Pattern Recognition Letters 108 (2018) 15–22

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Robust feature selection via $l_{2,1}$-norm in finite mixture of regression

Xiangrui Li, Dongxiao Zhu*

Department of Computer Science, Wayne State University, Detroit 48202, USA

ARTICLE INFO

Article history:
Received 30 May 2017
Available online 26 February 2018

MSC:
41A05
41A10
65D05
65D17

Keywords:
Finite mixture of regression
Feature selection
Non-convex optimization

ABSTRACT

Finite mixture of Gaussian regression (FMR) is a widely-used modeling technique in supervised learning problems. In cases where the number of features is large, feature selection is desirable to enhance model interpretability and to avoid overfitting. In this paper, we propose a robust feature selection method via $l_{2,1}$-norm penalized maximum likelihood estimation (MLE) in FMR, with extension to sparse $l_{2,1}$ penalty by combining l_1-norm with $l_{2,1}$-norm for increasing flexibility. To solve the non-convex and non-smooth problem of (sparse) penalized MLE in FMR, we develop an new EM-based algorithm for numerical optimization, with combination of block coordinate descent and majorizing-minimization scheme in M-step. We finally apply our method in six simulations and one real dataset to demonstrate its superior performance.

1. Introduction

Finite mixture of regression (FMR) is a flexible modeling technique for supervised learning problems. It extends uni-modal assumption of Generalized Linear Model (GLM) to multi-modal cases. This method is widely used in various applications such as biology, economics, and engineering [4,6,13]. Among these applications, many problems arise with high dimensionality yet the sample size is relatively small. Fitting FMR directly with maximum likelihood approach not only results in severe overfitting and poor generalization performance, but also makes the model hard to interpret. One viable approach to address those two issues, i.e., variance reduction and feature selection, is to fit FMR with sparsity-inducing penalty.

Various sparsity-inducing methods in GLM were extensively developed and successfully applied to classification and uni-modal regression problems. [8,16,22,25,30]. However, it is not the case for multi-modal regression problems, especially in high dimension. Initial works by [12,23] applied Lasso-typed (l_1-norm of model parameters) methods to finite mixture of Gaussian regression. l_1 penalty is capable of finding heterogeneous feature structure across mixture components. In FMR models, parameters of features are naturally structured, i.e., multiple parameters (one parameter from each mixture component) corresponding to one feature are grouped. The l_1 penalty, however, performs feature selection for each mixture component individually, hence it misses the similarity and relatedness among different mixture components. As we may expect that different mixture components of FMR share some common features, incorporating the group structure in modeling is beneficial. To this end, we seek a method that utilizes the grouping information in FMR as well as enjoy the flexibility of allowing features to be component-dependent.

We propose a novel penalized finite mixture of Gaussian regression model with structured feature selection that explicitly incorporates the information of parameter grouping via $l_{2,1}$-norm. $l_{2,1}$ penalty has one appealing property that it performs feature selection at the (parameter) group level, encouraging the same sparsity pattern across all mixture components. Our approach is more robust to noisy features and model noise as demonstrated in simulation studies. Moreover, we further extend $l_{2,1}$ to flexible sparse $l_{2,1}$ penalty by combing with l_1-norm. With incorporation of l_1-norm, sparse $l_{2,1}$ penalized FMR allows heterogeneous feature structures across mixture components while possesses robustness of $l_{2,1}$ penalty.

The resultant penalized MLE in FMR is formulated as a non-convex optimization problem. The standard approach for penalized FMR is EM-type algorithm; the non-smoothness of sparse $l_{2,1}$ penalty, however, poses substantial challenges in the M-step of EM algorithm. Inspired by a re-parameterization trick in [23], we combine block coordinate descent algorithm and majorizing-minimization scheme for numerical optimization in the M-step, with efficient closed-form updates in each iteration. Finally, we apply our method to evaluate its performance using simulation and real data sets.

* Corresponding author.
E-mail address: dzhu@wayne.edu (D. Zhu).

https://doi.org/10.1016/j.patrec.2018.02.021
0167-8655/

The rest of paper is organized as follows. In Section 2, we review related works in FMR and various sparsity-inducing methods. Section 3 describes the proposed penalized FMR. In Section 4, we present the EM-based optimization algorithm. Section 5 reports experimental results using simulation and real data. In Section 6, we conclude our paper.

2. Related work

Finite mixture of regression models can be viewed as an extension of Generalized Linear Model (GLM) and have been extensively studied [1,7,15,21,24]. In applications of FMR, "mixtures of experts" and its extension "hierarchical mixtures of experts", are widely used and have achieved great success in various applications [9–11]. See [15] for a comprehensive review for finite mixture models. While original FMR was developed mainly for low-dimensional data, as high dimensional data is common in recent years, fitting FMR directly is ill-suited due to the curse of dimensionality.

In terms of feature selection, various methods have been proposed for high dimensional data. They often improve model performances due to bias-variance trade-off. In GLM, the Lasso [25] is a penalized method by adding l_1 penalty in the maximum likelihood, for simultaneous parameter estimation and feature selection. In applications where features can be grouped by prior knowledge (for example, genes (features) are grouped into pathways), Group Lasso [16,30] extends Lasso for feature selection at the group level. Nie et al. [19] and Liu et al. [14] proposed the trace ratio criterion for best feature subset selection;

$l_{2,1}$ norm that is of the same functional form with Group Lasso is widely used as penalty for capturing the similarity and relatedness in feature structures. For example, Chang and Yang [2], Gong [5], Nie et al. [18], Xiang et al. [27] and Yang et al. [28] uses $l_{2,1}$ norm in the multi-task learning and multi-label classification; Nie et al. [20] and Yang et al. [29] proposed to use $l_{2,1}$-norm in unsupervised feature selection; Chen et al. [3] applied $l_{2,1}$-norm in semi-supervised multi-label learning.

In the context of FMR, Khalili et al. [12] and Städler et al. [23] are the pioneer works that combine FMR and feature selection via sparsity-inducing penalty. Both works consider l_1-norm penalized FMR and solve the penalized MLE using EM-algorithm, with challenges in the M-step due to the non-smoothness of l_1 penalty. In Khalili et al. [12], a differentiable approximation of l_1 penalty is used and M-step is numerically optimized by solving the system of equations given by the first order conditions. Städler et al. [23] differs from Khalili et al. [12] that a re-parameterization strategy is used, resulting in a convex problem in the M-step. This re-parameterization is beneficial as efficient algorithms such as coordinate descent in convex optimization can be applied. As mentioned in introduction, l_1 penalized FMR misses the grouping information of parameters in FMR, our approach of sparse $l_{2,1}$ penalized FMR improves this limit and extends l_1 penalized FMR, leading to group structured sparsity while keeping the ability of selecting component-dependent features.

