The Canadian Journal of Statistics
Vol. 45, No. 1, 2017, Pages 77–94
La revue canadienne de statistique

A new method for robust mixture regression

Chun YU[1]*, Weixin YAO[2] and Kun CHEN[3]

[1] *School of Statistics, Jiangxi University of Finance and Economics, Nanchang 330013, P. R. China*
[2] *Department of Statistics, University of California, Riverside, CA 92521, U.S.A.*
[3] *Department of Statistics, University of Connecticut, Storrs, CT 06269, U.S.A.*

Key words and phrases: EM algorithm; mixture regression models; outlier detection; penalized likelihood.

MSC 2010: Primary 62F35; secondary 62J99

Abstract: Finite mixture regression models have been widely used for modelling mixed regression relationships arising from a clustered and thus heterogenous population. The classical normal mixture model, despite its simplicity and wide applicability, may fail in the presence of severe outliers. Using a sparse, case-specific, and scale-dependent mean-shift mixture model parameterization, we propose a robust mixture regression approach for simultaneously conducting outlier detection and robust parameter estimation. A penalized likelihood approach is adopted to induce sparsity among the mean-shift parameters so that the outliers are distinguished from the remainder of the data, and a generalized Expectation–Maximization (EM) algorithm is developed to perform stable and efficient computation. The proposed approach is shown to have strong connections with other robust methods including the trimmed likelihood method and M-estimation approaches. In contrast to several existing methods, the proposed methods show outstanding performance in our simulation studies. *The Canadian Journal of Statistics* 45: 77–94; 2017 © 2016 Statistical Society of Canada

Résumé: Les modèles de régression à mélange fini sont largement utilisés pour modéliser la relation de régression mixte qui émerge de données par grappes issues de populations hétérogènes. Malgré sa simplicité et sa large applicabilité, le modèle de mélange normal classique peut échouer en présence de valeurs fortement aberrantes. Les auteurs proposent un modèle de mélange à décalage des moyennes dont la paramétrisation clairsemée et spécifique au cas dépend de l'échelle. Ils proposent une méthode robuste de régression par mélange qui détecte les valeurs aberrantes et estime les paramètres simultanément. Ils adoptent une approche par vraisemblance pénalisée qui force les paramètres de décalage à être clairsemés afin que les valeurs aberrantes se démarquent des autres données. Ils développent également un algorithme d'espérance-maximisation (EM) qui permet des calculs stables et efficaces. Les auteurs montrent que leur méthode possède de forts liens avec d'autres approches robustes, notamment la vraisemblance tronquée et les M-estimateurs. Ils présentent des simulations dans le cadre desquelles leur approche offre une performance exceptionnelle contrairement à de nombreuses méthodes existantes. *La revue canadienne de statistique* 45: 77–94; 2017 © 2016 Société statistique du Canada

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, where p is the number of independent variables and $\mathbb{R}$ is the set of real numbers. However in many applications the assumption that the regression relationship is homogeneous across all the observations $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ does not hold. Rather the observations may form several distinct clusters indicating mixed relationships between the response and the predictors. Such heterogeneity can be more appropriately modelled by a "finite mixture regression model" consisting of, say, m homogeneous linear regression components. Specifically it is assumed that a regression

* *Author to whom correspondence may be addressed.*
E-mail: chunyu@jxufe.edu.cn

model holds for each of the m components, that is, when $(y, \mathbf{x})$ belongs to the jth component ($j = 1, 2, \ldots, m$), $y = \mathbf{x}^\top \boldsymbol{\beta}_j + \epsilon_j$, where $\boldsymbol{\beta}_j \in \mathbb{R}^p$ is a fixed and unknown coefficient vector, and $\epsilon_j \sim N(0, \sigma_j^2)$ with $\sigma_j^2 > 0$. (The intercept term can be included by setting the first element of each $\mathbf{x}$ vector as 1). The conditional density of y given $\mathbf{x}$, is

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

where $\phi(\cdot\,; \mu, \sigma^2)$ denotes the probability density function (pdf) of the normal distribution $N(\mu, \sigma^2)$, the π_j are mixing proportions, and $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \sigma_1; \ldots; \pi_m, \boldsymbol{\beta}_m, \sigma_m)$ represents all of the unknown parameters.

As it was first introduced by Goldfeld & Quandt (1973) the above mixture regression model has been widely used in business, marketing, social sciences, etc; see, for example, Böhning (1999), Jiang & Tanner (1999), Hennig (2000), McLachlan & Peel (2000), Wedel & Kamakura (2000), Skrondal & Rabe-Hesketh (2004), and Frühwirth-Schnatter (2006). Maximum likelihood estimation (MLE) is commonly carried out to infer $\boldsymbol{\theta}$ in (1), that is,

$$\widehat{\boldsymbol{\theta}}_{mle} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi\big(y_i; \mathbf{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2\big) \right\}.$$

The $\widehat{\boldsymbol{\theta}}_{mle}$ does not have an explicit form in general and it is usually obtained by the Expectation–Maximization (EM) (algorithm Dempster, Laird, & Rubin, 1977).

Although the normal mixture regression approach has greatly enriched the toolkit of regression analysis due to its simplicity it can be very sensitive to the presence of gross outliers, and failing to accommodate these may greatly jeopardize both model estimation and inference. Many robust methods have been developed for mixture regression models. Markatou (2000) and Shen, Yang, & Wang (2004) proposed to weight each data point in order to robustify the estimation procedure. Neykov et al. (2007) proposed to fit the mixture model using the trimmed likelihood method. Bai, Yao, & Boyer (2012) developed a modified EM algorithm by adopting a robust criterion in the M-step. Bashir & Carter (2012) extended the idea of the S-estimator to mixture regression. Yao, Wei, & Yu (2014) and Song, Yao, & Xing (2014) considered robust mixture regression using a t-distribution and a Laplace distribution, respectively. There has also been extensive work in linear clustering; see, for example, Hennig (2002, 2003), Mueller & Garlipp (2005), and García-Escudero et al. (2009, 2010).

Motivated by She & Owen (2011), Lee, MacEachern, & Jung (2012), and Yu, Chen, & Yao (2015) we propose a "robust mixture regression via mean shift penalization approach (RM^2)" to conduct simultaneous outlier detection and robust mixture model estimation. Our method generalizes the robust mixture model proposed by Yu, Chen, & Yao (2015) and can handle more general supervised learning tasks. Under the general framework of mixture regression several new challenges are present for adopting the regularization methods. For example maximizing the mixture likelihood is a nonconvex problem, which complicates the computation; as the mixture components may have unequal variances even the definition of an outlier becomes ambiguous, as the scale of the outlying effect of a data point may vary across different regression components.

Several prominent features make our proposed RM^2 approach attractive. First instead of using a robust estimation criterion or complex heavy-tailed distributions to robustify the mixture regression model our method is built upon a simple normal mixture regression model so as to facilitate computation and model interpretation. Second we adopt a sparse and scale-dependent mean-shift parameterization. Each observation is allowed to have potentially different outlying effects across different regression components, which is much more flexible than the setup considered by Yu, Chen, & Yao (2015). An efficient thresholding-embedded generalized EM algorithm is developed to solve the nonconvex penalized likelihood problem. Third we establish connections between RM^2 and some familiar robust methods including the trimmed likelihood and modified M-estimation methods. The results provide justification for the proposed methods and, at the same time, shed light on their robustness properties. These connections also apply to special cases of mixture modelling. Compared to existing robust methods RM^2 allows an efficient solution via the celebrated penalized regression approach, and different information criteria (such as AIC and BIC) can then be used to adaptively determine the proportion of outliers. Through extensive simulation studies RM^2 is demonstrated to be highly robust to both gross outliers and high leverage points.

2. ROBUST MIXTURE REGRESSION VIA MEAN-SHIFT PENALIZATION

2.1. Model Formulation

We consider the robust mixture regression model

$$f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \gamma_i) = \sum_{j=1}^{m} \pi_j \phi\big(y_i; \mathbf{x}_i^\top \boldsymbol{\beta}_j + \gamma_{ij}\sigma_j, \sigma_j^2\big) \quad \text{for } i = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{\theta} = (\pi_1, \boldsymbol{\beta}_1, \sigma_1, \ldots, \pi_m, \boldsymbol{\beta}_m, \sigma_m)^\top$. Here, for each observation, a mean-shift parameter, γ_{ij}, is added to its mean structure in each mixture component; we refer to (2) as a mean-shifted normal mixture model (RM^2). Define $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{im})^\top$ as the mean-shift vector for the ith observation for $i = 1, \ldots, n$, and let $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_n^\top)^\top$ collect all the mean-shift parameters.

