# A Robust High-Dimensional Estimation of Multinomial Mixture Models

Azam Sabbaghi, Farzad Eskandari*, Hamid Reza Navabpoor

*Department of Statistics, Faculty of Mathematical Sciences and Computer, Allameh Tabataba'i University, Tehran, Iran*

## ARTICLE INFO

## ABSTRACT

In this paper, we are concerned with a robustifying high-dimensional (RHD) structured estimation in finite mixture of multinomial models. This method has been used in many applications that often involve outliers and data corruption. Thus, we introduce a class of the multinomial logistic mixture models for dependent variables having two or more discrete categorical levels. Through the optimization with the expectation maximization (EM) algorithm, we study two distinct ways to overcome sparsity in finite mixture of the multinomial logistic model; i.e., in the parameter space, or in the output space. It is shown that the new method is consistent for RHD structured estimation. Finally, we will implement the proposed method on real data.

## 1. INTRODUCTION

The Bernoulli mixture model (BMM) is applied for a binary dependent variable and showing how the model is estimated using the regularized maximum likelihood. The development and application of BMMs have gained increasing attention. Grantham [1] focused on BMMs for binary data clustering. Grilli *et al*. [2] used a binomial finite mixture to model the number of credits. Melkersson and Saarela [3] applied the binomial finite mixture model for nonzero counts. Brooks *et al*. [4] studied fetal deaths in litters with different types of finite mixture models including binomial finite mixture model.

The high-dimensional estimation under an additional sparse error vector for computing corrupted observations in recent studies has been widely considered (Wang *et al*. [5]; Nguyen and Tran [6]; Chen *et al*. [7]; Tibshirani and Manning [8]). Yang *et al*. [9] added an outlier error parameter for modeling the corrupted response. They applied two techniques for outlier modeling in GLMs. The first approach is in the parameter space, which is a convex optimization approach under stringent conditions. The second, which is in the output space, yields the nonconvex method with milder conditions. In this study, these two outlier modelings were used in binomial finite mixture modeling. We also used the multinomial logistic mixture models (MLMMs) to examine the problem of data corruption. Finally, the expectation maximization (EM) algorithm was considered for robust estimation.

The rest of this article is organized as follows: Section 2. introduces the Bernoulli finite mixture model (BMM) framework for binary data. Section 3. describes the MLMMs for dependent variables having two or more discrete categorical levels. In Section 4. we study properties of our approach by using the proposed method on real data.

## 2. MODELING OUTLIER ERRORS IN BMMs

In this paper, we examine the classification data set, the response variable $y_c$ of this set consisting of two classes -1 and 1, which are considered diametrically opposite such as pass/fail, win/lose, alive/dead. Bernoulli distribution is an effective method for studying the grouping variables. Assume that $P(Y_c = 1) = 1 - P(Y_c = -1) = p$. So, it can be expressed that $P(Y_c = y_c) = p^{I_1(Y_c)}(1 - p)^{1 - I_1(Y_c)}$.

A binary logistic model has a dependent variable with two possible values that are expressed by an indicator variable $y = I_1(y_c)$, where the two values are labeled as 0 and 1. Eskandari and Meshkani [10] presented the maximum likelihood equations from the probability distribution of the logistic regression and solved them using the Newton-Raphson method for nonlinear systems of equations. Biohning [11] applied the

---

Corresponding author. Email: askandari@atu.ac.ir

lower bound principle in the Newton-Raphson iteration instead of the Hessian matrix, which led to a monotonically converging sequence of iterates.

In real-world problems, we are interested in studying the logistic regression model in high-dimensional data problems with a small number of nonzero observations. Due to the presence of sparse parameters vector and outliers, the desirable theoretical properties of standard methods do not hold exactly. To solve this challenge and obtain a robust estimator for the lack of our model assumptions, we propose modeling outlier errors on BMM. Detail of the two approaches (i.e., modeling outlier errors in the parameter space and output space respectively) on Bernoulli mixture models will be explained later. Before that, we discuss the performance of the standard $\ell_1$ penalized BMMs over the uncorrupted version of the dataset. Let $y = (y_1, \cdots, y_n)$ be a random sample of binary vectors. We consider $y_i$ arises from a finite mixture density $p(y_i|\Psi) = \sum_{k=1}^{K} \pi_k p(y_i|p_{ik})$ of order K, where the mixture component density $p(y_i|p_{ik})$ is Bernoulli with success probability of $p_{ik}$:

$$p(y_i|\Psi) = \sum_{k=1}^{K} \pi_k \exp \left\{ y_i \ln \left( \frac{p_{ik}}{1 - p_{ik}} \right) + \ln(1 - p_{ik}) \right\} \tag{1}$$

where $\tilde{\Psi} = (\theta_1^*, ..., \theta_K^*, \pi)$ is a vector of mixture parameters and $\pi = (\pi_1, ...\pi_K)$ are mixing weights, such that $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$. In logistic regression, equate the logit function to the linear component of a covariate vector $x_i \in \mathbb{R}^p$ for the ith observation and the true regression parameter vector $\theta^* \in \mathbb{R}^p$ as follows:

$$\ln \left( \frac{p_{ik}}{1 - p_{ik}} \right) = \left\langle \theta_k^*, x_i \right\rangle = x_i^t \theta_k^*, \ i = 1, \cdots, n$$

Since the BMMs are presented in high-dimensional data problems, we assume $p$ is significantly larger than $n$. An $\ell_1$-penalized version of the maximum likelihood estimator (MLE) is defined to cope with observations that deviate from the true model. Suppose that $\{(x_i, y_i)\}_{i=1}^{n}$ is an independent sample of observations from (1). The negative log-likelihood function over $n$ values is given by

$$l_p = -\frac{1}{n} \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \exp \left\{ y_i \left\langle \theta_k^*, x_i \right\rangle - \ln \left( 1, +, \exp, \left\langle, \theta_k^*, ,, x_i, \right\rangle \right) \right\} \right\} \tag{2}$$

Now we estimate $\hat{\Theta} = (\hat{\theta}_1^*, \cdots, \hat{\theta}_K^*)$ by imposing the $\ell_1$ regularized maximum likelihood constrained proposed by Stadler *et al.* [12]:

$$\hat{\theta}_K^* \in \ \operatorname{argmin} \ l_p + \lambda_{n,\theta} \sum_{k=1}^{K} \pi_k \parallel \theta_k^* \parallel_1 \tag{3}$$
$$\parallel \theta_k^* \parallel_2 \leq a_0$$

where $a_0$ denotes a constant independent of $n$ and $p$. Note that $\parallel . \parallel_1$ for a vector is sum of absolute values and $\parallel . \parallel_2$ is the usual Euclidean norm. To compute this parameter, we proposed an iterative EM algorithm. At iteration $m$, the algorithm consists of Expectation (E-step) and Maximization (M-step) and seeks minimization of (2) using the complete negative log-likelihood function:

$$l^c(\tilde{\Psi}) = -\frac{1}{n} \sum_{n}^{i=1} \sum_{K}^{k=1} v_{ik} \left\{ \log \pi_k + y_i \left\langle \theta_k^*, x_i \right\rangle - \ln \left( 1 + \exp \left\langle \theta_k^*, x_i \right\rangle \right) \right\} + \lambda_{n,\theta} \sum_{K}^{k=1} \pi_k \parallel \theta_k^* \parallel_1 \tag{4}$$

