Shujie Ma[1] / Jian Huang[2] / Zhiwei Zhang[1] / Mingming Liu[3]

Exploration of Heterogeneous Treatment Effects via Concave Fusion

[1] Department of Statistics, University of California at Riverside, Riverside, California 92521, USA, E-mail: shujie.ma@ucr.edu, zhiwei.zhang@ucr.ed

[2] Department of Statistics and Actuarial Science, University of Iowa, Iowa City, USA, E-mail: jian-huang@uiowa.edu

[3] Department of Statistics, University of California at Riverside, Riverside, California, USA, E-mail: mliu034@ucr.edu

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 of achieving this goal is that we usually do not have *a priori* knowledge of the grouping information of patients with respect to treatment effect. To address this problem, we consider a heterogeneous regression model which allows the coefficients for treatment variables to be subject-dependent with unknown grouping information. We develop a concave fusion penalized method for estimating the grouping structure and the subgroup-specific treatment effects, and derive an alternating direction method of multipliers algorithm for its implementation. We also study the theoretical properties of the proposed method and show that under suitable conditions there exists a local minimizer that equals the oracle least squares estimator based on *a priori* knowledge of the true grouping information with high probability. This provides theoretical support for making statistical inference about the subgroup-specific treatment effects using the proposed method. The proposed method is illustrated in simulation studies and illustrated with real data from an AIDS Clinical Trials Group Study.

1   Introduction

Treatment effects are often heterogeneous, that is, the same treatment can have different effects on different patients [1, 2]. For instance, molecularly targeted cancer drugs are only effective for patients with tumors expressing targets [3], and the relative efficacy of antiretroviral drugs for treating human immunodeficiency virus infection sometimes depends on baseline viral load and CD4 count [4]. Understanding the heterogeneity of treatment effects (HTE) is critical to the eventual success of precision medicine, which seeks to tailor medical treatments to individual patients.

Possible HTE is usually assessed in a subgroup analysis [1, 5–8] or, more generally, a regression analysis relating the outcome of interest to treatment and a collection of baseline covariates [9]. Such a regression model can incorporate HTE as interactions between treatment and baseline covariates, and can be used to estimate covariate-specific treatment effects indirectly. Alternatively, a covariate-specific treatment effect model can be specified and estimated directly without relying on a regression model for the outcome [10, 11]. There is also a growing literature on HTE estimation using machine learning methods (e. g. [12–14]). All of these methods are based on observed baseline covariates; they do not address possible HTE due to unmeasured covariates.

The collection of observed baseline covariates is often limited, and thus may be insufficient for characterizing the true HTE across individual patients. The true HTE structure is not empirically identifiable unless all effect modifiers are measured, but can be explored under appropriate assumptions. Zhang et al. [4, 15] use random effect models to conduct sensitivity analyses concerning the joint distribution of two potential outcomes (for two different treatments). Shen and He [16] propose a two-group logistic-normal mixture model for the true HTE and develop a model-based procedure for testing the equivalence of the two groups. The model of Shen and He [16] includes a normality assumption for the outcome and a logistic regression model relating the latent group membership to a collection of observed covariates.

In this article, we propose a general latent class model for exploring the true HTE. Different from the aforementioned existing models, our model deals with an unspecified number of latent groups, without assuming normality for the outcome or a particular relationship between the latent group membership and the observed

---

covariates. In our model, we assume that the treatment coefficients are subject-specific and belong to different groups with unknown grouping information. The subgroups, therefore, can be driven by observed covariates, unobserved covariates or an arbitrary combination of both types of covariates. We recover the latent subgroups and estimate the model using concave fusion penalization, an unsupervised machine learning method, that applies a concave penalty, such as the smoothly clipped absolute deviations penalty [SCAD, [17]] or the minimax concave penalty [MCP, [18]], to pairwise differences of coefficients for treatment effects. The fusion penalized approach has been proposed for clustering analysis of grouping data objects [19] and grouping means of clinical outcomes with random errors [20], but not for exploring HTE. Ma and Huang [20] briefly mentioned potential applications of the fusion penalized approach to the estimation of subject-specific coefficient models without providing theoretical justifications and numerical studies.

We consider the present paper as the first work which applies the fusion learning approach to the investigation of HTE through subject-specific treatment coefficients and provides a theoretical analysis of the resulting estimator for the proposed general latent class model with subject-specific treatment effects. The consistency and asymptotic distributional properties of the estimators follow from the fact that, under appropriate conditions, the oracle least squares estimator based on *a priori* knowledge of the true group membership is a local minimizer of the objective function with high probability. Moreover, we derive the conditions on the number of subgroups and the number of covariates compared with the sample size as well as the lower bound of the minimum difference of coefficient values between the subgroups in order to identify the true subgroups of treatment effects. Computationally, we apply an alternating direction method of multipliers (ADMM) algorithm [21] for implementing the proposed approach. This algorithm has been used widely in convex optimization problems. We derive the convergence properties of the ADMM in the present setting. Another contribution is that using the proposed fusion penalized estimator, we further propose a bootstrapping procedure to test homogeneity of the treatment effects for confirmatory analysis.

The rest of this article is organized as follows. Section 2 introduces the proposed latent class model. Section 3 describes the concave fusion penalization method for model estimation. In Section 4 we establish the theoretical properties of the proposed estimator. Section 5 introduces a bootstrapping procedure for testing homogeneity. In Section 6 we evaluate the finite-sample properties of the proposed method via simulation studies, and apply the method to AIDS Clinical Trials Group Study 175. Concluding remarks are given in Section 7. The computational algorithm and the technical proofs are provided in the Appendix.

2 Model

Suppose the data consist of $(y_i, z_i, x_i), i = 1, \ldots, n$, where y_i is an outcome variable, z_i a q-vector of patient characteristics related to the outcome, and x_i a p-vector of treatment or exposure characteristics. The vector x_i may include discrete components (e. g. indicators of treatment groups or exposure status) and/or continuous components (e. g. dose of drug or radiation). Our goal is to understand how the causal effect of x_i may vary across individual subjects. Let $y_i(x)$ denote the potential outcome for the ith subject under treatment x [22]. We assume that treatment assignment is strongly ignorable [23] in the sense that

$$x_i \perp y_i(x) | z_i, \tag{1}$$

where $\perp$ denotes independence and x may be any treatment that could conceivably be given to the ith subject. From this assumption it follows that

$$E\{y_i(x) | z_i\} = E\{y_i(x) | z_i, x_i = x\} = E\{y_i | z_i, x_i = x\},$$

which is empirically identified for any possible treatment x for the ith subject. Assumption (1) is trivially true in a randomized clinical trial where x_i is independent of all baseline variables; in this case, adjustment for z_i is not strictly necessary but may improve efficiency.

Suppose the data generation mechanism can be described by the following linear regression model:

$$y_i = z_i^T \eta + x_i^T \beta_i + \varepsilon_i, i = 1, \ldots, n, \tag{2}$$

where η and β_i are unknown regression coefficients and the ε_i's are i.i.d. random errors with $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. We assume that the first entry in each x_i is 1 so the intercept is included in β_i. Under this model and the strongly ignorable treatment assignment assumption, the subject-specific regression coefficient β_i represents the causal effect of x_i on the ith subject, and the distribution of β_i across i represents the true HTE. With more unknown parameters than observations, model (2) is not estimable without additional assumptions.

A simple way to constrain model (2) and achieve identification is to assume that the $\boldsymbol{\beta}_i$ are all equal. This leads to the following model:

$$y_i = \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\eta} + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n, \tag{3}$$

which is commonly used in practice. Model (3) is quite restrictive as it rules out any HTE with respect to observed and unobserved covariates. The model can be expanded by including interaction terms between some or all components of $\mathbf{z}_i$ and $\mathbf{x}_i$, thereby allowing $\boldsymbol{\beta}_i$ to be a specified function of $\mathbf{z}_i$. The expanded model can accommodate a particular form of observed HTE, but it does not account for any latent HTE. Moreover, even if the HTE is only driven by the observed covariates $\mathbf{z}_i$, it is unclear in what specific form $\mathbf{z}_i$ causes the HTE. For instance, the common assumption that $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\Gamma} \mathbf{t}_i$, where $\mathbf{t}_i$ is a subset of $\mathbf{z}_i$, and $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ are coefficients implies a linear interaction structure which may be too stringent in practice. When many baseline variables are present, there can be many interaction terms. This increases the chance of false positive results in interaction tests and may lead to overstated and misleading conclusions [24].

Another approach is to treat $\boldsymbol{\beta}_i$ as a subject-specific random vector following a specified conditional distribution given $\mathbf{z}_i$. A prominent example in this category is the common linear mixed model where $\boldsymbol{\beta}_i$ is assumed to follow a normal distribution, independently of $\mathbf{z}_i$. Such a model may have identifiability issues in the present setting, where each subject contributes only one observation. Another example in this category, which achieves identifiability, is the two-group logistic-normal mixture model of Shen and He [16]. In the present notation, their model can be expressed as

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_i + \varepsilon_i, \text{ where } \boldsymbol{\beta}_i = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 w_i \text{ and } w_i = I(\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\gamma} + \epsilon > 0), \tag{4}$$

where ε is an error distributed by the standard logistic distribution, so that $P(w_i = 1|\mathbf{z}_i) = \exp(\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\gamma})/(1 + \exp(\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\gamma}))$. This model assumes that the $\boldsymbol{\beta}_i$ in (2) arise from a finite mixture model with two possible values ($\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2$), that ε_i is normally distributed, and that the latent group membership is related to observed covariates through a logistic regression model. Under these assumptions, model (4) can be estimated using a standard EM algorithm, and Shen and He [16] further develop an EM test for the null hypothesis $\boldsymbol{\alpha}_2 = 0$, which indicates the absence of (observed or latent) HTE. This approach, as a model-based way to estimate the true HTE, is limited by the two-group assumption and the modeling assumptions.

Here we consider a more flexible latent class model and propose a machine learning approach to recovering the subgroups. We assume that the $\boldsymbol{\beta}_i$ in (2) arise from a mixture model with an unspecified number of groups. Specifically, let $\mathscr{G} = (\mathscr{G}_1, \ldots, \mathscr{G}_K)$ be a mutually exclusive partition of $\{1, \ldots, n\}$. Suppose $\boldsymbol{\beta}_i = \boldsymbol{\alpha}_k$ for all $i \in \mathscr{G}_k$, where $\boldsymbol{\alpha}_k$ is the common value for the $\boldsymbol{\beta}_i$'s in group $\mathscr{G}_k$. In other words, we assume that

$$\boldsymbol{\beta}_i = \boldsymbol{\alpha}_1 w_{i1} + \boldsymbol{\alpha}_2 w_{i2} + \cdots + \boldsymbol{\alpha}_K w_{iK},$$

where $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kp})^{\mathrm{T}}$, and $w_{ik} \in (0, 1)$ is the (latent) indicator for the kth group $\mathscr{G}_k$, i. e. $w_{ik} = 1$ for $i \in \mathscr{G}_k$ and $w_{ik} = 0$ otherwise. Substituting this into model (2) yields the latent class model:

$$y_i = \mathbf{z}_i^{\mathrm{T}} \boldsymbol{\eta} + \mathbf{x}_i^{\mathrm{T}} (\boldsymbol{\alpha}_1 w_{i1} + \boldsymbol{\alpha}_2 w_{i2} + \cdots + \boldsymbol{\alpha}_K w_{iK}) + \varepsilon_i, \tag{5}$$

where the number of subgroups K is unknown and the group indicators w_{ik} are unobservable. Model (5) includes the case of no HTE as a special case (i. e. $K = 1$ and hence $w_{i1} \equiv 1$). Of note, we do not parameterize the distribution of ε_i, nor do we specify how the w_{ik}'s may be related to $\mathbf{z}_i$. Thus, our model is considerably more flexible than the model of Shen and He [16].