Without any constraints on the mean-shift parameters in (2) the model is over-parameterized. The essence of (2) lies in the additional sparsity structures which are imposed on the parameters γ_{ij}: we assume many of the γ_{ij} are in fact zero, corresponding to the typical observations; and only a few γ_{ij} are nonzero, corresponding to the outliers. Promoting sparsity of γ_{ij} in estimation provides a direct way for identifying and accommodating outliers in the mixture regression model. Also note that the outlying effect is made case-specific, component-specific, and scale-dependent, that is, the outlying effect of the ith observation to the jth component is modelled by $\gamma_{ij}\sigma_j$, depending directly on the scale of the jth component. This setup is thus much more flexible than the structure considered by Yu, Chen, & Yao (2015) in the context of a mixture model. In our model each γ_{ij} parameter becomes scale-free and can be interpreted as the number of standard deviations shifted from the mixture regression structure.

The model framework developed in (2) inherits the simplicity of the normal mixture model, and it allows us to take advantage of celebrated penalized estimation approaches (Tibshirani, 1996; Fan & Li, 2001; Zou, 2006; Huang, Ma, & Zhang, 2008) for achieving robust estimation. For a comprehensive account of penalized regression and variable selection techniques, see, for example, Bühlmann & van de Geer (2009) and Huang, Breheny, & Ma (2012). For model (2), we propose a penalized likelihood approach for estimation,

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Gamma}}) = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\Gamma}} J_n(\boldsymbol{\theta}, \boldsymbol{\Gamma}), \tag{3}$$

where

$$J_n(\boldsymbol{\theta}, \boldsymbol{\Gamma}) = l_n(\boldsymbol{\theta}, \boldsymbol{\Gamma}) - \sum_{i=1}^{n} P_\lambda(\boldsymbol{\gamma}_i),$$

$$l_n(\boldsymbol{\theta}, \boldsymbol{\Gamma}) = \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{m} \pi_j \phi\big(y_i - \gamma_{ij}\sigma_j - \mathbf{x}_i^\top \boldsymbol{\beta}_j; 0, \sigma_j^2\big)\right\}$$

is the log-likelihood function, and $P_\lambda(\cdot)$ is a penalty function chosen to induce either element- or vector-wise sparsity of its argument which is a vector with a tuning parameter λ controlling the degrees of penalization. Similar to that of the traditional mixture model (1) the above penalized log-likelihood is also unbounded. That is the penalized log-likelihood goes to infinity when $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_j + \gamma_{ij}\sigma_j$, and $\sigma_j \to 0$ (Hathaway, 1985, 1986; Chen, Tan, & Zhang, 2008; and Yao, 2010). To circumvent this problem, following Hathaway (1985, 1986), we restrict $(\sigma_1, \ldots, \sigma_m) \in \Omega_\sigma$, with Ω_σ defined as

$$\Omega_\sigma = \{(\sigma_1, \ldots, \sigma_m) : \sigma_j > 0 \text{ for } 1 \leq j \leq m; \text{ and } \sigma_j/\sigma_k \geq \epsilon \\ \text{for } j \neq k \text{ and } 1 \leq j, k \leq m\}, \tag{4}$$

where ϵ is a very small positive value. In the examples that follow we set $\epsilon = 0.01$. Accordingly we define the parameter space of $\boldsymbol{\theta}$ as

$$\Omega = \{(\pi_j, \boldsymbol{\beta}_j, \sigma_j), j = 1, \ldots, m : 0 \leq \pi_j \leq 1, \sum_{j=1}^m \pi_j = 1, (\sigma_1, \ldots, \sigma_m) \in \Omega_\sigma\}.$$

There are many choices for the penalty function in (3). For inducing vector-wise sparsity we may consider the group lasso penalty of the form $P_\lambda(\boldsymbol{\gamma}_i) = \lambda\|\boldsymbol{\gamma}_i\|_2$ and the group ℓ_0 penalty $P_\lambda(\boldsymbol{\gamma}_i) = \lambda^2 I(\|\boldsymbol{\gamma}_i\|_2 \neq 0)/2$, where $\|\cdot\|_q$ denotes the ℓ_q norm for $q \geq 0$, and $I(\cdot)$ is the indicator function. These penalty functions penalize the ℓ_2 norm of each $\boldsymbol{\gamma}_i$ vector to promote the entire vector to be zero. Alternatively one may take $P_\lambda(\boldsymbol{\gamma}_i) = \sum_{j=1}^m P_\lambda(|\gamma_{ij}|)$, where $P_\lambda(|\gamma_{ij}|)$ is a penalty function so as to induce element-wise sparsity. Some examples are the ℓ_1 norm penalty (Donoho & Johnstone, 1994; Tibshirani, 1996)

$$P_\lambda(\boldsymbol{\gamma}_i) = \lambda \sum_{j=1}^m |\gamma_{ij}|, \tag{5}$$

and the ℓ_0 norm penalty (Antoniadis, 1997)

$$P_\lambda(\boldsymbol{\gamma}_i) = \frac{\lambda^2}{2} \sum_{j=1}^m I(\gamma_{ij} \neq 0). \tag{6}$$

Other common choices include the SCAD penalty (Fan & Li, 2001) and the MCP penalty (Zhang, 2010). In this article we mainly focus on using the element-wise penalization methods in RM^2.

2.2. Thresholding-Embedded EM Algorithm for Penalized Estimation

In classical mixture regression problems the EM algorithm is commonly used to maximize the likelihood, in which case the unobservable component labels are treated as missing data. We here propose an efficient thresholding-embedded EM algorithm to maximize the proposed penalized log-likelihood criterion. Consider

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Gamma}}) = \arg \max_{\boldsymbol{\theta}\in\Omega, \boldsymbol{\Gamma}} \left\{ \sum_{i=1}^n \log \left\{ \sum_{j=1}^m \pi_j \phi\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \gamma_{ij}\sigma_j; 0, \sigma_j^2\big) \right\} - \sum_{i=1}^n \sum_{j=1}^m P_\lambda(|\gamma_{ij}|) \right\},$$

where $P_\lambda(\cdot)$ is either the ℓ_1 penalty function (5) or the ℓ_0 penalty function (6). The proposed method can be readily applied to other penalty forms such as group lasso and group ℓ_0 penalties; see the Appendix for more details.

Let

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

Denote the complete data set by $\{(\mathbf{x}_i, \mathbf{z}_i, y_i) : i = 1, 2, \ldots, n\}$, where the component labels $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{im})$ are not observable. The penalized complete log-likelihood function is

$$J_n^c(\boldsymbol{\theta}, \boldsymbol{\Gamma}) = l_n^c(\boldsymbol{\theta}, \boldsymbol{\Gamma}) - \sum_{i=1}^{n} \sum_{j=1}^{m} P_\lambda(|\gamma_{ij}|), \tag{7}$$

where the complete log-likelihood is given by $l_n^c(\boldsymbol{\theta}, \boldsymbol{\Gamma}) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log\{\pi_j \phi(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \gamma_{ij}\sigma_j; 0, \sigma_j^2)\}$.

In the E-step, given the current estimates $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\Gamma}^{(k)}$ (where k denotes the iteration number), the conditional expectation of the penalized complete log-likelihood (7) is computed as follows:

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\Gamma} \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}) = & \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \left\{ \log \pi_j + \log \phi\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \gamma_{ij}\sigma_j; 0, \sigma_j^2\big) \right\} \\ & - \sum_{i=1}^{n} \sum_{j=1}^{m} P_\lambda(|\gamma_{ij}|) \end{aligned} \tag{8}$$

where

$$p_{ij}^{(k+1)} = \mathbb{E}(z_{ij} | y_i; \boldsymbol{\theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}) = \frac{\pi_j^{(k)} \phi\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j^{(k)} - \gamma_{ij}^{(k)} \sigma_j^{(k)}; 0, \sigma_j^{2(k)}\big)}{\sum_{j=1}^{m} \pi_j^{(k)} \phi\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j^{(k)} - \gamma_{ij}^{(k)} \sigma_j^{(k)}; 0, \sigma_j^{2(k)}\big)}. \tag{9}$$

We then maximize (8) with respect to $(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ in the M-step. Specifically in the M-step, $\boldsymbol{\theta}$, and $\boldsymbol{\Gamma}$ are alternatingly updated until convergence. For fixed $\boldsymbol{\Gamma}$ and σ_j, each $\boldsymbol{\beta}_j$ can be solved explicitly from a weighted least squares procedure. For fixed $\boldsymbol{\Gamma}$ and $\boldsymbol{\beta}_j$, as each σ_j appears in the mean structure, it no longer has an explicit solution, but due to low dimension, it can be readily solved by standard nonlinear optimization algorithms in which an augmented Lagrangian approach can be used for handling the nonlinear constraints; an implementation is provided in the R package nloptr (Conn, Gould, & Toint, 1991). Also when ignoring the ratio constraints in (4) the optimization problem of the σ_j becomes separable and each σ_j can be updated more easily; in practice the constrained estimation is performed only when the above simple solutions violate the ratio condition in (4).

