High Dimensional Gaussian Graphical Regression Models with Covariates

Journal:	*Journal of the American Statistical Association*
Manuscript ID	JASA-T&M-2021-0223.R2
Manuscript Type:	Article – Theory & Methods
Keywords:	subject-specific Gaussian graphical model, Gaussian graphical model with covariates, non-asymptotic convergence rate, sparse group lasso, co-expression QTL

High Dimensional Gaussian Graphical Regression Models with Covariates

Abstract

Though Gaussian graphical models have been widely used in many scientific fields, relatively limited progress has been made to link graph structures to external covariates. We propose a Gaussian graphical regression model, which regresses both the mean and the precision matrix of a Gaussian graphical model on covariates. In the context of co-expression quantitative trait locus (QTL) studies, our method can determine how genetic variants and clinical conditions modulate the subject-level network structures, and recover both the population-level and subject-level gene networks. Our framework encourages sparsity of covariate effects on both the mean and the precision matrix. In particular for the precision matrix, we stipulate simultaneous sparsity, i.e., group sparsity and element-wise sparsity, on effective covariates and their effects on network edges, respectively. We establish variable selection consistency first under the case with known mean parameters and then a more challenging case with unknown means depending on external covariates, and establish in both cases the ℓ_2 convergence rates and the selection consistency of the estimated precision parameters. The utility and efficacy of our proposed method is demonstrated through simulation studies and an application to a co-expression QTL study with brain cancer patients.

Keywords: subject-specific Gaussian graphical model; Gaussian graphical model with covariates; non-asymptotic convergence rate; sparse group lasso; co-expression QTL.

1 Introduction

Gaussian graphical models, which shed light on the dependence structure among a set of response variables, have been applied to studies of, for example, gene regulatory networks from gene expression data (Fan et al., 2009; Li et al., 2012; Chen et al., 2016), brain connectivity networks from functional magnetic resonance imaging (fMRI) data (Li and Solea, 2018; Zhang et al., 2019), and firm-level financial networks from stock market data (Kolar et al., 2010). Most existing models consider a homogeneous population obeying a common graphical model (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008; Peng et al., 2009) or several stratified graphical models (Guo et al., 2011; Danaher et al., 2014).

In some applications, graph structures may depend on individuals' characteristics, leading to the notion of subject-specific graphical models. In gene expression networks, external covariates, such as genetic variants, clinical and environmental factors, may affect both the expression levels of individual genes and the co-expression relationships among genes. In biology, genetic variants that alter co-expression relationships are referred to as co-expression quantitative trait loci (QTLs), and identifying them is of keen scientific interest (Wang et al., 2012, 2013; van der Wijst et al., 2018a,b). Other factors such as cellular states and environmental conditions may also alter gene regulatory networks (Luscombe et al., 2004). With these relevant external covariates, a fundamental interest, therefore, is to ascertain how they modulate the subject-level network structures, and recover both the population-level and subject-level gene networks. Characterizing such gene regulatory networks is key in developing gene therapies that target specific gene or pathway disruptions (van der Wijst et al., 2018b).

Though the literature on graphical models has been steadily growing (for example, Mein-

shausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008; Peng et al., 2009; Fan et al., 2009; Xie et al., 2020), relatively few frameworks permit subject-specific graphical model estimation with theoretical justifications. Several works (Rothman et al., 2010; Yin and Li, 2011; Li et al., 2012; Lee and Liu, 2012; Cai et al., 2012; Lin et al., 2016; Chen et al., 2016) considered covariate-dependent Gaussian graphical models, wherein the mean of the nodes depends on covariates, while the network structure is constant across all of the subjects. Guo et al. (2011) and Danaher et al. (2014) jointly estimated several group-specific Gaussian graphical models, where the graph structure is allowed to vary with discrete covariates; Liu et al. (2010) proposed a graph-valued regression, which partitions the covariate space into several subspaces and fits separate Gaussian graphical models for each subspace using graphical lasso. As noted by Cheng et al. (2014), it may be difficult to interpret the relationship between the covariates and the graphical models, as even the adjacent covariate subspaces may differ much. Kolar et al. (2010) considered a nonparametric approach for conditional covariance estimation with continuous covariates. Cheng et al. (2014) considered a conditional Ising model for binary data where the log-odds is modeled as a linear function of external covariates. Ni et al. (2019) considered a conditional DAG model that allows the graph structure to vary with a finite number of discrete or continuous covariates, and assumed a known hierarchical ordering of the nodes. Such pre-knowledge may not always be available in practical settings.

We propose a Gaussian graphical regression model that allows the network structure to vary with external covariates (discrete and continuous) of high dimensions. Specifically, both the mean and the precision matrix are modeled as functions of covariates, enabling estimation of subject-specific graphical models; see Figure 1. To facilitate estimation, we show that our proposed model can be formulated as a sequence of linear regression models that include the interactions between response variables (e.g., gene expressions) and external covariates (e.g.,

genetic variants); Section 2.2. Our model accommodates the setting where both response variables and external covariates are high dimensional, which is frequently encountered in genetic studies, and includes the existing conditional mean Gaussian graphical model (e.g., Yin and Li, 2011) as a special case. To estimate coefficients in the covariate-dependent precision matrix, we impose a sparse group lasso penalty that encourages effective covariates to be sparse and their effects on edges to be sparse as well.

The simultaneously sparse structure leads to a parsimonious model with estimability and interpretability, and also brings considerable theoretical challenges that are to be tackled as follows. We first consider a simpler setting where the mean coefficients are known; this allows us to focus on estimating the precision matrix coefficients that are simultaneously sparse. Recent techniques developed for the sparse group lasso under the usual linear regression setting (Cai et al., 2019) may not be directly applicable, as the design matrix in our setting includes high-dimensional interaction terms and non sub-Gaussian rows. We then investigate a more challenging setting with unknown mean coefficients. In this case, estimating the precision matrix is more delicate with errors arising from the estimation of mean coefficients. For both cases, we derive the non-asymptotic rates of convergence in ℓ_2 norm and establish selection consistency, ensuring that we correctly select edges in both the population- and subject-level networks with probability going to 1.

Our work contributes to both methodology and theory. As to *methodology*, we propose a flexible subject-specific graphical model that depends on a large number of external covariates. We employ a combined sparsity structure that encourages effective covariates and the effect of effective covariates on the network to be simultaneously sparse. With respect to *theory*, we carry out a thorough investigation of the simultaneously sparse estimator, by deriving tight non-asymptotic estimation error bounds and establishing variable selection consistency. Our work addresses the theoretical challenges arising from regressing both the

means and the precision matrices on external covariates. Moreover, as the simultaneously sparse regularizer is non-decomposable, the existing techniques using decomposable regularizers and null space properties (Negahban et al., 2012) are not applicable; see Section 4. Thus, our techniques may advance high-dimensional regression with simultaneously sparse structures. Finally, though motivated by a biological application, our method provides a general regression framework of associating networks with external covariates and is broadly applicable to other scientific fields that involve networks.

The rest of the article is organized as follows. Section 2 introduces the Gaussian graphical regression model and Section 3 discusses model estimation with known mean coefficients. Section 4 investigates theoretical properties of the estimator from Section 3. Section 5 presents a two-step estimation procedure and the related theoretical properties with unknown mean coefficients. Section 6 reports the simulation results, and Section 7 conducts a co-expression QTL analysis using a brain cancer genomics data set. Section 8 concludes the paper with a brief discussion.

2 Graphical Regression Models

2.1 Notation and Preamble

We start with some notation. Given a vector $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we use $\|\boldsymbol{x}\|_0$, $\|\boldsymbol{x}\|_1$, $\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_\infty$ to denote the ℓ_0, ℓ_1, ℓ_2 and ℓ_∞ norms, respectively, and use $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle$ to denote the inner product of $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$. We write $[d] = \{1, 2, \ldots, d\}$. Given an index set $\mathcal{S} \in [d]$, we use $\boldsymbol{x}_\mathcal{S} \in \mathbb{R}^{|\mathcal{S}|}$ to denote the sub-vector of $\boldsymbol{x}$ corresponding to index $\mathcal{S}$. For a matrix $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$, we let $\|\boldsymbol{X}\|$ and $\|\boldsymbol{X}\|_{\max} = \max_{ij} X_{ij}$ denote the spectral norm and element-wise max norm, respectively. Given $\mathcal{S} \in [d_2]$, we use $\boldsymbol{X}_\mathcal{S} \in \mathbb{R}^{d_1 \times |\mathcal{S}|}$ to denote the sub-matrix with columns indexed in $\mathcal{S}$. We use $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ to denote the smallest and largest eigenvalues of a

matrix, respectively. For two positive sequences a_n and b_n, write $a_n \precsim b_n$ or $a_n = \mathcal{O}(b_n)$ if there exist $c > 0$ and $N > 0$ such that $a_n < cb_n$ for all $n > N$, and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$; write $a_n \asymp b_n$ if $a_n \precsim b_n$ and $b_n \precsim a_n$.

Suppose $\boldsymbol{X} = (X_1, \ldots, X_p) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Denote the precision matrix $\boldsymbol{\Sigma}^{-1}$ by $(\sigma^{ij})_{p\times p}$. Under a Gaussian distribution, $\sigma^{ij} \neq 0$ is equivalent to X_i and X_j being conditionally dependent given all other X variables (Lauritzen, 1996). Let $\boldsymbol{X}_{-j} = \{X_k : k \in [p], k \neq j\}$. Meinshausen and Bühlmann (2006) and Peng et al. (2009) related $(\sigma^{ij})_{p\times p}$ to the coefficients in this linear regression model:

$$X_j = \sum_{k\neq j}^{p} \beta_{jk} X_k + \epsilon_j, \quad j \in [p], \tag{1}$$

where ϵ_j is independent of $\boldsymbol{X}_{-j}$ if and only if $\beta_{jk} = -\sigma^{jk}/\sigma^{jj}$; for such defined β_{jk}, it holds that $\mathrm{Var}(\epsilon_j) = 1/\sigma^{jj}$. Consequently, estimating the conditional dependence structure (i.e., finding nonzero σ^{jk}'s) can be viewed as a model selection problem (i.e., finding nonzero β_{jk}'s) under the regression setting in (1).

Let $\boldsymbol{U} = (U_1, \ldots, U_q)^\top$ be a q-dimensional vector of covariates. One may consider a covariate-dependent Gaussian graphical model (Rothman et al., 2010; Yin and Li, 2011; Li et al., 2012; Lee and Liu, 2012; Cai et al., 2012; Chen et al., 2016):

$$\boldsymbol{X}|\boldsymbol{U} = \boldsymbol{u} \sim \mathcal{N}_p(\boldsymbol{\mu}(\boldsymbol{u}), \boldsymbol{\Sigma}), \tag{2}$$

where $\boldsymbol{\mu}(\boldsymbol{u}) = \boldsymbol{\Gamma}\boldsymbol{u}$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times q}$. In expression QTL studies, the jth row of $\boldsymbol{\Gamma}$ specifies how the q genetic regulators affect the expression level of the jth gene. Denote $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)^\top$. Similar to (1), we have that

$$X_j = \boldsymbol{u}^\top \boldsymbol{\gamma}_j + \sum_{k\neq j}^{p} \beta_{jk}(X_k - \boldsymbol{u}^\top \boldsymbol{\gamma}_k) + \epsilon_j, \quad j \in [p], \tag{3}$$

where ϵ_j is independent with $\boldsymbol{X}_{-j}$ if and only if $\beta_{jk} = -\sigma^{jk}/\sigma^{jj}$. With such defined β_{jk}, $\mathrm{Var}(\epsilon_j) = 1/\sigma^{jj}$.

2.2 High Dimensional Gaussian Graphical Regression with Covariates

With a p-dimensional response vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ and a q-dimensional covariate vector $\boldsymbol{U} = (U_1, \ldots, U_q)^\top$, we assume that

$$\boldsymbol{X}|\boldsymbol{U} = \boldsymbol{u} \sim \mathcal{N}_p(\boldsymbol{\mu}(\boldsymbol{u}), \boldsymbol{\Sigma}(\boldsymbol{u})), \tag{4}$$

where $\boldsymbol{\mu}(\boldsymbol{u}) = \boldsymbol{\Gamma}\boldsymbol{u}$ and $\boldsymbol{\Sigma}(\boldsymbol{u})$ are the conditional mean vector and covariance matrix, respectively, and $\boldsymbol{\Omega}(\boldsymbol{u}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{u})$ is the precision matrix linked to $\boldsymbol{u}$ via

$$\boldsymbol{\Omega}(\boldsymbol{u}) = \boldsymbol{B}_0 + \sum_{h=1}^{q} \boldsymbol{B}_h u_h.$$

Here, $\boldsymbol{B}_0, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_q$ are symmetric $p \times p$ coefficient matrices, where $\boldsymbol{B}_0$ characterizes the population level regulatory network, and $\boldsymbol{B}_h$ encodes the effect of u_h on the regulatory network. Specifically, for the (j,k)th entry, we have $\boldsymbol{\Omega}(\boldsymbol{u})_{jk} = [\boldsymbol{B}_0]_{jk} + \sum_{h=1}^{q}[\boldsymbol{B}_h]_{jk} \times u_h$, where $[\boldsymbol{B}_h]_{jk}$ denotes the (j,k)th entry of $\boldsymbol{B}_h$. We assume $\boldsymbol{\Omega}(\boldsymbol{u})_{jj} = \sigma^{jj}$ for any j and comment on it underneath (5). With this assumption, the partial correlation between Z_j and Z_k can be expressed as $\rho_{jk}(\boldsymbol{u}) = -\frac{\boldsymbol{\Omega}(\boldsymbol{u})_{jk}}{\sqrt{\sigma^{jj}\sigma^{kk}}}$. See sufficient conditions on $\boldsymbol{B}_h$'s and $\boldsymbol{u}$ in Section 8 for a positive definite $\boldsymbol{\Omega}(\boldsymbol{u})$. By specifying $\boldsymbol{\Omega}(\boldsymbol{u})_{jk}$'s to linearly depend on $\boldsymbol{u}$, the proposed model allows both the sparsity patterns and the strengths of dependence in $\boldsymbol{\Omega}(\boldsymbol{u})$ to vary with external covariates; see Figure 1. Model (4) is identifiable as long as the number of effective covariates (i.e., nonzero $\boldsymbol{B}_h$'s) is less than n (Wu and Wang, 2020).

