(\sigma^2\right)$ does not depend on γ, and

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

It follows that

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

Proof of Lemma 1

The penalized least squares has the following form:

$$g(\gamma) = \frac{1}{2}(\xi - \gamma)^2 + \frac{\sigma^2}{\hat{\sigma}^2} P_\lambda(\gamma) \tag{5.1}$$

where $\xi = \{\sum_{j=1}^{m} p_{ij}(y_i - \mu_j)\}/(\sum_{j=1}^{m} p_{ij})$. For simplicity, we have omitted the subscripts in γ_i and ξ_i. The first derivative of $g(\gamma)$ with respect to γ is

$$g'(\gamma) = \gamma - \xi + sgn(\gamma)\frac{\sigma^2}{\hat{\sigma}^2}P'_\lambda(\gamma).$$

We first discuss some possible solutions of (5.1) in three cases.

Case 1: when $|\gamma| \leq \lambda$, the problem becomes an ℓ_1 penalized problem, and the solution, if feasible, is given by $\hat{\gamma}_1 = \mathrm{sgn}(\xi)\left(|\xi| - \sigma^2\lambda/\hat{\sigma}^2\right)_+$.

Case 2: when $\lambda < |\gamma| \leq a\lambda$, $g''(\gamma) = 1 - \sigma^2/\hat{\sigma}^2/(a-1)$. The second derivative is positive if $\sigma^2/\hat{\sigma}^2 < a - 1$. The solution, if feasible, is given by

$$\hat{\gamma}_2 = \frac{\frac{\hat{\sigma}^2}{\sigma^2}(a-1)\xi - \mathrm{sgn}(\xi)a\lambda}{\frac{\hat{\sigma}^2}{\sigma^2}(a-1) - 1}.$$

Case 3: when $|\gamma| > a\lambda$, $g''(\gamma) = 1$. The solution, if feasible, is given by $\hat{\gamma}_3 = \xi$.

The above three cases indicate that the solution depends on the value $\sigma^2/\hat{\sigma}^2$ and ξ. Since equation (5.1) is symmetric about ξ and $\Theta(-\xi;\lambda) = -\Theta(\xi;\lambda)$, we shall only discuss the case $\xi \geq 0$.

We now derive the solution $\hat{\gamma}$ in the following scenarios.

Scenario 1: $\sigma^2/\hat{\sigma}^2 < a - 1$.

1. When $\xi > a\lambda$, γ satisfies Case 3. Then $\hat{\gamma} = \hat{\gamma}_3$.

2. When $\lambda + \sigma^2\lambda/\hat{\sigma}^2 < \xi \leq a\lambda$, γ satisfies Case 2. Then $\hat{\gamma} = \hat{\gamma}_2$.

3. When $\xi \leq \lambda + \sigma^2\lambda/\hat{\sigma}^2$, γ satisfies Case 1. Then $\hat{\gamma} = \hat{\gamma}_1$.

Scenario 2: $a - 1 \leq \sigma^2/\hat{\sigma}^2 \leq a + 1$. Case 2 is not feasible.

1. When $\xi \leq a\lambda$, based on Case 1, $\hat{\gamma} = \hat{\gamma}_1$.

2. When $a\lambda \leq \xi \leq \lambda + \sigma^2\lambda/\hat{\sigma}^2$. As $|\hat{\gamma}_1| \leq \lambda$ and $|\hat{\gamma}_3| \geq a\lambda$, they are both possible solutions. Define $d = g(\hat{\gamma}_1) - g(\hat{\gamma}_3)$. Then $\hat{\gamma} = \hat{\gamma}_3$ if $d > 0$ and $\hat{\gamma} = \hat{\gamma}_1$ if $d < 0$. It can be verified

that $d > 0$ if $\xi > \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda$, and $d < 0$ if $\xi < \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda$. When $\xi = \frac{a+1+\frac{\sigma^2}{\hat{\sigma}^2}}{2}\lambda$, both $\hat{\gamma}_1$ and $\hat{\gamma}_3$ are minimizers; in (2.11) we have taken $\hat{\gamma} = \hat{\gamma}_1$.

3. When $\xi > \lambda + \sigma^2\lambda/\hat{\sigma}^2$, then $\xi > a\lambda$. Based on Case 3, $\hat{\gamma} = \xi$.

Scenario 3: $\sigma^2/\hat{\sigma}^2 > a + 1$. Case 2 is not feasible.