Although the subgroups in model (5) are not (fully) ascertainable using observed covariates, an exploratory analysis based on model (5) can provide unique insights into the true HTE and helpful guidance for future research. For example, if the results indicate that a new treatment only benefits a small (and unidentified) proportion of the patient population, that finding might motivate scientists to collect more covariate data (e. g. biomarkers) in future studies and search for predictors of treatment benefit. Conversely, if there is no indication of clinically important HTE, that information would support a decision to (re)direct limited resources toward other, more promising areas of research.

3 Estimation

To identify the subgroups of the heterogeneous treatment effects, we first need to estimate model (5), i. e. we need to estimate the number of subgroups K, the coefficients $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$, and the group membership w_{ik} for each observation. Estimating model (5) is the same as estimation of model (2), since they are the same models with different notations. We propose a concave fusion method for model estimation described below.

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

$$Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\eta} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i)^2 + \sum_{1 \le i < j \le n} p(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|, \lambda), \tag{6}$$

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 $\lambda \ge 0$. For a given $\lambda > 0$, let

$$(\hat{\boldsymbol{\eta}}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda)) = \underset{\boldsymbol{\eta} \in \mathrm{IR}^q, \boldsymbol{\beta} \in \mathrm{IR}^{np}}{\operatorname{argmin}} Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}; \lambda). \tag{7}$$

We use sparsity-inducing penalties (to be discussed later) in (6). For a sufficiently large λ, the penalty shrinks some of $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|$ to zero. We partition the treatment effects into subgroups according to the unique values of $\hat{\boldsymbol{\beta}}$. Specifically, let $\hat{\lambda}$ be the value of the tuning parameter on the path selected based on a data-driven procedure such as the BIC. For simplicity, write $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\beta}}) \equiv (\hat{\boldsymbol{\eta}}(\hat{\lambda}), \hat{\boldsymbol{\beta}}(\hat{\lambda}))$. Let $\{\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_{\widehat{K}}\}$ be the distinct values of $\hat{\boldsymbol{\beta}}$, where $\widehat{K}$ is the number of the distinct values. These are the estimates of subgroup-specific treatment effects. The samples can then be divided into subgroups accordingly. Denote the set of indices of the kth subgroup by $\widehat{\mathscr{G}}_k = \{i : \hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\alpha}}_k, 1 \le i \le n\}, 1 \le k \le \widehat{K}$. Accordingly, we have $\widehat{w}_{ik} = 1$, if $i \in \widehat{\mathscr{G}}_k$ and $\widehat{w}_{ik} = 0$, otherwise.

A popular sparsity-inducing penalty is the L_1 or lasso penalty with $p_\gamma(t, \lambda) = \lambda|t|$ [25], but this penalty tends to produce too many subgroups [20]. Thus, we focus on two concave penalty functions: the smoothly clipped absolute deviation penalty [SCAD, [17]] and the minimax concave penalty [MCP, [18]]. 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 lead to nearly unbiased estimators of the parameters due to the fact that their derivatives equal zero at large (in magnitude) values of the parameter estimates. Moreover, they are more aggressive in enforcing a sparser solution. Thus, they are better suited for the current problem, where the number of subgroups may be expected to be much smaller than the sample size.

We compute $(\hat{\boldsymbol{\eta}}(\lambda), \hat{\boldsymbol{\beta}}(\lambda))$ given in (7) 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 positive number. We are particularly interested in the path $\{\hat{\boldsymbol{\beta}}(\lambda) : \lambda \in [\lambda_{\min}, \lambda_{\max}]\}$. The ADMM algorithm for computing the solution path on a grid of λ values is described in detail in Section A.1 of the Appendix. We also derive the convergence property of the ADMM algorithm with concave penalties. The result is given in Proposition A.1 of the Appendix.

Figure 1 illustrates the solution path for the estimates of the treatment coefficients $(\hat{\beta}_{21}(\lambda), \ldots, \hat{\beta}_{2n}(\lambda))$ against λ using MCP, SCAD and lasso penalties for simulated data in Example 1 of Section 6, which has two subgroups with 'treatment effects' 0 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 $\{\hat{\boldsymbol{\beta}}(\lambda), \lambda \in [\lambda_{\min}, \lambda_{\max}]\}$ as a *fusiongram*.

In Figure 1, 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 $(\beta_{21}, \ldots, \beta_{2n})$ merge into two values that are close to the true values 0 and 2. When the λ value exceeds 1.0, the estimates shrink to one value. For the lasso, the estimates of $(\beta_{21}, \ldots, \beta_{2n})$ merge to one value quickly at $\lambda = 0.05$ due to the overshrinkage of the L_1 penalty.
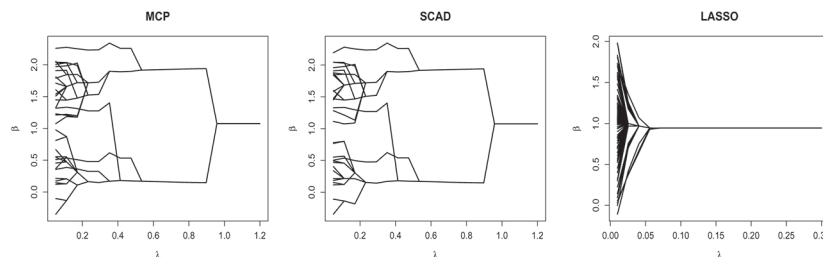

Figure 1: Solution paths for $(\hat{\beta}_{21}(\lambda), \ldots, \hat{\beta}_{2n}(\lambda))$ against λ with $n = 200$ for data from Example 1 in Section 6.

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 $i \in \mathscr{G}_k$ and $w_{ik} = 0$ otherwise. Let $\mathbf{W} = \widetilde{\mathbf{W}} \otimes \mathbf{I}_p$. Let $\mathscr{M}\mathscr{G} = \{\boldsymbol{\beta} \in \mathrm{IR}^{np}\} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_j, \text{for } i, j \in G_k, 1 \leq k \leq K$. For each $\boldsymbol{\beta} \in \mathscr{M}_{\mathscr{G}}$, it can be written as $\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \dots, \boldsymbol{\alpha}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_k$ is a $p \times 1$ vector of the kth subgroup-specific parameter for $k = 1, \dots, K$. Simple calculation shows

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

where $|\mathscr{G}_k|$ denotes the number of elements in $\mathscr{G}_k$. Denote the minimum and maximum group sizes by $|\mathscr{G}_{\min}| = \min_{1 \leq k \leq K} |\mathscr{G}_k|$ and $|\mathscr{G}_{\max}| = \max_{1 \leq k \leq K} |\mathscr{G}_k|$, 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 $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_s)^{\mathrm{T}} \in \mathrm{IR}^s$, let $\|\boldsymbol{\zeta}\|_\infty = \max_{1 \leq l \leq s} |\zeta_l|$. 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 $\mathbf{A} = \left(A_{ij}\right)_{i=1, j=1}^{s,t}$, denote $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq s} \sum_{j=1}^{t} |A_{ij}|$. Let $\mathbf{y} = (y_1, \dots, y_n)^{\mathrm{T}}$, $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)^{\mathrm{T}}$, and $\mathbf{X} = \mathrm{diag}(\boldsymbol{x}_1^{\mathrm{T}}, \dots, \boldsymbol{x}_n^{\mathrm{T}})$. Denote $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{W}$ 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 $\rho(t)$ is symmetric, non-decreasing and concave on $[0, \infty)$. It is constant for $t \geq a\lambda$ 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 $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$ has sub-Gaussian tails such that $P(|\mathbf{a}^{\mathrm{T}}\boldsymbol{\varepsilon}| > \|\mathbf{a}\|x) \leq 2\exp(-c_1 x^2)$ for any vector $\mathbf{a} \in \mathrm{IR}^n$ and $x > 0$, where $0 < c_1 < \infty$.

(C3) Assume $\sum_{i=1}^{n} z_{il}^2 = n$ for $1 \leq l \leq q$, and $\sum_{i=1}^{n} x_{ij}^2 1\{i \in \mathscr{G}_k\} = |\mathscr{G}_k|$ for $1 \leq j \leq p$, $\lambda_{\min}(\mathbf{U}^{\mathrm{T}}\mathbf{U}) \geq C_1 |\mathscr{G}_{\min}|$, $\sup_i \|\boldsymbol{x}_i\| \leq C_2\sqrt{p}$ and $\sup_i \|\mathbf{z}_i\| \leq C_3\sqrt{q}$ for some constants $0 < C_1 < \infty$, $0 < C_2 < \infty$ 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

$$\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}} = \mathrm{diag} \sum\nolimits_{i \in \mathscr{G}_k} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}, k = 1, \dots, K.$$

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

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

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

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 $\mathscr{G}_1, \dots, \mathscr{G}_K$ were known, the oracle estimator of $(\boldsymbol{\eta}, \boldsymbol{\beta})$ would be

$$(\widehat{\boldsymbol{\eta}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) = \underset{\boldsymbol{\eta} \in \mathrm{IR}^q, \boldsymbol{\beta} \in \mathscr{M}\mathscr{G}}{\mathrm{argmin}} \frac{1}{2} \| \mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta} \|^2. \tag{8}$$

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

$$(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or}) = \underset{\boldsymbol{\eta} \in \mathrm{IR}, \, \boldsymbol{\alpha} \in \mathrm{IR}^{Kp}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\alpha}\|^2 = (\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{y}.$$

Let $\boldsymbol{\alpha}_k^0$ be the true common coefficient vector for group $\mathscr{G}_k$, $k = 1,...,K$ 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 1

Suppose $|\mathscr{G}_{\min}| \gg \sqrt{(q + Kp)n \log n}$. Then under Conditions (C1)-(C3), we have with probability at least $1 - 2(Kp + q)n^{-1}$,

$$\left\| ((\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^T, (\hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^T)^T \right\| \leq \phi_n, \tag{9}$$

and

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

where

$$\phi_n = c_1^{-1/2} C_1^{-1} \sqrt{q + Kp} \, |\mathscr{G}_{\min}|^{-1} \sqrt{n \log n}. \tag{10}$$

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

$$\sigma_n(\mathbf{a}_n)^{-1}\mathbf{a}_n^T((\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^T, (\hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^T)^T \to_D N(0, 1), \tag{11}$$

where

$$\sigma_n(\mathbf{a}_n) = \sigma \left[\mathbf{a}_n^T (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{a}_n \right]^{1/2}. \tag{12}$$

Remark 1

Since $|\mathscr{G}_{\min}| \leq n/K$, by the condition $|\mathscr{G}_{\min}| \gg \sqrt{(q + Kp)n \log n}$, then q, K and p must satisfy $K\sqrt{q + Kp} = o\{\sqrt{n(\log n)^{-1}}\}$.

Remark 2

By letting $|\mathscr{G}_{\min}| = \delta n/K$ for some constant $0 < \delta \leq 1$, the bound (9) 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

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

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

Theorem 2

Suppose the conditions in Theorem 1 hold. If $b_n > a\lambda$ and $\lambda \gg \phi_n$, for some constant $a > 0$, where ϕ_n is given in (10), then there exists a local minimizer $(\hat{\boldsymbol{\eta}}(\lambda), \hat{\boldsymbol{\beta}}(\lambda))$ of the objective function $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}; \lambda)$ given in (6) satisfying

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

Remark 3

Remark 1 shows that the oracle estimator $(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ 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 $\hat{\boldsymbol{\alpha}}_k(\lambda) = \hat{\boldsymbol{\beta}}_i^{or}$ for $i \in \mathscr{G}_k$. This result holds given that $b_n \gg \phi_n$. As discussed in Remark Remark 2, when K, p and q are finite and fixed numbers and $|\mathscr{G}_{\min}| = \delta n/K$ for some constant $0 < \delta \leq 1$, $b_n \gg C^* \sqrt{\log n/n}$ for some constant $0 < C^* < \infty$.