3. Method

Notations: Scalars are denoted as lower case letters, vectors and matrices as bold letters with vectors viewed as one-column matrices. $||\cdot||_1$ represents the l_1-norm of a vector or a matrix, $||\cdot||_2$ the l_2-norm of a vector, $||\cdot||_F$ the Frobinius norm of a matrix, $||\cdot||_{2,1}$ the $l_{2,1}$-norm of a matrix. $\mathbf{R}_+$ is the set of positive reals. $\boldsymbol{v}'$ is the transpose of a vector $\boldsymbol{v}$ (or matrix). For a matrix $\boldsymbol{M}$, $\boldsymbol{M}_{i\cdot}$ and $\boldsymbol{M}_{\cdot j}$ are used to represent the ith row and the jth column respectively.

3.1. Finite mixture of regression

Let $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ be the set of independent observations, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})' \in \mathbf{R}^p$ is the p-dimensional vector of features for the ith observation, and $y_i \in \mathbf{R}$ is the response. $\boldsymbol{X}$ represents the $n \times p$ design matrix and $\boldsymbol{Y}$ the vector of observation responses. The finite mixture of regression model (with k mixture components) is given as follows:

$$y|x \sim \pi_1 \mathcal{N}(\beta_{01} + \boldsymbol{x}'\boldsymbol{\beta}_1, \sigma_1^2) + \cdots + \pi_k \mathcal{N}(\beta_{0k} + \boldsymbol{x}'\boldsymbol{\beta}_k, \sigma_k^2),$$

where π_j represents the mixture probability, σ_j $(\sigma_j > 0)$ the standard deviation, $(\beta_{0j}, \boldsymbol{\beta_j})$ the linear coefficients for the jth Gaussian component $\mathcal{N}(\beta_{0j} + \boldsymbol{x}'\boldsymbol{\beta}_j, \sigma_j^2)$ $(1 \le j \le k)$. π_j's satisfy $\pi_j > 0$ and $\sum_{j=1}^k \pi_j = 1$.

The finite mixture of regression model is studied from the maximum likelihood approach. The parameter of FMR,

$$\Theta = (\pi_1, \ldots, \pi_k, \beta_{01}, \ldots, \beta_{0k}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \sigma_1, \ldots, \sigma_k),$$

is estimated by minimizing the (scaled) negative log-likelihood:

$$L(\Theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} \frac{\pi_j}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(y_i - \beta_{0j} - \boldsymbol{x}_i'\boldsymbol{\beta}_j)^2}{2\sigma_j^2}\right)\right), \tag{1}$$

with the constraint $\sum_{j=1}^k \pi_j = 1,\ \sigma_j > 0$.

3.2. Feature selection in $l_{2,1}$-norm penalized FMR

Assume that for the jth $(1 \le j \le k)$ mixture component $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{pj})'$, and write $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$ as a $(p \times k)$ matrix of parameters. Let $\boldsymbol{\beta}_{l\cdot} = (\beta_{l1}, \ldots, \beta_{lk}) \in \mathbf{R}^k$ is the mth row of $\boldsymbol{\beta}$, corresponding to the lth feature in the finite mixture model. With this correspondence, $\boldsymbol{\beta}$ can be decomposed into p such parameter groups.

The $l_{2,1}$-norm of $\boldsymbol{\beta}$ is defined as:

$$||\boldsymbol{\beta}||_{2,1} = \sum_{l=1}^{p} \sqrt{\sum_{j=1}^{k} \beta_{lj}^2} = \sum_{l=1}^{p} ||\boldsymbol{\beta}_{l\cdot}||_2. \tag{2}$$

Using $l_{2,1}$-norm as a penalty, the penalized MLE for FMR is obtained by solving:

$$\arg\min_{\Theta} L(\Theta) + \lambda||\boldsymbol{\beta}||_{2,1}, \tag{3}$$

where $\lambda > 0$ is a tuning parameter.

The $l_{2,1}$ penalty selects features at the group level: with a large λ, one feature is unselected by dropping out the corresponding whole parameter group. For the selected parameter group in $l_{2,1}$, every element is non-zero. However, this property seems restrictive when the component-dependent features in FMR may exist. To this end, we further propose the sparse $l_{2,1}$ penalty in FMR by incorporating l_1-norm:

$$P(\boldsymbol{\beta}) = (1 - \alpha)\sqrt{k}||\boldsymbol{\beta}||_{2,1} + \alpha||\boldsymbol{\beta}||_1 \tag{4}$$

where $\alpha \in [0, 1]$ is a trade-off between l_1 and l_2 norm. $\sqrt{k}$ is used for a balance between l_1 and l_2 penalty. Notice that if $\alpha = 1$, (4) is exactly the l_1 penalty; while $\alpha = 0$, (4) is deduced to $l_{2,1}$ penalty.

With sparse $l_{2,1}$ penalty, (3) can be extended to

$$\arg\min_{\Theta} L(\Theta) + \lambda P(\boldsymbol{\beta}), \tag{5}$$

with constraints $\Theta \in \mathbf{R}_+^k \times \mathbf{R}^{k(p+1)} \times \mathbf{R}_+^k,\ \sum_{j=1}^k \pi_j = 1$.

3.3. Re-parameterization

The general framework for maximum likelihood approach in unpenalized FMR is to use EM-algorithm. In the penalized cases, existing methods [12,23] adopt EM approach with optimizing the expectation of penalized complete log-likelihood. With the conventional parameterization Θ, the subproblem of optimization in the M-step is a challenging task due to non-convexity of the negative expected complete log-likelihood and non-smoothness of the sparse $l_{2,1}$ penalty. However, the non-convexity can be tackled by a re-parameterization of model parameters [23], resulting in a convex optimization problem in the M-step. The same trick can be applied in the sparse $l_{2,1}$ penalized FMR. Although the non-smoothness induced by $l_{2,1}$- and l_1-norm still exists, as we show in the next section, re-parameterization is sufficient for an effective optimization scheme in the M-step of EM algorithm.

In FMR, the negative maximum likelihood function (1) can be re-parameterized, for $j = 1, \ldots, k$, as follows:

$$\tau_j = \sigma_j^{-1}, \quad \eta_{0j} = \beta_{0j}/\sigma_j, \quad \boldsymbol{\eta}_j = \boldsymbol{\beta}_j/\sigma_j.$$

Let $\Phi = (\pi_1, \ldots, \pi_k, \eta_{01}, \ldots, \eta_{0k}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k, \tau_1, \ldots, \tau_k)$. The re-parameterization yields an one-to-one mapping from Θ to Φ. Hence, (1) can be rewritten as:

$$L(\Phi) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} \pi_j \frac{\tau_j}{\sqrt{2\pi}} \exp\left(-\frac{(\tau_j y_i - \eta_{0j} - \boldsymbol{x}_i'\boldsymbol{\eta}_j)^2}{2}\right)\right). \tag{6}$$

Under the re-parameterization, the sparse $l_{2,1}$ penalized MLE Φ^* for FMR is given by the following optimization problem:

$$\Phi^* = \arg\min_{\Phi} L(\Phi) + \lambda P(\boldsymbol{\eta}), \tag{7}$$

where $\Phi \in \mathbf{R}_+^k \times \mathbf{R}^{k(p+1)} \times \mathbf{R}_+^k$, $\sum_{j=1}^k \pi_j = 1$. λ and α are the tuning parameters. Notice that η_{0j}'s are not penalized.

