A Concave Pairwise Fusion Approach to Subgroup Analysis

Shujie Ma & Jian Huang

Taylor & Francis
Taylor & Francis Group

A Concave Pairwise Fusion Approach to Subgroup Analysis

Shujie Ma[a] and Jian Huang[b]

[a]Department of Statistics, University of California Riverside, Riverside, CA; [b]Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA

ABSTRACT

An important step in developing individualized treatment strategies is correct identification of subgroups of a heterogeneous population to allow specific treatment for each subgroup. This article considers the problem using samples drawn from a population consisting of subgroups with different mean values, along with certain covariates. We propose a penalized approach for subgroup analysis based on a regression model, in which heterogeneity is driven by unobserved latent factors and thus can be represented by using subject-specific intercepts. We apply concave penalty functions to pairwise differences of the intercepts. This procedure automatically divides the observations into subgroups. To implement the proposed approach, we develop an alternating direction method of multipliers algorithm with concave penalties and demonstrate its convergence. We also establish the theoretical properties of our proposed estimator and determine the order requirement of the minimal difference of signals between groups to recover them. These results provide a sound basis for making statistical inference in subgroup analysis. Our proposed method is further illustrated by simulation studies and analysis of a Cleveland heart disease dataset. Supplementary materials for this article are available online.

1. Introduction

Precision medicine emphasizes the use of information available on individual patients to make treatment decisions. Developing individualized treatment strategies requires sophisticated analytic tools. One of the key statistical challenges is to correctly identify subgroups from a heterogeneous population, so that specific medical therapies can be prescribed for each subgroup. A popular method for analyzing data from a heterogeneous population is to view data as coming from a mixture of subgroups with their own sets of parameter values and then apply finite mixture model analysis (Everitt and Hand 1981). The mixture model approach has been widely used for data clustering and classification; see Banfield and Raftery (1993), Hastie and Tibshirani (1996), McNicholas (2010), and Wei and Kosorok (2013) for Gaussian mixture model approaches, Shen and He (2015) for a logistic-normal mixture model method, and Chaganty and Liang (2013) for a low-rank method for mixtures of linear regressions that provides a good initialization for the EM algorithm typically used in estimation of mixture models. The mixture model-based approach as a supervised clustering method needs to specify an underlying distribution for the data, and also requires specification of the number of mixture components in the population, often difficult to do in practice.

In this article, we propose a new approach to automatically detecting and identifying homogeneous subgroups based on a concave pairwise fusion penalty without knowledge of an a priori classification or a natural basis for separating a sample into subsets. Let y_i be the response variable for the ith subject. After adjusting for the effects of a set of covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\mathrm{T}$, we consider subgroup analysis for $\mathbf{y} = (y_1, \ldots, y_n)^\mathrm{T}$ with the heterogeneity driven by unknown or unobserved latent factors, which can be modeled through subject-specific intercepts in regression. Hence, we consider

$$y_i = \mu_i + \mathbf{x}_i^\mathrm{T}\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \qquad (1)$$

where μ_i's are unknown subject-specific intercepts, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathrm{T}$ is the vector of unknown coefficients for the covariates $\mathbf{x}_i$, and ϵ_i is the error term independent of $\mathbf{x}_i$ with $E(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$. In a biomedical study, y_i could be a certain phenotype associated with some disease such as the maximal heart rate, which is related to cardiac mortality or body mass index associated with obesity, and $\mathbf{x}_i$ is a set of observed covariates such as gender, age, race, etc. After adjusting for the effects of the covariates, the distribution of the response is still heterogeneous, as demonstrated for our heart disease application by multiple modes in the density plot shown in Figure 5. This heterogeneity can be the result of unobserved latent factors, so that it is modeled through the subject-specific intercepts μ_i's.

It is worth noting that if the factors contributing to this heterogeneity, for example, different treatments, become available, then μ_i can be written as $\mu_i = \mu + \mathbf{z}_i^\mathrm{T}\boldsymbol{\theta}$, where $\mathbf{z}_i$ are the observed variables for the treatments and $\boldsymbol{\theta}$ are the coefficients of $\mathbf{z}_i$. One interesting application in precision medicine is that the coefficients for $\mathbf{z}_i$ can be subject-specific, since the same treatment may have different effects on different patients. For this

case, we can consider the model with heterogeneous effects of some covariates given as

$$y_i = \mu + \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\theta}_i + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n. \quad (2)$$

Throughout this article, we focus on studying model (1) by considering that heterogeneity comes from unobserved latent factors. However, our proposed estimation method and the associated theoretical properties for model (1) can be extended to model (2) with some modifications. We provide the detailed estimation procedure for model (2) in Section A.5 of the online supplementary materials for interested readers. Assumptions of the structure are needed to estimate model (1). To this end, we assume that $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ are from K different groups with $K \geq 1$ and the data from the same group have the same intercept. In other words, let $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_K)$ be a partition of $\{1, \ldots, n\}$. We have $\mu_i = \alpha_k$ for all $i \in \mathcal{G}_k$, where α_k is the common value for the μ_i's from group $\mathcal{G}_k$. In practice, the number of groups K is unknown. However, it is usually reasonable to assume that K is much smaller than n. Our goal is to estimate K and identify the subgroups of outcomes. Moreover, accurately estimating the regression coefficient $\boldsymbol{\beta}$ is also crucial for subgroup analysis, since it will determine how well the effects of the covariates $\mathbf{x}_i$ are controlled for. As a consequence, it will affect the accuracy of subgroup analysis. We propose a concave pairwise fusion penalized least squares approach, and derive an alternating direction method of multipliers (ADMM, Boyd et al. 2011) algorithm for implementing the proposed approach. Our approach provides a useful tool for identifying subgroups of individuals along with controlling for effects of observed covariates. By dividing patients into homogeneous subgroups, treatment and follow-up can be tailored for each subgroup according to disease susceptibility and symptoms, the major goal of precision medicine.

Several authors have studied the problem of exploring homogeneous effects of covariates in the regression setting by assuming that the true coefficients are divided into a few clusters with common values. For instance, Tibshirani et al. (2005) proposed the fused lasso method that applies L_1 penalties to pairs of adjacent coordinates given that a complete ordering of covariates is available. Bondell and Reich (2008) proposed the OSCAR method where a special octagonal shrinkage penalty is applied to each pair of coordinates. Shen and Huang (2010) developed a group pursuit approach with truncated L_1 penalties to pairwise differences, and Ke, Fan, and Wu (2015) proposed a method called CARDS. All the above methods are about estimating homogeneous effects of covariates, which is different from our work aiming to identify subgroups among observations. Guo et al. (2010) proposed using a pairwise L_1 fusion penalty for identifying informative variables in the context of Gaussian model-based cluster analysis. In the unsupervised learning setting, a recent article (Chi and Lange 2015) considered the convex clustering problem and investigated the ADMM and the alternating minimization algorithms with the convex L_p ($p \geq 1$) penalties applied to pairwise differences of the data points.

The ADMM has good convergence properties for convex loss functions with the L_p, $p \geq 1$, penalties (Boyd et al. 2011; Chi and Lange 2015). Moreover, the L_1 penalty can shrink some pairwise differences of the parameter estimates to zero. However, the L_1

penalty generates large biases in the estimates in each iteration of the algorithm. As a result, it may not be able to identify subgroups, as illustrated in Figure 1. To address this issue, Chi and Lange (2015) proposed to multiply nonnegative weights to the L_1 norms to reduce the bias. However, the choice of weights can dramatically affect the quality of the clustering solution, and there is no clear rule for how to choose the weights. Thus, a penalty that can produce unbiased estimates is more desirable for identifying subgroups. We propose an ADMM algorithm using concave pairwise fusion penalties for estimation of model (1). The concave penalties in the optimization problem such as the smoothly clipped absolute deviations penalty (SCAD, Fan and Li 2001) and the minimax concave penalty (MCP, Zhang 2010) enjoy the unbiasedness property. We then derive the convergence properties of the ADMM algorithm. Moreover, we provide theoretical analysis of the proposed estimators. Specifically, we derive the order requirement of the minimum signal difference between groups to identify true subgroups. We also establish the oracle property such that under mild regularity conditions the oracle estimator is a local minimizer of the objective function with a high probability. The oracle estimator is obtained from least squares regression by assuming that the true group structure is known.

The rest of this article is organized as follows. In Section 2 we describe the proposed approach in detail. In Section 3 we derive an ADMM algorithm with concave penalties. We then state the theoretical properties of the proposed approach in Section 4. In Section 5 we evaluate the finite sample properties of the proposed procedures via simulation studies. Section 6 illustrates the proposed method using a data example. Some concluding remarks are given in Section 7. The estimation procedure for model (2) and all the technical proofs are provided in the online supplementary materials.

2. Subgroup Analysis Via Concave Pairwise Fusion

To estimate model (1), we propose a concave pairwise fusion penalized least squares approach. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$. The objective function is

$$Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}; \lambda) = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \mu_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} \right)^2$$
$$+ \sum_{1 \leq i < j \leq n} p(|\mu_i - \mu_j|, \lambda), \quad (3)$$

where $p(\cdot, \lambda)$ is a concave penalty function with a tuning parameter $\lambda \geq 0$.

For a given $\lambda > 0$, define

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

The penalty shrinks some of the pairs $\mu_j - \mu_k$ to zero. Based on this, we can partition the sample into subgroups. Specifically, let $\widehat{\lambda}$ be the value of the tuning parameter selected based on a data-driven procedure such as the Bayesian information criterion (BIC). For simplicity, write $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\beta}}) \equiv (\widehat{\boldsymbol{\mu}}(\widehat{\lambda}), \widehat{\boldsymbol{\beta}}(\widehat{\lambda}))$. Let $\{\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{\widehat{K}}\}$ be the distinct values of $\widehat{\boldsymbol{\mu}}$. Let $\widehat{\mathcal{G}}_k = \{i : \widehat{\mu}_i = \widehat{\alpha}_k, 1 \leq i \leq n\}, 1 \leq k \leq \widehat{K}$. Then $\{\widehat{\mathcal{G}}_1, \ldots, \widehat{\mathcal{G}}_{\widehat{K}}\}$ constitutes a partition of $\{1, \ldots, n\}$.
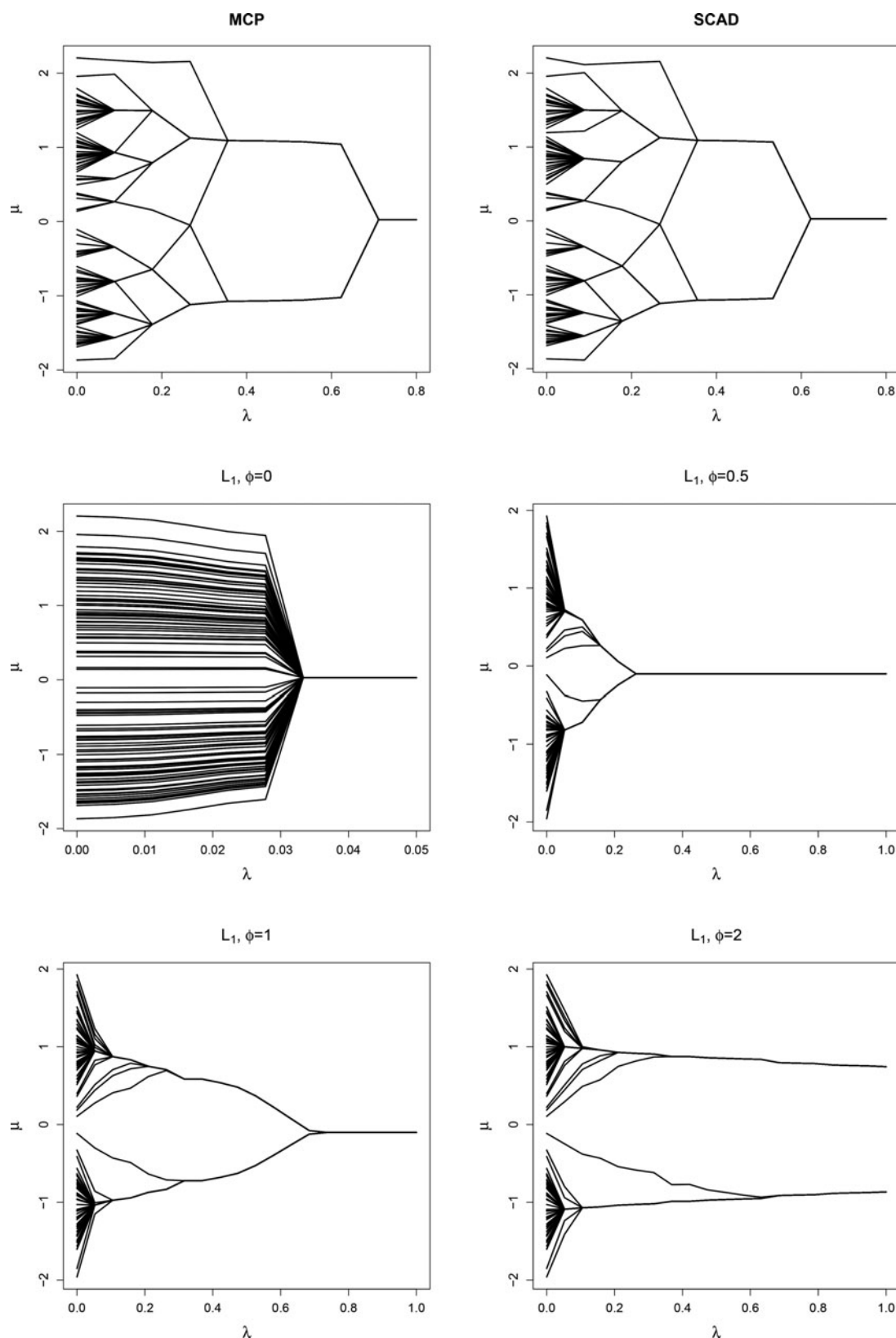
Figure 1. Solution paths for the estimated values of $(\mu_1, \ldots, \mu_n)$ against λ values by using MCP, SCAD, and various L_1 penalties, respectively, in Example 1.