Let $\hat{\boldsymbol{\alpha}}(\lambda) = (\hat{\boldsymbol{\alpha}}_1(\lambda)^{\mathrm{T}}, \ldots, \hat{\boldsymbol{\alpha}}_K(\lambda)^{\mathrm{T}})^{\mathrm{T}}$ be the estimated treatment effects such that $\hat{\boldsymbol{\alpha}}_k(\lambda) = \hat{\boldsymbol{\beta}}_i(\lambda)$ for $i \in \mathscr{G}_k$, where $k = 1,...,K$, and $\hat{\boldsymbol{\beta}}(\lambda) = \{\hat{\boldsymbol{\beta}}_i(\lambda)^{\mathrm{T}}, 1 \leq i \leq n\}^{\mathrm{T}}$ is the local minimizer given in Theorem 2. Based on the results in Theorem 1 and 2, we obtain the asymptotic distribution of $(\hat{\boldsymbol{\eta}}(\lambda)^{\mathrm{T}}, \hat{\boldsymbol{\alpha}}(\lambda)^{\mathrm{T}})^{\mathrm{T}}$ given in the following corollary.

Corollary 1

Under the conditions in Theorem 2, we have for any $\mathbf{a}_n \in \mathrm{IR}^{q+Kp}$ with $\|\mathbf{a}_n\| = 1$, as $n \to \infty$,

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

with $\sigma_n(\mathbf{a}_n)$ given in (12). As a result, we have for any vectors $\mathbf{a}_{n1} \in \mathrm{IR}^q$ with $\|\mathbf{a}_{n1}\| = 1$ and $\mathbf{a}_{n2} \in \mathrm{IR}^{Kp}$ $\|\mathbf{a}_{n2}\| = 1$, as $n \to \infty$,

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

where

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

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

Remark 4

From the oracle property in Theorem 2, we have that $P(\widehat{K} = K) \to 1$, $\widehat{K}$ is the estimated number of subgroups. Moreover, since $\hat{\boldsymbol{\beta}}^{or} = \widetilde{\mathbf{W}}\boldsymbol{\alpha}^{or}$ and $\hat{\boldsymbol{\beta}}(\lambda) = \widehat{\widetilde{\mathbf{W}}}\hat{\boldsymbol{\alpha}}(\lambda)$, then $P(\widehat{\widetilde{\mathbf{W}}} = \widetilde{\mathbf{W}}) \to 1$, where $\widehat{\widetilde{\mathbf{W}}} = \{\widehat{w}_{ik}\}$ with $\widehat{w}_{ik} = 1$ for $i \in \widehat{\mathscr{G}}_k$ and $w_{ik} = 0$ otherwise. Hence, the subgroup memberships can be recovered with a high probability.

Remark 5

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

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

respectively, where $z_{\alpha/2}$ is the $(1 - \alpha/2)100$ percentile of the standard normal, and $\hat{\sigma}_{n1}(\mathbf{a}_{n1})$ and $\hat{\sigma}_{n2}(\mathbf{a}_{n2})$ are estimates of $\sigma_{n1}(\mathbf{a}_{n1})$ and $\sigma_{n2}(\mathbf{a}_{n2})$ with σ^2 estimated by

$$\hat{\sigma}^2 = (n - q - \widehat{K}p)^{-1}\sum_{i=1}^n (y_i - \mathbf{z}_i^T\hat{\boldsymbol{\eta}} - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_i)^2.$$

4.3 Homogeneous model

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

$$(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or}) = \underset{\boldsymbol{\eta} \in \mathrm{IR}^q, \boldsymbol{\alpha} \in \mathrm{IR}^p}{\mathrm{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{x}\boldsymbol{\alpha}\|^2$$
$$= (\mathbf{U}^{*T}\mathbf{U}^*)^{-1}\mathbf{U}^{*T}\mathbf{y},$$

where $\mathbf{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T$ and $\mathbf{U}^* = (\mathbf{Z}, \mathbf{x})$. Let $\hat{\boldsymbol{\beta}}^{or} = (\hat{\boldsymbol{\beta}}_1^{orT}, \dots, \hat{\boldsymbol{\beta}}_n^{orT})^T$, where $\hat{\boldsymbol{\beta}}_i^{or} = \hat{\boldsymbol{\alpha}}^{or}$ 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 $1 \le j \le p$, $\lambda_{\min}(\mathbf{U}^{*T}\mathbf{U}^*) \ge C_1 n$, $\sup_i \|\mathbf{x}_i\| \le C_2\sqrt{p}$ and $\sup_i \|\mathbf{z}_i\| \le C_3\sqrt{q}$ for some constants $0 < C_1 < \infty$, $0 < C_2 < \infty$ and $0 < C_3 < \infty$.

Theorem 3

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

$$\|((\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^T, (\hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^T)^T\| \le \phi_n,$$
$$\sup_i \|\hat{\boldsymbol{\beta}}_i^{or} - \boldsymbol{\beta}_i^0\| \le \phi_n,$$

(13)

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 $\mathbf{a}_n \in \mathrm{IR}^{q+p}$ with $\|\mathbf{a}_n\| = 1$, as $n \to \infty$,*

$$\sigma_n(\mathbf{a}_n)^{-1}\mathbf{a}_n^{\mathrm{T}}((\hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^{\mathrm{T}})^{\mathrm{T}} \to N(0,1), \tag{14}$$

where

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

Moreover, if $\lambda \gg \phi_n$, then there exists a local minimizer $(\hat{\boldsymbol{\eta}}(\lambda), \hat{\boldsymbol{\beta}}(\lambda))$ of the objective function $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}; \lambda)$ given in (6) satisfying

$$P\left((\hat{\boldsymbol{\eta}}(\lambda), \hat{\boldsymbol{\beta}}(\lambda)) = (\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})\right) \to 1. \tag{15}$$

Remark 6

By Theorem 3, the local minimizer $\hat{\boldsymbol{\beta}}_i(\lambda) = \hat{\boldsymbol{\alpha}}(\lambda) = \hat{\boldsymbol{\alpha}}^{or}$ for all i with probability approaching 1. Then, we have for any vectors $\mathbf{a}_{n1} \in \mathrm{IR}^q$ with $\|\mathbf{a}_{n1}\| = 1$ and $\mathbf{a}_{n2} \in \mathrm{IR}^p$ with $\|\mathbf{a}_{n2}\| = 1$, as $n \to \infty$,

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

where

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

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

5   Testing of a homogeneous model

Next, we propose a residual bootstrapping procedure to test homogeneity of the treatment effects, i. e. to test whether the model is the homogeneous model, given as $y_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha} + \varepsilon_i, i = 1, \dots, n$. We consider this model as the reduced model, and the full model is given in (2). We estimate the reduced model by OLS and obtain the resulting estimators as $\hat{\boldsymbol{\eta}}^{\mathrm{R}}$ and $\hat{\boldsymbol{\alpha}}^{\mathrm{R}}$. We then estimate the full model by our proposed method and denote the resulting estimators as $\hat{\boldsymbol{\eta}}^{\mathrm{F}}$ and $\hat{\boldsymbol{\beta}}_i^{\mathrm{F}}$. Let the fitted values be $\hat{\mu}_i^{\mathrm{R}} = \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\eta}}^{\mathrm{R}} + \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\alpha}}^{\mathrm{R}}$ and $\hat{\mu}_i^{\mathrm{F}} = \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\eta}}^{\mathrm{F}} + \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i^{\mathrm{F}}$ for the reduced and full models, respectively. Borrowing the idea from [26], we use the integrated squared deviation between $\hat{\mu}_i^{\mathrm{R}}$ and $\hat{\mu}_i^{\mathrm{F}}$ as the test statistic, which would be $\mathcal{T}_n = \sum_{i=1}^n (\hat{\mu}_i^{\mathrm{F}} - \hat{\mu}_i^{\mathrm{R}})^2/n$. Let the residuals be $\hat{\epsilon}_i = y_i - \hat{\mu}_i^{\mathrm{F}}$ for $i = 1, \dots, n$. We obtain a randomly resampled residual $\hat{\epsilon}_i^*$ and then create synthetic response variables $y_i^* = \hat{\mu}_i^{\mathrm{R}} + \hat{\epsilon}_i^*$. Using $(\mathbf{z}_i, \mathbf{x}_i, y_i^*)$ as bootstrap observations, we obtain the fitted values $\hat{\mu}_i^{*\mathrm{R}}$ and $\hat{\mu}_i^{*\mathrm{F}}$, respectively, by refitting the reduced and full models, and then creat the bootstrapped version of the test statistic, denoted as $\mathcal{T}_n^* = \sum_{i=1}^n (\hat{\mu}_i^{*\mathrm{F}} - \hat{\mu}_i^{*\mathrm{R}})^2/n$. Using the Monte Carlo simulations to approximate the conditional distribution $\mathcal{L}^*(\mathcal{T}_n^*) = \mathcal{L}(\mathcal{T}_n^*|(\mathbf{z}_i, \mathbf{x}_i)_{i=1}^n)$, we obtain the $(1-\alpha)^{\mathrm{th}}$ quantile $\hat{t}_\alpha$ and reject the hypothesis of homogeneity if $\mathcal{T}_n > \hat{t}_\alpha$ at the significance level α. Alternatively, we can obtain the P-value p_{v} by finding the $(1-p_{\mathrm{v}})^{\mathrm{th}}$ quantile $\hat{t}_{p_{\mathrm{v}}}$ which satisfies $\hat{t}_{p_{\mathrm{v}}} = \mathcal{T}_n$.

6   Numerical studies

6.1   Simulation studies

We use the modified Bayes Information Criterion (BIC) [27] for high-dimensional data settings to select the tuning parameter by minimizing

$$\mathrm{BIC}(\lambda) = \log\left[\sum_{i=1}^n (y_i - \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\eta}}(\lambda) - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i(\lambda))^2/n\right] + C_n \frac{\log n}{n}(\hat{K}(\lambda)p + q), \tag{16}$$

where C_n is a positive number which can depend on n. When $C_n = 1$, the modified BIC reduces to the traditional BIC [28]. Following [29], we use $C_n = \log(np + q)$. We select λ by minimizing the modified BIC.

One important evaluation criterion for clustering methods is their ability to reconstruct the true underlying cluster structure. We, therefore, use the Rand Index measure [30] to evaluate the accuracy of the clustering results. The Rand Index is viewed as a measure of the percentage of correct decisions made by an algorithm. It is computed by using the formula:

$$\text{RI} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \tag{17}$$

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

Example 1

(Two subgroups). We simulate data from the heterogeneous model with two groups:

$$y_i = \mathbf{z}_i^{\text{T}}\boldsymbol{\eta} + \mathbf{x}_i^{\text{T}}\boldsymbol{\beta}_i + \varepsilon_i, i = 1,\dots,n, \text{ with } \boldsymbol{\beta}_i = \boldsymbol{\alpha}_1 w_{i1} + \boldsymbol{\alpha}_2 w_{i2}, \tag{18}$$

where $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})^{\text{T}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, in which $\boldsymbol{\Sigma} = \{\sigma_{jj'}\}$, $\sigma_{jj} = 1$ and $\sigma_{jj'} = 0.3$ for $j \neq j'$, and $\mathbf{x}_i = (1, x_i)^{\text{T}}$, in which x_i is simulated from centered and standardized binomial with probability 0.7 for one outcome. We simulate the error terms ε_i from independent $N(0, 0.5^2)$. Let $\boldsymbol{\eta} = (1,1,1)^{\text{T}}$, $\boldsymbol{\alpha}_1 = (2,2)^{\text{T}}$ and $\boldsymbol{\alpha}_2 = (0,0)^{\text{T}}$. Moreover, let $w_{i1} = I(z_{i1}^2 + u_i - 1 < 0)$ and $w_{i2} = 1 - w_{i1}$, where $u_i \sim \mathcal{N}(0,1)$.

Table 1: The sample mean, median and standard deviation (s.d.) of $\widehat{K}$, the Rand Index (RI) value and the percentage (per) of $\widehat{K}$ equaling to the true number of subgroups by MCP and SCAD based on 500 replications with $n = 200, 400$ in Example 1.

