| Statistica Sinica Preprint No: SS-2023-0075 | |
|---|---|
| **Title** | High-dimensional Subgroup Regression Analysis |
| **Manuscript ID** | SS-2023-0075 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202023.0075 |
| **Complete List of Authors** | Fei Jiang,<br>Tian Lu,<br>Jian Kang and<br>Lexin Li |
| **Corresponding Authors** | Fei Jiang |
| **E-mails** | homebovine@gmail.com |
| Notice: Accepted version subject to English editing. | |

# High-dimensional Subgroup Regression Analysis

Fei Jiang, Lu Tian, Jian Kang and Lexin LI

*University of California at San Francesco, Stanford University*

*University of Michigan, University of California at Berkeley*

*Abstract:* Classical regression generally assumes that all subjects follow a common model with the same set of parameters. With ever advancing capabilities of modern technologies to collect more subjects and more covariates, it has become increasingly common that there exist subgroups of subjects, and each group follows a different regression model with a different set of parameters. In this article, we propose a new approach for subgroup analysis in regression modeling. Specifically, we model the relation between a response and a set of primary predictors, while we explicitly model the heterogenous association given another set of auxiliary predictors, through the interaction between the primary and auxiliary variables. We introduce penalties to induce the sparsity and group structures within the regression coefficients, and to achieve simultaneous feature selection for both primary predictors that are significantly associated with the response, as well as the auxiliary predictors that define the subgroups. We establish the asymptotic guarantees in terms of parameter estimation consistency and cluster estimation consistency. We illustrate our method with an analysis of the functional magnetic resonance imaging data from the Adolescent Brain Cognitive Development Study. *Key words and phrases:* Adolescent Brain Cognitive Development Study; Functional magnetic resonance imaging; Group Lasso; High-dimensional

regressions; Subgroup analysis.

## 1. Introduction

Classical regression modeling generally assumes that all the subjects follow a *common* regression model with the same set of model parameters. In numerous applications, however, there may exist subgroups of subjects, and each group follows a *different* regression model with a different set of parameters. With ever advancing capabilities of modern technologies to collect more subjects and more covariates information, such data heterogeneity is becoming increasingly common. It thus becomes imperative to effectively identify subgroups of subjects and properly account for data heterogeneity in regression modeling (Ma and Huang, 2017).

Our motivation is the Adolescent Brain Cognitive Development (ABCD) Study, which plans to follow the brain development and health of over 10,000 children from their childhood through adolescence, and aims to understand biological and environmental factors that impact the brain development (Casey et al., 2018). Adolescence of a teenager is often characterized by substantial growth in cognitive skills, and those changes can be highly heterogeneous from individual to individual. Although the environmental factors contributing to the heterogeneity in cognitive development have been studied extensively (Luby et al., 2012; Mackey et al., 2017), it remains unclear whether associations between brain activation and cognitive ability vary across individuals, and if so,

how. The dataset we analyze is part of the baseline collection of the ABCD Study, which consists of the cross-sectional observations of 1,901 children from 9 to 11 years old, each with a working memory emotional n-back task functional magnetic resonance imaging (fMRI) scan, a cognitive score at the time of the scan, a psychological score that measures the mental well-being of the subject, and numerous demographic variables such as age, sex, race, and parental information. It is known that the psychological state can affect the cognitive behavior and possibly its association with brain activation (Sripada et al., 2020). Our study goal is to quantify the association pattern between the cognitive ability, which is measured by the cognitive score, and the brain region activation, which is measured by fMRI, and identify potential subgroups of subjects, which may be defined by the psychological score as well as the demographic variables.

Subgroup analysis is receiving constant attention, and has seen a surge of interest in recent years. A popular line of research that detects subgroups is to view the data as coming from a mixture of populations. For instance, Banfield and Raftery (1993); Hastie and Tibshirani (1996); Wei and Kosorok (2013) modeled the data as a mixture of Gaussian distributions to find different clusters of subjects. Another line aims to identify a subgroup of patients for an enhanced treatment effect in the setting of randomized clinical trials. Foster et al. (2011) considered the binary response case and proposed a virtual twins approach for

subgroup identification. Cai et al. (2011); Zhao et al. (2013) proposed parametric scoring methods based on the baseline covariates to rank treatment effects then identified patients who benefit most. Shen and He (2015) considered a linear logistic-normal mixture model to test for the existence of a subgroup and to score patients for treatment selection. Fan et al. (2017) developed a semiparametric change-plane approach to testing and identifying the subgroup with an enhanced treatment effect. The third line adopts a family of parametric models, typically linear association models, that include the subject-specific intercept or main effect terms, then employs various penalty functions to fuse the subject-specific terms to form subgroups. Notably, Ma and Huang (2017) considered a linear model with unobserved latent factors represented by subject-specific intercepts, and proposed a concave penalty to minimize the pairwise differences of the intercepts. Adopting similar penalization ideas for subgroup analysis, Zhu and Qu (2018) clustered the profiles of longitudinal data, Zhang et al. (2019) studied the robust median regression, Hu et al. (2021) studied the Cox proportional hazards model, and Wang et al. (2018), Tang et al. (2020), and Tang and Song (2021) addressed simultaneous subgroup identification and feature selection.

In this article, we propose a new approach to learn the heterogeneous association in a regression model. Specifically, we target the regression problem

between a response variable $Y$ and a set of *primary* predictors $\boldsymbol{X} \in \mathbb{R}^p$. Meanwhile, we explicitly model the potential heterogenous association given another set of *auxiliary* predictors $\boldsymbol{Z} \in \mathbb{R}^q$, through the interaction between $\boldsymbol{X}$ and $\boldsymbol{Z}$, then cluster the subjects into subgroups based on this interactive relation. In our motivating ABCD study, $Y$ is the cognitive score, $\boldsymbol{X}$ is the brain activation pattern from fMRI that is summarized in the form of a vector of measurements over a set of brain regions-of-interest, and $\boldsymbol{Z}$ is the vector of covariates consisting of the psychological score and demographic variables. Moreover, we introduce a number of penalty terms to induce the sparsity and group structures within the coefficients, and to achieve simultaneous feature selection for both primary predictors that are significantly associated with the response, as well as the auxiliary predictors that define the subgroups. We establish the corresponding asymptotic guarantees in terms of parameter estimation consistency and cluster estimation consistency.

Our proposal is related to but also clearly different from existing subgroup analysis solutions in multiple ways. First, our model essentially characterizes the heterogeneity through the interaction between the primary and auxiliary covariates and their joint effect on the response, and in this sense is similar in spirit to subgroup models that explicitly model the interaction between the covariates and the treatment assignment (Cai et al., 2011; Zhao et al., 2013; Tibshirani and

Friedman, 2020; Ma et al., 2019). Nevertheless, we target a high-dimensional auxiliary vector $\boldsymbol{Z}$, and the data can be observational, whereas the existing solutions mostly study a scalar treatment variable, which is independent of $\boldsymbol{Z}$ under a fully randomized trial (Cai et al., 2011; Zhao et al., 2013) or study a low dimensional $\boldsymbol{Z}$ without the theoretical justification of the estimation properties. Second, we employ special penalty functions to facilitate the model interpretation as well as feature selection, which is similar in spirit as the penalized subgroup solutions (Ma and Huang, 2017; Tang et al., 2020). However, we primarily focus on the group sparsity type penalties (Yuan and Lin, 2006), which can be efficiently implemented, enjoy desirable theoretical properties (Huang and Zhang, 2010), and are utterly different from the fusion type penalties currently used in subgroup modeling (Ma and Huang, 2017; Zhang et al., 2019). Finally, we adopt a linear regression model, but it can be straightforwardly extended to more flexible model forms through basis expansion approaches such as splines (De Boor, 1978) and reproducing kernels (Wahba, 1990). Besides, the choice of primary and auxiliary variables depends on the scientific interest, while it is flexible in terms of putting which variables into which set. We illustrate with the ABCD study, but the method is equally applicable to a wide range of modern scientific problems. In summary, our proposal addresses a critically important problem, and offers a new angle to subgroup analysis.

The rest of the article is organized as follows. Section 2 presents the model and the penalized optimization formulation. Section 3 develops the estimation algorithm, and Section 4 establishes the theoretical guarantees. Section 5 studies the finite-sample performance through simulations, and Section 6 revisits our motivating ABCD data analysis. Section 7 concludes the paper with a discussion. The Supplementary Appendix collects all technical proofs.

