Screened Pairwise Fusion for Subgroup Analysis

Jingyu Shao[a], Ruipeng Dong[a], Zemin Zheng[a]

[a]*International Institute of Finance, The School of Management, University of Science and Technology of China, Hefei, Anhui, 230026, China*

Abstract

Linear regression with subject-specific intercepts is an important tool for recovering the group structure in heterogeneous data analysis. Pairwise fusion on the intercepts with regularization methods has good performance in subgroup analysis when the sample size is moderate. Despite the recent theoretical as well as algorithmic advances, subgroup recovering and latent factor estimate is still a challenging task with the exponential growth of sample volume. Motivated by these concern, we suggest a screened pairwise fusion (SPF) approach based on sample information, to reduce the pairs of intercepts from large to a moderate scale. Then the pairwise fusion procedure can be accomplished by regularization methods, such as convex and concave penalty. Our proposed method not only has high performance in computing but also has ability to accurately identify the group structure as well as estimate the coefficients. With mild conditions, we show SPF enjoys consistency in estimation and the MSE can then be bounded accordingly. Simulation studies show that SPF-based approach outperforms the state-of-art methods. Meanwhile, the analysis of two real data sets contain COVID-19 clinical data and India liver patient data futhermore show the advantage of our approach in subgroup recovering capability.

Keywords: Pairwise fusion, Subgroup recovering, Sample information, Linear regression, Heterogeneous data

1. Introduction

In the field of heterogeneous data analysis, ranging from personalized medicine to economics and finance, it's of great importance to identify the groups, i.e.

the clusters of the population. For example, A popular model-based method for subgroup indentification is to view the data as a mixture of different subgroups with different sets of parameters and then apply the finite mixture model analysis through EM algorithm, see Everitt (2014); McLachlan et al., (2019). The finite mixture model based method always need the distrubution of data as well as the underlying number of components in the population, which is difficult to do in a variety of situations in real world application, such as the gaussian mixture model as well as the logistic-normal mixture model (Shen and He, 2015). Considering the limitation of mixture model based supervised learning methods, Ma and Huang (2017) proposed a new approach of concave pairwise fusion in subgroup analysis. The subject-specific intercepts was introduced to the linear regression, which represents the heterogeneity driven by unknown or unobserved latent factors. The pairwise fusion method is feasible to indentify the subgroups automatically without the knowledge of a priori classification. However, the pairwise fusion with penalty do the pairwise comparing in computing, which is of complexity $O(n^2)$. Therefore, the scalability issues on n make the pairwise fusion problem even harder to compute.

The groups indentifying of pairwise fusion combining with regularization methods are chosen to recover the subgroups. There are various of methods in the field of variable selection that yield sparse estimate of regression coefficients, such as lasso (Tibshirani, 1996), the group lasso (Yuan and Lin, 2006; Simon et al., 2013) and scaled lasso (Sun and Zhang, 2012). Non-convex regularization methods is also beneficial in high-demensional variable selection, see the smoothly clipped absolute deviation penalty (Fan and Li, 2001), the minimax concave penalty (Zhang et al., 2010) as well as the smooth integration of counting and absolute deviation penalty (Lv et al., 2009), respectively. A fast ADMM algorithm (Chi and Lange, 2015) can be apply to the pairwise fusion methods for estimation of parameters, which enjoys good convergence properties for convex loss functions with the L_p penalties, $p \geq 1$ (Chi and Lange, 2015).

In this article, we propose a framework of screened pairwise fusion (SPF) to significantly increase the scalability of computation in pairwise learning via

sample pairs screening. Since the information of the population is of great importance in supervised learning, we recognize that using all data can reduce information loss, but this still be a challenging computational problem. To improve this, we consider sample pairs screening under the premise of ensuring the quality of sample information. There are several works focus on taking random subsamples from the full data, see Ma and Sun (2015); Suchard et al. (2010); Ma et al. (2014) for linear regression model and Wang et al. (2018) for logistic regression. Wang et al. (2019) proposed an information-based optimal subdata selection framework in the case $n \gg p$, which enjoy the computing time of $O(np)$ rather than $O(np^2)$ in algorithmic leveraging (Ma et al., 2014). In the case $n \gg p$, pairwise fusion on the whole population will be challenging. In our framework, we screen the sample first to derive a subsample of size n_S, and $n_S = o(n)$ is allowed. And then we apply the pairwise fusion on the new population of size n_S. The pairwise fusion procedure has a complexity of $O(n^2)$ while the screened pairwise fusion enjoy time of $O(np + n_S^2)$. Our simulation show that the subgroup identifing ability is still strong after the initial screening of the sample since individual information is retained.

Our *sample pairs screening + pairwise fusion* procedure of SPF method is strongly motivatied by reducing redundant computation. Specifically, the pairwise fusion within one subgroup will not help us to find new subgroup but a lot of redundant compairing. Meanwhile, in large-scale clinical diagnosis, it is of great significance to speed up the quarantine procedure as well as make an arrangement of medical facilities. The pandemic COVID-19, which is caused by SARS-CoV-2 (Kim et al., 2020), has raging around the world. With the exponential increase in the number of cases, more and more overwhelm health systems around the world demand for intensive care unit (ICU) beds far above the existing capacity. It is a challenge for an overwhelm health system to perform the detection of SARS-CoV-2 by testing every case. However, the tests results could be delayed even if only a target subgroup would be test. Therefore, it is of great significance to have a clear understanding of the subgroups in the population with less costs. The SPF method ensures the high efficiency of

pairwise fusion as well as strong ability of subgroup identification.

The main contribution of our paper are fourfold. First, for the pairwise fushion problem, our insights upon information based screen approach are capable to recover the group structures and estimate parameters as well as the latent factor simultaneously. Second, our proposed SPF method is of great scalability in computing and SPF+ enable the pairwise fusion for whole population more efficiently by large scale data distributed computing approaches. Third, without additional assumptions, we bound the error as well as the mean square error for SPF estimate of both coefficients and latent factors. Fourth, our experiments and empirical analysis of COVID-19 clinical data and India liver patient data futher emphasis the advantages of SPF in subgroup analysis.

2. Model specification

Notations: Throughout this paper, we use bold letter to denote matrix and column vector. For any matrix $\boldsymbol{Z} = (z_{ij})$, denote by $||\boldsymbol{Z}||_F = \sqrt{\sum_{i,j} z_{ij}^2}$, $||\boldsymbol{Z}||_0 = \sum_{i,j} I(z_{ij} \neq 0)$ and $||\boldsymbol{Z}||_\infty = max_{ij}|z_{ij}|$ the Frobenius norm, entrywise L_0-norm and entrywise L_∞-norm. Denote the L_2 and L_0 norm of any vector $\boldsymbol{z}$ by $||\boldsymbol{z}||_2 = \sqrt{\sum_i z_i^2}$ and $|\boldsymbol{z}| = \sum_i I(z_i \neq 0)$. In this paper, $\boldsymbol{z}_j$ is the jth column vector of $\boldsymbol{Z}$, and z_{ij} is the ith element of vector $\boldsymbol{z}_j$ without special instructions.

Let y_i be the response for the ith sample. For a set of covariates $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$, we consider the heterogeneity of $\boldsymbol{y} = (y_1, ..., y_n)$ through subject-specific intercepts in regression

$$y_i = \mu_i + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \ i = 1, ..., n. \tag{1}$$

where μ_i are the unknown subject-specific intercepts, also can be viewed as unknown and unobserved latent factors, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is the vector of unknown coefficients, so the size of parameters is $n + p$. And the error term independent of $\boldsymbol{x}_i$ has $\mathbb{E}(\epsilon_i) = 0$ and $var(\epsilon_i) = \sigma^2$. It is worth noting that if the factors contain the information of heterogeneity, for example, different ICU and regular wards arrangements are available, then we can write $\mu_i = \mu + \boldsymbol{z}_i^T \boldsymbol{\theta}$,

where $\boldsymbol{z}_i$ are unobserved covriates for arrangement that could be correlated to the observed ones. and $\boldsymbol{\theta}$ are the coefficient of $\boldsymbol{z}_i$. From (1) we have that $y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} = \mu_i + \epsilon_i,\ i = 1, ..., n$. Let $\hat{\boldsymbol{\beta}}_{ols}$ be the least square estimate such that $\hat{\boldsymbol{\beta}}_{ols} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, where $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T$. Then we let

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ols} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y} \tag{2}$$

be the pseudo resonse which contains the information of heterogeneity in the subject-specific intercepts $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ instead of predictors $\boldsymbol{X}$. In the homogeneous model, where we have $\mu_1 = ... = \mu_n = \mu$, and we estimate the regular regression model

$$y_i = \mu + \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i,\ i = 1, ..., n. \tag{3}$$

The sample size is growing exponentially in various of application scenarios, which greatly increas the complexity of computing, especially when the algorithm lacks of scalability. In many cases, the sample has an unbalanced group structure (Sun et al., 2009), which inspired us to screen the sample to obtain a main subsample with a heterogeneous group structure, that is, the screened sample contains the individuals we are interested in. Inspired by the srue independence screening approach (Fan and Lv, 2008) in feature space, we hopefully expect to screen out a large number of individuals with the same as well as the analogous group labels, which contain lots of redundant computation in the pairwise fusion. In the screening stage of the sample, we use the homogeneous model (3) in order to facilitate the expression of information. We write $\boldsymbol{z}_i = (1, \boldsymbol{x}_i^T)^T$, $\boldsymbol{\beta}_f = (\mu, \boldsymbol{\beta}^T)^T$. Using the full data (i.e. information) and homogeneous model, the least square estimator of $\boldsymbol{\beta}_f$ has form $\hat{\boldsymbol{\beta}}_f = \left(\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T\right)^{-1}\sum_{i=1}^n \boldsymbol{z}_i y_i$, which is also the best linear unbiased estimator (BLUE). And when the ϵ_i's are i.i.d. normally distributed, then Fisher information matrix for $\boldsymbol{\beta}$ also known as the inverse of the covariance matrix of the unbiased estimation above is $\boldsymbol{I}_f = \frac{1}{\sigma^2}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^T$. Later we will show that the normality assumption of ϵ_i is not required.

To construct the Fisher information matrix for coefficients based on the

screened sample, we first let $S \subseteq \{1, ..., n\}$ be the index set of subsample. Without loss of generality, after reindexing, the screened sample $(\boldsymbol{x}_1^*, y_1^*), ..., (\boldsymbol{x}_{n_S}^*, y_{n_S}^*)$ of size $n_S = |S|$ has the information matrix for $\boldsymbol{\beta}$ as $\boldsymbol{I}_S = \frac{1}{\sigma^2} \sum_{i=1}^{n_S} \boldsymbol{z}_i^* \boldsymbol{z}_i^{*T}$. where $\boldsymbol{z}_i^* = (1, \boldsymbol{x}_i^{*T})^T$. Now we let $\delta_i = I\{i \in S\}$ be the indicator variable that signifies whether $(\boldsymbol{x}_i, y_i)$ is included in the screened sample. Then the information matrix with subsample of size n_S can be rewritten as

$$\boldsymbol{I}(\boldsymbol{\delta}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \delta_i \boldsymbol{z}_i \boldsymbol{z}_i^T, \tag{4}$$

where $\boldsymbol{\delta} = \{\delta_1, ..., \delta_n\}$ such that $\sum_{i=1}^{n} \delta_i = n_S$. Then for an optimality criterion function ψ, the main goal of our data screening procedure is to estimate the optimization problem

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmax}}\, \psi\{\boldsymbol{I}(\boldsymbol{\delta})\} \qquad \text{s.t.} \sum_{i=1}^{n} \delta_i = n_S. \tag{5}$$

Now we can construct a mathematical framework for our screened pairwise fusion method on heterogeneous model (1). A popular optimality criterion is the D-optimality criterion (Wang et al., 2019), which maximizes the determinant of $\boldsymbol{I}(\boldsymbol{\delta})$. Let $\det(\cdot)$ be the determinant of the matrix, then we use a modified D-optimality criterion $\psi\{\boldsymbol{I}(\boldsymbol{\delta})\} = \det(\boldsymbol{I}(\boldsymbol{\delta})) + \boldsymbol{\delta}^T \tilde{\boldsymbol{y}}$. We now give the objective function of our model on the whole population:

$$\begin{aligned} Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\delta}; \lambda) &= \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \sum_{1 \le i < j \le n} \delta_{ij} p(|\mu_i - \mu_j|, \lambda) \\ \text{s.t. } \boldsymbol{\delta} &= \underset{|\boldsymbol{\delta}_1| = n_S}{\operatorname{argmax}} \{\det(\boldsymbol{I}(\boldsymbol{\delta}_1)) + \boldsymbol{\delta}_1^T \tilde{\boldsymbol{y}}\}, \end{aligned} \tag{6}$$

where $\delta_{ij} = \delta_i \delta_j$, $\boldsymbol{I}(\boldsymbol{\delta})$ was defined in (4) and $p(., \lambda)$ is a penalty function with tuning parameter $\lambda \ge 0$. Then under the optimality criterion function ψ, the main goal of our screen method is to estimate $\boldsymbol{\delta}$, after which the procedure of screening on the full sample will be completed. The magnitude of n_S has a certain impact on the MSE of our estimator. We will give a mathematical description of the relationship between the screened sample size n_S and the MSE of our method in theorem 3.