	$n = 200$					$n = 400$				
	mean	median	s.d.	RI	per	mean	median	s.d.	RI	per
MCP	2.100	2.000	0.302	0.799	0.890	2.080	2.000	0.272	0.826	0.920
SCAD	2.100	2.000	0.303	0.799	0.890	2.080	2.000	0.273	0.825	0.920

We select the λ value by minimizing the modified BIC given in (16). Table 1 reports the sample mean, median and standard deviation (s.d.) of the estimated number of groups $\widehat{K}$, the average value of Rand Index (RI) defined in (17) for measuring clustering accuracy, and the percentage (per) of $\widehat{K}$ equaling to the true number of subgroups by the MCP and SCAD methods based on 500 simulation realizations with $n = 200,400$. 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 Rand Index (RI) value and the percentage of correctly selecting the number of subgroups become closer to 1 as n increases.

Table 2: The sample mean, median and asymptotic standard error (ASE) of the estimators $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$ by MCP and SCAD and oracle estimators $\hat{\boldsymbol{\alpha}}_1^{or}$ and $\hat{\boldsymbol{\alpha}}_2^{or}$ based on 500 replications with $n = 200, 400$ in Example 1.

		$n = 200$			$n = 400$		
		mean	median	ASE	mean	median	ASE
$\hat{\alpha}_{11}$	MCP	2.016	2.019	0.046	2.012	2.014	0.034
	SCAD	2.016	2.018	0.046	2.012	2.014	0.034
$\hat{\alpha}_{11}^{or}$		1.996	1.997	0.047	2.009	2.009	0.033
$\hat{\alpha}_{12}$	MCP	1.927	1.929	0.045	1.947	1.945	0.034
	SCAD	1.927	1.929	0.045	1.947	1.946	0.034
$\hat{\alpha}_{12}^{or}$		1.990	1.989	0.046	1.999	1.999	0.034
$\hat{\alpha}_{21}$	MCP	0.016	0.016	0.053	0.006	0.010	0.038
	SCAD	0.016	0.016	0.053	0.006	0.010	0.038
$\hat{\alpha}_{21}^{or}$		0.006	0.010	0.055	0.009	0.009	0.039
$\hat{\alpha}_{22}$	MCP	0.087	0.086	0.054	0.086	0.084	0.038

| | SCAD | 0.087 | 0.086 | 0.054 | 0.086 | 0.084 | 0.038 |
| $\hat{\alpha}_{22}^{or}$ | | -0.011 | -0.014 | 0.055 | 0.001 | 0.001 | 0.039 |

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 error (ASE) obtained according to Corollary Corollary 1 of the estimators $\hat{\boldsymbol{\alpha}}_1 = (\hat{\alpha}_{11}, \hat{\alpha}_{12})^T$ and $\hat{\boldsymbol{\alpha}}_2 = (\hat{\alpha}_{21}, \hat{\alpha}_{22})^T$ by the MCP and SCAD methods and oracle estimators $\hat{\boldsymbol{\alpha}}_1^{or} = (\hat{\alpha}_{11}^{or}, \hat{\alpha}_{12}^{or})^T$ and $\hat{\boldsymbol{\alpha}}_2^{or} = (\hat{\alpha}_{21}^{or}, \hat{\alpha}_{22}^{or})^T$ based on 500 replications with $n = 200$ and 400. The medians and means of $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$ are close to the true values 2 and 0 for all cases. Moreover, the asymptotic standard errors of the penalized estimators $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$ are close to those of the oracle estimators $\hat{\boldsymbol{\alpha}}_1^{or}$ and $\hat{\boldsymbol{\alpha}}_2^{or}$. This supports the oracle property established in Theorem 2.

Next, we calculate the mean squared error (MSE) of the estimates $\hat{\boldsymbol{\eta}}$ by using the formula $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\| / \sqrt{q}$. Figure 2 depicts the boxplots of the MSEs of $\hat{\boldsymbol{\eta}}$ by the MCP and SCAD, respectively, at $n = 200$ (white) and $n = 400$ (grey). The MCP and SCAD result in similar MSEs of $\hat{\boldsymbol{\eta}}$. The MSE values decrease as n increases for both MCP and SCAD.
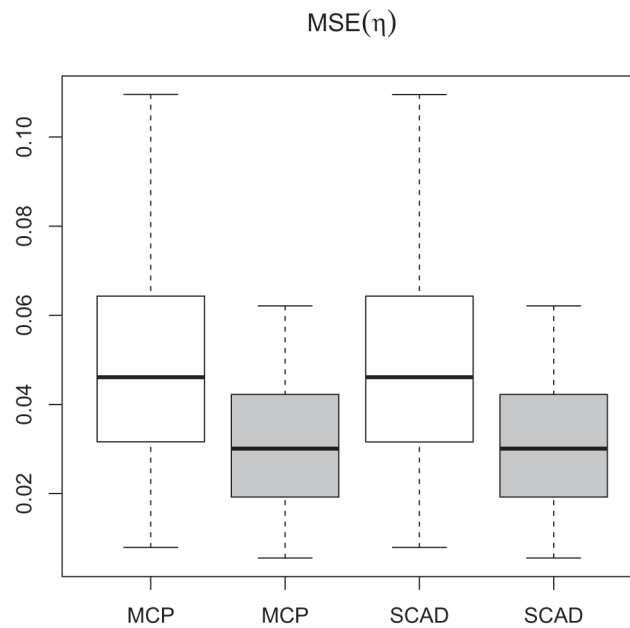

Figure 2: The boxplots of the MSEs of $\hat{\boldsymbol{\eta}}$ using MCP and SCAD, respectively, with $n = 200$ (white) and $n = 400$ (grey) in Example 1.

Lastly, we fit the two-group logistic-normal mixture model given in (4) using the R package "flexmix", and obtain the average values of RI, which are 0.516 and 0.518, for $n = 200$ and 400, respectively, based on 500 simulation realizations. We see that our method leads to higher RI values than the structured mixture modelling approach, even though our method does not need to specify the true number of subgroups beforehand. Table 3 reports the sample mean, median and empirical standard deviation (ESD) of the estimators $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$ by using the clustering result from the mixture modelling approach to fit regressions of the two groups. Clearly, the estimated values from our method given in Table 2 are closer to the true values $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ than those obtained from the mixture modelling approach given in Table 3. The mixture modelling method requires the indicator function which is $w_i = I(\mathbf{z}_i^T \boldsymbol{\gamma} + \epsilon > 0)$ given in (4) to be a parametric linear function of $\mathbf{z}_i$, so that it fits a mis-specified model for this example. As a result, it leads to biased estimates of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$.

Table 3: The sample mean, median and empirical standard deviation (ESD) of the estimators $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$ by the mixture modelling approach based on 500 replications in Example 1.

	$n = 200$			$n = 400$		
	mean	median	ESD	mean	median	ESD
$\hat{\alpha}_{11}$	1.834	1.809	0.266	1.819	1.794	0.179
$\hat{\alpha}_{12}$	1.254	1.262	0.137	1.263	1.266	0.090
$\hat{\alpha}_{21}$	0.809	0.842	0.419	0.786	0.796	0.343
$\hat{\alpha}_{22}$	0.949	0.977	0.210	0.940	0.957	0.151

Example 2

(Three subgroups). We simulate data from the heterogeneous model with three groups:

$$y_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i + \varepsilon_i, i = 1, \ldots, n, \text{ with } \boldsymbol{\beta}_i = \boldsymbol{\alpha}_1 w_{i1} + \boldsymbol{\alpha}_2 w_{i2} + \boldsymbol{\alpha}_3 w_{i3}, \tag{19}$$

where $\mathbf{z}_i$, ε_i and $\boldsymbol{\eta}$ are simulated in the same way as in Example 1. Let $\mathbf{x}_i = (1, x_i)^{\mathrm{T}}$, where x_i is generated from centered and standardized binomial with probability 0.5 for one outcome. Let $\boldsymbol{\alpha}_1 = (-c, -c)^{\mathrm{T}}$, $\boldsymbol{\alpha}_2 = (0, 0)^{\mathrm{T}}$ and $\boldsymbol{\alpha}_2 = (c, c)^{\mathrm{T}}$. Moreover, let $w_{i1} = I(|z_{i1} + z_{i2} + z_{i3}| < 0.9)$, $w_{i2} = I(z_{i1} + z_{i2} + z_{i3} \geq 0.9)$ and $w_{i3} = 1 - w_{i1} - w_{i2}$. Let $n = 200$.

Figure 3 displays the fusiongram for $(\hat{\beta}_{21}(\lambda), \ldots, \hat{\beta}_{2n}(\lambda))$, the elements of the second component in $\hat{\boldsymbol{\beta}}_i(\lambda)$'s, against λ values with $c = 2$. The fusiongrams for SCAD and MCP look similar. They generate three subgroups for λ in a certain interval, and the estimates of the treatment effects are nearly unbiased on this segment. For the LASSO, the estimates merge to a single value quickly due to the overshrinkage of the L_1 penalty.
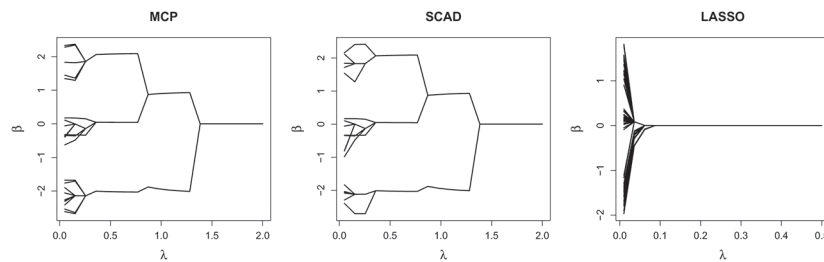

Figure 3: Fusiongram for $\hat{\beta}_{21}(\lambda), \ldots, \hat{\beta}_{2n}(\lambda)$, the second component in $\hat{\boldsymbol{\beta}}_i(\lambda)$'s, against λ with $c = 2$ in Example 2.

Table 4: The sample mean, median and standard deviation (s.d.) of $\widehat{K}$, the Rand Index (RI) value and the percentage (per) of $\widehat{K}$ equaling to the true number of subgroups by MCP and SCAD based on 500 replications for $c = 2, 3$ and 4 in Example 2.

	$c = 2$		$c = 3$		$c = 4$	
	MCP	**SCAD**	**MCP**	**SCAD**	**MCP**	**SCAD**
mean	3.450	3.470	3.400	3.410	3.350	3.370
median	3.000	3.000	3.000	3.000	3.000	3.000
s.d.	0.775	0.765	0.769	0.758	0.748	0.726
RI	0.651	0.651	0.658	0.657	0.672	0.672
per	0.630	0.600	0.680	0.660	0.800	0.760

We next conduct the simulations by selecting λ via minimizing the modified BIC given in (16). Table 4 reports the sample mean, median and standard deviation (s.d.) of the estimated number of groups $\widehat{K}$, the Rand Index (RI) defined in (17) for measuring clustering accuracy, and the percentage (per) of $\widehat{K}$ equaling to the true number of subgroups by the MCP and SCAD methods based on 500 simulation realizations for $c = 2, 3$ and 4. We observe that the median values of $\widehat{K}$ are 3, which is the true number of subgroups, for all cases. Moreover, as the c value becomes larger, the Rand Index (RI) value and the percentage of correctly selecting the number of subgroups increase.

Example 3

(No treatment heterogeneity). We generate data from a model with homogeneous treatment effects given by $y_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n$, where $\mathbf{z}_i$, $\mathbf{x}_i$, ε_i and $\boldsymbol{\eta}$ are simulated in the same way as in Example 1. Set $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathrm{T}} = (2, 2)$ and $n = 200$. We use our proposed penalized estimation method to fit the model. The sample mean of the estimated number of groups $\widehat{K}$ is 1.19 and 1.18, respectively, for the MCP and SCAD methods, and the sample median is 1 for both methods based on 500 replications.