## 2. Model and Penalized Formulation

### 2.1 Model

Consider the response $Y \in \mathbb{R}$, the primary predictor vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal \in \mathbb{R}^p$, and the auxiliary predictor vector $\boldsymbol{Z} = (Z_1, \ldots, Z_q)^\intercal \in \mathbb{R}^q$. Let $X_1 = 1$ and $Z_1 = 1$, which incorporate the intercept. Consider the observational data of $n$ i.i.d. copies of $\{\boldsymbol{X}, Y, \boldsymbol{Z}\}$ :, $\{\boldsymbol{X}_i, Y_i, \boldsymbol{Z}_i, i = 1, \ldots, n\}$. We posit a regression model,

$$Y_i = \boldsymbol{\alpha}(\boldsymbol{Z}_i)^\intercal \boldsymbol{X}_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\alpha}(\boldsymbol{Z}_i) = \{\alpha_1(\boldsymbol{Z}_i), \ldots, \alpha_p(\boldsymbol{Z}_i)\}^\intercal \in \mathbb{R}^p$ characterizes the effect of $\boldsymbol{X}_i$ on $Y_i$, and $\epsilon_i$ is a zero mean independent random error term. The coefficient of $\boldsymbol{X}_i$, i.e., $\boldsymbol{\alpha}(\boldsymbol{Z}_i)$, varies across individuals, and thus reflects the potentially heterogeneous effects of $\boldsymbol{X}_i$ on the outcome. We further assume that the data samples can be uniquely partitioned to $G$ groups, up to relabeling, with distinct

cluster centers $\boldsymbol{a}_{01}, \ldots, \boldsymbol{a}_{0K}$, such that

$$\mathbb{E}\left(\min_{g\in\{1,\ldots,G\}} \|\boldsymbol{\alpha}(\boldsymbol{Z}_i) - \boldsymbol{a}_{0g}\|_2^2\right)$$

is the smallest, where $\|\cdot\|_2$ is the vector $L_2$-norm, and $\mathbb{E}$ is with respect to $\boldsymbol{Z}_i$. Correspondingly, we divide our subgroup analysis into two main steps: we first estimate $\boldsymbol{\alpha}(\cdot)$, then carry out a clustering analysis based on the estimated $\boldsymbol{\alpha}(\cdot)$.

Next, we introduce a parametric form for $\boldsymbol{\alpha}(\boldsymbol{Z}_i)$, by setting $\alpha_j(\boldsymbol{Z}_i) = \sum_{k=1}^{q} \theta_{0jk} Z_{ik}$, $j = 1, \ldots, p$, and we set $m(\cdot)$ as the identity function. This leads to our final model,

$$\begin{aligned} Y_i &= \sum_{j=1}^{p}\left(\sum_{k=1}^{q} \theta_{0jk} Z_{ik}\right) X_{ij} + \epsilon_i = \sum_{j=1}^{p} \theta_{0j1} X_{ij} + \sum_{j=1}^{p}\sum_{k=2}^{q} \theta_{0jk} Z_{ik} X_{ij} + \epsilon_i \\ &= \boldsymbol{X}_i^{\top} \boldsymbol{\Theta}_0 \boldsymbol{Z}_i + \epsilon_i, \end{aligned} \tag{21}$$

where $\boldsymbol{\Theta}_0 = (\theta_{0jk})_{j=1,k=1}^{p,q} \in \mathbb{R}^{p\times q}$ collects all parameters of interest. We make some remarks about model (21). First of all, for the first column of $\boldsymbol{\Theta}_0$, i.e., $k = 1$, the parameters $\theta_{0j1}$ characterize the common baseline effect of $X_j$'s on $Y$, for $j = 1, \ldots, p$. For the rest of columns of $\boldsymbol{\Theta}_0$, i.e., $k = 2, \ldots, q$, the parameters $\theta_{0jk}$ capture the deviation from this baseline, or in other words, the heterogeneous effects introduced by $Z_k$'s. Second, when an entire row of $\boldsymbol{\Theta}_0$ equals zero, e.g., $\theta_{0j\cdot} = 0$, for some $j = 1, \ldots, p$, it implies that the $j$th primary predictor $X_j$ does not affect the response $Y$. Meanwhile, when an entire col-

umn of $\boldsymbol{\Theta}_0$ equals zero, e.g., $\theta_{0\cdot k} = 0$ for some $k = 2, \ldots, q$, it implies that the $k$th auxiliary predictor $Z_k$ does not introduce any additional heterogeneity to the association between $\boldsymbol{X}$ and $Y$. When all columns of $\boldsymbol{\Theta}_0$ except for the first column equal zero, model (21) reduces to the usual homogeneous linear regression. Finally, in a relatively straightforward fashion, model (21) can be generalized to more flexible settings, e.g., through spline basis expansions (De Boor, 1978) or reproducing kernel approaches (Wahba, 1990). In this article, we focus on the model form in (21), as it provides a foundation for those more flexible extensions.

### 2.2 Penalized optimization

Consider the least squares loss function and the penalty function,

$$\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\Theta}\boldsymbol{Z}_i\right)^2, \qquad \mathcal{R}(\boldsymbol{\Theta}) = \lambda_c\sum_{k=2}^{q}\sqrt{p}\|\boldsymbol{\Theta}_{\cdot k}\|_2 + \lambda_r\sum_{j=1}^{p}\sqrt{q}\|\boldsymbol{\Theta}_{j\cdot}\|_2.$$

We propose to estimate the parameter of interest $\boldsymbol{\Theta}_0$ via the following penalized optimization,

$$\mathcal{L}(\boldsymbol{\Theta}) + \mathcal{R}(\boldsymbol{\Theta}), \quad \text{subject to } \|\mathrm{vec}(\boldsymbol{\Theta})\|_2 \le \psi_1, \text{ and } \|\mathrm{vec}(\boldsymbol{\Theta})\|_1 \le \psi_2, \quad (22)$$

where $\boldsymbol{\Theta}_{j\cdot}$ denotes the $j$th row of $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}_{\cdot k}$ denotes the $k$th column of $\boldsymbol{\Theta}$, $\lambda_r, \lambda_c$ are the row and column penalty parameters, respectively, $\psi_1$ and $\psi_2$ are penalty constants and $\|\cdot\|_0$ and $\|\cdot\|_1$ are the vector $L_0$ and $L_1$-norm, respectively. We

again make some remarks about the penalized optimization formulation (22). First, the function $\mathcal{R}(\boldsymbol{\Theta})$ places the group Lasso penalties (Yuan and Lin, 2006) on both columns and rows of $\boldsymbol{\Theta}$. As such, it achieves *simultaneous* feature selection of both primary and auxiliary predictors. Second, the penalties placed on the rows and columns are different, reflecting the asymmetric roles of $\boldsymbol{X}$ and $\boldsymbol{Z}$ in the model. More specifically, we do not penalize the first column $\boldsymbol{\Theta}_{\cdot 1}$, since it corresponds to the baseline association between $\boldsymbol{X}$ and $Y$. Moreover, because each row and column intersect with each other, placing the group $L_1$ penalties on both row and column forms a penalty on their intersection, which in effect shrinks the small entries in $\boldsymbol{\Theta}$ to zero. Finally, the additional constraints $\|\mathrm{vec}(\boldsymbol{\Theta})\|_2 \leq \psi_1$ and $\|\mathrm{vec}(\boldsymbol{\Theta})\|_1 \leq \psi_2$ in (22) are only used to facilitate the theoretical derivation and have little empirical effect. Since $\boldsymbol{X}$ and $\boldsymbol{Z}$ are high-dimensional, i.e., $p, q > n$, when we treat each row of $\boldsymbol{\Theta}$ as a group, the number of elements in the group and the number of groups both grow with the sample size. Then these constraints limit the number of nonzero elements in $\boldsymbol{\Theta}$, which in turn ensures that the loss function is convex. By doing so, we avoid having to assume that the initial value is in a close neighborhood of the true parameter, or the step size of the parameter update is small enough so that the estimator in each iteration remains in a neighborhood of the truth. Similar constraints have been used in Yin et al. (2014), Jiang and Ma (2021), and Loh and Wainwright

(2012) too. In practice, we choose $\psi_1$, $\psi_2$ to take very large values, so that they would not affect the penalty function $\mathcal{R}(\boldsymbol{\Theta})$.

## 3. Estimation

### 3.1 Subgradient analysis

To solve the optimization problem (22), we first majorize the loss function $\mathcal{L}(\boldsymbol{\Theta})$ as,

$$\widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta}) = \mathcal{L}\big(\boldsymbol{\Theta}^{(t)}\big) + \left(\theta_{jk} - \theta_{jk}^{(t)}\right) \frac{\partial \mathcal{L}(\boldsymbol{\Theta}^{(t)})}{\partial \theta_{jk}} + \frac{1}{2\eta}\left(\theta_{jk} - \theta_{jk}^{(t)}\right)^2, \qquad (33)$$

for a given $(j,k)$, $j = 1,\ldots,p$, $k = 1,\ldots,q$, where $\boldsymbol{\Theta}^{(t)} = \big(\theta_{jk}^{(t)}\big)_{j=1,k=1}^{p,q}$ denotes the estimate of $\boldsymbol{\Theta} = (\theta_{jk})_{j=1,k=1}^{p,q}$ at the $t$th iteration, and $\eta$ is the step size. For a convex optimization problem, the minimizer of the surrogate $\widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta}) + \mathcal{R}(\boldsymbol{\Theta})$ is closer to the global minimizer of $\mathcal{L}(\boldsymbol{\Theta}) + \mathcal{R}(\boldsymbol{\Theta})$ than $\boldsymbol{\Theta}^{(t)}$ (Loh and Wainwright, 2012). Furthermore, finding the minimizer for $\widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta}) + \mathcal{R}(\boldsymbol{\Theta})$ instead of $\mathcal{L}(\boldsymbol{\Theta}) + \mathcal{R}(\boldsymbol{\Theta})$ avoids inverting a potentially large Hession matrix, which in turn reduces the computational burden. We next study the subgradient conditions that determine the rows, columns, and entries of $\boldsymbol{\Theta}$, respectively, that would be penalized to zero during the estimation. We present the detailed algorithm in the next section.

Observing that the solution to a convex optimization problem is a saddle

point at which the subgradient equals zero, we consider four different scenarios of the subgradient.

