Estimating subgroup-specific treatment effects via concave fusion

Shujie Ma[1] and Jian Huang[2]

[1]University of California, Riverside, USA
[2]University of Iowa, Iowa City, USA

Abstract

Understanding treatment heterogeneity is essential to the development of precision medicine, which seeks to tailor medical treatments to subgroups of patients with similar characteristics. One of the challenges to achieve this goal is that we usually do not have *a priori* knowledge of the grouping information of patients with respect to treatment. To address this problem, we consider a heterogeneous regression model by assuming that the coefficients for treatment variables are subject-dependent and belong to different subgroups with unknown grouping information. We develop a concave fusion penalized method and derive an alternating direction method of multipliers algorithm for its implementation. The method is able to automatically estimate the grouping structure and the subgroup-specific treatment effects. We show that under suitable conditions the oracle least squares estimator with *a priori* knowledge of the true grouping information is a local minimizer of the objective function with high probability. This provides a theoretical justification for the statistical inference about the subgroup structure and treatment effects. We evaluate the performance of the proposed method by simulation studies and illustrate its application by analyzing the data from the AIDS Clinical Trials Group Study.

Keywords: Fusiongram, oracle property, penalization, precision medicine, regression, treatment heterogeneity

Address for correspondence: Jian Huang, University of Iowa, Iowa City, IA 52242, Email: jian-huang@uiowa.edu

1 Introduction

In this paper, we consider the problem of estimating heterogeneous treatment effects in the context of linear regression models. We propose a new approach to estimating subgroup-specific treatment effects without knowing the group membership of the subjects in advance. It consists of two main ingredients: an individualized treatment effect model and a concave fusion penalized method. We develop a computational algorithm and study the theoretical properties of the approach.

Understanding treatment heterogeneity is critical to the eventual success of precision medicine, which seeks to develop medical treatments tailored to heterogeneous subpopulations of patients with similar characteristics. Treatment heterogeneity is present when the same treatment yields different results in different subpopulations. For many complex diseases, significant treatment heterogeneity exists among patients with different clinical characteristics (Sorensen, 1996). Heterogeneity of treatment effects reflects diversity of patients in genetic and environmental factors, responsiveness to treatment, vulnerability to adverse effects, among others. When treatment heterogeneity is present, the average effect can be misleading. Indeed, the modest benefit ascribed to many treatments in clinical trials can be misleading, since average effects may reflect a mixture of substantial benefits for some, little benefit for many, and harm for a few (Kravitz and Braslow, 2004).

The most commonly used approach to dealing with treatment heterogeneity is subgroup analysis, but the existing subgroup analysis methods lack a rigorous statistical framework and is prone to yielding misleading results (Kravitz and Braslow, 2004; Rothwell, 2005; Lagakos, 2006). Another commonly used approach in modeling heterogeneity is based on finite mixture models. Recently, this approach has been adapted to subgroup analysis (Shen and He, 2015). However, the mixture model approach requires specifying the number of components and a parametric assumption of the model, which is difficult to do in practice.

To estimate treatment effects in the presence of heterogeneous subgroups, a challenging problem is that the grouping information of patients with respect to treatment is unknown in advance. To address this problem, we consider a regression model with heterogeneous treatment effects by allowing the coefficients for treatment variables to be subject-dependent and assume these coefficients belong to different groups with unknown grouping information. We propose a concave fusion penalized method that applies a suitable penalty to pairwise differences of treatment effects. By using a data-driven procedure for determining the penalty parameter, the method is able to automatically estimate the grouping structure in the data and the subgroup-specific treatment effects. Because the number of subgroups is usually much smaller than the sample size, there is an underlying sparsity structure in subgroup analysis. This enables us to formulate the problem of subgroup analysis for treatment heterogeneity as a penalization problem. Thus our proposed method places the problem of

subgroup analysis on a solid theoretical footing based on a well defined objective function. As a result, statistical inference about subgroup structure and treatment effects can be carried out in a rigorous fashion.

Computationally, we devise an alternating direction method of multipliers algorithm (ADMM, Boyd et al., 2011) for implementing the proposed approach. This algorithm has been used for solving a large class of convex optimization problems. In this paper, we use concave penalties on the pairwise differences of the treatment effects. Such penalties include the smoothly clipped absolute deviations penalty (SCAD, Fan and Li, 2001) and the minimax concave penalty (MCP, Zhang, 2010). The main reason we use the concave penalties is that they enjoy certain attractive properties, in that under certain conditions they can correctly identify the number of subgroups and yield nearly unbiased estimates of treatment effects with high probability. In addition, the thresholding operators corresponding to these penalties have explicit expressions. This facilitates the implementation of the method in the framework of ADMM. We also derive the convergence properties of the ADMM algorithm.

Our theoretical analysis gives insights into the properties of the proposed method. In particular, we provide sufficient conditions under which the oracle least squares estimator with *a priori* knowledge of the true subgroups is a local minimizer of the objective function with high probability. Consequently, the approximate distributional properties of the estimator can be obtained. This gives theoretical support for using the method for making statistical inference about the treatment effects in the presence of heterogeneity. Moreover, we derive the lower bound of the minimum difference of coefficient values between groups in order to identify the true subgroups of treatment.

The basic idea of our proposed approach grew out of several existing methods, including the fused lasso (Tibshirani et al., 2005), the convex splitting method for clustering (Chi and Lange, 2015), and the concave pairwise fusion method for subgrouping in the presence of covariates (Ma and Huang, 2016). The fused lasso deals with the standard regression model with ordered coefficients, which is different from the problem we consider. The convex splitting method is developed for clustering analysis, not for regression problems. The method in Ma and Huang (2016) was proposed for regression models with subject-specific intercepts in the model, while the present work considers estimation of subgroup-specific effects of observed treatment variables.

For studying grouping effects of covariates, several penalization methods have been proposed. For example, the group and adaptive group LASSO methods using an L_2 norm of coefficients for groups of covariates have been widely applied in various studies (Yuan and Lin, 2006; van de Geer and Bühlmann, 2009; Huang et al., 2010; Breheny and Huang, 2015). The fused concave penalization methods have been considered (Guo et al., 2010; Shen and Huang, 2010; Ke et al., 2015) for grouping effects of covariates. Different from these stud-

ies on grouping effects of covariates, our work concerns the estimation of subgroup-specific treatment effects across subjects.

The rest of this paper is organized as follows. Section 2 describes the concave fusion penalization method. Section 3 presents the ADMM algorithm with concave penalties. In Section 4 we establish the theoretical properties of the proposed estimator. In Section 5 we evaluate the finite sample properties of the proposed method via simulation studies. In Section 6 we illustrate the proposed method by analyzing the data from the AIDS Clinical Trials Group Study. Concluding remarks are given in Section 7. The proofs are given in the Appendix.

2 The model and the method

2.1 A heterogeneous treatment effects model

Suppose the data consists of $(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i), i = 1, \ldots, n$, where y_i is a response, $\boldsymbol{z}_i$ is a q-dimensional covariate vector and $\boldsymbol{x}_i$ is a p-dimensional covariate vector of main interest. To motivate the proposed model and approach, first consider the standard linear regression model

$$y_i = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ are unknown regression coefficients and the ε_i's are i.i.d. random errors with $E(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. We assume that the first entry in each $\boldsymbol{z}_i$ is 1 so the intercept is included in $\boldsymbol{\eta}$. We are interested in the effects of $\boldsymbol{x}_i$ on the response in the presence of other nuisance covariates in $\boldsymbol{z}_i$ which also may be related to the response. For simplicity and to emphasize the main role of $\boldsymbol{x}_i$, we refer to $\boldsymbol{\beta}$ as the treatment effect. In this model, a key assumption is that $\boldsymbol{\beta}$ is the same for all individuals in the data. However, this homogeneity assumption in treatment is violated if the observations consist of subgroups and the effects are difference across the subgroups, that is, the treatment effects are subgroup-specific. Applying this model to data with subgroup structure can lead to misleading results.

To estimate subgroup-specific treatment effects, we propose a heterogenous treatment effects model given by

$$y_i = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i + \varepsilon_i, i = 1, \ldots, n. \tag{2.2}$$

The difference between (2.2) and (2.1) is that $\boldsymbol{\beta}_i$ can be individual-specific. This enables us to incorporate possible treatment heterogeneity in a natural way in regression modeling.

Clearly, it is impossible to estimate each individual-specific coefficient $\boldsymbol{\beta}_i$ without additional information or further assumptions on the structure of the parameters. Here we assume that there are K different subgroups and the treatment effects are the same within each subgroup. Specifically, let $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_K)$ be a mutually exclusive partition of $\{1, \ldots, n\}$.

Suppose $\boldsymbol{\beta}_i = \boldsymbol{\alpha}_k$ for all $i \in \mathcal{G}_k$, where $\boldsymbol{\alpha}_k$ is the common value for the $\boldsymbol{\beta}_i$'s from group $\mathcal{G}_k$. In practice, the number of groups K is unknown, and we also have no knowledge of which subjects belonging to which groups. Our task is to estimate K, identify the subgroups and estimate the underlying parameters $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$ and $\boldsymbol{\eta}$.

2.2 Concave fusion

For any vector $\mathbf{a}$, denote its L_2 norm by $\|\mathbf{a}\| = (\sum |a_i|^2)^{1/2}$. Consider the criterion

$$Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{2} \sum\nolimits_{i=1}^{n} (y_i - \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\eta} - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_i)^2 + \sum\nolimits_{1 \le i < j \le n} p(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|, \lambda), \tag{2.3}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_n^{\mathrm{T}})^{\mathrm{T}}$, and $p(\cdot, \lambda)$ is a penalty function with a tuning parameter $\lambda \ge 0$.

We use sparsity-inducing penalties in (2.3). For a sufficiently large λ, the penalty shrinks some of $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|$ to zero. We can partition the treatment effects into subgroups according to the unique values of $\widehat{\boldsymbol{\beta}}$. Specifically, let $\widehat{\lambda}$ be the value of the tuning parameter on the path selected based on a data-driven procedure such as the BIC. For simplicity, write $(\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{\beta}}) \equiv (\widehat{\boldsymbol{\eta}}(\widehat{\lambda}), \widehat{\boldsymbol{\beta}}(\widehat{\lambda}))$. Let $\{\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_{\widehat{K}}\}$ be the distinct values of $\widehat{\boldsymbol{\beta}}$. These are the estimates of subgroup-specific treatment effects. The samples can then be divided into subgroups accordingly. Let $\widehat{\mathcal{G}}_k = \{i : \widehat{\boldsymbol{\beta}}_i = \widehat{\boldsymbol{\alpha}}_k, 1 \le i \le n\}, 1 \le k \le \widehat{K}$. Then $\{\widehat{\mathcal{G}}_1, \ldots, \widehat{\mathcal{G}}_{\widehat{K}}\}$ constitutes a partition of $\{1, \ldots, n\}$.

A popular sparsity-inducing penalty is the L_1 or lasso penalty with $p_\gamma(t, \lambda) = \lambda |t|$ (Tibshirani, 1996). But this penalty tends to produce too many subgroups (Ma and Huang, 2016). So we focus on two concave penalty functions: the smoothly clipped absolute deviation penalty (SCAD, Fan and Li, 2001) and the minimax concave penalty (MCP, Zhang, 2010). The SCAD penalty is

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

The MCP has the form

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

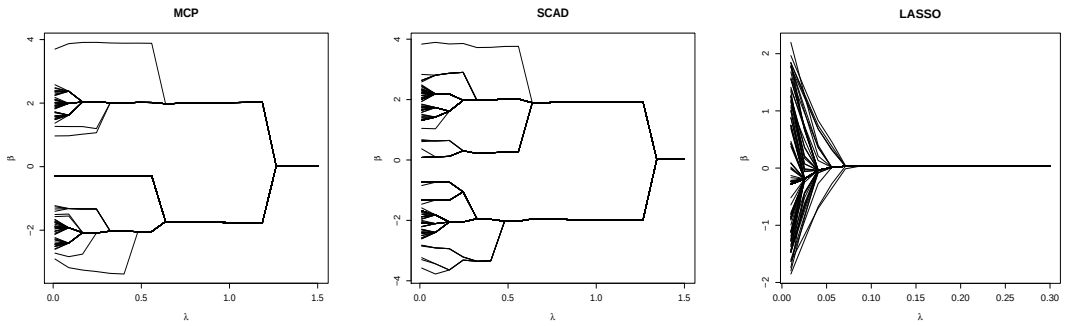
These penalties are nearly unbiased and are more aggressive in enforcing a sparser solution. Thus, they are better suited for the current problem, since the number of subgroups are usually much smaller than the sample size.

For a given $\lambda > 0$, let

$$(\widehat{\boldsymbol{\eta}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda)) = \underset{\boldsymbol{\eta} \in \mathbb{R}^q, \boldsymbol{\beta} \in \mathbb{R}^{np}}{\operatorname{argmin}} Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}; \lambda). \tag{2.4}$$

We compute $(\widehat{\boldsymbol{\eta}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$ for λ in a given interval $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\max}$ is the value that forces a constant $\hat{\boldsymbol{\beta}}$ solution, i.e., $\hat{\boldsymbol{\beta}}_j(\lambda_{\max}) = \hat{\boldsymbol{\beta}}_k(\lambda_{\max}), 1 \le j < k \le n$; $\lambda_{\min}$ is a small

Figure 1: *Solution paths for* $(\hat{\beta}_1(\lambda), \ldots, \hat{\beta}_n(\lambda))$ *against* λ *with* $n = 200$ *for data from Example 1 in Section 5.*



positive number. We are particularly interested in the path $\{\hat{\boldsymbol{\beta}}(\lambda) : \lambda \in [\lambda_{\min}, \lambda_{\max}]\}$. The algorithm for computing the solution path on a grid of λ values is described in detail in Section 3.

Figure 1 illustrates the solution path for $\widehat{\boldsymbol{\beta}}(\lambda) = (\widehat{\beta}_1(\lambda), \ldots, \widehat{\beta}_n(\lambda))$ against λ using MCP, SCAD and lasso penalties for data generated from the model in Example 1 in Section 5, in which there are two subgroups with 'treatment effects' 2 and -2, respectively. The path is calculated using a "bottom up" approach starting from $\lambda_{\min}$. It looks similar to the dendrogram for agglomerative hierarchical clustering. However, unlike the clustering algorithms which form the clusters based on a direct measure of dissimilarity, the fusion of the coefficients is based on solving the optimization problems along the solution path. We shall refer to the solution path $\{\widehat{\boldsymbol{\beta}}(\lambda), \lambda \in [\lambda_{\min}, \lambda_{\max}]\}$ as a *fusiongram*.

The fusiongrams for SCAD and MCP look similar. They both include a segment containing nearly unbiased estimates of the treatment effects. When the λ value reaches around 0.6, the estimates of $(\beta_1, \ldots, \beta_n)$ merge to the two true values 2 and -2, respectively. When the λ value exceeds 1.2, the estimates shrink to one value. For the lasso, we see that the estimates of $(\beta_1, \ldots, \beta_n)$ merge to one value quickly at $\lambda = 0.1$ due to the overshrinkage of the L_1 penalty.