As in (1) and (3), model (4) entails estimation of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}(\boldsymbol{u})$ via the following regression models, termed *Gaussian graphical regression*:

$$X_j = \boldsymbol{u}^\top\boldsymbol{\gamma}_j + \sum_{k\neq j}^{p} \beta_{jk0}(X_k - \boldsymbol{u}^\top\boldsymbol{\gamma}_k) + \sum_{k\neq j}^{p}\sum_{h=1}^{q} \beta_{jkh} \underbrace{u_h \times (X_k - \boldsymbol{u}^\top\boldsymbol{\gamma}_k)}_{\text{interaction term}} + \epsilon_j, \tag{5}$$

where $\beta_{jkh} = -[\boldsymbol{B}_h]_{jk}/\sigma^{jj}$and $\mathrm{Var}(\epsilon_j) = 1/\sigma^{jj}$, for all j, k and h. Model (5) provides a regression framework for estimating the mean and precision parameters in (4), by adding to

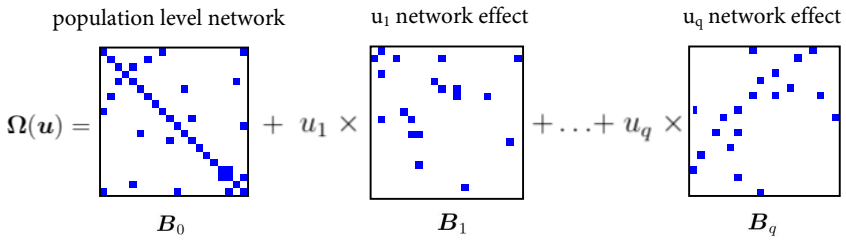

Figure 1: An illustration of the subject-specific Gaussian graphical model.

(1) or (3) the interactions between $\boldsymbol{X}_{-j}$ and $\boldsymbol{u}$. Correspondingly, the partial correlation between X_j and X_k, conditional on all other X variables, is modeled as a function of $\boldsymbol{u}$, forming the basis of Gaussian graphical regression. The diagonal elements of $\boldsymbol{\Omega}(\boldsymbol{u})$ (i.e., σ^{jj}'s) are connected to the residual variances in (5), that is, $\mathrm{Var}(\epsilon_j) = 1/\sigma^{jj}$. From this perspective, assuming σ^{jj} to be free of $\boldsymbol{u}$ may be viewed as assuming the residual variance of Z_j, after removing effects of $\boldsymbol{u}$, $\boldsymbol{Z}_{-j}$ and the interactions between $\boldsymbol{u}$ and $\boldsymbol{Z}_{-j}$, to not dependent on $\boldsymbol{u}$, which is plausible in the context of regression. However, as (3) is a regression-type representation of the precision matrix, caution must be exercised when comparing the residual terms in (3) to the error terms in a standard regression problem; see more discussions in Section 8. Obviously, model (5) includes models (1) and (3) as special cases with $\beta_{jkh} = 0$ for all j, k and h.

Given $\boldsymbol{U} = \boldsymbol{u}$, write $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{\Gamma}\boldsymbol{u} = (Z_1, \ldots, Z_p)$, and re-express (5) as

$$Z_j = \sum_{k \neq j}^{p} \beta_{jk0} Z_k + \sum_{k \neq j}^{p} \sum_{h=1}^{q} \beta_{jkh} u_h Z_k + \epsilon_j. \tag{6}$$

Denote $\boldsymbol{\beta}_j = (\boldsymbol{b}_{j0}, \boldsymbol{b}_{j1}, \ldots, \boldsymbol{b}_{jq})^\top \in \mathbb{R}^{(p-1)(q+1)}$, where $\boldsymbol{b}_{jh} = (\beta_{j1h}, \ldots, \beta_{jph}) \in \mathbb{R}^{p-1}$ for all h; see a more organizational and functional view of $\boldsymbol{\beta}_j$ below:

$$\boldsymbol{\beta}_j = (\ \underbrace{\beta_{j10}, \ldots, \beta_{jp0},}_{\boldsymbol{b}_{j0}:\ \textbf{population level edges of node } j\textbf{'s}} \ \underbrace{\beta_{j11}, \ldots, \beta_{jp1},}_{\boldsymbol{b}_{j1}:\ u_1\textbf{'s effect on edges of node } j\textbf{'s}} \ \ldots, \ \underbrace{\beta_{j1q}, \ldots, \beta_{jpq}}_{\boldsymbol{b}_{jq}:\ u_q\textbf{'s effect on edges of node } j\textbf{'s}} \ )^\top. \tag{7}$$

(group 0 / group 1 / group q)

When both p and q are large, to ensure the estimability of $\boldsymbol{\beta}_j$, we impose on it simultaneous group sparsity and element-wise sparsity. With groups illustrated in (7), we assume

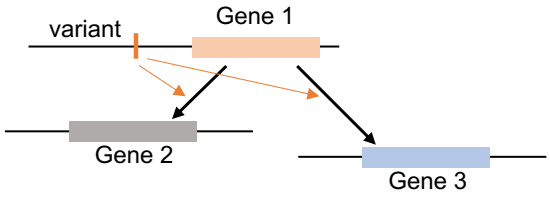

Figure 2: An illustration of gene co-expressions: the genetic variant is a trans-eQTL modulating co-expressions of pairs (1,2) and (1,3), where Gene 1 is an upstream gene and Genes 2 and 3 are downstream genes.

$\boldsymbol{\beta}_j$ is *group sparse*, stipulating that effective covariates are sparse, i.e., only a few covariates may impact edges and those impactful covariates are termed effective covariates. We further assume $\boldsymbol{\beta}_j$ is *element-wise sparse*. That is, effective covariates may influence only a few edges. These simultaneous sparsity assumptions are well supported by genetic studies (van der Wijst et al., 2018a). We exclude $\boldsymbol{b}_{j0}$ from the group sparsity constraint (but not the element-wise sparsity constraint), as it determines the population level regulatory network. With covariate $\boldsymbol{u}$ and sparsity on $\boldsymbol{\beta}_j$'s, it is possible that (X_j, X_k) and (X_k, X_s) are conditionally dependent, while (X_j, X_s) are conditionally independent. This type of structures is biological plausible. Consider as an example our motivating data application in co-expression QTL identification. It is possible for genetic variants located near a gene (say, Gene 1), called the trans-acting expression quantitative trait loci (trans-eQTLs) (Fehrmann et al., 2011), to alter how intensively Gene 1 may regulate (e.g. activate, inhibit) two downstream Genes 2 and 3, while not altering the coexpression between Genes 2 and 3. In fact, the two downstream Genes 2 and 3 can be independent conditional on the rest of the gene network, regardless of what the upstream trans-eQTLs might be (Kolberg et al., 2020; Gong et al., 2018; Brynedal et al., 2017); see Figure 2 for an illustration.

Model (6) can be viewed as an interaction model. Our later development does not abide by the common hierarchical principle for the inclusion of interactions, that is, an interaction

is allowed only if the main effects are present (Hao et al., 2018; She et al., 2018). This is because gene co-expressions may occur only for certain genetic variations (Wang et al., 2013; van der Wijst et al., 2018a), in which case, β_{jkh} (i.e., effect of u_h on edge (j,k)) can be nonzero while β_{jk0} is zero (i.e., population level edge (j,k)). Section 8 discusses modifications of our proposal if hierarchy is to be enforced.

To ease the exposition of key ideas, we first assume a known $\mathbf{\Gamma}$ in the ensuing development, and focus on the estimation of $\boldsymbol{\beta}_j$'s. In Section 5, we drop this assumption, develop an estimation procedure and derive theory when $\mathbf{\Gamma}$ is unknown.

3 Estimation

With n independent observations, denoted by $\mathcal{D} = \{(\boldsymbol{u}^{(i)}, \boldsymbol{x}^{(i)}), i \in [n]\} \in \mathbb{R}^p \times \mathbb{R}^q$, and $\boldsymbol{z}^{(i)} = \boldsymbol{x}^{(i)} - \mathbf{\Gamma}\boldsymbol{u}^{(i)}$. Also denote the samples of the jth $\boldsymbol{z}$ variable by $\boldsymbol{z}_j = (z_j^{(1)}, \ldots, z_j^{(n)})^\top$ for $j \in [p]$ and the samples of the hth $\boldsymbol{u}$ covariate by $\boldsymbol{u}_h = (u_h^{(1)}, \ldots, u_h^{(n)})^\top$ for $h \in [q]$. The Gaussian graphical regression model on the jth response variable can be written as

$$\boldsymbol{z}_j = \sum_{k \neq j}^{p} \beta_{jk0} \boldsymbol{z}_k + \sum_{k \neq j}^{p} \sum_{h=1}^{q} \beta_{jkh} \boldsymbol{u}_h \odot \boldsymbol{z}_k + \boldsymbol{\epsilon}_j, \tag{8}$$

where $\boldsymbol{\epsilon}_j \sim \mathcal{N}_p(\mathbf{0}, 1/\sigma^{jj}\boldsymbol{I})$ and $\odot$ denotes the element-wise product of two equal-length vectors. We partition the vector of $\boldsymbol{\beta}_j$ into $q+1$ blocks indexed by $(0), (1), \ldots, (q) \subset \{1, \ldots, (p-1)(q+1)\}$, such that $(\boldsymbol{\beta}_j)_{(0)} = \boldsymbol{b}_{j0}$ and $(\boldsymbol{\beta}_j)_{(h)} = \boldsymbol{b}_{jh}$, $h \in [q]$.

Denote the squared error loss function by

$$\ell_j(\boldsymbol{\beta}_j | \mathcal{D}) = \frac{1}{2n} \|\boldsymbol{z}_j - \boldsymbol{W}_{-j}\boldsymbol{\beta}_j\|_2^2,$$

where $\boldsymbol{W}_{-j} = [\boldsymbol{z}_1, \boldsymbol{z}_1 \odot \boldsymbol{u}_1, \ldots, \boldsymbol{z}_1 \odot \boldsymbol{u}_q, \ldots, \boldsymbol{z}_{j-1} \odot \boldsymbol{u}_q, \boldsymbol{z}_{j+1}, \boldsymbol{z}_{j+1} \odot \boldsymbol{u}_1 \ldots, \boldsymbol{z}_p \odot \boldsymbol{u}_q]$ is an $n \times (p-1)(q+1)$ matrix. To estimate $\boldsymbol{\beta}_j$, we consider

$$\ell_j(\boldsymbol{\beta}_j | \mathcal{D}) + \lambda \|\boldsymbol{\beta}_j\|_1 + \lambda_g \|\boldsymbol{\beta}_{j,-0}\|_{1,2}, \tag{9}$$

where $\|\boldsymbol{\beta}_{j,-0}\|_{1,2} = \sum_{h=1}^{q} \|(\boldsymbol{\beta}_j)_{(h)}\|_2$ and $\lambda, \lambda_g \geq 0$ are tuning parameters. The convex regularizing terms, $\|\boldsymbol{\beta}_j\|_1$ and $\|\boldsymbol{\beta}_{j,-0}\|_{1,2}$, encourage element- and group-wise sparsity, respectively, though the group sparse penalty is not applied to $(\boldsymbol{\beta}_j)_{(0)}$. The combined sparsity penalty in (9) is termed the *sparse group lasso* penalty (Simon et al., 2013; Li et al., 2015).

As (9) is convex, it can be optimized by using the existing gradient descent algorithms for sparse group lasso (Simon et al., 2013; Vincent and Hansen, 2014), even when both p and q are large. Since the optimizers do not guarantee the symmetry of $\boldsymbol{\Omega}(\boldsymbol{u})$, we propose a post-processing step, similar to Meinshausen and Bühlmann (2006) and Cheng et al. (2014). Denote by $\hat{\beta}^0_{jkh} = -\hat{\sigma}^{jj}\hat{\beta}_{jkh}$, where $\hat{\beta}_{jkh}$ is estimated from (9) and $\hat{\sigma}^{jj}$ from (16) for all j, k and h. With finite samples, we consider the following approach to enforce symmetry:

$$[\boldsymbol{B}_h]_{jk} = [\boldsymbol{B}_h]_{kj} = \hat{\beta}^0_{jkh} 1_{\{|\hat{\beta}^0_{jkh}| < |\hat{\beta}^0_{kjh}|\}} + \hat{\beta}^0_{kjh} 1_{\{|\hat{\beta}^0_{jkh}| > |\hat{\beta}^0_{kjh}|\}}. \tag{10}$$

Symmetrization can also be achieved via

$$[\boldsymbol{B}_h]_{jk} = [\boldsymbol{B}_h]_{kj} = \hat{\beta}^0_{jkh} 1_{\{|\hat{\beta}^0_{jkh}| \geq |\hat{\beta}^0_{kjh}|\}} + \hat{\beta}^0_{kjh} 1_{\{|\hat{\beta}^0_{jkh}| \leq |\hat{\beta}^0_{kjh}|\}}, \tag{11}$$

but it is less conservative as $[\hat{\boldsymbol{B}}_h]_{jk}$ is nonzero if either $\hat{\beta}^0_{jkh}$ or $\hat{\beta}^0_{kjh}$ is nonzero, compared to (10) wherein $[\hat{\boldsymbol{B}}_h]_{jk}$ is nonzero if both $\hat{\beta}^0_{jkh}$ and $\hat{\beta}^0_{kjh}$ are nonzero. Though both are asymptotically equivalent (see Theorem 2), (10) has a better finite sample performance (Meinshausen and Bühlmann, 2006), especially when p is large relative to n.

Two parameters λ and λ_g in (9) require tuning; in our procedure, they are jointly selected via L-fold cross validation. As in Simon et al. (2013) and Cai et al. (2019), we rewrite $\lambda = \alpha\lambda_0$ and $\lambda_g = (1-\alpha)\lambda_0$, where α reflects the weight of the lasso penalty relative to the group lasso penalty and λ_0 reflects the total amount of regularization. We assess a set of values for $\alpha \in [0, 1]$, with $\alpha = 0$ and 1 corresponding to lasso and group lasso, respectively; for each α, a sequence of λ_0 values are considered to obtain the whole regularization path (Vincent

and Hansen, 2014). Finally, we choose the combination of (α, λ_0) that minimizes the cross validation error. In our implementations, we consider $\alpha \in \{0, 0.1, 0.2, \ldots, 0.9, 1\}$ and $L = 5$, and note that the result is fairly robust to the choices of α (see Section 6).