Case I: Consider the saddle point with $\boldsymbol{\Theta}_{j\cdot} = 0$, $j = 1, \ldots, p$. It satisfies that,

$$-\frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}_{j\cdot}}\Big|_{\boldsymbol{\Theta}_{j\cdot}=\mathbf{0}} = \lambda_c\sqrt{p}\boldsymbol{\mu}' + \lambda_r\sqrt{q}\boldsymbol{\mu},$$

where $\boldsymbol{\mu} \in \mathbb{R}^q$ is the subgradient of $\|\boldsymbol{\Theta}_{j\cdot}\|_2$ with respect to $\boldsymbol{\Theta}_{j\cdot}$ evaluated at the saddle point, i.e.,

$$\boldsymbol{\mu} = \begin{cases} \frac{\boldsymbol{\Theta}_{\cdot k}}{\|\boldsymbol{\Theta}_{\cdot k}\|_2}, & \text{if } \|\boldsymbol{\Theta}_{j\cdot}\|_2 \neq 0 \\ \forall \boldsymbol{\mu}, \|\boldsymbol{\mu}\|_2 \leq 1, & \text{if } \|\boldsymbol{\Theta}_{j\cdot}\|_2 = 0, \end{cases}$$

and $\boldsymbol{\mu}' = (\mu_1', \cdots, \mu_q')^\intercal \in \mathbb{R}^q$ is the subgradient of $\sum_{k=2}^q \|\boldsymbol{\Theta}_{\cdot k}\|_2$ with respect to $\boldsymbol{\Theta}_{j\cdot}$ evaluated at the saddle point, i.e.,

$$\mu_k' = \begin{cases} 0, & \text{if } k = 1, \\ \frac{\theta_{jk}}{\|\boldsymbol{\Theta}_{\cdot k}\|_2}, & \text{if } \|\boldsymbol{\Theta}_{\cdot k}\|_2 \neq 0, k > 1, \\ \forall \mu, |\mu| \leq 1, & \text{if } \|\boldsymbol{\Theta}_{\cdot k}\|_2 = 0, k > 1. \end{cases}$$

When $\boldsymbol{\Theta}_{j\cdot} = \mathbf{0}$ and $\|\boldsymbol{\Theta}_{\cdot k}\|_2^2 \neq 0, k \geq 1$ , we have obtained the first order condition that

$$-\frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\boldsymbol{\Theta}_{j\cdot}=\mathbf{0}} = \lambda_r\sqrt{q}\mu_k.$$

On the other hand, following Simon et al. (2013), when $\boldsymbol{\Theta}_{\cdot k} = \mathbf{0}, k > 1$ and

$\boldsymbol{\Theta}_{j\cdot} = \mathbf{0}$, we have the first order condition that,

$$\left| \mathbb{S}\left( \frac{-\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\boldsymbol{\Theta}_{j\cdot}=\mathbf{0}},\ \lambda_c\sqrt{p} \right) \right| \leq \lambda_r \sqrt{q}|\mu_k|,$$

where $\mathbb{S}(a,\lambda) = \text{sign}(a)(|a|-\lambda)_+$ is the soft-thresholding function. Define $\boldsymbol{U}_j = (U_{j1}, \cdots, U_{jq})^\intercal \in \mathbb{R}^q$, $j = 1, \ldots, p$,

$$U_{jk} = \begin{cases} -\frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\boldsymbol{\Theta}_{j\cdot}=\mathbf{0}}, & \text{if } \boldsymbol{\Theta}_{\cdot k} = \mathbf{0}, \\ \mathbb{S}\left( \frac{-\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\boldsymbol{\Theta}_{j\cdot}=\mathbf{0}},\ \lambda_c\sqrt{p} \right), & \text{if } \boldsymbol{\Theta}_{\cdot k} = \mathbf{0} \text{ for } k > 1. \end{cases} \tag{34}$$

In summary, if $\|\boldsymbol{U}_j\|_2 \leq \lambda_r\sqrt{q}$, then the subgradient condition for $\boldsymbol{\Theta}_{j\cdot} = \mathbf{0}$ is satisfied.

Case II: Consider the saddle point with $\boldsymbol{\Theta}_{\cdot k} = \mathbf{0}$, $k = 2, \ldots, q$. It satisfies that,

$$-\frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}_{\cdot k}}\Big|_{\boldsymbol{\Theta}_{\cdot k}=\mathbf{0}} = \lambda_c\sqrt{p}\boldsymbol{\nu} + \lambda_r\sqrt{q}\boldsymbol{\nu}',$$

where $\boldsymbol{\nu} \in \mathbb{R}^p$ is the subgradient of $\|\boldsymbol{\Theta}_{\cdot k}\|_2$ with respect to $\boldsymbol{\Theta}_{\cdot k}$ evaluated at the saddle point, and $\boldsymbol{\nu}' = (\nu_1', \cdots, \nu_p')^\intercal \in \mathbb{R}^p$ is the subgradient of $\sum_{j=1}^p \|\boldsymbol{\Theta}_{j\cdot}\|_2$ with respect to $\boldsymbol{\Theta}_{\cdot k}$, i.e.,

$$\nu_j' = \begin{cases} \frac{\theta_{jk}}{\|\boldsymbol{\Theta}_{j\cdot}\|_2}, & \text{if } \|\boldsymbol{\Theta}_{j\cdot}\|_2 \neq 0, \\ \forall \nu, |\nu| \leq 1, & \text{if } \|\boldsymbol{\Theta}_{j\cdot}\|_2 = 0. \end{cases}$$

Following a similar argument as before, and define $\boldsymbol{V}_k = (V_{1k}, \cdots, V_{pk})^\intercal \in \mathbb{R}^p$,

$k = 2, \ldots, q$,

$$V_{jk} = \begin{cases} \frac{-\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\boldsymbol{\Theta}_{\cdot k}=\mathbf{0}}, & \text{if } \boldsymbol{\Theta}_{j\cdot} \neq \mathbf{0}, \\ \mathbb{S}\left(-\frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\boldsymbol{\Theta}_{\cdot k}=0},\ \lambda_r\sqrt{q}\right), & \text{if } \boldsymbol{\Theta}_{j\cdot} = \mathbf{0}. \end{cases} \tag{35}$$

If $\|\boldsymbol{V}_k\|_2 \leq \lambda_c\sqrt{p}$, then the first order condition for $\boldsymbol{\Theta}_{\cdot k} = \mathbf{0}$ is satisfied.

Case III: Consider the saddle point with both $\boldsymbol{\Theta}_{j\cdot} = \mathbf{0}$ and $\boldsymbol{\Theta}_{\cdot k} = \mathbf{0}$, $j = 1, \ldots, p, k = 1, \ldots, q$. It satisfies that,

$$-\frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\theta_{jk}=0} = \lambda_c\sqrt{p}\mu' + \lambda_r\sqrt{q}\nu'. \tag{36}$$

Therefore, if

$$\left|\frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\theta_{jk}=0}\right| \leq \lambda_r\sqrt{q}, \quad \text{or} \quad \left|\frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\theta_{jk}=0}\right| \leq \lambda_c\sqrt{p} + \lambda_r\sqrt{q},$$

for $k = 2, \ldots, q$, then the subgradient condition for $\theta_{jk} = 0$ is satisfied.

Case IV: Consider the saddle point with $\boldsymbol{\Theta}_{\cdot k} \neq \mathbf{0}$ or $\boldsymbol{\Theta}_{j\cdot} \neq \mathbf{0}$. The subgradient with respect to $\theta_{jk}$ evaluate at $\boldsymbol{\Theta}$ is,

$$-\frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}} = \lambda_c\sqrt{p}\frac{\theta_{jk}}{\|\boldsymbol{\Theta}_{\cdot k}\|_2} + \lambda_r\sqrt{q}\frac{\theta_{jk}}{\|\boldsymbol{\Theta}_{j\cdot}\|_2}. \tag{37}$$

Case I and Case II suggest that, if the $j$th row, or the $k$th column, of $\boldsymbol{\Theta}$ is zero, then correspondingly, $\|\boldsymbol{U}_j\|_2 \leq \lambda_r\sqrt{q}$, or $\|\boldsymbol{V}_k\|_2 \leq \lambda_c\sqrt{p}$. Henceforth, in an iterative update, we shrink the $j$th row or the $k$th column to $\mathbf{0}$, if the estimate

of $\boldsymbol{U}_j$ or $\boldsymbol{V}_k$ from the previous iteration satisfies those two inequalities. Case III suggests a strategy to shrink the $(j,k)$th entry of $\boldsymbol{\Theta}$ when the derivative is sufficiently small. Case IV guides the estimation of the nonzero entries of $\boldsymbol{\Theta}$. We present the complete iterative computational algorithm in the next section.

### 3.2 Algorithm

We develop an iterative optimization algorithm for parameter estimation with five main steps. The first four steps identify zero rows, zero columns, and update the nonzero entries of $\boldsymbol{\Theta}$, respectively, by utilizing the subgradient properties at the saddle points studied in Section 3.1. The last step employs a projection operation to make the estimate satisfy the constraint that $\|\mathrm{vec}(\boldsymbol{\Theta})\|_2 \leq \psi_1$ and $\|\mathrm{vec}(\boldsymbol{\Theta})\|_1 \leq \psi_2$.

When updating $\theta_{jk}$ at the $t$th iteration, we replace the parameters other than $\theta_{jk}$ by their most recent update at the $(t-1)$th iteration in (34), (35), (36), and (37). Since $\theta_{jk}$ is the only unknown parameter, we have that,

$$\frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\boldsymbol{\Theta}_{j\cdot}=0} = \frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\boldsymbol{\Theta}_{\cdot k}=0} = \frac{\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})}{\partial \theta_{jk}}\Big|_{\theta_{jk}=0} = \frac{\partial \mathcal{L}(\boldsymbol{\Theta}^{(t-1)})}{\partial \theta_{jk}} - \eta^{-1}\theta_{jk}^{(t-1)}. \quad (38)$$

We next describe the estimation algorithm step-by-step.

