Computational Statistics and Data Analysis 144 (2020) 106883

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Structured analysis of the high-dimensional FMR model

Mengque Liu [a], Qingzhao Zhang [b,c], Kuangnan Fang [b], Shuangge Ma [b,d,*]

[a] School of Journalism and New Media, Xi'an Jiaotong University, China
[b] School of Economics, Xiamen University, China
[c] Wang Yanan Institute for Studies in Economics, Xiamen University, China
[d] Department of Biostatistics, Yale University, United States of America

ARTICLE INFO

Article history:
Received 5 February 2019
Received in revised form 31 October 2019
Accepted 2 November 2019
Available online 13 November 2019

Keywords:
Finite mixture of regression model
Structure of covariate effect
High-dimensional data

ABSTRACT

The finite mixture of regression (FMR) model is a popular tool for accommodating data heterogeneity. In the analysis of FMR models with high-dimensional covariates, it is necessary to conduct regularized estimation and identify important covariates rather than noises. In the literature, there has been a lack of attention paid to the differences among important covariates, which can lead to the underlying structure of covariate effects. Specifically, important covariates can be classified into two types: those that behave the same in different subpopulations and those that behave differently. It is of interest to conduct structured analysis to identify such structures, which will enable researchers to better understand covariates and their associations with outcomes. Specifically, the FMR model with high-dimensional covariates is considered. A structured penalization approach is developed for regularized estimation, selection of important variables, and, equally importantly, identification of the underlying covariate effect structure. The proposed approach can be effectively realized, and its statistical properties are rigorously established. Simulation demonstrates its superiority over alternatives. In the analysis of cancer gene expression data, interesting models/structures missed by the existing analysis are identified.

1. Introduction

Heterogeneous data are not uncommon. As extensively explored in the literature, multiple factors can lead to heterogeneity. When only a small proportion of subjects behave differently and such subjects are not of interest, robust estimation can be conducted, focusing on the majority of homogeneous subjects. When all subjects are of interest, mixture models have been commonly adopted. In the context of regression analysis, finite mixture of regression (FMR) models have been popular and extensively used in biology, genetics, engineering, marketing, and other fields. For relevant discussions on the methodology and application of FMR, we refer to Wedel and Desarbo (1995, 2000), McLachlan and Peel (2000).

In early FMR studies, only a small number of covariates was present, and the focus was mostly on estimation and inference. With the fast development of data collection techniques, high-dimensional covariates are now routinely encountered. To accommodate high data dimensionality, regularized estimation is usually needed. In addition, among a large number of covariates, usually only a subset is relevant, and it is necessary to distinguish important covariates from noises. There have been extensive developments in high-dimensional estimation and selection in the past decades.

* Correspondence to: 60 College Street, LEPH 206, New Haven, CT, 06520, United States
E-mail address: shuangge.ma@yale.edu (S. Ma).

https://doi.org/10.1016/j.csda.2019.106883
0167-9473/

We refer to Hastie et al. (2008), Frommlet et al. (2016), and references therein. The analysis of FMR models with high-dimensional covariates can be more challenging because of the additional complexity brought by data heterogeneity, complex likelihood surface, and other factors (Khalili and Lin, 2013). In the literature, a relevant study is Khalili and Chen (2007), which proposed a class of penalty functions for variable selection and regularized estimation as well as an EM algorithm for numerical optimization. In this study, asymptotics were established assuming a fixed number of covariates. Städler et al. (2010) argued that a new parameterization could lead to more efficient computation with high-dimensional data. In this study, asymptotic results were established first assuming a fixed number of covariates. In addition, theoretical investigations were also conducted for Lasso-type penalties under high-dimensional settings with general nonconvex and smooth loss functions. Khalili and Lin (2013) further extended this to a general family of FMR models. In this work, the number of covariates was allowed to grow with the sample size in a polynomial manner, and a mixture of sparse penalties and a ridge penalty was proposed. Statistical investigations were conducted under regularity conditions weaker than those in Städler et al. (2010).

In the aforementioned and other high-dimensional studies, the focus of variable selection has been on distinguishing between important covariates and unimportant ones. Comparatively, there has been much less attention paid to the critical question of "*what causes different subpopulations to behave differently*?". Specifically, covariates can be classified into three categories: (i) those that are not associated with the response in any subpopulation; (ii) those that are associated with the response in the same manner in all subpopulations. In the context of regression analysis, they have the same nonzero regression coefficients in all regression models. With a slight abuse of terminology, we refer to such covariates as "homogeneous"; and (iii) those that are associated with the response in different manners in different subpopulations and directly cause heterogeneity. We refer to such covariates as "heterogeneous". Identifying such covariate categories is equivalent to quantifying the *structure of covariate effects*. Homogeneous and heterogeneous covariates have significantly different implications. Homogeneous covariates describe the shared properties (i.e., "commonality") of all subjects, whereas heterogeneous covariates determine the mixture of subjects and their differences. As such, identifying the structure of covariate effects can significantly advance our understanding of covariates and their associations with the response. In the literature, of relevance are a handful of recent studies on the structure of covariate effects in cure rate models (Fan et al., 2017). Penalization has been adopted for variable selection, estimation, and identification of the covariate effect structure. Cure rate models differ significantly from FMR models. More importantly, in the existing cure rate model studies, statistical properties have not been established.

In this article, we consider heterogeneous data with high-dimensional covariates that can be described using FMR models. Like the literature, we conduct regularized estimation and variable selection. Significantly advancing from the literature, our objective also includes identifying the underlying structure of covariate effects, that is, distinguishing the homogeneous covariates from the heterogeneous ones. This effort can greatly advance our understanding of covariates and their relationships with the response variable. Although sharing a related scheme with the recent cure rate model studies, the modeling, proposed approach, and computation in our study are significantly different. In addition, statistical properties are rigorously established, which can provide important insights into the proposed method as well as other high-dimensional FMR models. With an intuitive formulation, solid statistical basis, and satisfactory numerical performance, this study can provide a useful new venue for analyzing commonly encountered heterogeneous data.

The rest of the article is organized as follows. In Section 2, we describe the proposed method, its computational algorithm, and the statistical properties. Simulation in Section 3 and data analysis in Section 4 demonstrate the competitive practical performance of the proposed method. A brief discussion is provided in Section 5. Technical details and additional numerical results are provided in Appendix A.

2. Methods

2.1. Data, model, and estimation

Denote Y as the response variable and $\mathbf{X}$ as the length p vector of covariates. The conditional density of Y given $\mathbf{X}$ has the form

$$f_{\xi}(Y|\mathbf{X}) = \sum_{q=1}^{Q} \mu_q g\left(Y; h(\beta_{q0} + \mathbf{X}^T\beta_q), \sigma_q\right). \tag{1}$$

Here Q is the number of mixture components (subpopulations); μ_q's are the mixture weights and satisfy $\mu_q \geq 0$ and $\sum \mu_q = 1$; g is the known density function; h is the known link function; β_{q0} is the unknown intercept; $\beta_q = \left(\beta_{q1}, \ldots, \beta_{qp}\right)^T$ is the length p regression parameter vector; σ_q is an unknown parameter usually corresponding to variance; and $\xi = \left(\mu_1, \ldots, \mu_{Q-1}, \beta_{10}, \ldots, \beta_{Q0}, \beta_1^T, \ldots, \beta_Q^T, \sigma_1, \ldots, \sigma_Q\right)^T$ is the vector of all unknown parameters.