It's important that which penalty function should be chosen in problem (6). Since the L_1 penalty uses the same threshold for all pairs, it will produce biased estimates and may not be able to restore the subgroup structure. This situation is similar to the fact that Lasso tends to over-compress larger coefficients in variable selection. In numerical research, it is found that the L_1 penalty tends to filter out a large number of subgroups or no subgroups on the solution path. Therefore, a penalty term that can produce unbiased estimates is even more needed. Ma and Huang (2017) use concave penalty methods SCAD (Fan and Li, 2001) and MCP (Zhang et al., 2010). These methods are unbiased and are easier to produce sparse solutions, so the number of subgroups is usually much smaller than the sample size. The derivative of MCP penalty has form:

$$p_{\gamma}^{'}(t,\lambda) = \frac{(\gamma\lambda - t)_+}{\gamma}, \gamma > 1$$

The derivative of SCAD penalty has form:

$$p_{\gamma}^{'}(t,\lambda) = \lambda I(t \leq \lambda) + \frac{(\gamma\lambda - t)_+}{\gamma - 1} I(t > \lambda), \gamma > 2$$

where γ is a parameter that controls the concavity of the penalty function and $(x)_+ = \max\{x, 0\}$ for any x. In particular, when gamma tends to infinity, both penalties tend to L_1 penalties. In the following, we put γ in the subscript position to represent the relationship with L_1 penalty. According to Fan and Li (2001) and Zhang et al. (2010), we treat gamma as a fixed constant. These concave penalties are related to L1 penalty. Similar to L_1 penalty, concave penalties enjoy sparse solution, which means they can automatically generate zero estimates. More importantly, these concave penalty have unbiased properties and do not over-compress larger estimated parameters, so they are still unbiased in the iteration. This property is particularly important in the ADMM algorithm, because the deviation in the iteration will significantly affect the identification and determination of the subgroups.

For a given $\lambda > 0$, we estimate the optimization problem on the subsample as follows

$$(\hat{\boldsymbol{\mu}}(\lambda), \hat{\boldsymbol{\beta}}(\lambda), \hat{\boldsymbol{\delta}}(\lambda)) = \underset{\boldsymbol{\mu},\boldsymbol{\beta},\boldsymbol{\delta}}{\operatorname{argmin}} Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\delta}; \lambda).$$

it is known that this pairewise fusion method has a high computational complexity of $O(n^2)$ when the full sample participate directly in this kind of pairwise matching algorithm. However, for a subgroup that cover a large number of individuals, it is meaningless to pair and merge in this subgroup. Fortunately, we can use information based sample screen method to identify a subsample with heterogeneity, so as to perform subgroup analysis as well as outliar analysis. The pairwise fusion procedure process on $\delta_{ij}|\mu_i - \mu_j|$ instead of $|\mu_i - \mu_j|$ for $i, j \in [n]$, which significantly reduces the computational complexity of pairwise fusion.

By screening the sample, we can select the important subgroups and reduce significantly the sample pairs size n^2 to a much lower magnitude n_S^2, after which the pairwise fusion procedure become much more intuitive and interpretable, see figure 1 for an example.

3. Algorithm

It's difficult to compute the estimates directly by minimizing (6) because of the unseparable in $\boldsymbol{\mu}$, and the choice of $\boldsymbol{\delta}$ will also be a NP hard problem. Next, we will decompose problem (6) into two steps to solve. In the first step, we screen the sample via OBS and derive $\hat{\boldsymbol{\delta}}$ and a subsample which contain the most information among all subsamples of size n_S. Then our $\hat{\boldsymbol{\delta}}$ is an approximation solution for problem (5). In the second step, we apply the ADMM method to solve the pairwise fusion problem based on the estimate $\hat{\boldsymbol{\delta}}$.

3.1. Order based approximation for information

Obtaining an exact solution for (5) is computionally far too expensive. We choose $\psi(\boldsymbol{I}(\boldsymbol{\delta})) = \det(\boldsymbol{I}(\boldsymbol{\delta}))$. In order to have a nice approximate solution, we first introduce theorem 1 which is derived from theorem 1 in Wang et al. (2019) and the fact $\boldsymbol{\delta}^T\tilde{\boldsymbol{y}} \leq n_S||\tilde{\boldsymbol{y}}||_\infty$.

Theorem 1. *For subsample of size n_S represented by $\boldsymbol{\delta}$,*

$$\det(\boldsymbol{I}(\boldsymbol{\delta})) + \boldsymbol{\delta}^T\tilde{\boldsymbol{y}} \leq \frac{n_S^{p+1}}{4^p\sigma^{2(p+1)}}\prod_{j=1}^{p}(x_{(n)j} - x_{(1)j})^2 + n_S(|\tilde{\boldsymbol{y}}_{(1)}| + |\tilde{\boldsymbol{y}}_{(n)}|), \quad (7)$$

where $x_{(n)j} = max\{x_{1j}, ..., x_{nj}\}$ and $x_{(1)j} = min\{x_{1j}, ..., x_{nj}\}$ are the nth and first-order statistics of $x_{1j}, ..., x_{nj}$. If the subsample consists of the 2^p points $(a_1, ..., a_p)^T$ where $a_j = x_{(n)j}$ or $x_{(1)j}$, $j = 1, ..., p$, each occurring equally often, then equality holds in (7).

Motivated by the result in theorem 1, we give the order based screen (OBS) algorithm motivated by this approximation method. Suppose that $2pr \leq n_S \leq 4pr$ is an integer. Using a partition-based selection procedure, perform the following steps:

***step* 1** Estimate the pesudo response by (2). For $\tilde{\boldsymbol{y}} = (\tilde{y}_1, ..., \tilde{y}_n)^T$, include $2pr$ indexes with the pr largest values and pr smallest values of $\tilde{\boldsymbol{y}}$. Let $\mathcal{A}$ be the initial index set.

***step* 2** For x_{i1}, $1 \leq i \leq n$, include $2r$ sample points with the r largest values and r smallest values of x_{i1}. Add the selected indexes into $\mathcal{A}$.

***step* 3** For $j = 2, ..., p$, exclude sample points that were previously selected, and from the remainder select $2r$ sample points with the r largest values and r smallest values of x_{ij}, after which add them into $\mathcal{A}$.

***step* 4** Return $\hat{\boldsymbol{\delta}}$ according to $\mathcal{A}$ and the selected subsample $\boldsymbol{X}_S, \boldsymbol{y}_S$ with size $n_S = |\mathcal{A}|$, and let the left samples be the new $\boldsymbol{X}, \boldsymbol{y}$.

***step* 5** (Optional) Repeat step 2-4, derive a sequence of subsamples $(\boldsymbol{X}_{S_1}, \boldsymbol{y}_{S_1})$,..., $(\boldsymbol{X}_{S_m}, \boldsymbol{y}_{S_m})$, where $m = \lfloor \frac{n}{n_S} \rfloor$.

3.2. Screened Pairwise Fusion via ADMM

We reparameterize the pairwise difference by introducig $\eta_{ij} = \mu_i - \mu_j$, then (6) is equivalent to the constraint version:

$$S(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}) = \frac{1}{2}\sum_{i=1}^{n}(1-\hat{\delta}_i)(y_i - \mu_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \frac{1}{2}\sum_{i=1}^{n}\hat{\delta}_i(y_i - \mu_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \sum_{1\leq i<j\leq n}\hat{\delta}_{ij}p(|\eta_{ij}|, \lambda)$$

$$s.t.\ \eta_{ij} = \mu_i - \mu_j$$

By the augmented Lagrangian method, the parameters can be estimated by minimizing:

$$L(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\upsilon}) = S(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}) + \sum_{i<j} \hat{\delta}_{ij} \upsilon_{ij} (\mu_i - \mu_j - \eta_{ij}) + \frac{\vartheta}{2} \sum_{i<j} \hat{\delta}_{ij} (\mu_i - \mu_j - \eta_{ij})^2$$

Fortunately, although the objective function is nonconvex, it is convex if $\gamma > 1/\vartheta$ for MCP penalty and $\gamma > 1/\vartheta + 1$ for SCAD penalty. Moreover, for given $(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\upsilon})$, we can turn the prime problem into minimizing

$$\frac{\vartheta}{2} (\zeta_{ij}^{(m+1)} - \eta_{ij})^2 + p_\gamma(|\eta_{ij}|, \lambda)$$

with respect to η_{ij}, where $\zeta_{ij} = \mu_i - \mu_j + \upsilon_{ij}/\vartheta$. Let $\mathrm{ST}(t, \lambda) = \mathrm{sign}(t)(|t| - \lambda)_+$ be the soft thresholding rule. Then for MCP penalty with $\gamma > 1/\vartheta$, it has the closed form solution

$$\hat{\eta}_{ij} = \begin{cases} \frac{\mathrm{ST}(\zeta_{ij}, \lambda/\vartheta)}{1 - 1/(\gamma\vartheta)}, & |\zeta_{ij}| \le \gamma\lambda \\ \zeta_{ij}, & |\zeta_{ij}| > \gamma\lambda \end{cases}, \tag{8}$$

and for the SCAD penalty with $\gamma > 1/\vartheta + 1$, it has

$$\hat{\eta}_{ij} = \begin{cases} \mathrm{ST}(\zeta_{ij}, \lambda/\vartheta), & |\zeta_{ij}| \le \lambda + \lambda/\vartheta \\ \frac{\mathrm{ST}(\zeta_{ij}, \gamma\lambda/((\gamma-1)\vartheta))}{1 - 1/((\gamma-1)\vartheta)}, & \lambda + \lambda/\vartheta < |\zeta_{ij}| \le \gamma\lambda \\ \zeta_{ij}, & |\zeta_{ij}| > \gamma\lambda \end{cases}. \tag{9}$$

For L_1 penalty, it is $\hat{\eta}_{ij} = \mathrm{ST}(\zeta_{ij}, \lambda/\vartheta)$.