Step 1: Following the subgradient analysis in Case I, for row $j = 1, \ldots, p$, we set $\boldsymbol{\Theta}_{j\cdot}^{(t)} = \mathbf{0}$, if $\|\boldsymbol{U}_j\|_2 \leq \lambda_r\sqrt{q}$, where $\boldsymbol{U}_j$ is defined in (34) and $\partial \widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})/\partial\theta_{jk}$ is obtained via (38).

Step 2: Following the subgradient analysis in Case II, for column $k = 2, \ldots, q$, we set $\boldsymbol{\Theta}_{\cdot k}^{(t)} = \mathbf{0}$, if $\|\boldsymbol{V}_k\|_2 \leq \lambda_c\sqrt{p}$, where $\boldsymbol{V}_k$ is defined in (35) and $\partial\widetilde{\mathcal{L}}_{jk}(\boldsymbol{\Theta})/\partial\theta_{jk}$ is obtained via (38).

Step 3: Following the subgradient analysis in Case III, when $\sum_{l>1,l\neq k}(\theta_{jl}^{(t-1)})^2 = 0$ and $\sum_{u\neq j}(\theta_{uk}^{(t-1)})^2 = 0$, we update $\theta_{jk}^{(t)}$ as,

$$\theta_{jk}^{(t)} = \mathbb{S}\left(\eta^{-1}\theta_{jk}^{(t-1)} - \frac{\partial\mathcal{L}(\boldsymbol{\Theta}^{(t-1)})}{\partial\theta_{jk}},\ \lambda_c\sqrt{p} + \lambda_r\sqrt{q}\right),\ \text{ for }\ k \geq 2.$$

Also, since we place no penalty on the first column of $\boldsymbol{\Theta}$, we let

$$\theta_{j1}^{(t)} = \mathbb{S}\left(\eta^{-1}\theta_{j1}^{(t-1)} - \frac{\partial\mathcal{L}(\boldsymbol{\Theta}^{(t-1)})}{\partial\theta_{j1}},\ \lambda_r\sqrt{q}\right),$$

i.e., we set $\lambda_c = 0$ for $k = 1$.

Step 4: Following the subgradient analysis in Case IV, when $\sum_{l>1,l\neq k}(\theta_{jl}^{(t-1)})^2$ and $\sum_{u\neq j}(\theta_{uk}^{(t-1)})^2$ are not both zero, we update $\theta_{jk}^{(t)}$ as the root of the equation,

$$\eta^{-1}\theta_{jk} - \eta^{-1}\theta_{jk}^{(t-1)} + \frac{\partial\mathcal{L}(\boldsymbol{\Theta}^{(t-1)})}{\partial\theta_{jk}} + \lambda_r\sqrt{q}\frac{\theta_{jk}}{\left\{\sum_{u\neq j}(\theta_{uk}^{(t-1)})^2 + \theta_{jk}^2\right\}^{1/2}} + \lambda_c\sqrt{p}\frac{\theta_{jk}}{\left\{\sum_{l>1,l\neq k}(\theta_{jl}^{(t-1)})^2 + \theta_{jk}^2\right\}^{1/2}} = 0.$$

Again, we set $\lambda_c = 0$ when $k = 1$. In our implementation, we employ the Brent's method (Brent, 1973) to find the root, while other root finding methods can be applied as well.

Step 5: We project $\mathrm{vec}(\boldsymbol{\Theta}^{(t)})$ to $\mathbb{B}_1(\psi_2) = \left\{\boldsymbol{\Theta} \mid \|\mathrm{vec}(\boldsymbol{\Theta})\|_1 \leq \psi_2\right\}$, then $\mathbb{B}_2(\psi_1) = \left\{\boldsymbol{\Theta} \mid \|\mathrm{vec}(\boldsymbol{\Theta})\|_2 \leq \psi_1\right\}$, so that the final projection satisfies that

$\text{vec}(\boldsymbol{\Theta}^{(t)}) \in \mathbb{B}_1(\psi_2) \cap \mathbb{B}_2(\psi_1)$. We obtain the projection to $\mathbb{B}_1(\psi_2)$, a simplex in $\mathbb{R}^{pq}$, by adopting the linear programming method of Duchi et al. (2008).

We iterate the above steps until the algorithm converges, where we set the stopping criterion as $\|\text{vec}(\boldsymbol{\Theta}^{(t)}) - \text{vec}(\boldsymbol{\Theta}^{(t-1)})\|_2 \leq 10^{-4}$. In practice, we find our algorithm converges relatively fast.

For tuning parameters, we set $\lambda_c$ and $\lambda_r$ at their theoretical orders specified in the next section. Moreover, in real data analysis, we adopt the strategy of Chatterjee and Lahiri (2011) to further bootstrap the procedure to construct a confidence interval for each parameter after the sparse estimation, making the results less sensitive to the choice of $\lambda_c$ and $\lambda_r$. We choose $b_0$ and $b_1$ to be some constants much larger than $\|\text{vec}(\boldsymbol{\Theta})\|_2$ and $\|\text{vec}(\boldsymbol{\Theta})\|_1$ at their initial values. In addition, we find the method is not overly sensitive to the choice of $\eta$ in the majorization, as long as it is in a reasonable range.

Finally, after obtaining the final estimate $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(t)}$ through the iterative optimization, we employ a standard clustering algorithm, e.g., the $K$-means algorithm, to cluster the subjects into $G$ subgroups based on $\widehat{\boldsymbol{\Theta}}\boldsymbol{Z}_i$, for $i = 1, \ldots, n$. We choose the number of clusters using the Bayesian information criteria (BIC), while other criteria can be used as well (Wang, 2010).

## 4. Asymptotic Theory

### 4.1 Parameter estimation consistency

We first introduce a set of conditions needed for the parameter estimation consistency. Denote $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\intercal \in \mathbb{R}^n$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\intercal \in \mathbb{R}^n$, $\boldsymbol{W}_i = \boldsymbol{Z}_i \otimes \boldsymbol{X}_i \in \mathbb{R}^{pq}$, with $\otimes$ being the Kronecker product, and $\boldsymbol{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n)^\intercal \in \mathbb{R}^{n\times(pq)}$. Let $|\mathcal{A}|$ be the cardinality of a set $\mathcal{A}$, $\boldsymbol{a}_\mathcal{S}$ be the sub-vector of $\boldsymbol{a}$ with elements in the index set $\mathcal{S}$, and $\boldsymbol{M}_\mathcal{S}$ be the submatrix of $\boldsymbol{M}$ with columns in $\mathcal{S}$. Let $S_r \subset \{1, \ldots, p\}$ and $S_c \subset \{1, \ldots, q\}$ denote the set of nonzero rows and columns of $\boldsymbol{\Theta}_0$, respectively, and let $g_r = |S_r|$ and $g_c = |S_c|$ be the number of nonzero rows and columns, respectively. Let $d_0 = \|\boldsymbol{\Theta}_0\|_0$. Define

$$\begin{aligned}\rho_+(\omega) &= \max\{n^{-1}\|\boldsymbol{W}_\mathcal{S}\boldsymbol{u}_\mathcal{S}\|_2^2/\|\boldsymbol{u}_\mathcal{S}\|_2^2 : \boldsymbol{u} \in \mathbb{R}^{pq}, \mathcal{S} \subset \{1, \ldots, pq\}, |\mathcal{S}| \le \omega\},\\ \rho_-(\omega) &= \min\{n^{-1}\|\boldsymbol{W}_\mathcal{S}\boldsymbol{u}_\mathcal{S}\|_2^2/\|\boldsymbol{u}_\mathcal{S}\|_2^2 : \boldsymbol{u} \in \mathbb{R}^{pq}, \mathcal{S} \subset \{1, \ldots, pq\}, |\mathcal{S}| \le \omega\}.\end{aligned}$$

Intuitively, $\rho_+(\omega)$ and $\rho_-(\omega)$ are the maximum and minimal eigenvalues of $\boldsymbol{W}_\mathcal{S}^\intercal\boldsymbol{W}_\mathcal{S}$ for any set $\mathcal{S}$ satisfying that $|\mathcal{S}| \le \omega$. We impose the following regularity conditions.

(A1) Suppose $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are two sub-Gaussian random vectors, and $\epsilon_i$ is a mean zero sub-Gaussian random error. Furthermore, $\boldsymbol{X}_i, \boldsymbol{Z}_i$ and $\epsilon_i$ are independent, $i = 1, \ldots, n$.

(A2) Suppose $\|\mathrm{vec}(\mathbf{\Theta}_0)\|_2 \leq d_2 < \infty$. Denote the uniformity of the optimizer of (22) as $U_{\mathbf{\Theta}}$, where the uniformity of a matrix is defined as the ratio between the smallest nonzero entry and the largest one. Furthermore, define $b_1 \equiv \left\{\psi_2(U_{\mathbf{\Theta}}^{1/2} + U_{\mathbf{\Theta}}^{-1/2})/(2\psi_1)\right\}^2$, we assume $d_0 \leq b_1$.

