Title
A new method for robust mixture regression

Permalink
https://escholarship.org/uc/item/06v19071

Journal
CANADIAN JOURNAL OF STATISTICS-REVUE CANADIENNE DE STATISTIQUE, 45(1)

ISSN
0319-5724

Authors
Yu, C
Yao, W
Chen, K

Publication Date
2017-03-01

DOI
10.1002/cjs.11310

Peer reviewed
A New Method for Robust Mixture Regression

Abstract

Finite mixture regression models have been widely used for modeling mixed regression relationships arising from a clustered and thus heterogenous population. The classical normal mixture model, despite of 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 the M-estimation approaches. Contrast with several existing methods, the proposed methods show outstanding performance in our numerical studies.

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

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 $(y_1, x_1), \ldots, (y_n, x_n)$ 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 modeled by a **finite mixture regression model**, consisting of, say, \(m\) homogeneous linear regression components. Specifically, it is assumed that a regression model holds for each of the \(m\) components, i.e., when \((y, x)\) belongs to the \(j\)th component \((j = 1, \ldots, m)\), \(y = x^T \beta_j + \epsilon_j\), where \(\beta_j \in \mathbb{R}^p\) 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 \(x\) vector as one). The conditional density of \(y\) given \(x\), is

\[
f(y \mid x, \theta) = \sum_{j=1}^{m} \pi_j \phi(y; x^T \beta_j, \sigma_j^2),
\]

where \(\phi(\cdot; \mu, \sigma^2)\) denotes the probability density function (pdf) of the normal distribution \(N(\mu, \sigma^2)\), \(\pi_j\)'s are the mixing proportions, and \(\theta = (\pi_1, \beta_1, \sigma_1; \ldots; \pi_m, \beta_m, \sigma_m)\) collects all the unknown parameters.

Since first introduced by Goldfeld and Quandt (1973), the above mixture regression model has been widely used in business, marketing, social sciences, etc; see, e.g., Jiang and Tanner (1999), Böhning (1999), Wedel and Kamakura (2000), Henning (2000), McLachlan and Peel (2000), Skrondal and Rabe-Hesketh (2004), and Frühwirth-Schnatter (2006). Maximum likelihood estimation (MLE) is commonly carried out to infer \(\theta\) in (1.1), i.e.,

\[
\hat{\theta}_{mle} = \arg \max_{\theta} \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i; x_i^T \beta_j, \sigma_j^2) \right\}.
\]

The \(\hat{\theta}_{mle}\) does not have an explicit form in general and it is usually obtained by the Expectation-Maximization (EM) algorithm (Dempster et al., 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 the outlying effects may greatly jeopardize both model estimation and inference. Many robust methods have been developed for mixture regression models. Markatou (2000) and Shen et al. (2004) proposed to weight each data point to robustify the estimation procedure. Neykov et al. (2007) proposed to fit the mixture model using the trimmed likelihood method. Bai et al. (2012) developed a modified EM algorithm by adopting a robust criterion in the M-step. Bashir and Carter (2012) extended the idea of the S-estimator to mixture regression. Yao et al. (2014) and Song
et al. (2014) considered robust mixture regression using a $t$-distribution and a laplace distribution, respectively. There have also been extensive work in linear clustering; see, e.g., Henning (2002), Henning (2003), Mueller and Garlipp (2005), García-escudero et al. (2009), and García-escudero et al. (2010).

Motivated by She and Owen (2011), Lee et al. (2012) and Yu et al. (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 et al. (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 other robust estimation criterion or complex heavy-tailed distributions to robustify the mixture regression models, our method is built upon a simple normal mixture regression model to facilitate computation and model interpretation. Second, we adopt a sparse and scale-dependent mean-shift parameterization to robustify the normal mixture model. 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 et al. (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 thus provide justifications of the proposed methods and shed light on their robustness properties. These connections also apply to special cases of mixture modeling. Compared to existing robust methods, $RM^2$ allows an efficient solution via the celebrated penalized regression approach, and many information criteria (such as AIC and BIC) can then be used to data adaptively determine the proportion of outliers. With extensive numerical studies, $RM^2$ is demonstrated to be highly robust against 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

\[ f(y_i \mid x_i, \theta, \gamma_i) = \sum_{j=1}^{m} \pi_j \phi(y_i; x_i^T \beta_j + \gamma_{ij} \sigma_j), \quad i = 1, \ldots, n, \]  

(2.1)

where \( \theta = (\pi_1, \beta_1, \sigma_1, \ldots, \pi_m, \beta_m, \sigma_m)^T \). Here, for each observation, a mean-shift parameter, \( \gamma_{ij} \), is added to its mean structure in each mixture component; we refer to (2.1) as a mean-shifted normal mixture model (RM\(^2\)). Define \( \gamma_i = (\gamma_{i1}, \ldots, \gamma_{im})^T \) as the mean-shift vector for the \( i \)th observation for \( i = 1, \ldots, n \), and let \( \Gamma = (\gamma_1^T, \ldots, \gamma_n^T)^T \) collect all the mean-shift parameters.

Without any constraints on the mean-shift parameters in (2.1), the model is apparently over-parameterized. The essence of (2.1) lies in additional sparsity structures which are imposed on the parameters \( \gamma_{ij} \): we assume many \( \gamma_{ij} \)'s are in fact zero, corresponding to the typical observations; and only a few \( \gamma_{ij} \)'s are nonzero, corresponding to the outliers. Therefore, promoting sparsity of \( \gamma_{ij} \) 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, i.e., the outlying effect of the \( i \)th observation to the \( j \)th component is modeled by \( \gamma_{ij} \sigma_j \), depending directly on the scale of the \( j \)th component. This setup is thus much more flexible than the structure considered by Yu et al. (2015) in the context of mixture model. In our model, each \( \gamma_{ij} \) parameter becomes scale free, and can be understood as the number of standard deviations shifted from the mixture regression structure.