An important question is which penalty function should be used. The L_1 penalty with $p_\gamma(t, \lambda) = \lambda t$ applies the same thresholding to all pairs $|\mu_i - \mu_j|$. As a result, it leads to biased estimates and may not be able to correctly recover subgroups. This is similar to the situation in variable selection where the Lasso tends to over-shrink large coefficients. In our numerical studies, we found that the L_1 penalty tends to yield either a large number of subgroups or no subgroup on the solution path. Hence, a penalty that can produce unbiased estimates is more appealing. This motivates us to use the concave penalties including SCAD

(Fan and Li 2001) and MCP (Zhang 2010). These penalties are asymptotically unbiased and are more aggressive in enforcing a sparser solution. Thus, they are better suited for the current problem, since the number of subgroups is usually much smaller than the sample size.

The MCP has the form

$$p_\gamma(t, \lambda) = \lambda \int_0^t (1 - x/(\gamma\lambda))_+ dx, \gamma > 1,$$

and the SCAD penalty is

$$p_\gamma(t, \lambda) = \lambda \int_0^t \min\{1, (\gamma - x/\lambda)_+/(\gamma - 1)\} dx, \gamma > 2,$$

where γ is a parameter that controls the concavity of the penalty function. In particular, both penalties converge to the L_1 penalty as $\gamma \to \infty$. Here and in the rest of the article, we put γ in the subscript to indicate the dependence of these penalty functions on it. Following Fan and Li (2001) and Zhang (2010), we treat γ as a fixed constant. These concave penalties enjoy sparsity as the L_1 penalty, that is, they can automatically yield zero estimates. More importantly, the concave penalties have the unbiasedness property in that they do not shrink large estimated parameters, so that they remain unbiased in the iterations. This property is particularly essential in the ADMM algorithms since the biases in the iterations may significantly affect the search for subgroups.

3. Computation

It is difficult to compute the estimates directly by minimizing the objective function (3) because the penalty function is not separable in μ_i's. We reparameterize the criterion by introducing a new set of parameters $\eta_{ij} = \mu_i - \mu_j$. Then, the minimization of (3) is equivalent to the constraint optimization problem

$$S(\mu, \beta, \eta) = \frac{1}{2} \sum_{i=1}^n (y_i - \mu_i - \mathbf{x}_i^\mathrm{T}\beta)^2 + \sum_{i<j} p_\gamma(|\eta_{ij}|, \lambda),$$

$$\text{subject to } \mu_i - \mu_j - \eta_{ij} = 0,$$

where $\eta = \{\eta_{ij}, i < j\}^\mathrm{T}$. By the augmented Lagrangian method, the estimates of the parameters can be obtained by minimizing

$$L(\mu, \beta, \eta, \upsilon) = S(\mu, \beta, \eta) + \sum_{i<j} \upsilon_{ij}(\mu_i - \mu_j - \eta_{ij})$$
$$+ \frac{\vartheta}{2} \sum_{i<j} (\mu_i - \mu_j - \eta_{ij})^2, \quad (4)$$

where the dual variables $\upsilon = \{\upsilon_{ij}, i < j\}^\mathrm{T}$ are Lagrange multipliers and ϑ is the penalty parameter. We compute the estimates of $(\mu, \beta, \eta, \upsilon)$ through iterations by the ADMM.

It is noteworthy that by using the concave penalties, although the objective function $L(\mu, \beta, \eta, \upsilon)$ is not a convex function, it is convex with respect to each η_{ij} when $\gamma > 1/\vartheta$ for the MCP penalty and $\gamma > 1/\vartheta + 1$ for the SCAD penalty. Moreover, for given $(\mu, \beta, \eta, \upsilon)$, the minimizer of $L(\mu, \beta, \eta, \upsilon)$ with respect to η_{ij} is unique and has a closed-form expression for the L_1, MCP, and SCAD penalties, respectively. Specifically, for given $(\mu, \beta, \eta, \upsilon)$, the minimization problem is the same as

minimizing

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

with respect to η_{ij}, where $\delta_{ij} = \mu_i - \mu_j + \vartheta^{-1}\upsilon_{ij}$. Hence, the closed-form solution for the L_1 penalty is

$$\widehat{\eta}_{ij} = \mathrm{ST}(\delta_{ij}, \lambda/\vartheta), \quad (5)$$

where $\mathrm{ST}(t, \lambda) = \mathrm{sign}(t)(|t| - \lambda)_+$ is the soft thresholding rule, and $(x)_+ = x$ if $x > 0$, and $(x)_+ = 0$ otherwise. For the MCP penalty with $\gamma > 1/\vartheta$, it is

$$\widehat{\eta}_{ij} = \begin{cases} \frac{\mathrm{ST}(\delta_{ij}, \lambda/\vartheta)}{1 - 1/(\gamma\vartheta)} & \text{if } |\delta_{ij}| \leq \gamma\lambda \\ \delta_{ij} & \text{if } |\delta_{ij}| > \gamma\lambda. \end{cases} \quad (6)$$

For the SCAD penalty with $\gamma > 1/\vartheta + 1$, it is

$$\widehat{\eta}_{ij} = \begin{cases} \mathrm{ST}(\delta_{ij}, \lambda/\vartheta) & \text{if } |\delta_{ij}| \leq \lambda + \lambda/\vartheta \\ \frac{\mathrm{ST}(\delta_{ij}, \gamma\lambda/((\gamma-1)\vartheta))}{1 - 1/((\gamma-1)\vartheta)} & \text{if } \lambda + \lambda/\vartheta < |\delta_{ij}| \leq \gamma\lambda \\ \delta_{ij} & \text{if } |\delta_{ij}| > \gamma\lambda. \end{cases} \quad (7)$$

3.1. Algorithm

We now describe the computational algorithm based on the ADMM for minimizing the objective function (4). It consists of steps for iteratively updating μ, β, η, and υ. Denote the L_2 norm of any vector $\mathbf{a}$ by $||\mathbf{a}||$. The main ingredients of the algorithm are as follows.

First, for a given (η, υ), to obtain an update of μ and β, we set the derivatives $\partial L(\mu, \beta, \eta, \upsilon)/\partial\mu$ and $\partial L(\mu, \beta, \eta, \upsilon)/\partial\beta$ to zero, where

$$L(\mu, \beta, \eta, \upsilon) = \frac{1}{2} \sum_{i=1}^n (y_i - \mu_i - \mathbf{x}_i^\mathrm{T}\beta)^2$$
$$+ \frac{\vartheta}{2} \sum_{i<j} \{(e_i - e_j)^\mathrm{T}\mu - \eta_{ij} + \vartheta^{-1}\upsilon_{ij}\}^2 + C$$
$$= \frac{1}{2} \|\mu - \mathbf{y} + \mathbf{X}\beta\|^2 + \frac{\vartheta}{2} \|\Delta\mu - \eta + \vartheta^{-1}\upsilon\|^2 + C.$$

Here C is a constant independent of μ and β, $\mathbf{y} = (y_1, \ldots, y_n)^\mathrm{T}$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\mathrm{T}$, e_i is the $n \times 1$ vector whose ith element is 1 and the remaining ones are 0, and $\Delta = \{(e_i - e_j), i < j\}^\mathrm{T}$. Thus, for given $\eta^{(m)}$ and $\upsilon^{(m)}$ at the mth step, the updates $\mu^{(m+1)}$ and $\beta^{(m+1)}$, which are the minimizers of $L(\mu, \beta, \eta^{(m)}, \upsilon^{(m)})$, are

$$\mu^{(m+1)} = (\vartheta\Delta^\mathrm{T}\Delta + \mathbf{I}_n - \mathbf{Q_X})^{-1}\{(\mathbf{I}_n - \mathbf{Q_X})\mathbf{y} + \vartheta\Delta^\mathrm{T}(\eta^{(m)} - \vartheta^{-1}\upsilon^{(m)})\},$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\mathbf{Q_X} = \mathbf{X}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}$, and

$$\beta^{(m+1)} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}(\mathbf{y} - \mu^{(m+1)}).$$

Second, the update of η_{ij} at the $(m+1)$th iteration is obtained by the formula given in (5), (6), and (7), respectively, by the Lasso, MCP, and SCAD penalties with δ_{ij} replaced by $\delta_{ij}^{(m+1)} = \mu_i^{(m+1)} - \mu_j^{(m+1)} + \vartheta^{-1}\upsilon_{ij}^{(m)}$.

Finally, the estimate of υ_{ij} is updated as

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

Based on the above discussion, the algorithm consists of the following steps:

Step 1. Find initial estimates $\boldsymbol{\beta}^{(0)}$ from least squares regression by letting $\mu_i = \mu$ for all i. Let the initial estimates $\boldsymbol{\mu}^{(0)} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(0)}$, $\eta_{ij}^{(0)} = \mu_i^{(0)} - \mu_j^{(0)}$, and $\boldsymbol{v}^{(0)} = \mathbf{0}$.

Step 2. At iteration $m + 1$, compute $(\boldsymbol{\mu}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\eta}^{(m+1)}, \boldsymbol{v}^{(m+1)})$ by the methods described above.

Step 3. Terminate the algorithm if the stopping rule is met at step $m + 1$. Then $(\boldsymbol{\mu}^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\eta}^{(m+1)}, \boldsymbol{v}^{(m+1)})$ are our final estimates $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{v}})$. Otherwise, we go to Step 2.