(A3) Suppose there exists a constant $s \geq d_0 + 2b_1$, such that $\rho_-(s) \geq \epsilon_0 > 0$, and

$$\frac{\sqrt{\{\rho_+(s) - \rho_-(2s - d_0)\} \times \{\rho_+(s - d_0) - \rho_-(2s - d_0)\}}}{\rho_-(s)} \leq \frac{1}{3}\frac{\min(\lambda_r\sqrt{q}, \lambda_c\sqrt{p})}{\lambda_r\sqrt{q}g_r + \lambda_c\sqrt{p}g_c},$$

almost surely.

Condition (A1) is a fairly standard condition to ensure the distributions of the first and second order derivatives of the loss function vanish sufficiently fast at the tail. Condition (A2) ensures the convergence of the estimation procedure, and induces the sparsity on the estimator by shrinking in both the row and column directions, where $b_1$ is the upper bound of the sparseness of the optimizer by Lemma S1 in the appendix. It is introduced to deal with the challenge that, in our setting, the group size is either $p$ or $q$, which can be larger than the sample size $n$ and also diverges with $n$. In addition, we can always achieve $d_0 \leq b_1$ by increasing the value $\psi_2/\psi_1$. By contrast, the classical group $L_1$-based methods require the group size to be fixed to achieve the estimation consistency (Huang and Zhang, 2010; Tibshirani and Friedman, 2020). The same condition has been

used in Yin et al. (2014); Jiang and Ma (2021); Loh and Wainwright (2012). Condition (A3) is an eigenvalue condition. It restrict the $L_2$ norm of the off-diagonal submatrix of $n^{-1}\boldsymbol{W}^{\intercal}\boldsymbol{W}$, and ensures the second order derivatives of the loss function to be positive definite. Coupled with the lower bound on $\rho_{-}(s)$, the upper bound for $\rho_{+}(s)$ implies that the correlations among the $p+q$ covariates are bounded from above. Similar conditions have also been imposed by Huang and Zhang (2010).

**Theorem 1.** *Suppose Conditions (A1), (A2) and (A3) hold. Furthermore, there exit two positive constants* $c_1$ *and* $c_2$*, with* $c_1 < c_2$*, such that* $c_1 \leq \frac{\lambda_r}{\sqrt{b_1 \log(p)/(qn)}}$*, and* $\frac{\lambda_c}{\sqrt{b_1 \log(q)/(pn)}} \leq c_2$*. Then, with probability at least* $1-12\exp\{-b_1\log(p)\}-12\exp\{-b_1\log(q)\}$*, we have that,*

$$\|\mathrm{vec}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)\|_2 \leq C(g_r^2+g_c^2+1)^{1/2}(g_r+g_c)\sqrt{\frac{b_1\log(pq)}{n}},$$

*for some constant* $C>0$.

Theorem 1 utilizes the convexity of the objective function and the concentration inequality of sub-Gaussian variables, and establishes the statistical consistency of the proposed estimator when $p$ and $q$ are both in the order of $o\{\exp(n)\}$. Furthermore, when $b_1 = O(1)$, it implies that $\|\mathrm{vec}(\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_0)\|_2 = O_p\{\sqrt{\log(pq)/n}\}$, which is consistent with the conventional order of convergence in the group $L_1$ literature (Liu and Zhang, 2009; Huang and Zhang, 2010). Moreover, we can

achieve $b_1 = O(1)$ by selecting $R_2/R_1$ to be of order $O\{1/(U_{\boldsymbol{\Theta}}^{1/2} + U_{\boldsymbol{\Theta}}^{-1/2})\}$.

### 4.2 Clustering consistency

After obtaining the estimator $\widehat{\boldsymbol{\Theta}}$, we perform the clustering analysis based on $\widehat{\boldsymbol{\Theta}}\boldsymbol{Z}_i$ to identify the subgroups. We next introduce another set of regularity conditions needed for the clustering consistency. For a given parameter $\boldsymbol{\Theta}$, let $P_{\boldsymbol{\Theta}}$ denote the probability measure induced by $\boldsymbol{\Theta}\boldsymbol{Z}_i$, and $f_{\boldsymbol{\Theta}}(\cdot)$ the associated density function. We impose the following conditions.

(B1) Suppose the number of clusters $G$ is fixed, and suppose the true cluster centers,

$$\boldsymbol{a}_0 = (\boldsymbol{a}_{01}^\intercal, \ldots, \boldsymbol{a}_{0G}^\intercal)^\intercal = \underset{\boldsymbol{a}=(\boldsymbol{a}_1^\intercal, \cdots, \boldsymbol{a}_G^\intercal)^\intercal}{\arg\min} \mathbb{E}\left\{ \min_{1 \le g \le G} \|\boldsymbol{\Theta}_0 \boldsymbol{Z}_i - \boldsymbol{a}_g\|_2^2 \right\}$$

are unique up to relabeling, and $\mathbb{E}(\|\boldsymbol{\Theta}_0 \boldsymbol{Z}_i\|_2^2) < \infty$.

(B2) Suppose there is a function $g(\cdot)$, such that $f_{\boldsymbol{\Theta}_0}(\boldsymbol{v}) \le g(\|\boldsymbol{v}\|_2)$, and $\int_0^\infty v^l g(v) dv < \infty$, where $l = \|\boldsymbol{\Theta}_0 \boldsymbol{Z}\|_0$.

(B3) Suppose $|\int \|\boldsymbol{v}\|_2^2 \partial f_{\boldsymbol{\Theta}}(\boldsymbol{v}) / \partial \mathrm{vec}(\boldsymbol{\Theta})^\intercal \boldsymbol{e}_j d\boldsymbol{v}| < \infty$, $|\int \partial f_{\boldsymbol{\Theta}}(\boldsymbol{v}) / \partial \mathrm{vec}(\boldsymbol{\Theta})^\intercal \boldsymbol{e}_j d\boldsymbol{v}| < \infty$, and $\|\int \boldsymbol{v} \partial f_{\boldsymbol{\Theta}}(\boldsymbol{v}) / \partial \mathrm{vec}(\boldsymbol{\Theta})^\intercal \boldsymbol{e}_j d\boldsymbol{v}\|_2 < \infty$ at any $\boldsymbol{\Theta}$, with $\|\mathrm{vec}(\boldsymbol{\Theta})\|_2 < \infty$ and $\|\mathrm{vec}(\boldsymbol{\Theta})\|_0 \le 2b_1$, where $\boldsymbol{e}_j$ is the unit vector with the $j$th element being one and the rest zero, $j = 1, \ldots, pq$.

Conditions (B1) to (B3) are standard and mild conditions to establish the consistency of the $K$-means clustering in Pollard et al. (1982, Conditions (i), (ii), (iv)). In particular, Condition (B1) does not allow the number of clusters to diverge, and also ensures the uniqueness of the cluster membership. Condition (B2) ensures the contributions to the mean squared error made by the samples that are outside of a certain radius of each center to vanish sufficiently fast. Condition (B3) ensures that $\int \min_{1\le j\le K} \|\boldsymbol{v}-\boldsymbol{a}_j\|_2^2 dP_{\widehat{\boldsymbol{\Theta}}}(\boldsymbol{v})$ and $\int \min_{1\le j\le K} \|\boldsymbol{v}-\boldsymbol{a}_j\|_2^2 dP_{\boldsymbol{\Theta}_0}(\boldsymbol{v})$ are close. It is fairly mild, because it only requires that each element of $\partial f_{\boldsymbol{\Theta}}(\boldsymbol{v})/\partial \mathrm{vec}(\boldsymbol{\Theta})$, $\|\boldsymbol{v}\|_2^2 \partial f_{\boldsymbol{\Theta}}(\boldsymbol{v})/\partial \mathrm{vec}(\boldsymbol{\Theta})$, and $\boldsymbol{v}\partial f_{\boldsymbol{\Theta}}(\boldsymbol{v})/\partial \mathrm{vec}(\boldsymbol{\Theta})$ have finite integrals, rather than requiring the uniformly integrability of the entire vectors. This condition is easily satisfied, for instance, when $\boldsymbol{Z}_i$ is Gaussian or has a compact support.

**Theorem 2.** *Suppose the conditions in Theorem 1 along with (B1), (B2) and (B3) hold. Then,*

$$\|\widehat{\boldsymbol{a}}-\boldsymbol{a}_0\|_2 = O_p\left[\{b_1^2\log(pq)/n\}^{1/4} + n^{-1/2}(Gb_1)^{1/2}\right].$$

Theorem 2 establishes the consistency of the estimated cluster centers. There are two terms in the convergence rate. The first term comes from the estimation error of $\widehat{\boldsymbol{\Theta}}$, whereas the second term comes from the empirical distribution approximation to the distribution of $\boldsymbol{\Theta}\boldsymbol{Z}$. If $\boldsymbol{\Theta}_0$ were known, then the order of clustering

consistency is the same as that of Pollard et al. (1982). Moreover, the convergence rate depends on $b_1$, because the sparsity of $\widehat{\boldsymbol{a}}$ is bounded by $b_1$ as a result of the $L_1$ and $L_2$ constraints in our estimation. If $b_1 = O(1)$, $\log(pq) = o_p(n)$, then $\|\boldsymbol{a}_{0k} - \boldsymbol{a}_{0l}\|_2$ is bounded below by a constant, for $k \neq l$, and thus the clustering consistency is achieved as $n \to \infty$. It is also noteworthy that Theorem 2 does not require the number of identified subgroups to be the same as the true number of subgroups $G$, as long as $G$ is fixed. In fact, this theorem shows that the rate of convergence for clustering depends on $G$ asymptotically.