4 Theoretical Properties

In this section, we derive the non-asymptotic ℓ_2 convergence rate of the sparse group lasso estimator from (9) and establish variable selection consistency. Our theoretical investigation is challenged by several unique aspects of the model. First, as the design matrix $\boldsymbol{W}_{-j} \in \mathbb{R}^{n\times(p-1)(q+1)}$ includes high-dimensional interaction terms between $\boldsymbol{z}^{(i)}$ and $\boldsymbol{u}^{(i)}$, and the variance of $\boldsymbol{z}^{(i)}$ is a function of $\boldsymbol{u}^{(i)}$, characterizing the joint distribution of each row in $\boldsymbol{W}_{-j}$ is difficult and requires a delicate treatment. Second, as the combined penalty term $\lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}$ is not decomposable, the classic techniques for decomposable regularizers and null space properties (Negahban et al., 2012) are not applicable. Standard treatments of the stochastic term (Bickel et al., 2009; Lounici et al., 2011; Negahban et al., 2012) such as $\langle \boldsymbol{\epsilon}, \boldsymbol{W}_{-j}\Delta\rangle \leq \|\boldsymbol{W}_{-j}^\top\boldsymbol{\epsilon}\|_\infty\|\Delta\|_1$, where $\Delta \in \mathbb{R}^{(p-1)(q+1)}$, can only yield an ℓ_2 convergence rate comparable to that from the lasso or the group lasso. Utilizing the statistical properties and the computational optimality of the sparse group lasso estimator in (9), we derive two interrelated bounds on the stochastic term. The first bound characterizes the cardinality measure of the covariate space, while the second one utilizes the Karush–Kuhn–Tucker condition and properties of the combined regularizer. Combining these bounds, we give a sharp upper bound on the stochastic term, and show our proposed estimator possesses an improved ℓ_2 error bounds compared to the lasso and the group lasso when the true coefficients are simultaneously sparse; see Section S3.1.

Denote the true parameters by $\boldsymbol{\beta}_j$ for all j, though in some contexts we use them to

denote the corresponding arguments in functions. Let $\mathcal{S}_j$ be the element-wise support set and $\mathcal{G}_j$ be the group-wise support set of $\boldsymbol{\beta}_j$, i.e, $\mathcal{S}_j = \{l : (\boldsymbol{\beta}_j)_l \neq 0, l \in [(p-1)(q+1)]\}$ and $\mathcal{G}_j = \{h : (\boldsymbol{\beta}_j)_{(h)} \neq \mathbf{0}, h \in [q]\}$. Moreover, let $s_j = |\mathcal{S}_j|$, $s_{j,g} = |\mathcal{G}_j|$, and assume $s_j \geq 1$; it follows that $s_{j,g} \leq s_j$, $j \in [p]$. When there is no ambiguity, we write $\boldsymbol{W}$ without noting its dependence on j. Denote by $\sigma^2_{\epsilon_j} = 1/\sigma^{jj}$. We state a few regularity conditions and recall $\boldsymbol{\Sigma}(\boldsymbol{u}^{(i)}) = \mathrm{Cov}(\boldsymbol{z}^{(i)})$, $i \in [n]$.

Assumption 1. *Suppose $\boldsymbol{u}^{(i)}$ are i.i.d. mean zero random vectors with a covariance matrix satisfying $\lambda_{\min}(Cov(\boldsymbol{u}^{(i)})) \geq 1/\phi_0$ for some constant $\phi_0 > 0$. Moreover, there exists a constant $M > 0$ such that $|u_h^{(i)}| \leq M$ for all i and h.*

Assumption 2. *Suppose $\phi_1 \leq \lambda_{min}(Cov(\boldsymbol{z}^{(i)})) \leq \lambda_{max}(Cov(\boldsymbol{z}^{(i)})) \leq \phi_2$ for some constants $\phi_1, \phi_2 > 0$.*

Assumption 1 stipulates that the covariates are element-wise bounded, which is needed in characterizing the joint distribution of each row in $\boldsymbol{W}$. This condition is not restrictive as genetic variants are often coded to be $\{0, 1\}$ or $\{0, 1, 2\}$ (Chen et al., 2016). Assumptions 1 and 2 impose bounded eigenvalues on $\mathrm{Cov}(\boldsymbol{u}^{(i)})$ and $\mathrm{Cov}(\boldsymbol{z}^{(i)})$ as commonly done in the high-dimensional regression literature (Chen et al., 2016; Hao et al., 2018; Cai et al., 2019).

Assumption 3. *The dimensions p, q and sparsity s_j satisfy $\log p + \log q = \mathcal{O}(n^\delta)$ and $s_j = o(n^\delta)$ for $\delta \in [0, 1/6]$.*

Assumption 3 is a sparsity condition, allowing both $\log p$ and $\log q$ to grow at a polynomial order of n. Moreover, the number of nonzero entries s_j can also grow with n. This condition and $\delta \in [0, 1/6]$ are useful when establishing a restricted eigenvalue condition (Bickel et al., 2009) for $\boldsymbol{W}^\top \boldsymbol{W}/n$ and when bounding the stochastic term $\langle \boldsymbol{\epsilon}, \boldsymbol{W}\Delta \rangle$.

Let s_λ denote the number of nonzero entries in a candidate model such that $s_j < s_\lambda \leq n$. Given an s_λ satisfying the conditions in Theorem 1, we choose $\lambda_{\max}$ and $\lambda_{\min}$ to be the upper

and lower limits of λ_0 for each α, respectively corresponding to an empty model with no variables selected and a sparse model with s_λ variables selected.

Theorem 1. *Suppose that Assumptions 1-3 hold,* $s_\lambda(\log p + \log q) = \mathcal{O}(\sqrt{n})$ *and* $n \geq A_1\{s_j log(ep) + s_{j,g} log(eq/s_{j,g})\}$ *for some constant* $A_1 > 0$. *Then* $\hat{\boldsymbol{\beta}}_j$, $j \in [p]$, *in (9) with*

$$\lambda = C\sigma_{\epsilon_j}\sqrt{log(ep)/n + s_{j,g} log(eq/s_{j,g})/(ns_j)}, \quad \lambda_g = \sqrt{s_j/s_{j,g}}\lambda, \tag{12}$$

satisfies, with probability at least $1 - C_1 \exp[-C_2\{s_j log(ep) + s_{j,g} log(eq/s_{j,g})\}]$,

$$\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_2^2 \precsim \frac{\sigma_{\epsilon_j}^2}{n}\{s_j log(ep) + s_{j,g} log(eq/s_{j,g})\} + \frac{\sigma_{\epsilon_j}^2}{n}, \tag{13}$$

where C, C_1, *and* C_2 *are positive constants.*

Theorem 1 shows that our proposed estimator enjoys an improved ℓ_2 error bound over both the lasso and the group lasso under simultaneous sparsity. Specifically, given that the dimension of $\boldsymbol{\beta}_j$ is $(p-1)(q+1)$ and $s_{j,g} \leq s_j$, applying the regular lasso regularizer $\lambda\|\boldsymbol{\beta}_j\|_1$ alone would yield an error bound of $(s_j/n)\log(pq)$ (Negahban et al., 2012), which is slower than that in (13) when $\log p/\log q = o(1)$ and $s_{j,g}/s_j = o(1)$, corresponding to group sparsity. Moreover, when $p > n + 1$, estimating with the group lasso regularizer $\lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}$ alone, which excludes $(\boldsymbol{\beta}_j)_{(0)}$, is not feasible, because the dimension of the latter (i.e., $p-1$) exceeds n. If we utilize a group lasso regularizer $\lambda_g\|\boldsymbol{\beta}_j\|_{1,2}$ that includes $(\boldsymbol{\beta}_j)_{(0)}$, the estimator would have an ℓ_2 error bound of $(s_{j,g}/n)\log q + (s_{j,g}/n)p$ (Lounici et al., 2011), which is slower than that in (13) when $log q/p = o(1)$ and $s_j/s_{j,g} = o(p/\log p)$, corresponding to within-group sparsity. While the optimality of these error bounds warrants further investigation, the combined regularizer $\lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}$ may improve upon both the regular lasso and group lasso regularizers, when the true underlying coefficients are both element-wise and group sparse. In Theorem 1, the condition $s_\lambda(\log p + \log q) = \mathcal{O}(\sqrt{n})$ upper bounds the size

of candidate models, which in turn helps to bound $\langle \boldsymbol{\epsilon}, \boldsymbol{W}\Delta \rangle$. The parameter s_λ can be set to $c\sqrt{n}/\max\{\log p, \log q\})$ for some $c > 0$; by Assumption 3, it follows that $s_j = o(s_\lambda)$.

Some group lasso literature (Yuan and Lin, 2006; Lounici et al., 2011) noted that the grouped ℓ_1 penalty should compensate for the group size. It might be the case that λ_g is adjusted by $\sqrt{p-1}$, as each group in $\boldsymbol{\beta}_j$ is of size $p-1$. Indeed, with $(\boldsymbol{\beta}_j)_{(0)} = \mathbf{0}$ and no element-wise sparsity within the nonzero groups $\sqrt{s_j/s_{j,g}}$ becomes $\sqrt{p-1}$ in (12). Interestingly, our theoretical investigation reveals that, for the combined regularizer $\lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}$, $\lambda_g = \sqrt{s_j/s_{j,g}}\lambda$ suffices to suppress the noise term; see (S10).

We next show that our proposed sparse group lasso estimator achieves variable selection consistency under a mutual coherence condition. Let $\boldsymbol{\Sigma}_{\boldsymbol{W}} = \mathbb{E}(\boldsymbol{W}^\top \boldsymbol{W}/n)$.

Assumption 4 (Mutual coherence)**.** *Denote by* $\eta_j = 1 + \sqrt{s_j/s_{j,g}}$, $j \in [p]$. *We assume that for some positive constant* $c_0 > 6\phi_1/\phi_0$, *the covariance matrix* $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ *satisfies that*

$$\max_{k \neq l} |\boldsymbol{\Sigma}_{\boldsymbol{W}}(k,l)| \leq \frac{1}{c_0(1+8\eta_j)s_j},$$

where $\boldsymbol{\Sigma}_{\boldsymbol{W}}(k,l)$ *denotes the* (k,l)*th element of* $\boldsymbol{\Sigma}_{\boldsymbol{W}}$.

Assumption 4 specifies that the correlation between columns in $\boldsymbol{W}$ cannot be excessive.

Specifically, by the law of total probability, we write

$$\begin{aligned} \max_{k \neq l} |\boldsymbol{\Sigma}_{\boldsymbol{W}}(k,l)| &= \max_{\substack{l_1,l_2,l_3,l_4 \\ (l_1,l_2) \neq (l_3,l_4)}} \mathbb{E}\left\{\mathbb{E}\left(z_{l_1}^{(1)} z_{l_2}^{(1)} u_{l_3}^{(1)} u_{l_4}^{(1)} | \boldsymbol{u}^{(1)}\right)\right\} \\ &\leq \max_{l_1 \neq l_2} [\mathrm{Cov}(\boldsymbol{z}^{(1)})]_{l_1,l_2} \times \max_{l_3 \neq l_4} [\mathrm{Cov}(\boldsymbol{u}^{(1)})]_{l_3,l_4}. \end{aligned}$$

Hence, Assumption 4 holds when the correlations among $\{Z_j\}_{j\in[p]}$ and among $\{U_h\}_{h\in[q]}$ are not too large. A trivial sufficient condition is $\mathrm{Cov}(\boldsymbol{u}^{(1)}) = \boldsymbol{I}$. Furthermore, if $s_j = \mathcal{O}(1)$, Assumption 4 is satisfied when $\max_{l_1 \neq l_2}[\mathrm{Cov}(\boldsymbol{z}^{(1)})]_{l_1,l_2} \times \max_{l_3 \neq l_4}[\mathrm{Cov}(\boldsymbol{u}^{(1)})]_{l_3,l_4}$ is less than some positive constant. Similar correlation conditions include the neighborhood stability

condition (Meinshausen and Bühlmann, 2006) and the irrepresentability condition (Zhao and Yu, 2006); see Van De Geer and Bühlmann (2009) for a discussion of these relationships.

Theorem 2. *Suppose Assumptions 1-4 hold. If* $\log p \asymp \log q$ *and* $n \geq A_1\{s_j log(ep) + s_{j,g} log(eq/s_{j,g})\}$ *for some constant* $A_1 > 0$*, then for* $j \in [p]$*, the estimator* $\hat{\boldsymbol{\beta}}_j$ *in (9) with* λ *and* λ_g *as in (12) satisfies*

$$\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_\infty \leq \left\{3\phi_1\eta_j + \frac{18\phi_1^2(1+4\eta_j)^2\eta_j}{\phi_0(c_0\phi_0 - 2\phi_1)(1+8\eta_j)}\right\}\lambda, \tag{14}$$

with probability at least $1 - C_1' \exp(-C_2' log p)$*, where* C_1'*,* C_2' *are some positive constants. Define* $\hat{\mathcal{S}}_j = \left\{k : |(\hat{\boldsymbol{\beta}}_j)_k| > \left\{3\phi_1\eta_j + \frac{18\phi_1^2(1+4\eta_j)^2\eta_j}{\phi_0(c_0\phi_0-2\phi_1)(1+8\eta_j)}\right\}\lambda\right\}$*. In addition, if the minimum signal strength satisfies*

$$\min_{l\in\mathcal{S}} |(\boldsymbol{\beta}_j)_l| > 2\left\{3\phi_1\eta_j + \frac{18\phi_1^2(1+4\eta_j)^2\eta_j}{\phi_0(c_0\phi_0 - 2\phi_1)(1+8\eta_j)}\right\}\lambda, \tag{15}$$

we have that $\mathbb{P}(\hat{\mathcal{S}}_j = \mathcal{S}_j) \geq 1 - C_1' \exp(-C_2' log p)$*,* $j \in [p]$*.*

For the recovery of true signals in high-dimensional regression, minimum signal strength conditions such as (15) are necessary (Zhang, 2009). The condition of $\log p \asymp \log q$ allows p and q to grow at a polynomial rate relative to each other, ensuring a tighter bound on $\|\boldsymbol{W}^\top \boldsymbol{\epsilon}_j\|_\infty$; see Chen et al. (2016). Moreover, the selection consistency result in Theorem 2 holds for both estimates in (11) and (10), as (14) characterizes the relationship between the fitted values and the true parameters.

With $\hat{\boldsymbol{\beta}}_j$, a natural estimate of the variance $\sigma_{\epsilon_j}^2 = 1/\sigma^{jj}$ would be

$$\hat{\sigma}_{\epsilon_j}^2 = \frac{1}{n - \hat{s}_j}\|\boldsymbol{z}_j - \boldsymbol{W}\hat{\boldsymbol{\beta}}_j\|_2^2 = \frac{1}{n-\hat{s}_j}\boldsymbol{z}_j^\top\left(\boldsymbol{I}_{n\times n} - \mathcal{P}_{\hat{\mathcal{S}}_j}\right)\boldsymbol{z}_j, \tag{16}$$

where $\mathcal{P}_{\hat{\mathcal{S}}_j}$ is the projection matrix onto the column space of $\boldsymbol{W}_{\hat{\mathcal{S}}_j}$. The estimator in (16) can alternatively be written as $\hat{\sigma}_{\epsilon_j}^2 = \frac{1}{n-\hat{s}_j}(1-\gamma_n^2)\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$, where $\gamma_n^2 = \boldsymbol{\epsilon}^\top\mathcal{P}_{\hat{\mathcal{S}}_j}\boldsymbol{\epsilon}/\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$ represents the fraction of bias in $\hat{\sigma}_{\epsilon_j}^2$. Under conditions in Theorem 2 and using a result in (S10),

we get $\gamma_n^2 \asymp \sigma_{\epsilon_j}^2 \{s_j \log(ep) + s_{j,g} \log(eq/s_{j,g})\}/n$. Therefore, $\hat{\sigma}_{\epsilon_j}^2$ is consistent, provided that $\sigma_{\epsilon_j}^2 \{s_j \log(ep) + s_{j,g} \log(eq/s_{j,g})\}/n \to 0$.