3 Computation

3.1 ADMM with concave penalties

We derive an ADMM algorithm for computing the solution (2.4). The key idea is to introduce a new set of parameters $\boldsymbol{\delta}_{ij} = \boldsymbol{\beta}_i - \boldsymbol{\beta}_j$. Then, we can reformulate the problem of minimizing

(2.3) as that of minimizing

$$L_0(\boldsymbol{\eta},\boldsymbol{\beta},\boldsymbol{\delta}) = \frac{1}{2}\sum\nolimits_{i=1}^{n}(y_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i)^2 + \sum\nolimits_{i<j} p_\gamma(\|\boldsymbol{\delta}_{ij}\|,\lambda),$$
$$\text{subject to } \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij} = \mathbf{0}, \tag{3.1}$$

where $\boldsymbol{\delta} = \{\boldsymbol{\delta}_{ij}^{\mathrm{T}}, i<j\}^{\mathrm{T}}$. Let $\langle \mathbf{a},\mathbf{b}\rangle = \mathbf{a}^{\mathrm{T}}\mathbf{b}$ be the inner product of two vectors $\mathbf{a}$ and $\mathbf{b}$ with the same dimension. The augmented Lagrangian is

$$L(\boldsymbol{\eta},\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{\upsilon}) = L_0(\boldsymbol{\eta},\boldsymbol{\beta},\boldsymbol{\delta}) + \sum\nolimits_{i<j}\left\langle \boldsymbol{\upsilon}_{ij}, \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}\right\rangle + \frac{\vartheta}{2}\sum\nolimits_{i<j}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}\|^2, \tag{3.2}$$

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

For a given value of $\boldsymbol{\delta}^m$ and $\boldsymbol{\upsilon}^m$ at step m, the iteration goes as follows:

$$(\boldsymbol{\eta}^{m+1},\boldsymbol{\beta}^{m+1}) = \underset{\boldsymbol{\eta},\boldsymbol{\beta}}{\operatorname{argmin}} L(\boldsymbol{\eta},\boldsymbol{\beta},\boldsymbol{\delta}^m,\boldsymbol{\upsilon}^m), \tag{3.3}$$
$$\boldsymbol{\delta}^{m+1} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} L(\boldsymbol{\eta}^{m+1},\boldsymbol{\beta}^{m+1},\boldsymbol{\delta},\boldsymbol{\upsilon}^m), \tag{3.4}$$
$$\boldsymbol{\upsilon}_{ij}^{m+1} = \boldsymbol{\upsilon}_{ij}^m + \vartheta(\boldsymbol{\beta}_i^{m+1} - \boldsymbol{\beta}_j^{m+1} - \boldsymbol{\delta}_{ij}^{m+1}). \tag{3.5}$$

In (3.3), the problem is equivalent to the minimization of the function

$$f(\boldsymbol{\eta},\boldsymbol{\beta}) = \frac{1}{2}\sum\nolimits_{i=1}^{n}(y_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i)^2 + \frac{\vartheta}{2}\sum\nolimits_{i<j}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}^m + \vartheta^{-1}\upsilon_{ij}^m\|^2 + C,$$

where C is a constant independent of $(\boldsymbol{\eta},\boldsymbol{\beta})$. Let $\mathbf{y} = (y_1,\ldots,y_n)^{\mathrm{T}}$, $\mathbf{Z} = (\boldsymbol{z}_1,\ldots,\boldsymbol{z}_n)^{\mathrm{T}}$, and $\mathbf{X} = \mathrm{diag}(\boldsymbol{x}_1^{\mathrm{T}},\ldots,\boldsymbol{x}_n^{\mathrm{T}})$. Some algebra shows that we can write $f(\boldsymbol{\eta},\boldsymbol{\beta})$ as

$$f(\boldsymbol{\eta},\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{Z}\boldsymbol{\eta} + \mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 + \frac{\vartheta}{2}\|\mathbf{A}\boldsymbol{\beta} - \boldsymbol{\delta}^m + \vartheta^{-1}\boldsymbol{\upsilon}^m\|^2 + C, \tag{3.6}$$

where $\mathbf{A} = D \otimes \mathbf{I}_p$. Here $D = \{(e_i - e_j), i<j\}^{\mathrm{T}}$ with e_i being the ith unit $n\times 1$ vector whose ith element is 1 and the remaining ones are 0, $\mathbf{I}_p$ is a $p\times p$ identity matrix and $\otimes$ is the Kronecker product.

Thus for given $\boldsymbol{\delta}^m$ and $\boldsymbol{\upsilon}^m$ at the mth step, the updates $\boldsymbol{\beta}^{m+1}$ and $\boldsymbol{\eta}^{m+1}$ are

$$\begin{aligned}\boldsymbol{\beta}^{m+1} &= (\mathbf{X}^{\mathrm{T}}\mathbf{Q}_Z\mathbf{X} + \vartheta\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}[\mathbf{X}^{\mathrm{T}}\mathbf{Q}_Z\mathbf{y} + \vartheta\mathbf{A}^{\mathrm{T}}(\boldsymbol{\delta}^m - \vartheta^{-1}\boldsymbol{\upsilon}^m)],\\ \boldsymbol{\eta}^{m+1} &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{m+1}),\end{aligned} \tag{3.7}$$

where $\mathbf{Q}_Z = \mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}$. Since

$$\mathbf{A}^{\mathrm{T}}(\boldsymbol{\delta}^m - \vartheta^{-1}\boldsymbol{\upsilon}^m) = (D^{\mathrm{T}}\otimes\mathbf{I}_p)(\boldsymbol{\delta}^m - \vartheta^{-1}\boldsymbol{\upsilon}^m) = \mathrm{vec}((\boldsymbol{\Delta}^m - \vartheta^{-1}\boldsymbol{\Upsilon}^m)D),$$

where $\boldsymbol{\Delta}^m = \{\boldsymbol{\delta}_{ij}^m, i<j\}_{p\times n(n-1)/2}$ and $\boldsymbol{\Upsilon}^m = \{\boldsymbol{\upsilon}_{ij}^m, i<j\}_{p\times n(n-1)/2}$, then we have

$$\boldsymbol{\beta}^{m+1} = (\mathbf{X}^{\mathrm{T}}\mathbf{Q}_Z\mathbf{X} + \vartheta\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}[\mathbf{X}^{\mathrm{T}}\mathbf{Q}_Z\mathbf{y} + \vartheta\mathrm{vec}((\boldsymbol{\Delta}^m - \vartheta^{-1}\boldsymbol{\Upsilon}^m)D)]. \tag{3.8}$$

In (3.4), after discarding the terms independent of $\boldsymbol{\delta}$, we need to minimize

$$\frac{\vartheta}{2}\|\boldsymbol{\zeta}_{ij}^m - \boldsymbol{\delta}_{ij}\|^2 + p_\gamma(\|\boldsymbol{\delta}_{ij}\|, \lambda) \tag{3.9}$$

with respect to $\boldsymbol{\delta}_{ij}$, where $\boldsymbol{\zeta}_{ij}^m = \boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m + \vartheta^{-1}\boldsymbol{\upsilon}_{ij}^m$. This is a groupwise thresholding operator corresponding to p_γ.

For the L_1 penalty, the solution is

$$\boldsymbol{\delta}_{ij}^{m+1} = S(\boldsymbol{\zeta}_{ij}^m, \lambda/\vartheta), \tag{3.10}$$

where $S(\boldsymbol{z}, t) = (1 - t/\|\boldsymbol{z}\|)_+\boldsymbol{z}$ is the groupwise soft thresholding operator. Here $(x)_+ = x$ if $x > 0$ and $= 0$, otherwise.

For the MCP with $\gamma > 1/\vartheta$, the solution is

$$\boldsymbol{\delta}_{ij}^{m+1} = \begin{cases} \frac{S(\boldsymbol{\zeta}_{ij}^m, \lambda/\vartheta)}{1-1/(\gamma\vartheta)} & \text{if } \|\boldsymbol{\zeta}_{ij}^m\| \le \gamma\lambda, \\ \boldsymbol{\zeta}_{ij} & \text{if } \|\boldsymbol{\zeta}_{ij}^m\| > \gamma\lambda. \end{cases} \tag{3.11}$$

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

$$\boldsymbol{\delta}_{ij}^{m+1} = \begin{cases} S(\boldsymbol{\zeta}_{ij}^m, \lambda/\vartheta) & \text{if } \| \boldsymbol{\zeta}_{ij}^m\| \le \lambda + \lambda/\vartheta, \\ \frac{S(\boldsymbol{\zeta}_{ij}^m, \gamma\lambda/((\gamma-1)\vartheta))}{1-1/((\gamma-1)\vartheta)} & \text{if } \lambda + \lambda/\vartheta < \ \|\boldsymbol{\zeta}_{ij}^m\| \le \gamma\lambda, \\ \boldsymbol{\zeta}_{ij}^m & \text{if } \|\boldsymbol{\zeta}_{ij}^m\| > \gamma\lambda. \end{cases} \tag{3.12}$$

Finally, the update of $\boldsymbol{\upsilon}_{ij}$ is given in (3.5).

We summarize the above analysis in Algorithm 1.

Algorithm 1 ADMM for concave fusion

Require: Initialize $\boldsymbol{\delta}^0$, $\boldsymbol{\upsilon}^0$.

1: **for** $m = 0, 1, 2, \cdots$ **do**
2: Compute $\boldsymbol{\beta}^{m+1}$ using (3.8)
3: Compute $\boldsymbol{\eta}^{m+1}$ (3.7)
4: Compute $\boldsymbol{\delta}^{m+1}$ (3.10), (3.11) or (3.12)
5: Compute $\boldsymbol{\upsilon}^{m+1}$ using (3.5)
6: **if** convergence criterion is met, **then**
7: Stop and denote the last iteration by $(\widehat{\boldsymbol{\eta}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$,
8: **else**
9: $m = m + 1$.
10: **end if**
11: **end for**

Ensure: Output

Remark 3.1 *Our algorithm enables us to have* $\widehat{\boldsymbol{\delta}}_{ij} = \mathbf{0}$ *for a sufficiently large* λ. *We put observations* i *and* j *in the group with the same treatment effect if* $\widehat{\boldsymbol{\delta}}_{ij} = \mathbf{0}$. *As a result, we have* $\widehat{K}$ *estimated groups* $\widehat{\mathcal{G}}_1, \ldots, \widehat{\mathcal{G}}_{\widehat{K}}$. *The estimated treatment effect for the* k^{th} *group is* $\widehat{\boldsymbol{\alpha}}_k = |\widehat{\mathcal{G}}_k|^{-1} \sum_{i \in \widehat{\mathcal{G}}_k} \widehat{\boldsymbol{\beta}}_i$, *where* $|\widehat{\mathcal{G}}_k|$ *is the cardinality of* $\widehat{\mathcal{G}}_k$.

Remark 3.2 *In the algorithm, we require the invertibility of* $\mathbf{X}^T\mathbf{Q}_Z\mathbf{X} + \vartheta\mathbf{A}^T\mathbf{A}$. *It can be derived that* $\mathbf{A}^T\mathbf{A} = n\mathbf{I}_{np} - (1_n \otimes \mathbf{I}_p)(1_n \otimes \mathbf{I}_p)^T$. *For any nonzero vector* $\mathbf{a} = (a_{ij}, 1 \le i \le n, 1 \le j \le p)^T \in \mathrm{R}^{np}$, *we have* $\mathbf{a}^T(\vartheta\mathbf{A}^T\mathbf{A})\mathbf{a} \ge 0$ *and* $\mathbf{a}^T(\mathbf{X}^T\mathbf{Q}_Z\mathbf{X})\mathbf{a} \ge 0$. *Note that* $\mathbf{a}^T(\vartheta\mathbf{A}^T\mathbf{A})\mathbf{a} = 0$ *if and only if* $a_{ij} = a_j$ *for all* i. *When* $a_{ij} = a_j$ *for all* i, *we have* $\mathbf{a}^T(\mathbf{X}^T\mathbf{Q}_Z\mathbf{X})\mathbf{a} > 0$ *given that* $\lambda_{\min}(\sum_{i=1}^n (\boldsymbol{x}_i^T, \boldsymbol{z}_i^T)^T(\boldsymbol{x}_i^T, \boldsymbol{z}_i^T)) > 0$, *which is a common assumption that the design matrix needs to satisfy in linear regression. Therefore,* $\mathbf{X}^T\mathbf{Q}_Z\mathbf{X} + \vartheta\mathbf{A}^T\mathbf{A}$ *is invertible.*

Remark 3.3 *It is worth noting that the algorithm can be applied to find the estimate of parameter in the model* $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}_i + \varepsilon_i, i = 1, \ldots, n$ *by letting* $\mathbf{Q}_Z = \mathbf{I}_n$.

Remark 3.4 *We track the progress of the ADMM based on the primal residual* $\mathbf{r}^{m+1} = \mathbf{A}\boldsymbol{\beta}^{m+1} - \boldsymbol{\delta}^{m+1}$. *We stop the algorithm when* $\mathbf{r}^{m+1}$ *is close to zero such that* $\|\mathbf{r}^{m+1}\| < a$ *for some small value* a.

3.2 Initial value and computation of the solution path

To start the ADMM algorithm described above, it is important to find a reasonable initial value. For this purpose, we consider the ridge fusion criterion given by

$$L_R(\boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{Z}\boldsymbol{\eta} + \mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 + \frac{\lambda^*}{2} \sum_{1 \le i < j \le n} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|^2,$$

where λ^* is the tuning parameter having a small value. We use $\lambda^* = 0.001$ in our analysis. Then $L_R(\boldsymbol{\eta}, \boldsymbol{\beta})$ can be written as

$$L_R(\boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{Z}\boldsymbol{\eta} + \mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 + \frac{\lambda^*}{2}\|\mathbf{A}\boldsymbol{\beta}\|^2,$$

where $\boldsymbol{A}$ is defined in (3.6). The solutions are

$$\begin{aligned} \boldsymbol{\beta}_R(\lambda^*) &= (\boldsymbol{\beta}_{R,1}^{\mathrm{T}}(\lambda^*), \ldots, \boldsymbol{\beta}_{R,n}^{\mathrm{T}}(\lambda^*))^{\mathrm{T}} = (\mathbf{X}^{\mathrm{T}}\mathbf{Q}_Z\mathbf{X} + \lambda^*\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Q}_Z\mathbf{y}, \\ \boldsymbol{\eta}_R(\lambda^*) &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_R(\lambda^*)), \end{aligned}$$

where $\mathbf{Q}_Z$ is given in (3.8). Next, we assign the subjects to K^* groups by ranking the median values of $\boldsymbol{\beta}_{R,i}^{\mathrm{T}}(\lambda^*)$. We let $K^* = \lfloor n^{1/2} \rfloor$ to ensure that it is sufficiently large, where $\lfloor a \rfloor$ denotes the largest integer no greater than a. We then find the initial estimates $\boldsymbol{\eta}^0$ and

$\boldsymbol{\beta}^0$ from least squares regression with K^* groups. Let the initial estimates $\boldsymbol{\delta}_{ij}^0 = \boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0$ and $\boldsymbol{\upsilon}^0 = \mathbf{0}$.

To compute the solution path of $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ along the λ values, we use the warm start and continuation strategy to update the solutions. Let $[\lambda_{\min}, \lambda_{\max}]$ be the interval on which we compute the solution path, where $0 \leq \lambda_{\min} < \lambda_{\max} < \infty$. Let $\lambda_{\min} = \lambda_0 < \lambda_1 < \cdots < \lambda_K \equiv \lambda_{\max}$ be a grid of λ values in $[\lambda_{\min}, \lambda_{\max}]$. Compute $(\widehat{\boldsymbol{\eta}}(\lambda_0), \hat{\boldsymbol{\beta}}(\lambda_0))$ using $(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0)$ as the initial value. Then compute $(\widehat{\boldsymbol{\eta}}(\lambda_k), \hat{\boldsymbol{\beta}}(\lambda_k))$ using $(\widehat{\boldsymbol{\eta}}(\lambda_{k-1}), \hat{\boldsymbol{\beta}}(\lambda_{k-1}))$ as the initial value for each $k = 1, \ldots, K$.

Note that we start from the smallest λ value in computing the solution path. This is different from the coordinate descent algorithms for computing the solution path in penalized regression problems (Friedman et al., 2007), where the algorithms start at the λ value that forces all the coefficients to zero.

3.3 Convergence of the algorithm

We next derive the convergence properties of the ADMM algorithm. We show that the primal feasibility and dual feasibility are achieved by the algorithm.