## 5. Simulations

### 5.1 Estimation error

We first evaluate the estimation error and its convergence. We generate $Y_i$ based on model (21), generate the components of $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ from a standard Gaussian distribution, and generate the error $\epsilon_i$ from a Gaussian distribution with mean zero and variance 0.25. We set $\boldsymbol{\Theta}_0$ with its upper-left $5 \times 5$ submatrix equal to

$$\begin{pmatrix} 0.293 & 1.291 & 0.617 & 0.269 & -0.458 \\ -1.092 & -0.514 & -2.615 & 1.155 & 0.400 \\ 1.185 & -0.279 & 1.332 & 2.198 & 2.853 \\ 2.183 & -2.295 & 0.104 & -2.208 & 1.301 \\ -1.378 & -2.211 & -2.668 & -1.190 & -1.427 \end{pmatrix}, \tag{59}$$

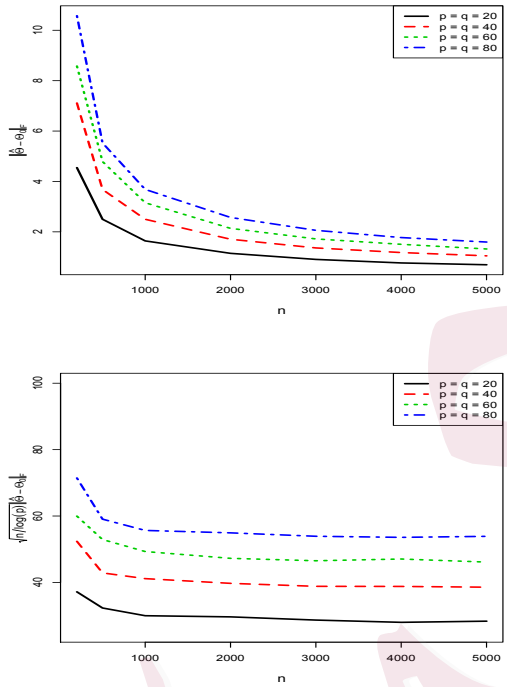


**Figure 1:** Estimation error $\|\mathrm{vec}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0)\|_2$ for different values of $p = q$ as $n$ increases.

where the entries are generated from Uniform$(-3, 3)$, and the rest entries of $\mathbf{\Theta}_0$ being zero. We set $p = q$, and vary this value along with the sample size $n$.

Figure 1 reports the estimation error $\|\mathrm{vec}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0)\|_2$ first in the original scale, then in the scale of $\sqrt{n/\log(pq)}\|\mathrm{vec}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0)\|_2$, for different values of $p, q$ as $n$ increases. The plot is based on 100 data replications. It is seen that the estimation error monotonically decreases as the sample size increases, and it remains a constant approximately at the scale of $\sqrt{n/\log(pq)}\|\mathrm{vec}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}_0)\|_2$, both of which agree with our theoretical result in Theorem 1.

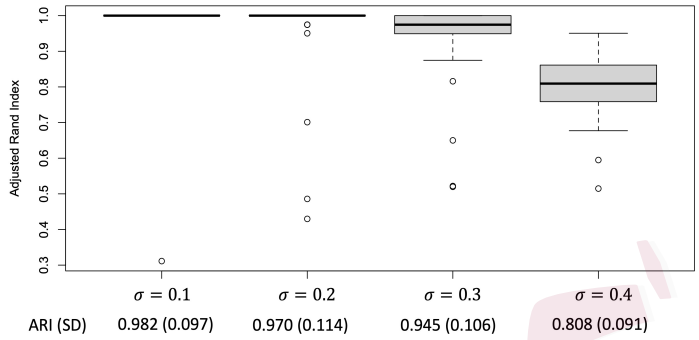


**Figure 2:** Average adjusted rand index between the clustering results using the estimated $\widehat{\boldsymbol{\Theta}}\boldsymbol{Z}$ and using the true $\boldsymbol{\Theta}_0\boldsymbol{Z}$, respectively, for different values of $\sigma$.

## 5.2 Clustering

We next investigate the clustering performance of our method. We adopt a similar setting as in Section 5.1, and we fix $p = 150$, $q = 200$, and $n = 100$. We set $\boldsymbol{\Theta}_0$ with its upper-left $5 \times 5$ submatrix equal to (59), and the rest entries being zero. We generate $Z_{ik}$ from a Gaussian distribution with mean $\mu_{ik} = \sum_{g=1}^{5}(g-1)\mathbb{I}\big\{20(g-1)+1 \le i \le 20g, 2 \le k \le 10\big\}$, where $\mathbb{I}(\cdot)$ is an indicator function, for $i = 1, \ldots, n, k = 1, \ldots, q$, and standard deviation $\sigma_k$. By design, there are $G = 5$ clusters, with 20 sample observations per cluster, that are determined by the clustering pattern in the linear combinations of $Z_{ik}$, for $k = 2, \ldots, 5$, where the linear combination coefficients are determined by the second to the fifth column of the submatrix in (59). Meanwhile, the remaining columns of $Z_{ik}$, for $k = 6, \ldots, q$, would not affect the subgroup structure, be-

cause the corresponding linear combination coefficients are all zero. We vary the variation level of the clusters as $\sigma_1 = \ldots = \sigma_5 = \sigma \in \{0.1, 0.2, 0.3, 0.4\}$, and fix $\sigma_k = 0.3$ for $6 \leq k \leq 10$, and $\sigma_k = 1$ for $k > 10$. We apply our method to estimate $\boldsymbol{\Theta}$, then apply $K$-means clustering with 5 groups based on the estimated $\widehat{\boldsymbol{\Theta}}\boldsymbol{Z}_i$. We also consider a benchmark solution by applying $K$-means on the true $\boldsymbol{\Theta}_0\boldsymbol{Z}_i$.

Figure 2 reports the average adjusted rand index (Sinnott et al., 2016), which provides a robust measure of the similarity between the two data clustering results. The plot is based on 100 data replications. It is seen that the clustering based on the estimated $\widehat{\boldsymbol{\Theta}}\boldsymbol{Z}_i$ works well, and generally agrees with the benchmark clustering based on the true $\boldsymbol{\Theta}_0\boldsymbol{Z}_i$. In addition, ARI decreases when the noise level $\sigma$ increases, which agrees with the expectation.

### 5.3 Sparsity

We next investigate the selection accuracy of the sparse estimation. We again adopt the similar setting as that in Section 5.1, except that we consider a sparse $\boldsymbol{\Theta}_0$, with $p = 91$, $q = 18$, and the sample size $n = 1901$, where $p, q, n$ are the same as the dimensions of $\boldsymbol{X}$, $\boldsymbol{Z}$ and the sample size in the real data in Section 6. More specifically, we set the entries of $\boldsymbol{\Theta}_0$ at the first and second columns and at rows 1, 11, 21, and 81 to be nonzero, and the rest zero. We set the

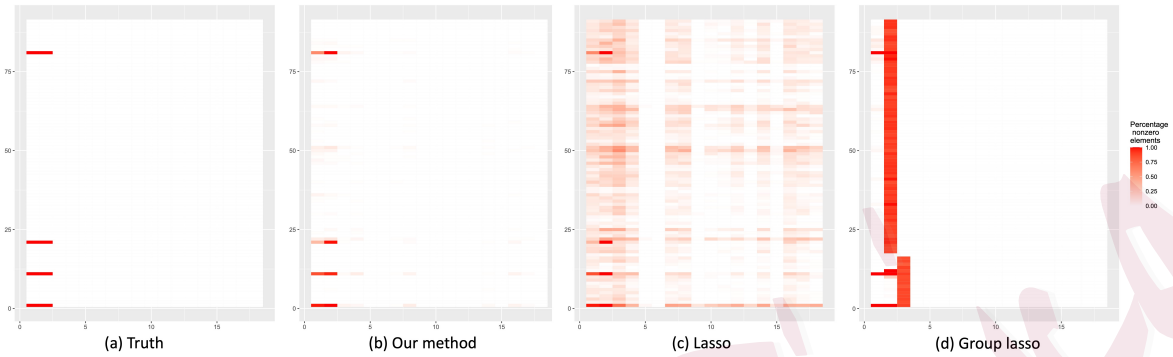


**Figure 3:** Heat map of the true nonzero entries (a), and the estimated nonzero entries by our method (b), the entry-wise Lasso (c), and column-wise group Lasso (d).

nonzero values of the first column as $0.088$, $-0.274$, $-0.093$, and $0.102$, which are randomly generated from a uniform distribution on $[-0.2, 0.2]$, and those of the second column as $0.691$, $-1.222$, $1.621$, and $1.575$, which are randomly generated from a uniform distribution on $[-3, 3]$. We also compare our method with two alternative solutions, the usual Lasso method applied to the individual entries of $\mathbf{\Theta}$, and the group Lasso method applied to each column of $\mathbf{\Theta}$, for $k = 2, \ldots, q$.

Figure 3 shows the heat map of the true nonzero entries and the estimated nonzero entries by the three methods. The plot is based on 100 data replications. It is seen that our method performs substantially better than the alternative solutions in sparse recovery of $\mathbf{\Theta}_0$.