4. Non-convex optimization

In this section, we present a EM-based algorithm for numerically optimizing sparse $l_{2,1}$ penalized FMR.

In the following sections, for notational ease, we have suppressed η_{0j} into $\boldsymbol{x}'\boldsymbol{\eta}_j$ except in (2b) and (2c) of M-step. $\langle \cdot, \cdot \rangle$ represents dot product and $\odot$ element-wise product for two matrices of the same dimension.

Let $l_c(\Phi)$ be the complete log-likelihood function:

$$l_c(\Phi) = \sum_{i=1}^{n}\sum_{j=1}^{k}\left[z_{ij}\log\left(\frac{\tau_j}{\sqrt{2\pi}}\exp\left(-\frac{(\tau_j y_i - \boldsymbol{x}_i'\boldsymbol{\eta}_j)^2}{2}\right)\right) + z_{ij}\log\pi_j\right],$$

where $\{z_{ij} : j = 1, \ldots, k\}$ is the latent membership vector for the ith observation belonging to which mixture component: $z_{ij} = 1$ and $z_{ir} = 0$ for $r \neq j$ indicating the i-th observation belongs to the jth component. The sparse $l_{2,1}$ penalized (scaled) negative complete log-likelihood is then

$$L_c(\Phi) = -\frac{1}{n} l_c(\Phi) + \lambda P(\boldsymbol{\eta}).$$

The EM algorithm minimizes $L_c(\Phi)$ by iterating between the E- and M-step as follows. Assume that $\Phi^{(m)}$ is the current parameter estimate:

E step. Given the current estimate $\Phi^{(m)}$, the E step computes the conditional expectation $Q(\Phi)$ of $L_c(\Phi)$ with respect to the latent variables z_{ij}'s. This is equivalent to compute the expectation $w_{ij}^{(m)}$ of z_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, k$:

$$w_{ij}^{(m)} = \frac{\pi_j^{(m)}\tau_j^{(m)}\exp\left(-\frac{1}{2}(\tau_j^{(m)}y_i - \boldsymbol{x}_i'\boldsymbol{\eta}_j^{(m)})^2\right)}{\sum_{r=1}^{k}\pi_r^{(m)}\tau_r^{(m)}\exp\left(-\frac{1}{2}(\tau_r^{(m)}y_i - \boldsymbol{x}_i'\boldsymbol{\eta}_r^{(m)})^2\right)}. \tag{8}$$

$$Q(\Phi) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}^{(m)}\left[\log\left(\frac{\tau_j}{\sqrt{2\pi}}\exp\left(-\frac{(\tau_j y_i - \boldsymbol{x}_i'\boldsymbol{\eta}_j)^2}{2}\right)\right) + \log\pi_j\right] + \lambda\left((1-\alpha)\sum_{l=1}^{p}\sqrt{k}||\boldsymbol{\eta}_{l\cdot}||_2 + \alpha\sum_{l=1}^{p}||\boldsymbol{\eta}_{l\cdot}||_1\right).$$

M step. The update $\Phi^{(m+1)}$ is obtained by minimizing $Q(\Phi)$ with respect to Φ, which is further equivalent to solve two independent minimization problems with respect to $(\pi_1, \ldots, \pi_k)$ and $(\eta_{01}, \ldots, \eta_{0k}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k, \tau_1, \ldots, \tau_k)$ respectively.

(1) Minimization with respect to $(\pi_1, \ldots, \pi_k)$. This is the same with mixture probability update in the usual EM algorithm:

$$\pi_j^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{ij}^{(m)}, \quad j = 1, \ldots, k. \tag{9}$$

(2) Update $\Phi_{-\pi} = (\eta_{01}, \ldots, \eta_{0k}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k, \tau_1, \ldots, \tau_k)$. After simplification of $Q(\Phi)$, we equivalently solve the following convex problem (each summand is convex):

$$\begin{aligned}\Phi_{-\pi}^{(m+1)} &= \min_{\Phi_{-\pi}} S(\Phi_{-\pi})\\ S(\Phi_{-\pi}) &= -\frac{1}{n}\sum_{j=1}^{k}\Big(\sum_{i=1}^{n} w_{ij}^{(m)}\Big)\log\tau_j\\ &\quad + \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}\frac{w_{ij}^{(m)}}{2}(\tau_j y_i - \boldsymbol{x}_i'\boldsymbol{\eta}_j)^2\\ &\quad + \lambda\Big((1-\alpha)\sum_{l=1}^{p}\sqrt{k}||\boldsymbol{\eta}_{l\cdot}||_2 + \alpha\sum_{l=1}^{p}||\boldsymbol{\eta}_{l\cdot}||_1\Big).\end{aligned} \tag{10}$$

The sparse $l_{2,1}$ penalty is separable between blocks of $\eta_{m\cdot}$'s [26], implying block coordinate algorithm is well suited for minimizing $S(\Phi_{-\pi})$.

In the block coordinate descent, we cyclically update each parameter (or parameter blocks) by approximately minimizing $S(\Phi_{-\pi})$, holding all except the current parameter fixed until convergence. This update strategy leads to the updates as follows. As a side note, Eq. (10) can be initialized by the current estimate $\Phi_{-\pi}^{(m)}$. For notational simplicity, we drop the iteration index of $\Phi_{-\pi}$ and use $\tilde{\Phi}_{-\pi} = (\tilde{\eta}_{01}, \ldots, \tilde{\eta}_{0k}, \tilde{\boldsymbol{\eta}}_1, \ldots, \tilde{\boldsymbol{\eta}}_k, \tilde{\tau}_1, \ldots, \tilde{\tau}_k)$ to represent the current value for $\Phi_{-\pi}$ in optimizing Eq. (10).

(2a) Update τ_j.

Since τ_j is not penalized and $S(\Phi_{-\pi})$ be differentiable with respect to τ_j, first order condition $\partial S(\Phi_{-\pi})/\partial\tau_j = 0$ results in the update for $j = 1, \ldots, k$:

$$\tilde{\tau}_j \leftarrow \frac{\langle \boldsymbol{W}_{\cdot j}^{(m)}, \boldsymbol{Y}\odot\hat{\boldsymbol{Y}}\rangle + \sqrt{\langle \boldsymbol{W}_{\cdot j}^{(m)}, \boldsymbol{Y}\odot\hat{\boldsymbol{Y}}\rangle^2 + 4s_j\langle \boldsymbol{W}_{\cdot j}^{(m)}, \boldsymbol{Y}\odot\boldsymbol{Y}\rangle}}{2\langle \boldsymbol{W}_{\cdot j}^{(m)}, \boldsymbol{Y}\odot\boldsymbol{Y}\rangle}, \tag{11}$$

where $\boldsymbol{W}_{\cdot j}^{(m)} = (w_{1j}^{(m)}, \ldots, w_{nj}^{(m)})'$, $\boldsymbol{Y} = (y_1, \ldots, y_n)'$, $\hat{\boldsymbol{Y}} = (\boldsymbol{x}_1'\tilde{\boldsymbol{\eta}}_j, \ldots, \boldsymbol{x}_n'\tilde{\boldsymbol{\eta}}_j)'$ and $s_j = \sum_{i=1}^n w_{ij}^{(m)}$.