Proposition 3.1 *Let* $\mathbf{r}^m = \mathbf{A}\boldsymbol{\beta}^m - \boldsymbol{\delta}^m$ *and* $\mathbf{s}^{m+1} = \vartheta \mathbf{A}^T(\boldsymbol{\delta}^{m+1} - \boldsymbol{\delta}^m)$ *be the primal residual and dual residual in the ADMM described above, respectively. It holds that* $\lim_{m\to\infty} \|\mathbf{r}^m\|^2 = 0$ *and* $\lim_{m\to\infty} \|\mathbf{s}^m\|^2 = 0$ *for the MCP and SCAD penalties.*

Proposition 3.1 shows that the ADMM algorithm 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

In this section, we study the theoretical properties of the proposed estimator. Specifically, we provide sufficient conditions under which there exists a local minimizer of the objective function equal to the oracle least squares estimator with *a priori* knowledge of the true groups with high probability. We also derive the lower bound of the minimum difference of coefficients between subgroups in order to be able to estimate the subgroup-specific treatment effects.

4.1 Notation and conditions

Let $\widetilde{\mathbf{W}} = \{w_{ik}\}$ be an $n \times K$ matrix with $w_{ik} = 1$ for $i \in \mathcal{G}_k$ and $w_{ik} = 0$ otherwise. Let $\mathbf{W} = \widetilde{\mathbf{W}} \otimes \mathbf{I}_p$. Let $\mathcal{M}_{\mathcal{G}} = \{\boldsymbol{\beta} \in \mathbb{R}^{np} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_j, \text{ for any } i, j \in \mathcal{G}_k, 1 \leq k \leq K\}$. For each

$\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}$, it can be written as $\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_k$ is a $p \times 1$ vector of the kth subgroup-specific parameter for $k = 1, \ldots, K$. Simple calculation shows

$$\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathrm{diag}(|\mathcal{G}_1|, \ldots, |\mathcal{G}_K|) \otimes \mathbf{I}_p,$$

where $|\mathcal{G}_k|$ denotes the number of elements in $\mathcal{G}_k$. Denote the minimum and maximum group sizes by $|\mathcal{G}_{\min}| = \min_{1 \le k \le K} |\mathcal{G}_k|$ and $|\mathcal{G}_{\max}| = \max_{1 \le k \le K} |\mathcal{G}_k|$, respectively. For any positive numbers a_n and b_n, let $a_n \gg b_n$ denote $a_n^{-1} b_n = o(1)$. For any vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_s)^{\mathrm{T}} \in R^s$, let $\|\boldsymbol{\zeta}\|_\infty = \max_{1 \le l \le s} |\zeta_l|$. For any symmetric matrix $\mathbf{A}_{s \times s}$, denote its L_2 norm by $\|\mathbf{A}\| = \max_{\boldsymbol{\zeta} \in R^s, \|\boldsymbol{\zeta}\| = 1} \|\mathbf{A}\boldsymbol{\zeta}\|$, and 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 $\|\mathbf{A}\|_\infty = \max_{1 \le i \le s} \sum_{j=1}^{t} |A_{ij}|$. Let $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{W}$ and $\mathbf{U} = (\mathbf{Z}, \mathbf{X}\mathbf{W})$. Finally, denote the scaled penalty function by

$$\rho(t) = \lambda^{-1} p_\gamma(t, \lambda).$$

We make the following basic assumptions.

(C1) The function $\rho_\gamma(t)$ is a symmetric, non-decreasing and concave on $[0, \infty)$. It is constant for $t \ge a\lambda$ for some constant $a > 0$, and $\rho(0) = 0$. In addition, $\rho'(t)$ exists and is continuous except for a finite number values of t and $\rho'(0+) = 1$.

(C2) The noise vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ has sub-Gaussian tails such that $P(|\mathbf{a}^{\mathrm{T}}\boldsymbol{\varepsilon}| > \|\mathbf{a}\| x) \le 2\exp(-c_1 x^2)$ for any vector $\mathbf{a} \in \mathbb{R}^n$ and $x > 0$, where $0 < c_1 < \infty$.

(C3) Assume $\sum_{i=1}^n z_{il}^2 = n$ for $1 \le l \le q$, and $\sum_{i=1}^n x_{ij}^2 1\{i \in \mathcal{G}_k\} = |\mathcal{G}_k|$ for $1 \le j \le p$, $\lambda_{\min}(\mathbf{U}^{\mathrm{T}}\mathbf{U}) \ge C_1 |\mathcal{G}_{\min}|$, $\lambda_{\max}(\mathbf{U}^{\mathrm{T}}\mathbf{U}) \le C_1' n$, $\sup_i \|\boldsymbol{x}_i\| \le C_2\sqrt{p}$ and $\sup_i \|\boldsymbol{z}_i\| \le C_3\sqrt{q}$ for some constants $0 < C_1 < \infty$, $0 < C_1' < \infty$, $0 < C_2 < \infty$ and $0 < C_3 < \infty$.

Conditions (C1) and (C2) are common assumptions in penalized regression in high-dimensional settings. The concave penalties such as MCP and SCAD satisfy (C1). In the literature, it is commonly assumed that the smallest eigenvalue of the transpose of the design matrix multiplied by the design matrix is bounded by $C_1 n$, which may not hold for $\mathbf{U}^{\mathrm{T}}\mathbf{U}$. By some calculation and $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{W}$, we have

$$\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}} = \mathrm{diag}(\textstyle\sum_{i \in \mathcal{G}_k} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}, k = 1, \ldots, K).$$

By assuming that $\lambda_{\min}(\sum_{i \in \mathcal{G}_k} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}) \ge c |\mathcal{G}_k|$ for some constant $0 < c < \infty$, we have $\lambda_{\min}(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}} \ge c |\mathcal{G}_{\min}|$. If $\mathbf{Z}^{\mathrm{T}}\widetilde{\mathbf{X}} = 0$ and $\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}) \ge Cn$, we have

$$\lambda_{\min}(\mathbf{U}^{\mathrm{T}}\mathbf{U}) = \min\{\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}), \lambda_{\min}(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})\} \ge \min(c |\mathcal{G}_{\min}|, Cn),$$

and $|\mathcal{G}_{\min}| \le n/K$. Therefore, we let the smallest eigenvalue in Condition (C3) be bounded below by $C_1 |\mathcal{G}_{\min}|$.

4.2 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, that is, $K \geq 2$. If the underlying groups $\mathcal{G}_1, \ldots, \mathcal{G}_K$ were known, the oracle estimator of $(\boldsymbol{\eta}, \boldsymbol{\beta})$ would be

$$(\widehat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or}) = \underset{\boldsymbol{\eta}\in\mathbb{R}^q, \boldsymbol{\beta}\in\mathcal{M}_{\mathcal{G}}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}\|^2. \tag{4.1}$$

Since $\boldsymbol{\beta} = \widetilde{\mathbf{W}}\boldsymbol{\alpha}$, the oracle estimators for the common coefficient $\boldsymbol{\alpha}$ and the coefficients $\boldsymbol{\eta}$ are

$$\begin{aligned}(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or}) &= \underset{\boldsymbol{\eta}\in\mathbb{R}^q, \boldsymbol{\alpha}\in\mathbb{R}^{Kp}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \widetilde{\mathbf{X}}\boldsymbol{\alpha}\|^2 \\ &= (\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{y}.\end{aligned}$$

Let $\boldsymbol{\alpha}_k^0$ be the true common coefficient vector for group $\mathcal{G}_k$, $k = 1, \ldots, K$ and $\boldsymbol{\alpha}^0 = ((\boldsymbol{\alpha}_k^0)^{\mathrm{T}}, k = 1, \ldots, K)^{\mathrm{T}}$. Of course, oracle estimators are not real estimators, they are theoretical constructions useful for stating the properties of the proposed estimators.

Theorem 4.1 *Suppose*

$$|\mathcal{G}_{\min}| \gg (q + Kp)^{1/2} n^{3/4}.$$

Then under Conditions (C1)-(C3), we have with probability at least $1 - 2(Kp + q + 1)n^{-1}$,

$$\left\|((\widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^T, (\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^T)^T\right\| \leq \phi_n, \tag{4.2}$$

and

$$\left\|\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0\right\| \leq \sqrt{|\mathcal{G}_{\max}|}\phi_n, \ \sup_i \left\|\widehat{\boldsymbol{\beta}}_i^{or} - \boldsymbol{\beta}_i^0\right\| \leq \phi_n,$$

where

$$\phi_n = c_1^{-1/2} C_1^{-1} \sqrt{q + Kp}\, |\mathcal{G}_{\min}|^{-1} \sqrt{n \log n}. \tag{4.3}$$

Moreover, for any vector $\mathbf{a}_n \in \mathbb{R}^{q+Kp}$ *with* $||\mathbf{a}_n|| = 1$, *we have as* $n \to \infty$,

$$\sigma_n(\mathbf{a}_n)^{-1}\mathbf{a}_n^T((\widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^T, (\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^T)^T \to_D N(0, 1), \tag{4.4}$$

where

$$\sigma_n(\mathbf{a}_n) = \sigma \left[\mathbf{a}_n^T(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{a}_n\right]^{1/2}. \tag{4.5}$$

Remark 4.1 *Since* $|\mathcal{G}_{\min}| \leq n/K$, *by the condition* $|\mathcal{G}_{\min}| \gg (q + Kp)^{1/2} n^{3/4}$, *then* q, K *and* p *must satisfy* $K\sqrt{q + Kp} = o\{(n)^{1/4}\}$, *and hence* $K = o(n^{1/6})$. *Thus in this theorem, the number of subgroups* K *is required to grow slower than* $n^{1/6}$.

Remark 4.2 *By letting* $|\mathcal{G}_{\min}| = \delta n/K$ *for some constant* $0 < \delta \leq 1$*, the bound (4.2) is* $\phi_n = c_1^{-1/2} C_1^{-1} \delta^{-1} K \sqrt{q + Kp} \sqrt{\log n/n}$*. Moreover, if* q*,* K *and* p *are fixed quantities, then* $\phi_n = C^* \sqrt{\log n/n}$ *for some constant* $0 < C^* < \infty$*.*

Let

$$b_n = \min_{i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}, k \neq k'} \|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\| = \min_{k \neq k'} \|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_{k'}^0\|$$

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

Theorem 4.2 *Suppose the conditions in Theorem 4.1 hold. If* $b_n > a\lambda$ *and* $\lambda \gg \phi_n$*, for some constant* $a > 0$*, where* ϕ_n *is given in (4.3), then there exists a local minimizer* $(\widehat{\boldsymbol{\eta}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$ *of the objective function* $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}; \lambda)$ *given in (2.3) satisfying*

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

Remark 4.3 *Theorem 4.2 shows that the oracle estimator* $(\widehat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ *is a local minimizer of the objective function with a high probability, and thus the true groups can be recovered with the estimated common value for group* k *given as* $\widehat{\boldsymbol{\alpha}}_k(\lambda) = \hat{\boldsymbol{\beta}}_i^{or}$ *for* $i \in \mathcal{G}_k$*. This result holds given that* $b_n \gg \phi_n$*. As discussed in Remark 4.2, when* K*,* p *and* q *are finite and fixed numbers and* $|\mathcal{G}_{\min}| = \delta n/K$ *for some constant* $0 < \delta \leq 1$*,* $b_n \gg C^* \sqrt{\log n/n}$ *for some constant* $0 < C^* < \infty$*.*

Let $\widehat{\boldsymbol{\alpha}}(\lambda) = (\widehat{\boldsymbol{\alpha}}_1(\lambda)^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\alpha}}_K(\lambda)^{\mathrm{T}})^{\mathrm{T}}$ be the estimated treatment effects such that $\widehat{\boldsymbol{\alpha}}_k(\lambda) = \widehat{\boldsymbol{\beta}}_i(\lambda)$ for $i \in \mathcal{G}_k$, where $k = 1, \ldots, K$, and $\widehat{\boldsymbol{\beta}}(\lambda) = \{\widehat{\boldsymbol{\beta}}_i(\lambda)^{\mathrm{T}}, 1 \leq i \leq n\}^{\mathrm{T}}$ is the local minimizer given in Theorem 4.2. Based on the results in Theorems 4.1 and 4.2, we obtain the asymptotic distribution of $(\widehat{\boldsymbol{\eta}}(\lambda)^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}(\lambda)^{\mathrm{T}})^{\mathrm{T}}$ given in the following corollary.

Corollary 4.1 *Under the conditions in Theorem 4.2, we have for any* $\mathbf{a}_n \in \mathbb{R}^{q+Kp}$ *with* $||\mathbf{a}_n|| = 1$*, as* $n \to \infty$*,*

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

with $\sigma_n(\mathbf{a}_n)$ *given in (4.5). As a result, we have for any vectors* $\mathbf{a}_{n1} \in \mathbb{R}^q$ *with* $||\mathbf{a}_{n1}|| = 1$ *and* $\mathbf{a}_{n2} \in \mathbb{R}^{Kp}$ $||\mathbf{a}_{n2}|| = 1$*, as* $n \to \infty$*,*

$$\sigma_{n1}^{-1}(\mathbf{a}_{n1}) \mathbf{a}_{n1}^T (\widehat{\boldsymbol{\eta}}(\lambda) - \boldsymbol{\eta}^0) \to_D N(0, 1) \text{ and } \sigma_{n2}^{-1}(\mathbf{a}_{n2}) \mathbf{a}_{n2}^T (\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}^0) \to_D N(0, 1),$$

where

$$\begin{aligned} \sigma_{n1}(\mathbf{a}_{n1}) &= \sigma \left[\mathbf{a}_{n1}^T [\mathbf{Z}^T \mathbf{Z} - \mathbf{Z}^T \widetilde{\mathbf{X}} (\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T \mathbf{Z}]^{-1} \mathbf{a}_{n1} \right]^{1/2}, \\ \sigma_{n2}(\mathbf{a}_{n2}) &= \sigma \left[\mathbf{a}_{n2}^T [\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \widetilde{\mathbf{X}}]^{-1} \mathbf{a}_{n2} \right]^{1/2}. \end{aligned}$$

Remark 4.4 *The asymptotic distribution of the penalized estimators provides a theoretical justification for further conducting statistical inference about heterogeneity. By the results in Corollary 4.1, for given* $\mathbf{a}_{n1} \in \mathbb{R}^q$ *and* $\mathbf{a}_{n2} \in \mathbb{R}^{Kp}$, $100(1-\alpha)\%$ *confidence intervals for* $\mathbf{a}_{n1}^T\boldsymbol{\eta}^0$ *and* $\mathbf{a}_{n2}^T\boldsymbol{\alpha}^0$ *are given by*

$$\mathbf{a}_{n1}^T\widehat{\boldsymbol{\eta}}(\lambda) \pm z_{\alpha/2}\widehat{\sigma}_{n1}(\mathbf{a}_{n1}) \quad \textit{and} \quad \mathbf{a}_{n2}^T\widehat{\boldsymbol{\alpha}}(\lambda) \pm z_{\alpha/2}\widehat{\sigma}_{n2}(\mathbf{a}_{n2}),$$

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*

$$\widehat{\sigma}^2 = (n - q - \widehat{K}p)^{-1}\sum\nolimits_{i=1}^{n}(y_i - \boldsymbol{z}_i^T\widehat{\boldsymbol{\eta}} - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_i)^2,$$

where $\widehat{K}$ *is the estimated number of subgroups satisfying* $P(\widehat{K} = K) \to 1$ *from the oracle property in Theorem 4.2.*

4.3 Homogeneous model

When the true model is the homogeneous model given as $y_i = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\alpha} + \varepsilon_i, i = 1, \ldots, n$, we have $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_n = \boldsymbol{\alpha}$ and $K = 1$. The penalized estimator $(\widehat{\boldsymbol{\eta}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$ of $(\boldsymbol{\eta}, \boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_n^{\mathrm{T}})^{\mathrm{T}}$, also has the oracle property given as follows. We define the oracle estimator for $(\boldsymbol{\eta}, \boldsymbol{\alpha})$ as

$$\begin{aligned}(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or}) &= \underset{\boldsymbol{\eta}\in\mathbb{R}^q, \boldsymbol{\alpha}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{x}\boldsymbol{\alpha}\|^2 \\ &= (\mathbf{U}^{*\mathrm{T}}\mathbf{U}^*)^{-1}\mathbf{U}^{*\mathrm{T}}\mathbf{y}.\end{aligned}$$

where $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$ and $\mathbf{U}^* = (\mathbf{Z}, \mathbf{x})$. Let $\widehat{\boldsymbol{\beta}}^{or} = (\widehat{\boldsymbol{\beta}}_1^{or\mathrm{T}}, \ldots, \widehat{\boldsymbol{\beta}}_n^{or\mathrm{T}})^{\mathrm{T}}$, where $\widehat{\boldsymbol{\beta}}_i^{or} = \widehat{\boldsymbol{\alpha}}^{or}$ for all i. Let $\boldsymbol{\eta}^0$ and $\boldsymbol{\alpha}^0$ be the true coefficient vectors. We introduce the following condition.