## 6. ABCD Data Analysis

### 6.1 Background and data

The Adolescent Brain Cognitive Development (ABCD) Study is the largest long-term study of brain development and child health in the United States. It aims to follow over 10,000 children from their childhood through adolescence, and to understand biological and environmental factors that impact their brain development and health (Casey et al., 2018). One of the key scientific questions is to investigate the association between cognitive ability, measured by the child's cognitive score or g-factor (Deary et al., 2007), and working memory brain activity, measured by the task fMRI. In the ABCD study, the working memory task fMRI is collected based on the emotional n-back tasks that engage the processes related to memory and emotion regulation (Casey et al., 2018). Our analysis in particular focuses on brain activity in response to the 0-back task that involves low memory load. At the beginning of fMRI scanning for the 0-back task, children are presented a target stimulus. Then during the scanning they are asked to hit a button for "match" when they see an identical picture and a button for "no match" when they see a different picture. Due to substantial growth and cognitive development during adolescence, the association between brain activity and cognitive ability can be highly heterogeneous from individual to individual. It is

thus important to account for potential subject heterogeneity in the association modeling. The dataset we analyze comes from the ABCD study, Release 1.0, which consists of $n = 1,901$ children from 9 to 11 years old. The response is the cognitive score. The primary predictors are the 0-back fMRI activation measurements over $p = 90$ automated anatomical labelling (AAL) regions (Tzourio-Mazoyer et al., 2002). In addition, there are a set of additional variables, including the psychological score or p-factor (Caspi et al., 2014), which measures general features of mental disorders, plus age, race, parental education, parental marital status, family size, and income. Together they form a set of $q = 18$ auxiliary predictors. We apply the proposed method to this dataset. To further amend the sparse estimation, we apply a bootstrap method to construct a $95\%$ confidence interval for each estimated parameter.

### 6.2 Subgroups and associations

We first examine the auxiliary predictors $\boldsymbol{Z}_i$ and the associated subgroups. Figure 4 plots the average effect of $\boldsymbol{Z}_i$ on $Y_i$, i.e., the $q$-dimensional estimate $n^{-1}\sum_{i=1}^{n}\widehat{\boldsymbol{\Theta}}^{\top}\boldsymbol{X}_i$, over 100 bootstrap replications. The left panel is based on our method, and the right panel based on the usual Lasso with $\boldsymbol{Z}_i$ as the predictor for comparison. It is seen that the psychological score, age, race, parental education whether or not having a postgraduate degree, parental marital status, and income, contribute

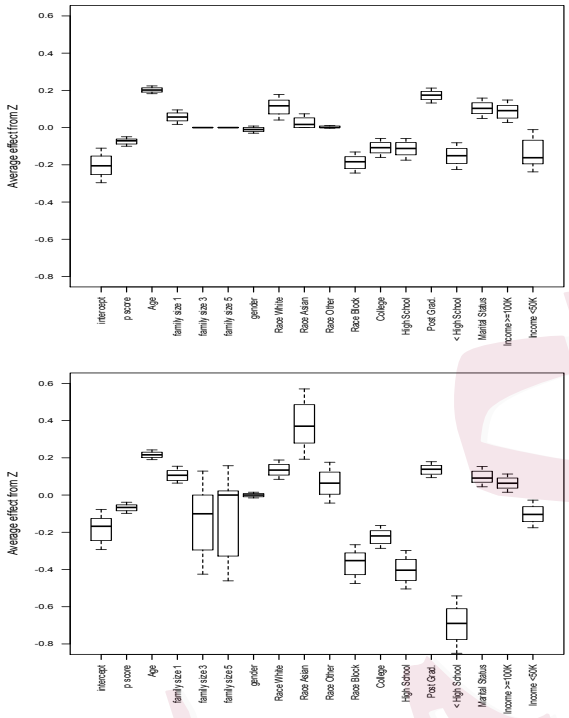


**Figure 4:** Average effect of the auxiliary predictors on the response, obtained from our method (left panel), and the usual Lasso (right panel). The error bars represent 95% confidence intervals obtained from 100 bootstraps.

significantly to clustering children into different subgroups. Meanwhile, our estimate has a much smaller variation compared to the usual Lasso estimate.

Next, we compare the analysis based on the entire study cohort to that based on the two groups of subjects divided by the psychological score. This is mainly motivated by the domain interest that the psychological score can be a major factor that influences the association between the cognitive outcome and 0-back (low memory load) brain activities. Figure 5, top row, plots the significantly

associated brain regions for the entire data, for those subjects whose psychological score is lower than 1.28, and for those subjects whose score is higher. The statistical significance of the association of interest is determined by the bootstrap method, i.e., whether $\geq 97.5\%$ of the bootstrapped regression coefficient estimates are positive or negative. The threshold value 1.28 is the largest psychological score in the low score group identified by the $K$-means clustering when applied to the psychological score variable. It is seen that, the association pattern for the low psychological score group is similar to that for the entire group, in that the left and right inferior frontal gyrus (the opercular part) have a positive effect on the cognitive outcome, while the right postcentral gyrus has a negative effect. By contrast, in the high score group, no AAL region is found significantly associated with the cognitive outcome, showing a heterogeneous relation between the brain activity and cognitive ability. The inferior frontal gyrus is related to several well known brain functions including language processing (Winhuisen et al., 2005), working memory (Liakakis et al., 2011) and spatial attention (Hartwigsen et al., 2019). Our results suggest that mental health issues of children may affect their brain activity in the inferior frontal gyrus and its related cognitive functions.

Next, instead of using the psychological score to form the clusters, we apply the $K$-means algorithm to the estimated $\widehat{\boldsymbol{\Theta}}\boldsymbol{Z}_i$. Based on the BIC criterion, we

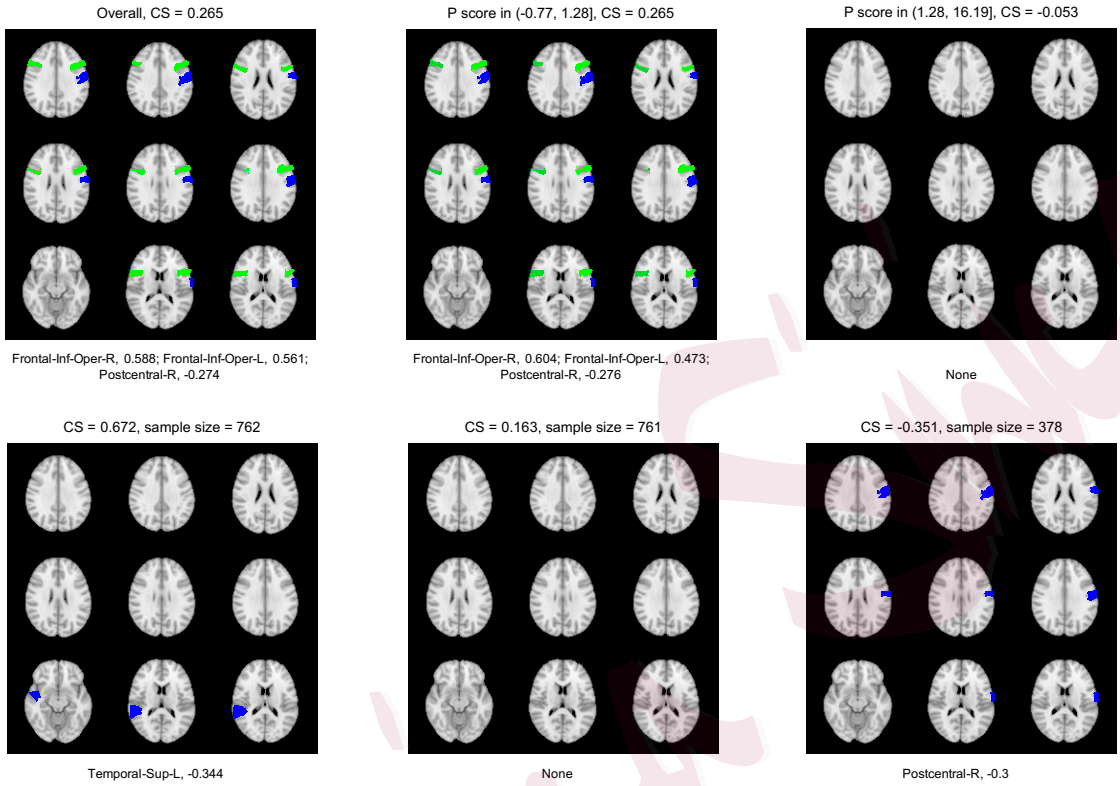


**Figure 5:** Association patterns between brain regions and cognitive outcome for different subgroups. The top row is for the entire study cohort, the group with the psychological score lower than the threshold, and the group higher than the threshold. The bottom row is for the three subgroups identified by $K$-means applied to $\widehat{\boldsymbol{\Theta}}\boldsymbol{Z}_i$. Significant positive and negative associations are represented by green and blue colors, respectively, for corresponding brain regions.

identify three subgroups, which happen to correspond to the high, medium, and low cognitive scores approximately. After identifying the subgroups, we refit the regression model of $Y$ on $X$ within each subgroup separately. We comment that this is mainly to simplify the analysis, but it may underestimate the variation in

the association between $Y$ and $\boldsymbol{X}$. Figure 5, bottom row, plots the significantly associated brain regions for the three subgroups. It is seen that, the superior temporal gyrus has a negative effect only for the high cognitive score group, while the right postcentral gyrus has a negative effect only for the low cognitive score group. This again demonstrates heterogenous patterns for different subgroups of subjects. It has been shown that the superior temporal gyrus (Mesgarani et al., 2014) involves the high-order auditory processing of speech, while the postcentral gyrus (Thomas et al., 1999) is related to spatial working memory task and visual processing. The identified heterogenous brain activation patterns between the high and low cognitive score groups provide new insights on how the different types of cognitive skills contribute differently to a general cognitive ability,

