Identifying Group-wise Mixtures in High-Dimensional Multi-Task Learning Model with an Application in Reconstructing Brain Subnetworks

Yisha Yao, Wei Dai, Shiying Wang, Zihuan Liu, Heping Zhang

Abstract

Existing methods for detecting communities, subnetworks, or clusters in biological/social systems are mostly based on observed information of each entity in the system. The performance of these approaches largely depend on how much the observable data manifest the underlying cluster/community/subnetwork structures. However, in many cases, the latent structures are not directly manifested by any observable features. Instead these structures are hidden beneath the interactions between the system and outside factors, or depend on the higher order information of the system itself. In this paper, we build a new statistical model that is tailored to this type of real-world problems. We also develop an efficient algorithm to fit our model. Both theoretical guarantees and empirical performances are shown to support our model and algorithm. Finally, we implement our method to a real problem in neuroscience, constructing brain subnetworks, and make some new discoveries.

1 Introduction

Inferring clusters, community structures, or subnetworks in biological systems, social and/or other settings are often of great interest, which motivates the development of many analytic methods. The existing approaches tend to be based on observed information of each entity in the system and then apply established statistical techniques, *e.g.*, graphical model (Chandrasekaran et al., 2010), community detection (Gao et al., 2018), and spectral clustering (Ng et al., 2001) to the observed data. These approaches have enjoyed successes in some cases, for example, differentiating cell populations in flow cytometry based on a set of intracellular or surface proteins markers (Zare et al., 2010), community detection of evolutionary modules of sub-domains in proteins (Hleap et al., 2013), and predicting direct couplings from protein sequence data (Barton et al., 2016). However, the performance of these approaches largely depend on how much the observable data manifest the underlying cluster/community/subnetwork structures.

In practice, the latent structures are not directly manifested by observable features. Instead these structures are hidden beneath the interactions between the system and outside factors, or depend on the higher order information of the system itself. For instance, constructing signaling transduction pathways relies on the dynamic changes in the concentrations of the biochemical species, which needs to be first modeled by ordinary differential equations (Sackmann et al., 2006). Protein-protein interaction networks are built on the evolving co-expression functions among the proteins, which are to be exyrtacted from the data (Wang et al., 2014).

An important and emerging example is brain subnetworks. It has been reported that genetic influences over brain vary by regions (Thompson et al., 2001; Smith et al., 2020). The brain regions in the same functional or structural subnetwork are more likely to have common genetic mechanisms (Wen et al., 2016; van der Meer et al., 2020): the structures of Broca's and Wernicke's language areas (Thompson et al., 2001) are more prone to genetic influences than other brain regions; genetic factors constrain the imaging-phenotypes of the visual, sensorimotor, basal ganglia regions more than those of the default-mode, executive control and attention regions (Fu et al., 2015); and the functional connectivity of default-mode network regions seem to be influenced by a specific set of genes (Glahn et al., 2010). These discoveries suggest that using genetic information in identifying brain subnetworks may help us better understand brain functions and gene-brain interactions. Reconstructing brain subnetworks as presented in Section 6 is indeed an objective of this study.

The rest of this paper is organized as follows. In Section 2, we propose a new statistical model that is tailored to address aforementioned biological problems. In Section 3, we develop an efficient algorithm to fit Model (2.1). Some theoretical guarantees are provided in Section 4 for the proposed method. We also show extensive simulation results in Section 5. Finally in Section 6, we apply Model (2.1) and the proposed algorithm to construct brain subnetworks based on genetic information.

2 Model

Motivated by the above biological problems, we propose a model below which integrates multiple linear models with group-specific mixture coefficients.

$$\begin{bmatrix}\boldsymbol{y}_1, \cdots, \boldsymbol{y}_p\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1, \cdots, \boldsymbol{X}_q\end{bmatrix} \begin{bmatrix}\boldsymbol{b}_{11} & \cdots & \boldsymbol{b}_{1p} \\ \vdots & & \vdots \\ \boldsymbol{b}_{q1} & \cdots & \boldsymbol{b}_{qp}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1, \cdots, \boldsymbol{\varepsilon}_p\end{bmatrix}, \tag{2.1}$$

where each $\boldsymbol{y}_j \in \mathbb{R}^n$ is n *i.i.d.* features of the j-th entity, $\boldsymbol{X}_\ell \in \mathbb{R}^{n \times d_\ell}$ is n *i.i.d.* measurements of the ℓ-th biological factor, $\boldsymbol{b}_{\ell j} \in \mathbb{R}^{d_\ell}$ is a vector of coefficients quantifying the effect of the ℓ-th factor on the feature of the j-th entity, and $\{\boldsymbol{\varepsilon}_j\}$ are *i.i.d.* $N(0, \sigma^2 \boldsymbol{I}_n)$ noise vectors.

For each $1 \leq \ell \leq q$, $\{\boldsymbol{b}_{\ell 1}, \ldots, \boldsymbol{b}_{\ell p}\}$ are assumed from a mixture distribution that consists of several spherical Gaussian components and an "all-zero" component, *i.e.*, a point mass on $\mathbf{0} \in \mathbb{R}^{d_\ell}$. The point mass on $\mathbf{0}$ promotes group sparsity in Model (2.1). Let K_ℓ be the number of Gaussian components in the ℓ-th mixture distribution, $w_{\ell 0}$ the weight of the "all-zero" component ($0 \leq w_{\ell 0} \leq 1$), $w_{\ell 1}, \ldots, w_{\ell K_\ell}$ the weights of the Gaussian components ($w_{\ell j} \geq 0$), $\boldsymbol{\mu}_{\ell 1}, \ldots, \boldsymbol{\mu}_{\ell K_\ell}$ the means of the Gaussian components, and $z_{\ell 1}, \ldots, z_{\ell p}$ the indicators that $\boldsymbol{b}_{\ell j}$ belongs to which component among $\{0, 1, \ldots, K_\ell\}$. We assume $\boldsymbol{b}_{\ell j} \sim N(\boldsymbol{\mu}_{\ell k}, \boldsymbol{I}_{d_\ell})$ given $z_{\ell j} = k$; and $\boldsymbol{b}_{\ell j} = \mathbf{0}$ with probability 1 given $z_{\ell j} = 0$. The total number of mixture components is $K_\ell + 1$ including the "all-zero" component.

Note that for some ℓ, K_ℓ can be zero meaning that the ℓ-th factor does not affect any of the p entities. To have a flexible model, we allow the mixture distributions to vary by ℓ. To avoid over-parametrization, we assume that $\{\boldsymbol{b}_{\ell j}\}_{\ell, j}$ are mutually independent. We also assume that the covariance matrices of all the Gaussian components are identity matrices. This is reasonable because if necessary, we can perform orthonormalization within each block $\boldsymbol{X}_\ell$ so that the corresponding coefficients are independent.

We need to estimate the group labels $z_{\ell 1}, \ldots, z_{\ell p}$ and the unknown Gaussian centers $\boldsymbol{\mu}_{\ell 1}, \ldots, \boldsymbol{\mu}_{\ell K_\ell}$. Before doing so, we rewrite Model (2.1) as follows

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}, \tag{2.2}$$

where $\boldsymbol{B}$ represents the entire coefficient matrix $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_{11} & \cdots & \boldsymbol{b}_{1p} \\ \vdots & & \vdots \\ \boldsymbol{b}_{q1} & \cdots & \boldsymbol{b}_{qp} \end{bmatrix}$, and $\boldsymbol{E} = \begin{bmatrix} \boldsymbol{\varepsilon}_1, \cdots, \boldsymbol{\varepsilon}_p \end{bmatrix}$. The coefficient matrix $\boldsymbol{B}$ is not observed and needs to be estimated. Due to the point mass at zero, $\boldsymbol{B}$ is groupwise sparse, *i.e.*, only a few groups of coefficients are nonzero. $\boldsymbol{Y}$ and $\boldsymbol{X}$ are observed with $\boldsymbol{X}$ being a fixed design matrix.

To elucidate Model (2.1), we use brain imaging data as an illustrative example. $\boldsymbol{y}_j$ is the brain phenotypes of n individuals in the j-th brain region, $\boldsymbol{X}_\ell$ represents the genotypes of n individuals at d_ℓ single nucleotide polymorphism (SNP) loci within the ℓ-th genomic region, and $\boldsymbol{b}_{\ell j}$ quantifies the specific genetic effect on the particular brain region's phenotype ($j = 1, \ldots, p,\ \ell = 1, \ldots, q$). Here p is the total number of brain regions, q is the total number of genomic regions. We group the SNPs within each genomic region as one factor since we are exploring the genetic effects at the "super-variant" level. We expect each factor to affect a small fraction of brain regions and each brain region to be influenced by a few factors, so $\boldsymbol{B}$ is group-wise sparse.

Our model builds upon the framework of multi-response regression model (Kim and Xing, 2012), but our key and novel idea is to incorporate group-specific mixture structures into the group-wise sparse coefficients so that we can perform both group selection and clustering. In addition, $\{\boldsymbol{b}_{\ell_1 1}, \ldots, \boldsymbol{b}_{\ell_1 p}\}$ and $\{\boldsymbol{b}_{\ell_2 1}, \ldots, \boldsymbol{b}_{\ell_2 p}\}$ may follow different mixture models for $\ell_1 \neq \ell_2$. In another word, the mixture models are group-specific and hence create flexibility in fitting real data.

3 Methodology

To fit Model (2.1) and eventually obtain the clustering labels, a statistician probably first thinks of Maximum Likelihood Estimator (MLE). However, the point mass component and the Gaussian components are not in the same probability measure space, a joint likelihood function including both probability mass and probability density is meaningless. Thus, we will detour via a more efficient way. Some analytic heuristics are given below, which eventually leads to an iterative approach.

Let $\mathcal{G}_j \subset [q]$ be the group index set of the nonzero groups in the true $\boldsymbol{b}_j$ (group support) and $\overline{\mathcal{G}_j}$ be its complement. Let $\omega_{\ell k} = w_{\ell k}/(1 - w_{\ell 0})$. If $\{\mathcal{G}_j\}$ is known, we can write out separately the log likelihood function for the discrete components (indicators $\left\{\mathbb{1}\{\boldsymbol{b}_{\ell j} = \mathbf{0}\}\right\}_{\ell j}$) using probability mass and the log likelihood function for the Gaussian components using probability density. The log likelihood for the indicators is

$$\sum_{\ell=1}^{q}\sum_{j=1}^{p}\Big[\mathbb{1}\{\ell \in \overline{\mathcal{G}_j}\}\log w_{\ell 0} + \mathbb{1}\{\ell \in \mathcal{G}_j\}\log(1 - w_{\ell 0})\Big], \tag{3.1}$$

and the maximum likelihood estimator (MLE) for $w_{\ell 0}$ would be $\hat{w}_{\ell 0} = \frac{\#\{j:\boldsymbol{b}_{\ell j}=\mathbf{0}\}}{p}$. The probability density for $\boldsymbol{b}_{\mathcal{G}_j j}$ given $\{\boldsymbol{b}_{\ell j} \neq \mathbf{0} : \ell \in \mathcal{G}_j\}$ is

$$g_{\boldsymbol{\theta}}(\boldsymbol{b}_{\mathcal{G}_j j}) = \prod_{\ell \in \mathcal{G}_j}\left(\sum_{k=1}^{K_\ell}\omega_{\ell k}\Phi_{\ell k}(\boldsymbol{b}_{\ell j})\right).$$

And hence the log likelihood for $(\boldsymbol{y}_j, \boldsymbol{b}_{\mathcal{G}_j j})$ is proportional to

$$\begin{aligned}\log p_{\boldsymbol{\theta}}(\boldsymbol{y}_j, \boldsymbol{b}_{\mathcal{G}_j j}) &\propto -\frac{1}{2\sigma^2}\left\|\boldsymbol{y}_j - \boldsymbol{X}_{\mathcal{G}_j}\boldsymbol{b}_{\mathcal{G}_j j}\right\|^2 + \sum_{\ell \in \mathcal{G}_j}\log\left(\sum_{k=1}^{K_\ell}\omega_{\ell k}\Phi_{\ell k}(\boldsymbol{b}_{\ell j})\right) \\ &= -\frac{1}{2\sigma^2}\left\|\boldsymbol{y}_j - \boldsymbol{X}_{\mathcal{G}_j}\boldsymbol{b}_{\mathcal{G}_j j}\right\|^2 + \sum_{\ell \in \mathcal{G}_j}\log\left(\sum_{k=1}^{K_\ell}\omega_{\ell k}e^{-\|\boldsymbol{b}_{\ell j} - \boldsymbol{\mu}_{\ell k}\|^2/2}\right).\end{aligned} \tag{3.2}$$

Here the parameter $\boldsymbol{\theta} = (w_{\ell 0}, \omega_{\ell 1}, \ldots, \omega_{\ell K_\ell}, \boldsymbol{\mu}_{\ell 1}, \ldots, \boldsymbol{\mu}_{\ell K_\ell})$. In this situation where the oracle information about $\{\mathcal{G}_j\}$ is available, we just need to maximize $\sum_{j=1}^{p}\log p_{\boldsymbol{\theta}}(\boldsymbol{y}_j, \boldsymbol{b}_{\mathcal{G}_j j})$ over $\boldsymbol{\theta}$. Nevertheless, in reality the true group supports $\{\mathcal{G}_j\}_j$ are unknown and need to be inferred from data. This analysis suggests that we alternate between identifying the group supports $\{\mathcal{G}_j\}_j$ and maximizing $\sum_{j=1}^{p}\log p_{\boldsymbol{\theta}}(\boldsymbol{y}_j, \boldsymbol{b}_{\mathcal{G}_j j})$.

Now we explore how to maximize the log likelihood in (3.2). The challenge of this task lies in two folds. First, the log likelihood function in (3.2) is highly non-convex in $\boldsymbol{\theta}$; second, the only information we have is $(\boldsymbol{Y}, \boldsymbol{X})$. We notice that the two terms in (3.2) can be optimized separately. We first obtain the maximizers $\{\hat{\boldsymbol{b}}_{\ell j}\}_{\ell j}$, and then maximize the second term based on $\{\hat{\boldsymbol{b}}_{\ell j}\}_{\ell j}$. Actually Model (2.1) can be perceived as an "atypical" mixture model with two sets of latent variables. Both $\{\boldsymbol{b}_{\ell j}\}_{\ell j}$ and $\{z_{\ell j}\}_{\ell j}$ are not observed. The $\{\boldsymbol{b}_{\ell j}\}_{\ell j}$ are sampled from mixture distributions of interest, and yet not directly observed. Their values are hidden beneath $(\boldsymbol{Y}, \boldsymbol{X})$. In this case, we need to unveil two layers of "mysteries".

In view of the above, we consider a generalized Expectation Maximization (EM) algorithm. The conventional EM algorithm (Dempster et al., 1977) has a long history in practically solving mixture models. It alternates between classification given the parameters and estimation given the labels. Its nice empirical performances have been widely acknowledged (Little and Rubin, 2019). A line of recent papers have also demonstrated its reasonable theoretical properties (Balakrishnan et al., 2017; Cai et al., 2019; Wu and Zhou, 2019).

In our atypical mixture model (2.1), three sets of quantities are unknown: the parameter $\boldsymbol{\theta}$, the labels $\{z_{\ell j}\}_{\ell,j}$, and the coefficients $\{\boldsymbol{b}_{\ell j}\}_{\ell,j}$. The proposed generalized EM algorithm iteratively estimates one set of unknown quantities given the rest two. More specifically, in Step 3 below we update $\{\boldsymbol{b}_{\ell j}^{t+1}\}_{\ell,j}$ given the parameters $\boldsymbol{\theta}^t = (\{w_{\ell k}^t\}_k, \{\boldsymbol{\mu}_{\ell k}^t\}_k)$ and labels $\{z_{\ell j}^t\}_{\ell,j}$; in Step 4 below we update the parameters $\boldsymbol{\theta}^{t+1}$ and labels $\{z_{\ell j}^{t+1}\}_{\ell,j}$ based on $\{\boldsymbol{b}_{\ell j}^{t+1}\}_{\ell j}$. When updating $\{\boldsymbol{b}_{\ell j}^{t+1}\}_{\ell,j}$, we sequentially update the on-support part $\{\boldsymbol{b}_{\ell j}^{t+1} : \ell \in \mathcal{G}_j^t\}$ in Step 3a and the group supports $\{\mathcal{G}_j^{t+1}\}_j$ in Step 3b&3c.

In view of the above, we resort to an iterative approach which alternates between updating $\boldsymbol{B}^t$ and clustering $\{\boldsymbol{b}_{\ell j}^t\}_{\ell,j}$ until convergence. Such iterations can be perceived as a generalized EM algorithm which deals with two sets of latent variables. Inside this general EM framework embeds a conventional EM (Step 4) to estimate the parameters and labels given the current $\{\boldsymbol{b}_{\ell j}^t\}_{\ell,j}$. The detailed steps are described below.

- Step 1, obtain an initial estimator $\boldsymbol{B}^0 = [\boldsymbol{b}_{\cdot 1}^0, \ldots, \boldsymbol{b}_{\cdot p}^0]$ by fitting p groupwise linear models, $\boldsymbol{y}_j \sim \boldsymbol{X}\boldsymbol{b}_{\cdot j}^0$ for $1 \leq j \leq p$. In fitting each groupwise linear model, we use group Lasso to select the nonzero groups and then compute the selected groups of coefficients via Minimax Concave Penalized linear unbiased selection (MC+).

- Step 2, cluster the estimates $\{\boldsymbol{b}_{\ell j}^0\}_{\ell,j}$. The estimated zero coefficient vectors automatically belong to the "all-zero" component. For each $1 \leq \ell \leq q$, fit a Gaussian mixture model on the nonzero coefficient vectors $\{\boldsymbol{b}_{\ell j}^0 \neq \mathbf{0}\}$ using EM algorithm, and obtain the estimated cluster number K_ℓ^0, the estimated cluster centers $\{\boldsymbol{\mu}_{\ell 1}^0, \ldots, \boldsymbol{\mu}_{\ell K_\ell^0}^0\}$, and the labels $\{z_{\ell j}^0\}_{\ell,j}$.

- Step 3, update $\boldsymbol{B}^t$ column by column.

- Step 4, cluster the updated $\{\boldsymbol{b}_{\ell j}^{t+1}\}_{\ell,j}$. The zero coefficient vectors automatically belong to the "all-zero" component since $\mathbb{P}_{\boldsymbol{\theta}^t}\{z_{\ell j}^{t+1} = 0 | \boldsymbol{b}_{\ell j}^{t+1} = \mathbf{0}\} = 1$. Fit a Gaussian mixture model on the nonzero coefficient vectors $\{\boldsymbol{b}_{\ell j}^{t+1} \neq \mathbf{0}\}$ for each ℓ by EM algorithm and obtain the estimated cluster number K_ℓ^{t+1}, cluster centers $\{\boldsymbol{\mu}_{\ell k}^{t+1}\}_{k=1}^{K_\ell^{t+1}}$, and labels $\{z_{\ell j}^{t+1}\}_j$.