(C3*) Assume $\sum_{i=1}^n z_{il}^2 = n$ for $1 \le l \le q$, and $\sum_{i=1}^n x_{ij}^2 = n$ for $1 \le j \le p$, $\lambda_{\min}(\mathbf{U}^{*\mathrm{T}}\mathbf{U}^*) \ge C_1 n$, $\lambda_{\max}(\mathbf{U}^{*\mathrm{T}}\mathbf{U}^*) \le C_1' n$, $\sup_i \|\mathbf{x}_i\| \le C_2\sqrt{p}$ and $\sup_i \|\boldsymbol{z}_i\| \le C_3\sqrt{q}$ for some constants $0 < C_1 \le C_1' < \infty$, $0 < C_2 < \infty$ and $0 < C_3 < \infty$.

Theorem 4.3 *Suppose Conditions (C1*), (C2) and (C3) hold. If* $p = o(n^{1/2})$ *and* $q = o(n^{1/2})$, *the oracle esitmator has the property that with probability at least* $1 - 2(p+q+1)n^{-1}$,

$$\begin{aligned}\left\|((\widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^T, (\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^T)^T\right\| &\le \phi_n, \\ \sup_i \left\|\widehat{\boldsymbol{\beta}}_i^{or} - \boldsymbol{\beta}_i^0\right\| &\le \phi_n, \end{aligned} \tag{4.6}$$

where

$$\phi_n = c_1^{-1/2}C_1^{-1}\sqrt{q+p}\sqrt{n^{-1}\log n},$$

in which c_1 and C_1 are given in Conditions (C2) and (C3), respectively, and for any vector $\mathbf{a}_n \in \mathbb{R}^{q+p}$ with $||\mathbf{a}_n|| = 1$, as $n \to \infty$,*

$$\sigma_n(\mathbf{a}_n)^{-1}\mathbf{a}_n^T((\widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^T, (\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^T)^T \to N(0,1), \tag{4.7}$$

where

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

Moreover, if $\lambda \gg \phi_n$, then there exists a local minimizer $(\widehat{\boldsymbol{\eta}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda))$ of the objective function $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}; \lambda)$ given in (2.3) satisfying

$$P\left((\widehat{\boldsymbol{\eta}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda)) = (\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or})\right) \to 1. \tag{4.8}$$

Remark 4.5 *By Theorem 4.3, the local minimizer $\widehat{\boldsymbol{\beta}}_i(\lambda) = \widehat{\boldsymbol{\alpha}}(\lambda) = \widehat{\boldsymbol{\alpha}}^{or}$ for all i. Then, we have for any vectors $\mathbf{a}_{n1} \in \mathbb{R}^q$ with $||\mathbf{a}_{n1}|| = 1$ and $\mathbf{a}_{n2} \in \mathbb{R}^p$ with $||\mathbf{a}_{n2}|| = 1$, as $n \to \infty$,*

$$\sigma_{n1}^{-1}(\mathbf{a}_{n1})\mathbf{a}_{n1}^T(\widehat{\boldsymbol{\eta}}(\lambda) - \boldsymbol{\eta}^0) \to_D N(0,1) \text{ and } \sigma_{n2}^{-1}(\mathbf{a}_{n2})\mathbf{a}_{n2}^T(\widehat{\boldsymbol{\alpha}}(\lambda) - \boldsymbol{\alpha}^0) \to_D N(0,1),$$

where

$$\begin{aligned}
\sigma_{n1}(\mathbf{a}_{n1}) &= \sigma \left[\mathbf{a}_{n1}^T[\mathbf{Z}^T\mathbf{Z} - \mathbf{Z}^T\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Z}]^{-1}\mathbf{a}_{n1}\right]^{1/2}, \\
\sigma_{n2}(\mathbf{a}_{n2}) &= \sigma \left[\mathbf{a}_{n2}^T[\mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{x}]^{-1}\mathbf{a}_{n2}\right]^{1/2}.
\end{aligned}$$

5 Simulation studies

We use the modified Bayes Information Criterion (BIC) (Wang et al., 2007) for high-dimensional data settings to select the tuning parameter by minimizing

$$\text{BIC}(\lambda) = \log[\textstyle\sum_{i=1}^n (y_i - \boldsymbol{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\eta}}(\lambda) - \boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_i(\lambda))^2/n] + C_n\frac{\log n}{n}(\widehat{K}(\lambda)p + q), \tag{5.1}$$

where C_n is a positive number which can depend on n. When $C_n = 1$, the modified BIC reduces to the traditional BIC (Schwarz, 1978). Following Lee et al. (2014), we use $C_n = \log(np + q)$. We select λ by minimizing the modified BIC.

Example 1 (One treatment variable). We simulate data from the heterogeneous model with one treatment variable:

$$y_i = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} + x_i\beta_i + \varepsilon_i, i = 1, \ldots, n, \tag{5.2}$$

where $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{i5})^{\mathrm{T}}$ with $z_{i1} = 1$ and $(z_{i2}, \ldots, z_{i5})^{\mathrm{T}}$ simulated from multivariate normal with mean 0, variance 1 and an exchangeable correlation $\rho = 0.3$, x_i is simulated from standard normal, and the error terms ε_i are from independent $N(0, 0.5^2)$. Let $\boldsymbol{\eta} =$

$(\eta_1, \ldots, \eta_5)^{\mathrm{T}}$ with η_k simulated from Uniform$[1, 2]$ for $k = 1, \ldots, 5$. We randomly assign the treatment coefficients to two groups with equal probabilities, i.e., we let $p(i \in \mathcal{G}_1) = p(i \in \mathcal{G}_2) = 1/2$, so that $\beta_i = \alpha_1$ for $i \in \mathcal{G}_1$ and $\beta_i = \alpha_2$ for $i \in \mathcal{G}_2$, where $\alpha_1 = 2$ and $\alpha_2 = -2$. We use different sample sizes by letting $n = 100, 200$. We fix $\vartheta = 1$ and $\gamma = 3$. We compare the performance of the estimators using the two concave penalties (MCP and SCAD) and the LASSO penalty.

Table 1: The sample mean, median and standard deviation (s.d.) of $\widehat{K}$ and the percentage (per) of $\widehat{K}$ equaling to the true number of subgroups by MCP and SCAD based on 100 replications with $n = 100, 200$ in Example 1.

	$n = 100$				$n = 200$			
	mean	median	s.d.	per	mean	median	s.d.	per
MCP	2.380	2.000	0.716	0.710	2.210	2.000	0.520	0.790
SCAD	2.340	2.000	0.708	0.710	2.210	2.000	0.541	0.800

We select the λ value by minimizing the modified BIC given in (5.1). Table 1 reports the sample mean, median and standard deviation (s.d.) of the estimated number of groups $\widehat{K}$ and the percentage of $\widehat{K}$ equaling to the true number of subgroups by the MCP and SCAD methods based on 100 simulation realizations with $n = 100, 200$. The median of $\widehat{K}$ is 2 which is the true number of subgroups for all cases. As n increases, the mean gets closer to 2 and the standard deviation becomes smaller. Moreover, the percentage of correctly selecting the number of subgroups becomes larger as n increases.

Without considering the possible heterogeneity in treatment effects, the estimates from ordinary least squares (OLS) can be misleading. To demonstrate this point, in Figure 2, we plot the values of $x_i\beta_i$ (black solid lines), $x_i\widehat{\beta}_i$ (red dashed lines) and $x_i\widehat{\beta}^{\mathrm{ols}}$(blue dotted lines) against values of x_i by using the 79 replications which have two estimated groups by the MCP method for $n = 200$, where β_i are the true values, $\widehat{\beta}_i$ are the estimated values by MCP and $\widehat{\beta}^{\mathrm{ols}}$ is the estimated value from OLS. We observe that the fitted lines by the MCP are close to the true lines. However, the fitted lines by the OLS center around the line $y = 0$, which are far away from the true lines.

To further study the estimation accuracy and evaluate the asymptotic properties stated in Section 4, Table 2 presents the sample mean, median and asymptotic standard deviation (ASD) obtained according to Corollary 4.1 of the estimators $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ by the MCP and SCAD methods and oracle estimators $\widehat{\alpha}_1^{or}$ and $\widehat{\alpha}_2^{or}$ based on 100 replications with $n = 100$ and 200. The medians of $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ are close to the true values 2 and -2 for all cases, and the means are closer to the true values as n increases. Moreover, the asymptotic standard

Figure 2: *Plots of $x_i\beta_i$ (black solid lines), $x_i\widehat{\beta}_i$ (red dashed lines) and $x_i\widehat{\beta}^{ols}$ (blue dotted lines) against values of x_i, where β_i are the true values, $\widehat{\beta}_i$ are the estimated values by MCP and $\widehat{\beta}^{ols}$ is the estimated value from OLS in Example 1.*

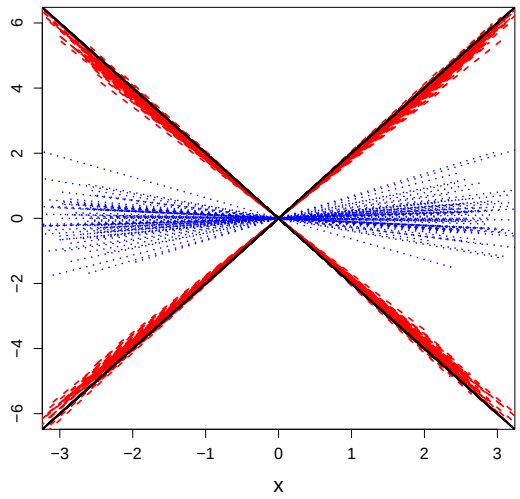
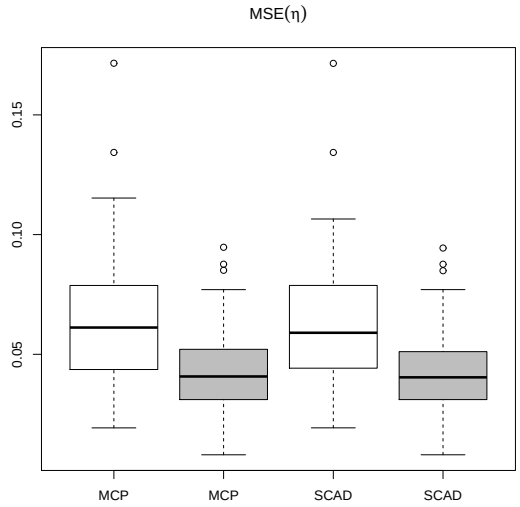


Table 2: The sample mean, median and asymptotic standard deviation (ASD) of the estimators $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ by MCP and SCAD and oracle estimators $\widehat{\alpha}_1^{or}$ and $\widehat{\alpha}_2^{or}$ based on 100 replications with $n = 100, 200$ in Example 1.

		$n = 100$			$n = 200$		
		mean	median	ASD	mean	median	ASD
$\widehat{\alpha}_1$	MCP	1.884	1.928	0.077	1.907	1.963	0.055
	SCAD	1.874	1.964	0.078	1.899	1.928	0.057
$\widehat{\alpha}_1^{or}$		1.993	1.998	0.072	1.998	1.999	0.051
$\widehat{\alpha}_2$	MCP	−1.783	−1.929	0.078	−1.823	−1.959	0.071
	SCAD	−1.770	−1.954	0.078	−1.778	−1.921	0.071
$\widehat{\alpha}_2^{or}$		−1.993	−1.988	0.073	−2.001	−2.005	0.052

deviations of the penalized estimates $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ are close to those of the oracle estimators $\widehat{\alpha}_1^{or}$ and $\widehat{\alpha}_2^{or}$. This supports the oracle property established in Theorem 4.2.

Figure 3: *The boxplots of the MSEs of $\widehat{\boldsymbol{\eta}}$ using MCP and SCAD, respectively, with $n = 100$ (white) and $n = 200$ (grey) in Example 1.*



Next, we calculate the mean squared error (MSE) of the estimates $\widehat{\boldsymbol{\eta}}$ by using the formula $\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|/\sqrt{q}$. Figure 3 depicts the boxplots of the MSEs of $\widehat{\boldsymbol{\eta}}$ by the MCP and SCAD, respectively, at $n = 100$ (white) and $n = 200$ (grey). The MCP and SCAD result in similar MSEs of $\widehat{\boldsymbol{\eta}}$. The MSE values decrease as n increases for both MCP and SCAD.

Lastly, we consider inferences about treatment heterogeneity between groups based on the asymptotic distribution of the resulting estimators established in Corollary 4.1. We test the hypothesis $\mathcal{H}_0 : \alpha_1 = \alpha_2$, i.e., $\mathcal{H}_0 : \mathbf{L}\boldsymbol{\alpha} = \mathbf{0}$, where $\mathbf{L} = [1, -1]$. According to the asymptotic normality given in Corollary 4.1, we use the F-test statistic

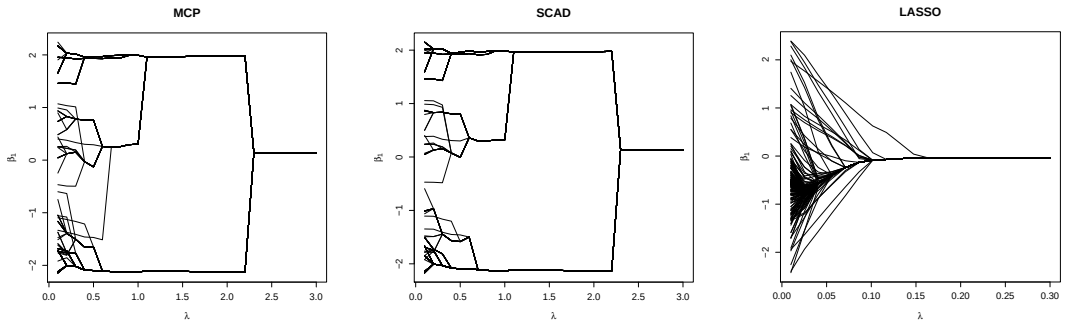
$$F = (\mathbf{L}\widehat{\boldsymbol{\alpha}}(\lambda))^{\mathrm{T}}(\widehat{\sigma}^2\mathbf{L}\widehat{\Sigma}_n^{-1}\mathbf{L}^{\mathrm{T}})^{-1}\mathbf{L}\widehat{\boldsymbol{\alpha}}(\lambda)/p$$

where

$$\widehat{\Sigma}_n = (\mathbf{X}\widehat{\mathbf{W}})^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{W}} - ((\mathbf{X}\widehat{\mathbf{W}})^{\mathrm{T}}\mathbf{Z})(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}(\mathbf{Z}^{\mathrm{T}}\mathbf{X}\widehat{\mathbf{W}}),$$

and $\widehat{\mathbf{W}}$ is defined in the same way as $\mathbf{W}$ given in Section 4 by replacing the true groups with the estimated groups, so that F asymptotically follows the F-distribution with degrees of freedom $(p, n - p\widehat{K} - q - 1)$. The estimates $\widehat{\boldsymbol{\alpha}}(\lambda)$ are obtained by the MCP and SCAD. The sample mean and median of the p-values are less than 0.01 in all cases. Thus the existence of treatment heterogeneity is further supported.