Based on the following Algorithm 1 for BMM, Steps 3 marks the E-step of the algorithm, where $\omega_{ik}^{(m)}$ is updated by $E[v_{ik}|x,y]$ where unobserved imaginary indicator variables $v_{ik}$ show the component membership of the ith observation in the model. The conditional expectation of $l^c(\tilde{\Psi})$ with respect to $v_{ik}$ is

$$Q(\tilde{\Psi}, \tilde{\Psi}^{(m)}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \omega_{ik}^{(m)} \left\{ y_i \left\langle \theta_k^*, x_i \right\rangle - \ln \left( 1 + \exp \left\langle \theta_k^*, x_i \right\rangle \right) \right\} - \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \omega_{ik}^{(m)} \log \pi_k + \lambda_{n,\theta} \sum_{k=1}^{K} \pi_k \parallel \theta_k^* \parallel_1 \tag{5}$$

Steps 4 and 5 show the M-step, where $\Psi^{(m+1)}$ is obtained by minimizing (5) with respect to $\Psi$. Algorithm 1:

To motivate robust high-dimensional estimators, we begin with modeling the outlier errors approach on the parameter space in the next section.

**Algorithm 1:** EM Algorithm for BMM

step 1: Begin with initial values $\theta_k^{*0}$ and $\pi_k^0$ for all $k$.

step 2: Compute $P_k^{(m)} = (p_{1k}^{(m)}, \cdots, p_{nk}^{(m)})$ for all $i$'s and $k$'s as follows:

$$p_{ik}^{(m)} = \frac{\exp\left\{\left\langle \theta_k^{*(m)}, x_i \right\rangle\right\}}{1 + \exp\left\{\left\langle \theta_k^{*(m)}, x_i \right\rangle\right\}}$$

step 3: Compute $\omega_k^{(m)} = (\omega_{1k}^{(m)}, \cdots, \omega_{nk}^{(m)})$ for all $i$'s and $k$'s as follows:

$$\omega_{ik}^{(m)} = \frac{\pi_k \exp\left\{ y_i \left\langle \theta_k^{*(m)}, x_i \right\rangle - \ln\left(1 + \exp\left\langle \theta_k^{*(m)}, x_i \right\rangle\right)\right\}}{\sum\limits_{l=1}^{K} \pi_l \exp\left\{ y_i \left\langle \theta_l^{*(m)}, x_i \right\rangle - \ln\left(1 + \exp\left\langle \theta_l^{*(m)}, x_i \right\rangle\right)\right\}}$$

step 4: Compute $\hat{\theta}_k^*$ via the EM algorithm and the following equations:

$$\theta_k^{(m+1)} = \theta_k^{(m)} + (X'(diag(\omega_k^{(m)})P_k^{(m)}(1 - P_k^{(m)}))X)^{-1} X' diag(\omega_k^{(m)})(y - P_k^{(m)}) \tag{6}$$

where $X = (x_1', \cdots, x_n')$

step 5: Determine $\pi_k^{(m+1)}$ via under formula (for all $k$).

$$\pi_k^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} \omega_{ik}^{(m)}$$

step 6: Assign $m + 1 \leftarrow m + 2$ and iterate steps 2, 3 and 4 until reaching a predefined convergence criterion.

## 2.1. Parameter Space

Based on the $i$-th response $y_i$ that is drawn from (1) and the reformulation of logit function with a corrupted parameter $\left\langle \theta_k^*, x_i \right\rangle + \sqrt{n}e_{ik}^*$, we propose the robust estimators for general high-dimensional problems by modeling outlier errors in the parameter space. We can then write down the negative log-likelihood as

$$l_p = -\frac{1}{n} \sum_{i=1}^{n} \log\left\{ \sum_{k=1}^{K} \pi_k \exp\left\{ y_i \left\langle \theta_k^*, x_i \right\rangle + \sqrt{n}e_{ik}^*) - \ln\left(1 + \exp\left\{ \left\langle \theta_k^*, x_i \right\rangle + \sqrt{n}e_{ik}^* \right\}\right)\right\}\right\} \tag{7}$$

The robust estimator problem can be solved with the following constrained $\ell_1$ regularized maximum likelihood where $a_0, b_0$ are constants independent of $n$ and $p$.

$$(\hat{\theta}_k^*, \hat{e}_k^*) \in \; \arg\min \; l_p + \lambda_{n,\theta} \sum_{k=1}^{K} \pi_k \| \theta_k^* \|_1 + \lambda_{n,e} \sum_{k=1}^{K} \pi_k \| e_k^* \|_1$$
$$\| \theta_k^* \|_2 \le a_0$$
$$\| e_k^* \|_2 \le \frac{b_0}{\sqrt{n}} \tag{8}$$

We now focus on the EM algorithm and provide the complete negative log-likelihood function as follows:

$$l^c(\tilde{\Psi}) = \lambda_{n,\theta} \sum_{K}^{k=1} \pi_k \| \theta_k^* \|_1 + \lambda_{n,e} \sum_{K}^{k=1} \pi_k \| e_k^* \|_1 - \frac{1}{n}$$
$$\sum_{n}^{i=1} \sum_{K}^{k=1} v_{ik} \left\{ \log \pi_k + y_i \left( \left\langle \theta_k^*, x_i \right\rangle + \sqrt{n}e_{ik}^* \right) - \ln\left(1 + \exp\left\{ \left\langle \theta_k^*, x_i \right\rangle + \sqrt{n}e_{ik}^* \right\}\right)\right\} \tag{9}$$

In the E-step, the conditional expectation of $l^c(\tilde{\Psi})$ with respect to $v_{ik}$ given the data $(x_i, y_i)$ is

$$Q(\tilde{\Psi}, \tilde{\Psi}^{(m)}) = \lambda_{n,\theta} \sum_{k=1}^{K} \pi_k \parallel \theta_k^* \parallel_1 + \lambda_{n,e} \sum_{k=1}^{K} \pi_k \parallel e_k^* \parallel_1 - \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \omega_{ik}^{(m)} \log \pi_k$$
$$- \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \omega_{ik}^{(m)} \left\{ y_i \left( \langle \theta_k^*, x_i \rangle + \sqrt{n} e_{ik}^* \right) - \ln \left( 1 + \exp \left\{ \langle \theta_k^*, x_i \rangle + \sqrt{n} e_{ik}^* \right\} \right) \right\}$$

(10)

By adding the estimation of outlier errors parameter, Algorithm 2 is obtained. In the M-step, the estimates of $\tilde{\boxtimes} = (e_1^*, \cdots, e_K^*, \theta_1^*, ..., \theta_K^*, \pi)$ are updated by Steps 4–6.