- Alternate between Step 3 and Step 4 until convergence.

Table 1: The algorithm for block coordinate descent minimization

```
Let $\mathbf{0}$ be the initial estimator, $\tilde{\boldsymbol{y}}_j = \boldsymbol{y}_j - \boldsymbol{X}_{\mathcal{G}_j^t}\boldsymbol{b}^{t+1}_{\mathcal{G}_j^t j}$, and $\tilde{\boldsymbol{X}} = \boldsymbol{X}_{\overline{\mathcal{G}_j^t}}$.
For $\ell \in \overline{\mathcal{G}_j^t}$
    if $\frac{1}{\sigma^2 n}\|\boldsymbol{X}_\ell^\top(\tilde{\boldsymbol{y}}_j - \tilde{\boldsymbol{X}}_{-\ell}\boldsymbol{v}_{-\ell})\| \le \lambda_\ell$, $\boldsymbol{v}_\ell \leftarrow \mathbf{0}$
    else, $\boldsymbol{v}_\ell \leftarrow \arg\min_{\boldsymbol{u}\in\mathbb{R}^{d_\ell}} \frac{1}{2\sigma^2 n}\|\tilde{\boldsymbol{y}}_j - \tilde{\boldsymbol{X}}_{-\ell}\boldsymbol{v}_{-\ell} - \boldsymbol{X}_\ell\boldsymbol{u}\|^2 + \lambda_\ell\|\boldsymbol{u}\|$
end
Repeat the For-loop until convergence
```

For the initialization in Step 1, we combine group Lasso with MC+ to obtain $\{\boldsymbol{b}^0_{\ell j}\}_{\ell,j}$. Group Lasso has nice selection power but produces biased estimates (Meier et al., 2008), while MC+ complements it by generating nearly unbiased estimates (Zhang, 2010). We utilize both the good selection power of group Lasso and the accuracy of MC+ to obtain good starting points. The group Lasso and MC+ are implemented by R packages "grplasso" (Meier et al., 2008) and "plus" (Zhang, 2010), respectively. The penalty level for group Lasso is set to be $\lambda^0_\ell = A^0(\sqrt{d_\ell/n} + \sqrt{\log(pq)/n})$ with A^0 being a tuning parameter. We will justify this choice of λ^0_ℓ in Proposition 1. The penalty level for MC+ is the default value in the R package "plus" (Zhang, 2010).

The conventional EM algorithm in Step 2 and Step 4 is implemented by R package "Mclust" (Scrucca et al., 2016). "Mclust" chooses the cluster number K^t_ℓ based on BIC. It computes the BIC values for a range of cluster numbers, and pick the one that maximizes BIC.

The optimization problem (**??**) in Step 3a has an explicit form

$$\boldsymbol{b}^{t+1}_{\mathcal{G}_j^t j} = \left(\boldsymbol{X}^\top_{\mathcal{G}_j^t}\boldsymbol{X}_{\mathcal{G}_j^t} + \sigma^2\boldsymbol{I}\right)^{-1}\left(\boldsymbol{X}^\top_{\mathcal{G}_j^t}\boldsymbol{y}_j + \sigma^2\bar{\boldsymbol{\mu}}^t_{\mathcal{G}_j^t}\right),$$

where $\bar{\boldsymbol{\mu}}^t_{\mathcal{G}_j^t}$ is the concatenate of $\{\bar{\boldsymbol{\mu}}^t_\ell\}_{\ell\in\mathcal{G}_j^t}$ and $\bar{\boldsymbol{\mu}}^t_\ell = \sum_{k=1}^{K^t_\ell}\mathbb{P}_{\boldsymbol{\theta}^t}(z_{\ell j} = k|\boldsymbol{b}^t_{\ell j})\boldsymbol{\mu}^t_{\ell k}$. In Step 3b, we employ block coordinate descent (Tseng, 2001; Tseng and Yun, 2009) to solve the optimization problem (**??**). The general procedure of block coordinate descent is depicted in Table 1. It cycles through the parameter groups $\{\boldsymbol{v}_\ell\}_{\ell\in\overline{\mathcal{G}_j^t}}$, and minimizes the objective function over the current group while keeping all the other groups fixed. The penalty level in (**??**) is set to be $\lambda^t_\ell = \max\{0.9^t\lambda^0_\ell, \lambda^*_\ell\}$ with λ^*_ℓ a tuning parameter. A decreasing sequence of λ^t_ℓ would grant some chance for the off-support groups in the previous iteration to enter the current support, and thus maintaining vibrant support-exchange along the iterations.

The complete log likelihood for $(\boldsymbol{y}_j, \boldsymbol{b}_{\mathcal{G}_j j}, \{z_{\ell j}\}_{\ell\in\mathcal{G}_j})$ given $\{\boldsymbol{b}_{\ell j} \neq \mathbf{0} : \ell \in \mathcal{G}_j\}$ is proportional to

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{y}_j, \boldsymbol{b}_{\mathcal{G}_j j}, \{z_{\ell j}\}_{\ell\in\mathcal{G}_j}) \propto -\frac{1}{2\sigma^2}\|\boldsymbol{y}_j - \boldsymbol{X}\boldsymbol{b}_{\cdot j}\|^2 + \sum_{\ell\in\mathcal{G}_j}\sum_{k=1}^{K_\ell} \mathbb{1}\{z_{\ell j} = k\}\Big(\log\omega_{\ell k} + \log\Phi_{\ell k}(\boldsymbol{b}_{\ell j})\Big). \tag{3.3}$$

Step 3a is actually maximizing the conditional expectation of (3.3) with respect to the conditional distribution of $z_{\ell j}$ given $z_{\ell j} \neq 0$ under current $\boldsymbol{\theta}^t$ and $\mathcal{G}_j^t$, which is

$$\begin{aligned} Q_{\boldsymbol{\theta}^t,\mathcal{G}_j^t}(\boldsymbol{b}_{\cdot j}\big|\boldsymbol{b}_{\cdot j}^t) &\equiv \sum_{\boldsymbol{z}\in\mathcal{Z}} g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\mathcal{G}_j^t j})\log p_{\boldsymbol{\theta}^t}(\boldsymbol{y}_j, \boldsymbol{b}_{\mathcal{G}_j^t j}, \boldsymbol{z}) \\ &= \sum_{\boldsymbol{z}\in\mathcal{Z}} g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\mathcal{G}_j^t j})\Big(\log g_{\boldsymbol{\theta}^t}(\boldsymbol{z}, \boldsymbol{b}_{\mathcal{G}_j^t j}) + \log p_{\boldsymbol{\theta}^t}(\boldsymbol{y}_j\big|\boldsymbol{b}_{\mathcal{G}_j^t j})\Big) \\ &= \log p_{\boldsymbol{\theta}^t}(\boldsymbol{y}_j\big|\boldsymbol{b}_{\mathcal{G}_j^t j}) + \sum_{\boldsymbol{z}\in\mathcal{Z}} g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\mathcal{G}_j^t j})\log g_{\boldsymbol{\theta}^t}(\boldsymbol{z}, \boldsymbol{b}_{\mathcal{G}_j^t j}), \end{aligned} \tag{3.4}$$

where $\boldsymbol{z}$ is the vector with elements $\{z_{\ell j}\}_{\ell\in\mathcal{G}_j^t}$ and $\mathcal{Z}$ is the space of all possible values of $\boldsymbol{z}$. The high level idea is that given $\mathcal{G}_j^t$ we improve $\boldsymbol{b}^{t+1}_{\mathcal{G}_j^t j}$ (realized in Step 3a), and given $\boldsymbol{b}^{t+1}_{\mathcal{G}_j^t j}$ we improve $\mathcal{G}_j^{t+1}$ (realized in Step 3b and Step 3c). The rigorous arguments are stated in Theorem 2.

4 Theoretical results

In this section, we provide some theoretical guarantees for our algorithm in the special case of three-component mixtures. In this simple case, for each ℓ, $\{\boldsymbol{b}_{\ell 1}, \ldots, \boldsymbol{b}_{\ell p}\}$ are assumed *i.i.d* from a mixture distribution that consists of one "all-zero" component with weight $w_{\ell 0}$ and two Gaussian components with weights $w_{\ell 1}$, $w_{\ell 2}$, where $w_{\ell 1} + w_{\ell 2} + w_{\ell 0} = 1$ and $0 < w_{\ell 0} \leq 1$. Correspondingly, each $z_{\ell j}$ can take values in $\{0, 1, 2\}$. If $w_{\ell 0} < 1$ (there are Gaussian components), $w_{\ell 1}, w_{\ell 2} \in (c_0, 1 - w_{\ell 0} - c_0)$ for some $0 < c_0 < (1 - w_{\ell 0})/2$, otherwise it degenerates to one Gaussian component. We denote the two Gaussian centers by $\boldsymbol{\mu}_{\ell 1}$ and $\boldsymbol{\mu}_{\ell 2}$, which are bounded away from $\mathbf{0}$. Though seemingly restrictive, this simple case is widely observed in biology: the ℓ-th factor up-regulates some entities, down-regulate other entities, and exerts no effects on the rest.

We actually show some nice simulation results in Section 5 for general cases with $K_\ell > 2$ Gaussian components. However, theoretical study for the general case will be left as an open problem since it is mathematically challenging.

Before diving into the technicalities, we introduce some notations, definitions and impose some mild assumptions. Let $\mathcal{A}_\ell \subset \{1, \ldots, d\}$ be the set of coordinate indices within the ℓ-th group, *i.e.*, $\mathcal{A}_1 = \{1, \ldots, d_1\}$, $\mathcal{A}_2 = \{d_1 + 1, \ldots, d_2\}$, ... For any $\mathcal{A} \subset \{1, \ldots, d\}$, define $\boldsymbol{X}_{\mathcal{A}} \in \mathbb{R}^{n\times|\mathcal{A}|}$ to be the submatrix composed of the columns whose indices are in $\mathcal{A}$. For simplicity, we abbreviate $\boldsymbol{X}_{\mathcal{A}_\ell}$ as $\boldsymbol{X}_\ell$ whenever it does not cause any confusion. For any set of group indices $\mathcal{G} \subset \{1, \ldots, q\}$, define $\boldsymbol{X}_{\bigcup_{\ell \in \mathcal{G}} \mathcal{A}_\ell} = \Big(\boldsymbol{X}_\ell, \ell \in \mathcal{G}\Big) \in \mathbb{R}^{n \times \sum_{\ell \in \mathcal{G}} d_\ell}$ to be the submatrix stacked with blocks whose indices are in $\mathcal{G}$. We also abbreviate $\boldsymbol{X}_{\bigcup_{\ell \in \mathcal{G}} \mathcal{A}_\ell}$ as $\boldsymbol{X}_{\mathcal{G}}$ whenever it does not cause any confusion. The j-th column of the coefficient matrix $\boldsymbol{B}$ in (2.2) is denoted by $\boldsymbol{b}_{\cdot j}$. For each j, let $\mathcal{G}_j \equiv \{1 \leq \ell \leq q : z_{\ell j} \neq 0\}$ be the set of group indices corresponding to the groups from Gaussian components. The groups with indices in its complement $\overline{\mathcal{G}}_j$ are zeros. In this sense $g_j = |\mathcal{G}_j|$ measures the group sparsity of $\boldsymbol{b}_{\cdot j}$. Let $\mathcal{S}_j \equiv supp(\boldsymbol{b}_{\cdot j})$, and $s_j = |\mathcal{S}_j| = \|\boldsymbol{b}_{\cdot j}\|_0$ measures the sparsity of $\boldsymbol{b}_{\cdot j}$ at the coordinate level. Let $\tilde{s}_j = \sum_{\ell \in \mathcal{G}_j} d_\ell$ be the sum of dimensions of nonzero groups in $\boldsymbol{b}_{\cdot j}$, and $\tilde{s} = \max_j \tilde{s}_j$. Clearly, $s_j \leq \tilde{s}_j \leq \tilde{s}$. For any $\mathcal{A} \subset \{1, \ldots, d\}$, define

$$\rho_-(\mathcal{A}) = \inf\left\{\frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta}\|^2/\|\boldsymbol{\beta}\|^2 : supp(\boldsymbol{\beta}) \subset \mathcal{A}\right\},$$

$$\rho_+(\mathcal{A}) = \sup\left\{\frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta}\|^2/\|\boldsymbol{\beta}\|^2 : supp(\boldsymbol{\beta}) \subset \mathcal{A}\right\}.$$

Further define for any $1 \leq s \leq d$,

$$\rho_-(s) = \inf\left\{\rho_-\Big(\bigcup_{\ell \in \mathcal{G}} \mathcal{A}_\ell\Big) : \mathcal{G} \subset \{1, \ldots, q\}, \sum_{\ell \in \mathcal{G}} d_\ell \leq s\right\},$$

$$\rho_+(s) = \inf\left\{\rho_+\Big(\bigcup_{\ell \in \mathcal{G}} \mathcal{A}_\ell\Big) : \mathcal{G} \subset \{1, \ldots, q\}, \sum_{\ell \in \mathcal{G}} d_\ell \leq s\right\}.$$

They are the maximal and minimal eigenvalues of the sub gram matrices with groups whose dimensions sum up to some dimension s. We denote $\mathcal{G}_j^t$, $\mathcal{S}_j^t$, $\boldsymbol{b}_{\ell j}^t$, $z_{\ell j}^t$, $\boldsymbol{\mu}_{\ell 1}^t$, $\boldsymbol{\mu}_{\ell 2}^t$ as the t-th iterates (estimates) for $\mathcal{G}_j$, $\mathcal{S}_j$, $\boldsymbol{b}_{\ell j}$, $z_{\ell j}$, $\boldsymbol{\mu}_{\ell 1}$, $\boldsymbol{\mu}_{\ell 2}$, respectively.

Throughout the paper, we assume $C_* d_0 \leq \min_\ell d_\ell \leq \max_\ell d_\ell \leq C^* d_0$ for some absolute constants $0 < C_* < C^*$ (even group sizes). Note the "even group sizes" assumption is not necessary for our proofs. It is just for simpler notations and readers' conveniences. We also assume $\rho_* \leq \min_{1 \leq \ell \leq q} \rho_-(\mathcal{A}_\ell) \leq \max_{1 \leq \ell \leq q} \rho_+(\mathcal{A}_\ell) \leq \rho^*$ for some constants $0 < \rho_* \leq \rho^*$. In fact, we could impose $\rho_* = \rho^* = 1$ as orthonormalization can be performed within each block $\boldsymbol{X}_\ell$. Since we are selecting at the group level, the representation scheme within each group will does not affect subsequent steps.

For any two real series $\{a_n\}_n$ and $\{b_n\}_n$, we write $a_n = O(b_n)$ if $\limsup_{n \to \infty} \frac{a_n}{b_n} < \infty$, $a_n = o(b_n)$ if $\limsup_{n \to \infty} \frac{a_n}{b_n} = 0$, $a_n = \Omega(b_n)$ if $\limsup_{n \to \infty} \frac{a_n}{b_n} > 0$, and $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$.

Below we list some technical assumptions which are required in the proofs.

- Assumption 1 (separation of the mixture components): for each $1 \leq \ell \leq q$ such that $w_{\ell 1}, w_{\ell 2} > 0$, we have $\min\{\|\boldsymbol{\mu}_{\ell 1}\|, \|\boldsymbol{\mu}_{\ell 2}\|\} \geq M_1$ and $\|\boldsymbol{\mu}_{\ell 1} - \boldsymbol{\mu}_{\ell 2}\| \geq M_2$ for some constants $M_1, M_2 > 0$. It essentials requires that the mixture components are separated to some extent. If they are almost overlapping, it is unlikely to achieve good clustering results.

- Assumption 2 (sparsity at the group level and non-sparsity within each nonzero group, sparsity at the coordinate level):
  $\max_{1\le j\le p}|\mathcal{G}_j| \le g \ll q$ and $\min\{\|\boldsymbol{\mu}_{\ell 1}\|_0, \|\boldsymbol{\mu}_{\ell 2}\|_0\} \approx d_\ell$ for $\ell \in \mathcal{G}_j$ and $1 \le j \le p$. It essentially means there are only a few nonzero groups, and within each nonzero group the coefficients are not sparse. It also implies that each $\boldsymbol{b}_{\cdot j}$ is sparse at the coordinate level, *i.e.*, $\|\boldsymbol{b}_{\cdot j}\|_0 = s_j \le s$ for all $1 \le j \le p$. In summary, we have $s_j \asymp \tilde{s}_j$, $s_j \le s$, $\tilde{s}_j \le \tilde{s}$, and $s \asymp \tilde{s}$.

- Assumption 3 (little correlation between the noise and the covariates): there exist nonnegative constants a, b such that for any fixed $1 \le j \le p$, $1 \le \ell \le q$, and $\epsilon_1 \in (0,1)$, with probability at least $1-\epsilon_1$,
  $$\left\|(\boldsymbol{X}_\ell^\top \boldsymbol{X}_\ell)^{-1/2}\boldsymbol{X}_\ell^\top \boldsymbol{\varepsilon}_j\right\| \le a\sqrt{d_\ell} + b\sqrt{\log(1/\epsilon_1)}.$$

- Assumption 4 (group sparse eigenvalue condition): the design matrix $\boldsymbol{X}$ satisfies
  $$\frac{\rho_+(\tilde{s}) - \rho_-(2\tilde{s})}{\rho_-(\tilde{s})} \le \kappa$$
  for some constant $\kappa > 0$.

- Assumption 5 (sparse Riesz Condition): the design $\boldsymbol{X}$ satisfies for some constants $0 < c_* \le c^* < \infty$ and rank s_0 such that $s < s_0 \ll d$,
  $$c_* \le \min_{|\mathcal{A}|\le s_0} \rho_-(\mathcal{A}) \le \max_{|\mathcal{A}|\le s_0} \rho_+(\mathcal{A}) \le c^*.$$

- Assumption 6 (high-dimensional setting and sample size requirements): d, d_0, q, g, $\tilde{s}$, p are allowed to go to ∞ as $n \to \infty$. Particularly, their magnitudes satisfy $n \ll d$, $g \ll q$, $g \ll p$, $d = O(d_0 \cdot q)$, $\tilde{s} = O(d_0 \cdot g)$, $d_0 \ll p$, $(\log p + \log g)^2 \ll d_0$. The sample size satisfies $n \gg \max\left\{\tilde{s} + g\log(qp/\epsilon_1), s\log(pd/\epsilon_2)\right\}$ for fixed small ϵ_1 and ϵ_2.

We also list some technical lemmas which we will use in subsequent proofs. These lemmas are established in existing literature.