The model framework developed in (2.1) inherits the simplicity of the normal mixture model, and it allows us to take advantage of the celebrated penalized estimation approaches (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006; Huang et al., 2008) for realizing robust estimation. For a comprehensive account of the penalized regression and variable selection techniques, see, e.g., Bühlmann and van de Geer (2009) and Huang et al. (2012). With the model in (2.1), we propose a penalized likelihood approach for conducting model estimation,

\[ (\hat{\theta}, \hat{\Gamma}) = \arg \max_{\theta, \Gamma} J_n(\theta, \Gamma), \]  

(2.2)
where

\[ J_n(\theta, \Gamma) = l_n(\theta, \Gamma) - \sum_{i=1}^{n} P_\lambda(\gamma_i), \]

\[ l_n(\theta, \Gamma) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i - \gamma_{ij}\sigma_j - x_i^T \beta_j; 0, \sigma_j^2) \right\} \]

is the log-likelihood function, and \( P_\lambda(\cdot) \) is a penalty function chosen to induce either element-wise or vector-wise sparsity of the enclosed vector with \( \lambda \) being a tuning parameter which controls the degrees of penalization. Similar to the traditional mixture model (1.1), the above penalized log-likelihood is also unbounded. That is, the penalized log-likelihood goes to infinity when \( y_i = x_i^T \beta + \gamma_{ij}\sigma_j \), and \( \sigma_j \to 0 \) (Hathaway, 1985, 1986; Chen et al., 2008; Yao, 2010). To circumvent this problem, following Hathaway (1985, 1986), we restrict \((\sigma_1, \ldots, \sigma_m) \in \Omega_\sigma\), with \( \Omega_\sigma \) defined as

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

where \( \epsilon \) is a very small positive value. In our examples to follow, we set \( \epsilon = 0.01 \). Accordingly, we define the parameter space of \( \theta \) as

\[ \Omega = \{ (\pi_j, \beta_j, \sigma_j), j = 1, \ldots, m : 0 \leq \pi_j \leq 1, \sum_{j=1}^{m} \pi_j = 1, (\sigma_1, \ldots, \sigma_m) \in \Omega_\sigma \}. \]

There are many choices of the penalty function in (2.2). For inducing vector-wise sparsity, we may consider the group lasso penalty of the form \( P_\lambda(\gamma_i) = \lambda \|\gamma_i\|_2 \) and the group \( \ell_0 \) penalty \( P_\lambda(\gamma_i) = \lambda^2 I(\|\gamma_i\|_2 \neq 0)/2 \), where \( \|\cdot\|_q \) denotes the \( \ell_q \) norm for \( q \geq 0 \), and \( I(\cdot) \) is the indicator function. These penalty functions penalize the \( \ell_2 \) norm of each \( \gamma_i \) vector, to promote the entire vector to be a zero vector. Alternatively, one may take \( P_\lambda(\gamma_i) = \sum_{j=1}^{m} P_\lambda(|\gamma_{ij}|) \), where \( P_\lambda(|\gamma_{ij}|) \) is a penalty function to induce element-wise sparsity. Some examples are the \( \ell_1 \) norm penalty (Donoho and Johnstone, 1994a; Tibshirani, 1996),

\[ P_\lambda(\gamma_i) = \lambda \sum_{j=1}^{m} |\gamma_{ij}|, \quad (2.4) \]
and the $\ell_0$ norm penalty (Antoniadis, 1997)

$$P_\lambda(\gamma_i) = \frac{\lambda^2}{2} \sum_{j=1}^{m} I(\gamma_{ij} \neq 0).$$  \hfill (2.5)

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

### 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 the unobservable component labels are treated as missing data. Here, we propose an efficient thresholding-embedded EM algorithm to maximize the proposed penalized log-likelihood criterion. Consider

$$\left(\hat{\theta}, \hat{\Gamma}\right) = \arg\max_{\theta \in \Omega, \Gamma} \left\{ \sum_{i=1}^{n} \log \left( \sum_{j=1}^{m} \pi_j \phi(y_i - x_i^T \beta_j - \gamma_{ij} \sigma_j; 0, \sigma_j^2) \right) - \sum_{i=1}^{n} \sum_{j=1}^{m} P_\lambda(|\gamma_{ij}|) \right\},$$

where $P_\lambda(\cdot)$ is either the $\ell_1$ penalty function in (2.4) or the $\ell_0$ penalty function in (2.5). The proposed method can be readily applied to other penalty forms such as group lasso and group $\ell_0$ penalties; see the Appendix for more details.

Let

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

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

$$J_n^{c}(\theta, \Gamma) = l_n^{c}(\theta, \Gamma) - \sum_{i=1}^{n} \sum_{j=1}^{m} P_\lambda(|\gamma_{ij}|),$$  \hfill (2.6)

where the complete log-likelihood is given by $l_n^{c}(\theta, \Gamma) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log \left\{ \pi_j \phi(y_i - x_i^T \beta_j - \gamma_{ij} \sigma_j; 0, \sigma_j^2) \right\}$.