Figure 4: *Fusiongram for* $(\beta_{11}, \ldots, \beta_{1n})$*, the first component in* $\boldsymbol{\beta}_i$*'s in Example 2.*



Example 2 (Multiple treatment variables). We simulated data from the heterogeneous model with multiple treatment variables:

$$y_i = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i + \varepsilon_i, i = 1, \ldots, n, \tag{5.3}$$

where $\boldsymbol{z}_i$, ε_i and $\boldsymbol{\eta}$ are simulated in the same way as in Example 1. Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$ in which x_{i1} is simulated from standard normal and $(x_{i2}, x_{i3})^{\mathrm{T}}$ are from centered and standardized binomial with probability 0.7 for one outcome. We randomly assign the responses to two groups with equal probabilities, i.e., we let $p(i \in \mathcal{G}_1) = p(i \in \mathcal{G}_2) = 1/2$, so that $\boldsymbol{\beta}_i = \boldsymbol{\alpha}_1$ for $i \in \mathcal{G}_1$ and $\boldsymbol{\beta}_i = \boldsymbol{\alpha}_2$ for $i \in \mathcal{G}_2$, where $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \alpha_{13})$ and $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22}, \alpha_{23})$. Let $\alpha_{1j} = \alpha$ and $\alpha_{2j} = -\alpha$ for $j = 1, 2, 3$. We let $\alpha = 1, 2$ for different signal-noise ratios. Let $n = 200$.

Figure 4 displays the fusiongram for $(\beta_{11}, \ldots, \beta_{1n})$, the elements of the first component in $\boldsymbol{\beta}_i$'s for $\alpha = 2$. We obtain similar patterns as those in Figure 1 for Example 1. Again, the two concave penalties, MCP and SCAD, generate two subgroups for λ in a certain interval. For the LASSO, the estimates merges to a single value quickly.

Table 3: The sample mean, median and standard deviation (s.d.) of $\widehat{K}$ and the percentage (per) that $\widehat{K}$ equals to the true number of subgroups by MCP and SCAD based on 100 replications with $\alpha = 1, 2$ in Example 2.

	$\alpha = 1$				$\alpha = 2$			
	mean	median	s.d.	per	mean	median	s.d.	per
MCP	2.700	3.000	0.717	0.440	2.180	2.000	0.411	0.830
SCAD	2.690	3.000	0.706	0.440	2.190	2.000	0.419	0.820

We next conduct the simulations by selecting λ via minimizing the modified BIC given

in (5.1). Table 3 reports the mean, median and standard deviation (s.d.) of the estimated number of groups $\widehat{K}$ and the percentage that $\widehat{K}$ equals to the true number of subgroups by the MCP and SCAD methods based on 100 simulation realization. We observe that the mean and median values of $\widehat{K}$ get closer to 2, which is the true number of subgroups, as the α value becomes larger. Moreover, the percentage of correctly selecting the number of subgroups increases as the α value becomes larger.

Lastly, we conduct inferences on heterogeneity of treatments between groups. We conduct the hypothesis testing $\mathcal{H}_0 : \alpha_{1j} = \alpha_{2j}$ for $j = 1, 2, 3$, i.e., $\mathcal{H}_0 : \mathbf{L}\boldsymbol{\alpha} = \mathbf{0}$, where $\mathbf{L} = [\text{diag}(1,1,1), \mathbf{0}_{3\times 3}] - [\mathbf{0}_{3\times 3}, \text{diag}(1,1,1)]$. We use the F-test statistic as described in Example 1. The estimates $\widehat{\boldsymbol{\alpha}}(\lambda)$ are obtained by the MCP and SCAD methods. We obtain that the sample mean and median of the p-values based on the 100 realizations are less than 0.01 for all cases. This further confirms the heterogeneous treatment effects by the inference procedure.

Example 3 (No treatment heterogeneity). We generate data from a model with homogeneous treatment effects given by $y_i = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} + x_i\beta + \varepsilon_i, i = 1, \ldots, n$, where $\boldsymbol{z}_i$, x_i, ε_i and $\boldsymbol{\eta}$ are simulated in the same way as in Example 1. Set $\beta = 2$ and $n = 200$. We use our proposed penalized estimation approach to fit the model assuming the possible existence of treatment heterogeneity. The sample mean of the estimated number of groups $\widehat{K}$ is 1.49 and 1.48, respectively, for the MCP and SCAD methods, and the sample median is 1 for both methods based on 100 replications.

Table 4: The empirical bias (Bias) for the estimates of β and $\boldsymbol{\eta}$, and the average asymptotic standard deviation (ASD) calculated according to Corrollary 4.1 and the empirical standard deviation (ESD) for the MCP and SCAD methods and oracle estimator (ORACLE) in Example 3.

		β	η_1	η_2	η_3	η_4	η_5
MCP	Bias	−0.005	−0.002	0.007	0.003	0.002	0.001
	ASE	0.035	0.034	0.037	0.037	0.038	0.037
	ESE	0.034	0.041	0.038	0.041	0.042	0.038
SCAD	Bias	−0.004	−0.001	0.007	0.003	0.002	0.001
	ASE	0.035	0.034	0.037	0.037	0.037	0.037
	ESE	0.034	0.040	0.037	0.041	0.042	0.038
ORACLE	Bias	−0.004	−0.001	0.006	0.004	0.002	−0.001
	ASE	0.036	0.035	0.038	0.038	0.039	0.038
	ESE	0.036	0.039	0.034	0.039	0.041	0.037

To evaluate the asymptotic normality established in Corollary 4.1, Table 4 lists the empirical bias (Bias) for the estimates of β and $\boldsymbol{\eta}$, the average asymptotic standard deviation (ASD) calculated according to Corollary 4.1, the empirical standard deviation (ESD) for the MCP, SCAD as well as the oracle estimator (ORACLE). The bias, ASD and ESD for the estimates of β by the MCP and SCAD are calculated based on the replications with the estimated number of groups equal to one. For other cases, they are calculated based on the 100 replications. The biases are small relative to the standard errors. The ESD and ASD are similar for both MCP and SCAD, and they are also close to the corresponding values for the oracle estimator. These results indicate that the proposed method works well for the homogeneous model.

6 Empirical example

We apply our method to the AIDS Clinical Trials Group Study 175 (ACTG175). ACTG175 was a randomized clinical trial to compare zidovudine with other three therapies including zidovudine and didanosine, zidovudine and zalcitabine, and didanosine in adults infected with the human immunodeficiency virus type I. We randomly select 300 patients from the study to consist of our dataset. We use the log-transformed values of the CD4 counts at 20$\pm$5 weeks as the responses y_i (Tsiatis et al., 2007), and use binary variables for the three therapies as the predictors $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$. Moreover, we include 12 baseline covariates in the model, which are age (years), weight (kg), Karnofsky score, CD4 counts at baseline, CD8 counts at baseline, hemophilia (0 =no, 1 =yes), homosexual activity (0 =no, 1 =yes), history of intravenous drug use (0 =no, 1 =yes), race (0 =white, 1 =white), gender (0 =female, 1 =male), antiretroviral history (0 =naive, 1 =experienced) and symptomatic status (0 =asymptomatic, 1 =symptomatic).

To see possible heterogeneity in treatment effects, we first fit a linear regression model by using y_i as the response and the 12 baseline covariates as predictors and obtain the residuals, so that the effects of the 12 baseline covariates are controlled. In Figure 5, it shows the kernel density plot of the residuals for the patients treated with the therapy didanosine. We can see that the distribution has multiple modes for these patients, which indicates possible heterogeneous treatment effects.

Next, we use the data to fit the heterogeneous model $y_i = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i + \varepsilon_i, i = 1, \ldots, 300$, where $\boldsymbol{z}_i = (1, z_{i2} \ldots, z_{i13})^{\mathrm{T}}$ with the first component for intercept and other components being the 12 covariates described above. All of the predictors are centered and standardized before applying the regularization methods. We then identify two heterogenous groups of treatment effects by both MCP and SCAD methods. Figure 6 displays the fusiongram for $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1n})$, $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2n})$, and $\boldsymbol{\beta}_3 = (\beta_{31}, \ldots, \beta_{3n})$ by MCP. We obtain similar

Figure 5: *The kernel density plot of the residuals after controlling for the effects of the 12 baseline covariates for the patients treated with the therapy didanosine.*

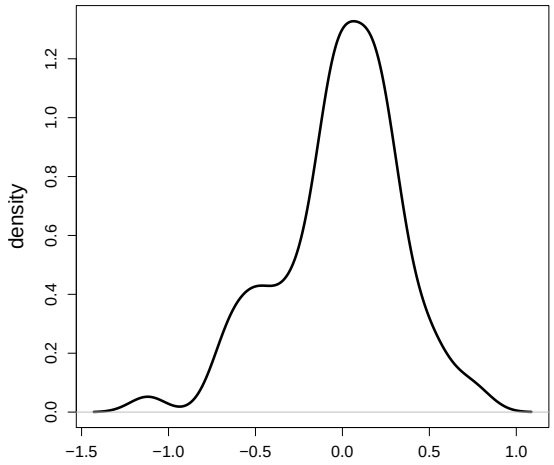
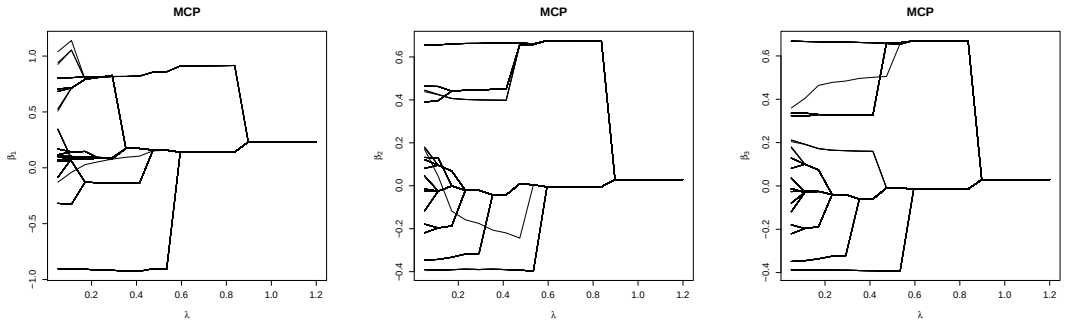


Figure 6: *Fusiongram for* $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1n})$, $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2n})$, *and* $\boldsymbol{\beta}_3 = (\beta_{31}, \ldots, \beta_{3n})$.



patterns by using SCAD. We see that the concave penalty generates two subgroups for λ in a certain interval.

We further conduct statistical inference for testing equality of the coefficients for the two identified groups by using the same method as described in Example 1 in Section 5, and then obtain the p-values < 0.01 by the MCP and SCAD methods. Thus, the heterogeneity of treatment effects is further confirmed by the inference procedure. Let $\widehat{\boldsymbol{\alpha}}_1 = (\widehat{\alpha}_{11}, \widehat{\alpha}_{12}, \widehat{\alpha}_{13})$ and $\widehat{\boldsymbol{\alpha}}_2 = (\widehat{\alpha}_{21}, \widehat{\alpha}_{22}, \widehat{\alpha}_{23})$ be the estimated coefficients for $\boldsymbol{x}_i$ in the two identified groups $\widehat{\mathcal{G}}_1$ and $\widehat{\mathcal{G}}_2$, respectively, so that $\widehat{\boldsymbol{\beta}}_i = \widehat{\boldsymbol{\alpha}}_1$ for $i \in \widehat{\mathcal{G}}_1$ and $\widehat{\boldsymbol{\beta}}_i = \widehat{\boldsymbol{\alpha}}_2$ for $i \in \widehat{\mathcal{G}}_2$.

In Table 5, we report the estimates (Est.), standard deviations (s.d.) and p-values (p-value) of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ by the MCP and SCAD methods, and those values of $\boldsymbol{\beta} = \boldsymbol{\alpha}_1$ by the OLS method. We see that only the first treatment has statistically significant effect by the

Table 5: The estimates (Est.), standard deviations (s.d.) and p-values (P-value) of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ by the MCP and SCAD methods, and those values of $\boldsymbol{\beta} = \boldsymbol{\alpha}_1$ by the OLS method.

		α_{11}	α_{12}	α_{13}	α_{21}	α_{22}	α_{23}
MCP	Est.	0.141	-0.011	-0.039	0.835	0.666	0.687
	s.d.	0.055	0.055	0.055	0.394	0.268	0.251
	p-value	0.010	0.841	0.478	0.034	0.013	0.006
SCAD	Est.	0.142	-0.010	-0.037	0.805	0.614	0.636
	s.d.	0.055	0.055	0.055	0.395	0.268	0.251
	p-value	0.010	0.855	0.501	0.041	0.022	0.011
OLS	Est.	0.212	0.035	0.036	—	—	—
	s.d.	0.060	0.058	0.058	—	—	—
	p-value	< 0.001	0.550	0.532	—	—	—

OLS method. By the MCP and SCAD methods, the first treatment still has significant effect. Moreover, the effects of the second and third treatments become significant for the second group by the MCP and SCAD methods. In conclusion, we can see that the treatments have heterogeneous effects on the two groups.

7 Discussion

The proposed heterogeneous model (2.2) looks similar to the linear mixed-effects model for studying correlated data such as longitudinal/panel data and data with repeated measurements. In the mixed-effects model, the $\boldsymbol{\beta}_i$'s are treated as random variables from a distribution, so that likelihood-based methods can be derived with the main interest of estimating the parameters $\{\boldsymbol{\eta}, \sigma^2\}$. In this paper, we aim to identify subgroups of treatment by applying a penalty function to the L_2 norms of the pairwise differences of $\boldsymbol{\beta}_i$'s. Our proposed method provides an automatic approach to identifying the subgroups.

The method can be applied to a general class of regression problems by using the criterion

$$\frac{1}{2}\sum_{i=1}^{n} \ell(y_i, \boldsymbol{z}_i^T\boldsymbol{\eta} + \boldsymbol{x}_i^T\boldsymbol{\beta}_i) + \sum_{1\le i<j\le n} p(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|; \lambda),$$

where ℓ is a given loss function. For example, for generalized linear models such as logistic regression, we take ℓ to be the negative log-likelihood function. For the Cox regression with censored data, we take the empirical loss function to be the negative partial likelihood. For these more complicated models, we can still use the basic idea of ADMM by approximating

the loss function locally using a quadratic function. However, further work is needed to study the theoretical properties.

In our theoretical results, we allow the number of the treatment variables and the confounding covariates, p and q, to diverge with the sample size n, but require them to be smaller than n. For high-dimensional problems with $p > n$ or $q > n$, a sparsity condition is needed on the coefficient to ensure the identifiability of the model. Further studies are needed to develop the computational algorithms and theoretical properties in the high-dimensional setting.