In the literature, multiple data distributions have been considered, including Binomial, Gaussian, Poisson, and others (Khalili and Lin, 2013). Here we use the popular Gaussian distribution as an example and note that the proposed analysis can also be conducted with other distributions. In the practical data analysis, Q, the number of mixture components, needs to be determined and may not be trivial. This problem has been examined multiple times. See, for example, Hafidi and Mkhadri (2010) and references therein. We refer to the literature for determining Q and will not

expand on it here. In what follows, we consider the representative case with $Q = 2$. In (1), the intercept terms are of less interest. In estimation, they are not subject to penalization. For the simplicity of notation, we omit them except when absolutely necessary.

Following Städler et al. (2010), we conduct a reparameterization of (1), which can lead to a penalized estimate scale-invariant and easier to compute. Define the new parameters

$$\phi_q = \beta_q/\sigma_q,\ \ \rho_q = \sigma_q^{-1},\ \ q = 1, 2.$$

Denote the new vector of unknown parameters as $\theta = (\phi_1, \phi_2, \rho_1, \rho_2, \mu_1)$. Assume n independent observations $\mathbb{Z} = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$. Then the log-likelihood function is

$$l(\theta; \mathbb{Z}) = n^{-1} \sum_{i=1}^{n} log \left(\sum_{q=1}^{2} \mu_q \frac{\rho_q}{\sqrt{2\pi}} e^{-\frac{1}{2}(\rho_q y_i - \mathbf{x}_i^T \phi_q)^2} \right).$$

We propose the penalized estimate

$$\check{\theta}_\lambda^\gamma = \arg\min \left\{ -l(\theta; \mathbb{Z}) + \lambda_1 \sum_{q=1}^{2} \mu_q^\gamma \left\| \phi_q \right\|_1 + \lambda_2 \sum_{j=1}^{p} I(\phi_{1j} \neq \phi_{2j}) \right\}, \tag{2}$$

where $\lambda = (\lambda_1, \lambda_2)$ are data-dependent tunings, γ is a parameter designed to accommodate (un)balance in data, $\mu_2 = 1 - \mu_1$, $\|\cdot\|_1$ is the ℓ_1 norm, and $I(\cdot)$ is the indicator function. Important and unimportant covariates correspond to the nonzero and zero components of $\check{\phi}_q$'s, respectively. If $\check{\phi}_{1j} = \check{\phi}_{2j} \neq 0$, then covariate j is a homogeneous one. Heterogeneous covariates can be identified accordingly. The mixture probability can be inferred from $\check{\mu}_1$. Specifically, for a certain subject, its posterior probability of belonging to a particular subpopulation can be calculated using the Bayesian rule. The formula is provided in (4) using the obtained estimates.

The indicator function is not continuous, making optimization challenging. To simplify computation, we further propose the estimate

$$\hat{\theta}_\lambda^\gamma = \arg\min \left\{ -l(\theta; \mathbb{Z}) + \lambda_1 \sum_{q=1}^{2} \mu_q^\gamma \left\| \phi_q \right\|_1 + \lambda_2 \sum_{j=1}^{p} \left[1 - e^{-\frac{(\phi_{1j} - \phi_{2j})^2}{\tau}} \right] \right\}, \tag{3}$$

where τ is a small positive number that controls the goodness of the approximation. Note that with this approximation, for homogeneous covariate j, $\hat{\phi}_{1j}$ and $\hat{\phi}_{2j}$ may not be exactly equal. We conclude that a covariate is homogeneous if its estimates in the two subpopulations are sufficiently close.

To accommodate cases with $Q > 2$, we propose further extending the second penalty as $\lambda_2 \sum_{j=1,\ldots,p;\, 1 \leq q_1 \leq q_2 \leq Q} I(\phi_{q_1 j} \neq \phi_{q_2 j})$, which can be approximated with $\lambda_2 \sum_{j=1,\ldots,p;\, 1 \leq q_1 \leq q_2 \leq Q} \left[1 - e^{-\frac{(\phi_{q_1 j} - \phi_{q_2 j})^2}{\tau}} \right]$. Note that here the definitions of homogeneous and heterogeneous covariates may get more complicated. For example, a covariate may have equal nonzero coefficients in some but not all Q datasets. That is, there is a possibility of "partially homogeneous" covariates. With $Q > 2$, as the newly added penalty has a pair-wise form, it is expected that the computational algorithm developed below can be applied with very minor revisions (details omitted).

Rationale Under the settings of Städler et al. (2010), the reparameterization, although seemingly simple and straightforward, has multiple advantages. Such advantages are "inherited" by the proposed estimate. In (2), the first penalty is Lasso, which can be replaced by other sparse penalties. The μ_q^γ term is included to accommodate data (un)balance. Specifically, when data is not too far from balance (that is, the two subpopulations have similar sizes), then $\gamma \sim 0$ (or strictly =0) can generate satisfactory results. For highly unbalanced data, $\gamma \in \{1/2, 1\}$ may generate better numerical results, at the price of a more challenging optimization problem.

The main advancement is the indicator penalty term, which encourages the two coefficients of a covariate to have exactly the same value, leading to the identification of homogeneous covariates. For heterogeneous covariates, it is of little interest how different their covariate effects are in the two subpopulations. As such, the penalty is designed to be not dependent on the magnitudes (if different). It may seem that the approximation makes the newly proposed penalty depend on magnitudes again. It is noted that when τ is small, the approximated penalty is very close to the indicator function and quite insensitive with regard to magnitudes. Shrinking the differences of regression coefficients has been pursued in fused penalization (Tibshirani et al., 2005), Laplacian penalization (Liu et al., 2013), and other studies. However, these penalties are all directly dependent on the magnitudes of differences (of parameters). They are not appropriate in the present context, as shrinking the magnitude differences for heterogeneous covariates is not obviously needed. There are also studies with sign/indicator-based penalties (Huang et al., 2017; Dicker et al., 2013). However, the contexts of these studies are significantly different from the present one. As such, development in this article is warranted. In (3), the approximation is developed to simplify computation. Note that several other approximations have also been proposed in the literature for the indicator function. The approximation is not "free" and requires an additional cutoff (for concluding equal coefficients). This price is shared by other approximation approaches. In the practical data analysis, we choose the cutoff to have a magnitude much smaller than that of the nonzero estimates, so that the main conclusions are not swayed.

Table 1
Toy example: estimation results.