For fixed $\boldsymbol{\theta}$, $\boldsymbol{\Gamma}$ is updated by maximizing

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

The problem is separable in each γ_{ij}, and after some algebra, it can be shown that γ_{ij} can be updated by minimizing

$$\frac{1}{2} \left(\gamma_{ij} - \frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j}{\sigma_j} \right)^2 + \frac{1}{p_{ij}^{(k+1)}} P_\lambda\left(|\gamma_{ij}|\right). \tag{10}$$

This one-dimensional problem admits an explicit solution. The solution of (10) when using the ℓ_1 penalty or the ℓ_0 penalty is given by a corresponding thresholding rule Θ_{soft} or Θ_{hard}, respectively:

$$\widehat{\gamma}_{ij} = \Theta_{soft}(\xi_{ij}; \lambda^*_{ij}) = \mathrm{sgn}(\xi_{ij})(|\xi_{ij}| - \lambda^*_{ij})_+, \tag{11}$$

$$\widehat{\gamma}_{ij} = \Theta_{hard}(\xi_{ij}; \lambda^*_{ij}) = \xi_{ij} I(|\xi_{ij}| > \lambda^*_{ij}), \tag{12}$$

where $\xi_{ij} = (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j)/\sigma_j$, $a_+ = \max(a, 0)$, λ^*_{ij} is taken as $\lambda / p_{ij}^{(k+1)}$ in Θ_{soft}, and λ^*_{ij} is set as $\lambda / \sqrt{p_{ij}^{(k+1)}}$ in Θ_{hard}. See the Appendix for details on handling group penalties on the $\boldsymbol{\gamma}_i$, such as the group ℓ_1 penalty and the group ℓ_0 penalty.

The proposed thresholding-embedded EM algorithm for any fixed tuning parameter λ is presented as follows:

Algorithm 1 Thresholding-Embedded EM algorithm for RM^2

Initialize $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\Gamma}^{(0)}$. Set $k \leftarrow 0$.
repeat
(1) E-step: Compute $Q(\boldsymbol{\theta}, \boldsymbol{\Gamma} \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{\Gamma}^{(k)})$ as in (8) and (9).
(2) M-step: Update $\pi_j^{(k+1)} = \left(\sum_{i=1}^n p_{ij}^{(k+1)} \right)/n$ and update the other parameters by maximizing $Q\left(\boldsymbol{\theta}, \boldsymbol{\Gamma} | \boldsymbol{\theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}\right)$, that is, start from $\left(\boldsymbol{\beta}^{(k)}, \sigma_j^{2(k)}, \boldsymbol{\Gamma}^{(k)}\right)$ and iterate the following steps until convergence to obtain $\left(\boldsymbol{\beta}^{(k+1)}, \sigma_j^{2(k+1)}, \boldsymbol{\Gamma}^{(k+1)}\right)$:

$$\text{(2.a)} \quad \boldsymbol{\beta}_j \leftarrow \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top p_{ij}^{(k+1)} \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i p_{ij}^{(k+1)} (y_i - \gamma_{ij}\sigma_j) \right), \; j = 1, \ldots, m,$$

$$\text{(2.b)} \quad (\sigma_1, \ldots, \sigma_m) \leftarrow \arg \max_{(\sigma_1, \ldots, \sigma_m) \in \Omega_\sigma} \sum_{i=1}^n \sum_{j=1}^m p_{ij}^{(k+1)} \log \phi(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \gamma_{ij}\sigma_j; 0, \sigma_j^2),$$

$$\text{(2.c)} \quad \gamma_{ij} \leftarrow \Theta(\xi_{ij}; \lambda^*_{ij}), \; i = 1, \ldots, n, \; j = 1, \ldots, m,$$

where Θ denotes one of the thresholding rules in (11–12) depending on the penalty form adopted.

$k \leftarrow k + 1$.
until convergence

The penalized log-likelihood does not decrease in any of the E- or M-step iterations, that is,

$$J_n\left(\widehat{\boldsymbol{\theta}}^{(k+1)}, \widehat{\boldsymbol{\Gamma}}^{(k+1)}\right) \geq J_n(\widehat{\boldsymbol{\theta}}^{(k)}, \widehat{\boldsymbol{\Gamma}}^{(k)})$$

for all $k \geq 0$. This property ensures the convergence of Algorithm 1.

The proposed algorithm can be readily modified to handle the special case of equal variances in model (1) with $\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$ for some $\sigma^2 > 0$. In the Algorithm σ_j shall be replaced by σ. The iterating steps stay the same with the exception of step (2.b) which becomes

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

The proposed EM algorithm is implemented for a 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 (e.g., Yi, Tan, & Li, 2015),

$$\mathrm{BIC}(\lambda) = -l(\lambda) + \log(n)\mathrm{df}(\lambda),$$

where $l(\lambda)$ is the mixture log-likelihood function evaluated at the solution of tuning parameter λ, and $df(\lambda)$ is the estimated model degrees of freedom. Following Zou (2006) we estimate the degrees of freedom using the sum of the number of nonzero elements in $\widehat{\boldsymbol{\Gamma}}$ and the number of component parameters in the mixture model. We fit the model for a certain number, say 100, of λ values which are equally spaced on the log scale in an interval $(\lambda_{\min}, \lambda_{\max})$, where $\lambda_{\min}$ is the smallest λ value for which roughly 50% of the entries in $\boldsymbol{\Gamma}$ are nonzero, and $\lambda_{\max}$ corresponds to the largest λ value for which $\boldsymbol{\Gamma}$ is estimated as a zero matrix. Other options for determining the optimal solution along the solution path are available as well (e.g., C_p, AIC, and GCV). For example one may discard a certain percentage of the observations as outliers if such prior knowledge is available. In the proposed model as the mean shift parameter of each observation can be interpreted as the number of standard deviations away from the observation to the component mean structure one may examine the magnitude of the mean-shift parameters in order to determine the number of outliers.

3. ROBUSTNESS OF RM^2

The outlier detection performance of RM^2 may depend on the choice of the penalty function. To understand the robustness properties of RM^2 we show, with a suitably chosen penalty function, that RM^2 has strong connections with some familiar robust methods including the trimmed likelihood and modified M-estimation methods. Our main results are summarized in Theorems 1 and 2 that follow. Their proofs are provided in the Appendix.

Theorem 1. *Consider RM^2 with a group ℓ_0 penalization, that is,*

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Gamma}}) = \arg \max_{\boldsymbol{\theta}\in\Omega,\boldsymbol{\Gamma}} \left[\sum_{i=1}^{n} \log\left\{ \sum_{j=1}^{m} \pi_j \phi\big(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j - \gamma_{ij}\sigma_j; 0, \sigma_j^2\big) \right\} - \frac{\lambda^2}{2} \sum_{i=1}^{n} I(\|\boldsymbol{\gamma}_i\|_2 \neq 0) \right]. \tag{13}$$

Denote $\widehat{\mathcal{S}} = \{i;\ \|\widehat{\boldsymbol{\gamma}}_i\| \neq 0\}$, and $h = n - |\widehat{\mathcal{S}}|$. Then

$$\begin{aligned}(\widehat{\boldsymbol{\theta}}, \widehat{\mathcal{S}}) = \arg \max_{\boldsymbol{\theta}\in\Omega,\mathcal{S}:|\mathcal{S}|=n-h} & \left[\sum_{i\in\mathcal{S}^c} \log\left\{ \sum_{j=1}^{m} \pi_j \phi\big(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j; 0, \sigma_j^2\big) \right\} \right. \\ & \left. + (n-h)\log\left\{ \sum_{j=1}^{m} \pi_j \phi\big(0; 0, \sigma_j^2\big) \right\} \right].\end{aligned}$$

In particular, when $\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$ and $\sigma^2 > 0$ is assumed known, the mean-shift penalization approach is equivalent to the trimmed likelihood method, that is,

$$(\widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\beta}}) = \arg \max_{\boldsymbol{\pi},\boldsymbol{\beta},\mathcal{S}:|\mathcal{S}|=n-h} \left[\sum_{i\in\mathcal{S}^c} \log\left\{ \sum_{j=1}^{m} \pi_j \phi\big(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_j; 0, \sigma^2\big) \right\} \right]. \tag{14}$$

In Theorem 1 we establish the connection between RM^2 and the trimmed likelihood method. In the special case of equal and known variances the two methods turn out to be entirely equivalent. This result partly explains the robustness property of RM^2 and shows that the trimmed likelihood estimation can be conveniently achieved with the proposed penalized likelihood approach.