In the E-step, given the current estimates $\theta^{(k)}$ and $\Gamma^{(k)}$, the conditional expectation of the
penalized complete log-likelihood (2.6) is computed,

\[
Q(\theta, \Gamma \mid \theta^{(k)}, \Gamma^{(k)})
= \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij}^{(k+1)} \left\{ \log \pi_j + \log \phi(y_i - x_i^T \beta_j - \gamma_{ij} \sigma_j; 0, \sigma_j^2) \right\} - \sum_{i=1}^{n} \sum_{j=1}^{m} P_\lambda(\gamma_{ij}) \tag{2.7}
\]

where

\[
P_{ij}^{(k+1)} = \mathbb{E}(z_{ij} \mid y_i; \theta^{(k)}, \Gamma^{(k)}) = \frac{\pi_j^{(k)} \phi(y_i - x_i^T \beta_j^{(k)} - \gamma_{ij}^{(k)} \sigma_j^{(k)}; 0, \sigma_j^{(k)^2}}{\sum_{j=1}^{m} \pi_j^{(k)} \phi(y_i - x_i^T \beta_j^{(k)} - \gamma_{ij}^{(k)} \sigma_j^{(k)}; 0, \sigma_j^{(k)^2})}. \tag{2.8}
\]

We then maximize (2.7) with respect to \((\theta, \Gamma)\) in the M-step. Specifically, in the M-step, \(\theta\) and \(\Gamma\) are alternatingly updated until convergence. For fixed \(\Gamma\) and \(\sigma_j\)'s, each \(\beta_j\) can be solved explicitly from a weighted least squares procedure. For fixed \(\Gamma\) and \(\beta_j\)'s, as each \(\sigma_j\) 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 et al., 1991). Also, when ignoring the ratio constraints in (2.3), the optimization problem of \(\sigma_j\)'s becomes separable and each \(\sigma_j\) can be updated more easily; in practice, the constrained estimation is performed only when the above simple solutions violate the ratio condition in (2.3).

For fixed \(\theta, \Gamma\) is updated by maximizing

\[
\sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij}^{(k+1)} \log \phi(y_i - x_i^T \beta_j - \gamma_{ij} \sigma_j; 0, \sigma_j^2) - \sum_{i=1}^{n} \sum_{j=1}^{m} P_\lambda(\gamma_{ij}).
\]

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

\[
\frac{1}{2} \left( \frac{\gamma_{ij} - y_i}{\sigma_j} \right)^2 + \frac{1}{P_{ij}^{(k+1)}} P_\lambda(\gamma_{ij}). \tag{2.9}
\]

This one dimensional problem admits an explicit solution. The solution of (2.9) for using the \(\ell_1\)
penalty or the $\ell_0$ penalty is given by a corresponding thresholding rule $\Theta_{\text{soft}}$ or $\Theta_{\text{hard}}$, respectively:

\begin{align}
\hat{\gamma}_{ij} & = \Theta_{\text{soft}}(\xi_{ij}; \lambda_{ij}^*) = \text{sgn}(\xi_{ij})(|\xi_{ij}| - \lambda_{ij}^*)_+, \\
\hat{\gamma}_{ij} & = \Theta_{\text{hard}}(\xi_{ij}; \lambda_{ij}^*) = \xi_{ij} I(|\xi_{ij}| > \lambda_{ij}^*),
\end{align}

where $\xi_{ij} = (y_i - x_i^T \beta_j)/\sigma_j$, $a_+ = \max(a, 0)$, $\lambda_{ij}^*$ is taken as $\lambda/p_i^{(k+1)}$ in $\Theta_{\text{soft}}$, and $\lambda_{ij}^*$ is set as $\lambda/\sqrt{p_i^{(k+1)}}$ in $\Theta_{\text{hard}}$. See the Appendix for the details on handling group penalties on $\gamma_i$’s, such as the group $\ell_1$ penalty and the group $\ell_0$ penalty.

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

**Algorithm 1** Thresholding-Embedded EM algorithm for RM²

1. **Initialize** $\theta^{(0)}$ and $\Gamma^{(0)}$. Set $k ← 0$.

2. **repeat**
   - **(1) E-Step:** Compute $Q(\theta, \Gamma | \theta^{(k)}, \Gamma^{(k)})$ based on (2.7) and (2.8).
   - **(2) M-Step:** Update $\pi^{(k+1)}_j = \frac{\sum_{i=1}^n p_{ij}^{(k+1)}}{n}$ and update other parameters by maximizing $Q(\theta, \Gamma | \theta^{(k)}, \Gamma^{(k)})$, i.e., start from $(\beta^{(k)}, \sigma_j^{2(k)}, \Gamma^{(k)})$ and iterate the following steps until convergence to obtain $(\beta^{(k+1)}, \sigma_j^{2(k+1)}, \Gamma^{(k+1)})$:
     
     \begin{align}
     (2.a) & \quad \beta_j \leftarrow \left( \sum_{i=1}^n x_i x_i^T p_{ij}^{(k+1)} \right)^{-1} \left( \sum_{i=1}^n x_i p_{ij}^{(k+1)} (y_i - \gamma_{ij} \sigma_j) \right), j = 1, \ldots, m, \\
     (2.b) & \quad (\sigma_1, \ldots, \sigma_m) \leftarrow \arg \max_{(\sigma_1, \ldots, \sigma_m) \in \Omega} \sum_{i=1}^n \sum_{j=1}^m p_{ij}^{(k+1)} \log \phi(y_i - x_i^T \beta_j - \gamma_{ij} \sigma_j; 0, \sigma_j^2), \\
     (2.c) & \quad \gamma_{ij} \leftarrow \Theta(\xi_{ij}; \lambda_{ij}^*), i = 1, \ldots, n, j = 1, \ldots, m,
     \end{align}

   where $\Theta$ denotes one of the thresholding rules in (2.10–2.11) depending on the penalty form adopted.

   $k ← k + 1$.

3. **until** convergence

The penalized log-likelihood does not decrease along each iteration of the E-step and M-step, i.e.,

$$J_n(\hat{\theta}^{(k+1)}, \hat{\Gamma}^{(k+1)}) \geq J_n(\hat{\theta}^{(k)}, \hat{\Gamma}^{(k)})$$

for all $k \geq 0$. This property ensures the convergence of of Algorithm 2.2.

The proposed algorithm can be readily modified to handle the special case of equal variances in
model (1.1) with $\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$ for some $\sigma^2 > 0$. In the Algorithm, $\sigma_j$ shall be replaced by $\sigma$.