We compute the estimates of $(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\upsilon})$ through iteractions by the ADMM. First of all, for a given $(\boldsymbol{\eta}, \boldsymbol{\upsilon})$, to obtain an update of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$, we need to minimize the following object function:

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\upsilon}) &= \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \frac{\vartheta}{2} \sum_{i<j} \hat{\delta}_{ij} \{(\boldsymbol{e}_i - \boldsymbol{e}_j)^T \boldsymbol{\mu} - \eta_{ij} + \vartheta^{-1} \upsilon_{ij}\}^2 + C \\ &= \frac{1}{2} ||\boldsymbol{\mu}_S - \boldsymbol{y}_S + \boldsymbol{X}_S \boldsymbol{\beta}||_2^2 + \frac{1}{2} ||\boldsymbol{\mu}_{S^c} - \boldsymbol{y}_{S^c} + \boldsymbol{X}_{S^c} \boldsymbol{\beta}||_2^2 \\ &+ \frac{\vartheta}{2} ||\boldsymbol{\Delta} \boldsymbol{\mu}_S - \boldsymbol{\eta} + \vartheta^{-1} \boldsymbol{\upsilon}||_2^2 + C. \end{aligned}$$

Here C is a constant, e_i is the $n \times 1$ column vector whose ith element is 1 while the remaining ones are 0, and $\boldsymbol{\Delta} = \{(\boldsymbol{e}_i - \boldsymbol{e}_j), i < j, \hat{\delta}_{ij} = 1\} = (\boldsymbol{e}_{i_1} - \boldsymbol{e}_{j_2}, ..., \boldsymbol{e}_{i_{n_S - 1}} -$

$\boldsymbol{e}_{j_{n_S}})^T$. Obviously, $||\boldsymbol{\Delta}||_0 < n_S(n_S-1)/2$. Denote the sreened sample $\boldsymbol{X}_S$ has size n_S, then $\boldsymbol{I}_S$ is the $n_S \times n_S$ identity matrix, $\boldsymbol{Q}_S = \boldsymbol{X}_S(\boldsymbol{X}_S^T\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^T$, $\boldsymbol{Q}_{S^c} = \boldsymbol{X}_{S^c}(\boldsymbol{X}_{S^c}^T\boldsymbol{X}_{S^c})^{-1}\boldsymbol{X}_{S^c}^T$. Without loss of generality, we write $\boldsymbol{\mu} = (\boldsymbol{\mu}_S^T, \boldsymbol{\mu}_{S^c}^T)^T$. Now we set the derivatives $\partial L/\partial \boldsymbol{\mu}_S$, $\partial L/\partial \boldsymbol{\mu}_{S^c}$ and $\partial L/\partial \boldsymbol{\beta}$ to zero, then we derive that, for given $\boldsymbol{\eta}^{(m)}$ and $\boldsymbol{\upsilon}^{(m)}$ at the mth step:

$$
\begin{aligned}
\boldsymbol{\mu}_{S^c}^{(m+1)} &= (\boldsymbol{I}_{S^c} - \boldsymbol{Q}_{S^c})\boldsymbol{y}_{S^c}, \\
\boldsymbol{\mu}_S^{(m+1)} &= (\vartheta\boldsymbol{\Delta}^T\boldsymbol{\Delta} + \boldsymbol{I}_S - \boldsymbol{Q}_S)^{-1}\{(\boldsymbol{I}_S - \boldsymbol{Q}_S)\boldsymbol{y}_S + \vartheta\boldsymbol{\Delta}^T(\boldsymbol{\eta}^{(m)} - \vartheta^{-1}\boldsymbol{\upsilon}^{(m)})\}, \\
\boldsymbol{\mu}^{(m+1)} &= ((\boldsymbol{\mu}_S^{(m+1)})^T, (\boldsymbol{\mu}_{S^c}^{(m+1)})^T)^T.
\end{aligned}
$$

The implementation of SPF is summarized in Algorithm 1.

According to the theorem 1, under certain assumptions, the preliminary screening of samples can effectively cover outliar samples, and sample compression can be performed on the basis of retaining the heterogeneous structure of the samples. Moreover, in some special cases, it is important to implement the subgroup analysis on the whole population. In these situations, we use Algorithm SPF+ that add step 4 of OBS and then perform ADMM iteration on subsamples in a parallel computing manner. In most cases, SPF is sufficient to recover the group structure.

Remark 1. *Since the outliers may be included in our subsample, the outlier diagnostic methods can be used on the subsample. It is worth mentioning that, our method tends to select individuals contain more information about the model, which should be used for subgroup analysis as well as parameter estimation.*

Remark 2. *The SPF+ method screen the samples sequentially until all the population are split into $m = n/n_S$ subsamples. We are inspired to give priority for the first screened batches of subsamples for the pairwise fusion and subgroup analysis. For the full population, SPF+ have a computing complexity of $O(nn_S)$.*

It's worth to mention that the subsample $(\boldsymbol{X}_{S_1}, \boldsymbol{y}_{S_1}), (\boldsymbol{X}_{S_2}, \boldsymbol{y}_{S_2}), ..., (\boldsymbol{X}_{S_m}, \boldsymbol{y}_{S_m})$ have an unbalanced allocation of information since the large or small x_{ij} which contain more information tend to be screened out early according to our OBS procedure.

Algorithm 1 Screened Pairwise Fusion (SPF)

Input: The full sample $(\boldsymbol{X}, \boldsymbol{y})$ and $\boldsymbol{X}$ is of size $n \times p$; tuning parameters λ and r; the tolerance τ.

$\tilde{\boldsymbol{y}} \leftarrow ((\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y})$

$\mathcal{A}_0 \leftarrow \{i : \text{the first } pr \text{ largest } \tilde{\boldsymbol{y}}_i \text{ and the first } pr \text{ smallest } \tilde{\boldsymbol{y}}_i\}, \mathcal{A}_1 = \mathcal{A}_2 \leftarrow \emptyset$

for $j = 1, ..., p$ **do**

  $\mathcal{A}_1 \leftarrow \mathcal{A}_1 \cup \{i : \text{the first } r \text{ largest } x_{ij} \text{ out of } \mathcal{A}_1 \cup \mathcal{A}_2\}$

  $\mathcal{A}_2 \leftarrow \mathcal{A}_2 \cup \{i : \text{the first } r \text{ smallest } x_{ij} \text{ out of } \mathcal{A}_1 \cup \mathcal{A}_2\}$

end for

$\boldsymbol{X}_S \leftarrow \boldsymbol{X}_{\mathcal{A}_0 \cup \mathcal{A}_1 \cup \mathcal{A}_2}, \boldsymbol{y}_S \leftarrow \boldsymbol{y}_{\mathcal{A}_0 \cup \mathcal{A}_1 \cup \mathcal{A}_2}, n_S \leftarrow |\mathcal{A}_0 \cup \mathcal{A}_1 \cup \mathcal{A}_2|$

$\boldsymbol{\beta}^{(0)} \leftarrow (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, $\boldsymbol{\mu}^{(0)} \leftarrow \tilde{\boldsymbol{y}}, \boldsymbol{\upsilon}^{(0)} \leftarrow \boldsymbol{0}, \boldsymbol{\eta}^{(0)} \leftarrow \boldsymbol{0}, \boldsymbol{\zeta}^{(0)} \leftarrow \boldsymbol{0}, m \leftarrow 0$

repeat

  $\boldsymbol{\mu}_S^{(m+1)} \leftarrow (\vartheta \boldsymbol{\Delta}^T\boldsymbol{\Delta} + \boldsymbol{I}_S - \boldsymbol{Q}_S)^{-1}\{(\boldsymbol{I}_S - \boldsymbol{Q}_S)\boldsymbol{y}_S + \vartheta\boldsymbol{\Delta}^T(\boldsymbol{\eta}^{(m)} - \vartheta^{-1}\boldsymbol{\upsilon}^{(m)})\}$

  $\boldsymbol{\mu}^{(m+1)} \leftarrow ((\boldsymbol{\mu}_S^{(m+1)})^T, ((\boldsymbol{I}_{S^c} - \boldsymbol{Q}_{S^c})\boldsymbol{y}_{S^c})^T)^T$

  $\boldsymbol{\beta}^{(m+1)} \leftarrow (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}^{(m+1)})$

  for $i = 1, ..., n_S$ **do**

    for $j = 1, ..., n_S$ **do**

      $\zeta_{ij}^{(m+1)} \leftarrow \mu_i^{(m+1)} - \mu_j^{(m+1)} + \vartheta^{-1}\upsilon_{ij}^{(m)}$

      $\eta_{ij}^{(m+1)} \leftarrow \underset{\eta_{ij}}{\text{argmin}} \left\{ \frac{\vartheta}{2}(\zeta_{ij}^{(m+1)} - \eta_{ij})^2 + p_\gamma(|\eta_{ij}|, \lambda) \right\}$

      $\upsilon_{ij}^{(m+1)} \leftarrow \upsilon_{ij}^{(m)} + \vartheta(\mu_i^{(m+1)} - \mu_j^{(m+1)} - \eta_{ij}^{(m+1)})$

    end for

  end for

  $m \leftarrow m + 1$

until $||\boldsymbol{\Delta}\boldsymbol{\mu}^{(m)} - \boldsymbol{\eta}^{(m)}||_2 < \tau$

Output: $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\upsilon}}) = (\boldsymbol{\mu}^{(m)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\eta}^{(m)}, \boldsymbol{\upsilon}^{(m)})$

4. Theoretical Properties

Suppose that the full sample $\boldsymbol{X}$ has groups $\mathcal{G}_1, ..., \mathcal{G}_K \subset \{1, ..., n\}$. Let $\mathcal{M}_{\mathcal{S}}$ be the subspace of $\mathbb{R}^{n_S}$, which is defined as

$$\mathcal{M}_{\mathcal{S}} = \{\boldsymbol{\mu} \in \mathbb{R}^{n_S} : \mu_i = \mu_j, \text{for any } i, j \in \mathcal{G}_k, 1 \leq k \leq K\} \tag{10}$$

We introduce a $n \times K$ matrix $\boldsymbol{Z} = \{z_{ik}\}$, which satisfies $z_{ik} = 1$ for $i \in \mathcal{G}_k$ and $z_{ik} = 0$ otherwise. Then we have $\boldsymbol{Z}^T\boldsymbol{Z} = \text{diag}(|\mathcal{G}_1|, ..., |\mathcal{G}_K|)$. Denote $\boldsymbol{\alpha}$ for a $K \times 1$ vector of parameters, then for any $\boldsymbol{\mu} \in \mathcal{M}_{\mathcal{G}}$, it has form $\boldsymbol{\mu} = \boldsymbol{Z}\boldsymbol{\alpha}$. Denote by $\boldsymbol{X}_S$ the screened sample matrix with n_S corresponding rows of copies in $\boldsymbol{X}$. And we define $\boldsymbol{Z}_S$ based on $\boldsymbol{Z}$ accordingly. The estimate based on the screened sample is $\hat{\boldsymbol{\mu}}_S$ which is of size n_S. Let $\boldsymbol{\Sigma}$ be the sample correlation matrix of subdata $\boldsymbol{X}_S$, and $\lambda_{min}(\boldsymbol{\Sigma}), \lambda_{max}(\boldsymbol{\Sigma})$ is the smallest and largest eigenvalue of $\boldsymbol{\Sigma}$.

We let $\boldsymbol{\alpha}^* = (\alpha_1^*, ..., \alpha_K^*)^T$ be the true common intercept for group $\mathcal{G}_1, ..., \mathcal{G}_K$ and $\boldsymbol{\beta}^* = (\beta_1^*, ..., \beta_p^*)^T$ be the true coefficient for regerssion. When the underlying group memberships $\mathcal{G}_1, ..., \mathcal{G}_K$ are known, then this leads to the oracle estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$

$$\begin{pmatrix} \hat{\boldsymbol{\mu}}_S^{or} \\ \hat{\boldsymbol{\beta}}_S^{or} \end{pmatrix} = \underset{\boldsymbol{\mu}_S \in \mathcal{M}_{\mathcal{S}}, \boldsymbol{\beta}_S \in \mathbb{R}^p}{\text{argmin}} ||\boldsymbol{y}_S - \boldsymbol{\mu}_S - \boldsymbol{X}_S\boldsymbol{\beta}_S||^2. \tag{11}$$

In equality, we have the oracle estimators for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as

$$\begin{aligned} \begin{pmatrix} \hat{\boldsymbol{\alpha}}_S^{or} \\ \hat{\boldsymbol{\beta}}_S^{or} \end{pmatrix} &= \underset{\boldsymbol{\alpha}_S \in \mathbb{R}^K, \boldsymbol{\beta}_S \in \mathbb{R}^p}{\text{argmin}} ||\boldsymbol{y}_S - \boldsymbol{Z}_S\boldsymbol{\alpha}_S - \boldsymbol{X}_S\boldsymbol{\beta}_S||^2 \\ &= [(\boldsymbol{Z}_S, \boldsymbol{X}_S)^T(\boldsymbol{Z}_S, \boldsymbol{X}_S)]^{-1}(\boldsymbol{Z}_S, \boldsymbol{X}_S)^T\boldsymbol{y}_S. \end{aligned}$$

Next we first bound the variance of estimation in our SPF method by a proposition.