---

**Algorithm 2:** EM Algorithm for modeling errors in the parameter space in BMM (PBMM)

step 1: Begin with initial values $\theta_k^{*0}$, $\pi_k^0$ and $e_k^{*0}$ for all $k$'s.

step 2: Compute $P_k^{(m)} = (p_{1k}^{(m)}, \cdots, p_{nk}^{(m)})$ at the mth itration for all $i$'s and $k$'s as follows:

$$p_{ik}^{(m)} = \frac{\exp \left\{ \langle \theta_k^{*(m)}, x_i \rangle + \sqrt{n} e_{ik}^* \right\}}{1 + \exp \left\{ \langle \theta_k^{*(m)}, x_i \rangle + \sqrt{n} e_{ik}^* \right\}}$$

step 3: Compute $\omega_k^{(m)} = (\omega_{1k}^{(m)}, \cdots, \omega_{nk}^{(m)})$ for all $i$'s and $k$'s:

$$\omega_{ik}^{(m)} = \frac{\pi_k \exp \left\{ y_i (\langle \theta_k^*, x_i \rangle + \sqrt{n} e_{ik}^*) - \ln(1 + \exp\left\{ \langle \theta_k^*, x_i \rangle + \sqrt{n} e_{ik}^* \right\}) \right\}}{\sum_{l=1}^{K} \pi_l \exp\{y_i(\langle \theta_l^*, x_i \rangle + \sqrt{n} e_{il}^*) - \ln(1 + \exp\{\langle \theta_l^*, x_i \rangle + \sqrt{n} e_{il}^*\})\}}$$

step 4: Obtain $\theta_k^{*(m+1)}$ from the EM algorithm and following equation:

$$\theta_k^{*(m+1)} = \theta_k^{*(m)} + (X'(diag(\omega_k^{(m)})P_k^{(m)}(1 - P_k^{(m)}))X)^{-1} X' diag(\omega_k^{(m)})(y - P_k^{(m)})$$

Where $X = (x_1', \cdots, x_n')$

step 5: Obtain $e_k^{*(m+1)}$ via the EM algorithm as follows:

$$e_k^{*(m+1)} = e_k^{*(m)} + (\sqrt{n} P_k^{(m)}(1 - P_k^{(m)}))^{-1} (y - P_k^{(m)})$$

step 6: Assign $\pi_k^{(m+1)} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \omega_{ik}^{(m)}$.

step 7: Consider $m + 1 \leftarrow m + 2$ and iterate steps 2-6 until reaching a predefined convergence criterion.

---

Although the optimization problem is convex, stringent conditions are needed to achieve a consistent estimator. The constraint $\parallel e_k^* \parallel_2 \leq \frac{b_0}{\sqrt{n}}$ ensures that the consistent optimum exists as discussed in Yang *et al.* [9]. To investigate the larger errors, we introduce modeling errors in the output space.

## 2.2. Output Space

In this section, we give statistical error directly in the response space of BMM (RBMM). Under the certain assumption, the group random variable $y_{ic} - \sqrt{n} e_i^*$ has two values $1 - \sqrt{n} e_i^*$ and $-1 - \sqrt{n} e_i^*$. Therefore, based on indicator variable $y_i' = I_{1-\sqrt{n} e_i^*}(y_{ic} - \sqrt{n} e_i^*)$ drawn from the conditional distribution in (1) with logit function $\langle \theta_k^*, x_i \rangle$ for each component, we have

$$p(y_i | x_i) = \sum_{k=1}^{K} \pi_k \exp \left\{ I_{1-\sqrt{n} e_i^*}(y_{ic} - \sqrt{n} e_i^*) \langle \theta_k^*, x_i \rangle - \ln \left( 1 + \exp \langle \theta_k^*, x_i \rangle \right) \right\}$$

Our starting point is the following negative log-likelihood function:

$$l_p = -\frac{1}{n} \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \exp \left\{ I_{1-\sqrt{n} e_i^*}(y_{ic} - \sqrt{n} e_i^*) \langle \theta_k^*, x_i \rangle - \ln \left( 1 + \exp \langle \theta_k^*, x_i \rangle \right) \right\} \right\}$$

(11)

The EM iteration alternates between performing E- and M-steps. The E-step creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters:

$$l^c(\tilde{\Psi}) = -\frac{1}{n} \sum_{i=1}^{L} \sum_{k=1}^{K} v_{ik} \left\{ \log \pi_k + y_i' \left\langle \theta_k^*, x_i \right\rangle - \ln\left(1 + \exp\left\langle \theta_k^*, x_i \right\rangle\right) \right\} \tag{12}$$

The M-step computes parameters maximizing the expected log-likelihood found on the E-step.

$$\begin{aligned}
Q = &-\frac{1}{n} \sum_{i=1}^{L} \sum_{k=1}^{K} \omega_{ik}^{(m)} \log \pi_k \\
&-\frac{1}{n} \sum_{i=1}^{L} \sum_{k=1}^{K} \omega_{ik}^{(m)} \left\{ y_i' \left\langle \theta_k^*, x_i \right\rangle - \ln\left(1 + \exp\left\langle \theta_k^*, x_i \right\rangle\right) \right\}
\end{aligned} \tag{13}$$

To this end, in an RBMM approach, iterate Algorithm 1 for each value in the set $L \in \{2, 3, \cdots, n\}$ and choose the parameters estimates that have the least empirical prediction errors.

## 3. MODELING OUTLIER ERRORS IN THE MLMM

In the MLMM, we consider vector $y_i = (y_{i1}, \cdots, y_{iJ})$, $i = 1, \cdots, n$, with $y_{ij} = 0$ for all $j$ except one $j'$ with $y_{ij'} = 1$ and corresponding probability $p_{ij'}$. Let $J$ represent the number of levels of the dependent variable. To simulate sparsity, we arbitrarily corrupt some of the observations $y_i$. For logistic regression models, the corrupted response $y_i$ is obtained by $y_i = (1 - \bar{y}_i)$, which suggests

$$E(y_i) = \begin{pmatrix} p_{i1} \\ p_{i2} \\ p_{i3} \end{pmatrix}, \quad Cov(y_i) = \begin{pmatrix} p_{i1}(1 - p_{i1}) & -p_{i1}p_{i2} & -p_{i1}p_{i3} \\ -p_{i1}p_{i2} & p_{i2}(1 - p_{i2}) & -p_{i2}p_{i3} \\ -p_{i1}p_{i3} & -p_{i2}p_{i3} & p_{i3}(1 - p_{i3}) \end{pmatrix}$$

Recall that $y_i$ follows a finite mixture of multinomial model of order K with the conditional density function as

$$p(y_i|x_i) = \sum_{k=1}^{K} \pi_k \exp\left\{ y_{i1} \ln\left(\frac{p_{i1k}}{p_{i3k}}\right) + y_{i2} \ln\left(\frac{p_{i2k}}{p_{i3k}}\right) + \ln(p_{i3k}) \right\} \tag{14}$$

where the multinomial logit-model is given by

$$\ln\left(\frac{p_{i1k}}{p_{i3k}}\right) = \left\langle \theta_{1k}^*, x_i \right\rangle, \quad \ln\left(\frac{p_{i2k}}{p_{i3k}}\right) = \left\langle \theta_{2k}^*, x_i \right\rangle \tag{15}$$

The conditional log-likelihood function of $\tilde{\Psi}$ has the form

$$l_p = -\frac{1}{n} \sum_{i=1}^{n} \log\left[ \sum_{k=1}^{K}, \pi_k, \exp, \left\{ y_{i1} \left\langle \theta_{1k}^*, x_i \right\rangle + yi1 \left\langle \theta_{2k}^*, x_i \right\rangle \text{-} \ln(1 + \exp\left\langle \theta_{1k}^*, x_i \right\rangle + \left\langle \theta_{2k}^*, x_i \right\rangle) \right\} \right] \tag{16}$$