Remark 1. In the computation, the covariates are all centered and standardized. It can be derived that $\boldsymbol{\Delta}^T \boldsymbol{\Delta} = n\mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n^T$, where $\mathbf{1}_n$ is the $n \times 1$ vector with ones. In the algorithm, we require $\vartheta \boldsymbol{\Delta}^T \boldsymbol{\Delta} + \mathbf{I}_n - \mathbf{Q}_X$ to be invertible. For any vector $\mathbf{a} \in R^n$ with $||\mathbf{a}|| = 1$, we have $\mathbf{a}^T(\vartheta \boldsymbol{\Delta}^T \boldsymbol{\Delta})\mathbf{a} \geq 0$ and $\mathbf{a}^T(\mathbf{I} - \mathbf{Q}_X)\mathbf{a} \geq 0$. Note that $\mathbf{a}^T(\vartheta \boldsymbol{\Delta}^T \boldsymbol{\Delta})\mathbf{a} = 0$ if and only if $\mathbf{a} = \mathbf{1}_n/\sqrt{n}$. Moreover, in the numerical analysis we let $\sum_{i=1}^n x_{ij}/n = 0$, $1 \leq j \leq p$. As a result, we have $(\mathbf{1}_n/\sqrt{n})^T \mathbf{Q}_X(\mathbf{1}_n/\sqrt{n}) = 0$, and thus $(\mathbf{1}_n/\sqrt{n})^T (\mathbf{I}_n - \mathbf{Q}_X)(\mathbf{1}_n/\sqrt{n}) = 1$. Hence, we have $\mathbf{a}^T(\vartheta \boldsymbol{\Delta}^T \boldsymbol{\Delta} + \mathbf{I}_n - \mathbf{Q}_X)\mathbf{a} > 0$ for any vector $\mathbf{a} \in R^n$ with $||\mathbf{a}|| = 1$. Therefore, $\vartheta \boldsymbol{\Delta}^T \boldsymbol{\Delta} + \mathbf{I}_n - \mathbf{Q}_X$ is invertible.

Remark 2. We track the progress of the ADMM based on the primal residual $\mathbf{r}^{(m+1)} = \boldsymbol{\Delta}\boldsymbol{\mu}^{(m+1)} - \boldsymbol{\eta}^{(m+1)}$. We stop the algorithm when $\mathbf{r}^{(m+1)}$ is close to zero such that $||\mathbf{r}^{(m+1)}|| < \epsilon$ for some small value $\epsilon > 0$.

Remark 3. The shrinkage step in this algorithm enables us to have $\widehat{\eta}_{ij} = 0$ for some values of λ. We put y_i and y_j in the same group if $\widehat{\eta}_{ij} = 0$. As a result, we have $\widehat{K}$ estimated groups $\widehat{\mathcal{G}}_1, \ldots, \widehat{\mathcal{G}}_{\widehat{K}}$ and let the estimated intercept for the kth group be $\widehat{\alpha}_k = |\widehat{\mathcal{G}}_k|^{-1} \sum_{i \in \widehat{\mathcal{G}}_k} \widehat{\mu}_i$, where $|\widehat{\mathcal{G}}_k|$ is the cardinality of $\widehat{\mathcal{G}}_k$.

3.2. Convergence of the Algorithm

We next consider the convergence properties of the ADMM algorithm.

Proposition 1. The primal residual $\mathbf{r}^{(m)} = \boldsymbol{\Delta}\boldsymbol{\mu}^{(m)} - \boldsymbol{\eta}^{(m)}$ and the dual residual $\mathbf{s}^{(m+1)} = \vartheta \boldsymbol{\Delta}^T(\boldsymbol{\eta}^{(m+1)} - \boldsymbol{\eta}^{(m)})$ of the ADMM satisfy that $\lim_{m\to\infty} ||\mathbf{r}^{(m)}||^2 = 0$ and $\lim_{m\to\infty} ||\mathbf{s}^{(m)}||^2 = 0$ for both MCP and SCAD penalties.

The proof of this result is given in the online supplementary materials. Proposition 1 shows that the primal feasibility and dual feasibility are achieved by the algorithm. Therefore, it converges to an optimal point. This optimal point may be a local minimum of the objective function when a concave penalty function is applied.

4. Theoretical Properties

4.1. Heterogeneous Model

In this section, we study the theoretical properties of the proposed estimator under the heterogeneous model in which there are at least two subgroups. We derive the order requirement of the minimum signal difference between groups to recover the true groups and the oracle property that under some regularity

conditions the oracle estimator is a local minimizer of the objective function with a high probability. Let $\mathcal{M}_{\mathcal{G}}$ be the subspace of R^n, defined as

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

For each $\boldsymbol{\mu} \in \mathcal{M}_{\mathcal{G}}$, it can be written as $\boldsymbol{\mu} = \mathbf{Z}\boldsymbol{\alpha}$, where $\mathbf{Z} = \{z_{ik}\}$ is the $n \times K$ matrix with $z_{ik} = 1$ for $i \in \mathcal{G}_k$ and $z_{ik} = 0$ otherwise, and $\boldsymbol{\alpha}$ is a $K \times 1$ vector of parameters. By matrix calculation, we have $\mathbf{D} = \mathbf{Z}^T\mathbf{Z} = \text{diag}(|\mathcal{G}_1|, \ldots, |\mathcal{G}_K|)$, where $|\mathcal{G}_k|$ denotes the number of elements in $\mathcal{G}_k$. Define $|\mathcal{G}_{\min}| = \min_{1 \leq k \leq K} |\mathcal{G}_k|$ and $|\mathcal{G}_{\max}| = \max_{1 \leq k \leq K} |\mathcal{G}_k|$. Let $\mathbf{X} = (\mathbf{X}_1, \ldots \mathbf{X}_p)$, where $\mathbf{X}_j$ is the jth column of $\mathbf{X}$. Denote

$$\rho(t) = \lambda^{-1} p_\gamma(t, \lambda) \text{ and } \overline{\rho}(t) = \rho'(|t|)\text{sgn}(t).$$

For any vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_s)^T \in R^s$, denote $||\boldsymbol{\zeta}||_\infty = \max_{1 \leq l \leq s} |\zeta_l|$. For any symmetric matrix $\mathbf{A}_{s \times s}$, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ be the smallest and largest eigenvalues of $\mathbf{A}$, respectively. For any matrix $\mathbf{A} = (A_{ij})_{i=1, j=1}^{s,t}$, denote its L_2 norm as $||\mathbf{A}|| = \max_{\boldsymbol{\zeta} \in R^t, ||\boldsymbol{\zeta}||=1} ||\mathbf{A}\boldsymbol{\zeta}||$ and denote $||\mathbf{A}||_\infty = \max_{1 \leq i \leq s} \sum_{j=1}^t |A_{ij}|$. For any positive numbers a_n and b_n, let $a_n \asymp b_n$ denote $\lim_{n\to\infty} a_n/b_n = c$, for a positive constant c, and $a_n \gg b_n$ denote $a_n^{-1}b_n = o(1)$. We introduce the following conditions.

(C1) Assume $||\mathbf{X}_j|| = \sqrt{n}$, for $1 \leq j \leq p$, $\lambda_{\min}[(\mathbf{Z}, \mathbf{X})^T(\mathbf{Z}, \mathbf{X})] \geq C_1 |\mathcal{G}_{\min}|$, and $||\mathbf{X}||_\infty \leq C_2 p$ for some constants $0 < C_1 \leq 1$ and $0 < C_2 < \infty$.

(C2) $p_\gamma(t, \lambda)$ is a symmetric function of t, and it is nondecreasing and concave in t for $t \in [0, \infty)$. There exists a constant $0 < a < \infty$ such that $\rho(t)$ is a constant for all $t \geq a\lambda$, and $\rho(0) = 0$. $\rho'(t)$ exists and is continuous except for a finite number of t and $\rho'(0+) = 1$.

(C3) The noise vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ has sub-Gaussian tails such that $P(|\mathbf{a}^T\boldsymbol{\epsilon}| > ||\mathbf{a}||x) \leq 2\exp(-c_1 x^2)$ for any vector $\mathbf{a} \in R^n$ and $x > 0$, where $0 < c_1 < \infty$.

Conditions (C2) and (C3) are commonly assumed in high-dimensional settings. The concave penalties such as MCP and SCAD satisfy (C2). In the literature, it is commonly assumed that the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}$ is bounded by Cn for some constant $0 < C < \infty$, which may not hold for $(\mathbf{Z}, \mathbf{X})^T(\mathbf{Z}, \mathbf{X})$. Note that $\lambda_{\min}(\mathbf{Z}^T\mathbf{Z}) = |\mathcal{G}_{\min}|$. By assuming $\lambda_{\min}(\mathbf{X}^T\mathbf{X}) = Cn$, we have $\lambda_{\min}[(\mathbf{Z}, \mathbf{X})^T(\mathbf{Z}, \mathbf{X})] \leq \min(|\mathcal{G}_{\min}|, Cn)$. The equality holds when $\mathbf{Z}^T\mathbf{X} = \mathbf{0}$. Hence, we assume $\lambda_{\min}[(\mathbf{Z}, \mathbf{X})^T(\mathbf{Z}, \mathbf{X})] \geq C_1 |\mathcal{G}_{\min}|$ for some constant $0 < C_1 \leq 1$.

When the true group memberships $\mathcal{G}_1, \ldots, \mathcal{G}_K$ are known, the oracle estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ are

$$(\widehat{\boldsymbol{\mu}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) = \arg \min_{\boldsymbol{\mu} \in \mathcal{M}_{\mathcal{G}}, \boldsymbol{\beta} \in R^p} \frac{1}{2}||\mathbf{y} - \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}||^2,$$

and correspondingly, the oracle estimators for the common intercepts $\boldsymbol{\alpha}$ and the coefficients $\boldsymbol{\beta}$ are given by

$$(\widehat{\boldsymbol{\alpha}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) = \arg \min_{\boldsymbol{\alpha} \in R^K, \boldsymbol{\beta} \in R^p} \frac{1}{2}||\mathbf{y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}||^2$$
$$= [(\mathbf{Z}, \mathbf{X})^T(\mathbf{Z}, \mathbf{X})]^{-1}(\mathbf{Z}, \mathbf{X})^T\mathbf{y}.$$

Let $\boldsymbol{\alpha}^0 = (\alpha_k^0, k = 1, \ldots, K)^T$, where α_k^0 is the underlying common intercept for group $\mathcal{G}_k$. Let $\boldsymbol{\beta}^0$ be the underlying regression coefficient.

Theorem 1. Suppose Conditions (C1)–(C3) hold. If $K = o(n)$, $p = o(n)$, and

$$|\mathcal{G}_{\min}| \gg \sqrt{(K+p)n \log n},$$

we have that with probability at least $1 - 2(K+p)n^{-1}$,

$$\left\| ((\widehat{\boldsymbol{\mu}}^{or} - \boldsymbol{\mu}^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^{\mathrm{T}})^{\mathrm{T}} \right\|_\infty \leq \phi_n, \qquad (8)$$

where

$$\phi_n = c_1^{-1/2} C_1^{-1} \sqrt{K+p}\, |\mathcal{G}_{\min}|^{-1} \sqrt{n \log n}, \qquad (9)$$

in which C_1 and c_1 are given in Conditions (C1) and (C3), respectively. Moreover, for any vector $\mathbf{a}_n \in R^{K+p}$, we have as $n \to \infty$,

$$\sigma_n^{-1}(\mathbf{a}_n) \mathbf{a}_n^{\mathrm{T}} ((\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^{\mathrm{T}})^{\mathrm{T}} \to N(0, 1), \quad (10)$$

where

$$\sigma_n(\mathbf{a}_n) = \sigma [\mathbf{a}_n^{\mathrm{T}} \{(\mathbf{Z}, \mathbf{X})^{\mathrm{T}} (\mathbf{Z}, \mathbf{X})\}^{-1} \mathbf{a}_n]^{1/2}. \qquad (11)$$

The proof of this theorem is given in the online supplementary materials.

Remark 4. Since $|\mathcal{G}_{\min}| \leq n/K$, by the condition $|\mathcal{G}_{\min}| \gg \sqrt{(K+p)n \log n}$, K and p must satisfy $K\sqrt{(K+p)} = o(\sqrt{n}(\log n)^{-1})$. Hence, $K = o(n^{1/3}(\log n)^{-1/3})$. By letting $|\mathcal{G}_{\min}| = \delta n/K$ for some $\delta \in (0, 1]$, the bound in (8) is $c_1^{-1/2} C_1^{-1} \delta^{-1} K \sqrt{K+p} \sqrt{\log n/n}$. Moreover, when K and p are fixed numbers, the bound in (8) is $C^* \sqrt{\log n/n}$ for some constant $0 < C^* < \infty$.