Proposition 1. *If $\lambda_{min}(\boldsymbol{\Sigma}) > 0$, let $r \leq \frac{n_S}{2p}$, then the following variance bounds*

hold for estimators $\hat{\boldsymbol{\mu}}_S^{or}$ *and* $\hat{\boldsymbol{\beta}}_S^{or}$*, obtained from SPF:*

$$\begin{aligned} \frac{\sigma^2}{n_S} \leq V(\hat{\mu}_{iS}^{or}|\boldsymbol{X}) &\leq \frac{\sigma^2}{n_S - 1}\left(1 + \frac{c_0}{\lambda_{\min}(\boldsymbol{\Sigma})}\right),\ i \in S \\ \frac{4\sigma^2}{n_S \lambda_{\max}(\boldsymbol{\Sigma})(x_{(n)j} - x_{(1)j})^2} &\leq V(\hat{\beta}_{jS}^{or}|\boldsymbol{X}) \\ &\leq \frac{4p\sigma^2}{n_S \lambda_{\min}(\boldsymbol{\Sigma})(x_{(n-r+1)j} - x_{(r)j})^2},\ j = 1, ..., p. \end{aligned} \tag{12}$$

The variance bound is helpful since it's necessary to bound the MSE of our method. For $K \geq 2$, we let

$$b_n = \min_{i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}, k \neq k'} |\mu_i^* - \mu_j^*| = \min_{k \neq k'} |\alpha_k^* - \alpha_{k'}^*| \tag{13}$$

be the minimal gap between two groups. Before present our main results, we give 3 assumptions below.

Assumption 1. $\lambda_{\min}[(\boldsymbol{Z}, \boldsymbol{X})^T(\boldsymbol{Z}, \boldsymbol{X})] \geq C_0|\mathcal{G}_{\min}|$ *holds for some constant* $0 < C_0 < 1$ *and* $||\boldsymbol{X}||_\infty \leq C_1 p$ *holds for some constant* $0 < C_1 < \infty$*. The sample is scaled as* $||\boldsymbol{X}_j||_2 = \sqrt{n}$ *for* $1 \leq j \leq p$*.*

Assumption 2. *The noise vector* $\boldsymbol{\epsilon}$ *has sub-Gaussian tails such that* $\mathbb{P}(|\boldsymbol{a}^T\boldsymbol{\epsilon}| > ||\boldsymbol{a}||_2 t) \leq 2\exp(-c_1 t^2)$ *for any vector* $\boldsymbol{a} \in \mathbb{R}^n$ *and* $t > 0$*, where* $0 < c_1 < \infty$*.*

Assumption 3. $p_\gamma(t, \lambda)$ *is nondecreasing, concave and symmetric with repect to* $t \in [0, \infty)$*. Let* $\rho(t) = \lambda^{-1}p_\gamma(t, \lambda)$*, then* $\rho'(t)$ *is almost surely continuous and* $\rho(0) = 0$*,* $\rho'(0+) = 1$*.* $\rho(t)$ *is a constant for all* $t > a\lambda$*, where* $0 < a < \infty$ *is a constant.*

The signal condition in assumption 1 is met in many situations. Our assumptions 2 and 3 are common with high-dimensinal settings. Moreover, MCP as well as SCAD penalties both enjoy the properties in assumption 3.

Theorem 2. *Under assumption 1 and 2, if* $|\mathcal{G}_{\min}| \gg \sqrt{(K+p)n\log n}$ *holds and the screen sample size satisfy* $n_S \leq \frac{n}{2}$ *and* $p = o(n)$*,* $K = o(n)$*, then there exist* $0 < C_p < \infty$ *only depend on* p *and a constant* $0 < C_3 < \infty$*, with probability at*

least $1-2/n_S-2/n-2(K+p)/n^2$

$$
\begin{aligned}
&||((\hat{\boldsymbol{\alpha}}_S^{or}-\boldsymbol{\alpha}^*)^T,(\hat{\boldsymbol{\beta}}_S^{or}-\boldsymbol{\beta}^*)^T)^T||_\infty \le \phi_n+\psi_n,\\
&||((\hat{\boldsymbol{\mu}}^{or}-\boldsymbol{\mu}^*)^T,(\hat{\boldsymbol{\beta}}^{or}-\boldsymbol{\beta}^*)^T)^T||_\infty \le \phi_n
\end{aligned}
\tag{14}
$$

where $\phi_n=\frac{\sqrt{K+p}}{C_0|\mathcal{G}_{\min}|}\sqrt{\frac{2n\log n}{c_1}}$ *and* $\psi_n=C_3(\frac{1}{n_S}-\frac{1}{n})(\sqrt{\frac{2\log n_S}{c_1}}+C_p)$

Theorem 3. *If the conditions in theorem 2 and assumption 3 hold, with* ϕ_n, ψ_n *as defined in theorem 2 and* $b_n > a\lambda$, $\lambda \gg \phi_n$, $1 \le r < \frac{n}{4p}$ *then there exist a local minimizer* $(\hat{\boldsymbol{\alpha}}_S(\lambda,r)^T,\hat{\boldsymbol{\beta}}_S(\lambda,r)^T)^T$ *satisfies*

$$
P\big((\hat{\boldsymbol{\alpha}}_S(\lambda,r)^T,\hat{\boldsymbol{\beta}}_S(\lambda,r)^T)^T=((\hat{\boldsymbol{\alpha}}_S^{or})^T,(\hat{\boldsymbol{\beta}}_S^{or})^T)^T\big)\to 1,
$$

and for a given $\boldsymbol{X}$, *with probability at least* $1-2/n_S-2/n-2(K+p)/n^2$

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}}(\hat{\mu}_{iS}-\mu_i^*)^2 \le \phi_n^2+\psi_n^2+\frac{\sigma^2}{n_S-1}\left(1+\frac{c_0}{\lambda_{\min}(\boldsymbol{\Sigma})}\right),\\
&\mathbb{E}_{\boldsymbol{X}}(\hat{\beta}_{jS}-\beta_j^*)^2 \le \phi_n^2+\psi_n^2+\frac{4p\sigma^2}{n_S\lambda_{min}(\boldsymbol{\Sigma})(x_{(n-r+1)j}-x_{(r)j})^2}
\end{aligned}
$$

hold for $i \in S$ *and* $j=1,...,p$, *where* c_0 *is a constant.*

All the proofs of our main results delay to the appendices.

5. Simulation studies

5.1. Simulation 1

We first consider the heterogeneity with balanced groups, that is, the value of μ_i are uniformly generated from $\{\alpha,-\alpha\}$. And $\boldsymbol{\beta}$ are generated by a uniform distribution on $[0.5,1.5]$. The data matrix $\boldsymbol{X}$ of size $n\times p$ is generated from $N(\boldsymbol{0},\boldsymbol{\Sigma})$, where $(\boldsymbol{\Sigma})_{ij}=0.3^{|i-j|}$. We set $n=100,p=5$ in data matrix $\boldsymbol{X}$. The error items ϵ_i *iid* come from $N(0,0.5^2)$.

In tuning parameter λ and r selection, we choose the modified BIC (Wang et al., 2007) method, which minimize

$$
\mathrm{BIC}(\lambda,r)=\log\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu}_i(\lambda,r)-\boldsymbol{x_i}^T\hat{\boldsymbol{\beta}}(\lambda,r))^2\right]+\mathrm{C}_n\frac{\log n}{n}(\hat{K}(\lambda,r)+p),
$$

where C_n is a positive number depends on n. Particularly, we choose $C_n = c\log(\log(n+p))$, where c is a positive constant that can be chosen as 5.0 or 10.0 later in our simulation. From the relationship $2pr \leq n_S \leq 4pr < n$, we have $1 \leq r < \frac{n}{4p}$ and r in an integer.

From theorem 2, we choose $n_S = \omega p\sqrt{n}$. Let $\omega = 1.5$, then $n_S = 75$ in this simulation. We apply MCP and SCAD on the full data and SPF-based methods accordingly, which shows it is sufficient to identify the group structure on the screened data instead of the full data. We present the results in table 1.

From table 1 we can see that, the SPF-based method is of high ability to identify the group structure. The SPF-based method learn from a subsample while still estimate the group numbers $\hat{K}$ well, see the median are all equal to the true K. What's more surprising is that the SPF method can recover the group structure more clearly through the solution path, see Figure 1.

5.2. Simulation 2

To measure the accuracy of the results, we first declare several definitions. We let a true positive (TP) count two observations from the same underlying group to the same cluster, a true negative (TN) count two observations from different underlying groups to different clusters, a false positive (FP) count two observations from different underlying groups to the same cluster, a false negative (FN) count two observations from the same underlying group to different clusters. And then we use Rand Index (Rand, 1971)

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

to measure the accuracy. We also use false positive rate $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$ and false negative rate $\mathrm{FNR} = \mathrm{FN}/(\mathrm{FN} + \mathrm{TP})$ to evaluate the results.

In order to consider more diverse group structures, we provide 4 designs:

Design 1 In this case we consider two unbalanced groups. The data matrix $\boldsymbol{X}$, coefficients $\boldsymbol{\beta}$, error items ϵ and screened sample size n_S are set exactly as simulation 1. In this design, we let $\mathbb{P}(\mu_i = 2) = 0.3$ and $\mathbb{P}(\mu_i = -2) = 0.7$ for $i = 1, ..., n$.

Table 1: The mean, median and standard error of $\hat{K}$ in MCP, SCAD, SPF-MCP and SPF-SCAD methods, 100 replications.

c	α		MCP	SCAD	SPF-MCP	SPF-SCAD
5.0	1.0	Mean	1.87	1.88	1.79	2.04
		Median	2.00	2.00	2.00	2.00
		s.e.	0.42	0.36	0.56	0.52
	1.5	Mean	2.18	2.22	2.15	2.25
		Median	2.00	2.00	2.00	2.00
		s.e.	0.59	0.83	0.82	0.75
	2.0	Mean	1.96	1.89	1.95	1.97
		Median	2.00	2.00	2.00	2.00
		s.e.	0.60	0.64	0.85	0.73
10.0	1.0	Mean	1.84	1.79	1.73	1.85
		Median	2.00	2.00	2.00	2.00
		s.e.	0.34	0.28	0.67	0.51
	1.5	Mean	2.12	2.09	2.19	2.23
		Median	2.00	2.00	2.00	2.00
		s.e.	0.54	0.73	0.82	0.79
	2.0	Mean	1.95	1.84	1.85	1.83
		Median	2.00	2.00	2.00	2.00
		s.e.	0.74	0.83	0.84	0.91

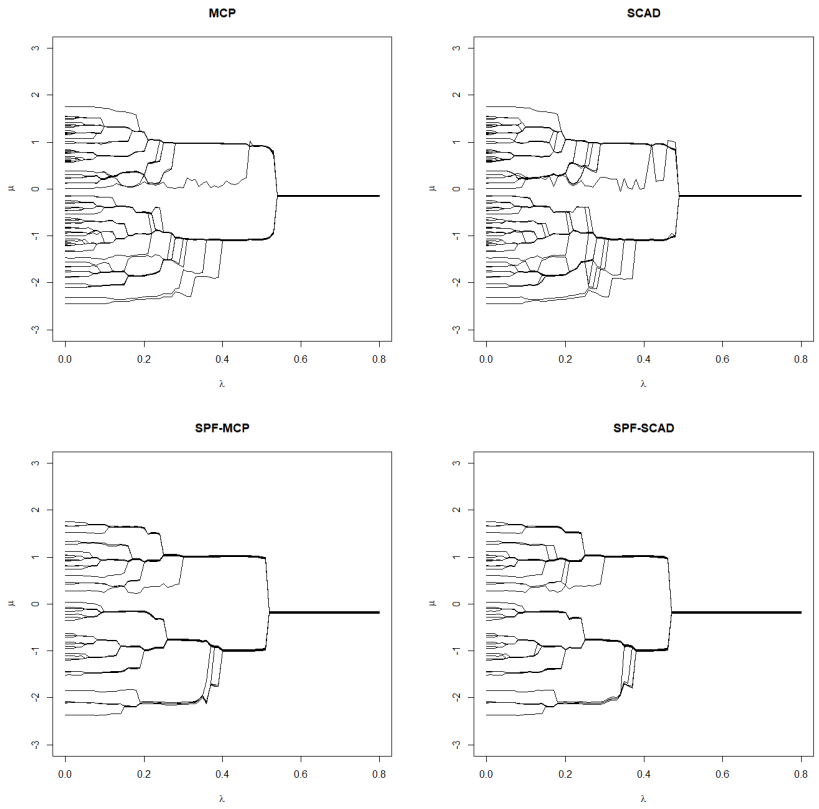

Figure 1: Balanced Groups: Comparison of the solution path in MCP, SCAD, SPF-MCP and SPF-SCAD methods. Here we choose $\alpha = 1,\ c = 5$.