Appendix

A.1 Proof of Proposition 3.1

In this section we show the results in Proposition 3.1. By the definition of $\boldsymbol{\delta}^{m+1}$, we have

$$L(\boldsymbol{\eta}^{m+1}, \boldsymbol{\beta}^{m+1}, \boldsymbol{\delta}^{m+1}, \boldsymbol{\upsilon}^m) \leq L(\boldsymbol{\eta}^{m+1}, \boldsymbol{\beta}^{m+1}, \boldsymbol{\delta}, \boldsymbol{\upsilon}^m)$$

for any $\boldsymbol{\delta}$. Define

$$\begin{aligned} f^{m+1} &= \inf_{\mathbf{A}\boldsymbol{\beta}^{m+1}-\boldsymbol{\delta}=\mathbf{0}} \{\frac{1}{2}\left\|\mathbf{y}-\mathbf{Z}\boldsymbol{\eta}^{m+1}-\mathbf{X}\boldsymbol{\beta}^{m+1}\right\|^2 + \sum\nolimits_{i<j} p_\gamma(|\boldsymbol{\delta}_{ij}|, \lambda)\} \\ &= \inf_{\mathbf{A}\boldsymbol{\beta}^{m+1}-\boldsymbol{\delta}=\mathbf{0}} L(\boldsymbol{\eta}^{m+1}, \boldsymbol{\beta}^{m+1}, \boldsymbol{\delta}, \boldsymbol{\upsilon}^m). \end{aligned}$$

Then

$$L(\boldsymbol{\eta}^{m+1}, \boldsymbol{\beta}^{m+1}, \boldsymbol{\delta}^{m+1}, \boldsymbol{\upsilon}^m) \leq f^{m+1}.$$

Let t be an integer. Since $\boldsymbol{\upsilon}^{m+1} = \boldsymbol{\upsilon}^m + \vartheta(\mathbf{A}\boldsymbol{\beta}^{m+1}-\boldsymbol{\delta}^{m+1})$, then we have

$$\boldsymbol{\upsilon}^{m+t-1} = \boldsymbol{\upsilon}^m + \vartheta \sum\nolimits_{i=1}^{t-1} (\mathbf{A}\boldsymbol{\beta}^{m+i}-\boldsymbol{\delta}^{m+i}),$$

and thus

$$\begin{aligned} & L(\boldsymbol{\eta}^{m+t}, \boldsymbol{\beta}^{m+t}, \boldsymbol{\delta}^{m+t}, \boldsymbol{\upsilon}^{m+t-1}) \\ = \; & \frac{1}{2}\left\|\mathbf{y}-\mathbf{Z}\boldsymbol{\eta}^{m+t}-\mathbf{X}\boldsymbol{\beta}^{m+t}\right\|^2 + (\boldsymbol{\upsilon}^{m+t-1})^{\mathrm{T}}(\mathbf{A}\boldsymbol{\beta}^{m+t}-\boldsymbol{\delta}^{m+t}) \\ & + \frac{\vartheta}{2}\|\mathbf{A}\boldsymbol{\beta}^{m+t}-\boldsymbol{\delta}^{m+t}\|^2 + \sum\nolimits_{i<j} p_\gamma(|\boldsymbol{\delta}_{ij}^{m+t}|, \lambda) \\ = \; & \frac{1}{2}\left\|\mathbf{y}-\mathbf{Z}\boldsymbol{\eta}^{m+t}-\mathbf{X}\boldsymbol{\beta}^{m+t}\right\|^2 + (\boldsymbol{\upsilon}^m)^{\mathrm{T}}(\mathbf{A}\boldsymbol{\beta}^{m+t}-\boldsymbol{\delta}^{m+t}) \\ & + \vartheta \sum\nolimits_{i=1}^{t-1} (\mathbf{A}\boldsymbol{\beta}^{m+i}-\boldsymbol{\delta}^{m+i})^{\mathrm{T}}(\mathbf{A}\boldsymbol{\beta}^{m+t}-\boldsymbol{\delta}^{m+t}) \\ & + \frac{\vartheta}{2}\|\mathbf{A}\boldsymbol{\beta}^{m+t}-\boldsymbol{\delta}^{m+t}\|^2 + p_\gamma(|\boldsymbol{\delta}_{ij}^{m+t}|, \lambda) \\ \leq \; & f^{m+t}. \end{aligned}$$

Since the objective function $L(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\upsilon})$ is differentiable with respect to $(\boldsymbol{\eta}, \boldsymbol{\beta})$ and is convex with respect to $\boldsymbol{\delta}$, by applying the results in Theorem 4.1 of (Tseng, 2001), the sequence $(\boldsymbol{\eta}^m, \boldsymbol{\beta}^m, \boldsymbol{\delta}^m)$ has a limit point, denoted by $(\boldsymbol{\eta}^*, \boldsymbol{\beta}^*, \boldsymbol{\delta}^*)$. Then we have

$$f^* = \lim_{m\to\infty} f^{m+1} = \lim_{m\to\infty} f^{m+t} = \inf_{\mathbf{A}\boldsymbol{\beta}^* - \boldsymbol{\delta} = \mathbf{0}} \{\frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta}^* - \mathbf{X}\boldsymbol{\beta}^*\|^2 + \sum\nolimits_{i<j} p_\gamma(|\boldsymbol{\delta}_{ij}|, \lambda)\},$$

and for all $t \geq 0$

$$\begin{aligned} & \lim_{m\to\infty} L(\boldsymbol{\mu}^{m+t}, \boldsymbol{\beta}^{m+t}, \boldsymbol{\eta}^{m+t}, \boldsymbol{\upsilon}^{m+t-1}) \\ = \;& \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta}^* - \mathbf{X}\boldsymbol{\beta}^*\|^2 + \sum\nolimits_{i<j} p_\gamma(|\boldsymbol{\delta}_{ij}^*|, \lambda) + \lim_{m\to\infty} (\boldsymbol{\upsilon}^m)^{\mathrm{T}}(\mathbf{A}\boldsymbol{\beta}^* - \boldsymbol{\delta}^*) + (t - \frac{1}{2})\vartheta\|\mathbf{A}\boldsymbol{\beta}^* - \boldsymbol{\delta}^*\|^2 \\ \leq \;& f^*. \end{aligned}$$

Hence $\lim_{m\to\infty} \|\mathbf{r}^m\|^2 = r^* = \|\mathbf{A}\boldsymbol{\beta}^* - \boldsymbol{\delta}^*\|^2 = 0$.

Since $\boldsymbol{\beta}^{m+1}$ minimizes $L(\boldsymbol{\eta}^m, \boldsymbol{\beta}, \boldsymbol{\delta}^m, \boldsymbol{\upsilon}^m)$ by definition, we have that

$$L(\boldsymbol{\eta}^m, \boldsymbol{\beta}, \boldsymbol{\delta}^m, \boldsymbol{\upsilon}^m)/\partial\boldsymbol{\beta} = \mathbf{0},$$

and moreover,

$$\begin{aligned} & L(\boldsymbol{\eta}^m, \boldsymbol{\beta}^{m+1}, \boldsymbol{\delta}^m, \boldsymbol{\upsilon}^m)/\partial\boldsymbol{\beta} \\ = \;& \mathbf{X}^{\mathrm{T}}(\mathbf{Z}\boldsymbol{\eta}^m + \mathbf{X}\boldsymbol{\beta}^{m+1} - \mathbf{y}) + \mathbf{A}^{\mathrm{T}}\boldsymbol{\upsilon}^m + \vartheta\mathbf{A}^{\mathrm{T}}(\mathbf{A}\boldsymbol{\beta}^{m+1} - \boldsymbol{\delta}^m) \\ = \;& \mathbf{X}^{\mathrm{T}}(\mathbf{Z}\boldsymbol{\eta}^m + \mathbf{X}\boldsymbol{\beta}^{m+1} - \mathbf{y}) + \mathbf{A}^{\mathrm{T}}(\boldsymbol{\upsilon}^m + \vartheta(\mathbf{A}\boldsymbol{\beta}^{m+1} - \boldsymbol{\delta}^m)) \\ = \;& \mathbf{X}^{\mathrm{T}}(\mathbf{Z}\boldsymbol{\eta}^m + \mathbf{X}\boldsymbol{\beta}^{m+1} - \mathbf{y}) + \mathbf{A}^{\mathrm{T}}(\boldsymbol{\upsilon}^{m+1} - \vartheta(\mathbf{A}\boldsymbol{\beta}^{m+1} - \boldsymbol{\delta}^{m+1}) + \vartheta(\mathbf{A}\boldsymbol{\beta}^{m+1} - \boldsymbol{\delta}^m)) \\ = \;& \mathbf{X}^{\mathrm{T}}(\mathbf{Z}\boldsymbol{\eta}^m + \mathbf{X}\boldsymbol{\beta}^{m+1} - \mathbf{y}) + \mathbf{A}^{\mathrm{T}}\boldsymbol{\upsilon}^{m+1} + \vartheta\mathbf{A}^{\mathrm{T}}(\boldsymbol{\delta}^{m+1} - \boldsymbol{\delta}^m). \end{aligned}$$

Therefore,

$$\mathbf{s}^{m+1} = \vartheta\mathbf{A}^{\mathrm{T}}(\boldsymbol{\delta}^{m+1} - \boldsymbol{\delta}^m) = -(\mathbf{X}^{\mathrm{T}}(\mathbf{Z}\boldsymbol{\eta}^m + \mathbf{X}\boldsymbol{\beta}^{m+1} - \mathbf{y}) + \mathbf{A}^{\mathrm{T}}\boldsymbol{\upsilon}^{m+1}).$$

Since $\|\mathbf{A}\boldsymbol{\beta}^* - \boldsymbol{\delta}^*\|^2 = 0$,

$$\begin{aligned} & \lim_{m\to\infty} L(\boldsymbol{\eta}^m, \boldsymbol{\beta}^{m+1}, \boldsymbol{\delta}^m, \boldsymbol{\upsilon}^m)/\partial\boldsymbol{\beta} \\ = \;& \lim_{m\to\infty} \mathbf{X}^{\mathrm{T}}(\mathbf{Z}\boldsymbol{\eta}^m + \mathbf{X}\boldsymbol{\beta}^{m+1} - \mathbf{y}) + \mathbf{A}^{\mathrm{T}}\boldsymbol{\upsilon}^{m+1} \\ = \;& \mathbf{X}^{\mathrm{T}}(\mathbf{Z}\boldsymbol{\eta}^* + \mathbf{X}\boldsymbol{\beta}^* - \mathbf{y}) + \mathbf{A}^{\mathrm{T}}\boldsymbol{\upsilon}^* = \mathbf{0}. \end{aligned}$$

Therefore, $\lim_{m\to\infty} \mathbf{s}^{m+1} = \mathbf{0}$.

A.2 Proof of Theorem 4.1

In this section we show the results in Theorem 4.1. For every $\boldsymbol{\beta} \in \mathcal{M}_{\mathcal{G}}$, it can be written as $\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\alpha}$. Recall $\mathbf{U} = (\mathbf{Z}, \mathbf{XW})$. We have

$$\begin{pmatrix} \widehat{\boldsymbol{\eta}}^{or} \\ \widehat{\boldsymbol{\alpha}}^{or} \end{pmatrix} = \arg \min_{\boldsymbol{\eta} \in R^q, \boldsymbol{\alpha} \in R^{Kp}} \frac{1}{2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}\| = \arg \min_{\boldsymbol{\eta} \in R^q, \boldsymbol{\alpha} \in R^{Kp}} \frac{1}{2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{XW}\boldsymbol{\alpha}\|^2.$$

Thus

$$\begin{pmatrix} \widehat{\boldsymbol{\eta}}^{or} \\ \widehat{\boldsymbol{\alpha}}^{or} \end{pmatrix} = [(\mathbf{Z}, \mathbf{XW})^{\mathrm{T}}(\mathbf{Z}, \mathbf{XW})]^{-1}(\mathbf{Z}, \mathbf{XW})^{\mathrm{T}}\mathbf{y} = (\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{T}\mathbf{y}.$$

Then

$$\begin{pmatrix} \widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0 \\ \widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \end{pmatrix} = (\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon}.$$

Hence

$$\left\| \begin{pmatrix} \widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0 \\ \widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \end{pmatrix} \right\| \leq \left\| [(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\right\| \left\| \mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon} \right\|. \tag{A.1}$$

By Condition (C1), we have

$$\left\| [(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\right\| \leq C_1^{-1} \left| \mathcal{G}_{\min} \right|^{-1}. \tag{A.2}$$

Moreover

$$P(\left\| \mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon} \right\|_\infty > C\sqrt{n \log n}) \leq P(\left\| \mathbf{Z}^{\mathrm{T}}\boldsymbol{\varepsilon} \right\|_\infty > C\sqrt{n \log n}) + P(\left\| (\mathbf{XW})^{\mathrm{T}}\boldsymbol{\varepsilon} \right\|_\infty > C\sqrt{n \log n}),$$

for some constant $0 < C < \infty$. Since $\mathbf{XW} = \left[\mathbf{x}_i^{\mathrm{T}} 1\{i \in \mathcal{G}_k\}\right]_{i=1,k=1}^{n,K}$, we have

$$\left\| (\mathbf{XW})^{\mathrm{T}}\boldsymbol{\varepsilon} \right\|_\infty = \sup_{j,k} |\sum\nolimits_{i=1}^{n} x_{ij}\varepsilon_i 1\{i \in \mathcal{G}_k\}|$$

and by union bound, Condition (C1) that $\sum_{i=1}^n x_{ij}^2 1\{i \in \mathcal{G}_k\} = |\mathcal{G}_k|$ and Condition (C3),

$$\begin{aligned} & P\left(\left\| (\mathbf{XW})^{\mathrm{T}}\boldsymbol{\varepsilon} \right\|_\infty > C\sqrt{n \log n}\right) \\ \leq & \sum\nolimits_{j=1,k=1}^{p,K} P\left(|\sum\nolimits_{i=1}^{n} x_{ij}\varepsilon_i 1\{i \in \mathcal{G}_k\}| > C\sqrt{n \log n}\right) \\ \leq & \sum\nolimits_{j=1,k=1}^{p,K} P\left(|\sum\nolimits_{i=1}^{n} x_{ij}\varepsilon_i 1\{i \in \mathcal{G}_k\}| > \sqrt{|\mathcal{G}_k|}C\sqrt{\log n}\right) \\ \leq & 2Kp \exp(-c_1 C^2 \log n) = 2Kpn^{-c_1 C^2}. \end{aligned}$$

By union bound, Condition (C1) that $\|\mathbf{Z}_k\| = \sqrt{n}$, where $\mathbf{Z}_k$ is the kth column of $\mathbf{Z}$, and Condition (C3),

$$\begin{aligned} & P\left(\left\| \mathbf{Z}^{\mathrm{T}}\boldsymbol{\varepsilon} \right\|_\infty > C\sqrt{n \log n}\right) \\ \leq & \sum\nolimits_{k=1}^{q} P\left(|\mathbf{Z}_k^{\mathrm{T}}\boldsymbol{\varepsilon}| > \sqrt{n}C\sqrt{\log n}\right) \\ \leq & 2q \exp(-c_1 C^2 \log n) = 2qn^{-c_1 C^2}. \end{aligned}$$

It follows that

$$P(\left\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon}\right\|_{\infty} > C\sqrt{n\log n}) \leq 2(Kp+q)n^{-c_1C^2}.$$

Since $\left\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon}\right\| \leq \sqrt{q+Kp}\left\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon}\right\|_{\infty}$, then

$$P(\left\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon}\right\| > C\sqrt{q+Kp}\sqrt{n\log n}) \leq 2(Kp+q)n^{-c_1C^2}. \tag{A.3}$$

Therefore, by (A.1), (A.2) and (A.3), we have with probability at least $1-2(Kp+q)n^{-c_1C^2}$,

$$\left\| \begin{pmatrix} \widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0 \\ \widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \end{pmatrix} \right\| \leq CC_1^{-1}\sqrt{q+Kp}\,|\mathcal{G}_{\min}|^{-1}\sqrt{n\log n}.$$

The result (4.2) in Theorem 4.1 is proved by letting $C = c_1^{-1/2}$. Moreover,

$$\begin{aligned} \left\|\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0\right\|^2 &= \sum\nolimits_{k=1}^{K}\sum\nolimits_{i\in\mathcal{G}_k}\left\|\widehat{\boldsymbol{\alpha}}_k^{or} - \boldsymbol{\alpha}_k^0\right\|^2 \leq |\mathcal{G}_{\max}|\sum\nolimits_{k=1}^{K}\left\|\widehat{\boldsymbol{\alpha}}_k^{or} - \boldsymbol{\alpha}_k^0\right\|^2 \\ &= |\mathcal{G}_{\max}|\left\|\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0\right\|^2 \leq |\mathcal{G}_{\max}|\,\phi_n^2, \end{aligned}$$

and

$$\sup_i\left\|\widehat{\boldsymbol{\beta}}_i^{or} - \boldsymbol{\beta}_i^0\right\| = \sup_k\left\|\widehat{\boldsymbol{\alpha}}_k^{or} - \boldsymbol{\alpha}_k^0\right\| \leq \|\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0\| \leq \phi_n.$$

Let $\mathbf{U} = (\mathbf{U}_1, \ldots, \mathbf{U}_n)^{\mathrm{T}}$, and $\boldsymbol{\Xi}_n = \mathbf{U}^{\mathrm{T}}\mathbf{U}$. Then

$$\mathbf{a}_n^{\mathrm{T}}((\widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^{\mathrm{T}})^{\mathrm{T}} = \sum\nolimits_{i=1}^{n}\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i.$$