For $K \geq 2$, let

$$b_n = \min_{i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}, k \neq k'} |\mu_i^0 - \mu_j^0| = \min_{k \neq k'} |\alpha_k^0 - \alpha_{k'}^0|$$

be the minimal difference of the common values between two groups.

Theorem 2. Suppose the conditions in Theorem 1 hold and $K \geq 2$. If $b_n > a\lambda$ and $\lambda \gg \phi_n$, where a is given in Condition (C2) and ϕ_n is given in (9), then there exists a local minimizer $(\widehat{\boldsymbol{\mu}}(\lambda)^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}(\lambda)^{\mathrm{T}})^{\mathrm{T}}$ of the objective function $Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}; \lambda)$ given in (3) satisfying

$$P\left((\widehat{\boldsymbol{\mu}}(\lambda)^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}(\lambda)^{\mathrm{T}})^{\mathrm{T}} = ((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}} \right) \to 1.$$

The proof of this theorem is given in the online supplementary materials.

Remark 5. The conditions $b_n > a\lambda$ and $\lambda \gg \phi_n$ imply that $b_n \gg \phi_n$. As discussed in Remark 5, when K is a finite and fixed number and $|\mathcal{G}_{\min}| = \delta n/K$ for some constant $0 < \delta \leq 1$, then we require $b_n \gg C^* \sqrt{\log n/n}$ for some constant $0 < C^* < \infty$. Moreover, Theorem 2 shows that the oracle estimator $((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ is a local minimizer of the objective function with probability approaching 1. Let $\widehat{\boldsymbol{\alpha}}(\lambda)$ be the distinct values of $\widehat{\boldsymbol{\mu}}(\lambda)$. Also $\widehat{\boldsymbol{\alpha}}^{or}$ consists of the distinct values of $\widehat{\boldsymbol{\mu}}^{or}$. By the oracle property in Theorem 2, we have $P\{\widehat{\boldsymbol{\alpha}}(\lambda) = \widehat{\boldsymbol{\alpha}}^{or}\} \to 1$. This result together with the asymptotic normality given in Theorem 1 directly leads to the asymptotic distribution of $(\widehat{\boldsymbol{\alpha}}(\lambda)^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}(\lambda)^{\mathrm{T}})^{\mathrm{T}}$ presented in the following corollary.

Corollary 1. Under the conditions in Theorem 2, we have for any vector $\mathbf{a}_n \in R^{K+p}$, as $n \to \infty$,

$$\sigma_n^{-1}(\mathbf{a}_n) \mathbf{a}_n^{\mathrm{T}} ((\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^0)^{\mathrm{T}})^{\mathrm{T}} \to N(0, 1),$$

with $\sigma_n(\mathbf{a}_n)$ given in (11). As a result, we have for any vectors $\mathbf{a}_{n1} \in R^K$ and $\mathbf{a}_{n2} \in R^p$, as $n \to \infty$, $\sigma_{n1}^{-1}(\mathbf{a}_{n1}) \mathbf{a}_{n1}^{\mathrm{T}} (\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}^0) \to N(0, 1)$ and $\sigma_{n2}^{-1}(\mathbf{a}_{n2}) \mathbf{a}_{n2}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^0) \to N(0, 1)$, where

$$\sigma_{n1}(\mathbf{a}_{n1}) = \sigma \left[\mathbf{a}_{n1}^{\mathrm{T}} \{ \mathbf{Z}^{\mathrm{T}} \mathbf{Z} - (\mathbf{Z}^{\mathrm{T}} \mathbf{X})(\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} (\mathbf{X}^{\mathrm{T}} \mathbf{Z}) \}^{-1} \mathbf{a}_{n1} \right]^{1/2},$$

$$\sigma_{n2}(\mathbf{a}_{n2}) = \sigma \left[\mathbf{a}_{n2}^{\mathrm{T}} \{ \mathbf{X}^{\mathrm{T}} \mathbf{X} - (\mathbf{X}^{\mathrm{T}} \mathbf{Z})(\mathbf{Z}^{\mathrm{T}} \mathbf{Z})^{-1} (\mathbf{Z}^{\mathrm{T}} \mathbf{X}) \}^{-1} \mathbf{a}_{n2} \right]^{1/2}.$$

Remark 6. By Corollary 1, for given $\mathbf{a}_{n1} \in R^K$ and $\mathbf{a}_{n2} \in R^p$, $100(1 - \alpha)\%$ confidence intervals for $\mathbf{a}_{n1}^{\mathrm{T}} \boldsymbol{\alpha}^0$ and $\mathbf{a}_{n2}^{\mathrm{T}} \boldsymbol{\beta}^0$ are given as $\mathbf{a}_{n1}^{\mathrm{T}} \widehat{\boldsymbol{\alpha}}(\lambda) \pm z_{\alpha/2} \widehat{\sigma}_{n1}(\mathbf{a}_{n1})$ and $\mathbf{a}_{n2}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}(\lambda) \pm z_{\alpha/2} \widehat{\sigma}_{n2}(\mathbf{a}_{n2})$, respectively, where $z_{\alpha/2}$ is the $(1 - \alpha/2)100$ percentile of the standard normal, and $\widehat{\sigma}_{n1}(\mathbf{a}_{n1})$ and $\widehat{\sigma}_{n2}(\mathbf{a}_{n2})$ are estimates of $\sigma_{n1}(\mathbf{a}_{n1})$ and $\sigma_{n2}(\mathbf{a}_{n2})$ with σ^2 estimated by $\widehat{\sigma}^2 = (n - \widehat{K} - p)^{-1} \sum_{i=1}^n (y_i - \widehat{\mu}_i - \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}})^2$, where $\widehat{K}$ is the number of distinct values in $\widehat{\boldsymbol{\mu}}(\lambda)$. However, although Corollary 1 provides support for using the asymptotic normality in conducting statistical inference, it is only proved for the fixed λ values satisfying the conditions in Theorem 2 and requires that the group structure be recovered with probability approaching 1.

Remark 7. It is worth noting that for a given λ, the estimators $\widehat{\boldsymbol{\alpha}}(\lambda)$ and $\widehat{\boldsymbol{\beta}}(\lambda)$ can have different convergence rates. To see this, we assume $\mathbf{Z}^{\mathrm{T}} \mathbf{X} = \mathbf{0}$ for convenience of mathematical derivations. Then, by the results in Corollary 1, for any vectors $\mathbf{a}_{n1} \in R^K$ and $\mathbf{a}_{n2} \in R^p$ satisfying $||\mathbf{a}_{n1}|| = 1$ and $||\mathbf{a}_{n2}|| = 1$, we have that $|\mathbf{a}_{n1}^{\mathrm{T}} (\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}^0)| = O_p[\{\lambda_{\max}((\mathbf{Z}^{\mathrm{T}} \mathbf{Z})^{-1})\}^{1/2}] = O_p(|\mathcal{G}_{\min}|^{-1/2})$ and $|\mathbf{a}_{n2}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^0)| = O_p[\{\lambda_{\min}(\mathbf{X}^{\mathrm{T}} \mathbf{X})\}^{-1/2}] = O_p(n^{-1/2})$ by assuming that $\lambda_{\min}(\mathbf{X}^{\mathrm{T}} \mathbf{X}) = Cn$. Hence, $\widehat{\alpha}_k(\lambda) - \alpha_k^0 = O_p(|\mathcal{G}_{\min}|^{-1/2})$ for every $1 \leq k \leq K$, where $\widehat{\alpha}_k(\lambda)$ and α_k^0 are the kth component of $\widehat{\boldsymbol{\alpha}}(\lambda)$ and $\boldsymbol{\alpha}^0$, respectively, and $\widehat{\beta}_j(\lambda) - \beta_j^0 = O_p(n^{-1/2})$ for every $1 \leq j \leq p$, where $\widehat{\beta}_j(\lambda)$ and β_j^0 are the jth component of $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\boldsymbol{\beta}^0$, respectively. Since $|\mathcal{G}_{\min}| \leq n/K$ and K is allowed to diverge with n, the rate of convergence of $\widehat{\boldsymbol{\alpha}}$ can be slower than that of $\widehat{\boldsymbol{\beta}}$.

4.2. Homogeneous Model

When the true model is the homogeneous model given as $y_i = \mu + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i$, $i = 1, \ldots, n$, we have $\mu_1 = \cdots = \mu_n = \mu = \alpha$ and $K = 1$. The penalized estimator $(\widehat{\boldsymbol{\mu}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$ of $(\boldsymbol{\mu}, \boldsymbol{\beta})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, also has the oracle property given as follows. We define the oracle estimator for $(\alpha, \boldsymbol{\beta})$ as

$$(\widehat{\alpha}^{or}, \widehat{\boldsymbol{\beta}}^{or}) = \arg \min_{\alpha \in R, \boldsymbol{\beta} \in R^p} \frac{1}{2} ||\mathbf{y} - \alpha \mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}||^2$$

$$= \left[(\mathbf{1}_n, \mathbf{X})^{\mathrm{T}} (\mathbf{1}_n, \mathbf{X}) \right]^{-1} (\mathbf{1}_n, \mathbf{X})^{\mathrm{T}} \mathbf{y}.$$

Let $\widehat{\boldsymbol{\mu}}^{or} = \widehat{\alpha}^{or} \mathbf{1}_n$. Let α^0 be the underlying common intercept, $\boldsymbol{\mu}^0 = \alpha^0 \mathbf{1}_n$, and $\boldsymbol{\beta}^0$ be the underlying regression coefficient. We introduce the following condition.

(C1)* Assume $\|\mathbf{X}_j\| = \sqrt{n}$, for $1 \leq j \leq p$, $\lambda_{\min}[(\mathbf{1}_n, \mathbf{X})^{\mathrm{T}} (\mathbf{1}_n, \mathbf{X})] \geq C_1 n$, and $||\mathbf{X}||_\infty \leq C_2 p$ for some constants $0 < C_1 < \infty$ and $0 < C_2 < \infty$.

Theorem 3. Suppose Conditions (C1*), (C2), and (C3) hold. If $p = o(n(\log n)^{-1})$, the oracle estimator has the property that with probability at least $1 - 2(1 + p)n^{-1}$,

$$\left\| ((\widehat{\boldsymbol{\mu}}^{or} - \boldsymbol{\mu}^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^{\mathrm{T}})^{\mathrm{T}} \right\|_\infty \leq \phi_n, \quad (12)$$

where $\phi_n = c_1^{-1/2} C_1^{-1} \sqrt{1 + p} \sqrt{n^{-1} \log n}$, in which C_1 and c_1 are given in Conditions (C1*) and (C3), respectively, and for any vector $\mathbf{a}_n \in R^{1+p}$, as $n \to \infty$,

$$\sigma_n^{-1}(\mathbf{a}_n) \mathbf{a}_n^{\mathrm{T}} ((\widehat{\alpha}^{or} - \alpha^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^{\mathrm{T}})^{\mathrm{T}} \to N(0, 1), \quad (13)$$

where

$$\sigma_n(\mathbf{a}_n) = \sigma \left[\mathbf{a}_n^{\mathrm{T}} \{ (\mathbf{1}_n, \mathbf{X})^{\mathrm{T}} (\mathbf{1}_n, \mathbf{X}) \}^{-1} \mathbf{a}_n \right]^{1/2}.$$

Moreover, if $\lambda \gg \phi_n$, then there exists a local minimizer $(\widehat{\boldsymbol{\mu}}(\lambda)^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}(\lambda)^{\mathrm{T}})^{\mathrm{T}}$ of the objective function $Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}; \lambda)$ given in (3) satisfying

$$P \left((\widehat{\boldsymbol{\mu}}(\lambda)^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}(\lambda)^{\mathrm{T}})^{\mathrm{T}} = ((\widehat{\boldsymbol{\mu}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}} \right) \to 1. \quad (14)$$

Remark 8. By Theorem 3, the local minimizer $\widehat{\boldsymbol{\mu}}(\lambda)$ can be written as $\widehat{\boldsymbol{\mu}}(\lambda) = \widehat{\alpha}(\lambda) \mathbf{1}_n$. Then, we have as $n \to \infty$, $\sigma_{n1}^{-1}(\widehat{\alpha}(\lambda) - \alpha^0) \to N(0, 1)$, and for any vector $\mathbf{a}_{n2} \in R^p$, $\sigma_{n2}^{-1}(\mathbf{a}_{n2}) \mathbf{a}_{n2}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^0) \to N(0, 1)$, where

$$\sigma_{n1} = \sigma \left\{ n - (\mathbf{1}_n^{\mathrm{T}} \mathbf{X}) (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} (\mathbf{X}^{\mathrm{T}} \mathbf{1}_n) \right\}^{-1/2},$$

$$\sigma_{n2}(\mathbf{a}_{n2}) = \sigma \left[\mathbf{a}_{n2}^{\mathrm{T}} \{ \mathbf{X}^{\mathrm{T}} \mathbf{X} - n^{-1} (\mathbf{X}^{\mathrm{T}} \mathbf{1}_n) (\mathbf{1}_n^{\mathrm{T}} \mathbf{X}) \}^{-1} \mathbf{a}_{n2} \right]^{1/2}.$$

5. Simulation Studies

In this section, we conduct simulation experiments to investigate the numerical performance of our proposed estimators.