1. When $\xi > \sigma^2\lambda/\hat{\sigma}^2$, it is easy to see that $\hat{\gamma} = \xi$.

2. When $0 \leq \xi \leq \sigma^2\lambda/\hat{\sigma}^2$, $\hat{\gamma}_1 = 0$ and $d = g(\hat{\gamma}_1) - g(\hat{\gamma}_3) = \xi^2/2 - \sigma^2(a+1)\lambda^2/(2\hat{\sigma}^2)$. It follows that $d > 0$ if $\xi > \sqrt{\frac{\sigma^2(a+1)}{\hat{\sigma}^2}}\lambda$, $d < 0$ if $\xi < \sqrt{\frac{\sigma^2(a+1)}{\hat{\sigma}^2}}\lambda$. When $\xi = \sqrt{\frac{\sigma^2(a+1)}{\hat{\sigma}^2}}\lambda$, both $\hat{\gamma}_1 = 0$ and $\hat{\gamma}_3 = \xi$ are minimizers; in (2.12) we have taken $\hat{\gamma} = \hat{\gamma}_1 = 0$.

Combining the three senarios leads to the modified SCAD thresholding rule in Lemma 1. We note that in practice, as $\sigma^2/\hat{\sigma}^2$ is close to one, Senarios 2 and 3 are highly unlikely tp occur.

References

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

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

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

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

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.

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.

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

Dalalyan, A. S. and Keriven, R.(2012). Robust estimation for an inverse problem arising in multiview geometry. *Journal of Mathematical Imaging and Vision*, 43, 10-23.

Dalalyan, A. S. and Chen, Y. (2012). Fused sparsity and robust estimation for linear models with unknown variance. *In Advances in Neural Information Processing Systems*, 25: NIPS, 1268-1276

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.

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

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

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.

Fujisawa, H. and Eguchi S. (2005). Robust estimation in the normal mixture model. *Journal of Statistical Planning and Inference*, 1-23.

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

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

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

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

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

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

Khalili, A., Chen, J.H. (2007) Variable Selection in Finite Mixture of Regression Models. *Journal of the American Statistical Association*, 102, 1025–1038.

Lee, Y., MacEachern, S. N., and Jung, Y. (2012), Regularization of Case-Specific Parameters for Robustness and Efficiency. *Statistical Science*, 27, 350-372.

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.

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

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

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

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.

Nguyen, N. H. and Tran, T. D. (2013). Robust Lasso With Missing and Grossly Corrupted Observations. *IEEE Transactions on Information Theory*, 59, 2036-2058.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution *Statistics and Computing*, 10, 339-348.

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.

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

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.

She, Y. and Owen, A. (2011). Outlier Detection Using Nonconvex Penalized Regression. *Journal of the American Statistical Association*, 106(494), 626-639.

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.

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

Song, W., Yao, W., and Xing Y. (2014). Robust mixture regression model fitting by laplace distribution. *Computational Statistics and Data Analysis*, 71, 128-137.

Stadler, N., Buhlmann, P., and van de Geer, S. (2010). ℓ_1-penalization for mixture regression models. *Test*, 19(2), 209–256.

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

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

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.

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

Yao, W. (2012b). Bayesian mixture labeling and clustering. *Communications in Statistics - Theory and Methods*, 41, 403-421.

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

Yao, W., Wei, Y., and Yu, C. (2014). Robust Mixture Regression Using T-Distribution. *Computational Statistics and Data Analysis*, 71, 116-127.

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

Zhang, C.-H. (2010). Nearly unbiased Variable Selection Under Minimax Concave Penalty. *Ann. Statist.* 38 (2), 894–942.

Table 1: Simulation results for the case of equal variances with $n = 200$ and $p_{\mathcal{O}} = 5\%$.

	Hard	Hard$_{\mathcal{O}}$	SCAD	SCAD$_{\mathcal{O}}$	Soft	Soft$_{\mathcal{O}}$	TLE$_{0.05}$	TLE$_{0.10}$	MLE
M%	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.06	–
S%	0.27	0.02	0.99	0.03	0.42	0.03	1.04	3.34	–
JD%	100.00	100.00	100.00	100.00	100.00	100.00	99.44	99.44	–
Mis%	0.26	0.02	0.94	0.02	0.40	0.03	0.07	5.01	15.53
MeSE(π)	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.002
MSE(π)	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.030
MeSE(μ)	0.018	0.017	0.035	0.052	0.055	0.065	0.017	0.031	0.293
MSE(μ)	0.023	0.022	0.041	0.063	0.061	0.071	0.022	0.038	11.150
MeSE(σ)	0.009	0.010	0.067	0.191	0.176	0.231	0.008	0.064	0.952
MSE(σ)	0.016	0.016	0.088	0.191	0.198	0.242	0.012	0.071	25.478

Table 2: Simulation results for the case of equal variances with $n = 200$ and $p_{\mathcal{O}} = 10\%$.