The iterating steps mostly stay the same, except that step (2.b) becomes

\[
(2.b) \quad \sigma^2 \leftarrow \arg \max_{\sigma^2 > 0} \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - x_i^T \beta_j - \gamma_{ij}\sigma; 0, \sigma^2).
\]

The proposed EM algorithm is implemented for a fixed tuning parameter $\lambda$. In practice, we need to choose an optimal $\lambda$ and hence an optimal set of parameter estimates. We construct a Bayesian information criterion (BIC) for tuning parameter selection (e.g., Yi et al. (2015)),

\[
\text{BIC}(\lambda) = -l(\lambda) + \log(n)df(\lambda),
\]

where $l(\lambda)$ is the mixture log-likelihood function evaluated at the solution of tuning parameter $\lambda$, 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 $\hat{\Gamma}$ and the number of component parameters in the mixture model. We fit the model for a certain number, say 100, of $\lambda$ values which are equally spaced at the log scale in an interval $(\lambda_{\text{min}}, \lambda_{\text{max}})$, where $\lambda_{\text{min}}$ is the smallest $\lambda$ value for which roughly 50% of the entries in $\Gamma$ are nonzero, and $\lambda_{\text{max}}$ corresponds to the largest $\lambda$ value for which $\Gamma$ is estimated as a zero matrix. Other ways are possible as well (e.g., $C_p$, AIC and GCV) to determine the optimal solution along the solution path. For example, one may discard a certain percentage of the observations as outliers if such prior knowledge is available. In the proposed model, since 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 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, 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 1–2 below. The proofs are provided in the Appendix.
Theorem 1. Consider RM$^2$ with a group $\ell_0$ penalization, i.e.,

$$
(\hat{\theta}, \hat{\Gamma}) = \arg \max_{\theta \in \Omega, \Gamma} \left\{ \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i - x_i^T \beta_j - \gamma_{ij} \sigma_j; 0, \sigma_j^2) \right\} - \frac{\lambda^2}{2} \sum_{i=1}^{n} I(\|\gamma_i\|_2 \neq 0) \right\}.
$$

(3.1)

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

$$
(\hat{\theta}, \hat{S}) = \arg \max_{\theta \in \Omega, S; |S| = n-h} \left[ \sum_{i \in S^c} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i - x_i^T \beta_j; 0, \sigma_j^2) \right\} + (n - h) \log \left\{ \sum_{j=1}^{m} \pi_j \phi(0; 0, \sigma_j^2) \right\} \right].
$$

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, i.e.,

$$
(\hat{\pi}, \hat{\beta}) = \arg \max_{\pi, \beta, S; |S| = n-h} \left[ \sum_{i \in S^c} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i - x_i^T \beta_j; 0, \sigma^2) \right\} \right].
$$

(3.2)

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 completely equivalent. This result partly explains the robust property of RM$^2$, and shows that the trimmed likelihood estimation can be conveniently achieved by the proposed penalized likelihood approach.

In the classical EM algorithm for solving the normal mixture model, the regression coefficients are updated based on the weighted least squares. A natural idea to robustify the normal mixture model is then to replace the weighted least squares by some robust estimation criterion, such as using M-estimation. Bai et al. (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,

$$
(\hat{\theta}, \hat{\Gamma}) = \arg \max_{\theta \in \Omega, \Gamma} \left\{ \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{m} \pi_j \phi(y_i - x_i^T \beta_j - \gamma_{ij} \sigma_j; 0, \sigma_j^2) \right\} - \sum_{i=1}^{n} \sum_{j=1}^{m} P_\lambda(\|\gamma_{ij}\|) \right\}.
$$

From the thresholding-embedded EM algorithm, define $\hat{W}_j = \text{diag}(\hat{\rho}_{ij}, \ldots, \hat{\rho}_{nj})$, and $\hat{w}_j = (\lambda_{1j}^*, \ldots, \lambda_{nj}^*)^T$. Here for simplicity we omit the superscript $(k)$ which denotes the iteration number. Then the pa-
Parameter estimates satisfy

\[
\hat{\gamma}_j = \Theta\left(\frac{1}{\hat{\sigma}_j}(y - \hat{X}\hat{\beta}_j), \hat{w}_j\right), \quad \text{and} \quad \hat{\beta}_j = (X^T\hat{W}_jX)^{-1}X^T\hat{W}_j(y - \hat{\sigma}_j\hat{\gamma}_j),
\]

(3.3)

where \(\Theta\) is defined componentwise.

**Theorem 2.** Consider RM² with an element-wise sparsity-inducing penalization, and define \((\hat{\theta}, \hat{\Gamma})\) as in (3.3). Then the parameter estimates satisfy

\[
X^T\hat{W}_j\psi\left(\frac{1}{\hat{\sigma}_j}(y - \hat{X}\hat{\beta}_j), \hat{w}_j\right) = 0, \quad j = 1, \ldots, m,
\]

(3.4)

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

In Theorem 2, the score equation (3.4) defines exactly an M-estimator. Interestingly, as shown by She and 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 \(\Theta_{soft}\), the corresponding \(\psi\) function is the well-known Huber’s \(\psi\). Similarly, \(\Theta_{hard}\) 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 \(\psi\) function is preferable, which corresponding to the use of a nonconvex penalty in RM². In Bai et al. (2012), the criterion parameter \(\lambda\) in \(\psi(t; \lambda)\) is a prespecified value, and it stays the same for any input \((y_i - x_i^T\hat{\beta}_j)/\hat{\sigma}_j\). In contrast, the criterion parameter becomes adaptive in RM², and its overall magnitude is determined by the penalization parameter, whose choice is data-driven and based on certain information criterion.