True		MLasso		Proposed	
β_1	β_2	β_1	β_2	β_1	β_2
−1	−1	−0.72	−0.91	**−0.93**	**−0.93**
−1	−1	−0.72	−0.66	−0.84	−0.68
−1	−1	−0.65	−0.76	**−0.79**	**−0.79**
4	−1	3.76	−0.97	3.91	−0.98
4	−1	3.72	−0.82	3.83	−0.86
0	0	0	0	0	0
0	0	0	0	0	0
.	.	.	.	.	.
.	.	.	.	.	.

A toy example To better appreciate the working characteristics of the proposed method, we simulate one dataset with $n = 100$. There are two subpopulations with equal sizes. For each subject, we simulate 20 covariates from a multivariate normal distribution. The response variable of each subject is generated from a linear regression model. The two subpopulations satisfy different regression models, and the true regression coefficients are presented in Table 1. Out of the 20 covariates, five are important, and of those there are three homogeneous and two heterogeneous ones. Beyond the proposed approach, we also consider the approach "FMRLasso" proposed by Städler et al. (2010), which applies Lasso to an FMR model, as an alternative (referred to as "MLasso" for simplicity). It is noted that this approach is just the proposed approach with $\lambda_2 = 0$. This comparison may directly establish the merit of the newly added indicator-based penalty. The estimation results are shown in Table 1. Both approaches can correctly identify the important covariates for this specific simulation replicate. The proposed approach identifies two out of three homogenous covariates, while MLasso fails to distinguish between the homogenous and heterogeneous ones. The proposed approach also seems to have more accurate estimation. More definitive results based on larger-scale simulations are presented in Section 3.

2.2. Computation

We develop a GEM (generalized Expectation–Maximization) algorithm for optimization. In the "standard" EM algorithm, sometimes the complete-data maximum likelihood estimation can be overly complicated. One way to reduce computational complexity is to increase the value of the objective function rather than maximizing it in each M-step, leading to the GEM technique (Dempster et al., 1977). It is noted that the procedure in Lloyd-Jones et al. (2018), which is constructed based on the minorization–maximization technique, may also be adapted and applied here.

For subject $i(= 1, \ldots, n)$, denote $(\Delta_{i,1}, \Delta_{i,2})$ as the unobserved mixture membership indicators. Specifically, $\Delta_{i,1} = 1$ if subject i belongs to the first subpopulation, and 0 otherwise. $\Delta_{i,2}$ is defined for the second subpopulation in the same way. Denote $\Delta = \{(\Delta_{i,1}, \Delta_{i,2}) : i = 1, \ldots, n\}$. The complete-data log-likelihood function is

$$l_c(\theta; \mathbb{Z}, \Delta) = \sum_{i=1}^{n}\sum_{q=1}^{2}\left\{\Delta_{i,q}\log\left(\frac{\rho_q}{\sqrt{2\pi}}e^{-\frac{1}{2}(\rho_q y_i - \mathbf{x}_i^T\phi_q)^2}\right) + \Delta_{i,q}\log(\mu_q)\right\}.$$

Consider the expectation

$$W(\theta|\theta') = -n^{-1}E_{\theta'}[l_c(\theta; \mathbb{Z}, \Delta)|\mathbb{Z}],$$

and its penalized counterpart

$$W_{pen}(\theta|\theta') = W(\theta|\theta') + \lambda_1\sum_{q=1}^{2}\mu_q^{\gamma}\left\|\phi_q\right\|_1 + \lambda_2\sum_{j=1}^{p}\left[1 - e^{-\frac{(\phi_{1j}-\phi_{2j})^2}{\tau}}\right].$$

The proposed algorithm iterates between the following E- and M-steps. Denote the parameter estimate at the mth iteration as $\theta^{(m)}$. To obtain the initial value $\theta^{(0)}$, a simple clustering can be conducted, and Lasso estimation is then conducted for each subpopulation separately.

E-Step: For $q = 1, 2,\ i = 1, \ldots, n$, compute

$$\hat{\delta}_{i,q} = E_{\theta^{(m)}}[\Delta_{i,q}|\mathbb{Z}] = \frac{\mu_q^{(m)}\rho_q^{(m)}e^{-\frac{1}{2}\left(\rho_q^{(m)}y_i - \mathbf{x}_i^T\phi_q^{(m)}\right)^2}}{\mu_1^{(m)}\rho_1^{(m)}e^{-\frac{1}{2}\left(\rho_1^{(m)}y_i - \mathbf{x}_i^T\phi_1^{(m)}\right)^2} + \mu_2^{(m)}\rho_2^{(m)}e^{-\frac{1}{2}\left(\rho_2^{(m)}y_i - \mathbf{x}_i^T\phi_2^{(m)}\right)^2}}. \quad (4)$$

Generalized M-Step: Optimize $W_{pen}(\theta|\theta^{(m)})$ with respect to θ.

(a) Optimize with respect to $\mu = (\mu_1, \mu_2)$:

Fix ϕ_q's at the current estimates $\phi_q^{(m)}$'s, and optimize

$$-n^{-1}\sum_{i=1}^{n}\sum_{q=1}^{2}\hat{\delta}_{i,q}\log(\mu_q) + \lambda_1\sum_{q=1}^{2}\mu_q^{\gamma}\left\|\phi_q^{(m)}\right\|_1. \tag{5}$$

Denote by $\bar{\mu}_q^{(m+1)} = \frac{\sum_{i=1}^{n}\hat{\delta}_{i,q}}{n}$. Update μ_1 as $\mu_1^{(m+1)} = \mu_1^{(m)} + t^{(m)}(\bar{\mu}_1^{(m+1)} - \mu_1^{(m)})$, where $t^{(m)} \in (0, 1]$. In practice, $t^{(m)}$ is chosen to be the largest value in the grid $\{\zeta^{0,1,2,\cdots}\}$ $(0 < \zeta < 1)$ such that the value of (5) does not increase. For the special case with $\gamma = 0$, minimization with respect to μ_q is achieved with $\mu_q^{(m+1)} = \frac{\sum_{i=1}^{n}\hat{\delta}_{i,q}}{n}$. That is, $t^{(m)} = 1$.

(b) Coordinate descent optimization with respect to ϕ and ρ. Consider

$$\sum_{q=1}^{2}-\frac{n_q}{n}\log(\rho_q) + \frac{1}{2n}\left\|\rho_q\tilde{y} - \tilde{\mathbf{x}}^T\phi_q\right\|^2 + \lambda_1\sum_{q=1}^{2}\mu_q^{\gamma}\left\|\phi_q\right\|_1 + \lambda_2\sum_{j=1}^{p}\left[1 - e^{-\frac{(\phi_{1j}-\phi_{2j})^2}{\tau}}\right], \tag{6}$$

where $\tilde{y}$ and $\tilde{\mathbf{x}}$ are composed of $(\tilde{y}_i, \tilde{\mathbf{x}}_i)$'s and $(\tilde{y}_i, \tilde{\mathbf{x}}_i) = \sqrt{\hat{\delta}_{i,q}}(y_i, \mathbf{x}_i)$. Also $\|\cdot\|^2$ represents the l_2-norm, and $n_q = \sum_{i=1}^{n}\hat{\delta}_{i,q}$, $q = 1, 2$.