We use the modified BIC (Wang, Li, and Tsai 2007) for high-dimensional data settings to select the tuning parameter by minimizing

$$\mathrm{BIC}(\lambda) = \log \left[\sum_{i=1}^n (y_i - \widehat{\mu}_i(\lambda) - \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}(\lambda))^2 / n \right] + C_n \frac{\log n}{n} (\widehat{K}(\lambda) + p) \quad (15)$$

with respect to λ, where C_n is a positive number that can depend on n. When $C_n = 1$, the modified BIC reduces to the traditional BIC (Schwarz 1978). Wang, Li, and Leng (2009) used $C_n = \log(\log(d))$ in their simulation study when the number of predictors d diverges with sample size. In this article, we adopt the same strategy and let $C_n = c \log(\log(d))$, where $d = n + p$ and c is a positive constant. In our analysis, we select λ by minimizing the modified BIC and use a fixed value for ϑ and γ.

Example 1. We simulate data from the model

$$y_i = \mu_i + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \quad (16)$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{i5})^{\mathrm{T}}$ are generated from the multivariate normal distribution with mean 0, variance 1, and an exchangeable correlation $\rho = 0.3$, and the error terms ϵ_i are from independent $N(0, 0.5^2)$. We simulate $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_5)^{\mathrm{T}}$ from independent Uniform[0.5, 1]. We generate μ_i from two different values $-\alpha$ and α with equal probabilities, that is, we generate

them from the distribution: $p(\mu_i = -\alpha) = p(\mu_i = \alpha) = 1/2$, so that there are two intercepts $\alpha_1 = -\alpha$ and $\alpha_2 = \alpha$. In our simulation studies, we take different values of α to illustrate our proposed method. It is noteworthy that for smaller value of α, it is more difficult to identify the two groups.

In our analysis, we choose to fix $\vartheta = 1$ and $\gamma = 3$. We compare the performance of the estimators with the ADMM algorithm by using the two concave penalties (MCP and SCAD) and using a weighted L_1 penalty

$$p_\gamma(|\mu_i - \mu_j|, \lambda) = \lambda \omega_{ij} |\mu_i - \mu_j|,$$

which requires specification of the weights ω_{ij}. As discussed in Chi and Lange (2015), the choice of the weights can dramatically affect the quality of results in cluster analysis. In the regression context such as in our study, it is even more challenging to select the weights.

For the L_1 penalty, we let the weight be $\omega_{ij} = \exp(-\phi(y_i - y_j)^2)$, which is a Gaussian kernel defined based on the distance of two points. The constant ϕ is nonnegative. When $\phi = 0$, it corresponds to the Lasso penalty. Note that it is unclear what weights we need to apply to obtain optimal results. We here use the Gaussian kernel as the weight to illustrate this point by using different values for ϕ.

Figure 1 shows the solution paths for the estimated values of $(\mu_1, \ldots, \mu_n)$ against λ values by using MCP and SCAD, and L_1 penalties of $\phi = 0, 0.5, 1, 2$, respectively, based on one sample with $n = 100$ and $\alpha = 1$. We observe that MCP and SCAD have similar solution paths as shown in Figure 1. For these two penalties, the estimated values of $\boldsymbol{\mu}$ converge to two different values around -1 and 1, which are the true values for the intercepts of the two groups, when λ reaches certain value (around 0.38 for both MCP and SCAD). They eventually converge to one value when λ exceeds 0.6. The various L_1 penalties, however, show solution paths substantially different from those of MCP and SCAD, and the solution paths appear quite different for different values of ϕ, showing how the choice of weights can dramatically affect estimation results. When $\phi = 0$, which corresponds to the Lasso penalty, we see that the estimated values for μ_i's converge quickly as the λ value increases until they converge to a common point around 0 when λ reaches 0.035. As a result, it cannot effectively identify the groups of the $\boldsymbol{\mu}$ value. By looking at the plots for $\phi = 0.5, 1, 2$, we observe that as the ϕ value increases, the estimated values converge to one point more slowly.

Next, we conduct the simulations by selecting λ via minimizing the modified BIC given in (15). Recall that we let $C_n = c \log(\log(n + p))$, where $n + p$ is the number of components in $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$, and c is a positive constant. We use different c values by letting $c = 5, 10$ in our estimation procedure. We consider different values for α by letting $\alpha = 1, 1.5, 2$, so that the difference of the true common values between the two groups varies from 2 to 4. Table 1 reports the mean, median, and standard error (s.e.) of the estimated number of groups $\widehat{K}$ by the MCP, SCAD, and L_1 methods with $\phi = 1$ and 2 based on 100 simulation realizations at $n = 100$. Moreover, to study the estimation accuracy, we report in Table 2 the average value and standard error (shown in the parentheses) of the square root of the mean squared error (MSE) for the estimated values of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ for the MCP, SCAD, and L_1 estimators. The square roots of the MSE for $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$

Table 1. The mean, median, and standard error (s.e.) of $\widehat{K}$ by the MCP, SCAD, and L_1 methods with $\phi = 1.0$ and 2.0 based on 100 realizations at $n = 100$ in Example 1.

c	α	1.0			1.5			2.0		
		Mean	Median	s.e.	Mean	Median	s.e.	Mean	Median	s.e.
5.0	MCP	2.57	2.00	0.90	2.41	2.00	0.93	2.10	2.00	0.44
	SCAD	2.58	2.00	0.96	2.37	2.00	0.90	2.18	2.00	0.63
	$L_1(\phi = 1.0)$	1.76	1.00	0.99	2.71	3.00	0.88	2.50	2.00	0.82
	$L_1(\phi = 2.0)$	3.03	3.00	1.16	3.13	3.00	1.19	3.25	3.00	1.00
10.0	MCP	2.10	2.00	0.33	2.04	2.00	0.20	2.01	2.00	0.11
	SCAD	2.11	2.00	0.35	2.04	2.00	0.20	2.02	2.00	0.14
	$L_1(\phi = 1.0)$	1.40	1.00	0.65	5.10	4.00	3.00	3.75	3.00	1.60
	$L_1(\phi = 2.0)$	2.29	2.00	0.78	3.03	3.00	1.02	3.25	3.00	1.00

are, respectively, defined as $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| / \sqrt{n}$ and $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}||/\sqrt{p}$ for each realization.

In Table 1, for both MCP and SCAD methods, we observe that the median value of $\widehat{K}$ among the 100 replications is 2 for all cases, which is the true number of groups in our model. We also observe that the mean values of $\widehat{K}$ are close to 2 for different values of α. For larger values of α, it is easier to detect the subgroups, so correspondingly we observe that the mean values of $\widehat{K}$ are closer to 2 for larger values of α. The MCP and SCAD can identify the groups for both values of c, although they perform better at $c = 10$ by having smaller standard errors. The L_1 penalties with both $\phi = 1$ and $\phi = 2$ in general have worse performance than the MCP and SCAD penalties. They have larger standard errors, and the mean and median values for $\widehat{K}$ are further away from 2. Moreover, the performance of the L_1 penalty is not stable. The L_1 penalty with $\phi = 1$ tends to select fewer than two groups for $\alpha = 1.0$ and more than two groups for $\alpha = 1.5, 2.0$, while the L_1 penalty with $\phi = 2$ tends to select more groups in general. For $\alpha = 1.5, 2.0$ and $c = 5.0$, the L_1 penalty with $\phi = 1$ performs better than the L_1 penalty with $\phi = 2$, since it yields smaller standard errors and has the $\widehat{K}$ values closer to 2. However, for other cases, the L_1 penalty with $\phi = 2$ seems to perform better. Thus, we see that different weights applied to the L_1 penalty may significantly affect the performance of the resulting estimator. Also, there is no clear rule on what weight to be used in the general situation. Table 2 shows that the MCP and SCAD methods have smaller MSE values than the L_1 penalty methods in general since they

have more accurate selection results and produce less biased estimates.

Next, we let $c = 10$ in the modified BIC method for tuning parameter selection. To evaluate the asymptotic normality given in Corollary 1, Table 3 lists the empirical bias (Bias) for the estimates of α_1 and α_2, and it also presents the average asymptotic standard error (ASE) calculated according to Corollary 1 and the empirical standard error (ESE) based on 100 replications for the MCP and SCAD methods as well as the oracle estimator (ORACLE). The biases are around zero for all cases. Moreover, we observe that the asymptotic standard errors for the MCP and SCAD methods are similar to those for the ORACLE estimator. This result supports our asymptotic normality result in Corollary 1.

Finally, we conduct statistical inference on the difference between groups. Table 4 presents the average p-values for testing $\mathcal{H}_0 : \alpha_1 = \alpha_2$ based on the 100 simulation realizations. We use $\sigma_{n1}(\mathbf{a})^{-1}(\widehat{\alpha}_1(\lambda) - \widehat{\alpha}_2(\lambda))$, $\mathbf{a} = (1, -1)$, as the test statistic that has the asymptotic normal distribution given in Corollary 1, and the estimates $\widehat{\alpha}_1(\lambda)$ and $\widehat{\alpha}_2(\lambda)$ are obtained by the MCP and SCAD methods. We obtain the p-values close to zero for all cases, so the difference between the groups is further confirmed by the inference procedure.

Example 2. We simulate data from model (16) with the predictors, the error terms, and the coefficients $\boldsymbol{\beta}$ generated from the same distributions as given in Example 1. We simulate μ_i's from three different values -2, 0, and 2 with equal probabilities. We

Table 2. The mean and standard error (s.e.) (shown in parentheses) of the square root of the MSE for the estimated values of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ for the MCP, SCAD, and L_1 ($\phi = 1.0$, 2.0) estimators based on 100 realizations at $n = 100$ in Example 1.

c	α	$\boldsymbol{\mu}$			$\boldsymbol{\beta}$		
		1.0	1.5	2.0	1.0	1.5	2.0
5.0	MCP	0.409	0.246	0.132	0.043	0.076	0.062
		(0.108)	(0.192)	(0.151)	(0.034)	(0.045)	(0.038)
	SCAD	0.414	0.240	0.158	0.091	0.075	0.065
		(0.116)	(0.190)	(0.168)	(0.036)	(0.044)	(0.040)
	$L_1(\phi = 1.0)$	0.874	0.370	0.185	0.118	0.084	0.066
		(0.202)	(0.237)	(0.173)	(0.040)	(0.047)	(0.036)
	$L_1(\phi = 2.0)$	0.637	0.274	0.167	0.106	0.076	0.064
		(0.226)	(0.180)	(0.153)	(0.040)	(0.041)	(0.035)
10.0	MCP	0.407	0.230	0.154	0.086	0.069	0.062
		(0.139)	(0.178)	(0.164)	(0.035)	(0.034)	(0.030)
	SCAD	0.409	0.234	0.155	0.086	0.069	0.061
		(0.138)	(0.178)	(0.163)	(0.034)	(0.035)	(0.030)
	$L_1(\phi = 1.0)$	0.946	0.265	0.203	0.121	0.075	0.069
		(0.138)	(0.142)	(0.169)	(0.039)	(0.038)	(0.038)
	$L_1(\phi = 2.0)$	0.769	0.287	0.167	0.113	0.078	0.064
		(0.215)	(0.210)	(0.153)	(0.039)	(0.046)	(0.035)

Table 3. The empirical bias (Bias) for the estimates of α_1 and α_2, the average asymptotic standard error (ASE) calculated according to Corollary 1, and the empirical standard error (ESE) based on 100 replications for the MCP, SCAD, and oracle (ORACLE) estimators in Example 1.