The $\ell_1$ penalized version of the classical MLE is as follows:

$$\begin{aligned}
(\hat{\theta}_{1k}^*, \hat{\theta}_{2k}^*) \in \ &\operatorname{argmin} l_p + \lambda_{n,\theta_1} \sum_{k=1}^{K} \pi_k \parallel \theta_{1k}^* \parallel_1 + \lambda_{n,\theta_2} \sum_{k=1}^{K} \pi_k \parallel \theta_{2k}^* \parallel_1 \\
&\parallel \theta_{1k}^* \parallel_2 \leq a_{10} \\
&\parallel \theta_{2k}^* \parallel_2 \leq a_{20}
\end{aligned} \tag{17}$$

In this situation, the joint estimation of $(\theta_{1k}^*, \theta_{2k}^*)$ is achieved through the complete-case framework required by the EM algorithm.

$$l^c(\tilde{\Psi}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} v_{ik} \left\{ \log \pi_k + y_{i1} \left\langle \theta_{1k}^*, x_i \right\rangle + y_{i2} \left\langle \theta_{2k}^*, x_i \right\rangle - \ln\left(1 + \exp\left\langle \theta_{1k}^*, x_i \right\rangle + \exp\left\langle \theta_{2k}^*, x_i \right\rangle\right) \right\} \tag{18}$$

where $\tilde{\Psi} = (\theta_{11}^*, ..., \theta_{1K}^*, \theta_{21}^*, ..., \theta_{2K}^*, \pi)$. By using (18) we can rewrite the new conditional expectation of $l^c(\tilde{\Psi})$:

$$Q(\tilde{\Psi}, \tilde{\Psi}^{(m)}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}^{(m)}\log \pi_k - \frac{1}{n}\sum_{i=1}^{n}$$

$$\sum_{k=1}^{K}\omega_{ik}^{(m)}\left\{y_{i1}\left\langle \theta_{1k}^*, x_i\right\rangle + y_{i2}\left\langle \theta_{2k}^*, x_i\right\rangle - \ln\left(1 + \exp\left\langle \theta_{1k}^*, x_i\right\rangle + \exp\left\langle \theta_{2k}^*, x_i\right\rangle\right)\right\}$$

The following Algorithm 3 illustrates the procedure employed for estimating the parameters:

---

**Algorithm 3:** EM Algorithm for multinomial logistic mixture models

---

step 1: Begin with initial values $(\theta_{1k}^{*\ 0}, \theta_{2k}^{*\ 0})$ and $\pi_k^0$ for all $k$.

step 2: Compute $(P_{1k}^{(m)}, P_{2k}^{(m)}, P_{3k}^{(m)})$ for all $i$'s and $k$'s as follow:

$$p_{irk}^{(m)} = \frac{\exp\left\{\left\langle \theta_{rk}^{*\ (m)}, x_i\right\rangle\right\}}{1 + \sum_{r=1}^{2}\exp\left\{\left\langle \theta_{rk}^{*\ (m)}, x_i\right\rangle\right\}}, \quad r = 1, 2$$

$$p_{i3k}^{(m)} = \frac{1}{1 + \sum_{r=1}^{2}\exp\left\{\left\langle \theta_{rk}^{*\ (m)}, x_i\right\rangle\right\}}$$

step 3: Compute $\omega_k^{(m)} = (\omega_{1k}^{(m)}, \cdots, \omega_{nk}^{(m)})$ for all $i$'s and $k$'s:

$$\omega_{ik}^{(m)} = \frac{\pi_k^{(m)}\exp\left\{\sum_{r=1}^{2}y_{ir}\left\langle \theta_{rk}^{*\ (m)}, x_i\right\rangle - \ln\left(1 + \sum_{r=1}^{2}\exp\left\langle \theta_{rk}^{*\ (m)}, x_i\right\rangle\right)\right\}}{\sum_{l=1}^{K}\pi_l^{(m)}\exp\left\{\sum_{r=1}^{2}y_{ir}\left\langle \theta_{rl}^{*(m)}, x_i\right\rangle - \ln\left(1 + \sum_{r=1}^{2}\exp\left\langle \theta_{rl}^{*(m)}, x_i\right\rangle\right)\right\}}$$

step 4: Compute $(\hat{\theta}_{1k}^*, \hat{\theta}_{2k}^*)$ from the EM algorithm and the following equations:

$$\begin{pmatrix}\theta_{1k}^{(m+1)} \\ \theta_{2k}^{(m+1)}\end{pmatrix} = \begin{pmatrix}\theta_{1k}^{(m)} \\ \theta_{2k}^{(m)}\end{pmatrix} + \begin{pmatrix}X'W_1^{(m)}X & X'V_{1.2}^{(m)}X \\ X'V_{1.2}^{(m)}X & X'W_2^{(m)}X\end{pmatrix}^{-1}\begin{pmatrix}X'diag(\omega_k^{(m)})(y_{1k} - P_{1k}^{(m)}) \\ X'diag(\omega_k^{(m)})(y_{2k} - P_{2k}^{(m)})\end{pmatrix}$$

where $X = (x_1', \cdots, x_n')$ and

$$W_1^{(m)} = diag(\omega_k^{(m)}P_{1k}^{(m)}(1 - P_{1k}^{(m)}))$$

$$W_2^{(m)} = diag(\omega_k^{(m)}P_{2k}^{(m)}(1 - P_{2k}^{(m)}))$$

$$V_{1.2}^{(m)} = diag(-\omega_k^{(m)}P_{1k}^{(m)}P_{2k}^{(m)})$$

step 5: Determine $\pi_k^{(m+1)}$ from the following formula (for all $k$).

$$\pi_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n}\omega_{ik}^{(m)}$$

step 6: Assign $m + 1 \leftarrow m + 2$ and iterate steps 2, 3 and 4 until reaching a predefined convergence criterion.

---

We studied the standard $\ell_1$ penalized finite mixture of the multinomial logistic model over the corrupted data. We also introduce two other methods, modeling outlier errors in the parameter space and output space, respectively. Finally, we will compare the performance of these three methods.