Table 5: The empirical bias (Bias) for the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, and the average asymptotic standard error (ASE) calculated according to Corollary 1 and the empirical standard error (ESE) for the MCP and SCAD methods and oracle estimator (ORACLE) in Example 3.

	β_1	β_2	η_1	η_2	η_3

	Bias	−0.005	0.003	−0.002	0.007	0.003
MCP	ASE	0.035	0.037	0.036	0.037	0.037
	ESE	0.034	0.036	0.041	0.038	0.041
	Bias	−0.004	0.002	−0.001	0.007	0.003
SCAD	ASE	0.035	0.036	0.036	0.037	0.037
	ESE	0.034	0.036	0.040	0.037	0.041
	Bias	−0.004	0.002	−0.001	0.006	0.004
ORACLE	ASE	0.036	0.036	0.037	0.038	0.038
	ESE	0.036	0.037	0.039	0.034	0.039

To evaluate the asymptotic normality established in Corollary 1, Table 5 lists the empirical bias (Bias) for the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, the average asymptotic standard error (ASE) calculated according to Corollary 1, the empirical standard error (ESE) for the MCP, SCAD as well as the oracle estimator (ORACLE). The bias, ASE and ESE for the estimates of $\boldsymbol{\beta}$ 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 500 replications. The biases are small for all cases. The ESE and ASE 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.

6.2 Empirical example

We apply our method to the AIDS Clinical Trials Group Study 175. This study 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 use the log-transformed values of the CD4 counts at 20± 5 weeks as the responses y_i [31]. For illustration of our method, we use didanosine as the treatment variable and use a binary variable x_i for this treatment. Thus, the coefficient of x_i represents the difference in the treatment effect between the two therapies: didanosine and zidovudine. We randomly select 500 patients from the study to consist of our dataset. Moreover, we include 5 baseline covariates in the model, which are age (years), weight (kg), Karnofsky score, log-transformed CD8 counts at baseline, and gender (0 = female, 1 = male).

To see possible heterogeneity in treatment effects, we first fit the homogeneous linear model (3) by using y_i as the response and the 5 baseline covariates and the treatment variable as predictors and obtain the residuals, so that the effects of the baseline covariates are controlled. In Figure 4 (a), it shows the kernel density plot of the residuals. We can see that the distribution has multiple modes for these patients, which indicates possible heterogeneous treatment effects.

Figure 4: The kernel density plot of the residuals (a) by fitting the homogeneous linear model (3) and (b) by fitting the heterogeneous linear model (2).

Next, we fit the heterogeneous model $y_i = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i + \varepsilon_i, i = 1, \ldots, n$, where $\boldsymbol{\beta}_i = (\beta_{1i}, \beta_{2i})^{\mathrm{T}}$, $\boldsymbol{z}_i = (z_{i1} \ldots, z_{i5})^{\mathrm{T}}$ which are the 5 covariates described above, and $\boldsymbol{x}_i = (1, x_i)^{\mathrm{T}}$, in which x_i is the binary variable for the treatment didanosine. All of the predictors are centered and standardized before applying the regularization methods. Figure 5 displays the fusiongrams for the estimated treatment coefficients, $(\hat{\beta}_{21}(\lambda), \ldots, \hat{\beta}_{2n}(\lambda))$, by MCP. We obtain similar patterns by using SCAD. The fusiongrams suggest the existence of heterogeneity in the treatment effects. In particular, the modified BIC criterion selected the λ value in the region that gives two subgroups with different treatment effects. Moreover, Figure 4 (b) shows the kernel density plot of the residuals by fitting the heterogeneous linear model (2). It shows a uni-modal distribution.
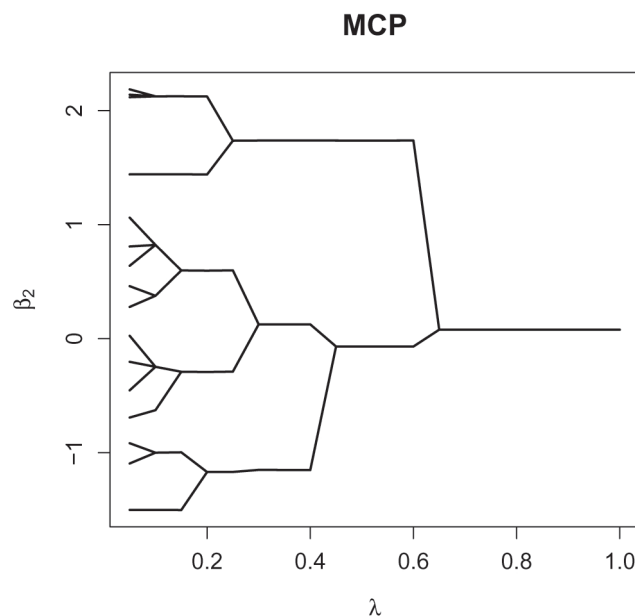
MCP

Figure 5: Fusiongram for $\hat{\boldsymbol{\beta}}_2(\lambda) = \widehat{\beta}_{21}(\lambda), \ldots, \widehat{\beta}_{2n}(\lambda)$ against λ.

Table 6: The estimates (Est.), their standard errors (s.e.) and p-values (P-value) for testing significance 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}(intercept)	α_{12}(trt)	α_{21} (intercept)	α_{22}(trt)
MCP	Est.	5.809	-0.065	5.735	1.687
	s.e.	0.016	0.041	0.055	0.110
	p-value	<0.001	0.112	<0.001	<0.001
SCAD	Est.	5.809	-0.065	5.737	1.688
	s.e.	0.016	0.042	0.055	0.110
	p-value	<0.001	0.122	<0.001	<0.001
OLS	Est.	5.787	0.077	–	–
	s.e.	0.019	0.040	–	–
	p-value	<0.001	0.054	–	–

Let $\hat{\boldsymbol{\alpha}}_1 = (\hat{\alpha}_{11}, \hat{\alpha}_{12})^{\mathrm{T}}$ and $\hat{\boldsymbol{\alpha}}_2 = (\hat{\alpha}_{21}, \hat{\alpha}_{22})^{\mathrm{T}}$ be the estimated coefficients for $\boldsymbol{x}_i$ in the two identified groups $\widehat{\mathscr{G}}_1$ and $\widehat{\mathscr{G}}_2$, respectively, so that $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\alpha}}_1$ for $i \in \widehat{\mathscr{G}}_1$ and $\hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\alpha}}_2$ for $i \in \widehat{\mathscr{G}}_2$. In Table 6, we report the estimates (Est.) and their standard errors (s.e.) of $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22})^{\mathrm{T}}$ by the MCP and SCAD methods, and those values of $\boldsymbol{\beta} = \boldsymbol{\alpha}_1$ by the OLS method. Note that the first and second components represent the coefficients for intercept and the treatment variable (trt), respectively. We also report the p-values (p-value) for testing whether each coefficient is zero or not. By the OLS method, the p-value for testing whether the coefficient of the therapy (didanosine), α_{12}, is zero or not is greater than 0.05. This result indicates that the effect of the therapy (didanosine) is not significantly different from that of the standard treatment (zidovudine) by assuming that the treatment effect is homogeneous. In comparison, by the proposed concave fusion approach, the p-value for testing significance of the treatment coefficient, α_{12}, for one group is greater than 0.05, but the p-value for the other group is less than 0.001. This result implies that although the treatment has no statistically significant effect on one group of patients, its effect on the other group is prominent.

Next, we employ the bootstrapping method given in Section 5 to test homogeneity of the treatment effect. We obtain the p-value <0.01. Thus, the heterogeneity of the treatment effect is further confirmed by the inference procedure. Lastly, we apply the method to model (2) with all three therapies by letting $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$, where $x_{i1} =$ zidovudine and didanosine (trt1), $x_{i2} =$ zidovudine and zalcitabine (trt2), and $x_{i3} =$ didanosine (trt3). Let $\hat{\boldsymbol{\alpha}}_1 = (\hat{\alpha}_{11}, \hat{\alpha}_{12}, \hat{\alpha}_{13}, \hat{\alpha}_{14})^{\mathrm{T}}$ and $\hat{\boldsymbol{\alpha}}_2 = (\hat{\alpha}_{21}, \hat{\alpha}_{22}, \hat{\alpha}_{23}, \hat{\alpha}_{24})^{\mathrm{T}}$ be the estimated coefficients for $\boldsymbol{x}_i$ in the two identified groups $\widehat{\mathscr{G}}_1$ and $\widehat{\mathscr{G}}_2$, respectively. In Table 7, we report the estimates (Est.) and their standard errors (s.e.) of $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{14})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{24})^{\mathrm{T}}$ by the MCP and SCAD methods, and those values of $\boldsymbol{\beta} = \boldsymbol{\alpha}_1$ by the OLS method. We obtain two subgroups as well by our proposed method. Moreover, the two methods,

MCP and SCAD, have similar results, and the effects of the three therapies are more significantly different from the standard therapy in one group than they are in the other group.

Table 7: The estimates (Est.), their standard errors (s.e.) and p-values (P-value) for testing significance of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ by MCP and SCAD, and those values obtained by OLS.

		α_{11}	α_{12}	α_{13}	α_{14}	α_{21}	α_{22}	α_{23}	α_{24}
MCP	Est.	3.985	0.114	0.104	0.126	5.525	1.116	1.200	1.214
	s.e.	0.086	0.049	0.051	0.052	0.074	0.103	0.106	0.107
	p-value	<0.001	0.020	0.041	0.015	<0.001	<0.001	<0.001	<0.001
SCAD	Est.	3.985	0.114	0.105	0.127	5.524	1.116	1.199	1.214
	s.e.	0.086	0.047	0.051	0.052	0.074	0.102	0.106	0.107
	p-value	<0.001	0.015	0.040	0.015	<0.001	<0.001	<0.001	<0.001
OLS	Est.	5.846	0.069	0.067	0.065	–	–	–	–
	s.e.	0.019	0.024	0.024	0.024	–	–	–	–
	p-value	<0.001	0.004	0.005	0.007	–	–	–	–

7 Discussion

It will be of interest to extend the proposed method to a more general class of regression problems including generalized linear and Cox regression models. Moreover, it is also possible to relax the linearity assumption on the covariates $\mathbf{z}_i$ by considering a nonparametric or semiparametric functional form of $\mathbf{z}_i$. For these more complicated models, the ADMM algorithm can still be employed with some modifications. However, further work is needed to study the theoretical properties. We refer to Section A.2 for detailed discussions on these extensions. Also, we have assumed that the number of treatment variables p and the number of confounding variables q are both smaller than the sample size n, although we allow p and q to diverge with n. For high-dimensional problems with $p > n$ or $q > n$, a sparsity condition on the coefficients would be needed to ensure identifiability of the model. Further studies are needed to develop computational algorithms and theoretical results for the high-dimensional setting.

Funding

A Appendix

In the Appendix, we give the computational details, the convergence property of the ADMM algorithm and the technical proofs for Theorem 1–Theorem 3.