As opposed to fully optimizing (6), we minimize it in a coordinate-wise manner, update one coordinate, and hold the other coordinates at their current estimates. The closed-form coordinate updates can be obtained as

$$\rho_q^{(m+1)} = \frac{\left\langle\tilde{y}, \tilde{\mathbf{x}}^T\phi_q^{(m)}\right\rangle + \sqrt{\left\langle\tilde{y}, \tilde{\mathbf{x}}^T\phi_q^{(m)}\right\rangle^2 + 4\left\|\tilde{y}\right\|^2 n_q}}{2\left\|\tilde{y}\right\|^2},\ q = 1, 2,$$

where $\langle\cdot,\cdot\rangle$ represents the inner product.

Denote $\tilde{\boldsymbol{x}}_j$ and $\tilde{\boldsymbol{x}}_r$ as the jth and rth columns of $\tilde{\mathbf{x}}$, respectively. Then we have:

$$\begin{aligned}
\phi_{1,j}^{(m+1)} &= \begin{cases}(-M_{1,j} - n\lambda_1(\mu_1^{(m+1)})^{\gamma} + L\phi_{2,j}^{(m)})/(\left\|\tilde{\boldsymbol{x}}_j\right\|^2 + L) & if \quad L\phi_{2,j}^{(m)} > M_{1,j} + n\lambda_1(\mu_1^{(m+1)})^{\gamma},\\ (-M_{1,j} + n\lambda_1(\mu_1^{(m+1)})^{\gamma} + L\phi_{2,j}^{(m)})/(\left\|\tilde{\boldsymbol{x}}_j\right\|^2 + L) & if \quad L\phi_{2,j}^{(m)} < M_{1,j} - n\lambda_1(\mu_1^{(m+1)})^{\gamma},\\ 0 & otherwise,\end{cases}\\
\phi_{2,j}^{(m+1)} &= \begin{cases}(-M_{2,j} - n\lambda_1(\mu_1^{(m+1)})^{\gamma} + L\phi_{1,j}^{(m)})/(\left\|\tilde{\boldsymbol{x}}_j\right\|^2 + L), & if \quad L\phi_{1,j}^{(m)} > M_{2,j} + n\lambda_1(\mu_1^{(m+1)})^{\gamma}\\ (-M_{2,j} + n\lambda_1(\mu_1^{(m+1)})^{\gamma} + L\phi_{1,j}^{(m)})/(\left\|\tilde{\boldsymbol{x}}_j\right\|^2 + L), & if \quad L\phi_{1,j}^{(m)} < M_{2,j} - n\lambda_1(\mu_1^{(m+1)})^{\gamma}\\ 0 & otherwise,\end{cases}
\end{aligned} \tag{7}$$

where $L = 2n\lambda_2 e^{-\frac{(\phi_{1,j}^{(m)}-\phi_{2,j}^{(m)})^2}{\tau}}/\tau$, $M_{q,j}(q = 1, 2)$ is defined as

$$M_{q,j} = -\rho_q^{(m+1)}\left\langle\tilde{\boldsymbol{x}}_j, \tilde{y}\right\rangle + \sum_{r<j}\phi_{q,r}^{(m+1)}\left\langle\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{x}}_r\right\rangle + \sum_{r>j}\phi_{q,r}^{(m)}\left\langle\tilde{\boldsymbol{x}}_j, \tilde{\boldsymbol{x}}_r\right\rangle,$$

and $j = 1, \ldots, p$.

Convergence is examined in Appendix A. It is established that the proposed algorithm converges to a stationary point. In each numerical analysis, convergence is satisfactorily achieved with a moderate number of iterations. (λ_1, λ_2) are chosen using a modified BIC criterion with the degree of freedom defined as the effective number of parameters (Pan and Shen, 2007), that is, $df = 2 + (2 - 1) + \sum_{j=1,\ldots,p;q=1,2} I\left(\hat{\phi}_{q,j} \neq 0\right)$. For τ, our simulation suggests that results are not sensitive to its value. We set $\tau = 0.01$ and note that, to be cautious, other values may also be considered in practice. Notably, the proposed computational algorithm is affordable. For one simulation replicate (details described below), computation can be accomplished within four minutes on a regular desktop. To facilitate data analysis within and beyond this study, we have developed an R program that implements the proposed approach and algorithm. The program is publicly available at www.github.com/shuanggema.

2.3. Statistical properties

Denote $\phi = (\phi_1, \phi_2)$ and the true value of $\theta = (\phi_1, \phi_2, \rho_1, \rho_2, \mu_1)$ as θ^*. Define the sets $J_1 = \left\{j : \phi_{1,j}^* \neq \phi_{2,j}^*\right\}$, $J_2 = \left\{j : \phi_{1,j}^* = \phi_{2,j}^* \neq 0\right\}$, $J_3 = \left\{j : \phi_{1,j}^* = \phi_{2,j}^* = 0\right\}$, and $S = \left\{(q, j) : \phi_{q,j}^* \neq 0\right\}$. J_1, J_2, and J_3 are disjoint sets and contain the indexes of heterogeneous, homogeneous, and irrelevant covariates, respectively. S is the active set. Let $j_1 = |J_1|, j_2 = |J_2|, j_3 = |J_3|$, and $s = |S|$, where $|\cdot|$ denotes cardinality. Consider the estimate defined in (3), with $\gamma = 0$. The following conditions are assumed.

Condition 1. *The parameter space is* $\Theta = \{\theta : \sup_i \max_q |\mathbf{x}_i^T \phi_q| \leq K, \max_q |\log \rho_q| \leq K, -K \leq \log \mu_1 < 0\}$, *where* K *is a fixed constant.*

Condition 2. *There exists a constant* $k \geq 1$ *such that, for all* ϕ *satisfying* $\|\phi_{S^c}\|_1 \leq 6\|\phi_S\|_1$, $\|\phi_S\|_2^2 \leq k^2 \sum_{q=1}^2 \phi_q^T \Sigma_n \phi_q$, *where* S^c *is the complement of* S *and* $\Sigma_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$.

Recall that the conditional density of Y given $\mathbf{X}$ in (1) depends on $\mathbf{X}$ through the parameter θ. As such, we can write it as $f_\theta(Y|\mathbf{X})$. Define the excess risk as

$$\varepsilon(\theta|\theta^*) = -\int \log\left[\frac{f_\theta(Y|\mathbf{X})}{f_{\theta^*}(Y|\mathbf{X})}\right] f_{\theta^*(\mathbf{X})}(Y)dY,$$

which is the Kullback–Leibler distance between the true and estimated density functions. Further define the empirical excess risk as

$$\bar{\varepsilon}(\theta|\theta^*) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i(\theta|\theta^*).$$

Condition 1 assumes that the parameter space Θ is a bounded subset of some finite-dimensional space. It automatically holds for FMR models. In the Basic Inequality that is based on the fact that the proposed estimator is a penalized excess risk minimizer, the excess risk, which stands for prediction performance, is bounded by the empirical process involving the random error. We introduce Lasso to overrule the empirical process, which can be bounded in terms of the l_1-norm of the parameters involved. The l_1 error can be bounded by the l_2 one. Condition 2 is the restricted eigenvalue condition and puts constraints on the Gram matrix that provides the bound for the l_2 error. It was formalized by Bickel et al. (2009) and is among the weakest and most general conditions that can be imposed on the Gram matrix in order to achieve satisfactory properties under Lasso. Van De Geer and Bühlmann (2009) showed that even for fixed designs (where empirical covariance matrices are singular), the collection of cases under which Condition 2 holds is quite large. We can establish the following finite-sample oracle results for the proposed approach.