## 3.1. Parameter Space

As in the previous section, we assume that $y_i$ follows the conditional distribution of (14) with the new multinomial logit-model:

$$\ln\left(\frac{p_{i1k}}{p_{i3k}}\right) = \langle \theta_{1k}^*, x_i \rangle + \sqrt{n}e_{ik}^*, \quad \ln\left(\frac{p_{i2k}}{p_{i3k}}\right) = \langle \theta_{2k}^*, x_i \rangle + \sqrt{n}e_{ik}^*$$

We can rewrite the log-likelihood function as follows:

$$l_p = -\frac{1}{n}\sum_{i=1}^{n}\log\left[\sum_{k=1}^{K}\right.$$
$$\left.\pi_k, \exp\left\{\sum_{r=1}^{2}y_{ir}\left(\langle\theta_{rk}^*, x_i\rangle + \sqrt{n}e_{ik}^*\right) - \ln(1 + \sum_{r=1}^{2}\exp\left\{\langle\theta_{rk}^*, x_i\rangle + \sqrt{n}e_{ik}^*\right\})\right)\right\}$$

(19)

A penalized log-likelihood function is defined as

$$(\hat{\theta}_{1k}^*, \hat{\theta}_{2k}^*, \hat{e}_k^*) \in$$
$$\operatorname*{argmin} l_p + \lambda_{n,\theta_1}\sum_{k=1}^{K}\pi_k \parallel \theta_{1k}^* \parallel_1 + \lambda_{n,\theta_2}\sum_{k=1}^{K}\pi_k \parallel \theta_{2k}^* \parallel_1 + \lambda_{n,e}\sum_{k=1}^{K}\pi_k \parallel e_k^* \parallel_1$$
$$\parallel \theta_{1k}^* \parallel_2 \leq a_{10}$$
$$\parallel \theta_{2k}^* \parallel_2 \leq a_{20}$$
$$\parallel e_k^* \parallel_2 \leq \frac{b_0}{\sqrt{n}}$$

(20)

where $\ell_1$-norm penalty function is

$$p_n(\tilde{\Psi}) = \lambda_{n,\theta_1}\sum_{k=1}^{K}\pi_k \parallel \theta_{1k}^* \parallel_1 + \lambda_{n,\theta_2}\sum_{k=1}^{K}\pi_k \parallel \theta_{2k}^* \parallel_1 + \lambda_{n,e}\sum_{k=1}^{K}\pi_k \parallel e_k^* \parallel_1$$

The complete log-likelihood function, after substituting $v_{ik}$, is

$$l^c(\tilde{\Psi}) = p_n(\tilde{\Psi}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}v_{ik}$$
$$\left\{\log\pi_k + \sum_{r=1}^{2}\left(y_{ir}\langle\theta_{rk}^*, x_i\rangle + \sqrt{n}e_{ik}^*\right) - \ln\left(1 + \sum_{r=1}^{2}\exp\{\langle\theta_{1k}^*, x_i\rangle + \sqrt{n}e_{ik}^*\}\right)\right\}$$

(21)

where $\tilde{\Psi} = (e_1^*, \cdots, e_K^*, \theta_{11}^*, ..., \theta_{1K}^*, \theta_{21}^*, ..., \theta_{2K}^*, \pi)$. Note that after taking the conditional expectation of (21), we have

$$Q(\tilde{\Psi}, \tilde{\Psi}^{(m)}) = p_n(\tilde{\Psi}) - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\omega_{ik}^{(m)}\log\pi_k - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}$$
$$\omega_{ik}^{(m)}\left\{\sum_{r=1}^{2}\left(y_{ir}\langle\theta_{rk}^*, x_i\rangle + \sqrt{n}e_{ik}^*\right) - \ln\left(1 + \sum_{r=1}^{2}\exp\left\{\langle\theta_{1k}^*, x_i\rangle + \sqrt{n}e_{ik}^*\right\}\right)\right\}$$

(22)

To obtain the EM algorithm 4, we use the Newton-Raphson method, which involves calculating the first and second derivatives of (22).

## 3.2. Output Space

In the response space corrupted data for MLMM (RMLMM), the dependent variable $y_{ic} - \sqrt{n}e_i^*$ has three levels. Therefore, we have the conditional distribution as follows:

$$p(y_i|x_i) = \sum_{k=1}^{K}\pi_k \exp$$
$$\{I_{1-\sqrt{n}e_i^*}(y_{ic} - \sqrt{n}e_i^*)\ln(\frac{p_{i1k}}{p_{i3k}}) + I_{-1-\sqrt{n}e_i^*}(y_{ic} - \sqrt{n}e_i^*)\ln(\frac{p_{i2k}}{p_{i3k}}) + \ln(p_{i3k})\}$$

(23)

---

**Algorithm 4:** EM Algorithm for modeling errors in the parameter space on multinomial logistic mixture models (PMLMM)

step 1: Begin with initial values $(\theta_{1k}^{*\,0}, \theta_{2k}^{*\,0}, e_k^{*0})$ and $\pi_k^0$ for all $k$'s.

step 2: Compute $(P_{1k}^{(m)}, P_{2k}^{(m)}, P_{3k}^{(m)})$ for all $i$'s and $k$'s as follows:

$$p_{irk}^{(m)} = \frac{\exp\left\{ \left\langle \theta_{rk}^{*\,(m)}, x_i \right\rangle + \sqrt{n} e_{ik}^* \right\}}{1 + \sum_{r=1}^{2} \exp\left\{ \left\langle \theta_{rk}^{*\,(m)}, x_i \right\rangle + \sqrt{n} e_{ik}^* \right\}}, \quad r = 1, 2$$

$$p_{i3k}^{(m)} = \frac{1}{1 + \sum_{r=1}^{2} \exp\left\{ \left\langle \theta_{rk}^{*\,(m)}, x_i \right\rangle + \sqrt{n} e_{ik}^* \right\}}$$

step 3: Compute $\omega_k^{(m)} = (\omega_{1k}^{(m)}, \cdots, \omega_{nk}^{(m)})$ for all $i$'s and $k$'s as follows:

$$\omega_{ik}^{(m)} = \frac{\pi_k^{(m)} \exp\left\{ \sum_{r=1}^{2} y_{ir} \left\langle \theta_{rk}^{*\,(m)}, x_i \right\rangle - \ln\left( 1 + \sum_{r=1}^{2} \exp\left\langle \theta_{rk}^{*\,(m)}, x_i \right\rangle \right) \right\}}{\sum_{l=1}^{K} \pi_l^{(m)} \exp\left\{ \sum_{r=1}^{2} y_{ir} \left\langle \theta_{rl}^{*(m)}, x_i \right\rangle - \ln\left( 1 + \sum_{r=1}^{2} \exp\left\langle \theta_{rl}^{*(m)}, x_i \right\rangle \right) \right\}}$$

step 4: Compute $(\hat{\theta}_{1k}^*, \hat{\theta}_{2k}^*)$ via the EM algorithm and following equations:

$$\begin{pmatrix} \theta_{1k}^{(m+1)} \\ \theta_{2k}^{(m+1)} \end{pmatrix} = \begin{pmatrix} \theta_{1k}^{(m)} \\ \theta_{2k}^{(m)} \end{pmatrix} + \begin{pmatrix} X' W_1^{(m)} X & X' V_{1.2}^{(m)} X \\ X' V_{1.2}^{(m)} X & X' W_2^{(m)} X \end{pmatrix}^{-1} \begin{pmatrix} X' diag(\omega_k^{(m)})(y_{1k} - P_{1k}^{(m)}) \\ X' diag(\omega_k^{(m)})(y_{2k} - P_{2k}^{(m)}) \end{pmatrix}$$

where $X = (x_1', \cdots, x_n')$ and

$$W_1^{(m)} = diag(\omega_k^{(m)} P_{1k}^{(m)}(1 - P_{1k}^{(m)}))$$

$$W_2^{(m)} = diag(\omega_k^{(m)} P_{2k}^{(m)}(1 - P_{2k}^{(m)}))$$

$$V_{1.2}^{(m)} = diag(-\omega_k^{(m)} P_{1k}^{(m)} P_{2k}^{(m)})$$

step 5: Obtaining $e_k^{*(m+1)}$ from the EM algorithm as follows:

$$e_k^{*(m+1)} = e_k^{*(m)} + (\sqrt{n} P_{3k}^{(m)}(1 - P_{3k}^{(m)}))^{-1}(y_{3k} - P_{3k}^{(m)})$$

step 6: Determine $\pi_k^{(m+1)}$ for all $k$ from the following formula.

$$\pi_k^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} \omega_{ik}^{(m)}$$

step 7: Assign $m + 1 \leftarrow m + 2$ and iterate steps 2, 3 and 4 until reaching a predefined convergence criterion.