In the classical EM algorithm for solving the normal mixture model the regression coefficients are updated based on weighted least squares. A natural idea with which to robustify the normal mixture model is hence to replace weighted least squares with some robust estimation criterion, such as using M-estimation. Bai, Yao, & Boyer (2012) pursued this idea and proposed a modified EM algorithm which was robust. Interestingly our RM^2 approach is closely connected to this modified EM algorithm. Consider RM^2 with an element-wise sparsity-inducing penalty,

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Gamma}}) = \arg \max_{\boldsymbol{\theta} \in \Omega, \boldsymbol{\Gamma}} \left\{ \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \gamma_{ij}\sigma_j; 0, \sigma_j^2\big) \right\} - \sum_{i=1}^{n}\sum_{j=1}^{m} P_\lambda(|\gamma_{ij}|) \right\}.$$

From the thresholding-embedded EM algorithm, define $\widehat{\mathbf{W}}_j = \operatorname{diag}(\widehat{p}_{1j}, \ldots, \widehat{p}_{nj})$, and $\widehat{\mathbf{w}}_j = (\lambda^*_{1j}, \ldots, \lambda^*_{nj})^\top$. Here for simplicity we omit the superscript (k) which denotes the iteration number. Then the parameter estimates satisfy

$$\widehat{\boldsymbol{\gamma}}_j = \Theta(\tfrac{1}{\widehat{\sigma}_j}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_j), \widehat{\mathbf{w}}_j), \text{ and } \widehat{\boldsymbol{\beta}}_j = (\mathbf{X}^\top \widehat{\mathbf{W}}_j \mathbf{X})^{-1}\mathbf{X}^\top \widehat{\mathbf{W}}_j(\mathbf{y} - \widehat{\sigma}_j\widehat{\boldsymbol{\gamma}}_j), \tag{15}$$

where Θ is defined element-wise.

Theorem 2. *Consider RM^2 with an element-wise sparsity-inducing penalization, and define $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Gamma}})$ as in (15). Then the parameter estimates satisfy*

$$\boldsymbol{X}^\top \widehat{\boldsymbol{W}}_j \psi\left(\frac{1}{\widehat{\sigma}_j}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j), \widehat{\boldsymbol{w}}_j\right) = 0, \qquad j = 1, \ldots, m, \tag{16}$$

where $\psi(t;\lambda) = t - \Theta(t;\lambda)$.

In Theorem 2 the score Equation (16) defines an M-estimator. Interestingly, as shown by She & Owen (2011), there is a general correspondence between the thresholding rules and the criteria used in M-estimation. It can be easily verified that for Θ_{soft}, the corresponding ψ function is the well-known Huber's ψ. Similarly Θ_{hard} corresponds to the Skipped Mean loss, and the SCAD thresholding corresponds to a special case of the Hampel loss. For robust estimation it is well understood that a redescending ψ function is preferable, which corresponds to the use of a nonconvex penalty in RM^2. In Bai, Yao, & Boyer (2012) the criterion parameter λ in $\psi(t;\lambda)$ is a prespecified value, and it stays the same for any input $(y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}_j)/\widehat{\sigma}_j$. In contrast the criterion parameter becomes adaptive in RM^2, with its overall magnitude determined by the penalization parameter, whose choice is data-driven and based on certain information criterion.

4. SIMULATION STUDY

4.1. Simulation Design

We consider two mixture regression models in which the observations are contaminated with additive outliers. We evaluate the finite sample performance of RM^2 and compare it with several existing methods. As we mainly focus on investigating outlier detection performance we have set $p = 2$ to keep the regression components relatively simple.

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

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

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

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

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

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

We consider two proportions of outliers, either 5 or 10%. The absolute value of any nonzero mean-shift parameter, $|\gamma_{ij}|$, is randomly generated from a uniform distribution between 11 and 13. Specifically in Model 1 we first generate $n = 400$ observations according to Model 1 with all γ_{ij} set to zero; when there are 5% (or 10%) outliers, 5 (or 10) observations from the first component are then replaced with $y_i = 1 - x_{i1} + x_{i2} - |\gamma_{i1}|\sigma + \epsilon_{i1}$ where $\sigma = 1, x_{i1} = 2$, and $x_{i2} = 2$, and 15 (or 30) observations from the second component are replaced with $y_i = 1 + 3x_{i1} + x_{i2} + |\gamma_{i2}|\sigma + \epsilon_{i2}$ where $\sigma = 1$, $x_{i1} = 2$, and $x_{i2} = 2$. In Model 2 the additive outliers are generated in the same fashion as in Model 1, except that the additive mean-shift terms become $-|\gamma_{i1}|\sigma_1$ or $|\gamma_{i2}|\sigma_2$, with $\sigma_1 = 1$ and $\sigma_2 = 2$. For each setting we repeat the simulation 200 times.

We compare our proposed RM^2 estimator when using ℓ_1 and ℓ_0 penalties, denoted as $\mathrm{RM}^2(\ell_1)$ and $\mathrm{RM}^2(\ell_0)$, respectively, to several existing robust regression estimators and the MLE of the classical normal mixture regression model. To examine the true potential of the RM^2 approaches, we report an "oracle" estimator for each penalty form, which is defined as the solution whose number of selected outliers is equal to (or is the smallest number greater than) the number of true outliers on the solution path. This is the penalized regression estimator we would have obtained if the true number of outliers was known a priori. All the estimators considered are listed below:

1. The MLE of the classical normal mixture regression model (MLE).
2. The trimmed likelihood estimator (TLE) proposed by Neykov et al. (2007), with the percentage of trimmed data set to either 5% ($\mathrm{TLE}_{0.05}$) or 10% ($\mathrm{TLE}_{0.10}$). We note that $\mathrm{TLE}_{0.05}$ ($\mathrm{TLE}_{0.10}$) can be regarded as the oracle TLE estimator when there are 5% (10%) outliers.
3. The robust estimator based on modified EM algorithm with bisquare loss (MEM-bisquare) proposed by Bai, Yao, & Boyer (2012).
4. The MLE of a mixture linear regression model that assumes a t-distributed error (Mixregt) as proposed by Yao, Wei, & Yu. (2014).
5. The RM^2 element-wise estimators using the ℓ_0 penalty ($\mathrm{RM}^2(\ell_0)$) and the ℓ_1 penalty ($\mathrm{RM}^2(\ell_1)$), and their oracle counterparts $\mathrm{RM}^2_O(\ell_0)$ and $\mathrm{RM}^2_O(\ell_1)$.