Theorem 1. *Denote* c_1, c_2, *and* c_3 *as constants depending on* K. *Assume that Conditions 1 and 2 hold, and that* $\lambda_1 \geq 2T\lambda_0$ *with* $\lambda_0 = c_1\sqrt{\log^3 n \log(p \vee n)/n}$ *and* $T \geq 1$. *Then, with probability at least* $1 - c_2\exp(-\log^2 n \log(p \vee n)) - n^{-1}$, *for all* $n \geq c_3$,

$$\bar{\varepsilon}(\hat{\theta}_\lambda|\theta^*) + 2(\lambda_1 - T\lambda_0)\|\hat{\phi}_{S^c}\|_1 \leq 4(\lambda_1 + T\lambda_0)^2 c_0^2 k^2 s + 4\lambda_2 j_1\{1 - \exp(-\max_j |\phi_{1j}^* - \phi_{2j}^*|^2/\tau)\},$$

where c_0 *is the constant defined in Lemma 1 of Städler et al. (2010).*

With this result, we can obtain $\bar{\varepsilon}(\hat{\theta}_\lambda|\theta^*) \leq 4(\lambda_1 + T\lambda_0)^2 c_0^2 k^2 s + 4\lambda_2 j_1\{1 - \exp(-\max_j |\phi_{1j}^* - \phi_{2j}^*|^2/\tau)\}$, which suggests that the prediction error is of the order $O(s\lambda_0^2 + j_1\lambda_2)$. Prediction consistency can be achieved with proper tunings (for example, when s is bounded and taking $\lambda_1 = 2T\lambda_0$, $\lambda_2 = O(\lambda_1)$). In addition, the noises in S^c have $\|\hat{\phi}_{S^c}\|_1 \leq 2(\lambda_1 + T\lambda_0)c_0^2 k^2 s + 2\lambda_2/(\lambda_1 - T\lambda_0)j_1\{1 - \exp(-\max_j |\phi_{1j}^* - \phi_{2j}^*|^2/\tau)\}$. That is, their estimates are small.

Remarks. This theorem establishes the convergence rate result of the proposed approach. This scheme is similar to that in Städler et al. (2010). This result is "re-assuring" in that, with the newly added penalty, the main properties of the approach in Städler et al. (2010) and others are preserved. We acknowledge that in some high-dimensional penalization studies, asymptotic estimation and variable selection consistency properties are established. However, we note that the strategy in such studies and that targets convergence rate results, as in this study, are often different and not "mixed together". We take the strategy of Städler et al. (2010) and others and do not explore asymptotic variable selection properties. It is noted that, as a limitation of this study, there is a lack of theoretical results that directly establish the advantage of the new approach. This limitation is also shared by, for example, some fused Lasso, Laplacian penalization, and other studies. Still, the satisfactory theoretical results along with the superior numerical results (shown below) can well demonstrate the merit of the proposed approach.

3. Simulation

We simulate heterogeneous data from model (1) with sample size $n = 200$ and a varying p. The design matrix $\mathbf{X}$ has a distribution of $N_p(0, \Sigma)$, where Σ is block-diagonal with block sizes $p/5$. The covariates in different blocks are independent, and the covariates i and j within the same block have correlation coefficient $\varrho^{|i-j|}$. We consider different levels of sparsity, with S1–S3 having four important variables in each model and S4–S9 having six. In addition, in S1–S3, the homogeneous variables have strong effects, whereas in S4–S9, the homogeneous effects are weak. In S1–S6, important variables occupy the same "positions" in the two regression models, whereas in S7–S9, there are two variables that are only important in the first model, and two others that are only important in the second model. More detailed information on regression coefficients is provided in Table 5 (Appendix A). Three values of ϱ are considered, representing no, moderate,

Table 2
Simulation results for S1–S3 and the balanced design. In each cell, mean(sd). CIR_{homo}: correct identification rate of homogeneous covariates.

p	Method	S1		S2		S3	
		AUC	CIR_{homo}	AUC	CIR_{homo}	AUC	CIR_{homo}
15	Threshold	1.00(0.00)	1(0)	1.00(0.00)	1(0)	1.00(0.00)	1(0.00))
	Kmeans	0.79(0.02)	0.86(0.23)	0.84(0.01)	0.76(0.34)	0.85(0.03)	0.51(0.38)
	MLasso	0.94(0.01)	1.00(0.00)	1.00(0.02)	1.00(0.00)	1.00(0.00)	1.00(0.00)
	Proposed	0.95(0.00)	1.00(0.00)	0.99(0.00)	1.00(0)	1.00(0.00)	1.00(0.00)
55	Threshold	0.97(0.01)	0.70(0.32)	0.97(0.01)	0.53(0.50)	0.99(0.00)	0.56(0.42)
	Kmeans	0.76(0.02)	0.70(0.35)	0.79(0.03)	0.45(0.39)	0.82(0.02)	0.58(0.36)
	MLasso	0.97(0.01)	0.92(0.23)	0.99(0.00)	0.83(0.27)	1.00(0.00)	0.8(0.31)
	Proposed	0.98(0.00)	0.98(0.09)	1.00(0.00)	0.98(0.09)	1.00(0.00)	0.98(0.09)
75	Threshold	0.91(0.04)	0.46(0.36)	0.93(0.01)	0.45(0.42)	0.92(0.03)	0.41(0.42)
	Kmeans	0.74(0.03)	0.83(0.41)	0.78(0.02)	0.47(0.46)	0.79(0.02)	0.40(0.35)
	MLasso	0.94(0.02)	0.95(0.2)	0.99(0.00)	0.8(0.28)	0.99(0.00)	0.72(0.34)
	Proposed	0.98(0.00)	1.00(0)	1.00(0.00)	1.00(0)	1(0.00)	0.98(0.09)
105	Threshold	0.86(0.04)	0.70(0.37)	1.00(0.00)	0.55(0.36)	1.00(0.00)	0.58(0.37)
	Kmeans	0.82(0.01)	0.77(0.35)	0.87(0.02)	0.41(0.36)	0.88(0.02)	0.57(0.40)
	MLasso	0.94(0.02)	0.52(0.09)	0.99(0.00)	0.53(0.32)	0.99(0.00)	0.49(0.40)
	Proposed	0.98(0.00)	1.00(0.00)	1.00(0.00)	0.84(0.28)	1.00(0.00)	0.87(0.26)
300	Threshold	0.74(0.02)	0.41(0.44)	0.78(0.03)	0.44(0.41)	0.68(0.04)	0.54(0.32)
	Kmeans	0.69(0.03)	0.43(0.33)	0.73(0.04)	0.35(0.37)	0.68(0.01)	0.68(0.4)
	MLasso	0.8(0.02)	0.27(0.29)	0.98(0.02)	0.15(0.3)	0.98(0.01)	0.67(0.56)
	Proposed	0.97(0.00)	1.00(0.00)	1.00(0.00)	0.93(0.22)	1.00(0.00)	0.78(0.38)
500	Threshold	0.88(0.05)	0.53(0.31)	0.82(0.04)	0.44(0.43)	0.69(0.05)	0.36(0.35)
	Kmeans	0.61(0.03)	0.56(0.31)	0.69(0.06)	0.32(0.31)	0.69(0.04)	0.40(0.35)
	MLasso	0.85(0.04)	0.7(0.42)	0.96(0.01)	0.54(0.35)	0.97(0.01)	0.43(0.31)
	Proposed	0.96(0.01)	0.93(0.25)	0.99(0.00)	0.59(0.46)	0.99(0.00)	0.67(0.4)