	Hard	$\text{Hard}_{\mathcal{O}}$	SCAD	$\text{SCAD}_{\mathcal{O}}$	Soft	$\text{Soft}_{\mathcal{O}}$	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$	MLE
M%	0.00	0.00	0.00	0.00	12.11	0.00	24.53	0.00	–
S%	0.32	0.04	2.89	0.04	0.80	0.04	0.19	1.19	–
JD%	100.00	100.00	100.00	100.00	72.78	100.00	2.78	100.00	–
Mis%	0.29	0.05	2.61	0.03	1.93	0.04	5.94	0.09	22.28
MeSE(π)	0.001	0.001	0.001	0.001	0.001	0.001	0.004	0.001	0.003
MSE(π)	0.002	0.002	0.002	0.002	0.002	0.002	0.009	0.002	0.053
MeSE(μ)	0.020	0.019	0.061	0.183	0.171	0.212	0.840	0.019	0.918
MSE(μ)	0.023	0.024	0.066	0.209	0.230	0.231	1.093	0.023	14.125
MeSE(σ)	0.012	0.010	0.120	0.700	0.590	0.815	9.164	0.010	2.648
MSE(σ)	0.016	0.014	0.139	0.698	0.742	0.809	6.345	0.012	12.599

Table 3: Simulation results for the case of unequal variances with $n = 200$ and $p_{\mathcal{O}} = 5\%$.

	Hard	$\text{Hard}_{\mathcal{O}}$	SCAD	$\text{SCAD}_{\mathcal{O}}$	Soft	$\text{Soft}_{\mathcal{O}}$	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$	MLE
M%	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.06	–
S%	0.13	0.04	1.12	0.23	1.32	0.29	0.73	3.12	–
JD%	100.00	100.00	100.00	100.00	100.00	100.00	93.89	99.44	–
Mis%	0.51	0.44	1.48	1.35	2.24	1.87	3.88	6.22	44.82
MeSE(π)	0.001	0.001	0.001	0.003	0.004	0.006	0.001	0.001	0.024
MSE(π)	0.002	0.002	0.002	0.005	0.004	0.007	0.008	0.002	0.148
MeSE(μ)	0.038	0.042	0.051	0.081	0.063	0.087	0.042	0.056	77.214
MSE(μ)	0.048	0.051	0.068	0.115	0.080	0.134	3.060	0.073	141.426
MeSE(σ)	0.022	0.019	0.149	0.730	1.121	2.133	0.026	0.112	7.711
MSE(σ)	0.028	0.024	0.177	1.474	1.121	2.345	0.172	0.121	10.154

Table 4: Simulation results for the case of unequal variances with $n = 200$ and $p_{\mathcal{O}} = 10\%$.

	Hard	$\text{Hard}_{\mathcal{O}}$	SCAD	$\text{SCAD}_{\mathcal{O}}$	Soft	$\text{Soft}_{\mathcal{O}}$	$\text{TLE}_{0.05}$	$\text{TLE}_{0.10}$	MLE
M%	0.08	0.00	18.72	1.70	55.67	1.90	24.44	1.11	–
S%	0.10	0.07	2.49	0.83	0.20	0.94	0.06	0.77	–
JD%	98.33	100.00	66.67	68.67	5.56	65.33	1.11	83.89	–
Mis%	0.46	0.42	6.14	4.35	11.48	4.82	23.96	7.65	47.99
MeSE(π)	0.001	0.002	0.002	0.019	0.030	0.023	0.024	0.002	0.112
MSE(π)	0.002	0.003	0.008	0.019	0.032	0.025	0.066	0.049	0.168
MeSE(μ)	0.036	0.037	0.095	0.165	0.212	0.193	10.861	0.044	79.288
MSE(μ)	0.044	0.046	0.136	0.222	0.265	0.239	17.001	21.439	193.846
MeSE(σ)	0.029	0.024	0.613	7.306	11.553	7.734	11.059	0.028	13.128
MSE(σ)	0.035	0.033	3.416	6.088	11.396	7.482	10.261	0.917	16.203

Table 5: Comparison of average computation times in seconds. To make fair comparison, each reported time is the average computation time per each tuning parameter and simulated dataset.