(2b) Update η_{0j}.

Due to differentiability of $S(\Phi_{-\pi})$ respect to η_{0j}, the first order condition $\partial S(\Phi_{-\pi})/\partial\eta_{0j} = 0$ gives the minimizer

$$\tilde{\eta}_{0j} \leftarrow \frac{\langle \boldsymbol{W}_{\cdot j}^{(m)}, \hat{\boldsymbol{H}}_{0j}\rangle}{s_j}, \quad j = 1, \ldots, k, \tag{12}$$

where $\hat{\boldsymbol{H}}_{0j} = (\tilde{\tau}_j y_1 - \boldsymbol{x}_1'\tilde{\boldsymbol{\eta}}_j, \ldots, \tilde{\tau}_j y_n - \boldsymbol{x}_n'\tilde{\boldsymbol{\eta}}_j)$.

(2c) Update parameter block $\boldsymbol{\eta}_{l\cdot} = (\eta_{l1}, \ldots, \eta_{lk})$.

Minimizing $S(\Phi_{-\pi})$ with respect to $\boldsymbol{\eta}_{l\cdot}$ is equivalent to minimizing (with matrix notation)

$$\min_{\boldsymbol{\eta}_{l\cdot}} \frac{1}{2n}||\sqrt{\boldsymbol{W}^{(m)}} \odot (\tilde{\boldsymbol{R}} - \sum_{h\neq l} \boldsymbol{X}_{\cdot h}\tilde{\boldsymbol{\eta}}_{h\cdot} - \boldsymbol{X}_{\cdot l}\boldsymbol{\eta}_{l\cdot})||_F^2 + \lambda_1||\boldsymbol{\eta}_{l\cdot}||_1 + \lambda_2||\boldsymbol{\eta}_{l\cdot}||_2, \quad (13)$$

where $\sqrt{\boldsymbol{W}^{(m)}} = (\sqrt{w_{ij}^{(m)}})_{n\times k}$, $\tilde{\boldsymbol{R}} = (\tilde{r}_{ij})_{n\times k}$ with $\tilde{r}_{ij} = \tilde{\tau}_j y_i - \tilde{\eta}_{0j}$, $\lambda_1 = \lambda\alpha$, $\lambda_2 = \lambda(1-\alpha)\sqrt{k}$. Since the problem in Eq. (13) is convex, an efficient gradient decent type algorithm can be used that combines majorizing-minimization scheme with the first order condition (in terms of subgradient).

Details of step (2c) We denote the quadratic term in Problem (13) (ignoring $1/2n$) as

$$M(\boldsymbol{\eta}_{l\cdot}) = ||\boldsymbol{A} - \sqrt{\boldsymbol{W}^{(m)}} \odot \boldsymbol{X}_{\cdot l}\boldsymbol{\eta}_{l\cdot}||_F^2,$$

where $\boldsymbol{A} = \sqrt{\boldsymbol{W}^{(m)}} \odot (\tilde{\boldsymbol{R}} - \sum_{h\neq l} \boldsymbol{X}_{\cdot h}\tilde{\boldsymbol{\eta}}_{h\cdot})$. Simple calculation gives us that for $j = 1, \ldots, k$:

$$\frac{\partial M}{\partial \eta_{lj}} = 2\Big[||\sqrt{\boldsymbol{W}^{(m)}}_{\cdot j} \odot \boldsymbol{X}_{\cdot l}||_2^2\eta_{lj} - (\sqrt{\boldsymbol{W}^{(m)}}_{\cdot j} \odot \boldsymbol{X}_{\cdot l})'\boldsymbol{A}_{\cdot j}\Big]$$

$$\frac{\partial^2 M}{\partial \eta_{lj}\partial \eta_{lr}} = \begin{cases} 2||\sqrt{\boldsymbol{W}^{(m)}}_{\cdot j} \odot \boldsymbol{X}_{\cdot l}||_2^2 & \text{if } j = r \\ 0 & \text{if } j \neq r, \end{cases}$$

implying the Hessian matrix $\boldsymbol{H}_M$ of M is diagonal. If we let

$$t = 2\max_{1\le j\le k} ||\sqrt{\boldsymbol{W}^{(m)}}_{\cdot j} \odot \boldsymbol{X}_{\cdot l}||_2^2,$$

we see that $t\boldsymbol{I}$ will dominate Hessian $\boldsymbol{H}_M$ of M. That is, $t\boldsymbol{I} - \boldsymbol{H}_M$ is positive semi-definite. Consequently, we have a majorization function M_m by replacing $\boldsymbol{H}_M$ with $t\boldsymbol{I}$ in the second order Taylor expansion of M (which is M itself) at the current value $\tilde{\boldsymbol{\eta}}_{l\cdot}$ of $\boldsymbol{\eta}_{l\cdot}$:

$$\begin{aligned} M_m(\boldsymbol{\eta}_{l\cdot}) &= M(\tilde{\boldsymbol{\eta}}_{l\cdot}) + (\boldsymbol{\eta}_{l\cdot} - \tilde{\boldsymbol{\eta}}_{l\cdot})\nabla M(\tilde{\boldsymbol{\eta}}_{l\cdot}) + \tfrac{t}{2}||\boldsymbol{\eta}_{l\cdot} - \tilde{\boldsymbol{\eta}}_{l\cdot}||_2^2, \\ M_m(\boldsymbol{\eta}_{l\cdot}) &\ge M(\boldsymbol{\eta}_{l\cdot}), \end{aligned}$$

where ∇M is the Jacobian of M.

With the help of the majorizing function $M_m(\boldsymbol{\eta}_{l\cdot})$, we approximately solve Problem (10) by solving:

$$\min_{\boldsymbol{\eta}_{l\cdot}} \frac{1}{2n}M_m(\boldsymbol{\eta}_{l\cdot}) + \lambda_1||\boldsymbol{\eta}_{l\cdot}||_1 + \lambda_2||\boldsymbol{\eta}_{l\cdot}||_2,$$

which is further equivalent to

$$\min_{\boldsymbol{\eta}_{l\cdot}} \frac{t}{4n}||\boldsymbol{\eta}_{l\cdot} - (\tilde{\boldsymbol{\eta}}_{l\cdot} - \nabla M(\tilde{\boldsymbol{\eta}}_{l\cdot})/t)||_2^2 + \lambda_1||\boldsymbol{\eta}_{l\cdot}||_1 + \lambda_2||\boldsymbol{\eta}_{l\cdot}||_2 \quad (14)$$

For Problem (14), by the first order condition given by subgradient, the closed form of exact solution can be calculated and given as follows:

$$\boldsymbol{\eta}_{l\cdot}^{\star} = \begin{cases} 0, & ||T_{\lambda_1}(\frac{\boldsymbol{\mu}}{u})||_2 \le \lambda_2 \\ \Big[1 - \dfrac{\lambda_2 u}{||T_{\lambda_1 u}(\boldsymbol{\mu})||_2}\Big] \cdot T_{\lambda_1 u}(\boldsymbol{\mu}), & ||T_{\lambda_1}(\frac{\boldsymbol{\mu}}{u})||_2 > \lambda_2, \end{cases} \quad (15)$$

where $u = 2n/t$, $\boldsymbol{\mu} = \tilde{\boldsymbol{\eta}}_{l\cdot} - \nabla M(\tilde{\boldsymbol{\eta}}_{l\cdot})/t$, $S(u, v) = \text{sign}(u)\max\{|u| - v, 0\}$ $(u \in \mathbf{R}, v \ge 0)$ is the soft-thresholding operator and $T_v(\boldsymbol{\mu}) = \big(S(\mu_1, v), \ldots, S(\mu_d, v)\big)$ is the element-wise soft-thresholding. Iteratively applying this update solves Problem (13).