5 Estimation with Unknown $\boldsymbol{\Gamma}$: a Two-Step Procedure

We present a two-step estimation procedure when $\boldsymbol{\Gamma}$ is unknown, followed by its theoretical properties. Assuming a sparse $\boldsymbol{\Gamma}$, Step 1 estimates $\boldsymbol{\Gamma}$ using an ℓ_1-penalized regression; in Step 2, we approximate each $\boldsymbol{z}^{(i)}$ with $\hat{\boldsymbol{z}}^{(i)} = \boldsymbol{x}^{(i)} - \hat{\boldsymbol{\Gamma}}\boldsymbol{u}^{(i)}$, where $\hat{\boldsymbol{\Gamma}}$ is estimated from the first step, and estimate $\boldsymbol{\beta}_j$ based on $\hat{\boldsymbol{z}}^{(1)}, \ldots, \hat{\boldsymbol{z}}^{(n)}$ by using the procedure described in Section 3. The two-step procedure is computationally feasible, particularly when both p and q are large, and has been considered for covariate-adjusted Gaussian graphical models (Cai et al., 2012; Yin and Li, 2013; Chen et al., 2016).

Step 1. Denote the covariate matrix by $\boldsymbol{H} = [\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(n)}]^\top$ and the sample of the jth variable by $\boldsymbol{x}_j = (x_j^{(1)}, \ldots, x_j^{(n)})^\top$. We first estimate $\boldsymbol{\Gamma}$ with

$$\hat{\boldsymbol{\gamma}}_j = \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^q} \frac{1}{2n} \|\boldsymbol{x}_j - \boldsymbol{H}\boldsymbol{\gamma}\|_2^2 + \lambda_1 \|\boldsymbol{\gamma}\|_1, \tag{17}$$

and denote the estimates by $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_p)^\top$.

Step 2. With $\hat{\boldsymbol{\Gamma}}$ obtained from Step 1, we calculate $\hat{\boldsymbol{z}}^{(i)} = \boldsymbol{x}^{(i)} - \hat{\boldsymbol{\Gamma}}\boldsymbol{u}^{(i)}$ and estimate $\boldsymbol{\beta}_j$ via

$$\hat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j \in \mathbb{R}^{(p-1)(q+1)}} \frac{1}{2n} \|\hat{\boldsymbol{z}}_j - \hat{\boldsymbol{W}}_{-j}\boldsymbol{\beta}_j\|_2^2 + \lambda \|\boldsymbol{\beta}_j\|_1 + \lambda_g \|\boldsymbol{\beta}_{j,-0}\|_{1,2}, \tag{18}$$

where $\hat{\boldsymbol{z}}_j = (\hat{z}_j^{(1)}, \ldots, \hat{z}_j^{(n)})^\top$ and $\hat{\boldsymbol{W}}_{-j} = [\hat{\boldsymbol{z}}_1 \odot \boldsymbol{u}_1, \ldots, \hat{\boldsymbol{z}}_1 \odot \boldsymbol{u}_q, \ldots, \hat{\boldsymbol{z}}_{j-1} \odot \boldsymbol{u}_q, \hat{\boldsymbol{z}}_{j+1} \odot \boldsymbol{u}_1 \ldots, \hat{\boldsymbol{z}}_p \odot \boldsymbol{u}_q]$, with $\hat{\boldsymbol{z}}_j$ and $\hat{\boldsymbol{W}}_{-j}$ respectively approximating $\boldsymbol{z}_j$ and $\boldsymbol{W}_{-j}$. When there is no ambiguity, we write $\hat{\boldsymbol{W}}$ without emphasizing its dependence on j. Both (17) and (18) are convex, and can be optimized efficiently (Simon et al., 2013; Vincent and Hansen, 2014).

Step 1 poses a regular lasso penalty on $\boldsymbol{\Gamma}$, as commonly done in the covariate-adjusted Gaussian graphical model literature (Rothman et al., 2010; Cai et al., 2012; Yin and Li,

2013; Chen et al., 2016). Note that each regression in Step 1 is of dimension q. When q is large, it may be necessary to consider a sparse group penalty (as in Step 2) that encourages $\mathbf{\Gamma}$ to be both element-wise and group sparse.

The two-step procedure involves three parameters λ_1, λ and λ_g that need to be tuned. We tune λ_1 in Step 1 via L-fold cross validation, and then tune λ and λ_g jointly in Step 2 using the same procedure as in Section 3. Sequential tuning is common (e.g. Danaher et al., 2014), and gives a good numerical performance in our experiments.

The theoretical development for the two-step procedure is challenging. Step 1 involves a regularized regression $\boldsymbol{x}_j = \boldsymbol{H}\boldsymbol{\gamma}_j + \boldsymbol{z}_j$ with heteroskedastic errors as the variance of $z_j^{(i)}$ is a function of $\boldsymbol{u}^{(i)}$. In Step 2, both the response vector $\hat{\boldsymbol{z}}_j$ and the design matrix $\hat{\boldsymbol{W}}$ inherit approximation errors from $\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}$, which further complicates the analysis of the sparse group lasso estimator. We show that $\hat{\boldsymbol{\beta}}_j$ in (18) can achieve the same convergence rate as that in Theorem 1 (i.e., the noiseless case), and thus enjoys the oracle property (as if $\mathbf{\Gamma}$ were known).

Assumption 5. *There exists a constant $M_2 > 0$ such that $\|\boldsymbol{\beta}_j\|_1 \leq \sigma_{\epsilon_j} M_2$ for all j. Moreover, there exists a constant $\phi_0' > 0$ such that $\lambda_{\max}(Cov(\boldsymbol{u}^{(i)})) \leq \phi_0'$.*

The boundedness of $\|\boldsymbol{\beta}_j\|_1$ controls the approximation errors in $\hat{\boldsymbol{z}}_j$ when analyzing the second step of the estimation procedure. Similar conditions have been considered in other two-step procedures (Cai et al., 2012; Chen et al., 2016). This assumption can be relaxed to allow M_2 to diverge, in which case M_2 appears in the convergence rates in Theorem 3 and 4.

Theorem 3. *Suppose that conditions in Theorem 1 and Assumption 5 are satisfied, $t = o(n^{1/3})$ and $n \geq A_2\{s_j log(ep) + s_{j,g} log(eq/s_{j,g})\}$ for some constant $A_2 > 0$. Let $\lambda_1 = 14\phi_2\sqrt{\tau_1 log\, q/n}$ for any $\tau_1 > 0$. The minimizer $\hat{\boldsymbol{\gamma}}_j$ in (17) satisfies*

$$\|\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\|_2^2 \precsim \frac{t log\, q}{n}, \quad \frac{1}{n}\|\hat{\boldsymbol{z}}_j - \boldsymbol{z}_j\|_2^2 \precsim \frac{t log\, q}{n}, \tag{19}$$

with probability at least $1-3\exp(-\tau_1 log\, q)$, $j \in [p]$. *The minimizer* $\hat{\boldsymbol{\beta}}_j$ *in* (18) *with* λ *and* λ_g *as in* (12) *satisfies with probability at least* $1-C_3\exp[C_4\{log\, p-(\tau_1-1)log\, q\}]$,

$$\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\|_2^2 \precsim \frac{\sigma_{\epsilon_j}^2}{n}\{s_j log(ep)+s_{j,g}log(eq/s_{j,g})\}+\frac{\sigma_{\epsilon_j}^2}{n}, \tag{20}$$

for some positive constants C_3, C_4.

When $\boldsymbol{\Gamma}$ is unknown, as opposed to the oracle regression equation $\boldsymbol{z}_j=\boldsymbol{W}\boldsymbol{\beta}_j+\boldsymbol{\epsilon}_j$, we only have access to the noisy equation $\hat{\boldsymbol{z}}_j=\hat{\boldsymbol{W}}\boldsymbol{\beta}_j+\boldsymbol{E}_j$, where $\boldsymbol{E}_j=\boldsymbol{\epsilon}_j+(\hat{\boldsymbol{z}}_j-\boldsymbol{z}_j)+(\boldsymbol{W}-\hat{\boldsymbol{W}})\boldsymbol{\beta}_j$. The condition $t=o(n^{1/3})$ is needed to control the errors from the estimation in the first step, which in turn controls the error $(\hat{\boldsymbol{z}}_j-\boldsymbol{z}_j)+(\boldsymbol{W}-\hat{\boldsymbol{W}})\boldsymbol{\beta}_j$. It is seen that the rate in (20) is the same as that derived in the oracle case (as if $\boldsymbol{\Gamma}$ were known) in (13).

Theorem 4. *Suppose that Assumptions 1-5 hold,* $\lambda_1=14\phi_2\sqrt{\tau_1 log\, q/n}$ *for any* $\tau_1>0$, $n\geq A_3\{s_j log(ep)+s_{j,g}log(eq/s_{j,g})\}$ *for some constant* $A_3>0$, $log\, p \asymp log\, q$ *and* $t=o(n^{1/3})$. *For* $j\in[p]$, *the sparse group lasso estimator* $\hat{\boldsymbol{\beta}}_j$ *in* (18) *with* λ *and* λ_g *as in* (12) *satisfies,*

$$\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\|_\infty \leq \frac{9}{2}\left\{\phi_1\eta_j+\frac{12\phi_1^2(1+3\eta_j)^2}{\phi_0(c_0\phi_0-6\phi_1)(1+8\eta_j)}\right\}\lambda, \tag{21}$$

with probability at least $1-C_5\exp[C_6\{log\, p-(\tau_1-1)log\, q\}]$, *for some positive constants* C_5, C_6. *Define* $\hat{\mathcal{S}}_j=\left\{k:|\hat{\boldsymbol{\beta}}_{j,k}|>\frac{9}{2}\left\{\phi_1\eta_j+\frac{12\phi_1^2(1+3\eta_j)^2}{\phi_0(c_0\phi_0-6\phi_1)(1+8\eta_j)}\right\}\lambda\right\}$. *If, in addition, the minimum signal strength satisfies*

$$\min_{k\in\mathcal{S}}|\boldsymbol{\beta}_{j,k}|>9\left\{\phi_1\eta_j+\frac{12\phi_1^2(1+3\eta_j)^2}{\phi_0(c_0\phi_0-6\phi_1)(1+8\eta_j)}\right\}\lambda, \tag{22}$$

then $\mathbb{P}(\hat{\mathcal{S}}_j=\mathcal{S}_j)\geq 1-C_5\exp[C_6\{log\, p-(\tau_1-1)log\, q\}]$, $j\in[p]$.

Compared to the minimal signal strength condition (15) in the noiseless case, the condition in (22) is slightly stronger. Similar to the case where $\boldsymbol{\Gamma}$ is known, (S36) leads to that $\hat{\sigma}_{\epsilon_j}^2=\frac{1}{n-\hat{s}_j}\|\hat{\boldsymbol{z}}_j-\hat{\boldsymbol{W}}\hat{\boldsymbol{\beta}}_j\|_2^2$ is consistent, if $\sigma_{\epsilon_j}^2\{s_j\log(ep)+s_{j,g}\log(eq/s_{j,g})\}/n\to 0$.

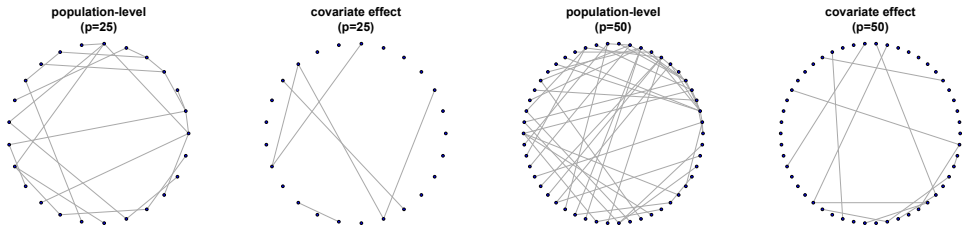

Figure 3: Graphs corresponding to the population-level effects and covariate effects with $p = 25$ or 50. When illustrating covariate effects, we randomly pick only one of the q_e effective covariates.

6 Simulations

We investigate the finite sample performance of our proposed method by comparing it with some competing solutions. Specifically, we evaluate three competing methods. We first consider our proposed Gaussian graphical model regression method defined in (18), referred to as `RegGMM` hereafter. We also consider a lasso estimator

$$\arg \min_{\boldsymbol{\beta}_j \in \mathbb{R}^{(p-1)(q+1)}} \frac{1}{2n} \|\hat{\boldsymbol{z}}_j - \hat{\boldsymbol{W}}_{-j} \boldsymbol{\beta}_j\|_2^2 + \lambda \|\boldsymbol{\beta}_j\|_1, \tag{23}$$

and a group lasso estimator

$$\arg \min_{\boldsymbol{\beta}_j \in \mathbb{R}^{(p-1)(q+1)}} \frac{1}{2n} \|\hat{\boldsymbol{z}}_j - \hat{\boldsymbol{W}}_{-j} \boldsymbol{\beta}_j\|_2^2 + \lambda_g (\|(\boldsymbol{\beta}_j)_{(0)}\|_1 + \sqrt{p-1} \|\boldsymbol{\beta}_{j,-0}\|_{1,2}), \tag{24}$$

where the total number of groups is $(p-1)+q$.

We simulate n samples $\{(\boldsymbol{u}^{(i)}, \boldsymbol{x}^{(i)}), i \in [n]\}$ from (4), where each sample has $\boldsymbol{x}^{(i)} \in \mathbb{R}^p$ (e.g., genes) and external covariate $\boldsymbol{u}^{(i)} \in \mathbb{R}^q$ (e.g., SNPs), including discrete and continuous covariates. Discrete covariates are generated from $\{0, 1\}$ with equal probabilities, and continuous covariates are generated from Uniform$[0, 1]$. For $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times q}$, we randomly set $s_{\boldsymbol{\Gamma}}$ of its entries to 0.25, and the rest to zero.