Example	$p_{\mathcal{O}}$	Hard	SCAD	Soft	$\text{TLE}_{0.05}$	$\text{TLE}_{0.1}$	MLE
1	5%	0.039	0.041	0.042	0.041	0.042	0.016
1	10%	0.043	0.043	0.046	0.089	0.045	0.025
2	5%	0.081	0.128	0.166	0.083	0.076	0.008
2	10%	0.084	0.112	0.201	0.179	0.088	0.007

Table 6: Parameter estimation in Acidity data analysis.

	#outlier	π_1	π_2	π_3	μ_1	μ_2	μ_3	σ
MLE	0	0.589	0.138	0.273	4.320	5.682	6.504	0.365
	1	0.327	0.324	0.349	4.455	4.455	6.448	0.687
	3	0.503	0.478	0.019	5.105	5.105	12.00	1.028
Hard	0	0.588	0.157	0.255	4.333	5.720	6.545	0.336
	1	0.591	0.157	0.252	4.333	5.723	6.548	0.334
	3	0.597	0.157	0.246	4.333	5.729	6.553	0.331

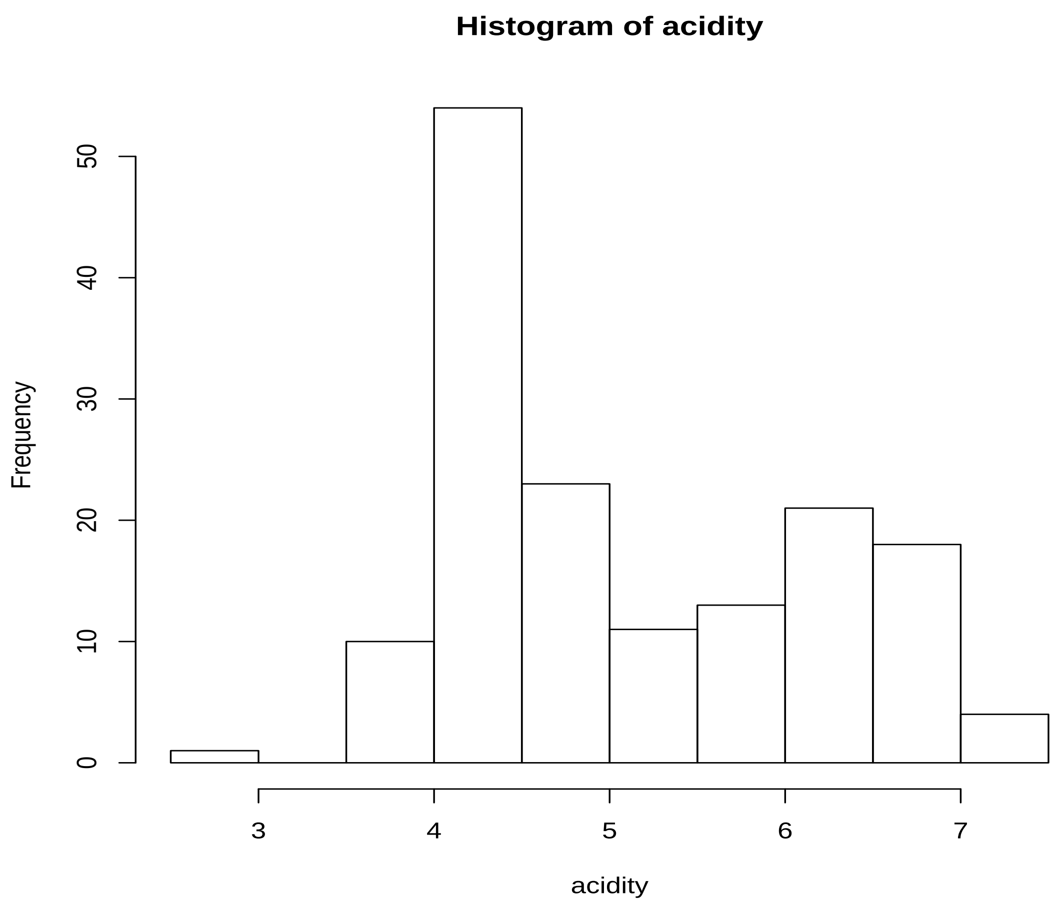

Figure 1: Histogram for Acidity data