Convergence and scalability The convergence of EM algorithm and block coordinate descent (BCD) [26] is non-trivial and well established. Hence, our algorithm nesting BCD in the M-step of EM still converges. In the E-step, the computational cost is $\mathcal{O}(nkp)$; in the M-step, one round of cyclically updating each parameter block is also $\mathcal{O}(nkp)$, leading to total cost $\mathcal{O}(Rnkp)$, where R is the number of iterations for BCD to converge with sub-linear convergence [17]. In summary, one iteration of EM has cost $\mathcal{O}(Rnkp)$.

5. Experiments

5.1. Simulation

In this section, we set up six models (M1-M6) for simulation to investigate performances of (sparse) $l_{2,1}$ penalized FMR. M1, M2, M3 and M4 are designed to evaluate $l_{2,1}$ penalized FMR ($\alpha = 0$), with comparison to l_1 penalized FMR[1] ($\alpha = 1$). In M5 and M6, we evaluate the sparse $l_{2,1}$ penalized FMR ($0 < \alpha < 1$). For model evaluation, we report average results on 50 independent runs.

In all models, design matrix $\boldsymbol{X}$ is generated from multivariate normal distribution with zero mean and a diagonal covariance matrix. Response $\boldsymbol{Y}$ is then generated from finite mixture of k Gaussians. Details of model setup are specified in Table 1.

In each run of simulations, we partitioned simulated data (n observations) to three equal subsets: training, validation and testing sets. A sequence of models corresponding to a λ-sequence is trained on training data; the optimal value of tuning parameter λ along with its corresponding trained model is selected as the one minimizing NLogloss (see below) on validation data. Finally, predictive NLogLoss are evaluated and reported on testing data.

Performance metrics We assess the performance of penalized FMR from different perspectives. For predictive performance, we used negative log-likelihood loss (NLogLoss) in trained model:

$$-\sum_{i=1}^{n}\log\left(\sum_{j=1}^{k}\hat{\pi}_j\frac{1}{\sqrt{2\pi}\hat{\sigma}_j}\exp\left(-\frac{(y_i - \hat{\beta}_{0j} - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_j)^2}{2\hat{\sigma}_j^2}\right)\right).$$

To measure sparsity, we first reported "true positive rate" (TPR) and "false positive rate" (FPR):

$$\text{TPR} = \frac{\#\text{active parameters selected}}{\#\text{active parameters}},$$

$$\text{FPR} = \frac{\#\text{inactive parameters selected}}{\#\text{inactive parameters}}.$$

To further quantify accuracy of feature selection, we used root mean squared error (RMSE) of estimators:

$$\text{RMSE} = \sqrt{\frac{1}{kp}||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||_F^2},$$

where β is the matrix of true values and $\hat{\beta}$ is the estimation matrix.

Results on M1-M4 In M1-M4, we specifically investigate performances of $l_{2,1}$ FMR ($\alpha = 0$), in comparison with l_1 FMR. M1 and M2 have the same model setups with five active parameters, except that M2 is noisier than M1: true standard deviation is 0.5 *vs* 1.5. M3 is slightly more complicated with three mixture components. M4 has similar setups with M1 but differs in the sparsity pattern of active parameters. There are two different sparsity patterns in four models: M1, M2 and M3 are of the same sparsity pattern that all mixture components have active parameters corresponding to a same set of features; on the contrary, mixture components in M4 are of heterogeneous feature structure: none of features are active in both two components.

We fitted the models with $p = 50$ and $p = 100$, keeping true parameters unchanged while doubling noisy parameters. Since $l_{2,1}$ penalty performs parameter selection at the group level, it implies an homogeneous assumption of active features for all mixture components. Therefore, $l_{2,1}$ FMR is expected to perform better in M1, M2 and M3, when the underlying assumption is satisfied, but worse than l_1 in heterogeneous case of M4. Simulation results confirm our expectations on $l_{2,1}$ FMR.

[1] l_1 FMR fitted with R package "fmrlasso".

Table 1
Model parameter specification.

	M1	M2	M3	M4	M5	M6
n	300	300	450	300	300	300
p	50, 100	50, 100	50, 100	50, 100	50, 100	50,100
k	2	2	3	2	2	2
π	(0.5,0.5)	(0.5,0.5)	(1/3,1/3,1/3)	(0.5,0.5)	(0.5,0.5)	(0.5,0.5)
σ	(0.5,0.5)	(1.5,1.5)	(0.5,0.5)	(0.5,0.5)	(1,1)	(0.3,0.3)
β_1	(4,4,4,4,4)	(4,4,4,4,4)	(10,10,10,10,10,)	(4,4,4,4,0,0,0,0)	(4,4,4,4,4,4,0,0,0)	(4,4,4,4,4,4,0,0,0)
β_2	$(-1,-1,-1,-1,-1)$	$(-1,-1,-1,-1,-1)$	(3,3,3,3,3)	(0,0,0,0,$-1,-1,-1,-1$)	$(-1,-1,-1,0,0,0,-1,-1,-1)$	$(-1,-1,-1,0,0,0,-1,-1,-1)$
β_3	–	–	$(-1,-1,-1,-1,-1)$	–	–	–

Table 2
M1-M4: predictive negative log-likelihood (smaller is better) with standard deviation for $l_{2,1}$ and l_1 penalized FMR.

p		M1	M2	M3	M4
50	$l_{2,1}$	**186.35** (10.80)	**267.94** (15.48)	**399.06** (24.86)	192.01 (17.35)
	l_1	190.21 (16.47)	277.30 (20.45)	402.28 (57.28)	**180.09** (16.41)
100	$l_{2,1}$	**201.36** (10.30)	**282.05** (27.33)	**429.35** (60.04)	219.65 (24.32)
	l_1	229.42 (31.94)	302.63 (23.87)	493.83 (54.44)	**198.43** (18.97)

Table 3
M5-M6: predictive negative log-likelihood (smaller is better) with standard deviation for sparse $l_{2,1}$ FMR, fitted with different α's.

p	α	M5	M6
50	0 ($l_{2,1}$)	258.45 (21.63)	188.35 (17.44)
	0.25	256.26 (21.56)	183.75 (17.83)
	0.50	**255.37** (22.06)	182.99 (18.21)
	0.75	256.22 (26.43)	181.72 (19.87)
	1 (l_1)	256.49 (25.96)	**166.82** (20.61)
100	0 ($l_{2,1}$)	281.24 (25.89)	224.08 (33.43)
	0.25	277.89 (26.87)	**216.77** (27.20)
	0.50	**277.56** (26.87)	218.72 (35.07)
	0.75	279.07 (26.92)	231.14 (55.18)
	1 (l_1)	289.84 (35.66)	236.09 (51.30)