The population level network is assumed to follow a scale-free network model, with the degrees of nodes generated from a power-law distribution (Clauset et al., 2009) with

parameter 2.5. We randomly select q_e out of q covariates to have nonzero effects, and the graphs for these q_e covariates follow an Erdos-Renyi model with edge probability v_e; see the graph structures in Figure 3. We set $\sigma^{jj} = 1$ for $j \in [p]$. The initial nonzero coefficients β_{jkh} are generated from Uniform($[-0.5, -0.35] \cup [0.35, 0.5]$). For each j, we rescale $\{\beta_{jkh}\}_{k \neq j \in [p], h \in \{0\} \cup [q]}$ by dividing each entry by $\sum_{k \neq j \in [p], h \in \{0\} \cup [q]} |\beta_{jkh}|$. After rescaling, for each j, k and h, we use the average of β_{jkh} and β_{kjh} to fill the entries at jkh and kjh. This process results in symmetry with diagonal dominance and, thus, ensures the positive definiteness of the precision matrices. We set $s_{\mathbf{\Gamma}} = 125$, $q_e = 5$, $v_e = 0.01$, and consider $n = 200, 400$, $p = 25, 50$ and $q = 50, 100$, with 1,224 to 4,949 parameters to estimate.

For each simulation configuration, we generate 200 independent data sets, within each of which we randomly set half of the q covariates to be discrete and the rest continuous. Given $\boldsymbol{u}^{(i)}$, we are able to determine $\mathbf{\Omega}(\boldsymbol{u}^{(i)})$ and $\mathbf{\Sigma}(\boldsymbol{u}^{(i)})$; the ith sample $\boldsymbol{x}^{(i)}$ is generated from $\mathcal{N}(\mathbf{\Gamma}\boldsymbol{u}^{(i)}, \mathbf{\Sigma}(\boldsymbol{u}^{(i)}))$, $i \in [n]$. When comparing the estimates of $\boldsymbol{\beta}_j$'s obtained by the competing methods, we report the results after post-processing as in (10). For a fair comparison, tuning parameters in all of the methods are selected via 5-fold cross validation.

To evaluate the estimation accuracy, we report the estimation errors $\|\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}\|_F$ (the Frobenius norm) and $\sum_{j=1}^{p} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_2$, where $\hat{\boldsymbol{\beta}}_j$'s, with a slight overuse of notation, denote the estimates of $\boldsymbol{\beta}_j$'s obtained by various methods. Also reported is the average estimation error of the precision matrix defined as $\sum_{i=1}^{n} \|\hat{\mathbf{\Omega}}_i - \mathbf{\Omega}_i\|_{F,\text{off}}^2/n$, where $\mathbf{\Omega}_i = \boldsymbol{B}_0 + \sum_{h=1}^{q} \boldsymbol{B}_h \boldsymbol{u}_h^{(i)}$ and $\hat{\mathbf{\Omega}}_i$ is estimated from a given method. For the selection accuracy, we report the true positive rate (TPR) and false positive rate (FPR). Results for estimating the mean coefficient $\mathbf{\Gamma}$ are also good, and are given in Section S1.2 in the interest of space.

Table 1 reports the average criteria for estimating $\boldsymbol{\beta}_j$'s, with standard errors in the parentheses, over 200 data replications. It shows that the proposed `RegGMM` outperforms the competing methods in both estimation accuracy and selection accuracy for different sample

n	p, q	Method	$\text{TPR}_{\boldsymbol{\beta}}$	$\text{FPR}_{\boldsymbol{\beta}}$	Error of $\boldsymbol{\beta}$	Error of $\boldsymbol{\Omega}$
200	$p = 25$ $q = 50$	`RegGMM`	0.817 (0.004)	**0.003** (0.000)	**1.378** (0.006)	**2.011** (0.018)
		lasso	**0.820** (0.005)	**0.003** (0.000)	1.541 (0.007)	2.500 (0.022)
		group lasso	0.756 (0.004)	0.030 (0.002)	2.101 (0.005)	5.130 (0.033)
	$p = 25$ $q = 100$	`RegGMM`	**0.777** (0.005)	**0.002** (0.000)	**1.417** (0.006)	**2.147** (0.016)
		lasso	0.753 (0.005)	**0.002** (0.000)	1.622 (0.005)	2.791 (0.018)
		group lasso	0.721 (0.004)	0.013 (0.000)	2.103 (0.006)	5.023 (0.039)
	$p = 50$ $q = 50$	`RegGMM`	**0.624** (0.004)	0.003 (0.000)	**2.228** (0.005)	**5.036** (0.024)
		lasso	0.546 (0.005)	**0.002** (0.000)	2.396 (0.006)	5.827 (0.029)
		group lasso	0.579 (0.002)	0.030 (0.000)	4.219 (0.008)	26.652 (0.163)
	$p = 50$ $q = 100$	`RegGMM`	**0.597** (0.003)	**0.001** (0.000)	**2.292** (0.005)	**5.332** (0.020)
		lasso	0.473 (0.004)	**0.001** (0.000)	2.514 (0.005)	6.412 (0.023)
		group lasso	0.550 (0.002)	0.013 (0.000)	4.220 (0.008)	26.331 (0.210)
400	$p = 25$ $q = 50$	`RegGMM`	**0.983** (0.001)	**0.003** (0.000)	**0.907** (0.003)	**0.893** (0.006)
		lasso	**0.983** (0.001)	**0.003** (0.000)	1.016 (0.003)	1.118 (0.006)
		group lasso	0.928 (0.002)	0.033 (0.000)	1.555 (0.002)	2.556 (0.010)
	$p = 25$ $q = 100$	`RegGMM`	0.959 (0.002)	**0.001** (0.000)	**0.997** (0.003)	**1.069** (0.007)
		lasso	**0.960** (0.002)	0.002 (0.000)	1.113 (0.003)	1.329 (0.008)
		group lasso	0.900 (0.003)	0.016 (0.000)	1.616 (0.003)	2.754 (0.011)
	$p = 50$ $q = 50$	`RegGMM`	**0.900** (0.002)	**0.002** (0.000)	**1.632** (0.003)	**2.741** (0.011)
		lasso	0.892 (0.002)	**0.002** (0.000)	1.736 (0.003)	3.096 (0.012)
		group lasso	0.769 (0.002)	0.042 (0.000)	3.107 (0.004)	10.755 (0.033)
	$p = 50$ $q = 100$	`RegGMM`	**0.894** (0.002)	**0.001** (0.000)	**1.690** (0.003)	**2.935** (0.011)
		lasso	0.876 (0.002)	**0.001** (0.000)	1.826 (0.003)	3.419 (0.011)
		group lasso	0.714 (0.002)	0.018 (0.000)	3.148 (0.004)	10.984 (0.033)

Table 1: Estimation accuracy of $\boldsymbol{\beta}_j$'s in simulations with varying sample size n, network size p and covariate dimension q. The three methods are `RegGMM`, the lasso estimator in (23) and the group lasso estimator in (24). Marked in boldface are those achieving the best evaluation criteria in each setting.

sizes n, network sizes p and covariate dimensions q. This is consistent with our theoretical findings. Moreover, the estimation errors of `RegGMM` decrease as n increases, or as p and q decrease, confirming the theoretical results in Theorem 3. For additional analyses, we evaluate some higher dimensional cases by increasing p, q to 300 or 400, and present the

results in Section S1.3. In Section S1.4, we compare several benchmark solutions, including the standard Gaussian graphical model estimated using the neighborhood selection method (Meinshausen and Bühlmann, 2006) and the graphical lasso estimation method (Friedman et al., 2008), the conditional mean Gaussian graphical model (Cai et al., 2012) and the stratified Gaussian graphical model (Danaher et al., 2014).

Next, we present several ROC curves, plotting the true positive rate against the false positive rate across a fine grid of tuning parameters. In each curve, the true positive and false positive rates are averaged over p regressions and over 200 data replicates. Specifically, to compare various methods in the accuracy of selecting coefficients for the precision matrices, Figure 4 shows the ROC curves for `RegGMM` with $\alpha = 0.25, 0.5, 0.75$, lasso and the group lasso (grlasso). We also compare BIC and cross validation for selecting the optimal tuning parameter. The penalty term in BIC is $\log n \times \hat{s}_j$, with $\hat{s}_j$ being the number of nonzero elements in $\hat{\boldsymbol{\beta}}_j$. It appears that cross validation strikes a reasonable balance between the true and false positive rates, especially when p, q are large; `RegGMM` performs better than the lasso and group lasso estimator; and selection in the precision coefficient estimation is not overly sensitive to α, which characterizes the weight of the lasso penalty relative to the group lasso penalty.

Finally, we investigate the computation cost that may occur during the tuning process. Table 2 shows the five-fold cross validation computation time for one node (or gene) and all p nodes for a given α. The simulations were run on an iMac with a 3.6 GHz Intel Core i9 processor. As the number of parameters is $\mathcal{O}(p^2 q)$, the total computing cost is expected to be roughly quadratic in p and linear in q, as seen in Table 2. Our method enables a parallel implementation over the p node-wise regressions and the working values of α, in which case the computing time on each core is, for example, 16 seconds when $p = 50$ and $q = 100$.

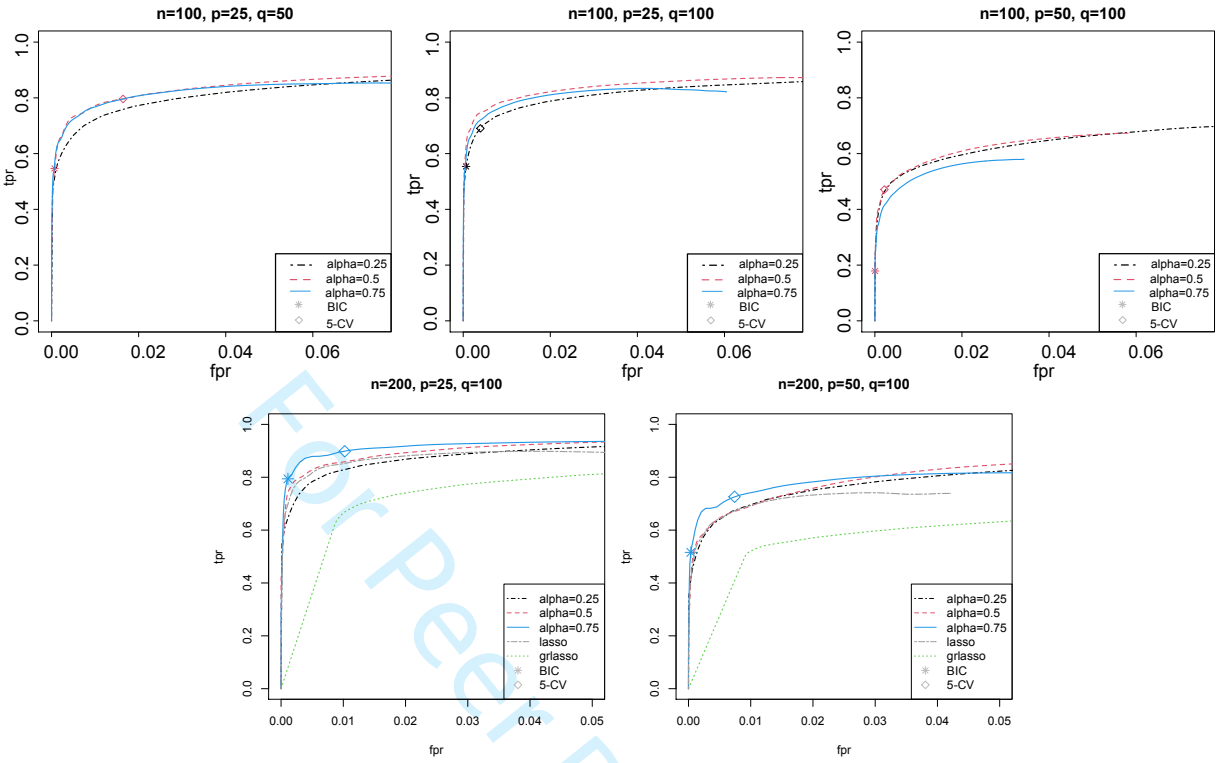

Figure 4: The ROC curves for `RegGMM` under $\alpha = 0.25, 0.5, 0.75$, lasso and the group lasso (grlasso). For `RegGMM`, the values selected by BIC and five-fold cross validation are marked on the curves.

n	(p, q)	computation time (s) one node	computation time (s) all nodes
200	(25,50)	2.819 (0.058)	70.475 (1.450)
	(25,100)	5.372 (0.116)	134.300 (2.900)
	(50,50)	8.343 (0.118)	140.950 (5.901)
	(50,100)	15.550 (0.231)	777.500 (11.552)

Table 2: Five-fold cross validation computation time for one node (or gene) and all p nodes for a given α.

7 Co-expression QTL Analysis

Our application focuses on glioblastoma multiforme (GBM), the most aggressive and fatal subtype of brain cancer (Bleeker et al., 2012), as featured in the REMBRANDT trial (GSE108476) with a subcohort of $n = 178$ GBM patients. Since existing therapies remain

largely ineffective (Bleeker et al., 2012), it is imperative to explore more effective treatment, such as new gene therapies (Kwiatkowska et al., 2013). Understanding the molecular underpinning of the disease is the key. In the study, all of these 178 patients had undergone microarray and single nucleotide polymorphism (SNP) chip profiling, with both gene expression and SNP data available for analysis. Specifically, the extracted RNA from each tumor sample was processed using microarrays with 23,521 genes assayed on each array. Genomic DNA from each sample was hybridized to SNP chips, which covers 116,204 SNP loci with a mean intermarker distance of 23.6kb. The raw data were pre-processed and normalized using standard pipelines; see Gusev et al. (2018) for more details.

We study a set of $p = 73$ genes (response variables) that belong to the human glioma pathway in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000); see Figure S1. The covariates include local SNPs (i.e., SNPs that fall within 2kb upstream and 0.5kb downstream of the gene) residing near these 73 genes, resulting in a total of 118 SNPs. SNPs are coded with "0" indicating homozygous in the major allele and "1" otherwise. For each patient, age and gender are included in analysis. Consequently, there are $q = 120$ covariates, bringing a total of $73 \times 36 \times 121 = 317,988$ model parameters (including intercepts). Our main objective is to recover both the population-level and subject-level gene networks, and to examine if and how age, gender and SNPs modulate the subject-level networks.