**4 Simulation**

**4.1 Simulation Setups**

We consider two mixture regression model setups, in which the observations are contaminated with additive outliers. We evaluate the finite sample performance of RM² and compare it with several existing methods. As we mainly focus on investigating the 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 by

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

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

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

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

where $z_{i1}$ is a component indicator generated from Bernoulli distribution with $P(z_{i1} = 1) = 0.3$; $x_{i1}$ and $x_{i2}$ are independently generated from $N(0, 1)$, and the error terms $\epsilon_{i1}$ and $\epsilon_{i2}$ are independently generated from $N(0, \sigma^2_1)$ and $N(0, \sigma^2_2)$, respectively, with $\sigma^2_1 = 1$ 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 $\gamma_{ij}$s set to be zero; when there are 5% (or 10%) outliers, 5 (or 10) observations from the first component are then replaced by $y_i = 1 - x_{i1} + x_{i2} - |\gamma_{i1}|\sigma + \epsilon_{i1}$ with $\sigma = 1$, $x_{i1} = 2$ and $x_{i2} = 2$, and 15 (or 30) observations from the second component are replaced by $y_i = 1 + 3x_{i1} + x_{i2} + |\gamma_{i2}|\sigma + \epsilon_{i2}$ with $\sigma = 1$, $x_{i1} = 2$, 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 $\sigma_2 = 2$. We repeat the simulation for 200 times in each setting.

4.2 Methods and Evaluation Measures

We compare our proposed RM² estimators with $\ell_1$ and $\ell_0$ penalties, denoted as RM²($\ell_1$) and RM²($\ell_0$), respectively, to several existing robust regression estimators and the MLE of the classical normal mixture regression model. To examine the true potential of the RM² 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 the true outliers on the solution path. This is the penalized regression estimator we would have obtained if the true number of outliers is 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 5% (TLE_{0.05}) or 10% (TLE_{0.10}). We note that TLE_{0.05} (TLE_{0.10}) can be regarded as the oracle TLE estimator when there are 5% (10%) outliers.

3. the robust estimator based on a modified EM algorithm with a bisquare loss (MEM-bisquare) proposed by Bai et al. (2012).

4. the MLE in mixture linear regression assuming a t-distributed error (Mixregt) proposed by Yao et al. (2014).

5. the RM^2 element-wise estimators using the ℓ_0 penalty (RM^2(ℓ_0)) and the ℓ_1 penalty (RM^2(ℓ_1)), and their oracle counterparts RM^2_0(ℓ_0) and RM^2_1(ℓ_1).

For fitting mixture models, there are well-known label switching issues (Celeux et al., 2000; Stephens, 2000; Yao and Lindsay, 2009; Yao, 2012). In our simulation study, since the underlying truth is known, the labels are determined by minimizing the Euclidean distance from the true parameter values. To evaluate the estimation performance, we report the median squared errors (MeSE) and the mean squared errors (MSE) of the parameter estimates. To evaluate the outlier detection performance, we report three measures: the average proportion of masking (M), i.e., the fraction of undetected outliers; the average proportion of swamping (S), i.e., the fraction of good points labeled as outliers; and the joint detection rate (JD), i.e., the proportion of simulations with 0 masking.

4.3 Simulation Results

The simulation results of 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 other robust methods. In the case of 5% outliers, all methods (except MLE) perform well in
outlier detection. For parameter estimation, RM^2(ℓ_1) and RM^2_0(ℓ_1) perform much worse than other methods. In the case of 10% outliers, RM^2(ℓ_0) and TLE_{0.10} work well, whereas RM^2(ℓ_1), TLE_{0.05}, MEM-bisquare, and Mixregt have much lower joint outlier detection rates and hence larger MeSE or MSE. The non-robustness of RM^2(ℓ_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 penalized regression point of view. While the ℓ_1 regularization induces sparsity, it also results in a heavy shrinkage effect. Consequently, the method tends to accommodate the outlying effects in the model, leading to biased and severely distorted estimation results. In contrast, the ℓ_0 penalization does not offer any shrinkage, so it is much harder for an outlier to stay in the model 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 simulated data set in Model 1, with two regression planes estimated by RM^2(ℓ_0). The outliers are marked in blue. It can be seen that the regression planes fit the bulk of good observations from two components quite well and the estimates are not influenced by the outliers.

Table 2 reports the simulation results of Model 2 (unequal variances case). The conclusions are similar to those from the equal variances case. Briefly, with 5% outliers, all methods (except MLE) have high joint outlier detection rates. When there are 10% outliers, RM^2(ℓ_0) and the trimmed likelihood methods continue to perform the best. In either case, the estimation accuracy of RM^2(ℓ_1) is much lower than that of RM^2(ℓ_0). Also, a 3-dimensional scatter plot of one typical simulated data set in Model 2 is shown in Figure 2 which clearly demonstrates that the estimates of RM^2(ℓ_0) are not influenced by the outliers.

In summary, TLE_{0.10} has good results in terms of outliers detection in all cases but has larger MSE for 5% outliers case. TLE_{0.05} fails to work in the case of 10% outliers due to 5% outliers remain in the data.

RM^2(ℓ_0) has comparable performance to the oracle TLE and RM^2_0(ℓ_0) in terms of both outlier detection and MeSE in most cases excepted for the 10% outlier case with a small |γ|. RM^2(ℓ_1) has comparable results to RM^2(ℓ_0) and TLE in terms of outlier detection with a large |γ| but fails to work with a small |γ|. As expected, MLE is sensitive to outliers.
5 Tone Perception Data Analysis

We apply the proposed robust procedure to tone perception data (Cohen, 1984). In the tone perception experiment of Cohen (1984), a pure fundamental tone with electronically generated overtones added was played to a trained musician. The experiment recorded 150 trials from the same musician. The overtones were determined by a stretching ratio, which is the ratio between 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 $\text{RM}^2(\ell_0)$ 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 dataset. Table 3 reports the parameter estimates. For the original data which contain no outliers, the proposed $\text{RM}^2(\ell_0)$ estimator has similar parameter estimates to that of the traditional MLE. This result shows that our proposed $\text{RM}^2(\ell_0)$ method performs as well as the traditional MLE. If there are outliers in the data, the proposed $\text{RM}^2(\ell_0)$ estimator is not influenced by the outlier and gives similar parameter estimates to the case of no outliers. However, MLE gives totally wrong parameter estimates.