Table 2 presents the predictive negative log-likelihood. In either case of $p = 50$ and $p = 100$, $l_{2,1}$ FMR has better predictive power than l_1 FMR in M1, M2 and M3, but less in M4. In evaluation of feature selection, Fig. 1 shows the results. We found that with respect to estimation accuracy and active parameter selection, RMSE

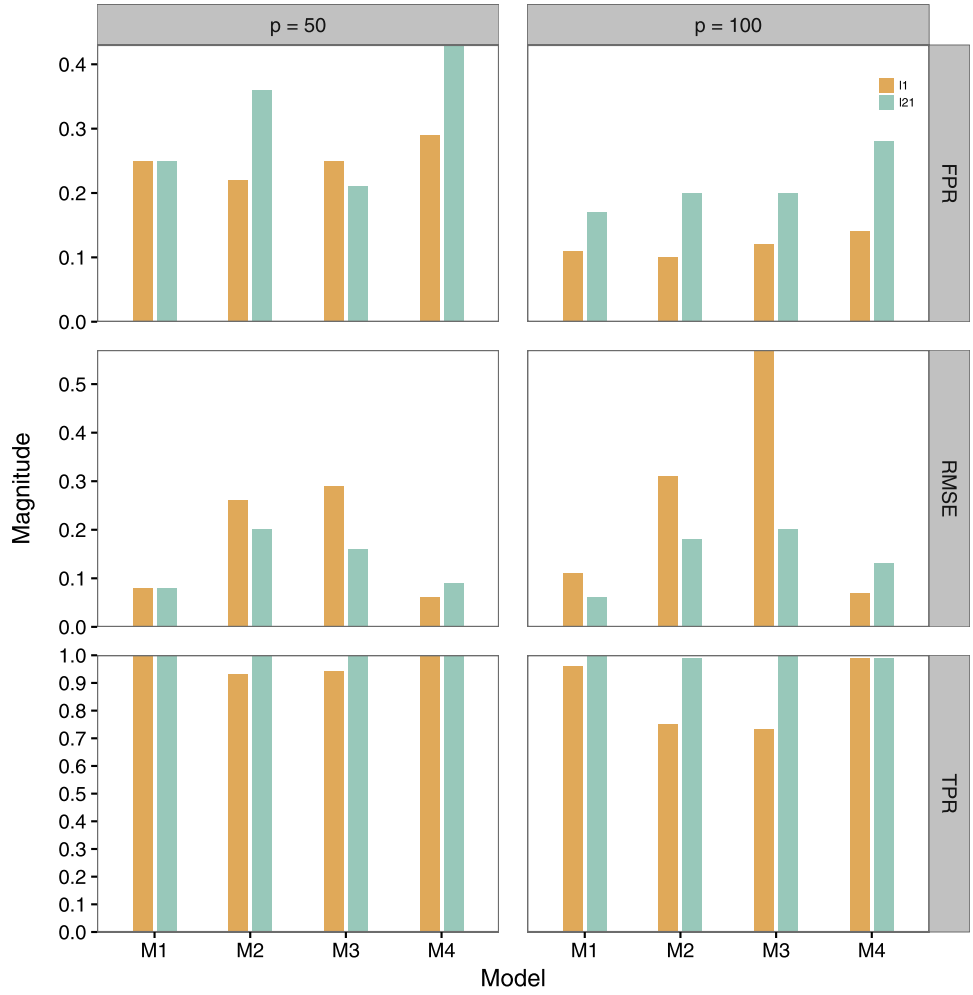

Fig. 1. RMSE, TPR and FPR for $l_{2,1}$- and l_1 penalty on M1-M4. For RMSE and FPR, smaller is better; for TPR, larger is better.

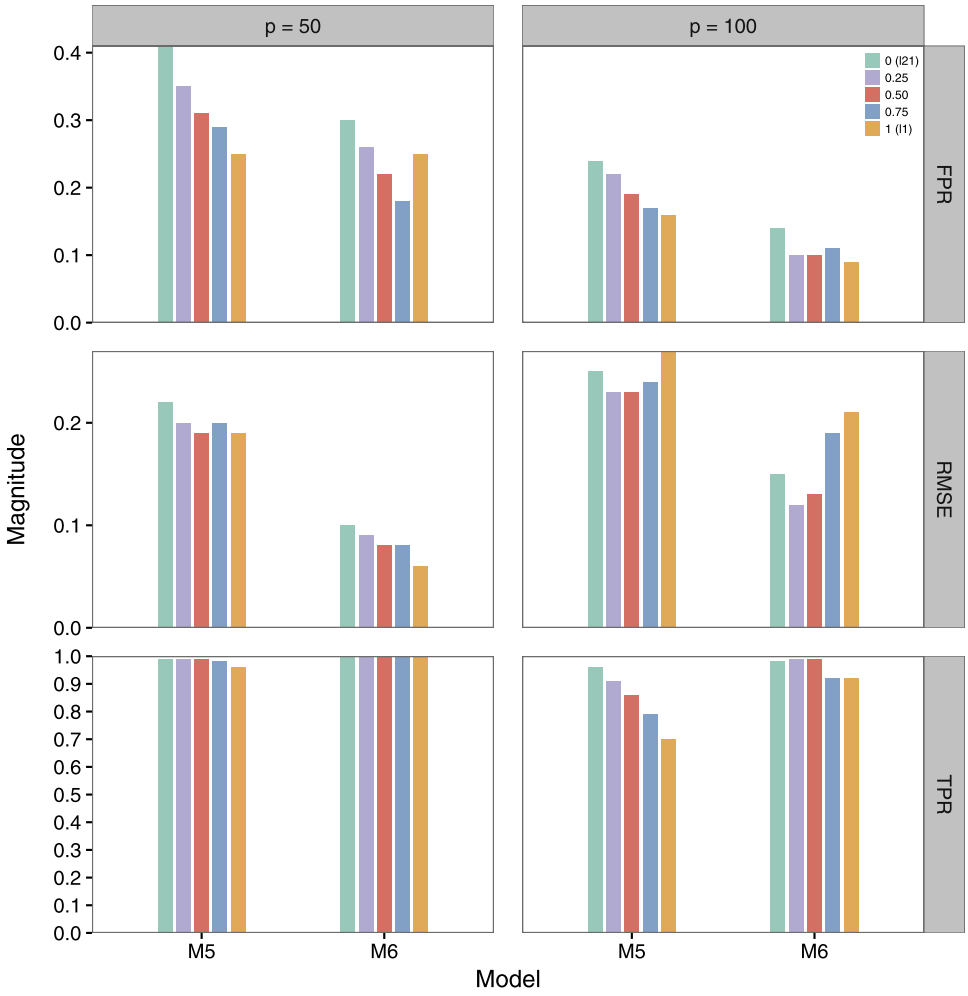

Fig. 2. RMSE, TPR and FPR for sparse $l_{2,1}$ penalty FMR fitted with different α values {0, 0.25, 0.50, 0.75, 1} on M5 and M6. Note that $\alpha = 0$ is $l_{2,1}$ penalty, $\alpha = 1$ is l_1 penalty.

and TPR display the same trend: $l_{2,1}$ is overall better than l_1 in M1, M2 and M3, but worse in M4. Interestingly in terms of FPR, l_1 penalty performs overall better than $l_{2,1}$ in M1-M4. This is possibly due to the effect of group selection in $l_{2,1}$ penalty. However, even with a larger FPR in M1, M2 and M3, the smaller RMSE of $l_{2,1}$ along with a larger TPR indicates that the false positives (i.e. inactive parameters selected) in $l_{2,1}$ are estimated accurately as insignificant numbers. It is particularly impressive that in M2 and M3, adding more noisy features, $l_{2,1}$ remains stable and robust while l_1 has a significant loss in TPR and RMSE.