We have evaluated several benchmark methods in Section S1.4 of the Supplementary Materials; however, these methods are not designed to and cannot detect eQTL variants. Therefore, we have elected to apply the proposed two-step procedure in Section 5 to this data set. It is common in penalized regressions to standardize predictors to ensure they be on the same scale (Tibshirani, 1997). For example, the covariates in the model are standardized to have mean zero and variance one. The scheme does not alter interpretations of the model;

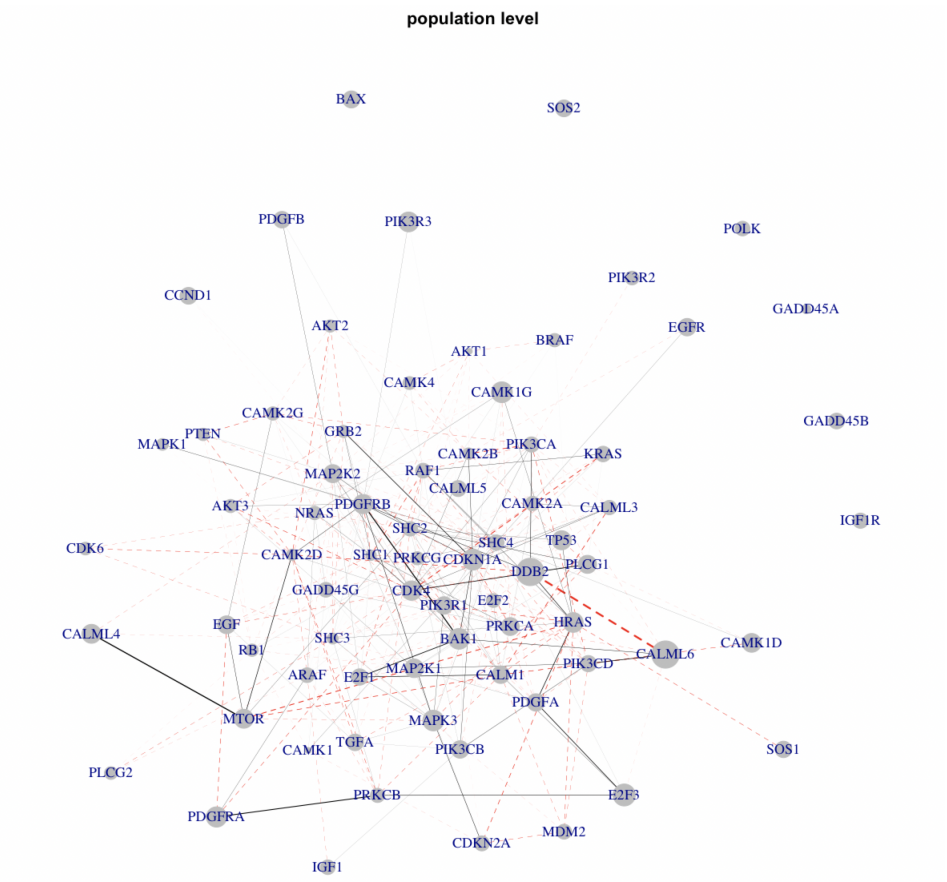

Figure 5: The graph corresponding to the population-level effect. The node sizes are proportional to mean expression levels and the edge weights are proportional to $\hat{\boldsymbol{B}}_0$. Edges with positive (negative) effects on partial correlations are shown in red dashed (black solid) lines.

see discussions in Section S1.5. Tuning parameters in both steps of the estimation procedure are selected via 5-fold cross validation, and post-processing, as in (10), generates the final estimates. Out of the 120 covariates considered, 9 SNPs are estimated to have nonzero effects on the network.

We first examine the population level network. Most of the well-connected genes in Figure 5 are known to be associated with cancer. For example, PIK3CA is a protein coding gene and is one of the most highly mutated oncogenes identified in human cancers (Samuels and Velculescu, 2004); mutations in the PIK3CA gene are found in many types of cancer, including cancer of the brain, breast, ovary, lung, colon and stomach (Samuels and Velculescu,

name	genes	references
PI3K/ AKT/MTOR signaling pathway	PIK3CA, PIK3CB, PIK3CD, PIK3R3, PTEN, AKT1, AKT2, AKT3, MTOR, IGF1, PRKCA	Network et al. (2008)
Ras-Raf-MEK-ERK signaling pathway	EGF, EGFR, GRB2, SOS1, SOS2, IGF1 SHC1, SHC2, SHC3, SCH4 MAPK1, MAPK3, MAP2K1, MAP2K2 HRAS, KRAS, NRAS, RAF1, ARAF, BRAF, PRKCA	Brennan et al. (2013)
calcium (Ca^{+2}) signaling pathway	CALM1,CALML3, CALML4, CALML5, CALML6, CAMK1,CAMK4, CAMK1D, CAMK1G,CAMK2A, CAMK2B, CAMK2D,CAMK2G, PRKCA	Maklad et al. (2019)
p53 signaling pathway	TP53, MDM2, DDB2, PTEN, IGF1 CDK4, CDK6, CDKN1A, CDKN2A	Network et al. (2008)

Table 3: Pathways and genes involves in each pathway.

2004). The PIK3CA gene is a part of the PI3K/AKT/MTOR signaling pathway, which is one of the core pathways in human GBM and other types of cancer (Network et al., 2008). TP53 is also a highly connected gene in the estimated network. This gene encodes a tumor suppressor protein containing transcriptional activation, and is the most frequently mutated gene in human cancer; the P53 signaling pathway is also one of the core pathways in human GBM and other types of cancer (Network et al., 2008). In Figure 5, we can identify several core pathways in human GBM including the PI3K/ AKT/MTOR, Ras-Raf-MEK-ERK, calcium and p53 signaling pathways; see Table 3 for genes included in each pathway. These findings are in agreement to the existing literature on GBM genes and pathways (Network et al., 2008; Brennan et al., 2013; Maklad et al., 2019).

We next examine the covariate effects on the network. Identified are nine co-expression QTLs, namely, `rs6701524`, `rs10509346`, `rs10492975`, `rs723211`, `rs1347069`, `rs473698`,

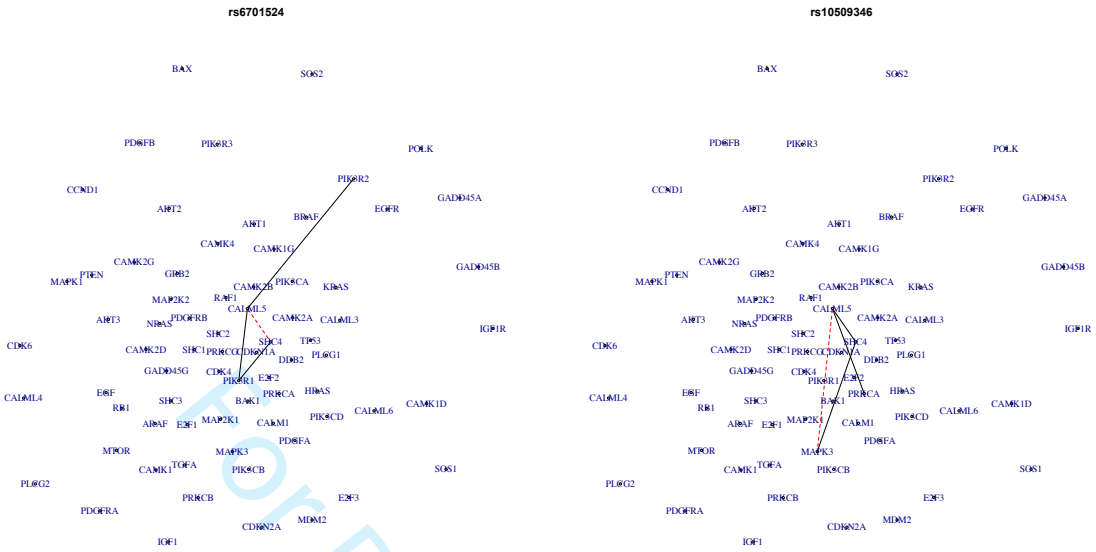

Figure 6: Graphs depending on each covariate (i.e., different SNPs). Edges that have positive (negative) effects on partial correlations are shown in red dashed (black solid) lines.

rs4118334, rs882664 and rs1267622. The network effects of rs6701524 are shown in Figure 6 (left panel). This SNP, residing in MTOR, is found to affect CALML5's co-expression with PIK3R1, and also with PIK3R2 and SHC4. This is an interesting finding as PI3K/MTOR is a key pathway in GBM development and progression, and inhibition of PI3K/MTOR signaling was found effective in increasing survival with GBM tumor (Batsios et al., 2019). This co-expression QTL can potentially regulate the co-expressions of CALML5, PIK3R1, PIK3R2, MTOR, and play an important role in activating the PI3K/MTOR pathway.

Shown in Figure 6 (right panel) are the network effects of rs10509346, a variant of CAMK2G. The figure indicates that this SNP affects the co-expressions of CAMK2G with genes in the Ras-Raf-MEK-ERK pathway. This agrees to the finding that the Ras-Raf-MEK-ERK pathway is modulated by Ca^{+2} and calmodulin (Agell et al., 2002). Moreover, based on our analysis, rs10492975 regulates the co-expressions of CALML5, PIK3R2 and CAMK1; rs723211 is associated with the co-expressions of CALML5 and other genes; rs1347069 in-

fluences the co-expressions of SHC4 and CDKN2A; `rs473698` may modify the co-expressions of PRKCG and CAMK1; `rs4118334` modulates the co-expressions of SHC2 and CAMK1; `rs882664` influences the co-expressions of PRKCA and CAMK1; `rs1267622` may alter the co-expressions of SHC3 and RAF1. See details in Table S6. As co-expression QTL identification has sparked recent interest, these findings warrant more in-depth investigation.

8 Discussion

As the off-diagonal entries in the precision matrix $\boldsymbol{\Omega}(\boldsymbol{u})$ are covariate dependent, a natural sufficient condition for positive definiteness, derived from diagonal dominance, is $\max(1, \|\boldsymbol{u}\|_\infty)\|\boldsymbol{\beta}_j\|_1 < 1$. With Assumption 1 stipulating $|u_h^{(i)}| \leq M$, positive definiteness holds when $\|\boldsymbol{\beta}_j\|_1 < 1/\max(1, M)$, $j \in [p]$. Assuming $u_h \in [-1, 1]$ (if not, rescale first), this sufficient condition can be simplified to $\|\boldsymbol{\beta}_j\|_1 < 1$ for all j (note that $\boldsymbol{\beta}_j$ is sparse), suggesting that, to satisfy diagonal dominance, the "effect sizes" of $\boldsymbol{u}$ (i.e., $\|\boldsymbol{\beta}_j\|_1$) on partial correlations cannot be too large. If the true covariance/precision matrix is positive definite, it then follows from Theorems 1-4 that the estimated precision is asymptotically positive definite. However, for finite sample cases, it may be desirable to ensure the positive definiteness of the final estimator. A post-hoc rescaling procedure seems to work well as in Section S1.6.

In our estimation procedure, we could estimate $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ jointly by combining p loss functions, and minimizing $\sum_{j=1}^p \ell_j(\boldsymbol{\beta}_j|\mathcal{D}) + \lambda \sum_j \|\boldsymbol{\beta}_j\|_1 + \lambda_g \sum_j \|\boldsymbol{\beta}_{j,-0}\|_{1,2}$. It would have the benefit of preserving symmetry by restricting $[\boldsymbol{B}_h]_{jk} = [\boldsymbol{B}_h]_{kj}$ and possibly permitting additional dimension reduction via low-rankness (Zhang and Xia, 2018). However, this would be much more computationally intensive than (9) by optimizing with respect to $\mathcal{O}(p^2 q)$ parameters simultaneously. Moreoever, we can modify our method to accommodate the

hierarchy between main effects and interaction terms by re-organizing $\boldsymbol{\beta}_j$ as

$$\boldsymbol{\beta}_j = (\overbrace{\underbrace{\beta_{j10}}_{\text{main effect}}, \underbrace{\beta_{j11}, \ldots, \beta_{j1q}}_{\text{interactions}}}^{\text{group 1}}, \ldots, \overbrace{\underbrace{\beta_{jp0}}_{\text{main effect}}, \underbrace{\beta_{jp1} \ldots, \beta_{jpq}}_{\text{interactions}}}^{\text{group } p}), \tag{25}$$

and imposing a modified sparse group lasso penalty $\lambda\|\boldsymbol{\beta}_j^{-0}\|_0 + \lambda_g\|\boldsymbol{\beta}_j\|_{1,2}$, where $\boldsymbol{\beta}_j^{-0}$ is $\boldsymbol{\beta}_j$ after leaving out the main effects $\{\beta_{j10} \ldots, \beta_{jp0}\}$ and groups in $\|\boldsymbol{\beta}_j\|_{1,2}$ are as defined in (25). The penalty is designed in such a way that the element-wise sparsity is not imposed on the main effects, and interactions, if selected, will enter the model with non-zero main effects; a similar regularizer was adopted by (She et al., 2018) for penalized interaction models. With slight modifications, our established theoretical framework can still be used to study the theoretical properties of this modified regularizer.

In model (4), we had assumed σ^{jj} to be free of $\boldsymbol{u}$. Our empirical investigations show that our method is not sensitive to this assumption (see simulations in Section S1.1). It is possible to further extend our framework to allow the residual variances in (5) (or correspondingly, the diagonal elements σ^{jj}'s) to depend on the covariate $\boldsymbol{u}$. To proceed, we consider $\sigma^{jj}(\boldsymbol{u}) = g(\boldsymbol{\nu}_j^\top \boldsymbol{u})$, where $\boldsymbol{\nu}_j$ is the vector of unknown coefficients; a viable choice is $g(x) = \exp(x)$. In this case, the node-wise regression representation, by scaling the response by $g(\boldsymbol{\nu}_j^\top \boldsymbol{u})$, may be reformulated as

$$Z_j \times g(\boldsymbol{\nu}_j^\top \boldsymbol{u}) = \sum_{k\neq j}^{p} \theta_{jk0} Z_k + \sum_{k\neq j}^{p}\sum_{h=1}^{q} \theta_{jkh} u_h Z_k + \tilde{\epsilon}_j, \quad \text{Var}(\tilde{\epsilon}_j) = g(\boldsymbol{\nu}_j^\top \boldsymbol{u}) \tag{26}$$

where $\theta_{jkh} = -[\boldsymbol{B}_h]_{jk}$. To estimate $\boldsymbol{\nu}_j$ and $\boldsymbol{\theta}_j = (\theta_{j10}, \ldots, \theta_{jp0}, \ldots, \theta_{j1q}, \ldots, \theta_{jpq})$, we may consider the following loss function

$$\ell_j(\boldsymbol{\nu}_j, \boldsymbol{\theta}_j | \mathcal{D}) = \frac{1}{2n}\sum_{i=1}^{n} \|\boldsymbol{z}_j^{(i)} \times g(\boldsymbol{\nu}_j^\top \boldsymbol{u}^{(i)}) - \boldsymbol{w}_i \boldsymbol{\theta}_j\|_2^2,$$

where $\mathcal{D}$ denotes the observed data, $\boldsymbol{z}_j^{(i)}$ collects the samples of the jth variable and $\boldsymbol{w}_i$ is the ith row of the design matrix $\boldsymbol{W}_{-j}$; see their definitions in Section 3. Through $\ell_j(\boldsymbol{\nu}_j, \boldsymbol{\theta}_j | \mathcal{D})$,

$\boldsymbol{\nu}_j$ and $\boldsymbol{\theta}_j$ can be estimated with sparse penalties. This new iterative estimation procedure is more computationally demanding and requires a new theoretical analysis. We leave its full investigation as future research.