Lemma 1. *(Huang and Zhang, 2010) Let $\boldsymbol{v} \in \mathbb{R}^d$ be a vector of i.i.d. standard Gaussian variables: $v_i \sim N(0,1)$. Then $\forall \delta > 0$,*
$$\mathbb{P}\left\{\left|\|\boldsymbol{v}\| - \sqrt{d}\right| \ge \delta\right\} \le 3e^{-\delta^2/2}.$$

Lemma 1 essentially states that with high probability, high-dimensional isotropic Gaussian vectors are behaving like spherical shell of radius $\sqrt{d_\ell}$ and thickness $d_\ell^{1/4}$ around their centers, *i.e.*, $\left|\|\boldsymbol{v}\| - O(\sqrt{d})\right| = O(d^{1/4})$.

Lemma 2. *(Zhang, 2010) Define seminorms*

$$\zeta(\boldsymbol{v}, m, \mathcal{A}) \equiv \max\left\{ \|(\boldsymbol{P}_{\mathcal{A}'} - \boldsymbol{P}_{\mathcal{A}})\boldsymbol{v}\| \Big/ \Big(nm\Big)^{1/2} : \mathcal{A} \subset \mathcal{A}' \subset \{1, \ldots, d\}, |\mathcal{A}'| = m + |\mathcal{A}| \right\}$$

for $\boldsymbol{v} \in \mathbb{R}^n$, where $\boldsymbol{P}_{\mathcal{A}}$ is the orthogonal projection from $\mathbb{R}^n$ to the span of the columns of $\boldsymbol{X}_{\mathcal{A}}$. Let $\tilde{d}_{\epsilon_2} \geq \sqrt{e}$ be the solution of (4.2) with $|\mathcal{A}| = s$. Suppose $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$. Then

$$\mathbb{P}\left\{ \zeta(\boldsymbol{v}, m, \mathcal{A}) \geq \sigma\sqrt{2\log \tilde{d}_{\epsilon_2}/n} \right\} \leq \frac{\epsilon_2}{2p\sqrt{\log \tilde{d}_{\epsilon_2}}}.$$

Lemma 3. *(Zhang, 2010) Suppose Assumption 6 holds with $c^* \geq c_* \geq \kappa \geq 0$. Let $\boldsymbol{p}(t; \lambda)$ be a penalty function satisfying $\lambda(1 - \kappa t/\lambda)_+ \leq \dot{\boldsymbol{p}}(t; \lambda) \leq \lambda$ for all $t > 0$. Let $\widehat{\boldsymbol{\beta}}$ be the corresponding penalized least square estimator and $\widehat{\boldsymbol{\beta}}^{orac}$ be the oracular least square estimator with the true support known. If $\lambda \geq 2\sqrt{c^*}\zeta(\boldsymbol{y}, s_0 - s, \mathcal{S})$ with $|\mathcal{S}| = s$, then*

$$c_* \|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{orac}\| \leq \sqrt{\sum_{h \in \mathcal{S}} \Big(\dot{\boldsymbol{p}}(|\widehat{\beta}_h|; \lambda)\Big)^2} + 2\sqrt{c_*/c^*}\sqrt{s_0}\lambda.$$

Lemma 4. *(Cai et al., 2019) For two vectors $\boldsymbol{v}$ and $\hat{\boldsymbol{v}}$, if $\|\boldsymbol{v} - \hat{\boldsymbol{v}}\| \leq \|\boldsymbol{v}\|$ and $\|\boldsymbol{v}\| \geq C$ for some constant $C > 0$, then*

$$\boldsymbol{v}^\top \hat{\boldsymbol{v}} - \|\boldsymbol{v}\| \cdot \|\hat{\boldsymbol{v}}\| \asymp \|\boldsymbol{v} - \hat{\boldsymbol{v}}\|^2.$$

Now we are ready to state our theoretical results. Propositions 1 and 2, and Theorem 1 together address the goodness of the starting point $\Big(\boldsymbol{B}^0, \{\boldsymbol{\mu}_{\ell 1}^0, \boldsymbol{\mu}_{\ell 2}^0\}_\ell, \{z_{\ell j}^0\}_{\ell j}\Big)$. Proposition 1 addresses the selection sensitivity of group Lasso in Step 1; Proposition 2 guarantees the selection consistency and estimation accuracy of MC+ in Step 1; Theorem 1 bounds the error rate for estimating the Gaussian centers and the misclassification error.

Proposition 1. *Suppose that Assumptions 1-5 are valid. Further assume that the group Lasso penalty $\lambda_\ell^0 \geq 4\sqrt{\rho^*}\left[\Big(\frac{a}{\sqrt{n}} + \frac{a'}{\sqrt{q}}\Big)\sqrt{d_\ell} + \Big(b\sqrt{\frac{\log(qp/\epsilon_1)}{n}} + \frac{b'}{\sqrt{q}}\Big)\right]$ with (a, a', b, b') being absolute constants, then with probability at least $1 - \epsilon_1 - o(1)$, all the groups from Gaussian components $\{\boldsymbol{b}_{\ell j} : z_{\ell j} \neq 0\}$ are selected by group Lasso in Step 1.*

Proof of Proposition 1. Let $\widehat{\boldsymbol{b}}_{\cdot j}^{gL}$ be the group Lasso estimator for $\boldsymbol{b}_{\cdot j}$. By Assumptions 2-4, the conditions for Theorem 5.1 in (Huang and Zhang, 2010) are verified, and hence we have with probability at least $1 - \epsilon_1$,

$$\begin{aligned}\sup_{1 \leq j \leq p} \|\widehat{\boldsymbol{b}}_{\cdot j}^{gL} - \boldsymbol{b}_{\cdot j}\|^2 &\leq \frac{72\rho^*(1 + 0.25/\kappa)^2}{\Big(\rho_-(2\tilde{s})\Big)^2 n}\Big(a^2\tilde{s} + b^2 \cdot g\log(qp/\epsilon_1)\Big) \\ &= O\bigg(\Big(\tilde{s} + g\log(qp/\epsilon_1)\Big)/n\bigg). \end{aligned} \tag{4.1}$$

Since $n \gg \tilde{s}+g\log(qp/\epsilon_1)$ by Assumption 5, it implies that any group with $\|\boldsymbol{b}_{\ell j}\|^2 = \Omega(1)$ will be selected by group Lasso. Otherwise, for this particular j, $\|\widehat{\boldsymbol{b}}_{\cdot j}^{gL} - \boldsymbol{b}_{\cdot j}\|^2 \geq \|\boldsymbol{b}_{\ell j}\|^2 = \Omega(1)$, contradict to (4.1). Now for any fixed pair (ℓ, j) such that $z_{\ell j} = 1$, by Lemma 1, with probability at least $1 - 3e^{-\sqrt{d_\ell}/2}$,

$$\left| \|\boldsymbol{b}_{\ell j} - \boldsymbol{\mu}_{\ell 1}\| - \sqrt{d_\ell} \right| \leq d_\ell^{1/4}.$$

It further implies that with probability at least $1 - 3e^{-\sqrt{d_\ell}/2}$,

$$\max\{\|\boldsymbol{\mu}_{\ell 1}\| - \sqrt{d_\ell} - d_\ell^{1/4},\ -\|\boldsymbol{\mu}_{\ell 1}\| + \sqrt{d_\ell} - d_\ell^{1/4}\} \leq \|\boldsymbol{b}_{\ell j}\| \leq \|\boldsymbol{\mu}_{\ell 1}\| + \sqrt{d_\ell} + d_\ell^{1/4}.$$

As long as $\|\boldsymbol{\mu}_{\ell 1}\| \neq \sqrt{d_\ell}$, we have $\|\boldsymbol{b}_{\ell j}\| \gg \Omega(1)$ with dominating probability. The same holds for $z_{\ell j} = 2$. Then, we have

$$\sum_{(\ell,j):z_{\ell j}\neq 0} \mathbb{P}\{\|\boldsymbol{b}_{\ell j}\| = o(1)\} \leq 3gp \cdot e^{-\sqrt{d_\ell}/2}.$$

Note $3gp \cdot e^{-\sqrt{d_\ell}/2} = o(1)$ under Assumption 5. Therefore, $\{\boldsymbol{b}_{\ell j} : z_{\ell j} \neq 0\}$ will all be selected by group Lasso in Step 1 with probability at least $1 - \epsilon_1 - 3gp \cdot e^{-\sqrt{d_\ell}/2} = 1 - \epsilon_1 - o(1)$. □

After the initial selection by group Lasso, we have included all the true nonzero groups of coefficients with overwhelming probability. Next our procedure runs MC+ to fit $\boldsymbol{y}_j$ against the selected groups for every $1 \leq j \leq p$. This round of selection by MC+ would filter out the falsely selected groups and the zero coordinates inside the nonzero groups. Furthermore, MC+ would produce accurate estimates of $\{\boldsymbol{b}_{\ell j}\}_{\ell j}$. We state Proposition 2 below to characterize the selection and estimation accuracy of MC+ in Step 1.

Proposition 2. *Suppose Assumption 6 holds and $\|\boldsymbol{X}_{\cdot j}\| = n$ (normalized covariates). Further assume that $\max_{1\leq j\leq p} s_j \leq s \equiv \frac{s_0}{c^*/c_* + 1/2}$, and the the largest diagonal value of $\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{S}_j}^\top \boldsymbol{X}_{\mathcal{S}_j}\right)^{-1}$ is bounded by w_j for $1 \leq j \leq p$. For a small fixed $\epsilon_2 > 0$, let $\tilde{d}_{\epsilon_2}$ be such that*

$$2\log\tilde{d}_{\epsilon_2} - 1 - \log(2\log\tilde{d}_{\epsilon_2}) = \frac{2}{s_0 - s}\left[\log\binom{d-s}{s_0-s} + \log(p/\epsilon_2)\right]. \tag{4.2}$$

(i) *Define $\widehat{\boldsymbol{b}}_{\cdot j}^{orac}$ to be the LSE with oracular knowledge about the true support,*

$$\widehat{\boldsymbol{b}}_{\cdot j}^{orac} \equiv \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y}_j - \boldsymbol{X}\boldsymbol{\beta}\|^2 : supp(\boldsymbol{\beta}) \subset \mathcal{S}_j\}.$$

Let $\lambda_{1,\epsilon_2} = \sigma\sqrt{2\log\left(p(d-s)/\epsilon_2\right)/n}$ and $\lambda_{2,\epsilon_2} = \max\{2\sqrt{c^*}\sigma\sqrt{2\log\tilde{d}_{\epsilon_2}/n},\ \lambda_{1,\epsilon_2}\}$. If for all $1 \leq j \leq p$

$$\min_{h\in\mathcal{S}_j}|b_{hj}| \geq \sigma\sqrt{(2/c_*)\log(ps/\epsilon_2)/n} + \gamma\lambda_{2,\epsilon_2}$$

with γ being the regularization parameter in the minimax concave penalty, and the MC penalty satisfies $\lambda_{1,\epsilon_2} \leq \lambda^{MCP} \leq \lambda_{2,\epsilon_2}$, the MC+ estimators $\{\boldsymbol{b}^0_{\cdot j}\}_{j=1}^p$ would achieve

$$\sum_{1\leq j\leq p} \mathbb{P}\Big\{\boldsymbol{b}^0_{\cdot j} \neq \widehat{\boldsymbol{b}}^{orac}_{\cdot j} \text{ or } sgn(\boldsymbol{b}^0_{\cdot j}) \neq sgn(\widehat{\boldsymbol{b}}^{orac}_{\cdot j})\Big\} \leq (3/2 + 1/\sqrt{2})\epsilon_2. \tag{4.3}$$

(ii) *When $\lambda^{MCP} \geq 2\sqrt{c^*}\sigma\sqrt{2\log \tilde{d}_{\epsilon_2}/n}$, it holds with probability at least $1 - \epsilon_2\Big/(2\sqrt{\log \tilde{d}_{\epsilon_2}})$,*

$$\sup_{1\leq j\leq p} \|\boldsymbol{b}^0_{\cdot j} - \widehat{\boldsymbol{b}}^{orac}_{\cdot j}\| \leq \frac{3\sqrt{s}\lambda^{MCP}}{2c_*}. \tag{4.4}$$

Since $\|\widehat{\boldsymbol{b}}^{orac}_{\cdot j} - \boldsymbol{b}_{\cdot j}\| \asymp O(\sqrt{s/n}) = o(\sqrt{s}\lambda^{MCP})$, it follows with high probability,

$$\sup_{1\leq j\leq p} \|\boldsymbol{b}^0_{\cdot j} - \boldsymbol{b}_{\cdot j}\| \leq O(\sqrt{s\log \tilde{d}_{\epsilon_2}/n}).$$

Proof of Proposition 2. Note the MC+ fits $\boldsymbol{y}_j$ against only the groups of variables selected by group Lasso $\boldsymbol{X}_{\mathcal{G}^{gL}_j}$ for each $1 \leq j \leq p$. The property of MC+ is used to merely bound the estimation error on $\mathcal{G}^{gL}_j$ part, $\|\boldsymbol{b}^0_{\mathcal{G}^{gL}_j j} - \boldsymbol{b}_{\mathcal{G}^{gL}_j j}\|$. The estimation error outside $\mathcal{G}^{gL}_j$ is zero since $\mathcal{G}^{gL}_j$ covers the true group support $\mathcal{G}_j$ with high probability by Proposition 1. Therefore, hereinafter we do not distinguish between $\|\boldsymbol{b}^0_{\cdot j} - \boldsymbol{b}_{\cdot j}\|$ and $\|\boldsymbol{b}^0_{\mathcal{G}^{gL}_j j} - \boldsymbol{b}_{\mathcal{G}^{gL}_j j}\|$ for simpler notations.

(i) (4.3) immediately follows from Theorem 1 in (Zhang, 2010) with adjustment for multiple linear models, *i.e.*, replacing the ϵ in (Zhang, 2010) by ϵ_2/p.

(ii) Since for each j, $\zeta(\boldsymbol{X}\boldsymbol{b}_{\cdot j}, s_0 - s, \mathcal{S}_j) \leq \|(\boldsymbol{I}_n - \boldsymbol{P}_{\mathcal{S}_j})\boldsymbol{X}\boldsymbol{b}_{\cdot j}\|/\sqrt{n(s_0 - s)} = 0$, we have $\zeta(\boldsymbol{y}_j, s_0 - s, \mathcal{S}_j) \leq \zeta(\boldsymbol{\varepsilon}, s_0 - s, \mathcal{S}_j) + \zeta(\boldsymbol{X}\boldsymbol{b}_{\cdot j}, s_0 - s, \mathcal{S}_j) \leq \zeta(\boldsymbol{\varepsilon}, s_0 - s, \mathcal{S}_j)$. By Lemma 2, $\mathbb{P}\{2\sqrt{c^*}\zeta(\boldsymbol{y}_j, s_0 - s, \mathcal{S}_j) > \lambda^{MCP}\} \leq \frac{\epsilon_2}{2p\sqrt{\log \tilde{d}_{\epsilon_2}}}$. Further by the Lemma 3, when $2\sqrt{c^*}\zeta(\boldsymbol{y}_j, s_0 - s, \mathcal{S}_j) \leq \lambda^{MCP}$, it holds $\|\boldsymbol{b}^0_{\cdot j} - \widehat{\boldsymbol{b}}^{orac}_{\cdot j}\| \leq 3\sqrt{s}\lambda^{MCP}/(2c_*)$. Therefore,

$$\begin{aligned}\mathbb{P}\Big\{\sup_{1\leq j\leq p} \|\boldsymbol{b}^0_{\cdot j} - \widehat{\boldsymbol{b}}^{orac}_{\cdot j}\| > \frac{3\sqrt{s}\lambda^{MCP}}{2c_*}\Big\} &\leq \sum_{1\leq j\leq p} \mathbb{P}\Big\{\|\boldsymbol{b}^0_{\cdot j} - \widehat{\boldsymbol{b}}^{orac}_{\cdot j}\| > \frac{3\sqrt{s}\lambda^{MCP}}{2c_*}\Big\} \\ &\leq \epsilon_2\Big/(2\sqrt{\log \tilde{d}_{\epsilon_2}}).\end{aligned}$$

□

Remark 1. *The definition of $\tilde{d}_{\epsilon_2}$ seems elusive. Actually, it is bounded by*

$$\sigma\sqrt{2\log \tilde{d}_{\epsilon_2}/n} \leq \sigma\sqrt{2\Big(\log(pd) - \log(n\cdot R^2 s)\Big)\Big/n} + \epsilon_2\sigma/(p\sqrt{n})$$

if all the $\boldsymbol{b}_{\cdot j}$ are within a centered Euclidean ball with radius $R\sqrt{s}$. Interested readers are referred to the PROOF OF THEOREM 2 in (Zhang, 2010). Although we require the nonzero coordinates to be bounded below by $\max\left\{O(\sqrt{\log(ps)/n}), O(\sqrt{\log \tilde{d}_{\epsilon_2}/n})\right\}$ to achieve selection consistency, no such condition is needed to establish the estimation error rate (4.4). In fact, all we need is to get reasonably accurate estimates of $\{\boldsymbol{b}_{\ell j}\}_{\ell,j}$ so that subsequent steps can be carried out efficiently. Apparently, the error rate in (4.4) is minimax optimal in the setting of multiple high-dimensional linear models because in a single high-dimensional linear model the optimal rate is $O(\sqrt{s\log d/n})$. For readers' greater convenience, we do not give redundant proofs on the minimax lower bound.

After the two rounds of selection by group Lasso and MC+, we then implement EM algorithm on the estimated nonzero groups of coefficients. In conventional EM for estimating Gaussian mixture models, the sample set $\{\boldsymbol{b}_{\ell j}\}_j$ is observed. In our case, they are unknown and estimated, and then based on the estimates we fit Gaussian mixtures. This is a major challenge that existing literature has not addressed to the best of our knowledge. Ideally, we hope that subsequent estimations of the Gaussian centers $\{\boldsymbol{\mu}_{\ell 1}, \boldsymbol{\mu}_{\ell 2}\}_\ell$ and the labels $\{z_{\ell j}\}_{\ell,j}$ based on the estimates $\{\boldsymbol{b}^0_{\ell j}\}_{\ell,j}$ would achieve comparable or slightly worse efficiency than those based on the true sample set $\{\boldsymbol{b}_{\ell j}\}_{\ell,j}$. Obviously, accurate $\{\boldsymbol{b}^0_{\ell j}\}_{\ell,j}$ are necessary to achieve this goal. Proposition 1 and 2 above provide some clues on how accurate these estimates $\{\boldsymbol{b}^0_{\ell j}\}_{\ell,j}$ are.