---

We consider multinomial logit-model (15) and $y_i' = (y_{i1}', y_{i2}', y_{i3}')$, where

$$y_{i1}' = I_{1-\sqrt{n}e_i^*}(y_{ic} - \sqrt{n}e_i^*), \quad y_{i2}' = I_{-1-\sqrt{n}e_i^*}(y_{ic} - \sqrt{n}e_i^*)$$

Let the complete log-likelihood function be

$$l^c(\tilde{\Psi}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} v_{ik} \left\{ \log \pi_k + \sum_{r=1}^{2} y_{ir}' \left\langle \theta_{rk}^*, x_i \right\rangle - \ln\left( 1 + \sum_{r=1}^{2} \exp\left\langle \theta_{rk}^*, x_i \right\rangle \right) \right\} \tag{24}$$

Using the log-likelihood function, it is clear the errors in the model of the RMLMM will affect the total number the sum of observations present.

$$Q(\tilde{\Psi}, \tilde{\Psi}^{(m)}) = -\frac{1}{n} \sum_{i=1}^{L} \sum_{k=1}^{K} \omega_{ik}^{(m)} \left\{ \sum_{r=1}^{2} y_{ir} \left\langle \theta_{rk}^{*}, x_i \right\rangle - \ln \left( 1 + \exp \sum_{r=1}^{2} \left\langle \theta_{rk}^{*}, x_i \right\rangle \right) \right\} - \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \omega_{ik}^{(m)} \log \pi_k$$

Therefore, it is sufficient that we repeat Algorithm 3 with the different numbers of observations $L \in \{2, 3, \cdots, n\}$.

## 4. APPLICATION TO THE ANALYSIS OF REAL DATA

We consider two different finite mixture of logistic models: $M_1$ and $M_2$ models. $M_1$ model has two components whereas $M_2$ model has three components. We employ the real binary classification dataset from Yang *et al.* [9]. Australian dataset was obtained from LIBSVM (http//:www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/). The dataset consists of $n = 690$ units and the dimension of the true parameter is $p = 14$, therefore we can set $\lambda_{n,\theta} = 0$. That is, there is no need to add further sparsity $\ell_1$ regularization to the parameters according to $(p \langle n \rangle)$. We divide our dataset into two groups; 40% of data as the training and 60% as the test data. By scaling the number of corrupted samples $(r)$ of training examples $(m)$, we make various datasets.

We compare the performances of the standard $\ell_1$ penalized mixture models and modeling outlier errors in parameter and output space.

To evaluate the performance of our proposed method, we use the Ratio of Generalized Mean Square Error (RGMSE) index that is defined as the ratio of generalized mean square error (GMSE) of $M_1$ to $M_2$, i.e.,

$$\hat{G}MSE(\hat{\theta}_k^*) = (\hat{\theta}_k^* - \theta_k^*)\hat{E}(XX')(\hat{\theta}_k^* - \theta_k^*)$$

where $\hat{E}(XX') = \frac{1}{n} \sum_{i=1}^{n} x_i' x_i$. And, we have

$$\hat{R}GMSE_j = \frac{GMSE(M_{j1})}{GMSE(M_{j2})}, j = 1, ..., 2000$$

This quantity was computed in each run and then the mean of $\hat{R}GMSEs$ over 2,000 runs was reported in Table 1. The lower $\hat{R}GMSE$ shows that fitting the mixture of two populations $(M_1)$ is better than fitting it with 3 components $(M_2)$.

Figure 1 plots the RGMSE of the parameter estimates, against the number of samples $n$. We compare three methods:

(1) the standard $\ell_1$ penalized finite mixture of generalized linear models (FMGLM) over the corrupted data (BMM, $\ell_1$reg),

(2) our first M-estimator that models errors in the parameter space (error in parameters, PBMM),

(3) our second M-estimator, which models error in the output space (error in output, RBMM).

Each row shows different types of outliers on the dataset: (w/o) original dataset without adding outliers (i.e., Log, Sqrt and Linear), where the number of outliers $r$ scaled to three different ways as $(r = \log(m), \sqrt{m}, 0.1(m))$.

Each column shows three different fraction of training dataset: 25% (Left column), 50% (Center), and 100% (Right column).

Table 2 (model $M_1$) and Table 3 (model $M_2$) represent estimates and their standard deviations.

## 5. DISCUSSION

In this paper, for the modeling sparsity of the outlier response vector on the BMMs, we randomly have selected a small number of $r$ samples from $n$ observations, and corrupted them arbitrarily. We have considered the performance of the proposed method on Australian real binary

**Table 1** | Comparisons of the mean of RGMSEs under different models.

| $r$ | 25% | | | 50% | | | 1000% | | |
|---|---|---|---|---|---|---|---|---|---|
| | **BMM** | **PBMM** | **RBMM** | **BMM** | **PBMM** | **RBMM** | **BMM** | **PBMM** | **RBMM** |
| w/o | 0.6852 | 0.4112 | 0.3109 | 0.6135 | 0.4688 | 0.2969 | 0.0061 | 0.4629 | 0.4335 |
| Log | 0.0079 | 0.1498 | 0.2638 | 0.2970 | 0.3970 | 0.3467 | 0.2911 | 0.2920 | 0.1946 |
| Sqrt | 0.7193 | 0.0549 | 0.1168 | 0.2634 | 0.1067 | 0.2821 | 0.0673 | 0.2280 | 0.1181 |
| Linear | 0.1115 | 0.1503 | 0.1080 | 0.1055 | 0.1545 | 0.2747 | 0.0399 | 0.3961 | 0.1153 |