For fitting mixture models there are well known label switching issues (Celeux, Hurn, & Robert, 2000; Stephens, 2000; Yao & Lindsay, 2009; Yao, 2012). In our simulation study, as the truth is known, the labels are determined by minimizing the Euclidean distance from the true parameter values. To evaluate estimator performance we report the median squared errors (MeSE) and the mean squared errors (MSE) of the resulting parameter estimates. To evaluate the outlier detection performance we report three measures: the average proportion of masking (M), that is,

TABLE 1: Outlier detection results and MSE/MeSE of parameter estimates for Model 1

	5% outliers								
	M	S	JD	$\text{MSE}(\widehat{\boldsymbol{\pi}})_{(se)}$	$\text{MeSE}(\widehat{\boldsymbol{\pi}})$	$\text{MSE}(\widehat{\boldsymbol{\beta}})_{(se)}$	$\text{MeSE}(\widehat{\boldsymbol{\beta}})$	$\text{MSE}(\widehat{\sigma})_{(se)}$	$\text{MeSE}(\widehat{\sigma})$
$\text{RM}^2(\ell_0)$	0.000	0.001	1.000	$0.001_{(0.001)}$	0.001	$0.048_{(0.002)}$	0.039	$0.003_{(0.001)}$	0.001
$\text{RM}^2_O(\ell_0)$	0.000	0.000	1.000	$0.001_{(0.001)}$	0.001	$0.048_{(0.002)}$	0.038	$0.003_{(0.001)}$	0.001
$\text{RM}^2(\ell_1)$	0.000	0.006	1.000	$0.001_{(0.001)}$	0.001	$0.336_{(0.008)}$	0.327	$0.216_{(0.004)}$	0.223
$\text{RM}^2_O(\ell_1)$	0.000	0.003	1.000	$0.001_{(0.001)}$	0.001	$0.184_{(0.003)}$	0.143	$0.662_{(0.002)}$	0.666
$\text{TLE}_{0.05}$	0.000	0.007	1.000	$0.002_{(0.001)}$	0.001	$0.047_{(0.002)}$	0.037	$0.002_{(0.001)}$	0.001
$\text{TLE}_{0.10}$	0.000	0.003	1.000	$0.002_{(0.001)}$	0.001	$0.085_{(0.004)}$	0.067	$0.025_{(0.001)}$	0.023
MEM-bisquare	0.000	0.005	1.000	$0.002_{(0.001)}$	0.001	$0.050_{(0.002)}$	0.041	$0.007_{(0.001)}$	0.004
Mixregt	0.000	0.078	1.000	$0.003_{(0.001)}$	0.002	$0.090_{(0.004)}$	0.080	$0.123_{(0.002)}$	0.121
MLE	–	–	–	$0.470_{(0.002)}$	0.680	$17.20_{(0.658)}$	20.33	$2.912_{(0.023)}$	2.920
	10% outliers								
	M	S	JD	$\text{MSE}(\widehat{\boldsymbol{\pi}})_{(se)}$	$\text{MeSE}(\widehat{\boldsymbol{\pi}})$	$\text{MSE}(\widehat{\boldsymbol{\beta}})_{(se)}$	$\text{MeSE}(\widehat{\boldsymbol{\beta}})$	$\text{MSE}(\widehat{\sigma})_{(se)}$	$\text{MeSE}(\widehat{\sigma})$
$\text{RM}^2(\ell_0)$	0.000	0.001	1.000	$0.001_{(0.001)}$	0.001	$0.055_{(0.003)}$	0.044	$0.007_{(0.001)}$	0.005
$\text{RM}^2_O(\ell_0)$	0.000	0.000	1.000	$0.001_{(0.001)}$	0.001	$0.054_{(0.003)}$	0.044	$0.006_{(0.001)}$	0.005
$\text{RM}^2(\ell_1)$	0.615	0.005	0.335	$0.028_{(0.007)}$	0.001	$6.505_{(0.324)}$	7.702	$0.778_{(0.011)}$	0.695
$\text{RM}^2_0(\ell_1)$	0.007	0.001	0.750	$0.001_{(0.001)}$	0.001	$4.973_{(0.056)}$	4.894	$0.581_{(0.002)}$	0.569
$\text{TLE}_{0.05}$	0.749	0.050	0.000	$0.274_{(0.018)}$	0.046	$50.94_{(1.100)}$	49.08	$0.298_{(0.009)}$	0.275
$\text{TLE}_{0.10}$	0.000	0.007	1.000	$0.002_{(0.001)}$	0.001	$0.057_{(0.003)}$	0.046	$0.002_{(0.001)}$	0.001
MEM-bisquare	0.639	0.061	0.145	$0.279_{(0.020)}$	0.043	$39.81_{(1.306)}$	45.74	$0.143_{(0.008)}$	0.120
Mixregt	0.313	0.096	0.555	$0.212_{(0.019)}$	0.005	$18.05_{(1.465)}$	0.174	$0.058_{(0.002)}$	0.056
MLE	–	–	–	$0.075_{(0.010)}$	0.014	$11.55_{(0.262)}$	10.09	$4.462_{(0.027)}$	4.459

The standard errors (se) of the MSE values are reported in the subscripts.

the fraction of undetected outliers; the average proportion of swamping (S), that is, the fraction of good points labeled as outliers; and the joint detection rate (JD), that is, the proportion of simulations with 0 masking.

4.2. Simulation Results

The simulation results for Model 1 (equal variances case) are reported in Table 1. It is apparent that MLE fails miserably in the presence of severe outliers, so in the following we focus on discussing only the robust methods. In the case of 5% outliers all methods (except MLE) perform well in detecting outliers. For parameter estimation, $\text{RM}^2(\ell_1)$ and $\text{RM}^2_O(\ell_1)$ perform much worse than other methods. In the case of 10% outliers, $\text{RM}^2(\ell_0)$ and $\text{TLE}_{0.10}$ work well, whereas $\text{RM}^2(\ell_1)$, $\text{TLE}_{0.05}$, MEM-bisquare, and Mixregt have much lower joint outlier detection rates and hence larger MeSE or MSE. The non-robustness of $\text{RM}^2(\ell_1)$ is as expected, as it corresponds to using Huber's loss which is known to suffer from masking effects. This can also be seen from the penalized regression point of view. As the ℓ_1 regularization induces sparsity it also results in a heavy shrinkage effect. Consequently the method tends to accommodate the outliers in the model, leading to biased and severely distorted estimation results. In contrast the ℓ_0 penalization does

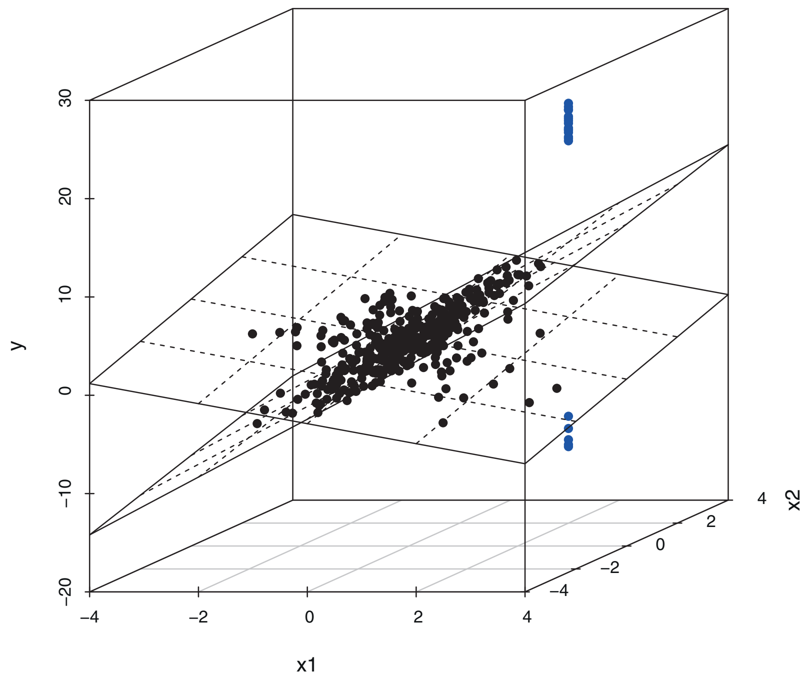

FIGURE 1: 3D scatter plot for one data set simulated from Model 1. Two regression planes estimated by $\mathrm{RM}^2(\ell_0)$ for the two different components are displayed and the 5% outliers are marked in blue.

not offer any shrinkage, so it is much harder for an outlier to be accommodated. Our results are consistent with the finding of She and Owen (2011) in the context of linear regression.

Figure 1 shows a 3-dimensional scatter plot of one typical data set simulated from Model 1, with two regression planes estimated by $\mathrm{RM}^2(\ell_0)$. The outliers are marked in blue. It can be seen that the regression planes fit the bulk of the good observations from the two components quite well and that the estimates are not influenced by the outliers.

Table 2 reports the simulation results for Model 2 (unequal variances case). The conclusions are similar to those for the equal variances case. Briefly, with 5% outliers, all methods (except MLE) have high joint outlier detection rates. When there are 10% outliers $\mathrm{RM}^2(\ell_0)$ and the trimmed likelihood methods continue to perform best. In either case the estimation accuracy of $\mathrm{RM}^2(\ell_1)$ is much lower than that of $\mathrm{RM}^2(\ell_0)$. Also a 3-dimensional scatter plot of one typical simulated data set is shown in Figure 2 and clearly demonstrates that the estimates of $\mathrm{RM}^2(\ell_0)$ are not influenced by the outliers.