Theorem 1 below provides some guarantees for the EM-based estimation of the Gaussian mixtures. Through EM we will obtain $\{w^0_{\ell 1}, w^0_{\ell 2}, \boldsymbol{\mu}^0_{\ell 1}, \boldsymbol{\mu}^0_{\ell 2}, z^0_{\ell 1}, \ldots, z^0_{\ell p}\}_{\ell j}$. Before stating Theorem 1, we introduce some notations. For each ℓ, define $\omega_\ell = w_{\ell 2}/(1-w_{\ell 0})$ and hence $1-\omega_\ell = w_{\ell 1}/(1-w_{\ell 0})$. Note the estimate $w^0_{\ell 0}$ is easily obtained by the output of Step 1, *i.e.*, $w^0_{\ell 0} = \#\{j : \boldsymbol{b}^0_{\ell j} = \boldsymbol{0}\}/p$. After obtaining $\{\omega^0_\ell\}_\ell$ in Step 2, we can calculate $w^0_{\ell 1} = (1-\omega^0_\ell)(1-w^0_{\ell 0})$ and $w^0_{\ell 2} = \omega^0_\ell(1-w^0_{\ell 0})$.

For simplicity and clarity, we carry out the following discussion in the context of two-component Gaussian mixture model. Let $\{c_0 < \omega < 1-c_0, \boldsymbol{\mu}_1 \in \mathbb{R}^{d_0}, \boldsymbol{\mu}_2 \in \mathbb{R}^{d_0}\}$ be the true weight and two centers, $\boldsymbol{b}$ be a general sample point to be clustered, and z be its true label. In the literature of clustering analysis, the optimal classification error rate is known to be

$$R_{opt} \equiv \mathbb{P}\{\hat{z} \neq z\} = (1-\omega)\Phi\left(\frac{\ln\tau - \|\boldsymbol{\delta}\|^2/2}{\|\boldsymbol{\delta}\|}\right) + \omega\left[1 - \Phi\left(\frac{\ln\tau + \|\boldsymbol{\delta}\|^2/2}{\|\boldsymbol{\delta}\|}\right)\right], \tag{4.5}$$

where Φ is the standard normal density, $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and $\tau = \omega/(1-\omega)$. In ideal scenarios where $\{\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$ are known, R_{opt} can be achieved by Fisher's linear discriminant rule. In our case, $\{\omega, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$ are unknown and estimated by EM. Based on $(\hat{\omega}, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2)$, the classification rule is

$$\hat{z} := \begin{cases} 1 & \text{if } (1-\hat{\omega})e^{-\|\boldsymbol{b}-\hat{\boldsymbol{\mu}}_1\|^2/2} \leq \hat{\omega}e^{-\|\boldsymbol{b}-\hat{\boldsymbol{\mu}}_2\|^2/2}, \\ 2 & \text{otherwise.} \end{cases}$$

It can be simplified to

$$\hat{z} := \begin{cases} 1 & \text{if } (\boldsymbol{b} - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\delta}} \leq \ln \hat{\tau}, \\ 2 & \qquad \text{otherwise,} \end{cases}$$

where $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$, $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$, and $\hat{\tau} = \hat{\omega}/(1 - \hat{\omega})$. The corresponding misclassification error rate is

$$R = (1-\omega)\Phi\left(\frac{\ln \hat{\tau} + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|}\right) + \omega\left[1 - \Phi\left(\frac{\ln \hat{\tau} + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|}\right)\right].$$

Theorem 1. *Define the event $\mathcal{E}^0 \equiv \{\mathcal{S}_j^0 = \mathcal{S}_j$ and $\boldsymbol{b}_{\cdot j}^0 = \widehat{\boldsymbol{b}}_{\cdot j}^{orac}$ for all $1 \leq j \leq p\}$. Suppose Assumption 1-6 hold, and the initialization of the EM procedure satisfies $\|\hat{\boldsymbol{\mu}}_{\ell 1}^{initial} - \boldsymbol{\mu}_{\ell 1}\| \leq M_2/4$ and $\|\hat{\boldsymbol{\mu}}_{\ell 2}^{initial} - \boldsymbol{\mu}_{\ell 2}\| \leq M_2/4$ with the same M_2 as in Assumption 1. Then the outputs of the EM procedure (Step 2 in Section 3) would achieve for some constant $C_1 > 0$,*

$$\mathbb{P}\left\{\sup_{\ell \in \mathcal{C}^0} \|\boldsymbol{\mu}_{\ell k}^0 - \boldsymbol{\mu}_{\ell k}\| \leq C_1\sqrt{d_\ell/p} \textit{ for } k = 1, 2 \middle| \mathcal{E}^0\right\} \geq 1 - \frac{g}{p}, \tag{4.6}$$

$$\mathbb{P}\left\{\sup_{\ell \in \mathcal{C}^0} \left(R(z_{\ell j}^0) - R_{opt}\right) \leq C_1 \frac{d_\ell}{p} \middle| \mathcal{E}^0\right\} \geq 1 - \frac{g}{p}, \tag{4.7}$$

provided that the EM iteration number is of order $O(\log(p/d_0))$.

Proof of Theorem 1. The argument below adopts some proof ideas and techniques from (Cai et al., 2019). We first prove (4.6). On the event $\mathcal{E}^0$, $\boldsymbol{b}_{\mathcal{S}_j j}^0 = \left(\boldsymbol{X}_{\mathcal{S}_j}^\top \boldsymbol{X}_{\mathcal{S}_j}\right)^{-1} \boldsymbol{X}_{\mathcal{S}_j}^\top \boldsymbol{y}_j$ for each j. By Assumption 6 and block matrix inversion formula, we have $\frac{\|\boldsymbol{b}_{\ell j}^0 - \boldsymbol{b}_{\ell j}\|^2}{\|\boldsymbol{b}_{\mathcal{S}_j j}^0 - \boldsymbol{b}_{\mathcal{S}_j j}\|^2} \propto \frac{d_\ell}{s_j}$. It is well-known that $\|\boldsymbol{b}_{\mathcal{S}_j j}^0 - \boldsymbol{b}_{\mathcal{S}_j j}\|^2 \leq C_2 s_j/n$ for some constant $C_2 > 0$ only depending on σ and $\boldsymbol{X}_{\mathcal{S}_j}$. Therefore, $\|\boldsymbol{b}_{\ell j}^0 - \boldsymbol{b}_{\ell j}\| \leq C_2'\sqrt{d_\ell/n}$ for all ℓ and j on the event $\mathcal{E}^0$. For each fixed ℓ and $k = 1, 2$, our stationary EM estimate for the Gaussian center would be

$$\begin{aligned} \boldsymbol{\mu}_{\ell k}^0 &= \left(1\Big/\sum_{j=1}^p I(\hat{z} = k)\right) \sum_{j=1}^p I(\hat{z} = k) \cdot \boldsymbol{b}_{\ell j}^0 \\ &= \left(1\Big/\sum_{j=1}^p I(\hat{z} = k)\right) \sum_{j=1}^p I(\hat{z} = k) \cdot (\boldsymbol{b}_{\ell j}^0 - \boldsymbol{b}_{\ell j}) + \left(1\Big/\sum_{j=1}^p I(\hat{z} = k)\right) \sum_{j=1}^p I(\hat{z} = k) \cdot \boldsymbol{b}_{\ell j}. \end{aligned}$$

So the estimation error can be written as

$$\begin{aligned} \|\boldsymbol{\mu}_{\ell k}^0 - \boldsymbol{\mu}_{\ell k}\| \leq &\left\| \left(1\Big/\sum_{j=1}^p I(\hat{z} = k)\right) \sum_{j=1}^p I(\hat{z} = k) \cdot (\boldsymbol{b}_{\ell j}^0 - \boldsymbol{b}_{\ell j}) \right\| + \\ &\left\| \left(1\Big/\sum_{j=1}^p I(\hat{z} = k)\right) \sum_{j=1}^p I(\hat{z} = k) \cdot \boldsymbol{b}_{\ell j} - \boldsymbol{\mu}_{\ell k} \right\|. \end{aligned}$$

The first term above is bounded by $C_2'\sqrt{d_\ell/n}$ on the event $\mathcal{E}_0$. The second term is equivalent to the estimation error of the standard EM, *i.e.*, implementing EM on the true set of samples $\{\boldsymbol{b}_{\ell j}\}$. Directly applying Theorem 4.1 in (Cai et al., 2019), the second term is bounded by $C_2''\sqrt{d_\ell/p}$ with probability at least $1-1/p$. Combining the bounds for both terms, we get $\mathbb{P}\Big\{\|\boldsymbol{\mu}^0_{\ell k}-\boldsymbol{\mu}_{\ell k}\| \leq C_1\sqrt{d_\ell/p}\ for\ k=1,2\Big|\mathcal{E}^0\Big\} \geq 1-1/p$ for any fixed ℓ. Applying a union bound leads to (4.6).

We then show (4.7). To bound the difference between $R(z^0_{\ell j})$ and R_{opt}, we first connect them with an intermediate quantity defined below

$$R_{int} \equiv (1-\omega)\Phi\left(\frac{\ln\tau - \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|}\right) + \omega\left[1-\Phi\left(\frac{\ln\tau + \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|}\right)\right].$$

We first show that $R_{int} - R_{opt} \lesssim d_\ell/p$. Writing the two terms in R_{int} in Taylor expansion at $\ln\tau/\|\boldsymbol{\delta}\| - \|\boldsymbol{\tau}\|/2$ and $\ln\tau/\|\boldsymbol{\delta}\| + \|\boldsymbol{\tau}\|/2$, respectively, leads to

$$\begin{aligned} R_{int} - R_{opt} =&(1-\omega)\left(\frac{\ln\tau}{\|\hat{\boldsymbol{\delta}}\|} - \frac{\ln\tau}{\|\boldsymbol{\delta}\|} + \frac{\|\boldsymbol{\delta}\|}{2} - \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|}\right)\cdot\Phi'\left(\frac{\ln\tau}{\|\boldsymbol{\delta}\|} - \frac{\|\boldsymbol{\delta}\|}{2}\right) \\ &- \omega\left(\frac{\ln\tau}{\|\hat{\boldsymbol{\delta}}\|} - \frac{\ln\tau}{\|\boldsymbol{\delta}\|} + \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|} - \frac{\|\boldsymbol{\delta}\|}{2}\right)\cdot\Phi'\left(\frac{\ln\tau}{\|\boldsymbol{\delta}\|} + \frac{\|\boldsymbol{\delta}\|}{2}\right) + O_p(d_\ell/p). \end{aligned} \tag{4.8}$$

The last term $O_p(d_\ell/p)$ is derived from the fact that $\Phi'' = O(1)$ and

$$\begin{aligned} \left|\frac{\ln\tau}{\|\hat{\boldsymbol{\delta}}\|} - \frac{\ln\tau}{\|\boldsymbol{\delta}\|} + \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|} - \frac{\|\boldsymbol{\delta}\|}{2}\right| &\leq \left|\frac{\ln\tau}{\|\hat{\boldsymbol{\delta}}\|} - \frac{\ln\tau}{\|\boldsymbol{\delta}\|}\right| + \left|\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|} - \frac{\|\boldsymbol{\delta}\|}{2}\right| \\ &\leq \left|\frac{\ln\tau}{\|\hat{\boldsymbol{\delta}}\|} - \frac{\ln\tau}{\|\boldsymbol{\delta}\|}\right| + \left|\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|} - \frac{\|\boldsymbol{\delta}\|^2}{2\|\hat{\boldsymbol{\delta}}\|}\right| + \left|\frac{\|\boldsymbol{\delta}\|^2}{2\|\hat{\boldsymbol{\delta}}\|} - \frac{\|\boldsymbol{\delta}\|}{2}\right| \\ &\lesssim \left|\|\hat{\boldsymbol{\delta}}\| - \|\boldsymbol{\delta}\|\right| + \left|\boldsymbol{\delta}^\top(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\right| \\ &\lesssim \|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\| \\ &\lesssim \sqrt{d_\ell/p}. \end{aligned}$$

We then re-write (4.8) to obtain

$$
\begin{aligned}
\frac{R_{int} - R_{opt}}{\sqrt{(1-\omega)\omega}} \asymp & \left(\frac{\ln\tau}{\|\hat{\boldsymbol{\delta}}\|} - \frac{\ln\tau}{\|\boldsymbol{\delta}\|} + \frac{\|\boldsymbol{\delta}\|}{2} - \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|}\right) e^{-\frac{1}{2}(\frac{\ln\tau}{\|\boldsymbol{\delta}\|} - \frac{\|\boldsymbol{\delta}\|}{2})^2 - \frac{\ln\tau}{2}} \\
& - \left(\frac{\ln\tau}{\|\hat{\boldsymbol{\delta}}\|} - \frac{\ln\tau}{\|\boldsymbol{\delta}\|} + \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|} - \frac{\|\boldsymbol{\delta}\|}{2}\right) e^{-\frac{1}{2}(\frac{\ln\tau}{\|\boldsymbol{\delta}\|} + \frac{\|\boldsymbol{\delta}\|}{2})^2 + \frac{\ln\tau}{2}} \\
& = e^{-\frac{(\ln\tau)^2}{2\|\boldsymbol{\delta}\|^2} - \frac{\|\boldsymbol{\delta}\|^2}{8}} \cdot \left(\|\boldsymbol{\delta}\| - \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|}\right) \\
& \lesssim \left|\|\boldsymbol{\delta}\| - \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|}\right|.
\end{aligned}
$$

By Assumption 1 and Lemma 4,

$$
\left|\|\boldsymbol{\delta}\| - \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|}\right| = \left|\frac{\|\boldsymbol{\delta}\|\|\hat{\boldsymbol{\delta}}\| - \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|}\right| \lesssim \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\|^2.
$$

So we obtain $R_{int} - R_{opt} \lesssim \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\|^2$. We then bound $R(z^0_{\ell j}) - R_{int}$. The Taylor series of $R(z^0_{\ell j})$ is

$$
\begin{aligned}
& R(z^0_{\ell j}) = \\
& (1-\omega)\left\{\Phi\left(\frac{\ln\tau - \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|}\right) + \frac{\ln\hat{\tau} - \ln\tau + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top\hat{\boldsymbol{\delta}} + \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|} \cdot \Phi'\left(\frac{\ln\tau - \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|}\right) + O_p(d_\ell/p)\right\} \\
& + \omega\left\{1 - \Phi\left(\frac{\ln\tau + \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|}\right) - \frac{\ln\hat{\tau} - \ln\tau + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|} \cdot \Phi'\left(\frac{\ln\tau + \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|}\right) + O_p(d_\ell/p)\right\}.
\end{aligned}
$$

This implies

$$\begin{aligned}
\frac{R_{int} - R(z^0_{\ell j})}{\sqrt{(1-\omega)\omega}} \lesssim & \sqrt{(1-\omega)/\omega} \cdot \frac{\ln\tau - \ln\hat{\tau} - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^\top \hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|} \cdot \Phi'\Big(\frac{\ln\tau - \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|}\Big) \\
& - \sqrt{\omega/(1-\omega)} \cdot \frac{\ln\tau - \ln\hat{\tau} - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\delta}} + \boldsymbol{\delta}^\top \hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|} \cdot \Phi'\Big(\frac{\ln\tau + \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|}\Big) \\
= & \frac{\ln\tau - \ln\hat{\tau} - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^\top \hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|} \cdot e^{-\frac{1}{2}(\frac{\ln\tau - \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|})^2 - \frac{\ln\tau}{2}} \\
& - \frac{\ln\tau - \ln\hat{\tau} - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\delta}} + \boldsymbol{\delta}^\top \hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|} \cdot e^{-\frac{1}{2}(\frac{\ln\tau + \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|})^2 + \frac{\ln\tau}{2}} \\
\lesssim & \Big|\frac{\ln\tau - \ln\hat{\tau} - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^\top \hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|}\Big| \cdot \Big| e^{-\frac{(\ln\tau - \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2)^2}{2\|\hat{\boldsymbol{\delta}}\|^2} - \frac{\ln\tau}{2}} - e^{-\frac{(\ln\tau + \boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2)^2}{2\|\hat{\boldsymbol{\delta}}\|^2} + \frac{\ln\tau}{2}} \Big| \\
= & \Big|\frac{\ln\tau - \ln\hat{\tau} - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^\top \hat{\boldsymbol{\delta}}/2}{\|\hat{\boldsymbol{\delta}}\|}\Big| \cdot e^{-\frac{(\ln\tau)^2 + (\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}/2)^2}{2\|\hat{\boldsymbol{\delta}}\|^2}} \cdot \Big| e^{\frac{\ln\tau\cdot\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|^2} - \frac{\ln\tau}{2}} - e^{-\frac{\ln\tau\cdot\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|^2} + \frac{\ln\tau}{2}} \Big| \\
\lesssim & \sqrt{d_\ell/p} \cdot O_p(1) \cdot \sqrt{d_\ell/p} \\
\lesssim & d_\ell/p.
\end{aligned}$$

The derivation of the last inequality uses the fact that $\Big|e^a - e^{-a}\Big| \lesssim a$ when $a = o(1)$, so that

$$\Big| e^{\frac{\ln\tau\cdot\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|^2} - \frac{\ln\tau}{2}} - e^{-\frac{\ln\tau\cdot\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{2\|\hat{\boldsymbol{\delta}}\|^2} + \frac{\ln\tau}{2}} \Big| \lesssim \Big| \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\delta}}}{\|\hat{\boldsymbol{\delta}}\|^2} - 1 \Big| \lesssim \sqrt{d_\ell/p}.$$

Finally, $R(z^0_{\ell j}) - R_{opt} \leq R(z^0_{\ell j}) - R_{int} + R_{int} - R_{opt} \lesssim d_\ell/p$. □

Next we explore the convergence property of the algorithm proposed in Section 3. Obviously the minimizer of (**??**) is well-defined and attainable as it is a strictly convex optimization problem. The definiteness and numerical convergence of Step 3b have been established by Proposition 1 in (Meier et al., 2008). Theorem 2 below shows that the algorithm is monotonically improving.

Theorem 2. *Through Step 3a, the on-support part of the estimate* $\boldsymbol{b}^{t+1}_{\mathcal{G}^t_j j}$ *is improving given* $\mathcal{G}^t_j$ *for all* j*; Step 3b and 3c would improve the support estimate* $\mathcal{G}^{t+1}_j$ *for all* j*; Step 4 monotonically improves* $\boldsymbol{\theta}^t$ *given* $\boldsymbol{B}^t$*. In summary, along the iterations between Step 3 and Step 4, the whole system* $(\boldsymbol{B}^t, \boldsymbol{\theta}^t)$ *is monotonically improving.*

Proof of Theorem 2. In Step 3a, the objective function (**??**) is actually the conditional expectation of (3.3) with respect to the conditional distribution of $\{z_{\ell j}\}_\ell$ given $\boldsymbol{b}_{\cdot j}$ under current $\boldsymbol{\theta}^t$ and $\mathcal{G}^t_j$, which is formulated in (3.4). The maximizer of (3.4) will be our $\boldsymbol{b}^{t+1}_{\mathcal{G}^t_j j}$. For simplicity, we do not distinguish $\boldsymbol{b}^t_{\mathcal{G}^t_j j}$ and $\boldsymbol{b}^t_{\cdot j}$ in the following analysis because the part of

$\boldsymbol{b}^t_{\cdot j}$ outside $\mathcal{G}^t_j$ is just $\mathbf{0}$. But readers should be aware that (4.9), (4.10), and (4.12) are only addressing the update on $\boldsymbol{b}_{\mathcal{G}^t_j j}$.