Acknowledgements

We are grateful to the Editor, the AE and three anonymous referees for their insightful comments that have substantially improved the quality and the presentation of the manuscript.

References

Agell, N., Bachs, O., Rocamora, N., and Villalonga, P. (2002), "Modulation of the Ras/Raf/MEK/ERK pathway by Ca2+, and calmodulin," *Cellular signalling*, 14, 649–654.

Batsios, G., Viswanath, P., Subramani, E., Najac, C., Gillespie, A. M., Santos, R. D., Molloy, A. R., Pieper, R. O., and Ronen, S. M. (2019), "PI3K/mTOR inhibition of IDH1 mutant glioma leads to reduced 2HG production that is associated with increased survival," *Scientific Reports*, 9, 1–15.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, 37, 1705–1732.

Bleeker, F. E., Molenaar, R. J., and Leenstra, S. (2012), "Recent advances in the molecular understanding of glioblastoma," *Journal of Neuro-oncology*, 108, 11–27.

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., et al. (2013), "The somatic genomic landscape of glioblastoma," *Cell*, 155, 462–477.

Brynedal, B., Choi, J., Raj, T., Bjornson, R., Stranger, B. E., Neale, B. M., Voight, B. F., and Cotsapas, C. (2017), "Large-scale trans-eQTLs affect hundreds of transcripts and mediate patterns of transcriptional co-regulation," *The American Journal of Human Genetics*, 100, 581–591.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2012), "Covariate-adjusted precision matrix estimation with an application in genetical genomics," *Biometrika*, 100, 139–156.

Cai, T. T., Zhang, A., and Zhou, Y. (2019), "Sparse Group Lasso: Optimal Sample Complexity, Convergence Rate, and Statistical Inference," in *arXiv preprint arXiv:1909.09851*.

Chen, M., Ren, Z., Zhao, H., and Zhou, H. (2016), "Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model," *Journal of the American Statistical Association*, 111, 394–406.

Cheng, J., Levina, E., Wang, P., and Zhu, J. (2014), "A sparse Ising model with covariates," *Biometrics*, 70, 943–953.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009), "Power-law distributions in empirical data," *SIAM review*, 51, 661–703.

Danaher, P., Wang, P., and Witten, D. M. (2014), "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 373–397.

Fan, J., Feng, Y., and Wu, Y. (2009), "Network exploration via the adaptive LASSO and SCAD penalties," *The annals of applied statistics*, 3, 521.

Fehrmann, R. S., Jansen, R. C., Veldink, J. H., Westra, H.-J., Arends, D., Bonder, M. J., Fu, J., Deelen, P., Groen, H. J., Smolonska, A., et al. (2011), "Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA," *PLoS genetics*, 7, e1002197.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., Feng, J., Liu, R., Diao, L., Guo, A.-Y., et al. (2018), "PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types," *Nucleic acids research*, 46, D971–D976.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011), "Joint estimation of multiple graphical models," *Biometrika*, 98, 1–15.

Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J.-C., Fine, H., and Madhavan, S. (2018), "The REMBRANDT study, a large collection of genomic data from brain cancer patients," *Scientific Data*, 5, 180158.

Hao, N., Feng, Y., and Zhang, H. H. (2018), "Model selection for high-dimensional quadratic regression via regularization," *Journal of the American Statistical Association*, 113, 615–625.

Kanehisa, M. and Goto, S. (2000), "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, 28, 27–30.

Kolar, M., Parikh, A. P., and Xing, E. P. (2010), "On sparse nonparametric conditional covariance selection," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 559–566.

Kolberg, L., Kerimov, N., Peterson, H., and Alasoo, K. (2020), "Co-expression analysis reveals interpretable gene modules controlled by trans-acting genetic variants," *Elife*, 9, e58705.

Kwiatkowska, A., Nandhu, M. S., Behera, P., Chiocca, E. A., and Viapiano, M. S. (2013), "Strategies in gene therapy for glioblastoma," *Cancers*, 5, 1271–1305.

Lauritzen, S. L. (1996), *Graphical Models*, vol. 17, Clarendon Press.

Lee, W. and Liu, Y. (2012), "Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood," *Journal of multivariate analysis*, 111, 241–255.

Li, B., Chun, H., and Zhao, H. (2012), "Sparse estimation of conditional graphical models with application to gene networks," *Journal of the American Statistical Association*, 107, 152–167.

Li, B. and Solea, E. (2018), "A nonparametric graphical model for functional data with application to brain networks based on fMRI," *Journal of the American Statistical Association*, 113, 1637–1655.

Li, Y., Nan, B., and Zhu, J. (2015), "Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure," *Biometrics*, 71, 354–363.

Lin, J., Basu, S., Banerjee, M., and Michailidis, G. (2016), "Penalized Maximum Likelihood Estimation of Multi-layered Gaussian Graphical Models," *Journal of Machine Learning Research*, 17, 1–51.

Liu, H., Chen, X., Wasserman, L., and Lafferty, J. D. (2010), "Graph-valued regression," in *Advances in Neural Information Processing Systems*, pp. 1423–1431.

Lounici, K., Pontil, M., Van De Geer, S., and Tsybakov, A. B. (2011), "Oracle inequalities and optimal inference under group sparsity," *The Annals of Statistics*, 39, 2164–2204.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004), "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, 431, 308–312.

Maklad, A., Sharma, A., and Azimi, I. (2019), "Calcium signaling in brain cancers: roles and therapeutic targeting," *Cancers*, 11, 145.

Meinshausen, N. and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, 34, 1436–1462.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, 27, 538–557.

Network, C. G. A. R. et al. (2008), "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, 455, 1061.

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2019), "Bayesian graphical regression," *Journal of the American Statistical Association*, 114, 184–197.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial correlation estimation by joint sparse regression models," *Journal of the American Statistical Association*, 104, 735–746.

Rothman, A. J., Levina, E., and Zhu, J. (2010), "Sparse multivariate regression with covariance estimation," *Journal of Computational and Graphical Statistics*, 19, 947–962.

Samuels, Y. and Velculescu, V. E. (2004), "Oncogenic mutations of PIK3CA in human cancers," *Cell Cycle*, 3, 1221–1224.

She, Y., Wang, Z., and Jiang, H. (2018), "Group regularized estimation under structural hierarchy," *Journal of the American Statistical Association*, 113, 445–454.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245.

Tibshirani, R. (1997), "The lasso method for variable selection in the Cox model," *Statistics in medicine*, 16, 385–395.

Van De Geer, S. A. and Bühlmann, P. (2009), "On the conditions used to prove oracle results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392.

van der Wijst, M. G., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., and Franke, L. (2018a), "Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs," *Nature Genetics*, 50, 493–497.

van der Wijst, M. G., de Vries, D. H., Brugge, H., Westra, H.-J., and Franke, L. (2018b), "An integrative approach for building personalized gene regulatory networks for precision medicine," *Genome Medicine*, 10, 96.

Vincent, M. and Hansen, N. R. (2014), "Sparse group lasso and high dimensional multinomial classification," *Computational Statistics & Data Analysis*, 71, 771–786.

Wang, L., Zheng, W., Zhao, H., and Deng, M. (2013), "Statistical analysis reveals co-expression patterns of many pairs of genes in yeast are jointly regulated by interacting loci," *PLoS Genet*, 9, e1003414.

Wang, Y., Joseph, S. J., Liu, X., Kelley, M., and Rekaya, R. (2012), "SNPxGE2: a database for human SNP–coexpression associations," *Bioinformatics*, 28, 403–410.

Wu, Y. and Wang, L. (2020), "A survey of tuning parameter selection for high-dimensional regression," *Annual Review of Statistics and Its Application*, 7, 209–226.

Xie, S., Li, X., McColgan, P., Scahill, R. I., Zeng, D., and Wang, Y. (2020), "Identifying disease-associated biomarker network features through conditional graphical model," *Biometrics*, 76, 995–1006.

Yin, J. and Li, H. (2011), "A sparse conditional gaussian graphical model for analysis of genetical genomics data," *The Annals of Applied Statistics*, 5, 2630.

— (2013), "Adjusting for high-dimensional covariates in sparse precision matrix estimation by ℓ_1-penalization," *Journal of Multivariate Analysis*, 116, 365–381.

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

— (2007), "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94, 19–35.

Zhang, A. and Xia, D. (2018), "Tensor SVD: Statistical and computational limits," *IEEE Transactions on Information Theory*, 64, 7311–7338.

Zhang, J., Sun, W. W., and Li, L. (2019), "Mixed-Effect Time-Varying Network Model and Application in Brain Connectivity Analysis," *Journal of the American Statistical Association*, 1–15.

Zhang, T. (2009), "Some sharp performance bounds for least squares regression with l1 regularization," *The Annals of Statistics*, 37, 2109–2144.

Zhao, P. and Yu, B. (2006), "On model selection consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

Supplementary Materials for "High Dimensional Gaussian Graphical Regression Models with Covariates"

S1 Additional Numerical Results

S1.1 Sensitivity analysis

We conduct a sensitivity analysis to examine the performance of our method when σ^{jj} is covariate dependent. Specifically, under the same setting as in Section 6, we set $\sigma^{jj}(\boldsymbol{u}) = 1 + \sum_{h=1}^{q} \beta_{jjh} u_h$. If u_h is an effective covariate with a nonzero effect on the graph, we set s_σ proportion of $\{\beta_{jjh}\}_{j\in[p]}$ to σ_0, where σ_0 reflects the strength of covariate dependence for $\sigma^{jj}(\boldsymbol{u})$, and set the rest of β_{jjh}'s to zero. We let $n = 200$, $p = 25$ and $s_\sigma = 0.1$. The tuning parameters are selected using cross validation. Table S1 reports the average evaluation criteria over 200 data replicates over $q = 25, 100$ and $\sigma_0 = 0, 0.1, 0.2$. Note that $\sigma_0 = 0$ corresponds to the case that σ^{jj}'s are not covariate dependent. When $\sigma_0 = 0.2$, we have $\sigma^{jj}(\boldsymbol{u}) \in [1, 2]$. Reported criteria are estimation errors of the mean $\boldsymbol{\mu}_{\text{error}} = \sum_{i=1}^{n} \|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|^2/n$ and the precision matrix $\boldsymbol{\Omega}_{\text{error}} = \sum_{i=1}^{n} \|\hat{\boldsymbol{\Omega}}_i - \boldsymbol{\Omega}_i\|_{F,\text{off}}^2/n$, where $\|\cdot\|_{F,\text{off}}$ denotes the off-diagonal Frobenius norm and $(\boldsymbol{\mu}_i, \boldsymbol{\Omega}_i)$ and $(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Omega}}_i)$ are the true and estimated values, respectively, the true positive rate (TPR) and false positive rate (FPR) in estimating the precision coefficient $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$. As seen from Table S1, under this misspecified setting with covariate dependent residual variances, our method still gives a reasonable performance.

S1.2 Estimation accuracy of $\boldsymbol{\Gamma}$

Table S2 shows that the first step of our estimation procedure achieves good performance, and the estimation error of $\boldsymbol{\Gamma}$ decreases as n increases, or as p and q decrease. Such observations conform to Theorem 3.

		$\boldsymbol{\mu}_{\text{error}}$	$\boldsymbol{\Omega}_{\text{error}}$	$\text{TPR}_{\boldsymbol{\beta}}$	$\text{FPR}_{\boldsymbol{\beta}}$
$q = 50$	$\sigma_0 = 0$	3.041	2.011	0.817	0.003
		(0.018)	(0.018)	(0.004)	(0.000)
	$\sigma_0 = 0.1$	2.871	2.167	0.700	0.004
		(0.022)	(0.015)	(0.004)	(0.000)
	$\sigma_0 = 0.2$	2.842	2.663	0.645	0.003
		(0.018)	(0.020)	(0.004)	(0.000)
$q = 100$	$\sigma_0 = 0$	3.603	2.147	0.777	0.002
		(0.019)	(0.016)	(0.005)	(0.000)
	$\sigma_0 = 0.1$	3.465	2.382	0.668	0.002
		(0.022)	(0.016)	(0.004)	(0.000)
	$\sigma_0 = 0.2$	3.370	2.878	0.627	0.002
		(0.023)	(0.018)	(0.003)	(0.000)

Table S1: Average evaluation criteria with varying covariate dimension q and covariate dependence parameter σ_0 with standard errors in parenthesis.

n	(p, q)	$\text{TPR}_{\boldsymbol{\Gamma}}$	$\text{FPR}_{\boldsymbol{\Gamma}}$	Error of $\boldsymbol{\Gamma}$
200	(25,50)	0.872 (0.003)	0.156 (0.002)	1.811 (0.006)
	(25,100)	0.804 (0.003)	0.095 (0.001)	1.980 (0.006)
	(50,50)	0.867 (0.003)	0.110 (0.001)	1.938 (0.006)
	(50,100)	0.789 (0.003)	0.072 (0.000)	2.162 (0.005)
400	(25,50)	0.996 (0.000)	0.172 (0.001)	1.273 (0.004)
	(25,100)	0.986 (0.001)	0.113 (0.001)	1.386 (0.005)
	(50,50)	0.992 (0.001)	0.122 (0.000)	1.351 (0.004)
	(50,100)	0.987 (0.001)	0.082 (0.001)	1.514 (0.004)

Table S2: Estimation accuracy of $\boldsymbol{\Gamma}$ in simulations with various sample sizes n, network sizes p and covariate dimensions q.

S1.3 Higher dimensional cases

We have furthered increased (p, q) to (25,400) and (300, 300). The simulation results over 200 data replicates are reported in Table S3. For both settings in Table S3, the mean coefficient $\boldsymbol{\Gamma}$ are set following Section 6. The non-sparse entries in $\{\boldsymbol{B}_h\}_{h=0}^{q}$ for $(p, q) = (25, 400)$ and $(p, q) = (300, 300)$ are set to be the same as $p = 25$ and $p = 50$ in Section 6, respectively. These settings involve an extremely large number of parameters; for example, with $(p, q) =$

$(300, 300)$, the total number of precision parameters is almost 27 million (or 26,999,700 precisely). It appears that our proposed method still achieves a reasonable performance in these high dimensional settings.

n	$(p,\ q)$	Error of $\boldsymbol{\Gamma}$	Error of $\boldsymbol{\beta}$	$\text{TPR}_{\boldsymbol{\beta}}$	$\text{FPR}_{\boldsymbol{\beta}}$
200	(25, 400)	2.195 (0.005)	1.534 (0.005)	0.721 (0.004)	0.0003 (0.0000)
	(300, 300)	2.525 (0.007)	2.461 (0.005)	0.527 (0.004)	0.0000 (0.0000)

Table S3: Estimation and selection accuracy of `RegGMM` with larger network sizes p and covariate dimensions q.