Design 2 Consider the situation of three balanced groups. We let $\mathbb{P}(\mu_i = 2) = \mathbb{P}(\mu_i = 0) = \mathbb{P}(\mu_i = -2) = 1/3$ and other settings the same as Design 1.

Design 3 In this design we consider three unbalanced groups. Let $\mathbb{P}(\mu_i = 2) = 0.2$, $\mathbb{P}(\mu_i = 0) = 0.3$ and $\mathbb{P}(\mu_i = -2) = 0.7$ with other settings the same as Design 1.

We implement the MCP, SCAD method on a random subsample of size n_S and SPF-based method on the full sample n that would be screened into n_S. To make more comparision we also consider the screened pairwise fusion

with the lasso penalty, that is, $p_\gamma(t, \lambda) = \lambda t$, which is also aviliable to shrink the difference of pairs: $|\mu_i - \mu_j|$ by applying the same thresholding. Another method we implement is the gaussian mixture model-based clustering method (Fraley and Raftery, 2002) from a R package MCLUST, which is widly used in determning the number of groups selected by BIC method. All the results of four designs are shown in Table 2. Through analyzing the results, our findings can be summarized as follows.

1. the SMSE in all Designs shows the SPF-based methods enjoy the consistance on estimate of factor $\boldsymbol{\mu}$ and outperform other methods. 2. The SPF-based methods are more accurate in estimating the number of heterogeneous groups K. 3. The concave pairwise fusion methods all enjoy high accuracy when $K = 2$, however, when $K = 3$ the SPF-based method perform much better than the random sampled MCP and SCAD. 4. Nonconcave pairwise fusion method SPF-Lasso and mixture models cluster method SPF-Mclust lost the ability for group structure recovering in our designs.

6. Real data example

6.1. COVID-19 Clinical Data

In this part, we implement our SPF-based method for mass clinical testing of COVID-19 as well as find an arrange toward ICU and regular wards. We use the COVID-19 clinical dataset in Kaggle, which is collected from the Hospital Israelita Albert Einstein, at So Paulo, Brazil, ranging from March 28 to April 3st, 2020. All clinical data were standardized to have a mean of zero and a unit standard deviation. After delating the missing rows and columns, there have $n = 599$ patients and $p = 14$ variables. Davies et al., (2020) shows that the age-dependent effects is significant in transmission and control of cases. Moreover, a large amount of literature shows that age is related to red blood cell volumn, platelets, hemoglobin, monocytes etc., see Reynolds et al., (2014) for an example. Then we are inspired to choose $\boldsymbol{y} = \textit{patient age quantile}$ and covariates as shown in table 3.

Table 2: Unbalanced Groups: The mean and standard error (in parentheses) of $\hat{K}$, square root of the MSE (SMSE) for the estimated $\boldsymbol{\mu}$, the clustering accuracy and cpu-times by different methods with 4 designs, 100 replications.

Method	$\hat{K}$	SMSE	Accuracy%	FPR%	FNR%
		Design 1			
MCP	2.390(0.540)	0.256(0.085)	99.27	0.72	0.74
SCAD	2.330(0.530)	0.251(0.085)	99.28	0.80	0.66
SPF-MCP	2.000(0.532)	0.186(0.067)	99.46	0.47	0.58
SPF-SCAD	1.840(0.464)	0.174(0.066)	99.49	0.48	0.52
SPF-Lasso	1.030(0.086)	1.437(0.044)	58.73	99.30	0.24
SPF-Mclust	1.470(0.297)	1.773(0.087)	67.50	82.18	17.93
		Design 2			
MCP	3.960(0.579)	0.687(0.097)	74.62	3.28	1.05
SCAD	3.730(0.548)	0.742(0.112)	73.72	3.42	1.03
SPF-MCP	3.500(0.495)	0.530(0.085)	86.93	3.22	1.09
SPF-SCAD	3.450(0.514)	0.560(0.090)	86.11	3.37	1.02
SPF-Lasso	1.000(0.000)	1.290(0.022)	48.82	93.68	1.81
SPF-Mclust	1.380(0.244)	1.799(0.081)	50.84	83.96	15.94
		Design 3			
MCP	3.810(0.522)	0.632(0.095)	77.46	3.23	0.68
SCAD	3.830(0.549)	0.672(0.106)	76.29	3.49	0.61
SPF-MCP	3.150(0.504)	0.513(0.079)	90.36	3.08	0.75
SPF-SCAD	3.120(0.518)	0.537(0.093)	89.55	3.21	0.74
SPF-Lasso	1.030(0.086)	1.219(0.034)	50.85	100	0.00
SPF-Mclust	1.500(0.279)	1.757(0.089)	54.91	79.64	20.37

Table 3: Covariates chosen in COVID-19 clinical data.

$\boldsymbol{x}_1$=Hemoglobin	$\boldsymbol{x}_8$=Basophils
$\boldsymbol{x}_2$=Platelets	$\boldsymbol{x}_9$=Mean corpuscular hemoglobin (MCH)
$\boldsymbol{x}_3$=Mean platelet volume	$\boldsymbol{x}_{10}$=Eosinophils
$\boldsymbol{x}_4$=Red blood Cells	$\boldsymbol{x}_{11}$=Mean corpuscular volume (MCV)
$\boldsymbol{x}_5$=Lymphocytes	$\boldsymbol{x}_{12}$=Monocytes
$\boldsymbol{x}_6$=Mean corpuscular hemoglobin concentration(MCHC)	$\boldsymbol{x}_{13}$=Red blood cell distribution width (RDW)
$\boldsymbol{x}_7$=Leukocytes	

We first obtained the OLS estimate $\hat{\boldsymbol{\beta}}_{OLS}$ and then we plot the density of $y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{OLS}$ as shown in the left panel of figure 2. We find that the response after adjusting for the effects of the covariates still has heterogenious group structure, which is caused by unobserved latent factors. Then we apply SPF-MCP and SPF-SCAD method on the full data after which we plot the density of $y_{iS} - \hat{\mu}_{iS} - \boldsymbol{x}_{iS}^T \hat{\boldsymbol{\beta}}_S$ in the right panel of figure 2. We only plot SPF-MCP method since two methods get nearly the same result. We also compare the estimate of partial coefficients by our SPF-based methods and OLS in table 4, from which we find MCHC and MCH significant by SPF-based methods yet ignored by OLS. Tuning of the model is based on the BIC methods, and with the best choice of λ, it can be found $\hat{K} = 3$ groups, denoted by $\hat{\mathcal{G}}_1$, $\hat{\mathcal{G}}_2$ and $\hat{\mathcal{G}}_3$ in the screened data $\boldsymbol{X}_S^*, \boldsymbol{y}_S^*$. Since the latent factor in our linear model can be contained in the object-specific intercepts, the information of group structure is shown to the identifiers by $\hat{\boldsymbol{\mu}}$. In order to have a clear measure for the interpretability of the factor, we let

$$\begin{aligned}
\mathcal{H}_1 &= \{i \in [n] : \textit{patient addmited to regular ward}\} \\
\mathcal{H}_2 &= \{i \in [n] : \textit{patient addmited to semi} - \textit{intensive unit}\} \\
\mathcal{H}_3 &= \{i \in [n] : \textit{patient addmited to intensive care unit}\} \\
\mathcal{T}_* &= \{i \in [n] : \mathrm{SARS} - \mathrm{Cov} - 2 \textit{ exam result shows positive}\}
\end{aligned}$$

which are also contained in the COVID-19 clinical dataset. Hence, to evaluate the information we recover from the SPF-based methods, we define the grouping

arrange accuracy (GAc) as

$$GAc(j) = \max_{i\in\{1,2,3\}} \left\{ \frac{|\hat{\mathcal{G}}_i \cap \mathcal{H}_j|}{|\hat{\mathcal{G}}_i|} \right\}, j \in \{1,2,3\} \quad GAc_* = \max_{i\in\{1,2,3\}} \left\{ \frac{|\hat{\mathcal{G}}_i \cap \mathcal{T}_*|}{|\hat{\mathcal{G}}_i|} \right\}.$$

After computing based on $\hat{\mathcal{G}}_1$, $\hat{\mathcal{G}}_2$ and $\hat{\mathcal{G}}_3$, we derive $GAc(1) = 0.83$, $GAc(2) = 0.85$, $GAc(3) = 0.82$ and $GAc_* = 0.76$, which shows the great interpretability of latent factors , that will provide the identifiers an suggestion on the arrangement of ICU. To futhermore evaluate the quality of the cluster method, we use the DaviesBouldin (DB) index. Denote by c_k and σ_k the centroid of cluster k and the average distance of $y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ to c_k in cluster k, then $DB = \hat{K}^{-1}\sum_{k=1}^{\hat{K}} \max_{k'\neq k}((\sigma_k+\sigma_{k'})/d(c_k, c_{k'}))$, where $d(c_k, c_{k'})$ is the distance between c_k and $c_{k'}$. It's known that the best method enjoy the smallest DB value. The DB values for SPF-MCP, SPF-SCAD, SPF-Lasso and SPF-Mclust are 0.326, 0.332, 0.574 and 0.565, which shows the superiority of concave penalty methods.

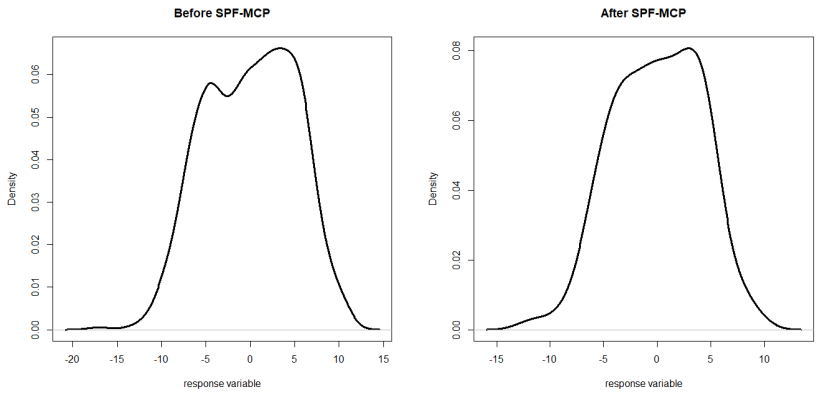

Figure 2: The density plot of the response variable after adjusting for the effects of the covariates before (left) and after (right) SPF-MCP method.

6.2. India Liver Patient Data

We apply the proposed methods to the Indian Liver Patient Dataset (ILPD) which has been studied in several literatures. This data set has $n = 582$ liver

Table 4: The estimated values (est) for the coefficients of $\boldsymbol{x}_3, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_9, \boldsymbol{x}_{13}$, their standard errors (s.e.), and the p-values for testing the significance of the coefficients by OLS, SPF-MCP, and SPF-SCAD, respectively.

Methods		$\boldsymbol{\beta}_3$	$\boldsymbol{\beta}_5$	$\boldsymbol{\beta}_6$	$\boldsymbol{\beta}_7$	$\boldsymbol{\beta}_9$	$\boldsymbol{\beta}_{13}$
SPF-MCP	est	0.788	-2.591	-4.289	-1.994	8.318	0.673
	s.e.	0.345	0.434	2.231	0.396	4.863	0.232
	p-value	0.157	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**
SPF-SCAD	est	0.737	-2.351	-4.529	-2.050	8.987	0.635
	s.e.	0.341	0.421	2.325	0.401	4.869	0.202
	p-value	0.205	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**
OLS	est	0.265	-1.097	-4.138	-1.047	8.015	1.096
	s.e.	0.229	0.235	2.225	0.278	4.850	0.255
	p-value	0.249	$<$ **0.001**	0.063	$<$ **0.001**	0.099	$<$ **0.001**

patients with $p = 10$ variables: age, gender, total bilirubin (TB), direct bilirubin (DB), alkaline phosphotase (Alkphos), alamine aminotransferase (Sgpt), aspartate aminotransferase (Sgot), total protiens (TP), albumin (ALB) and albumin and globulin ratio (A/G Ratio). And the response variable y values at 0 or 1 is used to split the data into two sets, which was labeled by experts. This response can also be viewed as the diagnostic suggestions from the experts.