BMM, Bernoulli mixture model; PBMM, parameter space in BMM; RBMM, response space of BMM.
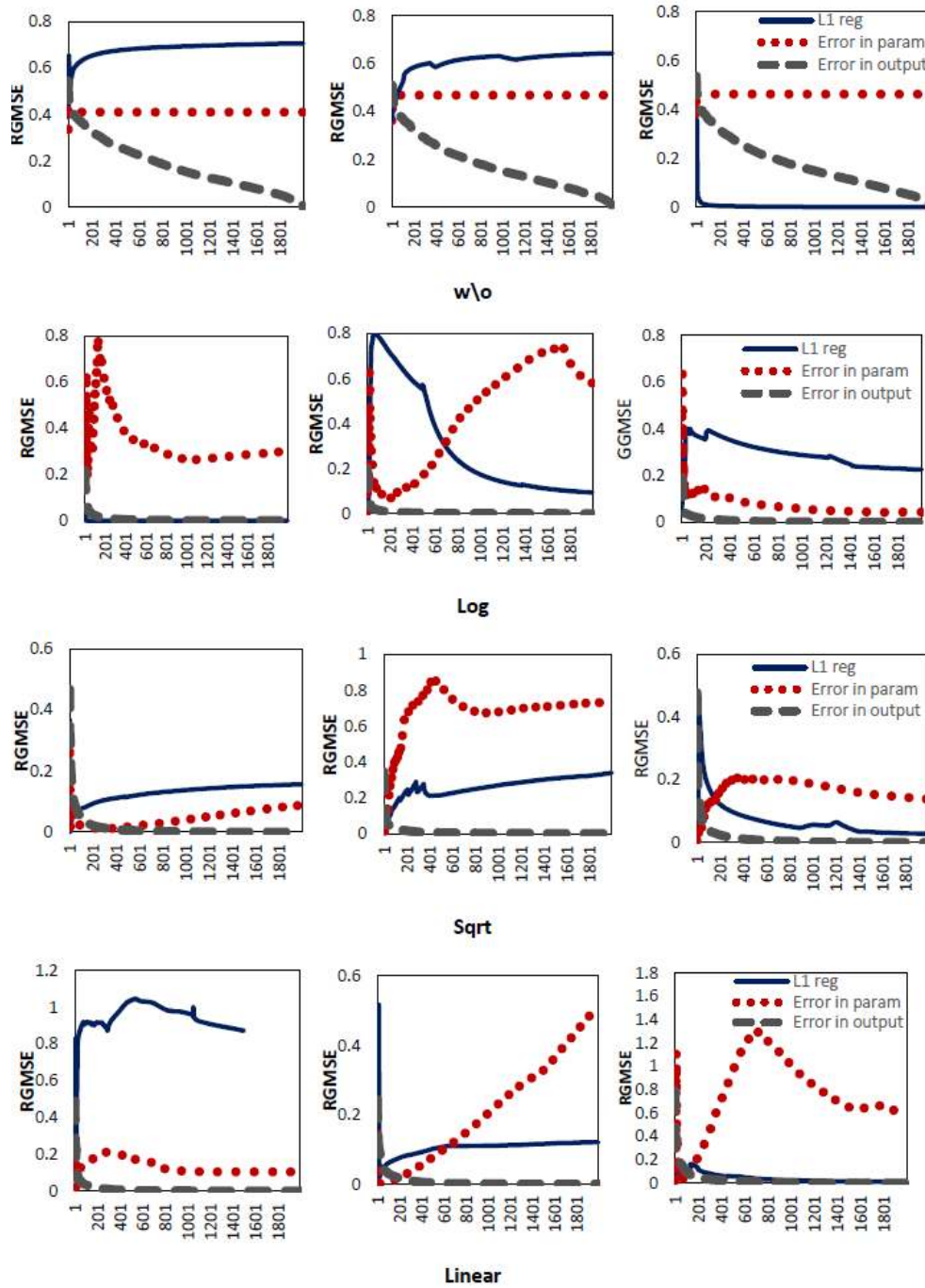
**Figure 1** | Comparison of RGMSE for different types of outliers; the share of the samples used in the training dataset: 25% (left column), 50% (Center), and 100%(right column).

**Table 2** | Estimates and their standard deviations based on 2,000 runs on the original dataset without adding artificial outliers.

| k | BMM | | PBMM | | OBMM | |
|---|---|---|---|---|---|---|
| | $\hat{\theta}$ | sd | $\hat{\theta}$ | sd | $\hat{\theta}$ | sd |
| | −1.1318 | 0.3898 | −2.4252 | 0.4315 | −1.1469 | 0.3962 |
| | −0.0339 | 0.0223 | −0.0819 | 0.0153 | −0.0343 | 0.0224 |
| | −0.0616 | 0.0385 | 0.0121 | 0.0324 | −0.0614 | 0.0386 |
| | 0.0186 | 0.4617 | 1.2399 | 0.3242 | 0.0285 | 0.4632 |

(*Continued*)

**Table 2** | Estimates and their standard deviations based on 2,000 runs on the original dataset without adding artificial outliers. (*Continued*)

| | | | | | | |
|---|---|---|---|---|---|---|
| | 0.1380 | 0.0491 | 0.2159 | 0.0457 | 0.1384 | 0.0490 |
| | −0.1279 | 0.0927 | 0.0399 | 0.0808 | −0.1278 | 0.0928 |
| 1 | 0.2291 | 0.0818 | 0.3160 | 0.0704 | 0.2301 | 0.0817 |
| | 2.0894 | 0.4519 | 4.6841 | 0.4094 | 2.1055 | 0.4628 |
| | −0.7119 | 0.6548 | −2.3642 | 0.5815 | −0.7503 | 0.6907 |
| | 0.1330 | 0.1288 | 0.3511 | 0.0894 | 0.1357 | 0.1307 |
| | −1.0179 | 0.3795 | −1.8398 | 0.3286 | −1.0262 | 0.3790 |
| | 0.2971 | 0.4797 | −1.0341 | 0.4926 | 0.2990 | 0.4802 |
| | −0.0016 | 0.0010 | −0.0034 | 0.0010 | −0.0016 | 0.0010 |
| | 0.0002 | 0.0009 | 0.0006 | 0.0002 | 0.0002 | 0.0009 |
| | 0.0113 | 0.3189 | −0.1207 | 0.3782 | 0.0113 | 0.3189 |
| | −0.0043 | 0.0148 | −0.0183 | 0.0177 | −0.0043 | 0.0148 |
| | −0.0023 | 0.0338 | 0.0392 | 0.0384 | −0.0023 | 0.0338 |
| | 0.0325 | 0.3526 | 1.0210 | 0.3784 | 0.0325 | 0.3526 |
| | 0.1339 | 0.0439 | 0.2152 | 0.0573 | 0.1339 | 0.0439 |
| | 0.1282 | 0.0820 | 0.4095 | 0.1049 | 0.1282 | 0.0820 |
| 2 | 0.0108 | 0.0560 | −0.0293 | 0.0704 | 0.0108 | 0.0559 |
| | 4.0875 | 0.4269 | 8.6061 | 0.8816 | 4.0875 | 0.4269 |
| | 1.2307 | 0.4384 | 1.0877 | 0.5550 | 1.2307 | 0.4384 |
| | 0.1042 | 0.0661 | 0.2849 | 0.1026 | 0.1043 | 0.0661 |
| | 0.16908 | 0.3102 | 0.1089 | 0.3531 | 0.1691 | 0.3102 |
| | −2.6615 | 0.4601 | −6.0474 | 0.7829 | −2.6615 | 0.4601 |
| | −0.0025 | 0.0012 | −0.0061 | 0.0013 | −0.0025 | 0.0012 |
| | 0.0009 | 0.0001 | 0.0019 | 0.0004 | 0.0009 | 0.0001 |

BMM, Bernoulli mixture model; PBMM, parameter space in BMM.