6 Discussion

We have proposed a robust mixture regression approach using a mean-shift normal mixture model parameterization, generalizing the work by She and Owen (2011), Lee et al. (2012), and Yu et al. (2015). The method is shown to have strong connections with several well-known robust methods. The proposed $\text{RM}^2$ method with the $\ell_0$ penalty has comparable performance to its oracle counterpart and the oracle Trimmed Likelihood Estimator (TLE).

There are several directions for future research. The oracle $\text{RM}^2$ estimators may have better performance than the BIC-tuned estimators in some cases; therefore, we can further improve the performance of $\text{RM}^2$ by improving tuning parameter selection. García-Escudero et al. (2010) showed that the traditional definition of breakdown point is not an appropriate measure to quantify the robustness of mixture regression procedures, since 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 mixture model setup. Although we do not discuss
the selection of the number of cluster components in this paper, it remains a pressing issue in
many mixture modeling problems. Finally, the proposed RM² approach can be further extended to
conduct simultaneous variable selection and outlier detection in mixture regression.

Appendix

Handling Group Penalties

The proposed EM algorithm can be readily modified to handle group lasso penalty and the group
ℓ₀ penalty on γ_i’s. The only change is in the way of updating Θ in the M step when θ is fixed.

For group lasso penalty, i.e., P_{λ}(γ_i) = λ∥γ_i∥₂, Θ is updated by maximizing

\[ \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(k+1)} \log \phi(y_i - x_i^T \beta_j - γ_{ij} \sigma_j; 0, \sigma_j^2) - \lambda \sum_{i=1}^{n} ∥γ_i∥₂. \]

The problem is separable in each γ_i. After some algebra, the problem for each γ_i has exactly the
same form as the problem considered in Qin et al. (2013),

\[ \hat{γ}_i = \arg \min_{γ_i} \frac{1}{2} γ_i^T W_i γ_i - a_i^T γ_i + λ∥γ_i∥₂, \] (6.1)

where W_i is an m × m diagonal matrix with diagonal elements \( \left\{ p_{ij}^{(k+1)} r_{ij} / σ_j \right\}, r_{ij} = y_i - x_i^T \beta_j, \) and \( a_i = \left( p_{i1}^{(k+1)} r_{i1} / σ_1, \ldots, p_{im}^{(k+1)} r_{im} / σ_m \right)^T. \) The detailed algorithm is given in Qin et al. (2013).

The solution of (6.1) can be expressed as

\[ \hat{γ}_i = \begin{cases} 0 & \text{if } ∥a_i∥₂ \leq λ; \\ Δ_i(Δ_i W_i + λI)^{-1} a_i & \text{if } ∥a_i∥₂ > λ, \end{cases} \]

in which Δ_i is the root of

\[ \phi(Δ_i) = 1 - \frac{1}{∥f(Δ_i)∥₂}, \]

where

\[ ∥f(Δ_i)∥₂^2 = \sum_{j=1}^{m} \left( \frac{p_{ij}^{(k+1)} r_{ij} / σ_j}{p_{ij}^{(k+1)}} Δ_i + λ \right)^2. \]
For group $\ell_0$ penalty, $\Gamma$ is updated by maximizing

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

The problem is separable in each $\gamma_i$, i.e,

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

$\gamma_i$ has a closed form solution. Define $\tilde{\gamma}_i$ such that $\tilde{\gamma}_{ij} = \frac{(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j)}{\hat{\sigma}_j}$, $j = 1, \ldots, m$. Then

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

**Proof of Theorem 1**

*Proof.* Recall $(\hat{\theta}, \hat{\Gamma})$ is the maximizer of the penalized log-likelihood problem (3.1). Then we have

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

The problem is separable in each $\gamma_i$, i.e,

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

If $\hat{\gamma}_i = 0$, (6.3) becomes

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

and if $\hat{\gamma}_i \neq 0$, it must be true that $\hat{\gamma}_{ij} = \frac{(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j)}{\hat{\sigma}_j}$, $j = 1, \ldots, m$, and (6.3) then becomes

$$
\log \left\{ \sum_{j=1}^{m} \hat{\pi}_j \phi(0; 0, \hat{\sigma}_j^2) \right\} - \frac{\lambda^2}{2}.
$$
It then follows that the maximum of the penalized log-likelihood (3.1) is

$$
\sum_{i \in \hat{S}} \log \left\{ \sum_{j=1}^{m} \hat{\pi}_j (y_i - x_i^T \hat{\beta}_j; 0, \hat{\sigma}_j^2) \right\} + \sum_{i \in \hat{S}} \log \left\{ \sum_{j=1}^{m} \hat{\pi}_j (0; 0, \hat{\sigma}_j^2) \right\} - \frac{\lambda^2}{2} (n - h). \tag{6.4}
$$

For a given tuning parameter $\lambda$, the number of nonzero $\hat{\gamma}_{ij}$ 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 (6.4) becomes $\sum_{i \in \hat{S}} \log \{1/(\sqrt{2\pi\sigma})\}$ and hence is a constant. It follows that maximizing (3.1) is equivalent to solving (3.2), in which $S$ is an index set with the same cardinality as $\hat{S}$. We recognize that (3.2) is exactly a trimmed likelihood problem. This completes the proof.

\[\square\]

**Proof of Theorem 2**

*Proof.* Consider element-wise penalization in (3.3). Based on the thresholding-embedded EM algorithm, we write

$$
\hat{\Gamma} = \arg \max_{\Gamma} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{p}_{ij} \log \phi(y_i - x_i^T \hat{\beta}_j - \gamma_{ij} \hat{\sigma}_j; 0, \hat{\sigma}_j^2) - \sum_{i=1}^{n} \sum_{j=1}^{m} P_{\lambda}(|\gamma_{ij}|) \right\}.
$$