### 6.3 Predictions

Finally, we evaluate the prediction performance of our proposed method in two ways. We first randomly split the data into two halves, and report the $R^2$ measure based on the testing data. We also compare our method with a number of alternative solutions, including the usual Lasso method applied to the individual entries of $\boldsymbol{\Theta}$, and the group Lasso method applied to each column of $\boldsymbol{\Theta}$ for $k = 2, \ldots, q$, similarly as in Section 5.3. In addition, we compare with the Lasso regression with $\boldsymbol{X}_i$ as the predictors, but separately for the three clusters iden-

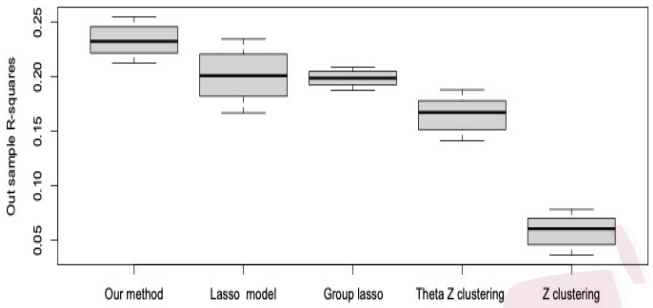


**Figure 6:** Prediction performance in terms of $R^2$ from 100 random data splits. The mean and the standard deviation are shown in the bottom row.

tified based on $\widehat{\boldsymbol{\Theta}}\boldsymbol{Z}_i$, and the Lasso regression separately for the three clusters identified based on $\boldsymbol{Z}_i$ directly. We repeat the random splits 100 times. Figure 6 shows the box plot of the $R^2$ values from 100 splits. It is clearly seen that our method achieves the best performance in terms of $R^2$. We also note that separate Lasso regression based on the clustering results using $Z_i$ directly performs poorly.

Moreover, we evaluate the prediction performance by using the Lasso model learnt from the samples in one cluster to predict the outcome of the samples in another cluster. Table 1 reports the corresponding mean squared prediction error. It is seen that the best prediction accuracy is achieved when predicting the same cluster of samples, suggesting that there is indeed data heterogeneity,

**Table 1:** Prediction performance. The mean squared prediction error for cross training and testing. All results are based on 100 random data splits.

| Groups | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 0.598 (0.059) | 0.901 (0.274) | 1.304 (0.339) |
| 2 | 0.948 (0.266) | 0.560 (0.048) | 0.901 (0.269) |
| 3 | 1.353 (0.370) | 0.903 (0.271) | 0.562 (0.054) |

and taking into account such heterogeneity actually helps improve the prediction performance.

## 7. Discussion

In this article, we have proposed a new approach for subgroup analysis in regression modeling. Our key idea is to treat the primary and auxiliary predictors separately, and model the heterogenous association through the interaction between the primary and auxiliary variables. In addition, we achieve simultaneous feature selection for both primary and auxiliary predictors through proper penalty functions. Our theoretical and numerical analyses demonstrate the effectiveness of the proposed method. Meanwhile, we remark that, although we adopt a linear type regression model and consider a continuous response variable, it is possible to extend our method to nonlinear models and other types of outcomes, by

adopting different link functions.

Our current work relies on a key underlying structure that $\Theta_0$ is sparse. Such a sparsity structure has several advantages. It makes the interpretation easier, as it allows us to identify and focus on the subset of auxiliary predictors that contribute to the heterogenous association between the response and the primary predictors. In addition, it leads to a low-dimensional $\Theta_0 Z$, which would in turn facilitate the downstream clustering analysis. As we have shown in Section 6.3, directly clustering based on the high-dimensional auxiliary covariates $Z$ would lead to a poor performance. Alternatively, one may consider another structure such that $\Theta_0$ is of a low rank. This can be achieved by introducing a nuclear norm type penalty in $\mathcal{R}(\Theta)$, and defining the subgroups according to the nonzero singular values of $\Theta_0 Z$. However, the low-rank structure is harder to interpret, and its effect on downstream clustering requires more investigations. We feel a full treatment is beyond the scope of this article, and we leave this extension as future research.

## Supplementary Materials

The supplementary document online includes the comprehensive proofs of all theoretic results. .

## Acknowledgements

This work was supported by NIH grants K25AG071840, R01AG061303, R01AG062542, R01DA048993, R01MH105561, NSF grant CIF-2102227 and IIS2123777, Resource Allocation Program (RAP) grant from UCSF,.

## References

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821.

Brent, R. (1973). An algorithm with guaranteedconvergenceforfindingazero ofa-function. *BrentRP. Algorithms forminimizationwithoutderivatives. Englewood Cliffs (NJ): Prentice-Hall*.

Cai, T., Tian, L., Wong, P. H., and Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282.

Casey, B., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., et al. (2018). The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience*, 32:43–54.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., et al. (2014).

The p factor: one general psychopathology factor in the structure of psychiatric disorders? *Clinical psychological science*, 2(2):119–137.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625.

De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag, New York.

Deary, I. J., Strand, S., Smith, P., and Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1):13–21.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279.

Fan, A., Song, R., and Lu, W. (2017). Change-plane analysis for subgroup detection and sample size calculation. *Journal of the American Statistical Association*, 112(518):769–778.

Foster, J., Taylor, J., and Ruberg, S. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30:2867–80.

Hartwigsen, G., Neef, N. E., Camilleri, J. A., Margulies, D. S., and Eickhoff, S. B. (2019). Functional segregation of the right inferior frontal gyrus: ev-

idence from coactivation-based parcellation. *Cerebral Cortex*, 29(4):1532–1546.

Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):155–176.

Hu, X., Huang, J., Liu, L., Sun, D., and Zhao, X. (2021). Subgroup analysis in the heterogeneous cox model. *Statistics in Medicine*, 40(3):739–757.

Huang, J. and Zhang, T. (2010). The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004.

Jiang, F. and Ma, Y. (2021). Poisson regression with error corrupted high dimensional features. *Statistica Sinica*, 0(0):00.

Liakakis, G., Nickel, J., and Seitz, R. (2011). Diversity of the inferior frontal gyrus—a meta-analysis of neuroimaging studies. *Behavioural brain research*, 225(1):341–347.

Liu, H. and Zhang, J. (2009). Estimation consistency of the group lasso and its applications. In *Artificial Intelligence and Statistics*, pages 376–383. PMLR.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with

noisy and missing data: provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Luby, J. L., Barch, D. M., Belden, A., Gaffrey, M. S., Tillman, R., Babb, C., Nishino, T., Suzuki, H., and Botteron, K. N. (2012). Maternal support in early childhood predicts larger hippocampal volumes at school age. *Proceedings of the National Academy of Sciences*, 109(8):2854–2859.

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

Ma, S., Huang, J., Zhang, Z., and Liu, M. (2019). Exploration of heterogeneous treatment effects via concave fusion. *The international journal of biostatistics*, 16(1):20180026.

Mackey, S., Chaarani, B., Kan, K.-J., Spechler, P. A., Orr, C., Banaschewski, T., Barker, G., Bokde, A. L., Bromberg, U., Büchel, C., et al. (2017). Brain regions related to impulsivity mediate the effects of early adversity on antisocial behavior. *Biological psychiatry*, 82(4):275–282.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010.

Pollard, D. et al. (1982). A central limit theorem for $k$-means clustering. *The Annals of Probability*, 10(4):919–926.

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

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245.

Sinnott, R., Duan, H., and Sun, Y. (2016). Chapter 15—a case study in big data analytics: exploring twitter sentiment analysis and the weather. *Big Data*, pages 357–388.

Sripada, C., Rutherford, S., Angstadt, M., Thompson, W. K., Luciana, M., Weigard, A., Hyde, L., and Heitzeg, M. (2020). Prediction of neurocognitive profiles in youth from resting state fmri. *Molecular Psychiatry*, 25:3413–3421.

Tang, L. and Song, P. X.-K. (2021). Poststratification fusion learning in longitudinal data analysis. *Biometrics*, 77(3):914–928.

Tang, X., Xue, F., and Qu, A. (2020). Individualized multidirectional variable selection. *Journal of the American Statistical Association*, 0(0):1–17.

Thomas, K. M., King, S. W., Franzen, P. L., Welsh, T. F., Berkowitz, A. L., Noll, D. C., Birmaher, V., and Casey, B. (1999). A developmental functional mri study of spatial working memory. *Neuroimage*, 10(3):327–338.

Tibshirani, R. and Friedman, J. (2020). A pliable lasso. *Journal of Computational and Graphical Statistics*, 29(1):215–225.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97(4):893–904.

Wang, W., Phillips, P. C., and Su, L. (2018). Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics*, 33(6):797–815.

Wei, S. and Kosorok, M. R. (2013). Latent supervised learning. *Journal of the American Statistical Association*, 108(503):957–970.

Winhuisen, L., Thiel, A., Schumacher, B., Kessler, J., Rudolf, J., Haupt, W. F.,

and Heiss, W. D. (2005). Role of the contralateral inferior frontal gyrus in recovery of language function in poststroke aphasia: a combined repetitive transcranial magnetic stimulation and positron emission tomography study. *Stroke*, 36(8):1759–1763.

Yin, P., Esser, E., and Xin, J. (2014). Ratio and difference of $l\_1$ and $l\_2$ norms and sparse representation with coherent dictionaries. *Communications in Information and Systems*, 14(2):87–109.

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

Zhang, Y., Wang, H. J., and Zhu, Z. (2019). Robust subgroup identification. *Statistica Sinica*, 29(4):1873–1889.

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L. J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, 108(502):527–539.

Zhu, X. and Qu, A. (2018). Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics*, 12(1):171–193.