The second term in (3.4) can be decomposed as

$$\begin{aligned}\sum_{\boldsymbol{z}\in\mathcal{Z}} g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j}) \log g_{\boldsymbol{\theta}^t}(\boldsymbol{z},\boldsymbol{b}_{\cdot j})\, d\boldsymbol{z} &= \sum_{\boldsymbol{z}\in\mathcal{Z}} g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j}) \log \frac{g_{\boldsymbol{\theta}^t}(\boldsymbol{z},\boldsymbol{b}_{\cdot j})}{g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})}\, d\boldsymbol{z} - \sum_{\boldsymbol{z}\in\mathcal{Z}} g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j}) \log \frac{1}{g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})}\, d\boldsymbol{z}\\ &= \sum_{\boldsymbol{z}\in\mathcal{Z}} g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j}) \log \frac{g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}_{\cdot j})\cdot g_{\boldsymbol{\theta}^t}(\boldsymbol{b}_{\cdot j})}{g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})}\, d\boldsymbol{z} - \mathcal{H}\Big(g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})\Big)\\ &= \log g_{\boldsymbol{\theta}^t}(\boldsymbol{b}_{\cdot j}) + \sum_{\boldsymbol{z}\in\mathcal{Z}} g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j}) \log \frac{g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}_{\cdot j})}{g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})}\, d\boldsymbol{z} - \mathcal{H}\Big(g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})\Big)\\ &= \log g_{\boldsymbol{\theta}^t}(\boldsymbol{b}_{\cdot j}) - D_{KL}\Big(g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})\Big|\Big|g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}_{\cdot j})\Big) - \mathcal{H}\Big(g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})\Big),\end{aligned} \tag{4.9}$$

where $D_{KL}(\cdot||\cdot)$ is the Kullback–Leibler (KL) divergence between two probability measures, and $\mathcal{H}(p(\cdot))$ is the entropy of the probability measure $p(\cdot)$. Plugging (4.9) into (3.4), we obtain

$$\log p_{\boldsymbol{\theta}^t}(\boldsymbol{y}_j,\boldsymbol{b}_{\cdot j}) = Q_{\boldsymbol{\theta}^t}(\boldsymbol{b}_{\cdot j}\big|\boldsymbol{b}^t_{\cdot j}) + D_{KL}\Big(g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})\Big|\Big|g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}_{\cdot j})\Big) + \mathcal{H}\Big(g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})\Big). \tag{4.10}$$

Since (4.10) holds for any $\boldsymbol{b}_{\cdot j}$, we have

$$\log p_{\boldsymbol{\theta}^t}(\boldsymbol{y}_j,\boldsymbol{b}^t_{\cdot j}) = Q_{\boldsymbol{\theta}^t}(\boldsymbol{b}^t_{\cdot j}\big|\boldsymbol{b}^t_{\cdot j}) + D_{KL}\Big(g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})\Big|\Big|g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})\Big) + \mathcal{H}\Big(g_{\boldsymbol{\theta}^t}(\boldsymbol{z}\big|\boldsymbol{b}^t_{\cdot j})\Big). \tag{4.11}$$

Subtracting (4.10) by (4.11) results in

$$\log p_{\boldsymbol{\theta}^t}(\boldsymbol{y}_j,\boldsymbol{b}_{\cdot j}) - \log p_{\boldsymbol{\theta}^t}(\boldsymbol{y}_j,\boldsymbol{b}^t_{\cdot j}) \geq Q_{\boldsymbol{\theta}^t}(\boldsymbol{b}_{\cdot j}\big|\boldsymbol{b}^t_{\cdot j}) - Q_{\boldsymbol{\theta}^t}(\boldsymbol{b}^t_{\cdot j}\big|\boldsymbol{b}^t_{\cdot j}) \tag{4.12}$$

because $D_{KL}(\cdot\big|\big|\cdot)$ is always non-negative. It essentially means that in every repeat of Step 3a the increment in the log-likelihood is larger than the increment in $Q_{\boldsymbol{\theta}^t}(\cdot\big|\boldsymbol{b}^t_{\cdot j})$. While $Q_{\boldsymbol{\theta}^t}(\cdot\big|\boldsymbol{b}^t_{\cdot j})$ is monotone increasing because we are maximizing $Q_{\boldsymbol{\theta}^t}(\boldsymbol{b}_{\cdot j}\Big|\boldsymbol{b}^t_{\cdot j})$ over $\boldsymbol{b}_{\cdot j}$. Equality will be attained in (4.12) if and only if $\boldsymbol{b}_{\cdot j} = \boldsymbol{b}^t_{\cdot j}$, *i.e.*, the algorithm has converged. Therefore, $\boldsymbol{b}^{t+1}_{\mathcal{G}^t_j j}$ is improving in the sense that it gets closer to the "MLE" of the log likelihood $\log p_{\boldsymbol{\theta}^t}(\boldsymbol{y}_j,\boldsymbol{b}_{\mathcal{G}^t_j j})$.

In Step 3b and 3c, we are updating the support estimate $\mathcal{G}^{t+1}_j$ given $\boldsymbol{b}^{t+1}_j$. For any $\ell\in\mathcal{G}^t_j$, by the KKT condition of (**??**), we have

$$\boldsymbol{X}^\top_\ell(\boldsymbol{y}_j - \boldsymbol{X}_{-\ell}\boldsymbol{b}^{t+1}_{-\ell j}) = \boldsymbol{X}^\top_\ell\boldsymbol{X}_\ell\boldsymbol{b}^{t+1}_{\ell j} + \sigma^2\boldsymbol{b}^{t+1}_{\ell j} - \sigma^2\bar{\boldsymbol{\mu}}^t_\ell = (n+\sigma^2)\boldsymbol{b}^{t+1}_{\ell j} - \sigma^2\bar{\boldsymbol{\mu}}^t_\ell.$$

If for some $\ell \in \mathcal{G}_j^t$ such that $\frac{1}{n}\|\boldsymbol{X}_\ell^\top(\boldsymbol{y}_j - \boldsymbol{X}_{-\ell}\boldsymbol{b}_{-\ell j}^{t+1})\| \leq \lambda_\ell^{t+1}$ with $\lambda_\ell^{t+1} = O(\sqrt{\log(pq)/n} + \sqrt{d_\ell/n})$, it implies that

$$\|(1+\sigma^2/n)\boldsymbol{b}_{\ell j}^{t+1}\| \leq (\sigma^2/n)\|\bar{\boldsymbol{\mu}}_\ell^t\| + \lambda_\ell^{t+1} \leq O(\sqrt{\log(pq)/n} + \sqrt{d_\ell/n}).$$

Since $\|\boldsymbol{b}_{\ell j} - \boldsymbol{b}_{\ell j}^0\| \leq O(\sqrt{d_\ell \log(pd)/n})$ by Proposition 2,

$$\|\boldsymbol{b}_{\ell j}\| \leq \|\boldsymbol{b}_{\ell j}^{t+1}\| + \|\boldsymbol{b}_{\ell j} - \boldsymbol{b}_{\ell j}^{t+1}\| \leq O(\sqrt{d_\ell \log(pd)/n}) \to \mathbf{0}.$$

Therefore, if $\frac{1}{n}\|\boldsymbol{X}_\ell^\top(\boldsymbol{y}_j - \boldsymbol{X}_{-\ell}\boldsymbol{b}_{-\ell j}^{t+1})\| \leq \lambda_\ell^{t+1}$, probably $\boldsymbol{b}_{\ell j}$ is exactly zero or negligibly small. It is reasonable to assign $\boldsymbol{b}_{\ell j}^{t+1} = \mathbf{0}$. While if for some $\ell \in \overline{\mathcal{G}_j^t}$ such that $\boldsymbol{b}_{\ell j}^{t+1}$ is nonzero after optimizing (**??**), by Theorem 5.1 in (Huang and Zhang, 2010), with high probability, the true $\boldsymbol{b}_{\ell j} \neq \mathbf{0}$. Therefore, with high probability, the support estimate $\mathcal{G}_j^{t+1}$ is approaching to wards the true support.

In Step 4, we are updating $\boldsymbol{\theta}^t$ and the cluster labels based on the current $\boldsymbol{B}^t$. It actually maximizes the second term in (3.2). This is accomplished via running a conventional EM procedure. Let t_o denote the iteration number of the main framework, and t_i denote the iteration number of the embedded (conventional) EM procedure in Step 4. It is well-known that the second part in (3.2) is monotonically increasing along the conventional EM trajectory, *i.e.*, $\sum_{\ell=1}^q \sum_{j:\boldsymbol{b}_{\ell j}^{t_o} \neq \mathbf{0}} \log\left(\sum_{k=1}^{K_\ell} \omega_{\ell k}^{t_i} \Phi_{\ell k}^{t_i}(\boldsymbol{b}_{\ell j}^{t_o})\right)$ is monotonically increasing as $t_i \to \infty$ for a fixed t_o. By assigning $w_{\ell 0}^t = (1/p)\sum_{j=1}^p \mathbb{1}\{\boldsymbol{b}_{\ell j}^t \neq \mathbf{0}\}$ we also maximize (3.1). □

Remark 2. *Although an explicit convergence rate is not provided, we have shown that our method could achieve minimax rates for both group selection and clustering. Actually the starting point already achieve rate optimality for three-component mixtures thanks to group Lasso, MC+, and EM. Subsequent iterations would further improve the estimators probably in term of the constant factors.*

5 Simulation

In the numerical simulation, we use some genotypic data from Human Connectome Project (HCP) (Van Essen et al., 2013) as covariates and some synthetic responses ($\boldsymbol{Y}$). The HCP dataset contains the SNPs on all 22 autosomes of 1056 individuals (n=1056). In each simulation experiment, we randomly pick a chromosome, and use the SNPs on that chromosome, *e.g.*, Chromosome 1, as predecessor covariates. Since we are interested in the genetic associations on the gene level, the SNPs within a gene or a biologically meaningful genetic region (*e.g.*, an intergenic regulatory region) are grouped together. In this way, all the SNPs on Chromosome 1 are grouped into non-overlapping blocks based on their physical locations using ANNOVAR (Wang et al., 2010), and each block represents a certain gene or regulatory region. To circumvent the extremely high correlations among SNPs, we further perform dimension reduction within each block, and use the principle components (PCs) of each block form to form the design matrix $\boldsymbol{X}$. The submatrix $\boldsymbol{X}_\ell$ are the PCs of the ℓ-th gene or regulatory region, and represents the genotypic measurements of this gene or region.

The true coefficient matrix $\boldsymbol{B}$ or $\{\boldsymbol{b}_{\ell j}\}_{1\leq\ell\leq q,1\leq j\leq p}$ is drawn from mixture models in (**??**). For each $1 \leq \ell \leq q$, the ℓ-th row block $[\boldsymbol{b}_{\ell 1}, \cdots, \boldsymbol{b}_{\ell p}]$ has probability $1-\delta$ to be all $\mathbf{0}$'s, and has probability δ to be from a mixture model with both Gaussian components and $\mathbf{0}$'s. The value of δ depends on the sample size n, dimensionality d, total number of blocks q. For an ℓ such that $\boldsymbol{b}_{\ell j} \neq \mathbf{0}$ for som j, we randomly pick a subset $\mathcal{M}_\ell \subset \{1, \ldots, p\}$ with $|\mathcal{M}_\ell| \in [p/3, p/2]$ such that $\boldsymbol{b}_{\ell j} = \mathbf{0}$ for every $j \in \mathcal{M}_\ell$. The rest $\boldsymbol{b}_{\ell j}$'s are drawn from a Gaussian mixture with the number of components $K_\ell \in \{2, 3, 4, 5, 6\}$ and the covariances being $\boldsymbol{I}_{d_\ell}$. The elements of the Gaussian centers $[\boldsymbol{\mu}_{\ell 1}, \cdots, \boldsymbol{\mu}_{\ell K_\ell}]$ are ranging from 0 to 10. The error matrix $\boldsymbol{E}$ has *i.i.d.* $N(0, 1)$ entries. The parameters λ^0, α and λ^* are tuned using an independently identically generated dataset. Throughout the simulation, $n = 1054$, $p = 200$, $\sigma = 1$.

In the tables below, $MSE = \frac{1}{d\cdot p}\|\hat{\boldsymbol{B}} - \boldsymbol{B}\|_2^2$, *selection accuracy* = (*true positive* + *true negative*)/*total*, "wrong classification proportion" equals $\frac{1}{q\cdot p}\sum_{\ell=1}^{q}\sum_{j=1}^{p}\mathbb{1}\{\hat{z}_{\ell j} \neq z_{\ell j}\}$, "wrong cluster numbers" equals $\frac{1}{2}\sum_{\ell=1}^{q}\mathbb{1}\{\hat{K}_\ell \neq K_\ell\}$, and "Gaussian center MSE" equals $\frac{1}{d\cdot p}\sum_{\ell=1}^{q}\sum_{j=1}^{p}\|\hat{\boldsymbol{\mu}}_{\ell j} - \boldsymbol{\mu}_{\ell j}\|_2^2$.

Table 2: Experiment 1 on Chr22: $d = 2267$, $q = 341$, $\boldsymbol{\Sigma}_{\ell k} = 0.2\boldsymbol{I}$, $\delta q/n = 0.11$, $\delta d/n = 0.71$.

ite	MSE	accuracy	wrong classification	wrong cluster numbers	Gaussian center MSE
0	0.3325	0.9572	0.1345	0.3783	0.7519
1	0.0969	0.9872	0.0817	0.3226	0.2501
2	0.0116	0.9946	0.0221	0.1466	0.0258
3	0.00563	0.9960	0.0161	0.1144	0.00655
4	0.00481	0.9961	0.0173	0.0997	0.00514
5	0.00461	0.9961	0.0161	0.1026	0.00483

Table 3: Experiment 2 on Chr1: $d = 9962$, $q = 1386$, $\boldsymbol{\Sigma}_{\ell k} = 0.2\boldsymbol{I}$, $\delta q/n = 0.12$, $\delta d/n = 0.85$.

ite	MSE	accuracy	wrong classification	wrong cluster numbers	Gaussian center MSE
0	0.4995	0.9629	0.0502	0.2835	0.4549
1	0.1782	0.9556	0.0615	0.5657	0.1078
2	0.04276	0.9687	0.0427	0.5043	0.1004
3	0.03452	0.9745	0.0360	0.4459	0.1083
4	0.03017	0.9780	0.0349	0.4149	0.1050
5	0.02778	0.9796	0.0319	0.3961	0.1037
6	0.02650	0.9809	0.0290	0.3773	0.1023
7	0.02559	0.9822	0.0270	0.3557	0.1021
8	0.02462	0.9840	0.0232	0.3175	0.1021
9	0.02269	0.9856	0.0229	0.2958	0.1021
10	0.02084	0.9867	0.0200	0.2720	0.1021
11	0.01981	0.9881	0.0178	0.2330	0.1021
12	0.01857	0.9899	0.0165	0.2020	0.1021
13	0.01631	0.9918	0.0135	0.1522	0.1021
14	0.01294	0.9940	0.0111	0.1118	0.1021
15	0.01049	0.9953	0.0098	0.0945	0.1021
16	0.00829	0.9961	0.0091	0.0671	0.1021
17	0.00717	0.9968	0.0086	0.0556	0.1021
18	0.00579	0.9972	0.0083	0.0505	0.1021
19	0.00515	0.9972	0.0080	0.0505	0.1021

Table 4: Experiment 3 on Chr19: $d = 4398$, $q = 919$, $\boldsymbol{\Sigma}_{\ell k} = 0.2\boldsymbol{I}$, $\delta q/n = 0.12$, $\delta d/n = 0.58$.

ite	MSE	accuracy	wrong classification	wrong cluster numbers	Gaussian center MSE
0	0.1827	0.9697	0.0596	0.2361	0.4368
1	0.0677	0.9842	0.0399	0.2361	0.1395
2	0.00557	0.9951	0.0132	0.0827	0.01185
3	0.00296	0.9978	0.0082	0.0501	0.00301
4	0.00242	0.9981	0.0057	0.0490	0.00258
5	0.00235	0.9981	0.0055	0.0479	0.00251

Table 5: Experiment 4 on Chr6: $d = 6523$, $q = 665$, $\boldsymbol{\Sigma}_{\ell k} = 0.2\boldsymbol{I}$, $\delta q/n = 0.082$, $\delta d/n = 0.80$.

ite	MSE	accuracy	wrong classification	wrong cluster numbers	Gaussian center MSE
0	0.3900	0.9723	0.0542	0.2226	0.4784
1	0.0754	0.9906	0.0325	0.1549	0.0453
2	0.00962	0.9956	0.0136	0.0947	0.00703
3	0.00646	0.9966	0.0105	0.0722	0.00476
4	0.00591	0.9966	0.0113	0.0767	0.00448
5	0.00588	0.9966	0.0124	0.0752	0.00446

Table 6: Experiment 5 on Chr12: $d = 6061$, $q = 704$, $\boldsymbol{\Sigma}_{\ell k} = 0.2\boldsymbol{I}$, $\delta q/n = 0.087$, $\delta d/n = 0.75$.

ite	MSE	accuracy	wrong classification	wrong cluster numbers	Gaussian center MSE
0	0.1861	0.9798	0.0397	0.1193	0.3894
1	0.03854	0.9939	0.0265	0.1179	0.0797
2	0.00315	0.9976	0.00695	0.0440	0.00245
3	0.00195	0.9979	0.00779	0.0341	0.00209
4	0.00183	0.9980	0.00802	0.0369	0.00204
5	0.00178	0.9980	0.00745	0.0327	0.00193

6 Real data analysis

In this section, we implement our Model (2.1) and algorithm on the problem of identifying brain subnetworks via clustering their genetic associations. The connectivity and segmentation of human brain have been a research priority to facilitate our understanding of cognitive behavior and neuropsychiatric disorders (Bullmore and Sporns, 2009). Functionally connected regions are generally referred to as a brain subnetwork or a functional module (Bullmore and Sporns, 2009). The composing regions of a brain subnetwork could spread widely over the brain or be spatially adjacent (Yeo et al., 2011). It is acknowledged that each brain subnetwork has specialized functions and multiple subnetworks coordinate with each other to manage complicated tasks (Smith et al., 2009). Abnormal connectivity of some brain regions are shown to be related to cognitive or neuropsychiatric disorders, such as schizophrenia (Marín, 2012), autism (Assaf et al., 2010), depression, Alzheimer's disease (Hahamy et al., 2015), and so on. With the emerging advanced medical techniques, such as Magnetic resonance imaging (MRI), functional MRI (fMRI) (Yeo et al., 2011), magnetoencephalography (MEG) and Electroencephalography (EEG), researchers are one-step closer to uncover the synergy, balance and connection among brain regions. However, these types of data are sometimes contaminated with large noises which could result in spurious discoveries or fail to detect the true connectivity. Moreover, the dimensionality of these datasets is intimidating, and analyzing them is computationally expensive.