We implement the SPF-SCAD and SPF-MCP procedure by a refit method. See Zheng et al. (2014) for more details about the idea of refitting. Since the response y is not a quantitative measure of patients, we need to convert this 0/1 variable to quantity firstly. Inspired by the idea of logit link function in generalized linear model, we fit the generalized linear model with homogeneous intercept

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \boldsymbol{x}_i^T\boldsymbol{\beta},$$

where $p_i = \mathbb{P}(y_i = 1|\boldsymbol{x}_i)$. After fitting the model by MLE, we get the prediction $y_{0i} = \hat{p}_i$ for $i = 1, ..., n$. Then we can refit our model using the heterogeneous

intercept on pseudo observations $\boldsymbol{y}_0 = (y_{01}, ..., y_{0n})^T$

$$y_{0i} = \mu_i + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i. \tag{15}$$

We then implement the SPF-SCAD and SPF-MCP procedure and derive the solution path on the heterogeneous intercept $\boldsymbol{\mu}$, see figure 3. We find that when $\lambda = 0.07 \pm 0.01$, two main subgroups can be indentified in both methods. Then the patients can be divided into general patients and more serious patients based on the experts diagnostic opinions and results of our methods. Now we are available to do hypothesis testing and clustering evaluation. For SPF-MCP, SPF-SCAD and OLS methods, the R^2 are 0.728, 0.702 and 0.610, and the corresponding DB values are 0.449, 0.456 and 0.528, which shows our methods can better identify the heterogeneity of the model.

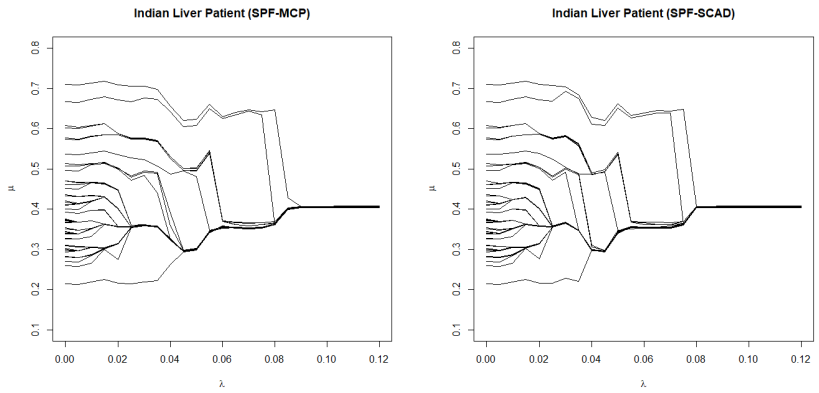

Figure 3: Solution path of SPF-MCP and SPF-SCAD. Here $n_S = p\log(n) = 54$.

7. Discussion

In this paper, we innovatively propose a preliminary sample screening strategy to greatly reduce the complexity of pairwise fushion in subgroup analysis, which efficiently recover the group structure and estimate the latent factor simultaneously. Our propsed SPF-based concave method outperform the other methods as shown in the experiment of comparsion. In the field of health data

analysis, our method also has efficient group recovering capability and the latent factor $\boldsymbol{\mu}$ is of great interpretability. At the same time, our method is scalable in computing, which provides a guarantee for the pairwise fusion in subgroup analysis with ultra-large sample size.

For future work, we provide some points of views. First of all, whether the screening method based on sample information has the guarantee of sure screen property. Moreover, our information based screen method can be combined with some other appealing methods in subgroup analysis which desire scalable computing. Some methods with shrinkage capacity (Zheng et al., 2014) can also be introduced to screened pairwise fusion. However, there are still challenging theoretical difficulties that need to be resolved.

Acknowledgments

Zheng's research is supported by National Natural Science Foundation of China (Grants 72071187, 11671374, 71731010, and 71921001) and Fundamental Research Funds for the Central Universities (Grants WK3470000017 and WK2040000027). The authors sincerely thank the Co-Editor, Associate Editor, and anonymous referees for their valuable and constructive comments that helped improve the article substantially.

Appendix A. Proof of Proposition 1

In the beginning, we give a useful lemma for the main proof. For $i = 1, ..., n, j = 1, ..., p$, denote by $\boldsymbol{x}_{(i)j}$ the ith order statistic for $\boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{nj}$ and $\boldsymbol{x}_{(i)jS}$ for $\boldsymbol{x}_{1jS}, ..., \boldsymbol{x}_{njS}$ accordingly.

Lemma 1. *For any $j = 1, ..., p$, the sample variance of $\boldsymbol{x}_{jS}$ has a lower bound as*

$$var(\boldsymbol{x}_{jS}) \geq \frac{r(x_{(n-r+1)j} - x_{(r)j})^2}{2(n_S - 1)}. \tag{A.1}$$

Proof of Lemma 1. Let $\bar{x}_{jS}$ be the sample mean of $\boldsymbol{x}_{jS}$ and

$$\begin{aligned}\bar{x}_j^{l+u} &= (\sum_{i=1}^{r} + \sum_{i=n-r+1}^{n})x_{(i)j}/(2r),\\ \bar{x}_j^{l} &= \sum_{i=1}^{r} x_{(i)j}/r,\\ \bar{x}_j^{u} &= \sum_{i=n-r+1}^{n} x_{(i)j}/r.\end{aligned} \tag{A.2}$$

For any given j, we have

$$\begin{aligned}(n_S-1)var(\boldsymbol{x}_{jS}) &= \sum_{i=1}^{n_S}(x_{ijS}-\bar{x}_{jS})^2\\ &= \left(\sum_{i=1}^{r}+\sum_{i=n-r+1}^{n}\right)(x_{(i)j}-\bar{x}_{jS})^2 + \sum_{l\neq j}\left(\sum_{i=1}^{r}+\sum_{i=n-r+1}^{n}\right)(x_j^{(i)l}-\bar{x}_{jS})^2\\ &\geq \left(\sum_{i=1}^{r}+\sum_{i=n-r+1}^{n}\right)(x_{(i)j}-\bar{x}_j^{l+u})^2\\ &= \sum_{i=1}^{r}(x_{(i)j}-\bar{x}_j^{l})^2 + \sum_{i=n-r+1}^{n}(x_{(i)j}-\bar{x}_j^{u})^2 + \frac{r}{2}(\bar{x}_j^{u}-\bar{x}_j^{l})^2\\ &\geq \frac{r}{2}(\bar{x}_j^{u}-\bar{x}_j^{l})^2 \geq \frac{r}{2}(x_{(n-r+1)j}-x_{(r)j})^2,\end{aligned}$$

where $x_j^{(i)l}$ satisfies that for $s=1,...,n$, if $x_{(i)l}=x_{sl}$ then $x_j^{(i)l}=x_{sj}$. □

Now we give the proof of Proposition 1.

Proof of Proposition 1. Since for every $\boldsymbol{\mu}_S \in \mathcal{M}_{\mathcal{G}}$, it can be written as $\boldsymbol{\mu}_S = \boldsymbol{Z}\boldsymbol{\alpha}_S$. Then we have

$$\begin{aligned}\begin{pmatrix}\hat{\boldsymbol{\alpha}}_S^{or}\\ \hat{\boldsymbol{\beta}}_S^{or}\end{pmatrix} &= \underset{\boldsymbol{\alpha}_S\in\mathbb{R}^K,\boldsymbol{\beta}_S\in\mathbb{R}^p}{argmin}\ \frac{1}{2}||\boldsymbol{y}_S-\boldsymbol{Z}_S\boldsymbol{\alpha}_S-\boldsymbol{X}_S\boldsymbol{\beta}_S||\\ &= [(\boldsymbol{Z}_S,\boldsymbol{X}_S)^T(\boldsymbol{Z}_S,\boldsymbol{X}_S)]^{-1}(\boldsymbol{Z}_S,\boldsymbol{X}_S)^T\boldsymbol{y}_S.\end{aligned}$$

Then the information matrix based on the screened data $(\boldsymbol{Z}_S,\boldsymbol{X}_S),\boldsymbol{y}_S$ can be written as

$$(\boldsymbol{Z}_S,\boldsymbol{X}_S)^T(\boldsymbol{Z}_S,\boldsymbol{X}_S) = \boldsymbol{B}^{-1}\begin{pmatrix} n_S\boldsymbol{I}_K & \boldsymbol{0}^T\\ \boldsymbol{0} & (n_S-1)\boldsymbol{\Sigma}\end{pmatrix}(\boldsymbol{B}^T)^{-1}, \tag{A.3}$$

where

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{A} & \boldsymbol{D} \end{pmatrix}, \tag{A.4}$$

here we have

$$\boldsymbol{D} = \begin{pmatrix} var(\boldsymbol{x}_{1S})^{-\frac{1}{2}} & & \\ & \ddots & \\ & & var(\boldsymbol{x}_{pS})^{-\frac{1}{2}} \end{pmatrix}$$

and $\boldsymbol{A} \in \mathbb{R}^{p\times K}$ satisfies $\boldsymbol{A}\boldsymbol{Z}_S^T = -\boldsymbol{D}\boldsymbol{X}_S^T$. Denote $\hat{\boldsymbol{\gamma}}_S^{or}$ for $\begin{pmatrix} \hat{\boldsymbol{\alpha}}_S^{or} \\ \hat{\boldsymbol{\beta}}_S^{or} \end{pmatrix}$, from (A.3) and (A.4),

$$\begin{aligned} V(\hat{\boldsymbol{\gamma}}_S^{or}|\boldsymbol{X}) &= \sigma^2[(\boldsymbol{Z}_S, \boldsymbol{X}_S)^T(\boldsymbol{Z}_S, \boldsymbol{X}_S)]^{-1} \\ &= \sigma^2 \boldsymbol{B}^T \begin{pmatrix} \frac{1}{n_S}\boldsymbol{I}_K & \boldsymbol{0}^T \\ \boldsymbol{0} & \frac{1}{n_S-1}\boldsymbol{\Sigma}^{-1} \end{pmatrix} \boldsymbol{B} \end{aligned}$$

Hence, we derive

$$V(\hat{\alpha}_{iS}^{or}) = \sigma^2\big(\frac{1}{n_S} + \frac{1}{n_S-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{e}_i\big),\ i = 1, ..., K.$$

and

$$V(\hat{\beta}_{jS}^{or}) = \frac{\sigma^2}{n_S-1}\frac{(\boldsymbol{\Sigma}^{-1})_{jj}}{var(\boldsymbol{x}_{jS})},\ j = 1, ..., p.$$

where $\boldsymbol{e}_i$ is an $K \times 1$ zero vector except the ith element is 1 and $(\boldsymbol{\Sigma}^{-1})_{jj}$ is the jth diagonal element of $\boldsymbol{\Sigma}^{-1}$.

Since $\boldsymbol{\Sigma}^{-1}$ is positive semidefine, we have $V(\hat{\alpha}_{iS}^{or}) \geq \sigma^2/n_S$ because $\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\boldsymbol{e}_i \geq 0$. Moreover, from the construct of $\boldsymbol{A}$ we know there exist a constant $0 < c_0 < \infty$ such that $\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{e}_i \leq c_0$, which leads to $V(\hat{\alpha}_{iS}^{or}) \leq \frac{\sigma^2}{n_S-1}(1 + c_0\lambda_{\min}^{-1}(\boldsymbol{\Sigma}))$. Then $\frac{\sigma^2}{n_S-1}(1 + c_0\lambda_{\min}^{-1}(\boldsymbol{\Sigma})) \geq V(\hat{\mu}_{iS}^{or}|\boldsymbol{X}) \geq \sigma^2/n_S$ accordingly.

We now use the spectral decompositon of $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$. It's known that $\boldsymbol{\Lambda}^{-1} \leq \lambda_{min}^{-1}(\boldsymbol{\Sigma})\boldsymbol{I}_p$, this leads to $\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T \leq \boldsymbol{U}\lambda_{min}^{-1}(\boldsymbol{\Sigma})\boldsymbol{I}_p\boldsymbol{U}^T = \lambda_{min}^{-1}(\boldsymbol{\Sigma})\boldsymbol{I}_p^T$, then we have $\boldsymbol{\Sigma}_{jj}^{-1} \leq \lambda_{min}^{-1}(\boldsymbol{\Sigma})$ for all j.