In summary, $\mathrm{TLE}_{0.10}$ yields good results in terms of outliers detection in all cases but has larger MSE for the 5% outliers case. $\mathrm{TLE}_{0.05}$ fails to work in the case of 10% outliers.

$\mathrm{RM}^2(\ell_0)$ is comparable to oracle TLE and $\mathrm{RM}^2_O(\ell_0)$ in terms of both outlier detection and MeSE in most cases except for the 10% outlier case with small $|\gamma|$. $\mathrm{RM}^2(\ell_1)$ with a large $|\gamma|$ is comparable to $\mathrm{RM}^2(\ell_0)$ and TLE in terms of outlier detection but fails to work with a small $|\gamma|$. As expected MLE is sensitive to outliers.

5. TONE PERCEPTION DATA ANALYSIS

We apply the proposed robust approach 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 by the same musician.

TABLE 2: Outlier detection results and MSE/MeSE of parameter estimates for Model 2

	5% outliers								
	M	S	JD	MSE($\widehat{\boldsymbol{\pi}}$)$_{(se)}$	MeSE($\widehat{\boldsymbol{\pi}}$)	MSE($\widehat{\boldsymbol{\beta}}$)$_{(se)}$	MeSE($\widehat{\boldsymbol{\beta}}$)	MSE($\widehat{\boldsymbol{\sigma}}$)$_{(se)}$	MeSE($\widehat{\boldsymbol{\sigma}}$)
RM$^2(\ell_0)$	0.001	0.001	0.995	$0.004_{(0.001)}$	0.002	$0.111_{(0.009)}$	0.088	$0.038_{(0.004)}$	0.024
RM$^2_O(\ell_0)$	0.001	0.001	0.995	$0.004_{(0.001)}$	0.002	$0.111_{(0.012)}$	0.087	$0.059_{(0.077)}$	0.024
RM$^2(\ell_1)$	0.000	0.005	1.000	$0.001_{(0.001)}$	0.001	$0.365_{(0.010)}$	0.344	$1.448_{(0.025)}$	1.436
RM$^2_O(\ell_1)$	0.000	0.003	1.000	$0.001_{(0.001)}$	0.001	$0.740_{(0.041)}$	0.699	$1.799_{(0.108)}$	1.706
TLE$_{0.05}$	0.004	0.008	0.915	$0.077_{(0.002)}$	0.002	$9.160_{(2.535)}$	0.096	$0.502_{(0.011)}$	0.023
TLE$_{0.10}$	0.008	0.032	0.845	$0.259_{(0.003)}$	0.007	$1.528_{(0.155)}$	0.219	$1.756_{(0.115)}$	0.655
MEM-bisquare	0.062	0.006	0.915	$0.087_{(0.001)}$	0.004	$9.835_{(2.242)}$	0.115	$0.637_{(0.091)}$	0.102
Mixregt	0.000	0.078	1.000	$0.008_{(0.001)}$	0.003	$0.421_{(0.154)}$	0.182	$0.683_{(0.010)}$	0.655
MLE	–	–	–	$0.761_{(0.001)}$	0.763	$43.20_{(0.715)}$	41.84	$186.2_{(1.014)}$	186.5
	10% outliers								
	M	S	JD	MSE($\widehat{\boldsymbol{\pi}}$)$_{(se)}$	MeSE($\widehat{\boldsymbol{\pi}}$)	MSE($\widehat{\boldsymbol{\beta}}$)$_{(se)}$	MeSE($\widehat{\boldsymbol{\beta}}$)	MSE($\widehat{\boldsymbol{\sigma}}$)$_{(se)}$	MeSE($\widehat{\boldsymbol{\sigma}}$)
RM$^2(\ell_0)$	0.000	0.000	1.000	$0.006_{(0.001)}$	0.005	$0.124_{(0.005)}$	0.106	$0.057_{(0.003)}$	0.052
RM$^2_O(\ell_0)$	0.000	0.000	1.000	$0.006_{(0.001)}$	0.005	$0.121_{(0.005)}$	0.106	$0.056_{(0.010)}$	0.043
RM$^2(\ell_1)$	0.030	0.012	0.955	$0.002_{(0.001)}$	0.001	$1.607_{(0.125)}$	1.255	$4.706_{(0.428)}$	3.489
RM$^2_O(\ell_1)$	0.001	0.001	0.970	$0.001_{(0.001)}$	0.001	$1.603_{(0.075)}$	1.546	$1.959_{(0.170)}$	1.959
TLE$_{0.05}$	0.656	0.018	0.000	$0.654_{(0.015)}$	0.679	$98.20_{(2.491)}$	90.68	$1.960_{(0.097)}$	1.970
TLE$_{0.10}$	0.003	0.008	0.900	$0.063_{(0.009)}$	0.002	$10.37_{(1.956)}$	0.125	$0.403_{(0.038)}$	0.018
MEM-bisquare	0.722	0.012	0.010	$0.622_{(0.009)}$	0.652	$94.93_{(1.956)}$	86.53	$2.397_{(0.062)}$	2.291
Mixregt	0.461	0.097	0.200	$0.516_{(0.018)}$	0.638	$70.46_{(2.632)}$	81.49	$0.968_{(0.028)}$	0.998
MLE	–	–	–	$0.593_{(0.010)}$	0.593	$40.89_{(0.743)}$	38.98	$188.1_{(2.879)}$	195.2

The standard errors (se) of the MSE values are reported in the subscripts.

The overtones were determined using a stretching ratio, which is the ratio between an 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 RM$^2(\ell_0)$ estimator and the traditional MLE after adding ten outliers $(1.5, a)$, where $a = 3 + 0.1i$, and $i = 1, 2, 3, 4, 5$ and $(3, b)$, where $b = 1 + 0.1i$, and $i = 1, 2, 3, 4, 5$ into the original data set. Table 3 reports the parameter estimates. For the original data which contained no outliers, the proposed RM$^2(\ell_0)$ estimator yields similar parameter estimates to that of the traditional MLE. This result shows that our proposed RM$^2(\ell_0)$ method performs as well as the traditional MLE. If there are outliers in the data the proposed RM$^2(\ell_0)$ estimator is not influenced by them and yields similar parameter estimates to those for the case of no outliers. However the MLE yields nonsensical parameter estimates.

6. DISCUSSION

We have proposed a robust mixture regression approach based on a mean-shift normal mixture model parameterization, generalizing the work of She & Owen (2011), Lee, MacEachern, &

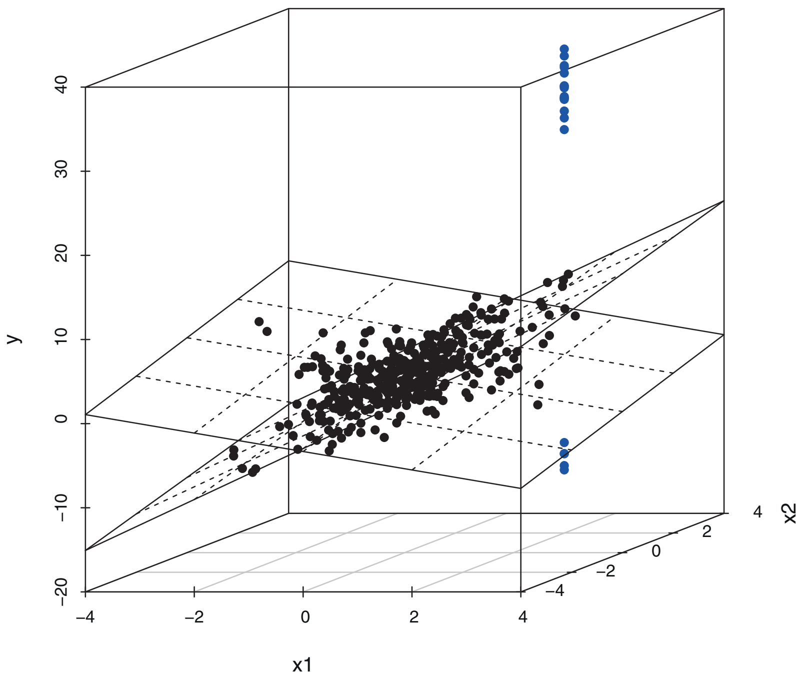

FIGURE 2: 3D scatter plot for one data set simulated from Model 2. Two regression planes estimated by RM^2 (ℓ_0) for the two different components are displayed and the 5% outliers are marked in blue.