On another hand, it has been reported that genes' influences vary over different brain regions (Thompson et al., 2001; Smith et al., 2020). The brain regions in the same functional or structural subnetwork are more likely to have common genetic mechanisms (Wen et al., 2016; van der Meer et al., 2020). For example, the structures of Broca's and Wernicke's language areas (Thompson et al., 2001) are more prone to genetic influences than other brain regions; genetic factors constrain the imaging-phenotypes of the visual, sensorimotor, basal ganglia regions more than those of the default-mode, executive control and attention regions (Fu et al., 2015); the functional connectivity of default-mode network regions seem to be influenced by a specific set of genes (Glahn et al., 2010).

Inspired by these phenomena, we propose to identify brain subnetworks via clustering their genetic associations. Each cluster would correspond to a potential brain subnetwork. Model (2.2) and the algorithm in Section 3 would be a perfect tool for this purpose. Herein after, let us denote the subnetworks identified by clustering the brain regions' associations with gene-A as "the subnetwork induced by gene-A". The proposed approach would have multiple benefits: if gene-A's biological annotation is largely known, the subnetworks induced by gene-A might correspond to some functional modules related to gene-A's biological annotation; conversely, if some regions within a gene-A-induced subnetwork are partially known, it is natural to conjecture gene-A's biological function based on these partially-known regions; one might also discover some new functions of a brain region based on which subnetwork it belongs to; it automatically predicts the possibly differential expression levels of gene-A over different brain subnetworks. To the best of our knowledge, this is the first attempt to identify brain subnetworks based on their genetic associations. Through our procedure, we could achieve four primary goals:

1. Identify the genes that are associated with the phenotypes of one or multiple brain regions;

2. Identify some new brain subnetworks based on the clustering pattern of the brain regions' genetic associations;

3. Predict some new functions of gene-A based on the subnetworks it induces;

4. Provide insights on the biological mechanisms of gene-A functioning by connecting its reported traits with its induced subnetworks.

The dataset to be analyzed is the Human Connectome Project (HCP) SNP dataset previously described in Section 4. The responses $\boldsymbol{Y}$ are the regional volumes of 68 brain regions of the 1056 individuals in the HCP Desikan-Killiany-Tourville (DKT) adult atlas (Desikan et al., 2006), *i.e.*, the i-th row of $\boldsymbol{Y} \in \mathbb{R}^{1056\times 68}$ records the 68 brain regional volumes of the i-th individual while the j-th column of $\boldsymbol{Y}$ records the j-th brain regional volumes of the 1056 individuals. In this case, $n = 1056$ and $p = 68$. The covariates $\boldsymbol{X} \in \mathbb{R}^{1056\times d}$ represent the genetic measures of these 1056 individuals on one particular autosome. We will fit $\boldsymbol{Y}$ against one autosome at a time, and perform analysis on all 22 autosomes in a successive manner. For example, when exploring the genetic association pattern between the genes on Chromosome 1 and the brain regions, we take the genetic measures on Chromosome 1 as $\boldsymbol{X}$ and fit Model

(2.2) on $(\boldsymbol{Y}, \boldsymbol{X})$. The similar procedure is repeated for all 22 autosomes. Theoretically, it is better to fit $\boldsymbol{Y}$ against all genetic measures on the whole-genome simultaneously to obtain the partial correlations. But in that case the design matrix will be ultra high-dimensional, and we would need to make the thresholding levels large enough to produce a meaningful estimate, which could result in mis-detection of some significant gene or subnetworks. To ensure high detection efficiency, we run separate analyses on individual chromosomes.

The way we obtain the genetic measures $\boldsymbol{X}$ is as follows. Taking Chromosome 1 as an example, the original covariate set is a 1056×87537 matrix recording totally 87537 SNPs on Chromosome 1 of the 1056 individuals. However, there are strong correlations (1 or -1) among the SNPs within the same gene or regulatory region. To circumvent this predicament, we use the principle components (PCs) within each gene or regulatory region. We first group the SNPs according to their physical locations on Chromosome 1. The SNPs within the same gene or genomic region are grouped together. Within each group we obtain the leading principle components whose cumulative energy exceeds 85%. The ℓ-th block of our covariate matrix, $\boldsymbol{X}_\ell$, represents the leading principle components of the SNPs within the ℓ-th gene. Note the complete set of PCs within one group form an orthogonal basis for the column space of this group. If $\boldsymbol{y}_j$ is not correlated with any of the leading PCs, $\boldsymbol{y}_j$ has negligible correlation with the original SNPs. Therefore, for the mere purpose of group selection (gene selection), using the leading PCs as our $\boldsymbol{X}$ would be a wise choice. After replacing the SNPs with PCs, there still exist a few overwhelming correlations between adjacent groups. This is expected since adjacent genes sometimes exhibit linkage disequilibrium (LD). So we further merge some adjacent gene pairs whose PCs are strongly correlated (≥ 0.8 or ≤ -0.8). We could interpret a pair of such genes as a heritable genetic unit.

To obtain robust results, stability selection (Meinshausen and Bühlmann, 2010) is employed in all 22 analyses. In each anlysis, we fit Model (2.2) 10 times on randomly subsampled datasets and also on the whole dataset, which makes 11 runs in each analysis. Every subsample has size of 80% of the whole dataset. A gene is eventually selected if it has been selected 10 times or more out of the total 11 times. For each selected gene, its induced clustering of the brain regions will be recorded if the clustering memberships are consistent in more than 6 times out of the total 11 times. For each brain region we identify a set of significant genes (ranging from 30 to 100) associated with its regional volume. Most of the selected genes are validated by existing literature. Due to limited space, we list part of the selected genes in Table S1 (in Supplementary materials). Among the selected genes, about 5% show evident clustering patterns among their genetic influences on these brain regions. Recall that for a particular gene-A, the brain regions with similar genetic associations with gene-A are clustered together resulting in one or more subnetworks. The gene-induced brain subnetworks are summarized in Table S2-S4.

6.1 Verification

To demonstrate the reliability of our proposed method, we find supporting literature for the selected genes (partially shown in Table S1) and perform two independent analyses to validate our identified brain subnetworks. First, we compare our identified brain subnetworks to the well-defined Yeo7 networks ((Yeo et al., 2011)) and compute the percentages of overlap. Second, we compare the expression levels of a gene inside the subnetwork induced by this gene and outside the subnetwork. In the first analysis, we analyze the composition of each identified subnetwork. The composition means how much percentages of our identified subnetwork belong to the Yeo7 networks, respectively: Visual, Somatomotor, Dorsal Attention, Salience/Ventral Attention, Limbic, Frontoparietal and Default. For example, subnetwork 1 has totally 10 regions with 5 belonging to Limbic, 4 belonging to Visual, and 1 belonging to Frontoparietal. Then its composition is $(0.4, 0.0, 0.0, 0.0, 0.5, 0.1, 0.0)$. The compositions of our identified brain subnetworks are partially summarized in Table 7 due to space limitation. We can see that most identified subnetworks have enrichment in one or two Yeo networks, indicating that our method does not just randomly pick up brain regions. It is also reasonable that our identified subnetworks do not align perfectly with Yeo7 networks. Each of our subnetworks is identified based on the variations of the genetic effects of one particular gene, while Yeo7-network system is drawn according to functional connectivity. Each of our identified subnetworks can be viewed as a module that is most relevant to the biological function of the gene that induces it. They can also be viewed as new parcellations of the human brain, which complement existing parcellations according to functional connectivity. In the second analysis, we use whole-genome expression provided by Allen Institute of Brain Science (AIBS) of six donors (H0351.1009, H0351.1016, H0351.1015, H0351.2002, H0351.1012 and H0351.2001) (Hawrylycz et al., 2014). The gene expression data is averaged across all probes and all samples from six donors, and then normalized via Z-transformation. The actual gene expression data is a 20737×68 matrix representing the Z-transformed expression values of the 20 737 genes across the 68 cortical regions. The details for preprocessing genome-wide expression data are provided in Supplementary Material, Section S1.

In this subsection, we just elaborate five most interesting brain subnetworks (listed on the top of Table 7) identified by our method. We will elaborate a few more in Subsection 6.2. For better visualization, we depict the 5 brain subnetworks in Figures 6.1-6.1.

The first brain subnetwork (6.1(a)) is induced by gene *ASTN2*. The subnetwork is overlapping with a large portion of the Default network and a few regions in the Limbic, Somatomotor, and Visual networks. We also observed differential gene expression levels within the identified subnetwork and outside the subnetwork (6.1(b)). It demonstrates that our method could group some functional related regions in the same cluster, and indeed the gene has differential expression levels inside and outside the cluster. To analyze the subnetwork more carefully, these brain regions serve to accomplish passive tasks (Buckner, 2022), sensory(Beheshtian et al., 2021), response (Haznedar et al., 2000). Gene *ASTN2* encodes a protein that is highly expressed in human brain. There has been reports associating

Table 7: Compositions of gene-induced subnetworks compared with the Yeo7 Networks

Gene	# of regions	Frontoparietal	Default	Dorsal Attention	Limbic	Salience	Somatomotor	Visu
CDH13	18	0.000	0.500	0.000	0.167	0.111	0.000	0.22
UMODL1	12	0.083	0.167	0.000	0.167	0.000	0.167	0.41
SGSM1	4	0.500	0.250	0.000	0.000	0.000	0.000	0.25
CNTNAP2	21	0.095	0.429	0.000	0.143	0.143	0.095	0.09
CNTNAP2	22	0.045	0.227	0.091	0.182	0.091	0.091	0.27
ASTN2	20	0.000	0.550	0.000	0.200	0.050	0.100	0.10
HCN2	11	0.000	0.636	0.000	0.000	0.364	0.000	0.00
ABR	21	0.047	0.429	0.000	0.238	0.095	0.000	0.19
SORCS1	24	0.042	0.208	0.000	0.250	0.167	0.125	0.20
SORCS1	19	0.053	0.158	0.053	0.316	0.157	0.106	0.15
MACROD2	22	0.000	0.273	0.000	0.273	0.091	0.136	0.22
MACROD2	22	0.136	0.318	0.091	0.045	0.091	0.136	0.18
KIF26B	9	0.000	0.444	0.000	0.222	0.222	0.000	0.11
NAALADL2	17	0.059	0.176	0.000	0.412	0.176	0.059	0.11
TENM3	7	0.000	0.571	0.000	0.286	0.000	0.000	0.14
MED10	13	0.000	0.385	0.000	0.385	0.076	0.154	0.00
MED10	14	0.000	0.429	0.000	0.000	0.214	0.071	0.28
LDLRAD3	15	0.000	0.467	0.000	0.200	0.200	0.000	0.13
CHST11	5	0.000	0.200	0.000	0.400	0.000	0.000	0.40
HMGB1	8	0.000	0.500	0.125	0.000	0.000	0.125	0.25
HMGB1	9	0.222	0.222	0.000	0.556	0.000	0.000	0.00
SETD3	10	0.000	0.500	0.000	0.300	0.200	0.000	0.00
ENSG268864	14	0.000	0.500	0.000	0.000	0.000	0.000	0.50
TOX2	6	0.000	0.500	0.000	0.333	0.000	0.167	0.00
NCAM2	12	0.167	0.417	0.000	0.000	0.000	0.083	0.33
C2CD2	7	0.000	0.429	0.000	0.571	0.000	0.000	0.00
TBC1D22A	11	0.000	0.455	0.000	0.182	0.000	0.363	0.00
CDC45	7	0.000	0.714	0.000	0.143	0.000	0.000	0.14

ASTN2 with vertex-wise sulcal depth (Van Der Meer et al., 2021), depression (Howard et al., 2019), brain regional volumes (Zhao et al., 2019), bipolar disorder (Coleman et al., 2020), and brain functional connectivity (Li et al., 2021). Although the function of *ASTN2* is largely unknown, studies show that it might regulate neural recognition molecules (Glessner et al., 2009) and synapse formation (Behesti et al., 2018). Besides, its related gene *ASTN1* encodes a neuronal adhesion molecule required for glial-guided migration of young postmitotic neuroblasts in cortical regions. Taking into account all these pieces of information, it is reasonable to conjecture *ASTN2* might be involved in neuron migration process. While these biological processes are most likely to be executed in the brain regions within the ASTN2-induced subnetwork.

The second subnetwork (6.1(a)) is induced by gene *SGSM1* whose full name is small G protein signaling modulator 1. It is highly expressed in brain according to the Genotype-Tissue Expression (GTEx) database. It is probably involved in activating GTPase and intracellular molecule transport (Yang et al., 2007), and has been associated with nicotine dependence and major depression (Zhou et al., 2018). Such information indicates that gene *SGSM1* is important for signal transduction in neurons and neuron-neuron interactions. The induced subnetwork is highly overlapped with the Frontoparietal network. It is known that Frontoparietal network serves to rapidly and instantiate new task states by flexibly interacting with other control and processing networks (Marek and Dosenbach, 2022). The regions in Frontoparietal network are generally active in signal transduction and neural transmitter exchange, which are consistent with the biological functions of gene *SGSM1*.

The third brain subnetwork is induced by gene *CDH13*. This subnetwork is enriched in the Default, Visual and Salience networks (6.1(a)). Gene *CDH13* encodes a member of the cadherin superfamily, which possibly acts as a negative regulator of axon growth during neural differentiation (Mossink et al., 2022). Such information indicates its potential role in neural stem cell preservation and differentiation in these regions. This potential function and the composition of the subnetwork are consistent with its reported traits in GWAS, educational attainment, cognitive ability, math ability (Davies et al., 2018), alcohol use disorder (Zhou et al., 2020). Moreover, functional connectivity within the Default network has been linked to genetic variants within gene *CDH13* in a study of bipolar disorder and schizophrenia (Meda et al., 2014).

The fouth subnetwork (6.1(a)) is induced by gene *CNTNAP2*. *CNTNAP2* encodes a member of the neurexin family which functions as cell adhesion molecules and receptors and is almost exclusively expressed in brain. It mediates interactions between neurons during nervous system development and glia and is also involved in localization of potassium channels within differentiating axons (Strauss et al., 2006). Gene *CNTNAP2* has been associated with autism (Alarcón et al., 2008), schizophrenia (Friedman et al., 2008), Alzheimer's disease (Jansen et al., 2019) and self-reported math ability (Lee et al., 2018). Variations in gene *CNTNAP2* are associated with altered functional connectivity in frontal lobe circuits (Scott-Van Zeeland et al., 2010), and homozygous loss of gene *CNTNAP2* results in reduced local and long-range prefrontal functional connectivity (Liska et al., 2018). Considering that medial prefrontal cortex is a part of the Default network, our results are consistent with the existing literature. *CNTNAP2* encodes a member of the neurexin family which functions as cell adhesion molecules and receptors and is almost exclusively expressed in brain. It mediates interactions between neurons during nervous system development and glia and is also involved in localization of potassium channels within differentiating axons (Strauss et al., 2006). Gene *CNTNAP2* has been associated with autism (Alarcón et al., 2008), schizophrenia (Friedman et al., 2008), Alzheimer's disease (Jansen et al., 2019) and self-reported math ability (Lee et al., 2018). Variations in gene *CNTNAP2* are associated with altered functional connectivity in frontal lobe circuits (Scott-Van Zeeland et al., 2010), and homozygous loss of gene *CNTNAP2* results in reduced local and long-range prefrontal functional connectivity (Liska et al., 2018). Considering that medial prefrontal cortex is a part of the Default network, our results are consistent with the existing literature.

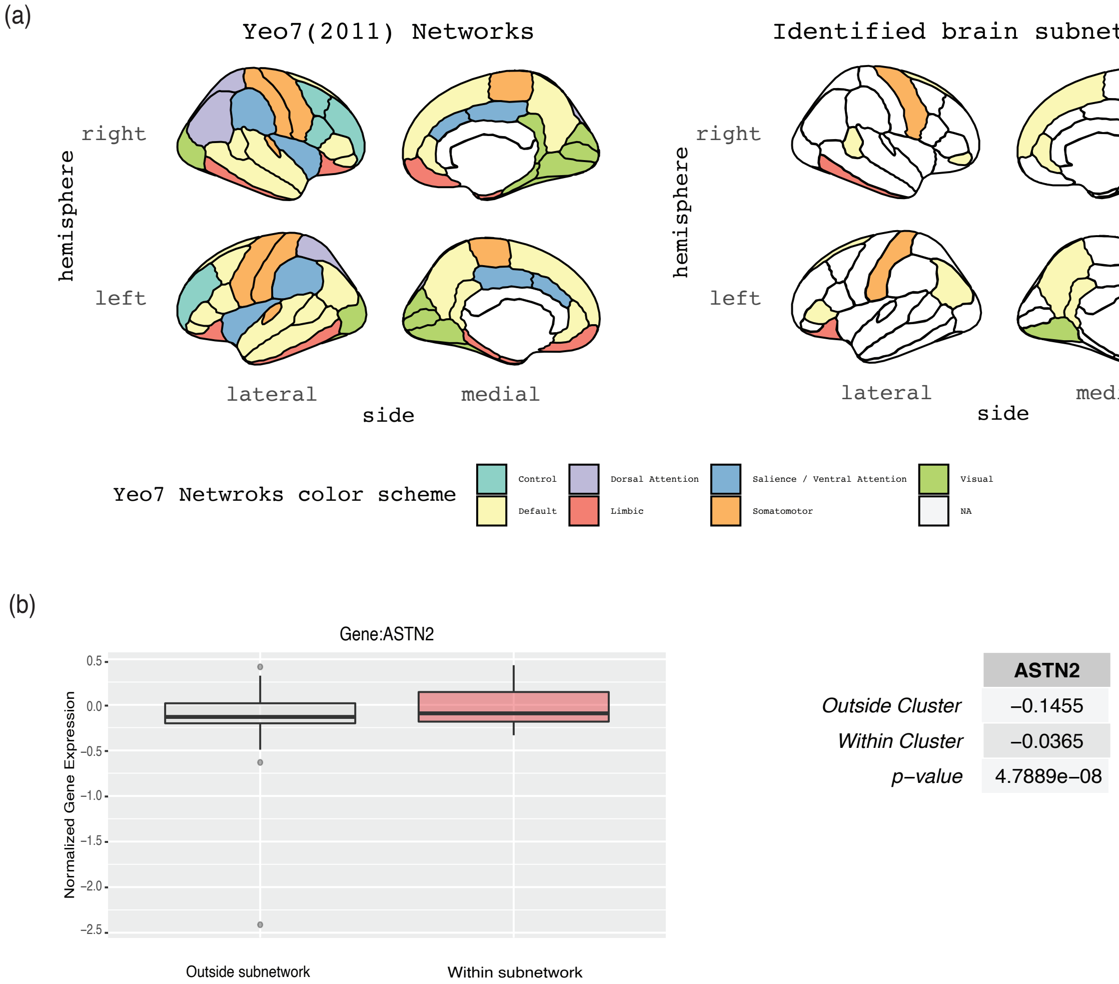
The fifth subnetwork (6.1(a)) is induced by gene *UMODL1*. The composition of this subnetwork is enriched in the Visual network. This is consistent with existing literature that *UMODL1* is associated with the susceptibility in myopia (Singh and Tyagi, 2018) and with brain vertex-wise sulcal depth (Van Der Meer et al., 2021).