From Lemma 1 we know

$$var(\boldsymbol{x}_{jS}) \geq \frac{r(x_{(n)j} - x_{(1)j})^2}{2(n_S - 1)} \left(\frac{x_{(n-r+1)j} - x_{(r)j}}{x_{(n)j} - x_{(1)j}} \right)^2.$$

Since we have the fact

$$\left| x_{ijS} - \frac{x_{(n)j} + x_{(1)j}}{2} \right| \leq \frac{x_{(n)j} - x_{(1)j}}{2}$$

for all $i = 1, ..., n_S$, which leads to the following inequality

$$var(\boldsymbol{x}_{jS}) \leq \frac{1}{n_S - 1} \sum_{i=1}^{n_S} \left(x_{ijS} - \frac{x_{(n)j} + x_{(1)j}}{2} \right)^2 \leq \frac{n_S}{4(n_S - 1)} (x_{(n)j} - x_{(1)j})^2.$$

From the facts above, we have

$$V(\hat{\beta}_{jS}^{or} | \boldsymbol{X}) = \frac{\sigma^2}{n_S - 1} \frac{(\boldsymbol{\Sigma}^{-1})_{jj}}{var(\boldsymbol{x}_{jS})} \leq \frac{4p\sigma^2}{n_S \lambda_{min}(\boldsymbol{\Sigma})(x_{(n-r+1)j} - x_{(r)j})^2}.$$

Similarly, we can derive the lower bound

$$V(\hat{\beta}_{jS}^{or} | \boldsymbol{X}) = \frac{\sigma^2}{n_S - 1} \frac{(\boldsymbol{\Sigma}^{-1})_{jj}}{var(\boldsymbol{x}_{jS})} \geq \frac{4p\sigma^2}{n_S \lambda_{min}(\boldsymbol{\Sigma})(x_{(n)j} - x_{(1)j})^2}.$$

□

Appendix B. Proof of Theorem 2

Proof of Theorem 2. Without loss of generality, we let the screened sample of size n_S concentrate on the first n_S rows of $\boldsymbol{X}$, that is $\boldsymbol{X}_S$ is non-zero in the first n_S rows with other rows zeros. When the underlying groups are known, we have $\hat{\boldsymbol{\mu}}_S^{or} = \boldsymbol{Z}_S \hat{\boldsymbol{\alpha}}_S^{or}$ and $\hat{\boldsymbol{\mu}}^{or} = \boldsymbol{Z} \hat{\boldsymbol{\alpha}}^{or}$ which implies that $((\hat{\boldsymbol{\mu}}^{or})^T, (\hat{\boldsymbol{\beta}}^{or})^T)^T = ((\boldsymbol{Z}\hat{\boldsymbol{\alpha}}^{or})^T, (\hat{\boldsymbol{\beta}}^{or})^T)^T$, then the proof has two main steps: first we bound $||\hat{\boldsymbol{\mu}}^{or} - \boldsymbol{\mu}^*||_\infty$ and then we bound $||\hat{\boldsymbol{\mu}}_S^{or} - \hat{\boldsymbol{\mu}}^{or}||_\infty$. Now we consider the full sample $\boldsymbol{X}$. We know that the OLS estimate of $((\hat{\boldsymbol{\alpha}}^{or})^T, (\hat{\boldsymbol{\beta}}^{or})^T)^T$ has form

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}^{or} \\ \hat{\boldsymbol{\beta}}^{or} \end{pmatrix} = [(\boldsymbol{Z}, \boldsymbol{X})^T (\boldsymbol{Z}, \boldsymbol{X})]^{-1} (\boldsymbol{Z}, \boldsymbol{X})^T \boldsymbol{y}.$$

Then for $\boldsymbol{\alpha}^* = (\alpha_1^*, ..., \alpha_K^*)$ and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$ we derive

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^* \\ \hat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^* \end{pmatrix} = [(\boldsymbol{Z}, \boldsymbol{X})^T (\boldsymbol{Z}, \boldsymbol{X})]^{-1} (\boldsymbol{Z}, \boldsymbol{X})^T \boldsymbol{\epsilon},$$

and accordingly we have

$$\left|\left|\begin{pmatrix}\hat{\boldsymbol{\alpha}}^{or}-\boldsymbol{\alpha}^{*}\\ \hat{\boldsymbol{\beta}}^{or}-\boldsymbol{\beta}^{*}\end{pmatrix}\right|\right|_{\infty}\leq |||[(\boldsymbol{Z},\boldsymbol{X})^{T}(\boldsymbol{Z},\boldsymbol{X})]^{-1}||_{\infty}||(\boldsymbol{Z},\boldsymbol{X})^{T}\boldsymbol{\epsilon}||_{\infty}$$

Under assumption 1 we have $|||[(\boldsymbol{Z},\boldsymbol{X})^{T}(\boldsymbol{Z},\boldsymbol{X})]^{-1}||_{2}\leq\frac{1}{C_{0}|\mathcal{G}_{\min}|}$ and then we derive $|||[(\boldsymbol{Z},\boldsymbol{X})^{T}(\boldsymbol{Z},\boldsymbol{X})]^{-1}||_{\infty}\leq\frac{\sqrt{K+p}}{C_{0}|\mathcal{G}_{\min}|}$. Futhermore, by union bound and assumption 2, for a constant $0<C<\infty$

$$\begin{aligned}&\mathbb{P}(||\boldsymbol{Z}^{T}\boldsymbol{\epsilon}||_{\infty}>C\sqrt{n\log n})\\&\leq\sum_{k=1}^{K}\mathbb{P}(|\sum_{i\in\mathcal{G}}\epsilon_{i}|>\sqrt{|\mathcal{G}_{k}|}C\sqrt{\log n})\leq 2K\exp(-c_{1}C^{2}\log n)=2Kn^{-c_{1}C^{2}}\end{aligned}$$

and accordingly,

$$\begin{aligned}&\mathbb{P}(||\boldsymbol{X}^{T}\boldsymbol{\epsilon}||_{\infty}>C\sqrt{n\log n})\\&\leq\sum_{j=1}^{p}\mathbb{P}(|\boldsymbol{X}_{j}^{T}\boldsymbol{\epsilon}|>\sqrt{n}C\sqrt{\log n})\leq 2p\exp(-c_{1}C^{2}\log n)=2pn^{-c_{1}C^{2}},\end{aligned}$$

thus we have

$$\begin{aligned}&\mathbb{P}(||(\boldsymbol{Z},\boldsymbol{X})^{T}\boldsymbol{\epsilon}||_{\infty}>C\sqrt{n\log n})\\&\leq\mathbb{P}(||\boldsymbol{Z}^{T}\boldsymbol{\epsilon}||_{\infty}>C\sqrt{n\log n})+\mathbb{P}(||\boldsymbol{X}^{T}\boldsymbol{\epsilon}||_{\infty}>C\sqrt{n\log n})\leq 2(K+p)n^{-c_{1}C^{2}}.\end{aligned}\tag{B.1}$$

Now we derive with probability at least $1-2(K+p)n^{-c_{1}C^{2}}$,

$$\left|\left|\begin{pmatrix}\hat{\boldsymbol{\mu}}^{or}-\boldsymbol{\mu}^{*}\\ \hat{\boldsymbol{\beta}}^{or}-\boldsymbol{\beta}^{*}\end{pmatrix}\right|\right|_{\infty}=\left|\left|\begin{pmatrix}\hat{\boldsymbol{\alpha}}^{or}-\boldsymbol{\alpha}^{*}\\ \hat{\boldsymbol{\beta}}^{or}-\boldsymbol{\beta}^{*}\end{pmatrix}\right|\right|_{\infty}\leq C\frac{\sqrt{K+p}}{C_{0}|\mathcal{G}_{\min}|}\sqrt{n\log n}\tag{B.2}$$

Until now we finish the first step of our proof. To start the second step, we use the form of OLS estimate as

$$\begin{aligned}||\hat{\boldsymbol{\alpha}}_{S}^{or}-\hat{\boldsymbol{\alpha}}^{or}||_{\infty}=&||(\boldsymbol{I}_{K},\boldsymbol{0})[(\boldsymbol{Z}_{S},\boldsymbol{X}_{S})^{T}(\boldsymbol{Z}_{S},\boldsymbol{X}_{S})]^{-1}(\boldsymbol{Z}_{S},\boldsymbol{X}_{S})^{T}\boldsymbol{y}_{S}\\&-(\boldsymbol{I}_{K},\boldsymbol{0})[(\boldsymbol{Z},\boldsymbol{X})^{T}(\boldsymbol{Z},\boldsymbol{X})]^{-1}(\boldsymbol{Z},\boldsymbol{X})^{T}\boldsymbol{y}||_{\infty}.\end{aligned}$$

From the proof of Proposition 1, we know that there exist

$$\boldsymbol{B}_1 = \begin{pmatrix} \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{A}_1 & \boldsymbol{D}_1 \end{pmatrix}, \ \boldsymbol{D}_1 = \text{diag}(var(\boldsymbol{x}_{1S})^{-\frac{1}{2}}, ..., var(\boldsymbol{x}_{pS})^{-\frac{1}{2}})$$

such that (A.3) holds, hence we have

$$\begin{aligned}
&(\boldsymbol{I}_K, \boldsymbol{0})[(\boldsymbol{Z}_S, \boldsymbol{X}_S)^T(\boldsymbol{Z}_S, \boldsymbol{X}_S)]^{-1}(\boldsymbol{Z}_S, \boldsymbol{X}_S)^T\boldsymbol{y}_S \\
&= (\boldsymbol{I}_K, \boldsymbol{0})\boldsymbol{B}_1^T \begin{pmatrix} \frac{1}{n_S}\boldsymbol{I}_K & \boldsymbol{0}^T \\ \boldsymbol{0} & \frac{1}{(n_S-1)}\boldsymbol{\Sigma}^{-1} \end{pmatrix} \boldsymbol{B}_1(\boldsymbol{Z}_S, \boldsymbol{X}_S)^T\boldsymbol{y}_S \\
&= (\boldsymbol{I}_K, \boldsymbol{0}) \begin{pmatrix} \boldsymbol{I}_K & \boldsymbol{A}_1^T \\ \boldsymbol{0}^T & \boldsymbol{D}_1^T \end{pmatrix} \begin{pmatrix} \frac{1}{n_S}\boldsymbol{I}_K & \boldsymbol{0}^T \\ \boldsymbol{0} & \frac{1}{n_S-1}\boldsymbol{\Sigma}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{A}_1 & \boldsymbol{D}_1 \end{pmatrix} (\boldsymbol{Z}_S, \boldsymbol{X}_S)^T\boldsymbol{y}_S \\
&= \left(\frac{1}{n_S}\boldsymbol{Z}_S^T + \frac{1}{n_S-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}_1\boldsymbol{Z}_S^T + \frac{1}{n_S-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{D}_1\boldsymbol{X}_S^T\right)\boldsymbol{y}_S.
\end{aligned}$$