The above problem is separable in each $\gamma_{ij}$,

$$
\hat{\gamma}_{ij} = \arg \max_{\gamma_{ij}} \left\{ \hat{p}_{ij} \log \phi(y_i - x_i^T \hat{\beta}_j - \gamma_{ij} \hat{\sigma}_j; 0, \hat{\sigma}_j^2) - P_{\lambda}(|\gamma_{ij}|) \right\} = \arg \min_{\gamma_{ij}} \frac{1}{2} \left( \gamma_{ij} - \frac{y_i - x_i^T \hat{\beta}_j}{\hat{\sigma}_j} \right)^2 + \frac{1}{\hat{p}_{ij}} P_{\lambda}(|\gamma_{ij}|),
$$

It can be easily shown that

$$
\hat{\gamma}_{ij} = \Theta \left( \frac{y_i - x_i^T \hat{\beta}_j}{\hat{\sigma}_j}, \lambda_{ij}^* \right),
$$

where the correspondence of $(P_{\lambda}(\cdot), \lambda_{ij}^*, \Theta)$ is discussed in Section 2.2. For example, using $\ell_1$ penalty leads to $\Theta = \Theta_{soft}$ and $\lambda_{ij}^* = \lambda/p_{ij}^{(k+1)}$; using $\ell_0$ penalty leads to $\Theta = \Theta_{hard}$ and $\lambda_{ij}^* = \lambda/\sqrt{p_{ij}^{(k+1)}}$. 

18
Define $\hat{W}_j = \text{diag}(\hat{p}_{1j}, \ldots, \hat{p}_{nj})$, and $\hat{w}_j = (\lambda_{1j}^*, \ldots, \lambda_{nj}^*)^T$. Then we can write

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

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

$$X^T\hat{W}_j \psi\left(\frac{1}{\hat{\sigma}_j}(y - X\hat{\beta}_j), \hat{w}_j\right) = X^T\hat{W}_j \left\{ \frac{1}{\hat{\sigma}_j}(y - X\hat{\beta}_j) - \Theta\left(\frac{1}{\hat{\sigma}_j}(y - X\hat{\beta}_j), \hat{w}_j\right) \right\}$$

$$= X^T\hat{W}_j \left\{ \frac{1}{\hat{\sigma}_j}y - \frac{1}{\hat{\sigma}_j}X(X^T\hat{W}_j X)^{-1}X^T\hat{W}_j(y - \hat{\sigma}_j \hat{\gamma}_j) - \hat{\gamma}_j \right\}$$

$$= \frac{1}{\hat{\sigma}_j}X^T\hat{W}_j y - \frac{1}{\hat{\sigma}_j}X^T\hat{W}_j(y - \hat{\sigma}_j \hat{\gamma}_j) - X^T\hat{W}_j \hat{\gamma}_j$$

$$= 0.$$

It follows that solving $\beta_j$ is equivalent to solving the score equation in (3.4), which completes the proof.