A.1 Computation

A.1.1 ADMM with concave penalties

We derive an ADMM algorithm for computing the solution (7). 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 (6) as that of minimizing

$$L_0(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{1}{2}\sum_{i=1}^n (y_i - \mathbf{z}_i^T\boldsymbol{\eta} - \mathbf{x}_i^T\boldsymbol{\beta}_i)^2 + \sum_{i<j} p_\gamma(\boldsymbol{\delta}_{ij}, \lambda),$$

$$\text{subject to } \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij} = \mathbf{0},$$

(20)

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

$$L(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{v}) = L_0(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}) + \sum_{i<j} \left\langle \boldsymbol{v}_{ij}, \boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij} \right\rangle + \frac{\vartheta}{2} \sum_{i<j} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j - \boldsymbol{\delta}_{ij}\|^2, \tag{21}$$

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

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

$$(\boldsymbol{\eta}^{m+1}, \boldsymbol{\beta}^{m+1}) = \underset{\boldsymbol{\eta}, \boldsymbol{\beta}}{\operatorname{argmin}} \, L(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}^m, \boldsymbol{v}^m), \tag{22}$$

$$\boldsymbol{\delta}^{m+1} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \, L(\boldsymbol{\eta}^{m+1}, \boldsymbol{\beta}^{m+1}, \boldsymbol{\delta}, \boldsymbol{v}^m), \tag{23}$$

$$\boldsymbol{v}_{ij}^{m+1} = \boldsymbol{v}_{ij}^m + \vartheta(\boldsymbol{\beta}_i^{m+1} - \boldsymbol{\beta}_j^{m+1} - \boldsymbol{\delta}_{ij}^{m+1}). \tag{24}$$

In (22), the problem is equivalent to the minimization of the function

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

where C is a constant independent of $(\boldsymbol{\eta}, \boldsymbol{\beta})$. Some algebra shows that we can write $f(\boldsymbol{\eta}, \boldsymbol{\beta})$ as

$$f(\boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{Z}\boldsymbol{\eta} + \mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 + \frac{\vartheta}{2}\|\mathbf{A}\boldsymbol{\beta} - \boldsymbol{\delta}^m + \vartheta^{-1}\boldsymbol{v}^m\|^2 + C, \tag{25}$$

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 $\boldsymbol{v}^m$ at the mth step, the updates $\boldsymbol{\beta}^{m+1}$ and $\boldsymbol{\eta}^{m+1}$ are

$$\begin{aligned}
\boldsymbol{\beta}^{m+1} &= (\mathbf{X}^{\mathrm{T}}\mathbf{Q}_Z\mathbf{X} + \vartheta\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}[\mathbf{X}^{\mathrm{T}}\mathbf{Q}_Z\mathbf{y} + \vartheta\mathbf{A}^{\mathrm{T}}(\boldsymbol{\delta}^m - \vartheta^{-1}\boldsymbol{v}^m)], \\
\boldsymbol{\eta}^{m+1} &= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{m+1}),
\end{aligned} \tag{26}$$

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

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

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

$$\boldsymbol{\beta}^{m+1} = (\mathbf{X}^{\mathrm{T}}\mathbf{Q}_Z\mathbf{X} + \vartheta\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}[\mathbf{X}^{\mathrm{T}}\mathbf{Q}_Z\mathbf{y} + \vartheta\operatorname{vec}((\boldsymbol{\Delta}^m - \vartheta^{-1}\mathbf{Y}^m)D)]. \tag{27}$$

In (23), after discarding the terms independent of $\boldsymbol{\delta}$, we need to minimize

$$\frac{\vartheta}{2}\|\boldsymbol{\zeta}_{ij}^m - \boldsymbol{\delta}_{ij}\|^2 + p_\gamma(\|\boldsymbol{\delta}_{ij}\|, \lambda) \tag{28}$$

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

For the L_1 penalty, the solution is

$$\boldsymbol{\delta}_{ij}^{m+1} = S(\boldsymbol{\zeta}_{ij}^m, \lambda/\vartheta), \tag{29}$$

where $S(\mathbf{z}, t) = (1 - t/\|\mathbf{z}\|)_+\mathbf{z}$ 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

$$\boldsymbol{\delta}_{ij}^{m+1} = \begin{cases} \dfrac{S(\boldsymbol{\zeta}_{ij}^m, \lambda/\vartheta)}{1 - 1/(\gamma\vartheta)} & \text{if } \|\boldsymbol{\zeta}_{ij}^m\| \le \gamma\lambda, \\ \boldsymbol{\zeta}_{ij} & \text{if } \|\boldsymbol{\zeta}_{ij}^m\| > \gamma\lambda. \end{cases} \tag{30}$$

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

$$\boldsymbol{\delta}_{ij}^{m+1} = \begin{cases} S(\boldsymbol{\zeta}_{ij}^m, \lambda/\vartheta) & \text{if } \|\boldsymbol{\zeta}_{ij}^m\| \le \lambda + \lambda/\vartheta, \\ \dfrac{S(\boldsymbol{\zeta}_{ij}^m, \gamma\lambda/((\gamma-1)\vartheta))}{1 - 1/((\gamma-1)\vartheta)} & \text{if } \lambda + \lambda/\vartheta < \|\boldsymbol{\zeta}_{ij}^m\| \le \gamma\lambda, \\ \boldsymbol{\zeta}_{ij}^m & \text{if } \|\boldsymbol{\zeta}_{ij}^m\| > \gamma\lambda. \end{cases} \tag{31}$$

Finally, the update of $\boldsymbol{v}_{ij}$ is given in (24).

We summarize the above analysis in Algorithm 1.

Algorithm 1 ADMM for concave fusion

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

 1:   **for** $m = 0, 1, 2, \cdots$ **do**
 2:      Compute $\boldsymbol{\beta}^{m+1}$ using (6)
 3:      Compute $\boldsymbol{\eta}^{m+1}$ (26)
 4:      Compute $\boldsymbol{\delta}^{m+1}$ (29), (30) or (13)
 5:      Compute $\boldsymbol{v}^{m+1}$ using (24)
 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 7

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 $\widehat{\mathscr{G}}_1, \ldots, \widehat{\mathscr{G}}_{\widehat{K}}$. The estimated treatment effect for the kth group is $\widehat{\boldsymbol{\alpha}}_k = |\widehat{\mathscr{G}}_k|^{-1} \sum_{i \in \widehat{\mathscr{G}}_k} \widehat{\boldsymbol{\beta}}_i$, where $|\widehat{\mathscr{G}}_k|$ is the cardinality of $\widehat{\mathscr{G}}_k$.

Remark 8

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

Remark 9

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.

A.1.2   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

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

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 (25). The solutions are

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

$$\boldsymbol{\eta}_R(\lambda^*) = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_R(\lambda^*)),$$

where $\mathbf{Q}_Z$ is given in (27). 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 $\boldsymbol{v}^0 = \mathbf{0}$.

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

$(\hat{\boldsymbol{\eta}}(\lambda_0), \hat{\boldsymbol{\beta}}(\lambda_0))$ using $(\boldsymbol{\eta}^0, \boldsymbol{\beta}^0)$ as the initial value. Then compute $(\hat{\boldsymbol{\eta}}(\lambda_k), \hat{\boldsymbol{\beta}}(\lambda_k))$ using $(\hat{\boldsymbol{\eta}}(\lambda_{k-1}), \hat{\boldsymbol{\beta}}(\lambda_{k-1}))$ 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 [32], where the algorithms start at the λ value that forces all the coefficients to zero.

A.1.3   Convergence of the algorithm

We next derive the convergence properties of the ADMM algorithm.

Proposition 1

Let $\mathbf{r}^m = \mathbf{A}\boldsymbol{\beta}^m - \boldsymbol{\delta}^m$ and $\mathbf{s}^{m+1} = \vartheta \mathbf{A}^{\mathrm{T}}(\boldsymbol{\delta}^{m+1} - \boldsymbol{\delta}^m)$ be the primal residual and the 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.

This proposition shows that the primal feasibility and the dual feasibility are achieved by the algorithm.

Proof. By the definition of $\boldsymbol{\delta}^{m+1}$, we have

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

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

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

$$= \inf_{\mathbf{A}\boldsymbol{\beta}^{m+1} - \boldsymbol{\delta} = \mathbf{0}} L(\boldsymbol{\eta}^{m+1}, \boldsymbol{\beta}^{m+1}, \boldsymbol{\delta}, \boldsymbol{v}^m).$$

Then

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

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

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

and thus

$$L(\boldsymbol{\eta}^{m+t}, \boldsymbol{\beta}^{m+t}, \boldsymbol{\delta}^{m+t}, \boldsymbol{v}^{m+t-1})$$

$$= \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta}^{m+t} - \mathbf{X}\boldsymbol{\beta}^{m+t}\|^2 + (\boldsymbol{v}^{m+t-1})^{\mathrm{T}}(\mathbf{A}\boldsymbol{\beta}^{m+t} - \boldsymbol{\delta}^{m+t})$$

$$+ \frac{\vartheta}{2}\|\mathbf{A}\boldsymbol{\beta}^{m+t} - \boldsymbol{\delta}^{m+t}\|^2 + \sum_{i<j} p_\gamma(|\boldsymbol{\delta}_{ij}^{m+t}|, \lambda)$$

$$= \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta}^{m+t} - \mathbf{X}\boldsymbol{\beta}^{m+t}\|^2 + (\boldsymbol{v}^m)^{\mathrm{T}}(\mathbf{A}\boldsymbol{\beta}^{m+t} - \boldsymbol{\delta}^{m+t})$$

$$+ \vartheta \sum_{i=1}^{t-1}(\mathbf{A}\boldsymbol{\beta}^{m+i} - \boldsymbol{\delta}^{m+i})^{\mathrm{T}}(\mathbf{A}\boldsymbol{\beta}^{m+t} - \boldsymbol{\delta}^{m+t})$$

$$+ \frac{\vartheta}{2}\|\mathbf{A}\boldsymbol{\beta}^{m+t} - \boldsymbol{\delta}^{m+t}\|^2 + p_\gamma(|\boldsymbol{\delta}_{ij}^{m+t}|, \lambda)$$

$$\leq f^{m+t}.$$

Since the objective function $L(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{v})$ 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 [33], 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

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

and for all $t \geq 0$

$$\lim_{m\to\infty} L(\boldsymbol{\mu}^{m+t}, \boldsymbol{\beta}^{m+t}, \boldsymbol{\eta}^{m+t}, \boldsymbol{v}^{m+t-1})$$

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

$$\leq f^*.$$

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

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

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

and moreover,

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

Therefore,

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

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

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

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

A.2 Extension to nonlinear models

In this paper, we focus on the linear regression model (2) to introduce our proposed method for exploring treatment heterogeneity. It is worth mentioning that our method can be readily extended to semi-parametric models by relaxing the linearity assumption on $\mathbf{z}_i$. Considering the model:

$$y_i = m(\mathbf{z}_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i + \varepsilon_i, i = 1, \ldots, n,$$

where $m(\cdot)$ is an unknown function of $\mathbf{z}_i$. This model has no constraint on the functional form of $\mathbf{z}_i$. Following the approach in Ma et al. [34], we can estimate $m(\cdot)$ by tensor-product regression splines weighted by categorical kernel functions, where the splines are used for the continuous predictors and the categorical kernels are for the discrete predictors, respectively. Then the objective function for obtaining the estimates consists of two parts similar as given in (6). The first part is a weighted least squares criterion as presented in eq. (2) of Ma et al. [34], and the second part contains the same penalty functions as given in (6). As a result, the proposed ADMM algorithm proposed in Section A.1 can be applied. Moreover, we can also assume semi-parametric structures on $m(\cdot)$ for further dimensionality reduction while retaining model flexibility. For example, when $\mathbf{z}_i$ is a set of continuous variables, we can assume the additive structure:

$$m(\mathbf{z}_i) = m_1(z_{i1}) + \cdots + m_q(z_{iq}),$$

where $m_k(\cdot)$ for $k = 1, \ldots, q$ are unknown functions. Also the partially linear additive structure is another popular semi-parametric model, given as

$$m(\mathbf{z}_i) = m_1(z_{i1}) + \cdots + m_{q_1}(z_{iq_1}) + \mathbf{z}_{i2}^{\mathrm{T}}\boldsymbol{\eta},$$

where $\mathbf{z}_i = (\mathbf{z}_{i1}^{\mathrm{T}}, \mathbf{z}_{i2}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{z}_{i1} = (z_{i1}, \ldots, z_{iq_1})^{\mathrm{T}}$ and $\mathbf{z}_{i2} = (z_{i,q_1+1}, \ldots, z_{iq})^{\mathrm{T}}$. If we use regression splines to approximate the unknown functions, the same ADMM algorithm given in Section A.1 can be applied to obtain the parameter estimators with the components in $\mathbf{z}_i$ replaced by their spline basis functions. We refer to Ma et al. [35] for the details of using regression splines to estimate unknown functions in these settings.