Jung (2012), and Yu, Chen, & Yao (2015). The method is shown to have strong connections with several well-known robust methods. The proposed RM^2 method with the ℓ_0 penalty has comparable performance to its oracle counterpart and the oracle Trimmed Likelihood Estimator (TLE).

There are several directions for future research. The oracle RM^2 estimators may have better performance than the BIC-tuned estimators in some cases; therefore we can further improve the performance of RM^2 by improving tuning parameter selection. García-Escudero et al. (2010) showed that the traditional definition of breakdown point is not an appropriate measure with which to quantify the robustness of mixture regression procedures, as the robustness of these procedures is not only data dependent but also cluster dependent. It is thus interesting to consider the construction and investigation of other robustness measures for a mixture model setup. Although we do not discuss the selection of the number of cluster components in this article it remains a pressing issue in many mixture modelling problems. The proposed RM^2 approach can be further extended to conduct simultaneous variable selection and outlier detection in mixture regression.

TABLE 3: Parameter estimation for the tone perception data

		π_1	π_2	β_{01}	β_{11}	β_{02}	β_{12}	σ
MLE	No outliers	0.326	0.674	−0.038	1.008	1.893	0.056	0.084
	10 outliers	0.082	0.918	5.250	−1.311	1.298	0.359	0.224
Hard	No outliers	0.311	0.689	0.049	0.947	1.904	0.079	0.100
	10 outliers	0.276	0.724	0.051	0.950	1.900	0.071	0.100

Finally our proposed robust regression approach is based on penalized estimation, for which statistical inference is an active research area; see, for example, Berk et al. (2013), Tibshirani et al. (2014), and Lee et al. (2016). It is pressing to investigate the inference problem in the context of robust estimation and outlier detection.

APPENDIX

Handling Group Penalties

The proposed EM algorithm can be readily modified to handle the group lasso penalty and the group ℓ_0 penalty on the $\boldsymbol{\gamma}_i$. The only change is in the way of updating $\boldsymbol{\Gamma}$ in the M step when $\boldsymbol{\theta}$ is fixed.

For the group lasso penalty, $P_\lambda(\boldsymbol{\gamma}_i) = \lambda \|\boldsymbol{\gamma}_i\|_2$, $\boldsymbol{\Gamma}$ is updated by maximizing

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \gamma_{ij}\sigma_j; 0, \sigma_j^2\right) - \lambda \sum_{i=1}^{n} \|\boldsymbol{\gamma}_i\|_2.$$

The problem is separable in each $\boldsymbol{\gamma}_i$. After some algebra the problem for each $\boldsymbol{\gamma}_i$ has exactly the same form as the problem considered in Qin, Scheinberg, & Goldfarb (2013),

$$\widehat{\boldsymbol{\gamma}}_i = \arg\min_{\boldsymbol{\gamma}_i} \frac{1}{2} \boldsymbol{\gamma}_i^\top \mathbf{W}_i \boldsymbol{\gamma}_i - \mathbf{a}_i^\top \boldsymbol{\gamma}_i + \lambda \|\boldsymbol{\gamma}_i\|_2, \tag{17}$$

where $\mathbf{W}_i$ is an $m \times m$ diagonal matrix with diagonal elements $\left\{p_{ij}^{(k+1)}, j = 1, \ldots, m\right\}$, $r_{ij} = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j$, and $\mathbf{a}_i = \left(\frac{p_{i1}^{(k+1)} r_{i1}}{\sigma_1}, \ldots, \frac{p_{im}^{(k+1)} r_{im}}{\sigma_m}\right)^\top$. The detailed algorithm is given in Qin, Scheinberg, & Goldfarb (2013). The solution of (17) can be expressed as

$$\widehat{\boldsymbol{\gamma}}_i = \begin{cases} \mathbf{0} & \text{if } \|_i\|_2 \le \lambda; \\ \Delta_i(\Delta_i \mathbf{W}_i + \lambda \mathbf{I})^{-1} \mathbf{a}_i & \text{if } \|\mathbf{a}_i\|_2 > \lambda, \end{cases}$$

in which Δ_i is the root of

$$\phi(\Delta_i) = 1 - \frac{1}{\|f(\Delta_i)\|_2},$$

where

$$\|f(\Delta_i)\|_2^2 = \sum_{j=1}^{m} \frac{\left(p_{ij}^{(k+1)} r_{ij}/\sigma_j\right)^2}{\left(p_{ij}^{(k+1)} \Delta_i + \lambda\right)^2}.$$

For group ℓ_0 penalty $\boldsymbol{\Gamma}$ is updated by maximizing

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \gamma_{ij}\sigma_j; 0, \sigma_j^2\right) - \frac{\lambda^2}{2} \sum_{i=1}^{n} I(\|\boldsymbol{\gamma}_i\|_2 \neq 0).$$

The problem is separable in each $\boldsymbol{\gamma}_i$, that is,

$$\widehat{\boldsymbol{\gamma}}_i = \arg\max_{\boldsymbol{\gamma}_i} \left\{ \sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j - \gamma_{ij}\sigma_j; 0, \sigma_j^2\right) - \frac{\lambda^2}{2} I(\|\boldsymbol{\gamma}_i\|_2 \neq 0) \right\}.$$

This has a closed-form solution. Define $\widetilde{\boldsymbol{\gamma}}_i$ such that $\widetilde{\gamma}_{ij} = (y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j)/\sigma_j$ for $j = 1, \ldots, m$. Then

$$\widehat{\boldsymbol{\gamma}}_i = \begin{cases} \mathbf{0} & \text{if } \sum_{j=1}^m p_{ij}^{(k+1)} \log \phi\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j; 0, \sigma_j^2\big) \geq \sum_{j=1}^m p_{ij}^{(k+1)} \log \phi\big(0; 0, \sigma_j^2\big) - \frac{\lambda^2}{2}; \\ \widetilde{\boldsymbol{\gamma}}_i & \text{if } \sum_{j=1}^m p_{ij}^{(k+1)} \log \phi\big(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j; 0, \sigma_j^2\big) < \sum_{j=1}^m p_{ij}^{(k+1)} \log \phi\big(0; 0, \sigma_j^2\big) - \frac{\lambda^2}{2}. \end{cases}$$

Proof of Theorem 1

Recall $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Gamma}})$ is the maximizer of the penalized log-likelihood problem (13). Then we have

$$\widehat{\boldsymbol{\Gamma}} = \arg\max_{\boldsymbol{\Gamma}} \left[\sum_{i=1}^n \log \left\{ \sum_{j=1}^m \widehat{\pi}_j \phi \left(y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j - \gamma_{ij}\widehat{\sigma}_j; 0, \widehat{\sigma}_j^2 \right) \right\} - \frac{\lambda^2}{2} \sum_{i=1}^n I(\|\boldsymbol{\gamma}_i\|_2 \neq 0) \right]. \tag{18}$$

The problem is separable in each $\boldsymbol{\gamma}_i$, that is,

$$\widehat{\boldsymbol{\gamma}}_i = \arg\max_{\boldsymbol{\gamma}_i} \left[\log \left\{ \sum_{j=1}^m \widehat{\pi}_j \phi \left(y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j - \gamma_{ij}\widehat{\sigma}_j; 0, \widehat{\sigma}_j^2 \right) \right\} - \frac{\lambda^2}{2} I(\|\boldsymbol{\gamma}_i\|_2 \neq 0) \right]. \tag{19}$$

If $\widehat{\boldsymbol{\gamma}}_i = \mathbf{0}$ (19) becomes

$$\log \left\{ \sum_{j=1}^m \widehat{\pi}_j \phi\big(y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j; 0, \widehat{\sigma}_j^2 \big) \right\},$$

and if $\widehat{\boldsymbol{\gamma}}_i \neq \mathbf{0}$, it must be true that $\widehat{\gamma}_{ij} = (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j)/\widehat{\sigma}_j$, $j = 1, \ldots, m$, and (19) then becomes

$$\log \left\{ \sum_{j=1}^m \widehat{\pi}_j \phi\big(0; 0, \widehat{\sigma}_j^2\big) \right\} - \frac{\lambda^2}{2}.$$

It then follows that the maximum of the penalized log-likelihood (13) is

$$\sum_{i \in \widehat{\mathcal{S}}^c} \log \left\{ \sum_{j=1}^m \widehat{\pi}_j \phi\big(y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j; 0, \widehat{\sigma}_j^2 \big) \right\} + \sum_{i \in \widehat{\mathcal{S}}} \log \left\{ \sum_{j=1}^m \widehat{\pi}_j \phi\big(0; 0, \widehat{\sigma}_j^2\big) \right\} - \frac{\lambda^2}{2}(n - h). \tag{20}$$

For a given tuning parameter λ the number of nonzero $\widehat{\boldsymbol{\gamma}}_i$ vectors is determined and hence h is a constant. This proves the first part of the theorem.