**Table 3** | Estimates and their standard deviations based on 2,000 runs on the original dataset without adding artificial outliers.

| k | BMM | | PBMM | | OBMM | |
|---|---|---|---|---|---|---|
| | $\hat{\theta}$ | *sd* | $\hat{\theta}$ | *sd* | $\hat{\theta}$ | *sd* |
| | −1.2010 | 0.3741 | −1.4134 | 0.5099 | −1.2166 | 0.3770 |
| | −0.0370 | 0.0223 | −0.0447 | 0.0239 | −0.0376 | 0.0212 |
| | −0.0422 | 0.0362 | 0.0279 | 0.0386 | −0.0426 | 0.0349 |
| | 0.2765 | 0.4392 | 0.4722 | 0.4717 | 0.3072 | 0.4206 |
| | 0.1517 | 0.0462 | 0.1629 | 0.0511 | 0.1494 | 0.0466 |
| | −0.0574 | 0.0874 | 0.0323 | 0.0946 | −0.0527 | 0.0864 |
| 1 | 0.1541 | 0.0756 | 0.1795 | 0.0852 | 0.1551 | 0.0730 |
| | 2.2690 | 0.4286 | 2.6910 | 0.9504 | 2.2529 | 0.4337 |
| | −1.2131 | 0.5075 | −1.4257 | 0.4287 | −1.2208 | 0.4859 |
| | 0.2220 | 0.0651 | 0.2452 | 0.0737 | 0.2192 | 0.0658 |
| | −1.1937 | 0.3391 | −1.3085 | 0.3758 | −1.2083 | 0.3281 |
| | 0.0563 | 0.4798 | −0.1952 | 0.6393 | 0.0511 | 0.4619 |
| | −0.0028 | 0.0009 | −0.0029 | 0.0009 | −0.0028 | 0.0009 |
| | 0.0003 | 0.0001 | 0.0003 | 0.0002 | 0.0003 | 0.0001 |
| | 0.1157 | 0.3198 | −0.0833 | 0.3290 | 0.1096 | 0.3275 |
| | −0.0104 | 0.0148 | −0.0125 | 0.0160 | −0.0103 | 0.0144 |
| | −0.0090 | 0.0351 | 0.0135 | 0.0365 | −0.0062 | 0.0329 |
| | 0.2715 | 0.3605 | 1.4226 | 0.4718 | 0.2767 | 0.3599 |
| | 0.1379 | 0.0446 | 0.1502 | 0.0555 | 0.1362 | 0.0447 |
| | 0.2050 | 0.0837 | 0.2431 | 0.1198 | 0.2042 | 0.0822 |
| 2 | 0.0534 | 0.0617 | −0.0488 | 0.0615 | −0.0521 | 0.0603 |
| | 4.4087 | 0.4109 | 5.1751 | 1.8911 | 4.3724 | 0.3989 |
| | 0.9351 | 0.4350 | 0.9548 | 0.4725 | 0.9199 | 0.3903 |
| | 0.1620 | 0.0558 | 0.1816 | 0.0829 | 0.1592 | 0.0572 |
| | 0.1720 | 0.3284 | 0.1467 | 0.3258 | 0.1577 | 0.3027 |

(*Continued*)

**Table 3** | Estimates and their standard deviations based on 2,000 runs on the original dataset without adding artificial outliers. (*Continued*)

|   |          |        |          |         |          |        |
|---|----------|--------|----------|---------|----------|--------|
|   | −2.9528  | 0.4649 | −3.5070  | 1.4356  | −2.9171  | 0.4669 |
|   | −0.0042  | 0.0011 | −0.0045  | 0.0013  | −0.0042  | 0.0010 |
|   | 0.0010   | 0.0002 | 0.0012   | 0.0004  | 0.0010   | 0.0002 |
|   | −0.2239  | 0.3374 | −0.2846  | 0.3565  | −0.2295  | 0.3403 |
|   | −0.0021  | 0.0146 | −0.0033  | 0.0155  | −0.0019  | 0.0146 |
|   | −0.0027  | 0.0347 | −0.0209  | 0.0365  | −0.0290  | 0.0335 |
|   | 0.4231   | 0.3952 | 0.2910   | 0.4358  | 0.4132   | 0.3933 |
|   | 0.1379   | 0.0491 | 0.1505   | 0.0570  | 0.1362   | 0.0492 |
|   | −0.0482  | 0.0862 | −0.03477 | 0.0934  | −0.0488  | 0.0855 |
| 3 | 0.2206   | 0.0773 | 0.2411   | 0.08556 | 0.2204   | 0.0755 |
|   | 4.0891   | 0.4534 | 4.7372   | 1.5019  | 4.0570   | 0.4535 |
|   | 1.5783   | 0.4913 | 1.6599   | 0.5687  | 1.5755   | 0.4643 |
|   | 0.0348   | 0.1026 | 0.0484   | 0.1032  | 0.0326   | 0.1033 |
|   | 0.0196   | 0.3324 | 0.1495   | 0.3233  | 0.1865   | 0.3130 |
|   | −2.1498  | 0.5279 | −2.6216  | 1.1565  | −2.1277  | 0.5287 |
|   | −0.0001  | 0.0009 | −0.0001  | 0.0009  | −0.0001  | 0.0009 |
|   | 0.0006   | 0.0002 | 0.0007   | 0.0003  | 0.0007   | 0.0002 |

BMM, Bernoulli mixture model; PBMM, parameter space in BMM.

classification dataset obtained from LIBSVM. We have obtained two distinct ways to analyze sparsity in the finite mixture of the generalized linear model (FMGLM); the parameter space of the GLM, and the space output. Using the EM algorithm that is a convenient approach for the optimization of finite mixture models, we have shown our performance is improved. Comparing results and figures in the paper, we saw that the proposed robust methods, as well as and are better than the finite mixture of the logistic regression with multiple components.

## CONFLICTS OF INTEREST

The authors declare of no conflicts of interest.

## AUTHORS' CONTRIBUTIONS

Prof. Eskandari with designed the model, idea and the computational framework and Mrs Sabbaghi analyzed the data. All authors discussed the results and contributed to the final manuscript.

## ACKNOWLEDGMENTS

## REFERENCES

1. N.S. Grantham, Clustering Binary Data with Bernoulli Mixture Models, 2014.
2. L. Grilli, R. Varriale, C. Rampichini, Commun. Stat. Theory Methods. 44 (2013), 4866–4879.
3. M. Melkersson, J. Saarela, J. Popul. Econ. 17 (2004), 409–431.
4. S.P. Brooks, B.J.T. Morgan, M.S. Ridout, S.E. Pack, Biometrics. 53 (1997), 1097–1115.
5. H. Wang, G. Li, G. Jiang, J. Bus. Econ. Stat. 25 (2007), 347–355.
6. N.H. Nguyen, T.D. Tran, IEEE Trans. Inf. Theory. 59 (2013), 2036–2058.
7. Y. Chen, C. Caramanis, S. Rampichini Mannor, in: The Proceedings of the International Conference on Machine Learning, 2013.
8. J. Tibshirani, C.D. Manning, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 124–129.
9. E. Yang, A. Tewari, P. Ravikumar, in: International Joint Conference on Artificial Intelligence, 2013, vol. 13, pp. 1834–1840.
10. F. Eskandari, M.R. Meshkani, J. Iran. Stat. Soc. 5 (2006), 9–24.
11. D. Biohning, Ann. Inst. Stat. Math. 44 (1992), 197–200.
12. N. Stadler, P. Buhlmann, S. van de Geer, TEST. 19 (2010), 209–256.