6.2 New discovery

In this subsection, we report mainly three folds of novel discoveries. First, we list some new brain subnetworks induced by genes that are well-studied in Table S2. These subnetworks have not been reported in existing literature to the best of our knowledge, and thus their modular functions or properties are unknown. Luckily, we know well about the genes that induce them. Based on the functional annotation of gene-A, we could conjecture the modular functions or properties of its induced subnetworks. Second, we list in Table S3 the subnetworks induced by genes whose functions are largely unknown. It is reasonable to predict some new functions of gene-A by checking the composition of its induced subnetworks. Third, we connect the reported traits of gene-A to the known functions/properties of its induced subnetworks (listed in Table S4). Incorporating reported traits of a gene and its induced subnetworks give insights on the biological mechanism by which this gene causes the reported traits.

Three typical example for the first fold of discovery are gene *ELFN1* on Chromosome 7, gene *ABR* on Chromosome 17, and gene *HCN2* on Chromosome 19. Studies have shown that *ELFN1* inhibits protein phosphatase, and is involved in synapse organization (Dunn et al., 2018). It is located in dendrite and excitatory synapse (Dunn et al., 2018). While its induced subnetwork contains regions belonging to Visual, Somatomotor, Salience, and Limbic networks. These regions serve to sensory(Beheshtian et al., 2021), selecting stimuli (Uddin, 2016), and response (Haznedar et al., 2000). It suggests that these regions could be a functional module that is active in synapse formation. *HCN2* encodes a hyperpolarization-activated cation channel involved in the generation of native pacemaker activity in the heart and in the brain (Thomas et al., 2019). It is well know that the native pacemakers in brain are

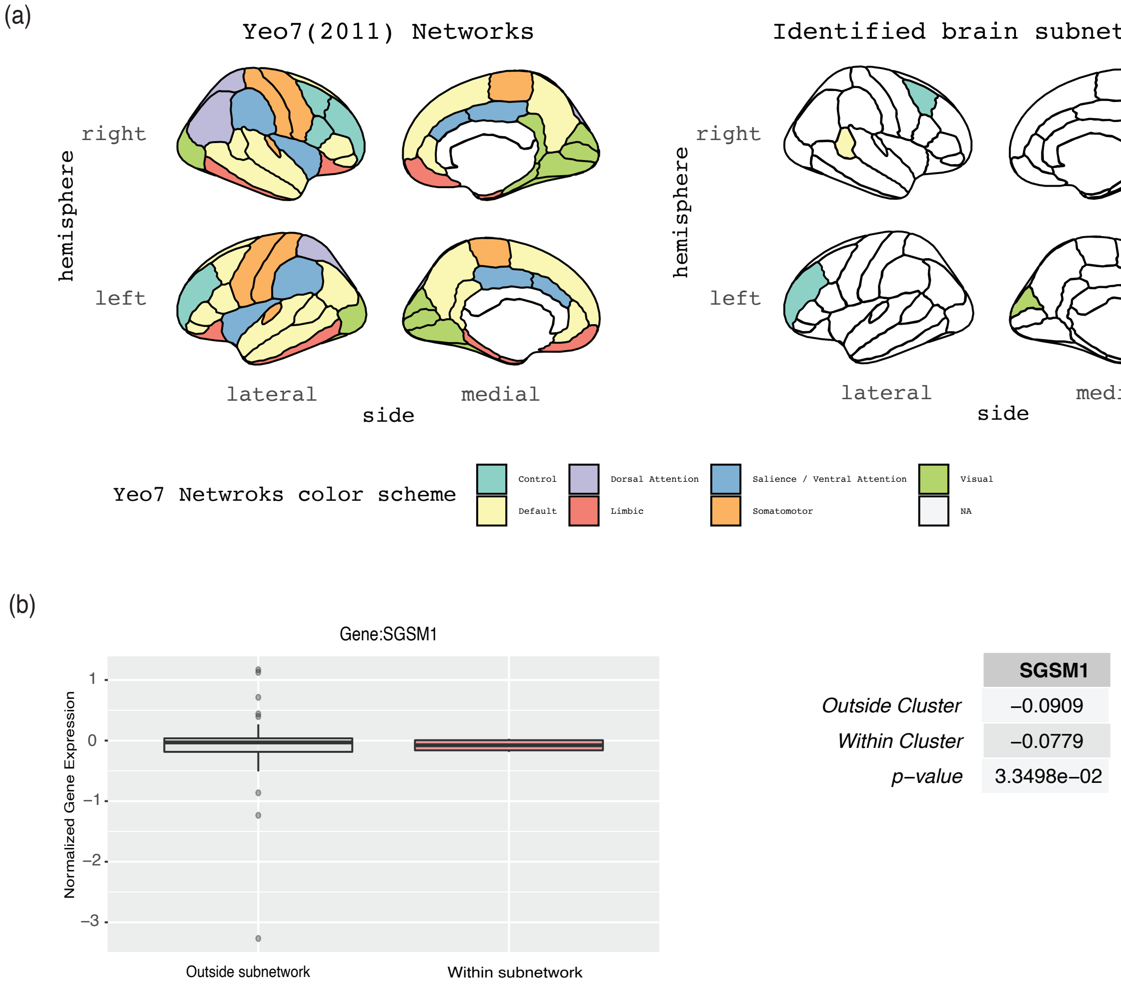
Figure 1: Subnetwork induced by gene *ASTN2*. (a) Identified subnetwork (right) and the Yeo7 network (left). Cyan: Control. Yellow: Default. Purple: Dorsal Attention. Red: Limbic. Blue: Salience / Ventral Attention. Orange: Somatomotor. Green: Visual. (b) Boxplots showing the gene expressions within the identified subnetwork (in blue) compared to those outside the subnetwork (in red). (c) The normalized average gene expression level within and outside the subnetwork, and the p-value for testing two group means.



	ASTN2
Outside Cluster	−0.1455
Within Cluster	−0.0365
p−value	4.7889e−08

vital for sensing and processing passive stimuli. And *HCN2* induced a subnetwork consists regions belonging to the Default and Somatomotor networks. It suggests that these brain regions belong to the functional module that senses and processes passive stimuli. *ABR* encodes a GTPase-activating protein, and thus regulates a wide range of signal transduction pathways, protein biosynthesis, cell proliferation/differentiation/movement (Heisterkamp et al., 1993). Studies in mice reveal its role in vestibular morphogenesis. Indeed, ABR-induced

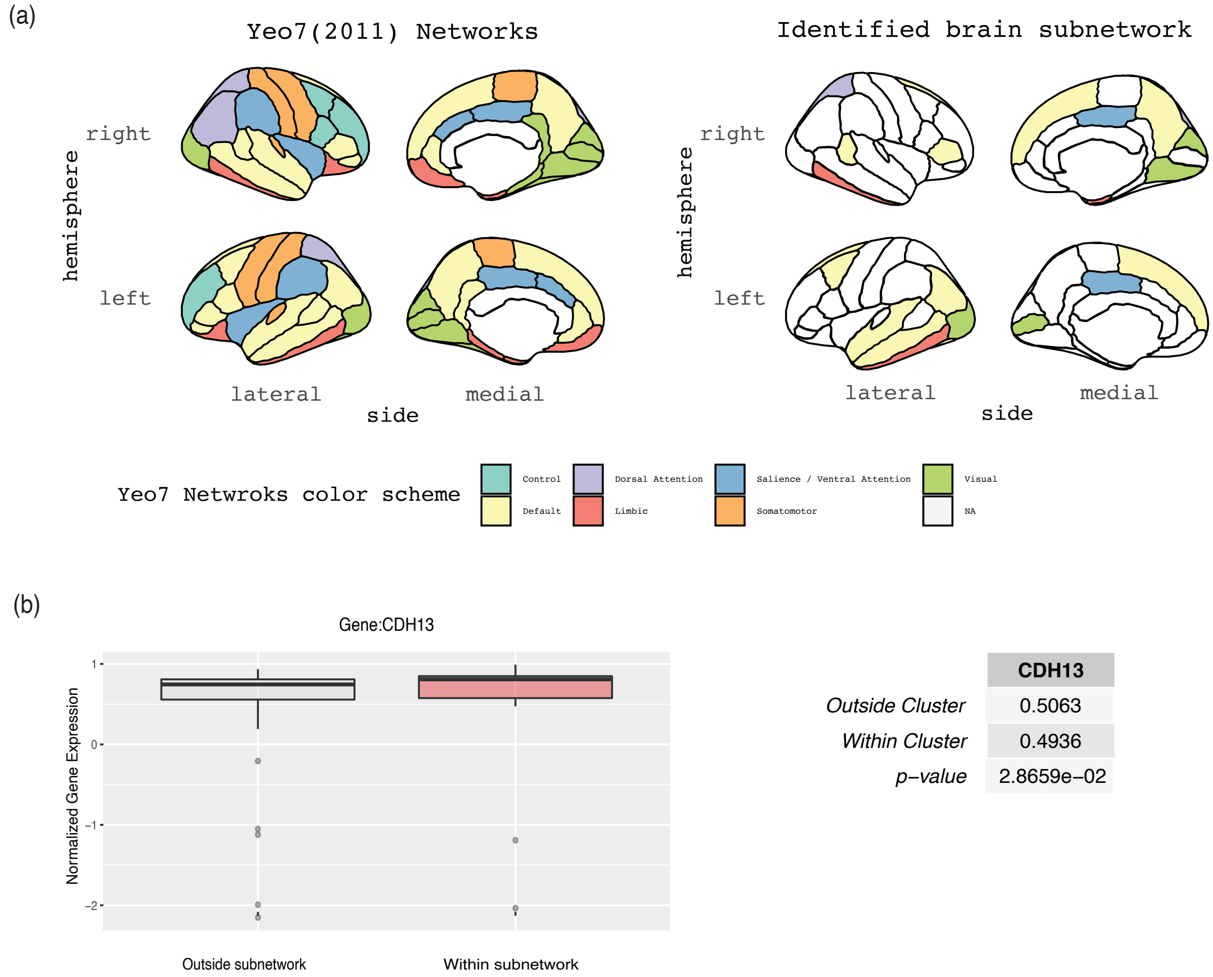
Figure 2: Subnetwork induced by gene *SGSM1*. (a) Identified subnetwork (right) and the Yeo7 network (left). Cyan: Control. Yellow: Default. Purple: Dorsal Attention. Red: Limbic. Blue: Salience / Ventral Attention. Orange: Somatomotor. Green: Visual. (b) Boxplots showing the gene expressions within the identified subnetwork (in blue) compared to those outside the subnetwork (in red). (c) The normalized average gene expression level within and outside the subnetwork, and the p-value for testing two group means.



	SGSM1
Outside Cluster	−0.0909
Within Cluster	−0.0779
p−value	3.3498e−02

brain subnetwork in our study includes almost all the regions in the vestibular system, which further supports the reliability of our method. ABR-induced networks also includes some regions in Limbic, Salience, and Visual networks, indicating these brain regions could be relatively more active in signaling transduction, synapse re-wiring, and cell renewal. It is acknowledged that Visual network processes the information seen by eyes, Salience network selects stimuli that are deserving of our attention, and Limbic network involves in processing

Figure 3: Subnetwork induced by gene *CDH13*. (a) Identified subnetwork (right) and the Yeo7 network (left). Cyan: Control. Yellow: Default. Purple: Dorsal Attention. Red: Limbic. Blue: Salience / Ventral Attention. Orange: Somatomotor. Green: Visual. (b) Boxplots showing the gene expressions within the identified subnetwork (in blue) compared to those outside the subnetwork (in red). (c) The normalized average gene expression level within and outside the subnetwork, and the p-value for testing two group means.

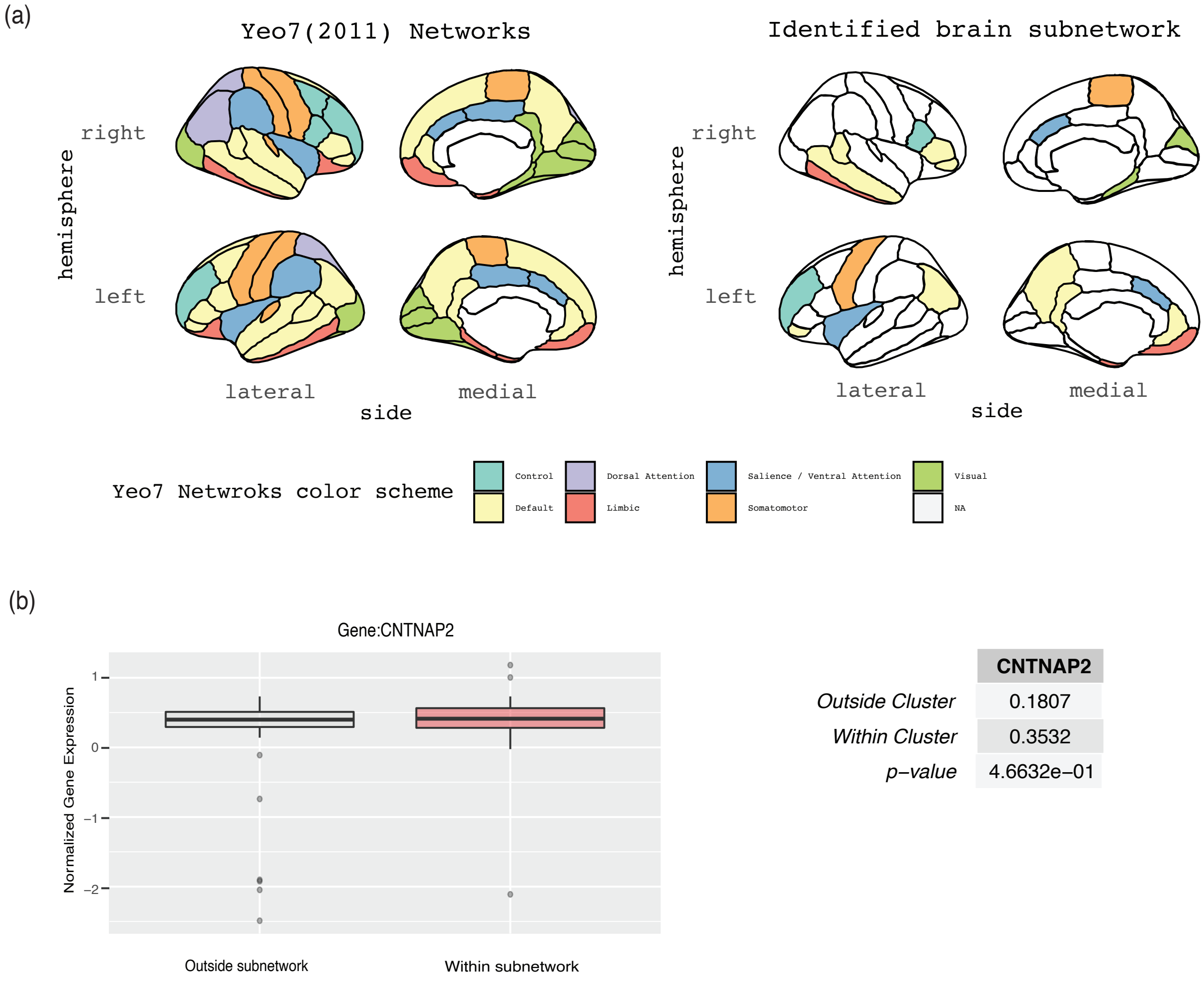


	CDH13
Outside Cluster	0.5063
Within Cluster	0.4936
p–value	2.8659e–02

memory and emotion. Taking into account such information, it immediately suggests that ABR-induced subnetwork might be a functional module: collecting what we see, selecting meaningful stimuli, and store the selected information.