Hence

$$E\{\mathbf{a}_n^{\mathrm{T}}((\widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^{\mathrm{T}})^{\mathrm{T}}\} = 0,$$

and for any vector $\mathbf{a}_n \in \mathbb{R}^{q+Kp}$ with $||\mathbf{a}_n|| = 1$, by Condition (C3), we have

$$\begin{aligned} &\mathrm{var}\{\mathbf{a}_n^{\mathrm{T}}((\widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^{\mathrm{T}})^{\mathrm{T}}\} \\ =\ &\sigma_n^2(\mathbf{a}_n) = \sigma^2\left[\mathbf{a}_n^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{a}_n\right] \geq \sigma^2(C_1')^{-1}n^{-1}. \end{aligned} \tag{A.4}$$

Moreover, for any $\epsilon > 0$,

$$\begin{aligned} &\sum\nolimits_{i=1}^{n}E[(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i)^2 \cdot 1_{\{|\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i| > \epsilon\sigma_n(\mathbf{a}_n)\}}] \\ \leq\ &\sum\nolimits_{i=1}^{n}\{E(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i)^4\}^{1/2}[P\{|\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i| > \epsilon\sigma_n(\mathbf{a}_n)\}]^{1/2}. \end{aligned}$$

Since $E(\varepsilon_i^4) \leq c$ for some constant $c \in (0,\infty)$ by Condition (C2), then

$$\begin{aligned} \{E(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i)^4\}^{1/2} &\leq ||\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}||^2||\mathbf{U}_i||^2\{E(\varepsilon_i^4)\}^{1/2} \\ &\leq c'||\boldsymbol{\Xi}_n^{-1}||^2(q+Kp) \leq c'C_1^{-2}\,|\mathcal{G}_{\min}|^{-2}\,(q+Kp) \end{aligned}$$

for some constant $c' \in (0,\infty)$, where the last inequality follows from Condition (C3). Similarly, $E(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i)^2 \le c''C_1^{-1}\left|\mathcal{G}_{\min}\right|^{-2}(q+Kp)$ for some constant $c'' \in (0,\infty)$. Thus,

$$\begin{aligned} & P\left\{|\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i| > \epsilon\sigma_n(\mathbf{a}_n)\right\} \\ \le\ & E(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i)^2/\{\epsilon^2\sigma_n^2(\mathbf{a}_n)\} \\ \le\ & c''C_1^{-1}C_1'\sigma^{-2}\epsilon^{-2}\left|\mathcal{G}_{\min}\right|^{-2}(q+Kp)n. \end{aligned}$$

Therefore, by the above results, we have

$$\begin{aligned} & \sigma_n^{-2}(\mathbf{a}_n)\sum\nolimits_{i=1}^{n} E[(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i)^2\cdot 1_{\{|\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i|>\epsilon\sigma_n(\mathbf{a}_n)\}}] \\ \le\ & \sigma^{-2}(C_1')n^2c'C_1^{-2}\left|\mathcal{G}_{\min}\right|^{-2}(q+Kp)c''C_1^{-1}C_1'\sigma^{-2}\epsilon^{-2}\left|\mathcal{G}_{\min}\right|^{-2}(q+Kp)n \\ =\ & O\{n^3\left|\mathcal{G}_{\min}\right|^{-4}(q+Kp)^2\} = o(1). \end{aligned}$$

The last equality follows from the assumption that $|\mathcal{G}_{\min}| \gg (q+Kp)^{1/2}n^{3/4}$. Then, the result (4.4) follows from Lindeberg–Feller Central Limit Theorem.

A.3 Proof of Theorem 4.2

In this section we show the results in Theorem 4.2. Define

$$\begin{aligned} L_n(\boldsymbol{\eta},\boldsymbol{\beta}) &= \frac{1}{2}\|\mathbf{y}-\mathbf{Z}\boldsymbol{\eta}-\mathbf{X}\boldsymbol{\beta}\|^2, P_n(\boldsymbol{\beta}) = \lambda\sum_{i<j}\rho(\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|), \\ L_n^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) &= \frac{1}{2}\|\mathbf{y}-\mathbf{Z}\boldsymbol{\eta}-\mathbf{X}\mathbf{W}\boldsymbol{\alpha}\|^2, P_n^{\mathcal{G}}(\boldsymbol{\alpha}) = \lambda\sum_{k<k'}|\mathcal{G}_k\|\mathcal{G}_{k'}|\rho(\|\boldsymbol{\alpha}_k-\boldsymbol{\alpha}_{k'}\|), \end{aligned}$$

and let

$$Q_n(\boldsymbol{\eta},\boldsymbol{\beta}) = L_n(\boldsymbol{\eta},\boldsymbol{\beta}) + P_n(\boldsymbol{\beta}), Q_n^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) = L_n^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) + P_n^{\mathcal{G}}(\boldsymbol{\alpha}).$$

Let $T : \mathcal{M}_{\mathcal{G}} \to R^{Kp}$ be the mapping that $T(\boldsymbol{\beta})$ is the $Kp\times 1$ vector consisting of K vectors with dimension p and its k^{th} vector component equals to the common value of $\boldsymbol{\beta}_i$ for $i \in \mathcal{G}_k$. Let $T^* : R^{np} \to R^{Kp}$ be the mapping that $T^*(\boldsymbol{\beta}) = \{|\mathcal{G}_k|^{-1}\sum_{i\in\mathcal{G}_k}\boldsymbol{\beta}_i^{\mathrm{T}}, k=1,\ldots,K\}^{\mathrm{T}}$. Clearly, when $\boldsymbol{\beta}\in\mathcal{M}_{\mathcal{G}}$, $T(\boldsymbol{\beta}) = T^*(\boldsymbol{\beta})$.

By calculation, for every $\boldsymbol{\beta}\in\mathcal{M}_{\mathcal{G}}$, we have $P_n(\boldsymbol{\beta}) = P_n^{\mathcal{G}}(T(\boldsymbol{\beta}))$ and for every $\boldsymbol{\alpha}\in R^K$, we have $P_n(T^{-1}(\boldsymbol{\alpha})) = P_n^{\mathcal{G}}(\boldsymbol{\alpha})$. Hence

$$Q_n(\boldsymbol{\eta},\boldsymbol{\beta}) = Q_n^{\mathcal{G}}(\boldsymbol{\eta},T(\boldsymbol{\beta})), Q_n^{\mathcal{G}}(\boldsymbol{\eta},\boldsymbol{\alpha}) = Q_n(\boldsymbol{\eta},T^{-1}(\boldsymbol{\alpha})). \tag{A.5}$$

Consider the neighborhood of $(\boldsymbol{\eta}^0,\boldsymbol{\beta}^0)$:

$$\Theta = \{\boldsymbol{\eta}\in R^q, \boldsymbol{\beta}\in R^{Kp}: \left\|\boldsymbol{\eta}-\boldsymbol{\eta}^0\right\| \le \phi_n, \sup_i\left\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^0\right\| \le \phi_n\}.$$

By Theorem 4.1, there exists an event E_1 in which

$$\left\|\widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0\right\| \leq \phi_n, \sup_i \left\|\widehat{\boldsymbol{\beta}}_i^{or} - \boldsymbol{\beta}_i^0\right\| \leq \phi_n$$

and $P(E_1^C) \leq 2(q + Kp)n^{-1}$. Hence $(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) \in \Theta$ in E_1. For any $\boldsymbol{\beta} \in R^{np}$, let $\boldsymbol{\beta}^* = T^{-1}(T^*(\boldsymbol{\beta}))$. We show that $(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ is a strictly local minimizer of the objective function (2.3) with probability approaching 1 through the following two steps.

(i). In the event E_1, $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) > Q_n(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for any $(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta$ and $((\boldsymbol{\eta})^{\mathrm{T}}, (\boldsymbol{\beta}^*)^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$.

(ii). There is an event E_2 such that $P(E_2^C) \leq 2n^{-1}$. In $E_1 \cap E_2$, there is a neighborhood of $((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, denoted by Θ_n such that $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) \geq Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*)$ for any $((\boldsymbol{\eta})^{\mathrm{T}}, (\boldsymbol{\beta}^*)^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$ for sufficiently large n.

Therefore, by the results in (i) and (ii), we have $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) > Q_n(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for any $(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$ and $((\boldsymbol{\eta})^{\mathrm{T}}, (\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ in $E_1 \cap E_2$, so that $((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ is a strict local minimizer of $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta})$ (2.3) over the event $E_1 \cap E_2$ with $P(E_1 \cap E_2) \geq 1 - 2(q + Kp + 1)n^{-1}$ for sufficiently large n.

In the following we prove the result in (i). We first show $P_n^{\mathcal{G}}(T^*(\boldsymbol{\beta})) = C_n$ for any $\boldsymbol{\beta} \in \Theta$, where C_n is a constant which does not depend on $\boldsymbol{\beta}$. Let $T^*(\boldsymbol{\beta}) = \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_K^{\mathrm{T}})^{\mathrm{T}}$. It suffices to show that $\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\| > a\lambda$ for all k and k'. Then by Condition (C2), $\rho(\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\|)$ is a constant, and as a result $P_n^{\mathcal{G}}(T^*(\boldsymbol{\beta}))$ is a constant. Since

$$\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\| \geq \|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_{k'}^0\| - 2\sup_k \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\|,$$

and

$$\begin{aligned}
\sup_k \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\|^2 &= \sup_k \left\| |\mathcal{G}_k|^{-1} \sum\nolimits_{i \in \mathcal{G}_k} \boldsymbol{\beta}_i - \boldsymbol{\alpha}_k^0 \right\|^2 = \sup_k \left\| |\mathcal{G}_k|^{-1} \sum\nolimits_{i \in \mathcal{G}_k} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0) \right\|^2 \\
&= \sup_k |\mathcal{G}_k|^{-2} \left\| \sum\nolimits_{i \in \mathcal{G}_k} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0) \right\|^2 \leq \sup_k |\mathcal{G}_k|^{-1} \sum\nolimits_{i \in \mathcal{G}_k} \left\| (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0) \right\|^2 \\
&\leq \sup_i \left\| \boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0 \right\|^2 \leq \phi_n^2, \qquad \text{(A.6)}
\end{aligned}$$

then for all k and k'

$$\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\| \geq \|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_{k'}^0\| - 2\sup_k \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\| \geq b_n - 2\phi_n > a\lambda.$$

where the last inequality follows from the assumption that $b_n > a\lambda \gg \phi_n$. Therefore, we have $P_n^{\mathcal{G}}(T^*(\boldsymbol{\beta})) = C_n$, and hence $Q_n^{\mathcal{G}}(\boldsymbol{\eta}, T^*(\boldsymbol{\beta})) = L_n^{\mathcal{G}}(\boldsymbol{\eta}, T^*(\boldsymbol{\beta})) + C_n$ for all $(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta$. Since $((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ is the unique global minimizer of $L_n^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha})$, then $L_n^{\mathcal{G}}(\boldsymbol{\eta}, T^*(\boldsymbol{\beta})) > L_n^{\mathcal{G}}(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ for all $(\boldsymbol{\eta}^{\mathrm{T}}, (T^*(\boldsymbol{\beta}))^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ and hence $Q_n^{\mathcal{G}}(\boldsymbol{\eta}, T^*(\boldsymbol{\beta})) > Q_n^{\mathcal{G}}(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or})$ for all $T^*(\boldsymbol{\beta}) \neq \widehat{\boldsymbol{\alpha}}^{or}$. By (A.5), we have $Q_n^{\mathcal{G}}(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\alpha}}^{or}) = Q_n(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ and $Q_n^{\mathcal{G}}(\boldsymbol{\eta}, T^*(\boldsymbol{\beta})) =$

$Q_n(\boldsymbol{\eta}, T^{-1}(T^*(\boldsymbol{\beta}))) = Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*)$. Therefore, $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) > Q_n(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for all $\boldsymbol{\beta}^* \neq \widehat{\boldsymbol{\beta}}^{or}$, and the result in (i) is proved.

Next we prove the result in (ii). For a posive sequence t_n, let $\Theta_n = \{\boldsymbol{\beta}_i : \sup_i \|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i^{or}\| \leq t_n\}$. For $(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$, by Taylor's expansion, we have

$$Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) - Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) = \Gamma_1 + \Gamma_2,$$

where

$$\begin{aligned} \Gamma_1 &= -(\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}^m)^{\mathrm{T}} \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ \Gamma_2 &= \sum_{i=1}^{n} \frac{\partial P_n(\boldsymbol{\beta}^m)}{\partial \boldsymbol{\beta}_i^{\mathrm{T}}} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*). \end{aligned}$$

and $\boldsymbol{\beta}^m = \alpha\boldsymbol{\beta} + (1-\alpha)\boldsymbol{\beta}^*$ for some constant $\alpha \in (0,1)$. Moreover,

$$\begin{aligned} \Gamma_2 &= \lambda \sum\nolimits_{\{j>i\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|^{-1} (\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m)^{\mathrm{T}} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) \\ &\quad + \lambda \sum\nolimits_{\{j<i\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|^{-1} (\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m)^{\mathrm{T}} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) \\ &= \lambda \sum\nolimits_{\{j>i\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|^{-1} (\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m)^{\mathrm{T}} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) \\ &\quad + \lambda \sum\nolimits_{\{i<j\}} \rho'(\|\boldsymbol{\beta}_j^m - \boldsymbol{\beta}_i^m\|) \|\boldsymbol{\beta}_j^m - \boldsymbol{\beta}_i^m\|^{-1} (\boldsymbol{\beta}_j^m - \boldsymbol{\beta}_i^m)^{\mathrm{T}} (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*) \\ &= \lambda \sum\nolimits_{\{j>i\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|^{-1} (\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m)^{\mathrm{T}} \{(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) - (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*)\} \end{aligned} \quad \text{(A.7)}$$

When $i, j \in \mathcal{G}_k$, $\boldsymbol{\beta}_i^* = \boldsymbol{\beta}_j^*$, and $\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m = \alpha(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)$. Thus,

$$\begin{aligned} \Gamma_2 &= \lambda \sum_{k=1}^{K} \sum_{\{i,j \in \mathcal{G}_k, i<j\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|^{-1} (\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m)^{\mathrm{T}} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_j) \\ &\quad + \lambda \sum_{k<k'} \sum_{\{i \in \mathcal{G}_k, j' \in \mathcal{G}_{k'}\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|^{-1} (\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m)^{\mathrm{T}} \{(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) - (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*)\}. \end{aligned}$$

Moreover,

$$\sup_i \|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i^0\|^2 = \sup_k \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\|^2 \leq \phi_n^2, \quad \text{(A.8)}$$

where the last inequality follows from (A.6). Since $\boldsymbol{\beta}_i^m$ is between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_i^*$,

$$\sup_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_i^0\| \leq \alpha \sup_i \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\| + (1-\alpha) \sup_i \|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i^0\| \leq \alpha\phi_n + (1-\alpha)\phi_n = \phi_n. \quad \text{(A.9)}$$

Hence for $k \neq k'$, $i \in \mathcal{G}_k, j' \in \mathcal{G}_{k'}$,

$$\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\| \geq \min_{i \in \mathcal{G}_k, j' \in \mathcal{G}_{k'}} \|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\| - 2 \max_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_i^0\| \geq b_n - 2\phi_n > a\lambda,$$

and thus $\rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) = 0$. Therefore,

$$\begin{aligned}\Gamma_2 &= \lambda \sum_{k=1}^{K} \sum_{\{i,j\in\mathcal{G}_k, i<j\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|)\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|^{-1}(\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m)^{\mathrm{T}}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j) \\ &= \lambda \sum_{k=1}^{K} \sum_{\{i,j\in\mathcal{G}_k, i<j\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|,\end{aligned}$$

where the last step follows from $\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m = \alpha(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)$. Furthermore, by the same reasoning as (A.6), we have

$$\sup_i \|\boldsymbol{\beta}_i^* - \widehat{\boldsymbol{\beta}}_i^{or}\| = \sup_k \|\boldsymbol{\alpha}_k - \widehat{\boldsymbol{\alpha}}_k^{or}\|^2 \le \sup_i \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_i^{or}\|.$$