		$\alpha = 1.0$		$\alpha = 1.5$		$\alpha = 2.0$	
		α_1	α_2	α_1	α_2	α_1	α_2
MCP	Bias	0.037	-0.066	0.031	-0.055	0.065	-0.083
	ASE	0.071	0.070	0.072	0.071	0.075	0.074
	ESE	0.104	0.117	0.085	0.092	0.085	0.092
SCAD	Bias	0.040	0.069	0.036	-0.060	0.067	-0.087
	ASE	0.071	0.070	0.072	0.071	0.075	0.074
	ESE	0.103	0.119	0.085	0.094	0.084	0.094
ORACLE	Bias	-0.009	-0.005	-0.010	-0.005	-0.010	-0.005
	ASE	0.072	0.072	0.072	0.072	0.072	0.072
	ESE	0.070	0.067	0.070	0.067	0.070	0.067

Table 4. The average p-values for testing $\mathcal{H}_0 : \alpha_1 = \alpha_2$ based on the 100 simulation realizations with the estimates $\widehat{\alpha}_1(\lambda)$ and $\widehat{\alpha}_2(\lambda)$ obtained by the MCP and SCAD methods in Example 1.

| α	1.0	1.5	2.0
MCP	<0.001	<0.001	<0.001
SCAD	<0.001	<0.001	<0.001

Table 5. The mean and median of the estimated number of subgroups $\widehat{K}$, and the percentage that one group is identified among the 100 simulation realizations in Example 3.

	Mean	Median	Percentage
MCP	1.070	1.000	0.960
SCAD	1.080	1.000	0.960

use the modified BIC to select the tuning parameter λ by letting $C_n = 5 \log(\log(n + p))$. Figure 2 shows the boxplots of $\widehat{K}$ and the square root of the MSE for the estimated values of μ and β, respectively, by the MCP, SCAD, and L_1 with $\phi = 1, 2$ methods based on 100 simulation realizations at $n = 100$. In the first plot, we observe that for the MCP and SCAD methods, the median value for $\widehat{K}$ is 3, which is the true number of groups in our model. For some replications, they select more groups than 3. For the L_1 penalty with $\phi = 1$, the median value for $\widehat{K}$ is 3 as well. However, some replications have more than 3 and others have less than 3 for the $\widehat{K}$ value. Moreover, for this example, the L_1 penalty with $\phi = 2$ tends to select more groups in all replications. The other two plots show that MCP and SCAD have much smaller MSE values than the two L_1 penalty methods.

Example 3. We generate data from a homogeneous model given as $y_i = \mu + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i, i = 1, \ldots, 100$. The predictors, the error term, and the coefficients are simulated in the same way as in Example 1. Let $\mu = 2$. We estimate the model by using our proposed method. We choose the value of λ by the modified BIC method with $c = 10$. Table 5 reports the mean and median of the estimated number of subgroups $\widehat{K}$, and the percentage that one group is identified among the 100 simulation realizations.

For this homogeneous design, the true number of subgroups is 1. We observe that the median value of $\widehat{K}$ is 1, and the mean value is close to 1 for both MCP and SCAD methods. Moreover, the proposed penalized method identifies one group for 96 replications out of 100 by both methods. In Figure 3, we plot the solution paths for the estimated values of $(\mu_1, \ldots, \mu_n)$ against λ values by MCP and SCAD, based on one sample. We see that the estimated values of μ_i's converge to the value close to 2, which is the true value of μ, when λ is greater than a thresholding value for both methods. Next, we illustrate our inference method in this homogeneous model by using a set of different values for the tuning parameter λ. To test on homogeneity, we formulate the hypothesis that $\mathcal{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_n = \mu$. For a given λ, let $\widehat{\boldsymbol{\alpha}}(\lambda)$ be the distinct values of the estimated intercepts. We define the test statistic as $T(\lambda) = (\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha})^{\mathrm{T}} \widehat{\Sigma}_n^{-1} (\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = \mu \mathbf{1}_{\widehat{K}}$, $\widehat{\Sigma}_n = \widehat{\sigma}^2 \{\widehat{\mathbf{Z}}^{\mathrm{T}} \widehat{\mathbf{Z}} - (\widehat{\mathbf{Z}}^{\mathrm{T}} \mathbf{X})(\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} (\mathbf{X}^{\mathrm{T}} \widehat{\mathbf{Z}})\}^{-1}$, and $\widehat{\mathbf{Z}} = \widehat{z}_{ik}$ is the $n \times \widehat{K}$ matrix with $\widehat{z}_{ik} = 1$ for $i \in \widehat{\mathcal{G}}_k$ and $z_{ik} = 0$ otherwise. At significance level of 0.05, we calculate the average Type I error based on 500 replications by using $\sum_{i=1}^{500} I(\widehat{T}_i(\lambda) > \chi_{0.95,\widehat{K}}^2)/500$, where $\widehat{T}_i(\lambda)$ is the observed value of $T(\lambda)$ for the ith replicate and $\chi_{0.95,\widehat{K}}^2$ is the 0.95th quantile of the χ^2 distribution with $\widehat{K}$ degrees of freedom.
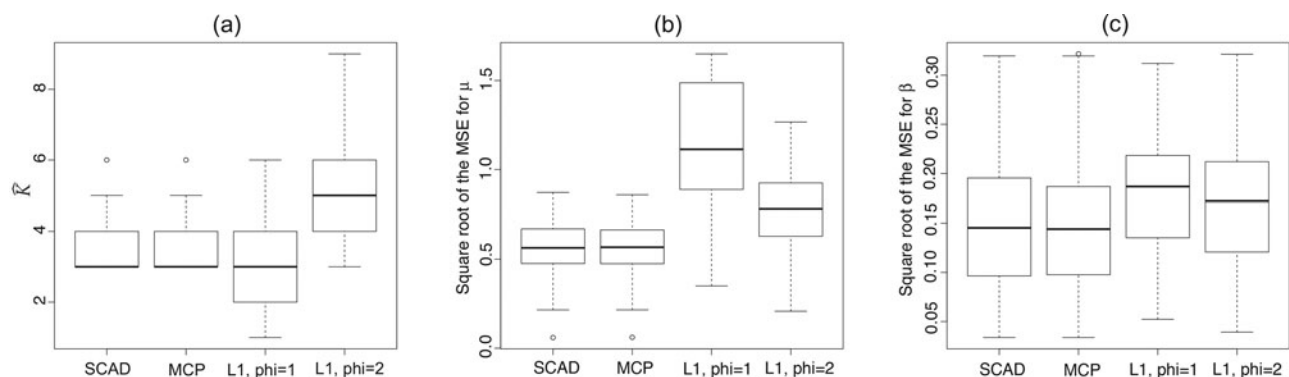
Figure 2. Boxplots of (a) $\widehat{K}$, (b) the square root of the MSE for the estimated values of μ, and (c) the square root of the MSE for the estimated values of β, by the MCP, SCAD, and L_1 with $\phi = 1, 2$ methods based on 100 simulation realizations at $n = 100$ in Example 2.
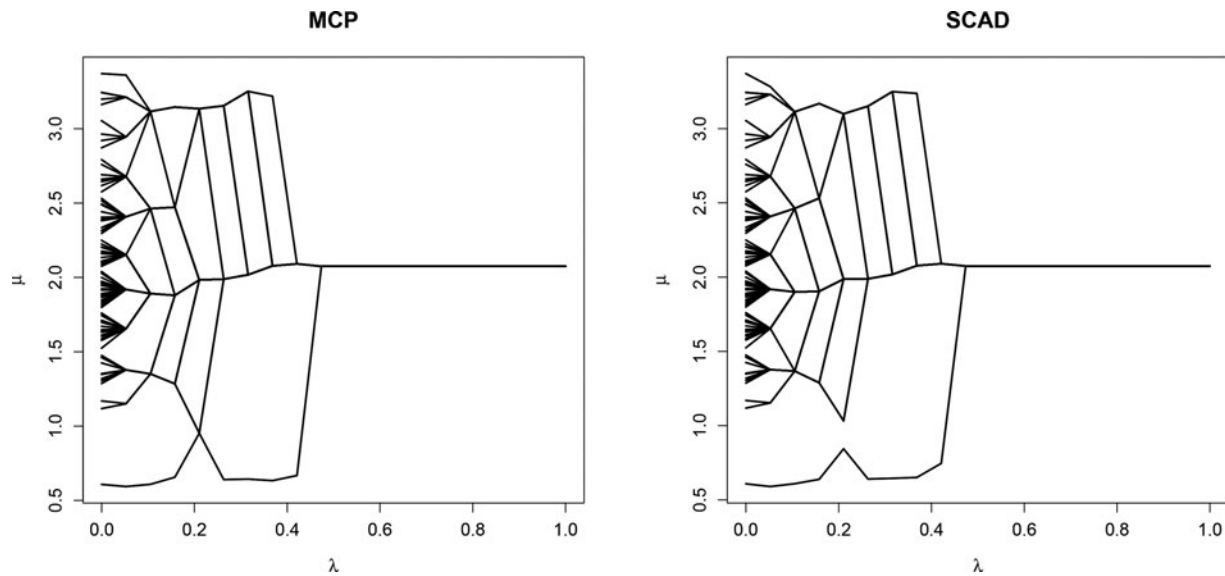
MCP

SCAD

Figure 3. Solution paths for the estimated values of $(\mu_1, \ldots, \mu_n)$ against λ values by using the MCP and SCAD penalties in Example 3.

Table 6 reports the average Type I errors based on 500 replications at different λ values. The last column shows the average Type I error at $\widehat{\lambda}_{\mathrm{BIC}}$, where $\widehat{\lambda}_{\mathrm{BIC}}$ is the λ value chosen by the modified BIC method. We observe that when λ is larger than some thresholding value, the average Type I error is very close to 0.05, which is the nominal significance level for both MCP and SCAD. This is because that for large λ values, one group is identified for most or all replications. Then the true structure is recovered with a high probability. When λ is chosen by the modified BIC, the average Type I error is reasonably controlled, but it is slightly higher than the nominal level. This suggests that using the λ value chosen by the modified BIC may affect the distribution of the test statistic. How to adjust such an effect is an interesting and challenging problem and requires further investigation.

Example 4.

   Case 1. We simulate data from the same data-generating process as that used in Example 2, so the data are generated from three groups with the same size (balanced groups). In this example, we aim to compare the performance of cluster analysis by using different penalties including MCP, SCAD, and truncated L_1 as well as by using the Gaussian mixture model-based clustering algorithm from the R package of MCLUST (Fraley and Raftery 2002). In our regression setting, we need to apply MCLUST to $y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ for cluster analysis. One simple way is to obtain the estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ by the ordinary least squares (OLS) first, and then apply MCLUST to the pseudo observations $y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$, which is adopted in our numerical analysis.

Table 6. The average Type I errors at different λ values based on 500 replications in Example 3.