Results on M5-M6 As we see the experimental results in M1-M4, the homogeneous assumption of active features is critical for $l_{2,1}$ penalty. In real cases, we believe this assumption is too restrictive to be satisfied. To alleviate the limits of $l_{2,1}$, we introduce sparse $l_{2,1}$ penalized FMR that is a weighted combination of l_1 norm and $l_{2,1}$ norm. As a result, the sparse $l_{2,1}$ penalty has a property of further selecting parameters within the selected parameter groups, which could potentially enhance the performance of $l_{2,1}$ penalty only.

With this purpose in mind, we designed M5 and M6 to show that sparse $l_{2,1}$ indeed benefits from incorporating within-group sparsity in $l_{2,1}$. M5 and M6 have the same setups except that M5 is noisier than M6. Notice that in M5 and M6, feature structures for mixture components are different yet with three active features in common. In these cases, we fit sparse $l_{2,1}$ penalized FMR with $\alpha \in (0, 1)$ and compare it with models of $l_{2,1}$ only and l_1 only.

In the sparse $l_{2,1}$ penalty, α could be treated as a tuning parameter and the optimal value can be selected by validation set or cross validation. In our experiment, to keep computational cost moderate, we pre-fix α on a grid of values {0, 0.25, 0.5, 0.75, 1}, where $\alpha = 0$ corresponds to $l_{2,1}$ penalty and $\alpha = 1$ to l_1 penalty. Predictive negative log-likelihoods are shown in Table 3. We see that in four cases, the sparse $l_{2,1}$ penalty (that is, $\alpha \neq 0$) indeed provides improvement over $l_{2,1}$ penalty only. In the setting of higher dimension $p = 100$ or larger noise $\sigma = 1$, it also performs better than l_1 penalty. Fig. 2 shows sparsity performance. One observation is that, due to group selection effect, $l_{2,1}$ overall has larger both the TPR (larger is better) and FPR (smaller is better); incorporating l_1 into $l_{2,1}$ alleviates this trade-off, leading that the sparse $l_{2,1}$ has better estimation RMSE than the single $l_{2,1}$ penalty. Similarly as predictive performance, it also has better accuracy than the single l_1 penalty in the noisier models.

5.2. Real data application

We apply the sparse $l_{2,1}$ penalized FMR in (1) Wisconsin breast cancer dataset (WBC) that is publicly available at UCI machine learning repository[2]; (2) NHL salaries dataset.[3] WBC contain 194 records of "time to recur" for patients with breast cancer (after removing 4 cases with missing values) with 32 features describing

[2] http://archive.ics.uci.edu/ml/index.html.
[3] https://www.kaggle.com/camnugent/predict-nhl-player-salaries.

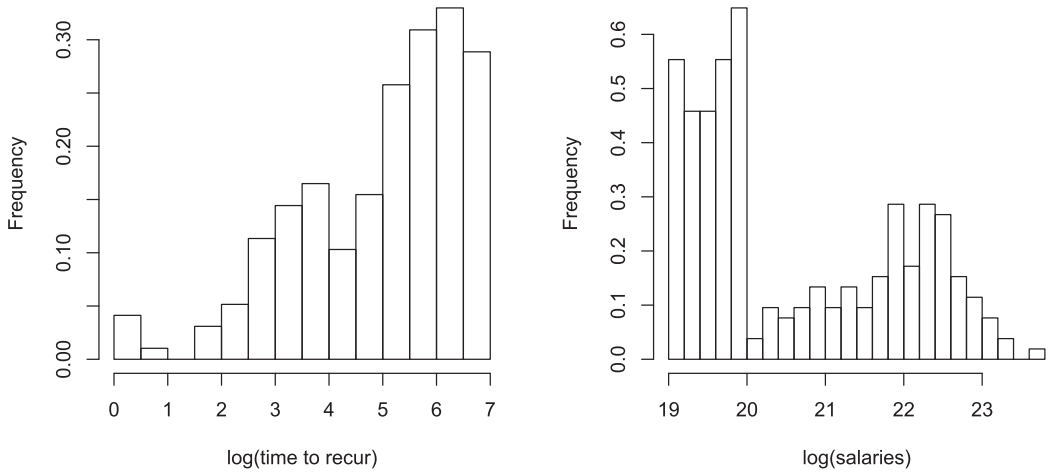

Fig. 3. Histogram of log(time to recur).

Table 4
Predictive performance. 10-fold mean predictive negative log-likelihood for unpenalized-, l_1 penalized- and sparse $l_{2,1}$ penalized FMR. Number of components varies from 1 to 3.

	Model/Component(s)	1	2	3
WBC	Unpenalized	1.87	3.10	15.70
	l_1	1.70	1.71	1.69
	Sparse $l_{2,1}$	–	1.74	**1.66**
NHL	Unpenalized	1.85	–	
	l_1	1.41	1.40	
	Sparse l_{21}	–	**1.38**	

Table 5
Parameter estimation for WBC dataset.

Feature	Component 1	Component 2	Component 3
Texture	−0.060	−0.010	0.028
Area	−0.004	−0.002	− 0.001
Area SE	0.012	−0.008	−0.013
Worst texture	−0.044	−0.041	0.045
Worst perimeter	0	−0.014	0.007
Largest area	0	0.002	0.001
Tumor size	0.077	−0.027	−0.015
Positive lymph nodes	0.047	0.011	−0.097

characteristics of the cell nuclei in the digitized image of breast cancer. NHL dataset used in experiments has 262 records of NHL players with 122 features measuring players' performances. In our experiment, we used the logarithm of the original targets as the targets due to the right-skewness of original scale. Fig. 3 shows the histograms of transformed targets.

The histogram in Fig. 3 demonstrates a highly unbalanced mixture of Gaussian distribution. Thus we fit FMR with $k = 1, 2, 3$ components for WBC and $k = 1, 2$ for NHL. Three methods, unpenalized-,[4] l_1 penalized- and sparse $l_{2,1}$ penalized FMR, were applied in the experiment. For penalized FMR, tuning parameters λ and α are selected by 10-fold cross-validation based on predictive mean negative log-likelihood (CV loss).