and strong correlations. We further consider both balanced ($\mu_1 = \mu_2 = 0.5$) and unbalanced ($\mu_1 = 0.2$) cases. Overall, the simulation settings comprehensively cover all important features of mixture regression models with sparsity.

Simulation I is first conducted to "reestablish" the necessity of mixture modeling. Specifically, consider simulation setting S1 and $p = 105, 300, 500$. We consider the proposed approach, MLasso (which conducts "FMRLasso"), and SLasso (which assumes homogeneity and applies Lasso for variable selection and estimation). For each approach, we examine a sequence of tuning parameter values, evaluate variable selection performance at each tuning parameter value, and construct the ROC (receiver operating characteristic) curves. With 100 replicates, the AUC (area under curve) values are computed as SLasso (0.7,0.63,0.61), MLasso (0.94,0.8,0.82), and Proposed (0.98,0.97,0.96) for $p = 105, 300, 500$, respectively. The two mixture modeling approaches have significantly better variable selection performance.

Simulation II is conducted to compare approaches that can accommodate heterogeneity. Beyond the proposed approach and MLasso, we also consider "Threshold" (which assumes the FMR model and applies hard thresholding for regularized estimation and variable selection) and "Kmeans" (which separates samples into two clusters using the Kmeans technique and then applies Lasso to each cluster). To compare performance, we first consider AUC, which is calculated the same way as described above and evaluates the overall variable selection performance. In addition, for all the approaches, we select tuning parameters using the BIC and compute the CIR_{homo}, the correct identification rate of homogeneous covariates. We have also examined other measures, for example, variable selection measures at the BIC-selected tunings, and reached similar conclusions. To avoid redundancy, these measures are not included.

Summary statistics based on 100 replicates for S1–S3 and the balanced design are shown in Table 2. Results for some settings are presented in Appendix A. All simulation settings lead to similar conclusions. Specifically, Kmeans, which separates the identification of heterogeneity and variable selection, has inferior performance. When the data dimension is low, Threshold and Kmeans have competitive performance. However, when the data dimension gets high, the superiority of the proposed approach becomes prominent. For example, in Table 2 with $p = 500$, under setting S1, the four approaches have AUC values of 0.88 (Threshold), 0.61 (Kmeans), 0.85 (MLasso), and 0.96 (Proposed). The CIR_{homo} values are 0.53 (Threshold), 0.56 (Kmeans), 0.7 (MLasso), and 0.93 (Proposed). In the Appendix A, we also present representative ROC plots, where the superiority of the proposed approach in variable selection is clearly evidenced. Other ROC plots are available from the authors.

Simulation III is further conducted to examine the proposed indicator-based penalty. Specifically, we additionally consider the fused Lasso-type penalty, which takes the form $\lambda_2 \sum_{j=1}^{p} |\phi_{1j} - \phi_{2j}|$ and directly penalizes the magnitudes of differences. Representative results for settings S7–S9, $p = 15$, 55, and 75, and the unbalanced design are shown in Fig. 3 (Appendix A). The superiority of the proposed approach over the fused Lasso is clearly shown, with higher ROC curves in almost all plots.

4. Data analysis

We conduct an analysis of two TCGA (The Cancer Genome Atlas) datasets. TCGA is a collective effort of multiple institutes organized by the NIH and has recently published high-quality profiling data on multiple cancer types. TCGA is observational in nature with no strict patient selection standards, which, along with the inherent heterogeneity of cancer, naturally lead to the heterogeneity of samples. The heterogeneity analysis of the TCGA data has already been conducted. See, for example, Lawrence et al. (2013).

4.1. Analysis of cutaneous melanoma (SKCM) data

We first consider the SKCM data. In the literature (Jiang et al., 2016; Chai et al., 2017), the regulation of Breslow thickness, which is an important biomarker for prognosis and other outcomes, by gene expressions has been explored. However, the existing studies have assumed homogeneity, and there has been a lack of attention paid to potential heterogeneity.

The data is downloaded from the TCGA data portal. We select the 170 samples with the AJCC pathologic tumor stages being II and III. For these samples, we download 18,947 gene expression measurements. More specifically, the processed level-3 gene expression data is used. We refer to the literature (Molony et al., 2009) for detailed information on the generation and processing of gene expression data. To improve interpretability, we further identify 4,243 genes with well-defined KEGG pathway information. Each gene expression is then normalized to have a mean of zero and a variance of one. In principle, it is possible to directly apply the proposed approach to all genes. However, with the small sample size and additional complexity brought by heterogeneity, such an analysis may not be reliable. We further conduct a marginal screening based on correlation and select the top 300 genes for downstream analysis.

The analyzed samples have different stages. It is noted that such heterogeneity is defined mainly using pathological characteristics. In our analysis, we are interested in the heterogeneity in the regulation of Breslow thickness by gene expressions. A literature search does not suggest whether such heterogeneity and the pathology-based one (and other types) are linked or not. As the proposed approach is designed to accommodate heterogeneity, mixing samples with different stages does not pose a problem. Also because it is unclear whether stage or another variable plays a role in the Breslow thickness-gene expression regulation, we do not adjust for other variables.

The data is analyzed using the proposed approach and alternatives. For all approaches, tuning parameters are selected using the modified BIC criterion described in Section 2.2. Detailed estimation results obtained by using the proposed approach are provided in Table 3. In particular, 49 genes are identified as associated with Breslow thickness, among which 32 are identified as heterogeneous. For this dataset, the heterogeneous effects have the same signs but different magnitudes. Some magnitudes (of the same gene effect) can be quite different. For example, gene CRELD2 has estimates of 0.11 and 0.23 in the two subpopulations. It is observed that different approaches lead to different findings. Summary comparison results are provided in Table 9 (Appendix A), and detailed estimation results obtained by using the alternatives are available from the authors.