References


Table 1: Outlier detection results and $MSE_{(se)}/MeSE$ of point estimates for Model 1.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>S</th>
<th>JD</th>
<th>$MSE(\pi)_{(se)}$</th>
<th>MeSE(\pi)</th>
<th>$MSE(\beta)_{(se)}$</th>
<th>MeSE(\beta)</th>
<th>$MSE(\sigma)_{(se)}$</th>
<th>MeSE(\sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RM$^2(\ell_0)$</td>
<td>0.000</td>
<td>0.001</td>
<td>1.000</td>
<td>0.001(0.001)</td>
<td>0.001</td>
<td>0.048(0.002)</td>
<td>0.039</td>
<td>0.003(0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>RM$^2(\ell_6)$</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.001(0.001)</td>
<td>0.001</td>
<td>0.048(0.001)</td>
<td>0.038</td>
<td>0.003(0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>RM$^2(\ell_1)$</td>
<td>0.000</td>
<td>0.006</td>
<td>1.000</td>
<td>0.001(0.001)</td>
<td>0.001</td>
<td>0.048(0.001)</td>
<td>0.037</td>
<td>0.002(0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>RM$^2(\ell_1)$</td>
<td>0.000</td>
<td>0.003</td>
<td>1.000</td>
<td>0.001(0.001)</td>
<td>0.001</td>
<td>0.048(0.001)</td>
<td>0.037</td>
<td>0.002(0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>TLE$0.05$</td>
<td>0.000</td>
<td>0.007</td>
<td>1.000</td>
<td>0.002(0.001)</td>
<td>0.001</td>
<td>0.047(0.002)</td>
<td>0.067</td>
<td>0.002(0.001)</td>
<td>0.004</td>
</tr>
<tr>
<td>TLE$0.10$</td>
<td>0.000</td>
<td>0.003</td>
<td>1.000</td>
<td>0.002(0.001)</td>
<td>0.001</td>
<td>0.085(0.004)</td>
<td>0.216(0.004)</td>
<td>0.666</td>
<td></td>
</tr>
<tr>
<td>MEM-bisquare</td>
<td>0.000</td>
<td>0.005</td>
<td>1.000</td>
<td>0.002(0.001)</td>
<td>0.001</td>
<td>0.500(0.002)</td>
<td>0.615</td>
<td>0.008(0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>Mixreg</td>
<td>0.000</td>
<td>0.078</td>
<td>1.000</td>
<td>0.003(0.001)</td>
<td>0.002</td>
<td>0.090(0.004)</td>
<td>0.080</td>
<td>0.123(0.002)</td>
<td>0.121</td>
</tr>
<tr>
<td>MLE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.470(0.002)</td>
<td>0.680</td>
<td>17.20(0.658)</td>
<td>20.33</td>
<td>2.912(0.023)</td>
<td>2.920</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>S</th>
<th>JD</th>
<th>$MSE(\pi)_{(se)}$</th>
<th>MeSE(\pi)</th>
<th>$MSE(\beta)_{(se)}$</th>
<th>MeSE(\beta)</th>
<th>$MSE(\sigma)_{(se)}$</th>
<th>MeSE(\sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10% outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RM$^2(\ell_0)$</td>
<td>0.000</td>
<td>0.001</td>
<td>1.000</td>
<td>0.001(0.001)</td>
<td>0.001</td>
<td>0.048(0.002)</td>
<td>0.039</td>
<td>0.003(0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>RM$^2(\ell_6)$</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.001(0.001)</td>
<td>0.001</td>
<td>0.048(0.001)</td>
<td>0.038</td>
<td>0.003(0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>RM$^2(\ell_1)$</td>
<td>0.000</td>
<td>0.006</td>
<td>1.000</td>
<td>0.001(0.001)</td>
<td>0.001</td>
<td>0.048(0.001)</td>
<td>0.037</td>
<td>0.002(0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>RM$^2(\ell_1)$</td>
<td>0.000</td>
<td>0.003</td>
<td>1.000</td>
<td>0.001(0.001)</td>
<td>0.001</td>
<td>0.048(0.001)</td>
<td>0.037</td>
<td>0.002(0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>TLE$0.05$</td>
<td>0.000</td>
<td>0.007</td>
<td>1.000</td>
<td>0.002(0.001)</td>
<td>0.001</td>
<td>0.047(0.002)</td>
<td>0.067</td>
<td>0.002(0.001)</td>
<td>0.004</td>
</tr>
<tr>
<td>TLE$0.10$</td>
<td>0.000</td>
<td>0.003</td>
<td>1.000</td>
<td>0.002(0.001)</td>
<td>0.001</td>
<td>0.085(0.004)</td>
<td>0.216(0.004)</td>
<td>0.666</td>
<td></td>
</tr>
<tr>
<td>MEM-bisquare</td>
<td>0.000</td>
<td>0.005</td>
<td>1.000</td>
<td>0.002(0.001)</td>
<td>0.001</td>
<td>0.500(0.002)</td>
<td>0.615</td>
<td>0.008(0.001)</td>
<td>0.001</td>
</tr>
<tr>
<td>Mixreg</td>
<td>0.000</td>
<td>0.078</td>
<td>1.000</td>
<td>0.003(0.001)</td>
<td>0.002</td>
<td>0.090(0.004)</td>
<td>0.080</td>
<td>0.123(0.002)</td>
<td>0.121</td>
</tr>
<tr>
<td>MLE</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.470(0.002)</td>
<td>0.680</td>
<td>17.20(0.658)</td>
<td>20.33</td>
<td>2.912(0.023)</td>
<td>2.920</td>
</tr>
</tbody>
</table>

*The standard errors (se) for MSE’s of the parameter estimation are shown in the parentheses.
Table 2: Outlier detection results and $MSE_{\text{se}}$/MeSE of point estimates for Model 2.

<table>
<thead>
<tr>
<th></th>
<th>$\pi_2$</th>
<th>$\beta_01$</th>
<th>$\beta_11$</th>
<th>$\beta_21$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RM2($\ell_0$)</td>
<td>0.001</td>
<td>0.674</td>
<td>-0.038</td>
<td>1.908</td>
<td>0.699</td>
</tr>
<tr>
<td>RM2($\ell_0$)</td>
<td>0.001</td>
<td>0.655</td>
<td>-0.038</td>
<td>1.908</td>
<td>0.699</td>
</tr>
<tr>
<td>RM2($\ell_1$)</td>
<td>0.000</td>
<td>0.674</td>
<td>-0.038</td>
<td>1.908</td>
<td>0.699</td>
</tr>
<tr>
<td>RM2($\ell_1$)</td>
<td>0.000</td>
<td>0.655</td>
<td>-0.038</td>
<td>1.908</td>
<td>0.699</td>
</tr>
<tr>
<td>TLE0.05</td>
<td>0.000</td>
<td>0.121</td>
<td>0.012</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>TLE0.10</td>
<td>0.000</td>
<td>0.121</td>
<td>0.012</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>MEM-bisque</td>
<td>0.000</td>
<td>0.121</td>
<td>0.012</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>Mixregt</td>
<td>0.000</td>
<td>0.121</td>
<td>0.012</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>MLE</td>
<td>0.000</td>
<td>0.121</td>
<td>0.012</td>
<td>0.005</td>
<td>0.005</td>
</tr>
</tbody>
</table>

The standard errors (se) for the MSE's of the parameter estimation are shown in the parentheses.

Table 3: Parameter Estimation in Tone Perception Data Analysis

<table>
<thead>
<tr>
<th></th>
<th>$\pi_1$</th>
<th>$\pi_2$</th>
<th>$\beta_01$</th>
<th>$\beta_11$</th>
<th>$\beta_21$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>0.326</td>
<td>0.674</td>
<td>-0.038</td>
<td>1.908</td>
<td>0.699</td>
<td>0.056</td>
</tr>
<tr>
<td>Hard</td>
<td>0.311</td>
<td>0.689</td>
<td>0.049</td>
<td>0.947</td>
<td>1.904</td>
<td>0.079</td>
</tr>
</tbody>
</table>

The standard errors (se) for the MSE's of the parameter estimation are shown in the parentheses.
Figure 1: 3D scatter plot of one simulated data set in Model 1. Two regression planes created by \( \text{RM}^2(\ell_0) \) from two different components are displayed and 5% outliers are marked in blue.
Figure 2: 3D scatter plot of one simulated data set in Model 2. Two regression planes created by $RM^2(\ell_0)$ from two different components are displayed and 5% outliers are marked in blue.