λ	0.5	0.6	0.8	1.0	1.2	$\widehat{\lambda}_{\mathrm{BIC}}$
MCP	0.120	0.068	0.054	0.054	0.054	0.074
SCAD	0.102	0.058	0.054	0.054	0.054	0.072

For the penalized methods, we apply the same iterative algorithm as described in Section 3.1 to obtain the parameter estimates by using different penalties. The same BIC method is applied to choose the tuning parameter as described in Example 2. It is worth noting that the remaining steps are the same except for the estimation of η_{ij}, which needs some modifications due to the use of different penalties. Specifically, for the truncated L_1 penalty that has the form $p(|t|, \lambda; \tau) = \lambda \min(|t|; \tau)$, where τ is the thresholding parameter, the estimate of η_{ij} is obtained by minimizing $h(\eta_{ij}) = \frac{\vartheta}{2}(\delta_{ij} - \eta_{ij})^2 + \lambda \min(|\eta_{ij}|; \tau)$. We then apply the difference of convex programming technique as given in Shen and Huang (2010) to obtain the minimizer of $h(\eta_{ij})$. In this algorithm, the function $h(\eta_{ij})$ needs to be decomposed into difference of two convex functions $h_1(\eta_{ij}) - h_2(\eta_{ij})$, where $h_1(\eta_{ij}) = \frac{\vartheta}{2}(\delta_{ij} - \eta_{ij})^2 + \lambda|\eta_{ij}|$ and $h_2(\eta_{ij}) = \lambda(|\eta_{ij}| - \tau)_+$. This enables us to approximate $h(\eta_{ij})$ by an upper convex function at iteration $m + 1$, which results in

$$\widehat{\eta}_{ij}^{(m+1)}$$
$$= \begin{cases} \widehat{\delta}_{ij}^{(m+1)} & \text{if } |\widehat{\eta}_{ij}^{(m)}| \geq \tau \\ \left(|\widehat{\delta}_{ij}^{(m+1)}| - \lambda/\vartheta\right)_+ \left(\widehat{\delta}_{ij}^{(m+1)}/|\widehat{\delta}_{ij}^{(m+1)}|\right) & \text{otherwise.} \end{cases}$$

One important evaluation criterion for clustering methods is their ability to reconstruct the true underlying cluster structure. We, therefore, use the Rand Index measure (Rand 1971) to evaluate the accuracy of the clustering results. The Rand Index is viewed as a measure of the percentage of correct decisions made by an algorithm. It is computed by using the formula

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

Table 7. The mean and standard error (s.e.) of $\widehat{K}$ and the square root of the MSE (SMSE) for the estimated μ as well as the clustering accuracy (Accuracy) by different methods based on 100 realizations with $n = 100$ for Case 1 of Example 4 with balanced groups.

		MCP	SCAD	Truncated L_1			MCLUST	MCLUST-MCP
				$\tau = 0.5$	$\tau = 1.0$	$\tau = 1.5$		
K	Mean	3.570	3.600	6.930	3.960	2.390	2.400	—
	s.e.	0.671	0.696	0.956	0.887	0.737	0.711	—
SMSE of μ	Mean	0.589	0.585	0.597	0.605	0.963	0.791	0.607
	s.e.	0.157	0.154	0.158	0.164	0.195	0.380	0.134
Accuracy	Mean	0.897	0.892	0.829	0.873	0.707	0.777	0.864
	s.e.	0.059	0.057	0.066	0.064	0.112	0.193	0.058

where a true positive (TP) decision assigns two observations from the same ground truth group to the same cluster, a true negative (TN) decision assigns two observations from different groups to different clusters, a false positive (FP) decision assigns two observations from different groups to the same cluster, and a false negative (FN) decision assigns two observations from the same group to different clusters. The Rand Index lies between 0 and 1. Higher values of the Rand Index indicate better performance of the algorithm.

Table 7 presents the mean and standard error (s.e.) of $\widehat{K}$, the square root of the MSE (SMSE) for the estimated values of μ, and the clustering accuracy (Accuracy) by different methods. For the truncated L_1, by taking the same strategy as Shen and Huang (2010), we use different values $\tau = 0.5, 1.0, 1.5$ for the thresholding parameter. In the MCLUST column, it shows the results by using the MCLUST package with the number of groups selected by the BIC method. This is the default method in the MCLUST package and is widely used for determining the number of clusters in practice. The MCLUST-MCP column shows the results by using the MCLUST package with the number of groups determined by our proposed penalized approach with MCP penalty.

From Table 7, we observe that the proposed concave fusion penalized methods, MCP and SCAD, have better performance than other methods. They have higher clustering accuracy rates and smaller SMSE values for $\widehat{\mu}$ than others. This result is further reflected by the boxplots in Figure 4 of accuracy rates for the MCP, SCAD, truncated L_1 with $\tau = 1.0$, and MCLUST methods. For the truncated L_1, the performance of $\tau = 1.0$ is the best among the three different values for τ. Moreover, the three penalized methods, MCP, SCAD, and truncated L_1 with $\tau = 1.0$, can identify cluster membership more correctly than the MCLUST method by having higher accuracy rates. MCP improves the accuracy rate by 15.4% compared to MCLUST. It is worth noting that to apply the Gaussian mixture model-based method, how many clusters to be used is always crucial. For MCLUST-MCP, instead of using the BIC, we use our proposed penalized MCP approach to determine the number of clusters and then apply MCLUST. We see that the accuracy rate is improved compared to MCLUST with the BIC method. This result indicates that our proposed concave penalized method also provides a possible tool to determine the number of clusters for the Gaussian mixture model-based method.

Case 2. In this setting, we generate data from three groups with different sizes (unbalanced groups). We consider two simulation designs: Design 1: μ_i's are generated from three different values -2, 0, and 2 with probabilities 0.2, 0.3, and 0.5, respectively, and Design 2: μ_i's are generated from -2, 0, and 2 with probabilities 0.1, 0.3, and 0.6, respectively. Other terms are simulated according to the same setting as Case 1. Table 8 presents the mean and standard error (s.e.) of $\widehat{K}$, the square root of the MSE (SMSE) for the estimated values of μ, and the clustering accuracy (Accuracy) by MCP, SCAD, truncated L_1, MCLUST, and MCLUST-MCP based on 100 realizations. We see that for the MCP, SCAD, and truncated L_1 methods, the performance for the two unbalanced designs is comparable to that for the balanced design in Case 1. Again, MCP and SCAD outperform the other methods. The performance of MCLUST-MCP shows improvement over MCLUST. For MCLUST, the estimated number of groups $\widehat{K}$ decreases as the design becomes more unbalanced. The smallest group is not successfully identified for most replications. However, for the penalized method, the $\widehat{K}$ values remain similar for different designs. Hence, MCLUST may be more sensitive to cluster sizes based on these simulation results.
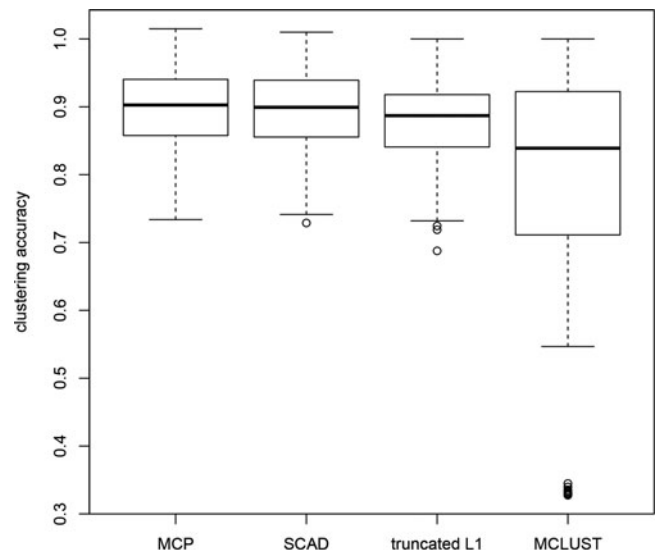

Figure 4. Boxplots of the clustering accuracy for the MCP, SCAD, truncated L_1, and MCLUST methods based on the 100 simulation realizations in Case 1 of Example 4.

Table 8. The mean and standard error (s.e.) of $\widehat{K}$, the square root of the MSE (SMSE) for the estimated values of $\boldsymbol{\mu}$, and the clustering accuracy (Accuracy) by different methods based on 100 realizations at $n = 100$ for Case 2 of Example 4 with unbalanced groups.

		MCP	SCAD	Truncated L_1 $\tau = 0.5$	$\tau = 1.0$	$\tau = 1.5$	MCLUST	MCLUST-MCP
				Design 1				
K	Mean	3.730	3.660	6.540	3.870	2.360	2.380	—
	s.e.	0.670	0.713	1.049	0.928	0.659	0.663	—
SMSE of $\boldsymbol{\mu}$	Mean	0.561	0.556	0.585	0.577	0.893	0.771	0.592
	s.e.	0.126	0.130	0.127	0.146	0.135	0.309	0.124
Accuracy	Mean	0.890	0.891	0.822	0.872	0.733	0.792	0.846
	s.e.	0.048	0.048	0.051	0.058	0.092	0.152	0.064
				Design 2				
K	Mean	3.700	3.730	6.350	3.960	2.690	2.230	—
	s.e.	0.717	0.709	0.880	1.197	1.473	0.679	—
SMSE of $\boldsymbol{\mu}$	Mean	0.488	0.487	0.522	0.502	0.852	0.763	0.579
	s.e.	0.121	0.120	0.122	0.127	0.135	0.220	0.148
Accuracy	Mean	0.898	0.899	0.823	0.877	0.713	0.793	0.818
	s.e.	0.048	0.047	0.056	0.054	0.118	0.123	0.097

6. Empirical Example

In this section, we use the Cleveland Heart Disease Dataset to illustrate our method. This dataset is available at the UCI machine learning repository. The dataset has 13 clinical measurements on 297 individuals. As described in Lauer et al. (1999), the maximum heart rate achieved (thalach) variable is related to cardiac mortality. In addition, some categorical variables are also used to check heart problems including chest pain type, exercise-induced angina indicator, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, the number of major vessels colored by fluoroscopy, and the heart status (normal = 3; fixed defect = 6; reversible defect = 7). We use the fitted value of thalach as the response variable by projecting it onto the linear space spanned by the categorical variables. Our interest is to conduct subgroup analysis for the fitted value of thalach as the response y after adjusting for the effects of the covariates: $x_1 =$ age in years; $x_2 =$ gender; $x_3 =$ resting blood pressure; $x_4 =$ serum cholesterol; $x_5 =$ fasting blood sugar indicator; and $x_6 =$ resting electrocardiographic results.

We first plot the kernel density estimates of $y_i - \mathbf{x}_i^\mathrm{T} \widehat{\boldsymbol{\beta}}^{\text{ols}}$ in Figure 5, where $\widehat{\boldsymbol{\beta}}^{\text{ols}}$ is obtained from OLS estimation. Clearly, we see that after adjusting for the effects of the covariates, the distribution in Figure 5 still shows multiple modes. The heterogeneity may be caused by some unobserved latent factors. Hence, it is not suitable to fit a standard linear regression model with a common intercept by using the response and the predictors. Instead, we fit the heterogeneous model $y_i = \mu_i + \mathbf{x}_i^\mathrm{T}\boldsymbol{\beta} + \epsilon_i$, $i = 1, \ldots, 297$, and we identify subgroups by our proposed ADMM algorithm. We select the tuning parameter by minimizing the modified BIC in a certain range by following the same rule as given in Example 2 of Section 5. As a result, two major groups are identified by both MCP and SCAD. Figure 6 displays the kernel density estimates of $y_i - \mathbf{x}_i^\mathrm{T}\widehat{\boldsymbol{\beta}}$ in each of the two identified subgroups for the SCAD method. The density plots for the MCP method show similar patterns, so we do not present them to save space. We see that the distribution is more homogeneous within each identified group than the distribution of all response values as shown in Figure 5. We also conduct inference by testing the difference of the intercepts for the two identified groups by using the asymptotic normality in Corollary 1, and

we find that the p-values are close to zero for both MCP and SCAD.