Predictive performance Table 4 shows the results of predictive negative log-likelihood. It turns out that the optimal value (1.66 and 1.38) is achieved with sparse $l_{2,1}$ FMR. An interesting observation for WBC is that for unpenalized FMR, a mixture of two or three components introduces much variance in modeling, resulting in a significant larger CV loss than a single component linear model (3.10 and 15.70 respectively compared with 1.87); for NHL data, unpenalized mixture of two Gaussians fails due to the high dimensionality (122 features). In contrast, l_1 and $l_{2,1}$ FMR are more robust when model complexity increases.

Feature selection In Table 5, the sparse $l_{2,1}$ FMR selects 8 features for WBC data. The selected features demonstrate heterogeneous effects in different mixture components, possibly due to high correlations among features. However, in terms of predictive performance, the sparse $l_{2,1}$ penalized FMR performs even better than the unpenalized uni-modal model (non-mixture) with an 11% improvement ((1.87-1.66)/1.87). Table 6 shows the results for NHL data. The selected features suggest that players who are able to bring more wins or scores (faceoff wins, penalty drawn and shifts) have higher salaries (Component 1 corresponds to the left peak and Component 2 to the right peak of the histogram in Fig. 3). This interpretation accords with our intuitions that better players have higher pays.

[4] Unpenalized FMR was fitted with R package "flexmix".

Table 6
Parameter estimation for NHL dataset.

Feature	Component 1	Component 2
Games played	0.057	0.016
Shifts	0.716	0.091
Setup passes	0.012	0
Shots on goal	0.005	0
Faceoff wins	1.371	0.070
Penalty drawn	0.657	0.042

6. Conclusion

In this paper, we have introduced the sparse $l_{2,1}$ penalized FMR model with feature selection. The $l_{2,1}$ penalty captures relatedness among mixture components. We have shown in the simulation studies $l_{2,1}$ is more robust than l_1 penalized FMR in noisy cases. Sparse $l_{2,1}$ improves $l_{2,1}$ allowing component-dependent feature structure. For model inference, we use EM algorithm and propose an efficient update scheme based on re-parameterization. Finally, sparse $l_{2,1}$ is applied to a real world dataset with an improvement over full linear model as well as feature selection. Future work includes extending the sparse $l_{2,1}$ to mixtures of logistic or Poisson regressions.

Acknowledgment

This work was supported by the US National Science Foundation grant NSF/CCF 1451316.

References

[1] A.T. Chaganty, P. Liang, Spectral experts for estimating mixtures of linear regressions., in: ICML, 2013, pp. 1040–1048.
[2] X. Chang, Y. Yang, Semisupervised feature analysis by mining correlations among multiple tasks, IEEE Trans. Neural Netw. Learn. Syst. (2017).
[3] X. Chen, F. Nie, G. Yuan, J.Z. Huang, Semi-supervised feature selection via rescaled linear regression,
[4] P.K. Dunstan, S.D. Foster, F.K. Hui, D.I. Warton, Finite mixture of regression modeling for high-dimensional count and biomass data in ecology, J. Agric. Biol. Environ. Stat. 18 (3) (2013) 357–375.
[5] C. Gong, Exploring commonality and individuality for multi-modal curriculum learning., in: AAAI, 2017, pp. 1926–1933.
[6] M. Henry, Y. Kitamura, B. Salanié, Partial identification of finite mixtures in econometric models, Quant. Econ. 5 (1) (2014) 123–144.
[7] M. Hurn, A. Justel, C.P. Robert, Estimating mixtures of regressions, J. Comput. Graph. Stat. 12 (1) (2003) 55–79.
[8] L. Jacob, G. Obozinski, J.-P. Vert, Group lasso with overlap and graph lasso, in: ICML, 2009, pp. 433–440.
[9] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton, Adaptive mixtures of local experts, Neural Comput. 3 (1) (1991) 79–87.
[10] W. Jiang, M.A. Tanner, Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation, Ann. Stat. (1999) 987–1011.
[11] M.I. Jordan, R.A. Jacobs, Hierarchical mixtures of experts and the em algorithm, Neural Comput. 6 (2) (1994) 181–214.
[12] A. Khalili, J. Chen, Variable selection in finite mixture of regression models, J. Am. Stat. Assoc. 12 (479) (2007) 1025–1038.
[13] J. Kim, H.S. Mahmassani, A finite mixture model of vehicle-to-vehicle and day–to-day variability of traffic network travel times, Transp. Res. Part C 46 (2014) 83–97.
[14] Y. Liu, F. Nie, J. Wu, L. Chen, Efficient semi-supervised feature selection with noise insensitive trace ratio criterion, Neurocomputing 105 (2013) 12–18.
[15] G. McLachlan, D. Peel, Finite Mixture Models, John Wiley & Sons, 2004.
[16] L. Meier, S. Van De Geer, P. Bühlmann, The group lasso for logistic regression, J. R. Stat. Soc. 70 (1) (2008) 53–71.
[17] Y. Nesterov, Efficiency of coordinate descent methods on huge-scale optimization problems, SIAM J. Optim. 22 (2) (2012) 341–362.
[18] F. Nie, H. Huang, X. Cai, C.H. Ding, Efficient and robust feature selection via joint 2, 1-norms minimization, in: NIPS, 2010, pp. 1813–1821.
[19] F. Nie, S. Xiang, Y. Jia, C. Zhang, S. Yan, Trace ratio criterion for feature selection., in: AAAI, 2, 2008, pp. 671–676.
[20] F. Nie, W. Zhu, X. Li, et al., Unsupervised feature selection with structured graph optimization., in: AAAI, 2016, pp. 1302–1308.
[21] R.E. Quandt, J.B. Ramsey, Estimating mixtures of normal distributions and switching regressions, J. Am. Stat. Assoc. 73 (364) (1978) 730–738.
[22] N. Rao, R. Nowak, C. Cox, T. Rogers, Classification with the sparse group lasso, IEEE Trans. Signal Process. 64 (2) (2016) 448–463.
[23] N. Städler, P. Bühlmann, S. Van De Geer, 1-penalization for mixture regression models, Test 19 (2) (2010) 209–256.
[24] Y. Sun, S. Ioannidis, A. Montanari, Learning mixtures of linear classifiers., in: ICML, 2014, pp. 721–729.
[25] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B (Methodol.) (1996) 267–288.
[26] P. Tseng, S. Yun, A coordinate gradient descent method for nonsmooth separable minimization, Math. Program. 117 (1) (2009) 387–423.
[27] S. Xiang, F. Nie, G. Meng, C. Pan, C. Zhang, Discriminative least squares regression for multiclass classification and feature selection, IEEE Trans. Neural. Netw. Learn. Syst. 23 (11) (2012) 1738–1754.
[28] Y. Yang, Z. Ma, A.G. Hauptmann, N. Sebe, Feature selection for multimedia analysis by sharing information among multiple tasks, IEEE Trans. Multimedia 15 (3) (2013) 661–669.
[29] Y. Yang, H.T. Shen, Z. Ma, Z. Huang, X. Zhou, l2, 1-norm regularized discriminative feature selection for unsupervised learning, in: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, 22, 2011, p. 1589.
[30] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J R. Stat. Soc. 68 (1) (2006) 49–67.