Accordingly, in the same manner for the full sample size, we have

$$\begin{aligned}
&(\boldsymbol{I}_K, \boldsymbol{0})[(\boldsymbol{Z}, \boldsymbol{X})^T(\boldsymbol{Z}, \boldsymbol{X})]^{-1}(\boldsymbol{Z}, \boldsymbol{X})^T\boldsymbol{y} \\
&= \left(\frac{1}{n}\boldsymbol{Z}^T + \frac{1}{n-1}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{A}_2\boldsymbol{Z}^T + \frac{1}{n-1}\boldsymbol{\Sigma}_f^{-1}\boldsymbol{D}_2\boldsymbol{X}^T\right)\boldsymbol{y},
\end{aligned}$$

where

$$\boldsymbol{B}_2 = \begin{pmatrix} \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{A}_2 & \boldsymbol{D}_2 \end{pmatrix}, \ \boldsymbol{D}_2 = \text{diag}(var(\boldsymbol{x}_1)^{-\frac{1}{2}}, ..., var(\boldsymbol{x}_p)^{-\frac{1}{2}})$$

and $\boldsymbol{\Sigma}_f$ is the sample covariance matrix of $\boldsymbol{X}$. Meanwhile, we apply the decomposition method of Gram matrix on $||\hat{\boldsymbol{\beta}}_S^{or} - \hat{\boldsymbol{\beta}}^{or}||_\infty$ accordingly. Since $\boldsymbol{A}_1\boldsymbol{Z}_S^T = -\boldsymbol{D}_1\boldsymbol{X}_S^T$ and $\boldsymbol{A}_2\boldsymbol{Z}^T = -\boldsymbol{D}_2\boldsymbol{X}^T$ is satisfied for the decomposition of $[(\boldsymbol{Z}_S, \boldsymbol{X}_S)^T(\boldsymbol{Z}_S, \boldsymbol{X}_S)]^{-1}$ and $[(\boldsymbol{Z}, \boldsymbol{X})^T(\boldsymbol{Z}, \boldsymbol{X})]^{-1}$. Denote by $[\boldsymbol{Z}]_S$ the $n \times K$ matrix with n_S corresponding rows of copies in $\boldsymbol{Z}$ and $n - n_S$ rows of zeros,

which is different from the definition of $\boldsymbol{Z}_S$. We then derive the bound

$$\begin{aligned}
\left|\left|\begin{pmatrix} \hat{\boldsymbol{\alpha}}_S^{or} - \hat{\boldsymbol{\alpha}}^{or} \\ \hat{\boldsymbol{\beta}}_S^{or} - \hat{\boldsymbol{\beta}}^{or} \end{pmatrix}\right|\right|_\infty &\leq \left|\left|\frac{1}{n_S}\boldsymbol{Z}_S^T\boldsymbol{y}_S - \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{y}\right|\right|_\infty \leq \left|\left|\left(\frac{1}{n_S}[\boldsymbol{Z}]_S^T - \frac{1}{n}\boldsymbol{Z}^T\right)\boldsymbol{y}\right|\right|_\infty \\
&\leq \left|\left|\left(\frac{1}{n_S}[\boldsymbol{Z}]_S^T - \frac{1}{n}\boldsymbol{Z}^T\right)\left(\boldsymbol{\mu}^* + \boldsymbol{X}\boldsymbol{\beta}^*\right)\right|\right|_\infty \\
&+ \left|\left|\left(\frac{1}{n_S}[\boldsymbol{Z}]_S^T - \frac{1}{n}\boldsymbol{Z}^T\right)\boldsymbol{\epsilon}\right|\right|_\infty \\
&= \Gamma_1 + \Gamma_2.
\end{aligned}$$

Now we consider Γ_1 and Γ_2 separately. Since $n_S \leq \frac{n}{2}$, which leads to $\frac{1}{n_S} - \frac{1}{n} \geq \frac{1}{n}$, then for Γ_1, by assumption 1 we have

$$\begin{aligned}
\Gamma_1 &= \left|\left|\left(\frac{1}{n_S}[\boldsymbol{Z}]_S^T - \frac{1}{n}\boldsymbol{Z}^T\right)\left(\boldsymbol{\mu}^* + \boldsymbol{X}\boldsymbol{\beta}^*\right)\right|\right|_\infty \\
&\leq \left(\frac{1}{n_S} - \frac{1}{n}\right)||\boldsymbol{\mu}^* + \boldsymbol{X}\boldsymbol{\beta}^*||_\infty \leq \left(\frac{1}{n_S} - \frac{1}{n}\right)(||\boldsymbol{\mu}^*||_\infty + ||\boldsymbol{X}||_\infty||\boldsymbol{\beta}^*||_1) \\
&\leq \left(\frac{1}{n_S} - \frac{1}{n}\right)C_p, \qquad \text{(B.3)}
\end{aligned}$$

where $C_p = (1 - C_1p^2)C_2$ and $||(\boldsymbol{\mu}^{*T}, \boldsymbol{\beta}^{*T})^T||_\infty = C_2$.

From assumption 2, $\boldsymbol{\epsilon}$ enjoys the sub-Gaussian tails bound, using constant C the same as above, we let

$$Y = \max\left\{C\left(\frac{1}{n_S} - \frac{1}{n}\right)\sqrt{\log n_S}, C\frac{\sqrt{\log n}}{n}\right\}.$$

Then the union bound and tails bound lead to

$$\begin{aligned}
&\mathbb{P}\left(\left(\frac{1}{n_S}[\boldsymbol{Z}]_S^T - \frac{1}{n}\boldsymbol{Z}^T\right)\boldsymbol{\epsilon}\right) \\
&\leq \sum_{i=1}^{n_S}\mathbb{P}\left(|\epsilon_i| > \frac{Y}{\frac{1}{n_S} - \frac{1}{n}}\right) + \sum_{i=n_S+1}^{n}\mathbb{P}\left(|\epsilon_i| > nY\right) \\
&\leq 2n_S\exp\left\{-c_1\left(\frac{Y}{\frac{1}{n_S} - \frac{1}{n}}\right)^2\right\} + 2(n - n_S)\exp\left\{-c_1(nY)^2\right\}.
\end{aligned}$$

For $Y_1 = C(\frac{1}{n_S} - \frac{1}{n})\sqrt{\log n_S}$ and $Y_2 = C\sqrt{\log n}/n$, we derive

$$\begin{aligned}
&\mathbb{P}\bigg(\bigg(\frac{1}{n_S}[\boldsymbol{Z}]_S^T - \frac{1}{n}\boldsymbol{Z}^T\bigg)\boldsymbol{\epsilon}\bigg) \\
&\leq 2n_S \exp\bigg\{-c_1\bigg(\frac{Y_1}{\frac{1}{n_S}-\frac{1}{n}}\bigg)^2\bigg\} + 2(n-n_S)\exp\bigg\{-c_1(nY_2)^2\bigg\} \\
&\leq 2n_S^{1-c_1C^2} + 2n^{1-c_1C^2}. \qquad \text{(B.4)}
\end{aligned}$$

Combine the bound of Γ_1 and Γ_2, there exist a constant $0 < C_3 < \infty$, with probability at least $1 - 2n_S^{1-c_1C^2} - 2n^{1-c_1C^2}$

$$\begin{aligned}
&\left\| \begin{pmatrix} \hat{\boldsymbol{\alpha}}_S^{or} - \hat{\boldsymbol{\alpha}}^{or} \\ \hat{\boldsymbol{\beta}}_S^{or} - \hat{\boldsymbol{\beta}}^{or} \end{pmatrix} \right\|_\infty \\
&\leq \max\bigg\{\bigg(\frac{1}{n_S}-\frac{1}{n}\bigg)(C\sqrt{\log n_S} + C_p), \bigg(\frac{1}{n_S}-\frac{1}{n}\bigg)C_p + C\frac{\sqrt{\log n}}{n}\bigg\} \\
&\leq C_3\bigg(\frac{1}{n_S}-\frac{1}{n}\bigg)(C\sqrt{\log n_S} + C_p). \qquad \text{(B.5)}
\end{aligned}$$

Finally we use the fact

$$\left\| \begin{pmatrix} \hat{\boldsymbol{\alpha}}_S^{or} - \boldsymbol{\alpha}^* \\ \hat{\boldsymbol{\beta}}_S^{or} - \boldsymbol{\beta}^* \end{pmatrix} \right\|_\infty \leq \left\| \begin{pmatrix} \hat{\boldsymbol{\alpha}}_S^{or} - \hat{\boldsymbol{\alpha}}^{or} \\ \hat{\boldsymbol{\beta}}_S^{or} - \hat{\boldsymbol{\beta}}^{or} \end{pmatrix} \right\|_\infty + \left\| \begin{pmatrix} \hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^* \\ \hat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^* \end{pmatrix} \right\|_\infty$$

and by letting $C = \sqrt{\frac{2}{c_1}}$, which complete the proof. □

Appendix C. Proof of Theorem 3

The proof of theorem 3 is mainly based on theorem 1, theorem 2 and lemma 2 below. in Ma and Huang, 2017

Lemma 2. *If the conditions in theorem 2 and assumption 3 hold, with ϕ_n, ψ_n as defined in theorem 2 and $b_n > a\lambda$, $\lambda \gg \phi_n$, then there exists a local minimizer $(\hat{\boldsymbol{\mu}}_S^T, \hat{\boldsymbol{\beta}}_S^T)^T$ of the objective function $Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}; \lambda)$ given in (6) satisfying*

$$\mathbb{P}\big((\hat{\boldsymbol{\mu}}_S^T, \hat{\boldsymbol{\beta}}_S^T)^T = ((\hat{\boldsymbol{\mu}}_S^{or})^T, (\hat{\boldsymbol{\beta}}_S^{or})^T)^T\big) \to 1.$$

The proof of lemma 2 is the same as theorem 2 in Ma and Huang, 2017, the only different is we apply the theorem on the screened sample $\boldsymbol{X}_S, \boldsymbol{y}_S$. Now we give the proof of theorem 3 based on the results above.

Proof of Theorem 3. From lemma 2 we know that both $\mathbb{E}_{\boldsymbol{X}}(\hat{\mu}_{iS}-\mu_i^*)^2 = \mathbb{E}_{\boldsymbol{X}}(\hat{\mu}_{iS}^{or}-\mu_i^*)^2$ and $\mathbb{E}_{\boldsymbol{X}}(\hat{\beta}_{iS}-\beta_i^*)^2 = \mathbb{E}_{\boldsymbol{X}}(\hat{\beta}_{iS}^{or}-\beta_i^*)^2$ hold almost surely. Using the fact that $(\psi_n+\phi_n)^2 \leq 2(\phi_n^2+\psi_n^2)$, and combining with the results in (12) and (14) we can derive

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}}(\hat{\mu}_{iS}-\mu_i^*)^2 &= \mathbb{E}^2_{\boldsymbol{X}}(\hat{\mu}_{iS}^{or}-\mu_i^*) + \mathrm{Var}_{\boldsymbol{X}}(\hat{\mu}_{iS}^{or}) \\
&\leq \phi_n^2+\psi_n^2+\frac{\sigma^2}{n_S-1}\left(1+\frac{c_0}{\lambda_{\min}(\boldsymbol{\Sigma})}\right) \\
\mathbb{E}_{\boldsymbol{X}}(\hat{\beta}_{iS}-\beta_i^*)^2 &= \mathbb{E}^2_{\boldsymbol{X}}(\hat{\beta}_{iS}^{or}-\beta_i^*) + \mathrm{Var}_{\boldsymbol{X}}(\hat{\beta}_{iS}^{or}) \\
&\leq \phi_n^2+\psi_n^2+\frac{4p\sigma^2}{n_S\lambda_{min}(\boldsymbol{\Sigma})(x_{(n-r+1)j}-x_{(r)j})^2},
\end{aligned}$$

which hold for $i \in S$. □

Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

Nicholas G Davies, Petra Klepac, Yang Liu, Kiesha Prem, Mark Jit, and Rosalind M Eggo. Age-dependent effects in the transmission and control of covid-19 epidemics. *Nature medicine*, 26(8):1205–1211, 2020.

Brian S Everitt. Finite mixture distributions. *Wiley StatsRef: Statistics Reference Online*, 2014.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

Dongwan Kim, Joo-Yeon Lee, Jeong-Sun Yang, Jun Won Kim, V Narry Kim, and Hyeshik Chang. The architecture of sars-cov-2 transcriptome. *Cell*, 181 (4):914–921, 2020.

Jinchi Lv, Yingying Fan, et al. A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, 37 (6A):3498–3528, 2009.

Ping Ma and Xiaoxiao Sun. Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):70–76, 2015.

Ping Ma, Michael Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. pages 91–99, 2014.

Shujie Ma and Jian Huang. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423, 2017.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Lindsay M Reynolds, Jackson R Taylor, Jingzhong Ding, Kurt Lohman, Craig Johnson, David Siscovick, Gregory Burke, Wendy Post, Steven Shea, David R Jacobs Jr, et al. Age-related variations in the methylome associated with gene expression in human monocytes and t cells. *Nature communications*, 5(1):1–8, 2014.

Juan Shen and Xuming He. Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association*, 110(509):303–312, 2015.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

Marc A Suchard, Quanli Wang, Cliburn Chan, Jacob Frelinger, Andrew Cron, and Mike West. Understanding gpu programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of computational and graphical statistics*, 19(2):419–438, 2010.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522): 829–844, 2018.

HaiYing Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525):393–405, 2019.

Hansheng Wang, Runze Li, and Chih-Ling Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Zemin Zheng, Yingying Fan, and Jinchi Lv. High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 627–649, 2014.