To complement the identification/estimation analysis, we also apply a resampling-based approach and evaluate prediction performance and stability. Specifically, the dataset is randomly divided into training and testing sets, with sizes 9:1. The parameters are estimated only using the training set and then used to make prediction for the samples in the testing set. In addition, the training set estimates are also used to evaluate stability. This approach has been extensively adopted in the literature to provide support for the validity of estimation. With a continuous response, we use the prediction mean squared error (PMSE) to evaluate prediction. The squared roots of the PMSEs are 4.723 (Threshold), 6.72 (Kmeans), 4.697 (MLasso), and 3.728 (Proposed), with the proposed approach having the best prediction. In the stability evaluation, we compute the OOI (observed occurrence index) for each gene. Briefly, the OOI is the probability of a specific gene identified across replicates and measures the stability of a finding. For the identified homogeneous genes, we find the mean OOI values to be 0 (Threshold), 0 (Kmeans), 0.19 (MLasso), and 0.32 (Proposed). The OOI values for individual genes are plotted in Fig. 4 (Appendix A), where the better stability of the proposed approach is clearly shown. The prediction and stability evaluation provide partial support for the superiority of the proposed approach.

4.2. Analysis of lung cancer data

We further conduct an analysis of lung cancer data. As the proposed approach can accommodate data heterogeneity, to increase sample size, we combine the lung adenocarcinoma (LUAD) and lung squamous cell (LUSC) data, both of which are non-small-cell lung cancers. Such data have been analyzed in other studies (Hammerman et al., 2012; Collisson et al., 2014), though under the assumption of homogeneity. In our analysis, we examine the FEV (forced expiratory volume), which measures lung capacity and is an important marker in lung cancer development. A total of 231 samples are available for analysis. Expression data is available for 20,531 genes. We conduct the same processing as with the SKCM data and analyze 300 gene expressions using the proposed and alternative approaches.

Estimation results obtained by using the proposed approach are provided in Table 4. Specifically, a total of 57 genes are identified, of which there are 10 homogeneous and 47 heterogeneous genes. Different from the SKCM analysis, conflicting signs (for the heterogeneous gene effects) are observed. For example, gene DFNA44 has estimates -0.05 and 0.01. This

Table 3
Analysis of SKCM data: estimated coefficients using the proposed approach.

Heterogeneous			Homogeneous	
Gene	β_1	β_2	Gene	$\beta_1(=\beta_2)$
CKMT2	0.04	0.08	KTGNR	0.09
ANKRD20A20P	0.10	0.21	CD55	0.10
LOC115165	0.04	0.05	TMEM244	0.09
C22ORF34	0.04	0.05	FLJ23058	0.08
CRELD2	0.11	0.23	LOC102723772	0.09
C20ORF166AS1	0.00	0.33	LOC155060	0.08
DLAT	0.04	0.08	C19ORF48	−0.09
LOC148922	0.08	0.16	CCDC172	0.11
RBBP8NL	0.05	0.08	BAALC	0.06
LOC51745	0.04	0.06	CENPJ	0.10
ZFAS1	−0.04	−0.07	C10orf114	0.10
LOC101928620	0.04	0.05	BCL9L	0.10
LOC100506602	0.05	0.10	CDKL3	−0.07
CPD	−0.04	−0.08	ARHGDIG	0.08
SMIM20	−0.09	−0.17	ARHGEF17	−0.07
TMEM248	0.03	0.07	BHLHE41	−0.09
TEN1	0.05	0.06	OGFOD3	0.07
ARHGAP31	−0.03	−0.07		
LOC399807	0.03	0.07		
ARSE	0.13	0.24		
CXORF38	0.03	0.07		
CDK2AP1	−0.03	−0.05		
DMXL1	0.08	0.19		
AKR1D1	−0.04	−0.09		
SMIM21	0.01	0.08		
CD151	0.04	0.09		
BCAR3	−0.04	−0.06		
CCDC36	0.04	0.09		
LACC1	0.09	0.20		
ARHGEF7	0.02	0.06		
C7ORF25	−0.10	−0.18		
NUTM1	−0.04	−0.07		

higher level of heterogeneity may be reasonable, as the dataset contains two cancer subtypes. It is noted that most gene effects have consistent signs. Table 10 (Appendix A) again suggests that different approaches lead to different findings. Estimation results obtained by using the alternatives are available from the authors. A prediction and stability evaluation is also conducted. The root PMSEs are 9.012 (Threshold), 10.05 (Kmeans), 8.928 (MLasso) and 7.925 (Proposed), with the proposed approach showing improved prediction. The OOI results are plotted in Fig. 5 (Appendix A). The mean OOI values for the identified homogeneous genes are 0.07 (Threshold), 0 (Kmeans), 0 (MLasso), and 0.25 (Proposed). The improved prediction and stability provide support for the validity of the proposed analysis.

5. Discussion

Under the high-dimensional FMR framework, this study has advanced from existing ones by focusing on the structure of covariate effects. As shown in our data analysis, the proposed approach can separate heterogeneous covariates from homogeneous ones. Such an analysis may have important implications. For example, in the analysis of SKCM data, gene C20ORF166AS1 has estimated effects of <0.01 and 0.33 for the two subpopulations. As such, targeting this gene as a way to affect the Breslow thickness may be effective for only some samples. In general, the heterogeneous covariates define the unique characteristics of subpopulations and may deserve additional attention. Besides taking a unique perspective, this study may have also advanced from the existing literature by rigorously establishing the finite-sample consistency results, which may shed light on other high-dimensional mixture modelings. For the simplicity of notation, we have described the proposed approach using two subpopulations. The proposed methodology and computation can be extended to multiple subpopulations with minor modifications. For practical data analysis under simpler (especially low-dimensional) settings, there are proposals for determining the number of mixtures. However, establishing the validity under high-dimensional settings is expected to be challenging and hence will be postponed to future research. It may also be of interest to examine asymptotic variable selection consistency and normality properties. As discussed in Section 2.3, that investigation would require an analysis scheme different from the present one and will be postponed to future research.

Acknowledgments

We thank the editor and reviewers for their careful reviews and insightful comments, which have led to a significant improvement of this article. This study was partly supported by the NIH, United States of America (CA241699, CA121974,

Table 4
Analysis of lung cancer data: estimated coefficients using the proposed approach.