It is also of interest to extend the proposed method to the case with discrete responses. For a general scenario, one may fit a generalized linear model:

$$E(y_i|\mathbf{z}_i, \mathbf{x}_i) = \mu_i = g^{-1}(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i), i = 1, \ldots, n,$$

where g is a known monotone link function. For obtaining the estimates of the parameters, we can consider the negative quasi-likelihood function

$$Q(\mu, y) = \int_{\mu}^{y} \{(y - \xi)/V(\xi)\}d\xi,$$

where $V(\cdot)$ is the conditional variance of y given $(\mathbf{z}, \mathbf{x})$. Then the parameter estimates can be obtained by minimizing

$$\sum_{i=1}^{n} Q(g^{-1}(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i), y_i) + \sum_{1 \le i < j \le n} p(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|, \lambda).$$

Because $Q(g^{-1}(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\eta} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}_i), y_i)$ is differentiable with respect to $(\boldsymbol{\eta}, \boldsymbol{\beta}_i)$, the ADMM algorithm given in Section A.1 can be straightforwardly applied.

A.3   Proof of Theorem 1

In this section we show the results in Theorem 1. For every $\boldsymbol{\beta} \in \mathcal{M}_{\mathscr{G}}$, it can be written as $\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\alpha}$. Recall $\mathbf{U} = (\mathbf{Z}, \mathbf{XW})$. We have

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

Thus

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

Then

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

Hence

$$\left\| \begin{pmatrix} \hat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0 \\ \hat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0 \end{pmatrix} \right\| \le \left\|(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\right\| \left\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon}\right\|. \tag{32}$$

By Condition (C1), we have

$$\left\|(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\right\| \le C_1^{-1}|\mathscr{G}_{\min}|^{-1}. \tag{33}$$

Moreover

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

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

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

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

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

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),

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

It follows that

$$P(\left\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon}\right\|_{\infty} > C\sqrt{n\log n}) \leq 2(Kp + q)n^{-c_1 C^2}.$$

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

$$P(\left\|\mathbf{U}^{\mathrm{T}}\boldsymbol{\varepsilon}\right\| > C\sqrt{q + Kp}\sqrt{n\log n}) \leq 2(Kp + q)n^{-c_1 C^2}. \tag{34}$$

Therefore, by (32), (33) and (34), we have with probability at least $1 - 2(Kp + q)n^{-c_1 C^2}$,

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

The result (9) in Theorem 1 is proved by letting $C = c_1^{-1/2}$. Moreover,

$$\left\|\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0\right\|^2 = \sum_{k=1}^{K}\sum_{i\in\mathscr{G}_k}\left\|\widehat{\boldsymbol{\alpha}}_k^{or} - \boldsymbol{\alpha}_k^0\right\|^2 \leq |\mathscr{G}_{\max}|\sum_{k=1}^{K}\left\|\widehat{\boldsymbol{\alpha}}_k^{or} - \boldsymbol{\alpha}_k^0\right\|^2$$
$$= |\mathscr{G}_{\max}|\left\|\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0\right\|^2 \leq |\mathscr{G}_{\max}|\phi_n^2,$$

and

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

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

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

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 $\mathbf{a}_n \in \mathrm{IR}^{q+Kp}$ with $\|\mathbf{a}_n\| = 1$,

$$\mathrm{var}\{\mathbf{a}_n^{\mathrm{T}}((\widehat{\boldsymbol{\eta}}^{or} - \boldsymbol{\eta}^0)^{\mathrm{T}}, (\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^{\mathrm{T}})^{\mathrm{T}}\}$$
$$= \sigma_n^2(\mathbf{a}_n) = \sigma^2\left[\mathbf{a}_n^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1}\mathbf{a}_n\right] \geq \sigma^2\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{a}_n.$$

Moreover, for any $\varepsilon > 0$,

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

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

$$\sum_{i=1}^{n}\{E(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i)^4\}^{1/2} \leq c^{1/2}\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{a}_n.$$

Moreover,

$$\max_i P\left\{|\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i| > \sigma_n(\mathbf{a}_n)\right\}$$
$$\leq \max_i E(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i)^2/\{{}^2\sigma_n^2(\mathbf{a}_n)\} \leq c'^{-2}(q + Kp)\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\boldsymbol{\Xi}_n^{-1}\mathbf{a}_n/(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{a}_n)$$

for some constant $c' \in (0, \infty)$. Therefore, by the above results, we have

$$\sigma_n^{-2}(\mathbf{a}_n)\sum_{i=1}^{n}E[(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i)^2 \cdot 1_{\{|\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{U}_i\varepsilon_i|>\in\sigma_n(\mathbf{a}_n)\}}]$$
$$\leq \{\sigma^2\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{a}_n\}^{-1}c^{1/2}\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{a}_n\{c'\in^{-2}(q + Kp)\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\boldsymbol{\Xi}_n^{-1}\mathbf{a}_n/(\mathbf{a}_n^{\mathrm{T}}\boldsymbol{\Xi}_n^{-1}\mathbf{a}_n)\}^{1/2}$$
$$\leq c^{1/2}c'^{1/2}C_1^{-1/2}\in^{-1}\sigma^{-1}(q + Kp)^{1/2}|\mathscr{G}_{\min}|^{-1/2} = o(1).$$

Then, the result (11) follows from Lindeberg–Feller Central Limit Theorem.

A.4 Proof of Theorem 2

In this section we show the results in Theorem 2. Define

$$L_n(\boldsymbol{\eta}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}\|^2, P_n(\boldsymbol{\beta}) = \lambda \sum_{i<j} \rho(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|),$$

$$L_n^{\mathscr{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\mathbf{W}\boldsymbol{\alpha}\|^2, P_n^{\mathscr{G}}(\boldsymbol{\alpha}) = \lambda \sum_{k<k'} |\mathscr{G}_k||\mathscr{G}_{k'}|\rho(\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\|),$$

and let

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

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

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

$$Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) = Q_n^{\mathscr{G}}(\boldsymbol{\eta}, T(\boldsymbol{\beta})), Q_n^{\mathscr{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha}) = Q_n(\boldsymbol{\eta}, T^{-1}(\boldsymbol{\alpha})). \tag{35}$$

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

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

By Theorem 1, there exists an event E_1 in which

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

and $P(E_1^C) \le 2(q + Kp)n^{-1}$. Hence $(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or}) \in \Theta$ in E_1. For any $\boldsymbol{\beta} \in R^{np}$, let $\boldsymbol{\beta}^* = T^{-1}(T^*(\boldsymbol{\beta}))$. We show that $(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ is a strictly local minimizer of the objective function (6) with probability approaching 1 through the following two steps.

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

ii. There is an event E_2 such that $P(E_2^C) \le 2n^{-1}$. In $E_1 \cap E_2$, there is a neighborhood of $((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, denoted by Θ_n such that $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) \ge Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*)$ 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 $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) > Q_n(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ for any $(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$ and $((\boldsymbol{\eta})^{\mathrm{T}}, (\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}} \ne ((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ in $E_1 \cap E_2$, so that $((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ is a strict local minimizer of $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta})$ (6) over the event $E_1 \cap E_2$ with $P(E_1 \cap E_2) \ge 1 - 2(q + Kp + 1)n^{-1}$ for sufficiently large n.

In the following we prove the result in (i). We first show $P_n^{\mathscr{G}}(T^*(\boldsymbol{\beta})) = C_n$ 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 $P_n^{\mathscr{G}}(T^*(\boldsymbol{\beta}))$ is a constant. Since

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

and

$$\begin{aligned}
\sup_k \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\|^2 &= \sup_k \left\||\mathscr{G}_k|^{-1}\sum_{i\in\mathscr{G}_k}\boldsymbol{\beta}_i - \boldsymbol{\alpha}_k^0\right\|^2 = \sup_k \left\||\mathscr{G}_k|^{-1}\sum_{i\in\mathscr{G}_k}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0)\right\|^2 \\
&= \sup_k |\mathscr{G}_k|^{-2}\left\|\sum_{i\in\mathscr{G}_k}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0)\right\|^2 \le \sup_k |\mathscr{G}_k|^{-1}\sum_{i\in\mathscr{G}_k}\left\|(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0)\right\|^2 \\
&\le \sup_i \left\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\right\|^2 \le \phi_n^2,
\end{aligned} \tag{36}$$

then for all k and k'

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

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 $((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ is the unique global minimizer of $L_n^{\mathcal{G}}(\boldsymbol{\eta}, \boldsymbol{\alpha})$, then $L_n^{\mathcal{G}}(\boldsymbol{\eta}, T^*(\boldsymbol{\beta})) > L_n^{\mathcal{G}}(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or})$ for all $(\boldsymbol{\eta}^{\mathrm{T}}, (T^*(\boldsymbol{\beta}))^{\mathrm{T}})^{\mathrm{T}} \neq ((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ and hence $Q_n^{\mathcal{G}}(\boldsymbol{\eta}, T^*(\boldsymbol{\beta})) > Q_n^{\mathcal{G}}(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or})$ for all $T^*(\boldsymbol{\beta}) \neq \hat{\boldsymbol{\alpha}}^{or}$. By (35), we have $Q_n^{\mathcal{G}}(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\alpha}}^{or}) = Q_n(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ 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, $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) > Q_n(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ for all $\boldsymbol{\beta}^* \neq \hat{\boldsymbol{\beta}}^{or}$, and the result in (i) is proved.

Next we prove the result in (ii). For a positive sequence t_n, let $\Theta_n = \{\boldsymbol{\beta}_i : \sup_i \|\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i^{or}\| \le t_n\}$. 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

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

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

$$
\begin{aligned}
\Gamma_2 &= \lambda \sum_{\{j>i\}} \rho'(\|\beta_i^m - \beta_j^m\|)\|\beta_i^m - \beta_j^m\|^{-1}(\beta_i^m - \beta_j^m)^{\mathrm{T}}(\beta_i - \beta_i^*) \\
&\quad + \lambda \sum_{\{j<i\}} \rho'(\|\beta_i^m - \beta_j^m\|)\|\beta_i^m - \beta_j^m\|^{-1}(\beta_i^m - \beta_j^m)^{\mathrm{T}}(\beta_i - \beta_i^*) \\
&= \lambda \sum_{\{j>i\}} \rho'(\|\beta_i^m - \beta_j^m\|)\|\beta_i^m - \beta_j^m\|^{-1}(\beta_i^m - \beta_j^m)^{\mathrm{T}}(\beta_i - \beta_i^*) \\
&\quad + \lambda \sum_{\{i<j\}} \rho'(\|\beta_j^m - \beta_i^m\|)\|\beta_j^m - \beta_i^m\|^{-1}(\beta_j^m - \beta_i^m)^{\mathrm{T}}(\beta_j - \beta_j^*) \\
&= \lambda \sum_{\{j>i\}} \rho'(\|\beta_i^m - \beta_j^m\|)\|\beta_i^m - \beta_j^m\|^{-1}(\beta_i^m - \beta_j^m)^{\mathrm{T}}\{(\beta_i - \beta_i^*) - (\beta_i - \beta_i^*)\}.
\end{aligned}
$$
(37)

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,

$$\sup_i \|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i^0\|^2 = \sup_k \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\|^2 \le \phi_n^2,$$
(38)

where the last inequality follows from (36). Since $\boldsymbol{\beta}_i^m$ is between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_i^*$,

$$\sup_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_i^0\| \le \alpha \sup_i \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0\| + (1 - \alpha) \sup_i \|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_i^0\| \le \alpha\phi_n + (1 - \alpha)\phi_n = \phi_n.$$
(39)