Two typical examples for the second fold of discovery are gene *ASTN2* on Chromosome 9 and gene *SORCS1* on Chromosome 10. *SORCS1* encodes a type I transmembrane receptor protein required for the sorting of the soluble vacuolar hydrolase carboxypeptidase Y (Xu

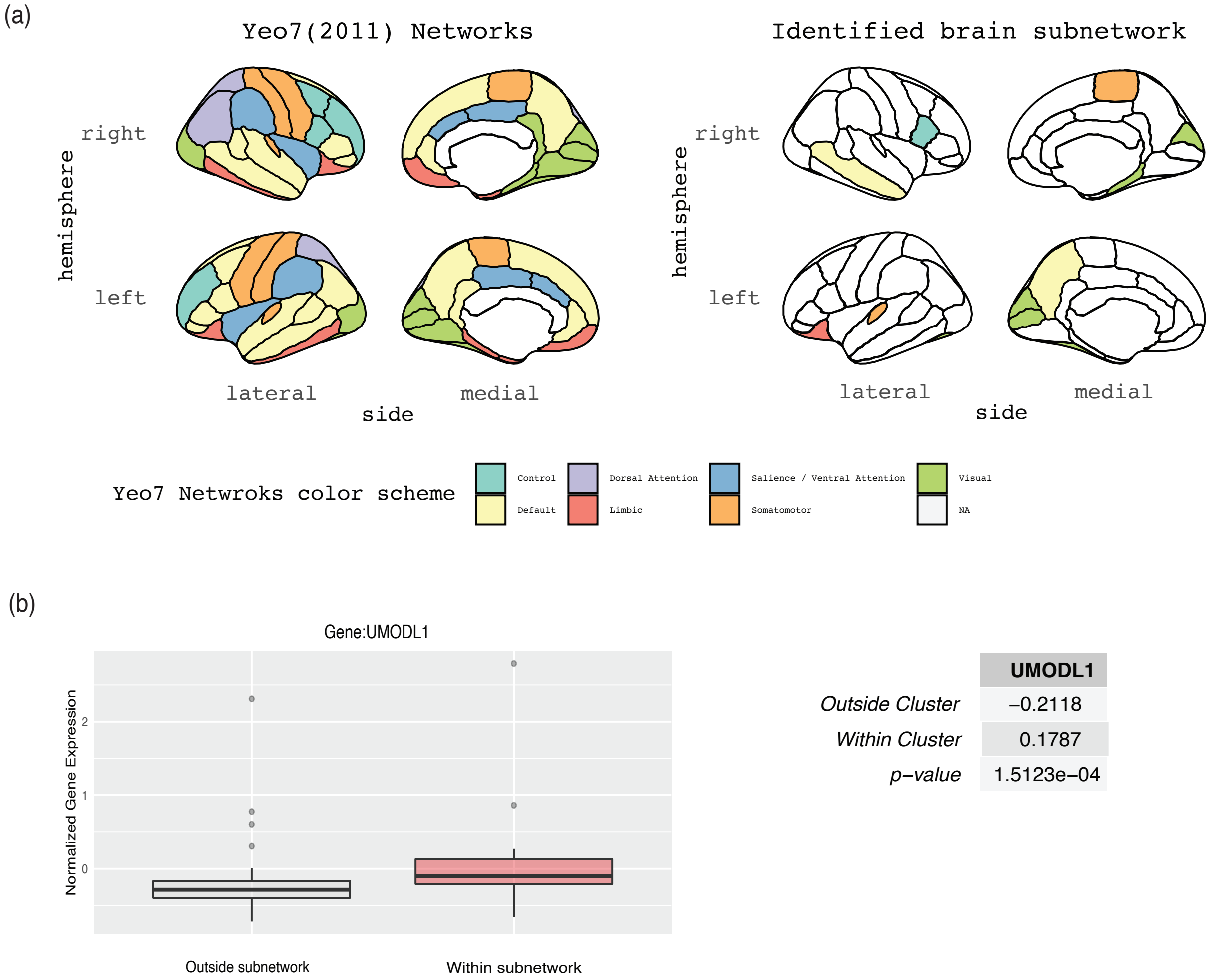
Figure 4: Subnetwork induced by gene *CNTNAP2*. (a) Identified subnetwork (right) and the Yeo7 network (left). Cyan: Control. Yellow: Default. Purple: Dorsal Attention. Red: Limbic. Blue: Salience / Ventral Attention. Orange: Somatomotor. Green: Visual. (b) Boxplots showing the gene expressions within the identified subnetwork (in blue) compared to those outside the subnetwork (in red). (c) The normalized average gene expression level within and outside the subnetwork, and the p-value for testing two group means.



	CNTNAP2
Outside Cluster	0.1807
Within Cluster	0.3532
p−value	4.6632e−01

et al., 2013). It is highly expressed in human thyroid and brain. Its induced subnetworks consist of regions spreading over all Yeo7 networks, which indicates its role in regulating exocytosis, endocytosis, and cellular transportation of signaling molecules. This is consistent with its high expressions in thyroid and brain: thyroid follicular cells secrete hormones and neurons release transmitters. The GWAS reported traits further support the above conjecture because high-level transmitter leads to insomnia and lack of transmitter could result in

Figure 5: Subnetwork induced by gene *UMODL1*. (a) Identified subnetwork (right) and the Yeo7 network (left). Cyan: Control. Yellow: Default. Purple: Dorsal Attention. Red: Limbic. Blue: Salience / Ventral Attention. Orange: Somatomotor. Green: Visual. (b) Boxplots showing the gene expressions within the identified subnetwork (in blue) compared to those outside the subnetwork (in red). (c) The normalized average gene expression level within and outside the subnetwork, and the p-value for testing two group means.



	UMODL1
Outside Cluster	−0.2118
Within Cluster	0.1787
p−value	1.5123e−04

Alzheimer's disease. *ASTN2* induced subnetwork has been elaborated in Section 6.1.

For the third fold of discovery we also list tow examples, gene *CDH13* on Chromosome 16 and gene *MACROD2* on Chromosome 20. In existing literature, gene *MACROD2* has been associated with autism spectrum disorder (Grove et al., 2019), bipolar disorder (Bigdeli et al., 2021), general cognitive ability (Davies et al., 2018), and left–right brain asymmetry (Sha et al., 2021). In our analysis, gene *MACROD2* induces two brain subnetworks. The first subnetwork mainly overlaps with the Default, Limbic, and Visual networks. While it has been shown that abnormalities of connectivity in Default network and Limbic network lead to bipolar disorder (Öngür et al., 2010; Liu et al., 2019) and autism (Jung et al., 2014; Haznedar et al., 2000). It immediately suggests that gene *MACROD2* function through regulating the regions in Default, Limbic, and Visual networks. The second subnetwork overlaps largely (75 %) with the Frontoparietal network. The Frontoparietal network is involved in a variety of cognitive functions such as working memory, attention, shifting, reasoning, and cognitive ability (Niendam et al., 2012). From these facts, we conjecture that gene *MACROD2* regulates cognitive ability via affecting the Frontoparietal network. *CDH13* induced subnetwork has been elaborated in Section 6.1.

In summary, we identified a set of brain subnetworks induced by individual genes. We could predict the modular functions of a particular subnetwork based the known functions of the inducing gene, conjecture new functions of a gene based on its induced subnetworks, and deduce the biological mechanisms of a gene by integrating its reported traits and induced subnetworks.

Acknowledgments

The authors would like to thank Linjun Zhang for enlightening discussions and thank Cun-Hui Zhang for pointing out related literature.

References

Alarcón, M., Abrahams, B. S., Stone, J. L., Duvall, J. A., Perederiy, J. V., Bomar, J. M., Sebat, J., Wigler, M., Martin, C. L., Ledbetter, D. H., et al. (2008). Linkage, association, and gene-expression analyses identify cntnap2 as an autism-susceptibility gene. *The American Journal of Human Genetics*, 82(1):150–159.

Assaf, M., Jagannathan, K., Calhoun, V. D., Miller, L., Stevens, M. C., Sahl, R., O'Boyle, J. G., Schultz, R. T., and Pearlson, G. D. (2010). Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients. *Neuroimage*, 53(1):247–256.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.

Barton, J. P., De Leonardis, E., Coucke, A., and Cocco, S. (2016). Ace: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics*, 32(20):3089–3097.

Beheshtian, E., Jalilianhasanpour, R., Modir Shanechi, A., Sethi, V., Wang, G., Lindquist, M. A., Caffo, B. S., Agarwal, S., Pillai, J. J., Gujar, S. K., et al. (2021). Identification of the somatomotor network from language task–based fmri compared with resting-state fmri in patients with brain lesions. *Radiology*, 301(1):178–184.

Behesti, H., Fore, T. R., Wu, P., Horn, Z., Leppert, M., Hull, C., and Hatten, M. E. (2018). Astn2 modulates synaptic strength by trafficking and degradation of surface proteins. *Proceedings of the National Academy of Sciences*, 115(41):E9717–E9726.

Bigdeli, T. B., Fanous, A. H., Li, Y., Rajeevan, N., Sayward, F., Genovese, G., Gupta, R., Radhakrishnan, K., Malhotra, A. K., Sun, N., et al. (2021). Genome-wide association studies of schizophrenia and bipolar disorder in a diverse cohort of us veterans. *Schizophrenia bulletin*, 47(2):517–529.

Buckner, R. L. (2022). The brain's default network: origins and implications for the study of psychosis. *Dialogues in clinical neuroscience*.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198.

Cai, T. T., Ma, J., and Zhang, L. (2019). Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2010). Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE.

Coleman, J. R., Gaspar, H. A., Bryois, J., Byrne, E. M., Forstner, A. J., Holmans, P. A., de Leeuw, C. A., Mattheisen, M., McQuillin, A., Pavlides, J. M. W., et al. (2020). The genetics of the mood disorder spectrum: genome-wide association analyses of more than 185,000 cases and 439,000 controls. *Biological psychiatry*, 88(2):169–184.

Davies, G., Lam, M., Harris, S. E., Trampush, J. W., Luciano, M., Hill, W. D., Hagenaars, S. P., Ritchie, S. J., Marioni, R. E., Fawns-Ritchie, C., et al. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature communications*, 9(1):1–16.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.

Dunn, H. A., Patil, D. N., Cao, Y., Orlandi, C., and Martemyanov, K. A. (2018). Synaptic adhesion protein elfn1 is a selective allosteric modulator of group iii metabotropic glutamate receptors in trans. *Proceedings of the National Academy of Sciences*, 115(19):5022–5027.

Friedman, J., Vrijenhoek, T., Markx, S., Janssen, I., Van Der Vliet, W., Faas, B., Knoers, N., Cahn, W., Kahn, R., Edelmann, L., et al. (2008). Cntnap2 gene dosage variation is associated with schizophrenia and epilepsy. *Molecular psychiatry*, 13(3):261–266.

Fu, Y., Ma, Z., Hamilton, C., Liang, Z., Hou, X., Ma, X., Hu, X., He, Q., Deng, W., Wang, Y., et al. (2015). Genetic influences on resting-state functional networks: A twin study. *Human brain mapping*, 36(10):3959–3972.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185.

Glahn, D., Winkler, A., Kochunov, P., Almasy, L., Duggirala, R., Carless, M. A., Curran, J., Olvera, R., Laird, A., Smith, S., et al. (2010). Genetic control over the resting brain. *Proceedings of the National Academy of Sciences*, 107(3):1223–1228.

Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., Zhang, H., Estes, A., Brune, C. W., Bradfield, J. P., et al. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, 459(7246):569–573.

Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O. A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics*, 51(3):431–444.

Hahamy, A., Behrmann, M., and Malach, R. (2015). The idiosyncratic brain: distortion of spontaneous connectivity patterns in autism spectrum disorder. *Nature neuroscience*, 18(2):302–309.

Hawrylycz, M., Ng, L., Feng, D., Sunkin, S., Szafer, A., and Dang, C. (2014). The allen brain atlas. *Springer Handbook of Bio-/Neuroinformatics*, pages 1111–1126.

Haznedar, M. M., Buchsbaum, M. S., Wei, T.-C., Hof, P. R., Cartwright, C., Bienstock, C. A., and Hollander, E. (2000). Limbic circuitry in patients with autism spectrum disorders studied with positron emission tomography and magnetic resonance imaging. *American Journal of Psychiatry*, 157(12):1994–2001.

Heisterkamp, N., Kaartinen, V., van Soest, S., Bokoch, G., and Groffen, J. (1993). Human abr encodes a protein with gaprac activity and homology to the dbl nucleotide exchange factor domain. *Journal of Biological Chemistry*, 268(23):16903–16906.

Hleap, J. S., Susko, E., and Blouin, C. (2013). Defining structural and evolutionary modules in proteins: a community detection approach to explore sub-domain architecture. *BMC structural biology*, 13(1):1–16.

Howard, D. M., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., Shirali, M., Coleman, J. R., Hagenaars, S. P., Ward, J., Wigmore, E. M., et al. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature neuroscience*, 22(3):343–352.

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

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nature genetics*, 51(3):404–413.

Jung, M., Kosaka, H., Saito, D. N., Ishitobi, M., Morita, T., Inohara, K., Asano, M., Arai, S., Munesue, T., Tomoda, A., et al. (2014). Default mode network in young male adults with autism spectrum disorder: relationship with autism spectrum traits. *Molecular autism*, 5(1):1–11.

Kim, S. and Xing, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *The Annals of Applied Statistics*, 6(3):1095–1117.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*, 50(8):1112–1121.

Li, T., Hu, J., Wang, S., and Zhang, H. (2021). Super-variants identification for brain connectivity. *Human brain mapping*, 42(5):1304–1312.

Liska, A., Bertero, A., Gomolka, R., Sabbioni, M., Galbusera, A., Barsotti, N., Panzeri, S., Scattoni, M. L., Pasqualetti, M., and Gozzi, A. (2018). Homozygous loss of autism-risk gene cntnap2 results in reduced local and long-range prefrontal functional connectivity. *Cerebral cortex*, 28(4):1141–1153.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Liu, C., Pu, W., Wu, G., Zhao, J., and Xue, Z. (2019). Abnormal resting-state cerebral-limbic functional connectivity in bipolar depression and unipolar depression. *BMC neuroscience*, 20(1):1–8.

Marek, S. and Dosenbach, N. U. (2022). The frontoparietal network: function, electrophysiology, and importance of individual precision mapping. *Dialogues in clinical neuroscience.*

Marín, O. (2012). Interneuron dysfunction in psychiatric disorders. *Nature Reviews Neuroscience*, 13(2):107–120.

Meda, S. A., Ruano, G., Windemuth, A., O'Neil, K., Berwise, C., Dunn, S. M., Boccaccio, L. E., Narayanan, B., Kocherla, M., Sprooten, E., et al. (2014). Multivariate analysis reveals genetic associations of the resting default mode network in psychotic bipolar disorder and schizophrenia. *Proceedings of the National Academy of Sciences*, 111(19):E2066–E2075.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Mossink, B., Van Rhijn, J.-R., Wang, S., Linda, K., Vitale, M. R., Zöller, J. E., van Hugte, E. J., Bak, J., Verboven, A. H., Selten, M., et al. (2022). Cadherin-13 is a critical regulator of gabaergic modulation in human stem-cell-derived neuronal networks. *Molecular psychiatry*, 27(1):1–18.

Ng, A., Jordan, M., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C., and Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience*, 12(2):241–268.

Öngür, D., Lundy, M., Greenhouse, I., Shinn, A. K., Menon, V., Cohen, B. M., and Renshaw, P. F. (2010). Default mode network abnormalities in bipolar disorder and schizophrenia. *Psychiatry Research: Neuroimaging*, 183(1):59–68.

Sackmann, A., Heiner, M., and Koch, I. (2006). Application of petri net based analysis techniques to signal transduction pathways. *BMC bioinformatics*, 7(1):1–17.

Scott-Van Zeeland, A. A., Abrahams, B. S., Alvarez-Retuerto, A. I., Sonnenblick, L. I., Rudie, J. D., Ghahremani, D., Mumford, J. A., Poldrack, R. A., Dapretto, M., Geschwind, D. H., et al. (2010). Altered functional connectivity in frontal lobe circuits is associated with variation in the autism risk gene cntnap2. *Science translational medicine*, 2(56):56ra80–56ra80.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289.

Sha, Z., Schijven, D., Carrion-Castillo, A., Joliot, M., Mazoyer, B., Fisher, S. E., Crivello, F., and Francks, C. (2021). The genetic architecture of structural left–right asymmetry of the human brain. *Nature human behaviour*, 5(9):1226–1239.

Singh, M. and Tyagi, S. C. (2018). Genes and genetics in eye diseases: a genomic medicine approach for investigating hereditary and inflammatory ocular disorders. *International journal of ophthalmology*, 11(1):117.

Smith, S. M., Elliott, L. T., Alfaro-Almagro, F., McCarthy, P., Nichols, T. E., Douaud, G., and Miller, K. L. (2020). Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. *Elife*, 9:e52677.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the national academy of sciences*, 106(31):13040–13045.

Strauss, K. A., Puffenberger, E. G., Huentelman, M. J., Gottlieb, S., Dobrin, S. E., Parod, J. M., Stephan, D. A., and Morton, D. H. (2006). Recessive symptomatic focal epilepsy and mutant contactin-associated protein-like 2. *New England Journal of Medicine*, 354(13):1370–1377.

Thomas, M., Ranjith, G., Radhakrishnan, A., and Anirudhan, V. A. (2019). Effects of hcn2 mutations on dendritic excitability and synaptic plasticity: a computational study. *Neuroscience*, 423:148–161.

Thompson, P. M., Cannon, T. D., Narr, K. L., Van Erp, T., Poutanen, V.-P., Huttunen, M., Lönnqvist, J., Standertskjöld-Nordenstam, C.-G., Kaprio, J., Khaledy, M., et al. (2001). Genetic influences on brain structure. *Nature neuroscience*, 4(12):1253–1258.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423.

Uddin, L. Q. (2016). *Salience network of the human brain*. Academic press.

van der Meer, D., Frei, O., Kaufmann, T., Shadrin, A. A., Devor, A., Smeland, O. B., Thompson, W. K., Fan, C. C., Holland, D., Westlye, L. T., et al. (2020). Understanding the genetic determinants of the brain with mostest. *Nature communications*, 11(1):1–9.

Van Der Meer, D., Kaufmann, T., Shadrin, A. A., Makowski, C., Frei, O., Roelfs, D., Monereo-Sánchez, J., Linden, D. E., Rokicki, J., Alnæs, D., et al. (2021). The genetic architecture of human cortical folding. *Science advances*, 7(51):eabj9446.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79.

Wang, J., Peng, X., Peng, W., and Wu, F.-X. (2014). Dynamic protein interaction network construction and applications. *Proteomics*, 14(4-5):338–352.

Wang, K., Li, M., and Hakonarson, H. (2010). Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164.

Wen, W., Thalamuthu, A., Mather, K. A., Zhu, W., Jiang, J., de Micheaux, P. L., Wright, M. J., Ames, D., and Sachdev, P. S. (2016). Distinct genetic influences on cortical and subcortical brain structures. *Scientific reports*, 6(1):1–11.

Wu, Y. and Zhou, H. H. (2019). Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in $o($\sqrt {n}$) iterations. arXiv preprint arXiv: 1908.10935$.

Xu, W., Xu, J., Wang, Y., Tang, H., Deng, Y., Ren, R., Wang, G., Niu, W., Ma, J., Wu, Y., et al. (2013). The genetic variation of sorcs1 is associated with late-onset alzheimer's disease in chinese han population. *PloS one*, 8(5):e63621.

Yang, H., Sasaki, T., Minoshima, S., and Shimizu, N. (2007). Identification of three novel proteins (sgsm1, 2, 3) which modulate small g protein (rap and rab)-mediated signaling pathway. *Genomics*, 90(2):249–260.

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*.

Zare, H., Shooshtari, P., Gupta, A., and Brinkman, R. R. (2010). Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC bioinformatics*, 11(1):1–16.

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

Zhao, B., Luo, T., Li, T., Li, Y., Zhang, J., Shan, Y., Wang, X., Yang, L., Zhou, F., Zhu, Z., et al. (2019). Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nature genetics*, 51(11):1637–1644.

Zhou, H., Cheng, Z., Bass, N., Krystal, J. H., Farrer, L. A., Kranzler, H. R., and Gelernter, J. (2018). Genome-wide association study identifies glutamate ionotropic receptor gria4 as a risk gene for comorbid nicotine dependence and major depression. *Translational psychiatry*, 8(1):1–7.

Zhou, H., Sealock, J. M., Sanchez-Roige, S., Clarke, T.-K., Levey, D. F., Cheng, Z., Li, B., Polimanti, R., Kember, R. L., Smith, R. V., et al. (2020). Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nature neuroscience*, 23(7):809–818.