We report in Table 9 the estimated coefficients $\widehat{\boldsymbol{\beta}}$, their standard errors (s.e.), and the p-values for testing the significance of the coefficients using MCP, SCAD, and OLS by assuming a common intercept. The standard error for the MCP and SCAD methods is calculated by the asymptotic formula given in Corollary 1. The age and gender variables show a strong effect by these three methods with p-values close to zero, while resting blood pressure and serum cholesterol show a very weak effect and have large p-values. Moreover, using the MCP and SCAD methods, the effects of fasting blood sugar indicator and resting electrocardiographic results are more significant than their effects by the OLS method. This result indicates that recovering the hidden heterogeneous structure of the data helps us identify useful variables that may have effects on the response. We also calculate the coefficient of determination R^2, and obtain $R^2 = 0.667, 0.704$, and 0.109 for MCP, SCAD, and OLS. We see that taking into account the subgroup structure leads to a significant improvement of the model fitting. Next we apply the
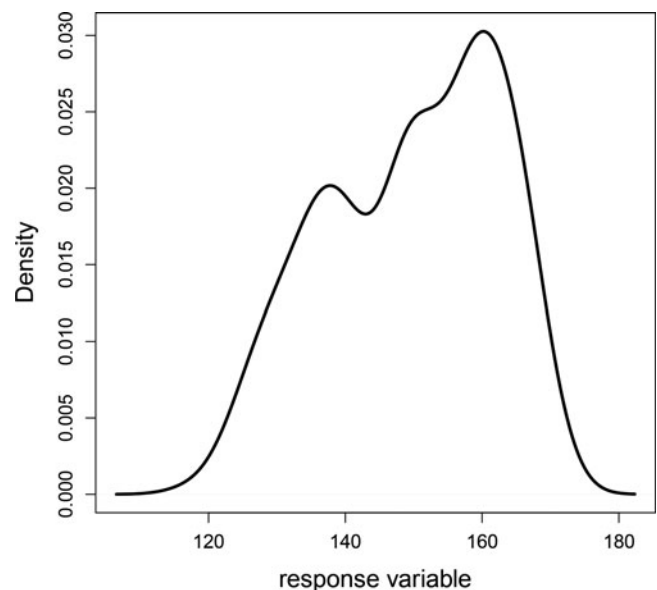

Figure 5. Density plot of the response variable after adjusting for the effects of the covariates for the empirical example.
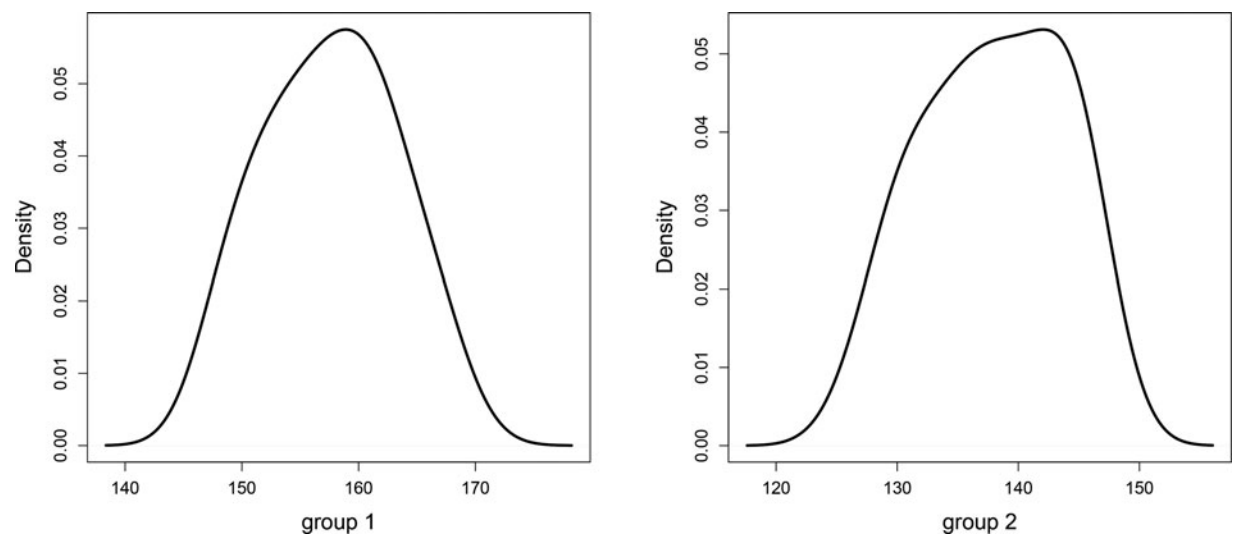
Figure 6. Density plots of the response variable after adjusting for the effects of the covariates in each of the two identified subgroups for the SCAD method.

Gaussian mixture model-based method to this dataset for cluster analysis. As described in Example 4 of the simulation section, we apply MCLUST to the pseudo observations $y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is obtained from OLS. As a result, two subgroups are identified. For the real data, since the true underlying cluster structure is unknown, we cannot use the external criterion, Rand Index measure, to evaluate and compare different methods. Instead, we use the internal criterion, the Davies–Bouldin index, to assess the quality of clustering algorithms, which is calculated by the formula: $\mathrm{DB} = \widehat{K}^{-1}\sum_{k=1}^{\widehat{K}}\max_{k'\neq k}((\sigma_k + \sigma_{k'})/d(c_k, c_{k'}))$, where $\widehat{K}$ is the estimated number of clusters, c_k is the centroid of cluster k, σ_k is the average distance of all observations $y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ in cluster k to centroid c_k, and $d(c_k, c_{k'})$ is the distance between centroids c_k and $c_{k'}$. The clustering algorithm that has the smallest Davies–Bouldin index is considered the best algorithm. The Davies–Bouldin index values for MCP, SCAD, and MCLUST are 0.469, 0.467, and 0.506, respectively, so MCP and SCAD outperform MCLUST based on this criterion.

7. Discussion

Model (1) is related to the Neyman-Scott models (Neyman and Scott 1948). In the terminology of Neyman and Scott, the μ_i's in (1) are called incidental parameters. In the literature, such parameters are usually treated as nuisance parameters, while the main interest lies in estimating the common parameter such as $\{\boldsymbol{\beta}, \sigma^2\}$ in (1) based on panel data (Lancaster

2000). The problem we consider here is different and we use the μ_i's to represent latent heterogeneity in the observations for the purpose of conducting subgroup analysis. Also, we do not assume that panel data are available, so model (1) is not identifiable without a constraint on the parameter space such as the subgroup structure considered in the present article.

It is also possible to adopt a random effects model approach by taking the μ_i's in (1) as random variables from a mixture distribution. Then, the estimation and inference can be carried out using a likelihood-based method. The main difficulty in applying this approach is that it requires specifying the number of subgroups, the parametric form of the mixture distribution, and an assumption on the error distribution. It is worth noting that the choice of the number of groups is always crucial in mixture model-based methods. Different methods on this topic have been proposed in the literature. Among them, the Bayesian model selection criteria (Fraley and Raftery 1998) are widely used, and the gap statistic proposed in Tibshirani, Walther, and Hastie (2001) is also an important tool. Our proposed penalized method provides another possible approach to automatically estimate the number of groups with reliable theoretical properties. By using MCLUST, our simulation studies show that the clustering accuracy is improved compared to the BIC method by using the proposed penalized method to select the number of groups.

The oracle property of the penalized estimators allows one to carry out statistical inference on the regression parameters,

Table 9. The estimated values (est) for the coefficients, their standard errors (s.e.), and the p-values for testing the significance of the coefficients by OLS, MCP, and SCAD, respectively.

		β_1	β_2	β_3	β_4	β_5	β_6
OLS	est	-0.345	-4.120	-0.028	-0.008	0.183	-1.359
	s.e.	0.083	1.534	0.042	0.0142	2.031	0.725
	p-value	<0.001	0.007	0.502	0.563	0.928	0.061
MCP	est	-0.355	-3.825	-0.007	-0.006	0.628	-1.849
	s.e.	0.040	0.752	0.021	0.007	1.016	0.354
	p-value	<0.001	<0.001	0.563	0.558	0.283	<0.001
SCAD	est	-0.358	-3.698	-0.012	-0.004	1.091	-2.129
	s.e.	0.040	0.743	0.021	0.007	1.005	0.351
	p-value	<0.001	<0.001	0.558	0.554	0.278	<0.001

given that the group structure is recovered with probability approaching 1, and the tuning parameter λ is a fixed value. We identify two possible future research topics including (1) studying accuracy of the resulting inference after penalized regression, that is, investigating the convergence rate of the distribution of the penalized estimator to its oracle limit, and (2) studying the properties of the penalized estimator with λ selected by data-driven methods such as BIC. These two topics are interesting albeit mathematically challenging, and need to be further investigated carefully. In our theoretical results, we allow p, the dimension of the regression parameter $\boldsymbol{\beta}$, to diverge with n, but require it to be smaller than n. For models and data with $p > n$, a sparsity condition needs to be imposed on $\boldsymbol{\beta}$ and an additional penalty term to enforce the sparsity is required. Computationally, we can still derive an algorithm within the framework. However, significant extra effort is needed to establish the theoretical properties of the estimators in this high-dimensional setting. This interesting and challenging technical problem deserves further investigation, but is beyond the scope of this article.

The proposed method can be extended to other models including generalized linear models and regression models for censored survival data. Although these extensions appear to be conceptually straightforward, it is a nontrivial task to develop computational algorithms and establish theoretical properties in these more complicated models.

Supplementary Materials

In the supplementary materials, we give the technical proofs for Proposition 1 and Theorems 1–3. We also provide a detailed estimation procedure for model (2) based on the ADMM algorithm.

Acknowledgments

Funding

References

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821. [410]

Bondell, H. D., and Reich, B. J. (2008), "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors With Oscar," *Biometrics*, 64, 115–123. [411]

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, 3, 1–122. [411]

Chaganty, A. T., and Liang, P. (2013), "Spectral Experts for Estimating Mixtures of Linear Regressions," *Proceedings of the 30th International Conference on Machine Learning*, 28, 1040–1048. [410]

Chi, E. C., and Lange, K. (2015), "Splitting Methods for Convex Clustering," *Journal of Computational and Graphical Statistics*, 24, 994–1013. [411,416]

Everitt, B., and Hand, D. J. (1981), *Finite Mixture Distributions*, New York: Chapman & Hall. [410]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [411,413]

Fraley, C., and Raftery, A. E. (1998), "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *The Computer Journal*, 41, 578–588. [422]

—— (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. [419]

Guo, F. J., Levina, E., Michailidis, G., and Zhu, J. (2010), "Pairwise Variable Selection for High-Dimensional Model-Based Clustering," *Biometrics*, 66, 793–804. [411]

Hastie, T., and Tibshirani, R. (1996), "Discriminant Analysis by Gaussian Mixtures," *Journal of the Royal Statistical Society*, Series B, 58, 155–176. [410]

Ke, T., Fan, J., and Wu, Y. (2015), "Homogeneity in Regression," *Journal of the American Statistical Association*, 110, 175–194. [411]

Lancaster, T. (2000), "The Incident Parameter Problem Since 1948," *Journal of Econometrics*, 95, 391–413. [422]

Lauer, M. S., Francis, G. S., Okin, P. M., Pashkow, F. J., Snader, C. E., and Marwick, T. H. (1999), "Impaired Chronotropic Response to Exercise Stress Testing as a Predictor of Mortality," *Journal of the American Medical Association*, 281, 524–529. [421]

McNicholas, P. D. (2010), "Model-Based Classification Using Latent Gaussian Mixture Models," *Journal of Statistical Planning and Inference*, 140, 1175–1181. [410]

Neyman, J., and Scott, E. L. (1948), "Consistent Estimation From Partially Consistent Observations," *Econometrica*, 16, 1–32. [422]

Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846–850. [420]

Schwarz, C. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [416]

Shen, J., and He, X. (2015), "Inference for Subgroup Analysis With a Structured Logistic-Normal Mixture Model," *Journal of the American Statistical Association*, 110, 303–312. [410]

Shen, X., and Huang, H. C. (2010), "Grouping Pursuit Through a Regularization Solution Surface," *Journal of the American Statistical Association*, 105, 727–739. [411,420]

Tibshirani, R., Walther, G., and Hastie, T. (2001), "Estimating the Number of Clusters in a Dataset via the Gap Statistic," *Journal of the Royal Statistical Society*, Series B, 63, 411–423. [422]

Tibshirani, S., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of Royal Statistical Society*, Series B, 67, 91–108. [411]

Wang, H., Li, B., and Leng, C. (2009), "Shrinkage Tuning Parameter Selection With a Diverging Number of Parameters," *Journal of Royal Statistical Society*, Series B, 71, 671–683. [416]

Wang, H., Li, R., and Tsai, C. L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568. [416]

Wei, S., and Kosorok, M. (2013), "Latent Supervised Learning," *Journal of the American Statistical Association*, 108, 957–970. [410]

Zhang, C. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [411,413]