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

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

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 (36), we have

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

Then

$$\sup_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\| \leq 2 \sup_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_i^*\| \leq 2 \sup_i \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\|$$

$$\leq 2(\sup_i \|\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i^{or}\| + \sup_i \|\boldsymbol{\beta}_i^* - \hat{\boldsymbol{\beta}}_i^{or}\|) \leq 4 \sup_i \|\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i^{or}\| \leq 4t_n.$$

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

$$\Gamma_2 \geq \sum_{k=1}^K \sum_{\{i,j \in \mathscr{G}_k, i<j\}} \lambda \rho'(4t_n) \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|. \tag{40}$$

Let

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

Then

$$\Gamma_1 = -\mathbf{Q}^T(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = -\sum_{k=1}^K \sum_{\{i,j \in \mathscr{G}_k\}} \frac{\mathbf{Q}_i^T(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)}{|\mathscr{G}_k|}$$

$$= -\sum_{k=1}^K \sum_{\{i,j \in \mathscr{G}_k\}} \frac{\mathbf{Q}_i^T(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)}{2|\mathscr{G}_k|} - \sum_{k=1}^K \sum_{\{i,j \in \mathscr{G}_k\}} \frac{\mathbf{Q}_i^T(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)}{2|\mathscr{G}_k|}$$

$$= -\sum_{k=1}^K \sum_{\{i,j \in \mathscr{G}_k\}} \frac{(\mathbf{Q}_j - \mathbf{Q}_i)^T(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)}{2|\mathscr{G}_k|}$$

$$= -\sum_{k=1}^K \sum_{\{i,j \in \mathscr{G}_k, i<j\}} \frac{(\mathbf{Q}_j - \mathbf{Q}_i)^T(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)}{|\mathscr{G}_k|}. \tag{41}$$

Moreover,

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

and then

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

By Condition (C1) that $\sup_i \|\mathbf{x}_i\| \leq C_2\sqrt{p}$ and $\sup_i \|\mathbf{z}_i\| \leq C_3\sqrt{q}$, (39) that $\sup_i \|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_i^m\| \leq \phi_n$ and $\|\boldsymbol{\eta}^0 - \boldsymbol{\eta}\| \leq \phi_n$, we have

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

By Condition (C3)

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

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

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

Then

$$|\frac{(\mathbf{Q}_j - \mathbf{Q}_i)^T(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)}{|\mathscr{G}_k|}|$$

$$\leq |\mathscr{G}_{\min}|^{-1}\mathbf{Q}_j - \mathbf{Q}_i\boldsymbol{\beta}_i - \boldsymbol{\beta}_j \leq |\mathscr{G}_{\min}|^{-1}2\sup_i \mathbf{Q}_i\boldsymbol{\beta}_i - \boldsymbol{\beta}_j \tag{42}$$

$$\leq 2C_2|\mathscr{G}_{\min}|^{-1}\sqrt{p}(\sqrt{2c_1^{-1}}\sqrt{\log n} + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|.$$

Therefore, by (40), (41) and (42), we have

$$Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) - Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*)$$

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

Let $t_n = o(1)$, then $\rho'(4t_n) \to 1$. Since $\lambda \gg \phi_n$, $p = o(n)$, and $|\mathscr{G}_{\min}|^{-1}p = o(1)$, then $\lambda \gg |\mathscr{G}_{\min}|^{-1}\sqrt{p}\sqrt{\log n}$, $\lambda \gg |\mathscr{G}_{\min}|^{-1}\sqrt{pq}$ and $\lambda \gg |\mathscr{G}_{\min}|^{-1}p\phi_n$. 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.

A.5 Proof of Theorem 3

In this section we show the results in Theorem 3. The proofs of (13) and (14) follow the same arguments as the proof of Theorem 1 by letting $\widetilde{\mathbf{X}} = \mathbf{x}$ and $|\mathcal{G}_{\min}| = n$, and thus they are omitted. Next, we will show (15). It follows similar procedures as the proof of Theorem 2 with the details given below. Define $\mathcal{M} = \{\boldsymbol{\beta} \in \mathrm{IR}^{np} : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_n\}$. 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 $(\boldsymbol{\eta}, \boldsymbol{\beta}^0)$:

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

where $\phi_n = c_1^{-1/2} C_1^{-1} \sqrt{q + p} \sqrt{n^{-1} \log n}$. By the result in (13), there exists an event E_1 such that on the event E_1,

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

and $P(E_1^C) \le 2(q + p)n^{-1}$. Hence $(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or}) \in \Theta$ on the event E_1. For any $\boldsymbol{\beta} \in R^{np}$, let $\boldsymbol{\beta}^* = T^{-1}(T^*(\boldsymbol{\beta}))$. We show that $(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ is a strictly local minimizer of the objective function (6) with probability approaching 1 through the following two steps.

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

(ii). There is an event E_2 such that $P(E_2^C) \le 2n^{-1}$. On $E_1 \cap E_2$, there is a neighborhood of $((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$, denoted by Θ_n such that $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) \ge Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*)$ 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 $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) > Q_n(\hat{\boldsymbol{\eta}}^{or}, \hat{\boldsymbol{\beta}}^{or})$ for any $(\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} \in \Theta_n \cap \Theta$ and $((\boldsymbol{\eta})^{\mathrm{T}}, (\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}} \ne ((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ in $E_1 \cap E_2$, so that $((\hat{\boldsymbol{\eta}}^{or})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}^{or})^{\mathrm{T}})^{\mathrm{T}}$ is a strict local minimizer of $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta})$ (6) on the event $E_1 \cap E_2$ with $P(E_1 \cap E_2) \ge 1 - 2(q + p + 1)n^{-1}$ for sufficiently large n.

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

Next we prove the result in (ii). For a positive sequence t_n, let $\Theta_n = \{\boldsymbol{\beta}_i : \sup_i \|\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i^{or}\| \le t_n\}$. 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

$$\Gamma_1 = -(\mathbf{y} - \mathbf{Z}\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}^m)^{\mathrm{T}}\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$$

$$\Gamma_2 = \sum_{i=1}^{n} \frac{\partial P_n(\boldsymbol{\beta}^m)}{\partial \boldsymbol{\beta}_i^{\mathrm{T}}}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*).$$

$P_n(\boldsymbol{\beta}) = \lambda \sum_{i<j} \rho\left(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|\right)$, and $\boldsymbol{\beta}^m = a\boldsymbol{\beta} + (1 - a)\boldsymbol{\beta}^*$ for some constant $a \in (0, 1)$. Moreover, by (37),

$$\Gamma_2 = \lambda \sum_{\{j>i\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|)\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|^{-1}(\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m)^{\mathrm{T}}\{(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) - (\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*)\}$$

$$= \lambda \sum_{\{j>i\}} \rho'(\|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\|)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|,$$

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 (38), we have

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

Then

$$\sup_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_j^m\| \le 2\sup_i \|\boldsymbol{\beta}_i^m - \boldsymbol{\beta}_i^*\| \le 2\sup_i \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\|$$

$$\le 2(\sup_i \|\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i^{or}\| + \sup_i \|\boldsymbol{\beta}_i^* - \hat{\boldsymbol{\beta}}_i^{or}\|) \le 4\sup_i \|\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i^{or}\| \le 4t_n.$$

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

$$\Gamma_2 \ge \sum_{\{i<j\}} \lambda \rho'(4t_n)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|. \tag{43}$$

Let

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

By the same reasoning as the proof for (41), we have

$$\Gamma_1 = -\mathbf{Q}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = -n^{-1}\sum_{\{i<j\}}(\mathbf{Q}_j - \mathbf{Q}_i)^{\mathrm{T}}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i). \tag{44}$$

By the same argument as the proof for (42), we have that there is an event E_2 such that $P(E_2^C) \le 2n^{-1}$, and on the event E_2,

$$\begin{aligned} &n^{-1}|(\mathbf{Q}_j - \mathbf{Q}_i)^{\mathrm{T}}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i) \\ &\le 2C_2 n^{-1}\sqrt{p}(\sqrt{2c_1^{-1}}\sqrt{\log n} + C_3\sqrt{q}\phi_n + C_2\sqrt{p}\phi_n)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|. \end{aligned} \tag{45}$$

Therefore, by (43), (44) and (45), we have

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

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, $Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}) - Q_n(\boldsymbol{\eta}, \boldsymbol{\beta}^*) \ge 0$ for sufficiently large n, so that the result in (ii) is proved.

References

[1] Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q. 2004;82:661–87.

[2] Sorensen T. Which patients may be harmed by good treatments? Lancet. 1996;348:351–2.

[3] Simon R. Clinical trials for predictive medicine: new challenges and paradigms. Clin Trials. 2010;7:516–24.

[4] Zhang Z, Nie L, Soon G, Liu A. The use of covariates and random effects in evaluating predictive biomarkers under a potential outcome framework. Ann Appl Stat. 2014;8:2336–55.

[5] Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics. 1985;41:361–72.

[6] Lagakos SW. The challenge of subgroup analysis: Reporting without distorting. N Engl J Med. 2006;354:1667–9.

[7] Rothwell PM. Subgroup analysis in randomized clinical trials: importance, indications and interpretation. Lancet. 365:176–86.

[8] Russek-Cohen E, Simon R. Evaluating treatments when a gender by treatment interaction may exist. Stat Med. 1998;16:455–64.

[9] Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011;12:270–82.

[10] Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A simple method for estimating interactions between a treatment and a large number of covariates. JASA. 2014;109:1517–32.

[11] Zhang Z, Qu Y, Zhang B, Nie L, Soon G. Use of auxiliary covariates in estimating a biomarker-adjusted treatment effect model with clinical trial data. Stat Meth Med Res. 2017;25:2103–19.

[12] Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomzied program evaluation. Ann Appl Stat. 2013;7:443-70.

[13] Su S, Meneses K, McNees P, Johnson WO. Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. J R Stat Soc, Ser C. 2011;60:457–74.

[14] Su X, Pena AT, Liu L, Levine RA. Random forests of interaction trees for estimating individualized treatment effects in randomized trials. Stat Med. 2018;37:2547–60.

[15] Zhang Z, Wang C, Nie L, Soon G. Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. J R Stat Soc - Ser C. 2013;62:687–704.

[16] Shen J, He X. Inference for subgroup analysis with a structured logistic-normal mixture model. J Am Stat Assoc. 2015;110:303–12.

[17] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–60.

[18] Zhang C. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38:894–942.

[19] Chi EC, Lange K. Splitting methods for convex clustering. J Comput Graphical Stat. 2015;24:994–1013.

[20] Ma S, Huang J. A concave pairwise fusion approach to subgroup analysis. J Am Stat Assoc. 2016 (Accepted for publication).

[21] Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends in Mach Learn. 2011;3:1–122.

[22] Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educational Psychol. 1974;66:688–701.

[23] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.

[24] Wang R, Lagakos S, Ware J, Hunter D, Drazen J. Reporting of subgroup analyses in clinical trials. N Engl J Med. 2007;357:2189–94.

[25] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B. 1996;58:267–88.

[26] Härdle W, Mammen E. Comparing nonparametric versus parametric regression fits. Ann Stat. 1993;21:1926-47.

[27] Wang H, Li R, Tsai CL. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika. 2007;94:553–68.

[28] Schwarz C. Estimating the dimension of a model. Ann Stat. 1978;6:461–4.

[29] Lee ER, Noh H, Park BU. Model selection via bayesian information criterion for quantile regression models. J Am Stat Assoc. 2014;109:216–29.

[30] Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Asso. 1971;66:846–50.

[31] Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Stat Med. 2007;27:4658–77.

[32] Friedman J, Hastie T, Höfling H, Tibshirani, R. Pathwise coordinate optimization. Ann Appl Stat. 2007;1:302–32.

[33] Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. J Opt Theo Appl. 2001;109:475–94.

[34] Ma S, Racine JS, Yang L. Spline regression in the presence of categorical predictors. J Appl Econometrics. 2015;30:705–17.

[35] Ma S, Song Q, Wang L. Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data. Bernoulli. 2013;19:252–74.