Then

$$\begin{aligned}\sup_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\| &\le 2\sup_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_i^*\| \le 2\sup_i \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\| \\ &\le 2(\sup_i \|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i^{or}\| + \sup_i \|\boldsymbol{\beta}_i^* - \widehat{\boldsymbol{\beta}}_i^{or}\|) \le 4\sup_i \|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i^{or}\| \le 4t_n.\end{aligned}$$

Hence $\rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) \ge \rho'(4t_n)$ by concavity of $\rho(\cdot)$. As a result,

$$\Gamma_2 \ge \sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k, i<j\}} \lambda\rho'(4t_n)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|. \tag{A.10}$$

Let

$$\mathbf{Q} = (\mathbf{Q}_1^{\mathrm{T}}, \ldots, \mathbf{Q}_n^{\mathrm{T}})^{\mathrm{T}} = [(\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}^m)^{\mathrm{T}}\mathbf{X}]^{\mathrm{T}}.$$

Then

$$\begin{aligned}\Gamma_1 &= -\mathbf{Q}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = -\sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k\}} \frac{\mathbf{Q}_i^{\mathrm{T}}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)}{|\mathcal{G}_k|} \\ &= -\sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k\}} \frac{\mathbf{Q}_i^{\mathrm{T}}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)}{2|\mathcal{G}_k|} - \sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k\}} \frac{\mathbf{Q}_i^{\mathrm{T}}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)}{2|\mathcal{G}_k|} \\ &= -\sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k\}} \frac{(\mathbf{Q}_j - \mathbf{Q}_i)^{\mathrm{T}}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)}{2|\mathcal{G}_k|} \\ &= -\sum\nolimits_{k=1}^{K} \sum\nolimits_{\{i,j\in\mathcal{G}_k, i<j\}} \frac{(\mathbf{Q}_j - \mathbf{Q}_i)^{\mathrm{T}}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)}{|\mathcal{G}_k|}.\end{aligned} \tag{A.11}$$

Moreover,

$$\mathbf{Q}_i = (y_i - \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i^m)\boldsymbol{x}_i = (\varepsilon_i + \mathbf{z}_i^{\mathrm{T}}(\boldsymbol{\eta}^0 - \boldsymbol{\eta}) + \boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_i^m))\boldsymbol{x}_i,$$

and then

$$\sup_i \|\mathbf{Q}_i\| \le \sup_i \left\{\|\boldsymbol{x}_i\|(\|\boldsymbol{\varepsilon}\|_\infty + \|\boldsymbol{z}_i\|\|\boldsymbol{\eta}^0 - \boldsymbol{\eta}\| + \|\boldsymbol{x}_i\|\|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_i^m\|)\right\}$$

By Condition (C1) that $\sup_i \|\boldsymbol{x}_i\| \le C_2\sqrt{p}$ and $\sup_i \|\boldsymbol{z}_i\| \le C_3\sqrt{q}$, (A.9) that $\sup_i \|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_i^m\| \le \phi_n$ and $\|\boldsymbol{\eta}^0 - \boldsymbol{\eta}\| \le \phi_n$, we have

$$\sup_i \|\mathbf{Q}_i\| \le C_2\sqrt{p}(\|\boldsymbol{\varepsilon}\|_\infty + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n).$$

By Condition (C3)

$$P(\|\boldsymbol{\varepsilon}\|_\infty > \sqrt{2c_1^{-1}}\sqrt{\log n}) \le \sum\nolimits_{i=1}^n P(|\varepsilon_i| > \sqrt{2c_1^{-1}}\sqrt{\log n}) \le 2n^{-1}.$$

Thus there is an event E_2 such that $P(E_2^C) \le 2n^{-1}$, and over the event E_2,

$$\sup_i \|\mathbf{Q}_i\| \le C_2\sqrt{p}(\sqrt{2c_1^{-1}}\sqrt{\log n} + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n).$$

Then

$$\begin{aligned} & |\frac{(\mathbf{Q}_j - \mathbf{Q}_i)^{\mathrm{T}}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)}{|\mathcal{G}_k|}| \\ \le\ & |\mathcal{G}_{\min}|^{-1}\|\mathbf{Q}_j - \mathbf{Q}_i\|\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\| \le |\mathcal{G}_{\min}|^{-1} 2 \sup_i \|\mathbf{Q}_i\|\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\| \\ \le\ & 2C_2|\mathcal{G}_{\min}|^{-1}\sqrt{p}(\sqrt{2c_1^{-1}}\sqrt{\log n} + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|. \qquad \text{(A.12)} \end{aligned}$$

Therefore, by (A.10), (A.11) and (A.12), we have

$$\begin{aligned} & Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) - Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) \\ \ge\ & \sum_{k=1}^K \sum_{\{i,j\in\mathcal{G}_k, i<j\}} \{\lambda\rho'(4t_n) - 2C_2|\mathcal{G}_{\min}|^{-1}\sqrt{p}(\sqrt{2c_1^{-1}}\sqrt{\log n} + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n)\}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|. \end{aligned}$$

Let $t_n = o(1)$, then $\rho'(4t_n) \to 1$. Since $\lambda \gg \phi_n$, $p = o(n)$, and $|\mathcal{G}_{\min}|^{-1}p = o(1)$, then $\lambda \gg |\mathcal{G}_{\min}|^{-1}\sqrt{p}\sqrt{\log n}$, $\lambda \gg |\mathcal{G}_{\min}|^{-1}\sqrt{pq}$ and $\lambda \gg |\mathcal{G}_{\min}|^{-1}p\phi_n$. Therefore, $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) - Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) \ge 0$ for sufficiently large n, so that the result in (ii) is proved.

A.4 Proof of Theorem 4.3

In this section we show the results in Theorem 4.3. The proofs of (4.6) and (4.7) follow the same arguments as the proof of Theorem 4.1 by letting $\widetilde{\mathbf{X}} = \mathbf{x}$ and $|\mathcal{G}_{\min}| = n$, and thus they are omitted. Next, we will show (4.8). It follows similar procedures as the proof of Theorem 4.2 with the details given below. Define $\mathcal{M} = \{\boldsymbol{\beta} \in \mathbb{R}^{np} : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_n\}$. For each $\boldsymbol{\beta} \in \mathcal{M}$, we have $\boldsymbol{\beta}_i = \boldsymbol{\alpha}$ for all i. Let $T : \mathcal{M} \to R^p$ be the mapping that $T(\boldsymbol{\beta})$ is the $p \times 1$ vector equal to the common vector $\boldsymbol{\alpha}$. Let $T^* : R^{np} \to R^p$ be the mapping that $T^*(\boldsymbol{\beta}) = \{n^{-1}\sum_{i=1}^n \boldsymbol{\beta}_i$. Clearly, when $\boldsymbol{\beta} \in \mathcal{M}$, $T(\boldsymbol{\beta}) = T^*(\boldsymbol{\beta})$. Consider the neighborhood of $(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0)$:

$$\Theta = \{\boldsymbol{\eta} \in R^q, \boldsymbol{\beta} \in R^p : \left\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\right\| \le \phi_n, \sup_i \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\right\| \le \phi_n\},$$

where $\phi_n = c_1^{-1/2} C_1^{-1} \sqrt{q+p} \sqrt{n^{-1} \log n}$. By the result in (4.6), there exists an event E_1 such that on the event E_1,

$$\left\| \widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0 \right\| \leq \phi_n, \sup_i \left\| \widehat{\boldsymbol{\beta}}_i^{or} - \boldsymbol{\beta}_i^0 \right\| \leq \phi_n,$$

and $P(E_1^C) \leq 2(q+p)n^{-1}$. Hence $(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) \in \Theta$ on the event E_1. For any $\boldsymbol{\beta} \in R^{np}$, let $\boldsymbol{\beta}^* = T^{-1}(T^*(\boldsymbol{\beta}))$. We show that $(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ is a strictly local minimizer of the objective function (2.3) with probability approaching 1 through the following two steps.

(i). On the event E_1, $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) > Q_n(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for any $(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta$ and $((\boldsymbol{\eta})^{\mathrm{T}}, (\boldsymbol{\beta}^*)^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$.

(ii). There is an event E_2 such that $P(E_2^C) \leq 2n^{-1}$. On $E_1 \cap E_2$, there is a neighborhood of $((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, denoted by Θ_n such that $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) \geq Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*)$ for any $((\boldsymbol{\eta})^{\mathrm{T}}, (\boldsymbol{\beta}^*)^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$ for sufficiently large n.

Therefore, by the results in (i) and (ii), we have $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) > Q_n(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$ for any $(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$ and $((\boldsymbol{\eta})^{\mathrm{T}}, (\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ in $E_1 \cap E_2$, so that $((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ is a strict local minimizer of $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta})$ (2.3) on the event $E_1 \cap E_2$ with $P(E_1 \cap E_2) \geq 1 - 2(q+p+1)n^{-1}$ for sufficiently large n.

By the definition of $((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, we have $\frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}^*\|^2 > \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\eta}}^{or} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{or}\|^2$ for any $((\boldsymbol{\eta})^{\mathrm{T}}, (\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}} \in \Theta$ and $((\boldsymbol{\eta})^{\mathrm{T}}, (\boldsymbol{\beta}^*)^{\mathrm{T}})^{\mathrm{T}} \neq ((\widehat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$. Moreover, since $p_\gamma(\|\widehat{\boldsymbol{\beta}}_i^{or} - \widehat{\boldsymbol{\beta}}_j^{or}\|, \lambda) = p_\gamma(\|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_j^*\|, \lambda) = 0$ for $1 \leq i, j \leq n$, we have $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) = \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}^*\|^2$ and $Q_n(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) = \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\eta}}^{or} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{or}\|^2$. Therefore, $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) > Q_n(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or})$.

Next we prove the result in (ii). For a positive sequence t_n, let $\Theta_n = \{\boldsymbol{\beta}_i : \sup_i \|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i^{or}\| \leq t_n\}$. For $(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$, by Taylor's expansion, we have

$$Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) - Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) = \Gamma_1 + \Gamma_2,$$

where

$$\begin{aligned} \Gamma_1 &= -(\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}^m)^{\mathrm{T}} \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ \Gamma_2 &= \sum_{i=1}^{n} \frac{\partial P_n(\boldsymbol{\beta}^m)}{\partial \boldsymbol{\beta}_i^{\mathrm{T}}} (\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*). \end{aligned}$$

$P_n(\boldsymbol{\beta}) = \lambda \sum_{i<j} \rho(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|)$, and $\boldsymbol{\beta}^m = a\boldsymbol{\beta} + (1-a)\boldsymbol{\beta}^*$ for some constant $a \in (0,1)$. Moreover, by (A.7),

$$\begin{aligned} \Gamma_2 &= \lambda \sum\nolimits_{\{j>i\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|^{-1} (\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m)^{\mathrm{T}} \{(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) - (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*)\} \\ &= \lambda \sum\nolimits_{\{j>i\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|, \end{aligned}$$

where the second equality holds due to the fact that $\boldsymbol{\beta}_i^* = \boldsymbol{\beta}_j^*$ and $\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m = a(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)$. Let $T^*(\boldsymbol{\beta}) = \boldsymbol{\alpha}$. Then, following the same argument as (A.8), we have

$$\sup_i \|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i^0\|^2 = \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^0\|^2 \leq \sup_i \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\|^2.$$

Then

$$\begin{aligned}\sup_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\| &\leq 2\sup_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_i^*\| \leq 2\sup_i \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\| \\ &\leq 2(\sup_i \|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i^{or}\| + \sup_i \|\boldsymbol{\beta}_i^* - \widehat{\boldsymbol{\beta}}_i^{or}\|) \leq 4\sup_i \|\boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_i^{or}\| \leq 4t_n.\end{aligned}$$

Hence $\rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|) \geq \rho'(4t_n)$ by concavity of $\rho(\cdot)$. As a result,

$$\Gamma_2 \geq \sum\nolimits_{\{i<j\}} \lambda\rho'(4t_n)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|. \tag{A.13}$$

Let

$$\mathbf{Q} = (\mathbf{Q}_1^{\mathrm{T}}, \ldots, \mathbf{Q}_n^{\mathrm{T}})^{\mathrm{T}} = [(\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}^m)^{\mathrm{T}}\mathbf{X}]^{\mathrm{T}}.$$

By the same reasoning as the proof for (A.11), we have

$$\Gamma_1 = -\mathbf{Q}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = -n^{-1}\sum\nolimits_{\{i<j\}} (\mathbf{Q}_j - \mathbf{Q}_i)^{\mathrm{T}}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i). \tag{A.14}$$

By the same argument as the proof for (A.12), we have that there is an event E_2 such that $P(E_2^C) \leq 2n^{-1}$, and on the event E_2,

$$\begin{aligned}& n^{-1}|(\mathbf{Q}_j - \mathbf{Q}_i)^{\mathrm{T}}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)| \\ \leq\ & 2C_2 n^{-1}\sqrt{p}(\sqrt{2c_1^{-1}}\sqrt{\log n} + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|.\end{aligned} \tag{A.15}$$

Therefore, by (A.13), (A.14) and (A.15), we have

$$\begin{aligned}& Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) - Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) \\ \geq\ & \sum\nolimits_{\{i<j\}} \{\lambda\rho'(4t_n) - 2C_2 n^{-1}\sqrt{p}(\sqrt{2c_1^{-1}}\sqrt{\log n} + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n)\}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|.\end{aligned}$$

Let $t_n = o(1)$, then $\rho'(4t_n) \to 1$. Since $\lambda \gg \phi_n$, $p = o(n)$, and $n^{-1}p = o(1)$, then $\lambda \gg n^{-1}\sqrt{p}\sqrt{\log n}$, $\lambda \gg n^{-1}\sqrt{pq}$ and $\lambda \gg n^{-1}p\phi_n$. Therefore, $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) - Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) \geq 0$ for sufficiently large n, so that the result in (ii) is proved.

References

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.

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors, *Statistics and Computing* **24**: 871–883.

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

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.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization, *Annals of Applied Statistics* **1**(2): 302–332.

Guo, F. J., Levina, E., Michailidis, G. and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering, *Biometrics* **66**: 793–804.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models, *Annals of Statistics* **38**: 2282–2313.

Ke, T., Fan, J. and Wu, Y. (2015). Homogeneity in regression, *Journal of the American Statistical Association* **110**: 175–194.

Kravitz, R. L., D. N. and Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages, *Milbank Q.* **82**: 661–687.

Lagakos, S. W. (2006). The chanllenge of subgroup analysis: Reporting without distorting, *New England Journal of Medicine* **354**: 1667–1669.

Lee, E. R., Noh, H. and Park, B. U. (2014). Model selection via bayesian information criterion for quantile regression models, *Journal of the American Statistical Association* **109**: 216–229.

Ma, S. and Huang, J. (2016). A concave pairwise fusion approach to subgroup analysis, *Accepted for publication by the Journal of American Statistical Association* .

Rothwell, P. M. (2005). Subgroup analysis in randomized clinical trials: importance, indications and interpretation, *Lancet* **365**: 176–186.

Schwarz, C. (1978). Estimating the dimension of a model, *The Annals of Statistics* **6**: 461–464.

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.

Shen, X. and Huang, H. C. (2010). Grouping pursuit through a regularization solution surface, *Journal of the American Statistical Association* **105**: 727–739.

Sorensen, T. (1996). Which patients may be harmed by good treatments?, *Lancet* **348**: 351–352.

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

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**: 91108.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization1, *Journal of Optimization Theory and Applications* **109**: 475–494.

Tsiatis, A. A., Davidian, M., Zhang, M. and Lu, X. (2007). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach, *Statistics in Medicine* **27**: 4658–4677.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso, *Electronic Journal of Statistics* **3**: 1360–1392.

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika* **94**: 553–568.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B* **68**: 4967.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics* **38**: 894–942.