When $\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$ and $\sigma^2 > 0$ is assumed known the second term in (20) becomes $\sum_{i \in \widehat{\mathcal{S}}} \log\{1/(\sqrt{2\pi}\sigma)\}$ and hence is a constant. It follows that maximizing (13) is equivalent to solving (14), in which $\mathcal{S}$ is an index set with the same cardinality as $\widehat{S}$. We recognize that (14) is exactly a trimmed likelihood problem. This completes the proof. ■

Proof of Theorem 2

Consider element-wise penalization in (15). Based on the thresholding-embedded EM algorithm we write

$$\widehat{\boldsymbol{\Gamma}} = \arg\max_{\boldsymbol{\Gamma}} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} \widehat{p}_{ij} \log \phi\big(y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j - \gamma_{ij}\widehat{\sigma}_j; 0, \widehat{\sigma}_j^2\big) - \sum_{i=1}^{n} \sum_{j=1}^{m} P_\lambda(|\gamma_{ij}|) \right\}.$$

The above problem is separable in each γ_{ij},

$$\begin{aligned}\widehat{\gamma}_{ij} &= \arg\max_{\gamma_{ij}} \left\{ \widehat{p}_{ij} \log \phi\big(y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j - \gamma_{ij}\widehat{\sigma}_j; 0, \widehat{\sigma}_j^2\big) - P_\lambda(|\gamma_{ij}|) \right\} \\ &= \arg\min_{\gamma_{ij}} \frac{1}{2} \left(\gamma_{ij} - \frac{y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j}{\widehat{\sigma}_j} \right)^2 + \frac{1}{\widehat{p}_{ij}} P_\lambda(|\gamma_{ij}|),\end{aligned}$$

It can be easily shown that

$$\widehat{\gamma}_{ij} = \Theta \left(\frac{y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_j}{\widehat{\sigma}_j}, \lambda_{ij}^* \right),$$

where the correspondence of $(P_\lambda(\cdot), \lambda_{ij}^*, \Theta)$ is discussed in Section 2.2. For example using the ℓ_1 penalty leads to $\Theta = \Theta_{soft}$ and $\lambda_{ij}^* = \lambda / p_{ij}^{(k+1)}$; using the ℓ_0 penalty leads to $\Theta = \Theta_{hard}$ and $\lambda_{ij}^* = \lambda / \sqrt{p_{ij}^{(k+1)}}$.

Define $\widehat{\mathbf{W}}_j = \operatorname{diag}(\widehat{p}_{1j}, \ldots, \widehat{p}_{nj})$, and $\widehat{\mathbf{w}}_j = (\lambda_{1j}^*, \ldots, \lambda_{nj}^*)^\top$. Then we can write

$$\widehat{\boldsymbol{\gamma}}_j = \Theta \left(\frac{1}{\widehat{\sigma}_j} \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_j \right), \widehat{\mathbf{w}}_j \right), \text{ and } \widehat{\boldsymbol{\beta}}_j = (\mathbf{X}^\top \widehat{\mathbf{W}}_j \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{W}}_j (\mathbf{y} - \widehat{\sigma}_j \widehat{\boldsymbol{\gamma}}_j).$$

Now, consider any $\psi(t; \lambda)$ function satisfying $\Theta(t; \lambda) + \psi(t; \lambda) = t$ for any t. We have

$$\begin{aligned}\mathbf{X}^\top \widehat{\mathbf{W}}_j \psi \left(\frac{1}{\widehat{\sigma}_j} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_j), \widehat{\mathbf{w}}_j \right) &= \mathbf{X}^\top \widehat{\mathbf{W}}_j \left\{ \frac{1}{\widehat{\sigma}_j} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_j) - \Theta \left(\frac{1}{\widehat{\sigma}_j} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_j), \widehat{\mathbf{w}}_j \right) \right\} \\ &= \mathbf{X}^\top \widehat{\mathbf{W}}_j \left\{ \frac{1}{\widehat{\sigma}_j} \mathbf{y} - \frac{1}{\widehat{\sigma}_j} \mathbf{X} (\mathbf{X}^\top \widehat{\mathbf{W}}_j \mathbf{X})^{-1} \mathbf{X}^\top \widehat{\mathbf{W}}_j (\mathbf{y} - \widehat{\sigma}_j \widehat{\boldsymbol{\gamma}}_j) - \widehat{\boldsymbol{\gamma}}_j \right\} \\ &= \frac{1}{\widehat{\sigma}_j} \mathbf{X}^\top \widehat{\mathbf{W}}_j \mathbf{y} - \frac{1}{\widehat{\sigma}_j} \mathbf{X}^\top \widehat{\mathbf{W}}_j (\mathbf{y} - \widehat{\sigma}_j \widehat{\boldsymbol{\gamma}}_j) - \mathbf{X}^\top \widehat{\mathbf{W}}_j \widehat{\boldsymbol{\gamma}}_j \\ &= 0.\end{aligned}$$

It follows that solving $\boldsymbol{\beta}_j$ is equivalent to solving the score equation in (16), which completes the proof. ■

ACKNOWLEDGEMENTS

We thank the two referees and the Associate Editor, whose comments and suggestions have helped us to improve the article significantly. Yu's research is supported by National Natural Science Foundation of China grant 11661038. Yao's research is partially supported by the U.S. National Science Foundation (DMS-1461677) and the Department of Energy (Award No. 10006272). Chen's research is partially supported by the U.S. National Institutes of Health (U01 HL114494) and the U.S. National Science Foundation (DMS-1613295).

BIBLIOGRAPHY

Antoniadis, A. (1997). Wavelets in statistics: A review (with discussion). *Journal of the Italian Statistical Association*, 6, 97–144.

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

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

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41, 802–837.

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

Bühlmann, P. & van de Geer, S. (2009). *Statistics for High-Dimensional Data*. Springer, New York.

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

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

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

Conn, A., Gould, N., & Toint, P. (1991). A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, 28, 545–572.

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

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, 81, 425–455.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348–1360.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models.* Springer, New York.

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

García-Escudero, L. A., Gordaliza, A., San Martín, R., Van Aelst, S., & Zamar, R. (2009). Robust linear clustering. *Journal of The Royal Statistical Society Series*, B71, 301–318.

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

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

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

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.

Huang, J., Breheny, P. & Ma, S. (2012). A selective review of group selection in high dimensional models. *Statistical Science*, 27, 481–499.

Huang, J., Ma, S. & Zhang, C.-H. (2008). Adaptive lasso for high-dimensional regression models. *Statistica Sinica*, 18, 1603–1618.

Jiang, W. & 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.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44, 907–927.

Lee, Y., MacEachern, S. N. & Jung, Y. (2012). Regularization of case-specific parameters for robustness and efficiency. *Statistical Science*, 27(3), 350–372.

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

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

Mueller, C. H. & 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.

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

Qin, Z., Scheinberg, K., & Goldfarb, D. (2013). Efficient block-coordinate descent algorithms for the Group Lasso. *Mathematical Programming Computation*, 5, 143–169.

She, Y. & Owen, A. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106, 626–639.

Shen, H., Yang, J., & Wang, S. (2004). Outlier detecting in fuzzy switching regression models. *Artificial Intelligence: Methodology, Systems, and Applications*, 208–215.

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

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

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

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of The Royal Statistical Society Series B*, 58, 267–288.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2014). Exact post-selection inference for sequential regression procedures. *arXiv preprint arXiv:1401.3889*.

Wedel, M. & Kamakura, W. A. (2000). *Market Segmentation: Conceptual and Methodological Foundations.* Springer, New York.

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

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

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

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

Yi, G. Y., Tan, X., & Li, R. (2015). Variable selection and inference procedures for marginal analysis of longitudinal data with missing observations and covariate measurement error. *The Canadian Journal of Statistics*, 43, 498–518.

Yu, C., Chen, K., & Yao, W. (2015). Outlier detection and robust mixture modelling using nonconvex penalized likelihood. *Journal of Statistical Planning and Inference*, 164, 27–38.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942.

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

Received 12 February 2016
Accepted 2 October 2016