Heterogeneous					
Gene	β_1	β_2	Gene	β_1	β_2
CCT7	−0.05	−0.02	CDCA7L	−0.06	−0.02
LOC130618	0.07	0.04	HECH	−0.10	−0.02
B3GALT1	−0.08	−0.04	LOC135662	0.17	0.12
B9D2	0.18	0.07	CENPL	0.05	0.03
ARL6IP	−0.13	−0.03	C21ORF2	0.00	0.05
ASCC1	−0.46	0.00	MGME1	0.07	0.00
C2ORF44	0.09	0.03	ADAM11	0.00	0.12
C2ORF42	−0.02	−0.07	PRO1331	−0.14	0.00
CTNNAL1	−0.25	−0.13	DEFA5	−0.21	−0.10
MGC5254	−0.03	−0.06	ACAA1	0.00	0.05
ALPL	0.79	−0.03	LOC196753	0.15	0.10
ARHGAP36	−0.08	−0.01	C6ORF62	0.06	0.04
CPM	0.05	0.04	CYP4V2	0.00	0.09
CBX1	0.06	0.00	ANKRD27	−0.20	−0.08
ARHGEF1	0.00	0.08	AK4	0.21	0.12
ASAH1	0.00	0.06	TIGAR	−0.07	−0.02
CHEK2	0.08	0.00	CCDC15	0.00	−0.14
COLEC10	−0.28	−0.18	LOC196411	0.21	0.08
LOC254023	0.20	0.12	CLDN22	−0.19	−0.11
CSTF2T	−0.07	−0.05	CD151	−0.05	−0.04
VSTM5	−0.08	0.00	LOC648756	0.08	0.02
CHAF1B	0.21	0.09	DFNA44	−0.05	0.01
ATP13A3	0.09	0.00	ARHGEF2	−0.07	0.00
DLG1	0.09	0.04			
Homogeneous					
Gene	$\beta_1(=\beta_2)$		Gene	$\beta_1(=\beta_2)$	
LOC100132855	0.09		C15ORF54	0.19	
C1orf67	−0.12		CACNA2D2	−0.09	
BUB1B	0.08		SPRTN	0.08	
MMS22L	0.08		CNOT10	−0.09	
DNAJB4	−0.06		FLJ32997	−0.19	

and CA196530), NSF, United States of America (1916251), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (19YJC910010), and Fundamental Research Funds for the Central Universities, China (20720171064 and 20720181003).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2019.106883.

References

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. 37 (4), 1705–1732. http://dx.doi.org/10.1214/08-aos620.

Chai, H., Shi, X., Zhang, Q., Zhao, Q., Huang, Y., Ma, S., 2017. Analysis of cancer gene expression data with an assisted robust marker identification approach. Genet. Epidemiol. 41 (6), 779–789. http://dx.doi.org/10.1002/gepi.22066.

Collisson, E., Campbell, J., Brooks, A., Berger, A., Lee, W., Chmielecki, J., Beer, D., Cope, L., Creighton, C., Danilova, L., et al., 2014. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511 (7511), 543–550. http://dx.doi.org/10.1038/nature13385.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1), 1–38. http://dx.doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Dicker, L., Huang, B., Lin, X., 2013. Variable selection and estimation with the seamless-L0 penalty. Statist. Sinica 23 (2), 929–962. http://dx.doi.org/10.5705/ss.2011.074.

Fan, X., Liu, M., Fang, K., Huang, Y., Ma, S., 2017. Promoting structural effects of covariates in the cure rate model with penalization. Stat. Methods Med. Res. 26 (5), 2078–2092. http://dx.doi.org/10.1177/0962280217708684.

Frommlet, F., Bogdan, M., Ramsey, D., 2016. Statistical Methods in High Dimensions. Springer, London, http://dx.doi.org/10.1007/978-1-4471-5310-8_3.

Hafidi, B., Mkhadri, A., 2010. The Kullback information criterion for mixture regression models. Statist. Probab. Lett. 79 (9), 807–815. http://dx.doi.org/10.1016/j.spl.2010.01.014.

Hammerman, P.S., Voet, D., Lawrence, M.S., Voet, D., Jing, R., Cibulskis, K., Sivachenko, A., Stojanov, P., McKenna, A., Lander, E.S., et al., 2012. Comprehensive genomic characterization of squamous cell lung cancers. Nature 489 (7417), 519–525. http://dx.doi.org/10.1038/nature11404.

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning Data Mining, Inference, and Prediction, second ed. Springer, New York, p. 192. http://dx.doi.org/10.1111/j.1751-5823.2009.00095_18.x.

Huang, Y., Zhang, Q., Zhang, S., Huang, J., Ma, S., 2017. Promoting similarity of sparsity structures in integrative analysis with penalization. J. Amer. Statist. Assoc. 112 (517), 342–350. http://dx.doi.org/10.1080/01621459.2016.1139497.

Jiang, Y., Shi, X., Zhao, Q., Krauthammer, M., Rothberg, B.E., Ma, S., 2016. Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. Genomics 107 (6), 223–230. http://dx.doi.org/10.1016/j.ygeno.2016.04.005.

Khalili, A., Chen, J., 2007. Variables selection in finite mixture of regression models. J. Amer. Statist. Assoc. 102 (479), 1025–1038. http://dx.doi.org/10.1198/016214507000000590.

Khalili, A., Lin, S., 2013. Regularization in finite mixture of regression models with diverging number of parameters. Biometrics 69 (2), 436–446. http://dx.doi.org/10.1111/biom.12020.

Lawrence, M.S., Petar, S., Paz, P., Kryukov, G.V., Kristian, C., Andrey, S., Carter, S.L., Chip, S., Mermel, C.H., Roberts, S.A., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499 (7457), 214–218. http://dx.doi.org/10.1038/nature12213.

Liu, J., Huang, J., Ma, S., 2013. Incorporating network structure in integrative analysis of cancer prognosis data. Genet. Epidemiol. 37 (2), 173–183. http://dx.doi.org/10.1002/gepi.21697.

Lloyd-Jones, L.R., Nguyen, H.D., Mclachlan, G.J., 2018. A globally convergent algorithm for lasso-penalized mixture of linear regression models. Comput. Statist. Data Anal. 119, 19–38. http://dx.doi.org/10.1016/j.csda.2017.09.003.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York, http://dx.doi.org/10.1007/978-1-4419-1698-3_100608.

Molony, C., Sieberts, S.K., Schadt, E.E., 2009. Processing Large-Scale, High-Dimension Genetic and Gene Expression Data. Springer, Berlin, Heidelberg, pp. 307–330. http://dx.doi.org/10.1007/978-3-540-69264-5_11.

Pan, W., Shen, X., 2007. Penalized model-based clustering with application to variable selection. 8, pp. 1145–1164. http://dx.doi.org/10.1115/1.2719785,

Städler, N., Bühlmann, P., Van De Geer, S., 2010. l_1-Penalization for mixture regression models. Test 19 (2), 280–285. http://dx.doi.org/10.1007/s11749-010-0197-z.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (1), 91–108. http://dx.doi.org/10.1111/j.1467-9868.2005.00490.x.

Van De Geer, S.A., Bühlmann, P., 2009. On the conditions used to prove oracle results for the lasso. Electron. J. Stat. 3, 1360–1392. http://dx.doi.org/10.1214/09-ejs506.

Wedel, M., Desarbo, W.S., 1995. A mixture likelihood approach for generalized linear models. J. Classification 12 (1), 21–55. http://dx.doi.org/10.1007/bf01202266.

Wedel, M., Desarbo, W.S., 2000. Mixture Regression Models. Springer, Boston, MA, pp. 366–382. http://dx.doi.org/10.1007/978-1-4615-4651-1_7.