S1.4 Comparison with benchmark methods

We compare our proposed method with several benchmark solutions, including the standard i.i.d Gaussian graphical model estimated by the neighborhood selection method (Meinshausen and Bühlmann, 2006), referred to as `MB`, and by the graphical lasso estimation method (Friedman et al., 2008), referred to as `glasso`, the conditional mean Gaussian graphical model (Cai et al., 2012), referred to as `cmGMM`, and the stratified Gaussian graphical model (Danaher et al., 2014), referred to as `stGMM`. Our evaluation criteria include the average estimation error of the mean defined as $\boldsymbol{\mu}_{\text{error}} = \sum_{i=1}^{n} \|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|^2/n$, where $\boldsymbol{\mu}_i = \boldsymbol{\Gamma}\boldsymbol{u}^{(i)}$ and $\hat{\boldsymbol{\mu}}_i$ is estimated from a given method, average estimation error of the precision matrix defined as $\boldsymbol{\Omega}_{\text{error}} = \sum_{i=1}^{n} \|\hat{\boldsymbol{\Omega}}_i - \boldsymbol{\Omega}_i\|_{F,\text{off}}^2/n$, where $\boldsymbol{\Omega}_i = \boldsymbol{B}_0 + \sum_{h=1}^{q} \boldsymbol{B}_h \boldsymbol{u}_h^{(i)}$ and $\hat{\boldsymbol{\Omega}}_i$ is estimated from a given method, and the average selection error of the precision matrix $\text{TPR}_{\boldsymbol{\Omega}} = \sum_{i=1}^{n} \text{TPR}(\hat{\boldsymbol{\Omega}}_i)/n$ and $\text{FPR}_{\boldsymbol{\Omega}} = \sum_{i=1}^{n} \text{FPR}(\hat{\boldsymbol{\Omega}}_i)/n$, where $\text{TPR}(\hat{\boldsymbol{\Omega}}_i)$ and $\text{TPR}(\hat{\boldsymbol{\Omega}}_i)$ are true and false positive errors calculated by comparing the entries in $\hat{\boldsymbol{\Omega}}_i$ and $\boldsymbol{\Omega}_i$, respectively.

With both `MB` and `glasso` stemming from an i.i.d. model, we have $\hat{\boldsymbol{\mu}}_i = \sum_{j=1}^{n} \boldsymbol{x}^{(j)}/n$ and $\hat{\boldsymbol{\Omega}}_i = \hat{\boldsymbol{\Omega}}$, where $\hat{\boldsymbol{\Omega}}$ is the output from `MB` or `glasso`. As `cmGMM` assumes heterogeneous

n	p, q	Method	$\boldsymbol{\mu}_{\text{error}}$	$\boldsymbol{\Omega}_{\text{error}}$	$\text{TPR}_{\boldsymbol{\Omega}}$	$\text{FPR}_{\boldsymbol{\Omega}}$
200	$p = 25$ $q = 50$	RegGMM	**3.042** (0.019)	**2.011** (0.018)	0.853 (0.004)	0.122 (0.002)
		MB	7.067 (0.010)	3.371 (0.011)	0.666 (0.003)	0.129 (0.002)
		glasso	7.067 (0.010)	10.824 (0.027)	0.706 (0.004)	0.158 (0.002)
		cmGMM	3.919 (0.040)	29.248 (0.443)	**0.982** (0.003)	0.477 (0.002)
		stGMM	7.219 (0.015)	8.807 (0.054)	0.492 (0.008)	**0.003** (0.001)
	$p = 25$ $q = 100$	RegGMM	**3.603** (0.019)	**2.147** (0.016)	0.809 (0.005)	0.121 (0.002)
		MB	6.916 (0.010)	3.168 (0.012)	0.642 (0.004)	0.107 (0.002)
		glasso	6.916 (0.010)	11.394 (0.040)	0.698 (0.004)	0.149 (0.002)
		cmGMM	5.724 (0.023)	15.236 (0.225)	**0.837** (0.005)	0.340 (0.004)
		stGMM	7.226 (0.011)	9.131 (0.047)	0.502 (0.007)	**0.028** (0.001)
	$p = 50$ $q = 50$	RegGMM	**3.753** (0.020)	**5.036** (0.024)	0.657 (0.004)	0.104 (0.002)
		MB	7.735 (0.010)	6.650 (0.014)	0.448 (0.003)	0.060 (0.001)
		glasso	7.735 (0.010)	15.661 (0.031)	0.511 (0.002)	0.090 (0.001)
		cmGMM	5.491 (0.027)	34.644 (0.232)	**0.890** (0.003)	0.430 (0.002)
		stGMM	8.414 (0.013)	11.345 (0.071)	0.182 (0.008)	**0.003** (0.000)
	$p = 50$ $q = 100$	RegGMM	**4.440** (0.019)	**5.332** (0.020)	0.629 (0.003)	0.094 (0.001)
		MB	7.791 (0.010)	6.656 (0.015)	0.447 (0.002)	0.061 (0.001)
		glasso	7.791 (0.010)	15.697 (0.039)	0.524 (0.003)	0.095 (0.001)
		cmGMM	6.199 (0.028)	29.146 (0.080)	**0.829** (0.003)	0.396 (0.000)
		stGMM	8.667 (0.013)	11.322 (0.077)	0.182 (0.008)	**0.004** (0.000)

Table S4: Evaluation criteria with varying network size p and covariate dimension q. Numbers achieving the best evaluation criteria in each setting are marked in boldface.

means but a common covariance, we let $\hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{\Gamma}}\boldsymbol{u}^{(i)}$ and $\hat{\boldsymbol{\Omega}}_i = \hat{\boldsymbol{\Omega}}$, where $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Omega}}$ are the output from `cmGMM`. For `stGMM` and given q binary covariates, there are 2^q groups (or conditions) to be considered, far exceeding the number of subjects n. To implement `stGMM`, we cluster the subjects into $K < n$ groups by applying K-means clustering to the covariates and denote the cluster label of subject i as $c_i \in \{1, \ldots, K\}$. We then implement `stGMM` and let $\hat{\boldsymbol{\mu}}_i = \sum_{j=1}^n \boldsymbol{x}^{(j)} \mathbf{1}_{\{c_j = c_i\}} / (\sum_{j=1}^n \mathbf{1}_{\{c_j = c_i\}})$ and $\hat{\boldsymbol{\Omega}}_i = \hat{\boldsymbol{\Omega}}_{c_i}$, where $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K$ are the output from `stGMM`. All benchmark methods are tuned using the tuning methods suggested in the respective papers.

Under the same setting as in Section 6 and let $p = 25, 50$, $q = 50, 100$, $n = 200$ and $K = 5$, Table S4 reports the average evaluation criteria over 200 data replicates, revealing that `RegGMM` improves upon the benchmark methods notably in estimation accuracy and selection accuracy. Specifically, because `MB`, `glasso` and `stGMM` are not designed to estimate the covariate dependent means or precision matrices well, they tend to incur large estimation errors and low selection accuracy. While `cmGMM` estimates well the covariate dependent means, it is not designed to accommodate covariate dependent precision matrices and, therefore, tends to overselect nonzero coefficients in $\hat{\boldsymbol{\Omega}}$ in the settings examined.

S1.5 Role of standardization

We comment on the model interpretability after standardizing covariates (Tibshirani, 1997) to have mean zero and variance one. As such, each graphical regression (8) is in effect

$$\hat{\boldsymbol{z}}_j = \sum_{k\neq j}^{p} \beta_{jk0} \frac{\hat{\boldsymbol{z}}_k}{\mathrm{sd}(z_k)} + \sum_{k\neq j}^{p}\sum_{h=1}^{q} \beta_{jkh} \frac{(\boldsymbol{u}_h - \bar{\boldsymbol{u}}_h)}{\mathrm{sd}(\boldsymbol{u}_h)} \odot \frac{\hat{\boldsymbol{z}}_k}{\mathrm{sd}(\hat{\boldsymbol{z}}_k)} + \epsilon_j,$$

where $\bar{\boldsymbol{u}}_h$ denotes the mean of $\boldsymbol{u}_h$, $\mathrm{sd}(\cdot)$ denotes the standard deviation, and $\hat{\boldsymbol{z}}_j$, residual from the first estimation step, has mean zero. The above equation can be re-written as

$$\hat{\boldsymbol{z}}_j = \sum_{k\neq j}^{p} \tilde{\beta}_{jk0} \hat{\boldsymbol{z}}_k + \sum_{k\neq j}^{p}\sum_{h=1}^{q} \tilde{\beta}_{jkh} \boldsymbol{u}_h \odot \hat{\boldsymbol{z}}_k + \epsilon_j, \tag{S1}$$

where $\tilde{\beta}_{jk0} = \frac{1}{\mathrm{sd}(\hat{\boldsymbol{z}}_k)} \left\{ \beta_{jk0} - \sum_{h=1}^{q} \frac{\bar{\boldsymbol{u}}_h}{\mathrm{sd}(\boldsymbol{u}_h)} \beta_{jkh} \right\}$ and $\tilde{\beta}_{jkh} = \frac{\beta_{jkh}}{\mathrm{sd}(\boldsymbol{u}_h)\mathrm{sd}(\boldsymbol{z}_k)}$. Notice $\tilde{\beta}_{jkh}$ and β_{jkh} only differs in scale by a positive scalar. Therefore, parameter estimates can be interpreted with non-standardized covariates after a re-calculation as in (S1).

S1.6 Positive definiteness

As our goal is to identify edges that are modulated by external covariates, we choose to employ the computationally efficient node-wise regressions and bypass the need to work with the entire covariance/precision matrix. If the true covariance/precision matrix is positive

	$p = 25$ $q = 50$	$p = 25$ $q = 100$	$p = 50$ $q = 50$	$p = 50$ $q = 100$
$\boldsymbol{\beta}_{\text{error}}$	1.378 (0.006)	1.417 (0.006)	2.228 (0.005)	2.291 (0.005)
$\boldsymbol{\beta}^r_{\text{error}}$	1.372 (0.005)	1.446 (0.006)	2.226 (0.005)	2.292 (0.004)

Table S5: Estimation accuracy of $\boldsymbol{\beta}_j$'s and $\boldsymbol{\beta}^r_j$'s in simulations with $n = 200$, varying network size p and covariate dimension q.

definite, it then follows from Theorems 1-4 that the estimated precision is asymptotically positive definite. However, for finite sample cases, it may be desirable to ensure the positive definiteness of the final estimator. We find that, assuming the true precision matrix is diagonal dominant and the covariates are in known and bounded intervals, we can adopt a post-hoc re-scaling step that gives the same estimator asymptotically as guaranteed by Theorems 1-4. Specifically, without loss of generality, assume $u_h \in [0, 1]$ (if not, rescale u_h first). For any j such that $\|\hat{\boldsymbol{\beta}}_j\|_1 > 1$, we set the final estimate to $[\hat{\boldsymbol{B}}_h]_{j\cdot}/\hat{\sigma}^{jj}$ and $[\hat{\boldsymbol{B}}_h]_{\cdot j}/\hat{\sigma}^{jj}$.

We have evaluated the performance of the above procedure under the same setting as in Section 6. Recall in our simulations, the true precision matrix is diagonal dominant and $u_h \in [0, 1]$. The results over 200 data replicates are reported in Table S5, where the rescaled $\hat{\boldsymbol{\beta}}_j$ is denoted as $\hat{\boldsymbol{\beta}}^r_j$. Table S5 suggests that the original and rescaled estimators have similar estimation errors.

S2 Technical Lemmas

We state the technical lemmas that will be used in our proofs.

Lemma 1 (Lemma 1 in Bellec et al. (2018))**.** *Let* $pen : \mathbb{R}^d \to \mathbb{R}$ *be any convex function and* $\hat{\boldsymbol{\beta}}$ *be defined by*

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \|\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}\|_2^2 + pen(\boldsymbol{\beta}) \right\},$$

where $\boldsymbol{W} \in \mathbb{R}^{n\times d}$, $\boldsymbol{y} \in \mathbb{R}^n$. *Then for* $\boldsymbol{\beta} \in \mathbb{R}^d$,

$$\|\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\beta}}\|_2^2 + pen(\hat{\boldsymbol{\beta}}) + \|\boldsymbol{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2 \leq \|\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}\|_2^2 + pen(\boldsymbol{\beta}).$$

Lemma 2 (Theorem F in Graybill and Marsaglia (1957))**.** *Let* $\boldsymbol{\epsilon}_j \sim \mathcal{N}_p(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ *and* $\boldsymbol{A}$ *be an* $p \times p$ *idempotent matrix with rank equals to* $r \leq p$. *Then,* $\boldsymbol{\epsilon}_j^\top \boldsymbol{A}\boldsymbol{\epsilon}_j/\sigma^2$ *follows a* χ^2 *distribution with* r *degrees of freedom.*

Lemma 3 (Lemma 1 in Laurent and Massart (2000))**.** *Suppose that* U *follows a* χ^2 *distribution with* r *degrees of freedom. For any* $x > 0$, *it holds that*

$$P(U - r \geq 2\sqrt{rx} + 2x) \leq \exp(-x).$$

Lemma 4 (Proposition 5.16 in Vershynin (2010))**.** *Let* $X_1, \ldots, X_n$ *be independent mean zero sub-exponential random variables. Let* $v_1 = \max_i \|X_i\|_{\psi_1}$, *where* $\|X_i\|_{\psi_1} = \sup_{d\geq 1} d^{-1}(E|X_i|^d)^{1/d}$ *denotes the sub-exponential norm. There exists a constant* c *such that, for any* $t > 0$,

$$P\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left\{-c\min\left(\frac{t^2}{v_1^2 n}, \frac{t}{z_1}\right)\right\}.$$

Lemma 5. *Consider independent vectors* $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ *in* $\mathbb{R} \times \mathbb{R}^p$ *such that* $y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \epsilon_i$ *and* $\boldsymbol{x}_i$ *is elementwise sub-Gaussian,* $i \in [n]$. *Let* ϵ_i*'s be independent Gaussian errors with non-constant variances and assume that* $\sup_{i\in[n]} Var(\epsilon_i)$ *is bounded by a constant* $K_1 > 0$. *Suppose* $\|\boldsymbol{\beta}\|_0 = k$ *and* $\lambda_{\min}(Cov(\boldsymbol{x}_1)) \geq 1/\phi_0$ *for some* $\phi_0 > 0$. *Let*

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta})^2 + \lambda\|\boldsymbol{\theta}\|_1.$$

When $\lambda = 14\sigma_n\sqrt{\tau_1 \log p/n} + \tau_1 K(\log p \log n)^{1/2}/n$ *for any* $\tau_1 > 0$, *the lasso estimate satisfies with probability at least* $1 - 3p^{-\tau_1}$ *that*

$$\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_2 \precsim \sigma_n\sqrt{\frac{k\log p}{n}} + \frac{k^{1/2}\log p}{n},$$

where $\sigma_n = \frac{1}{n}\sum_{i=1}^n Var(\epsilon_i)$.

The above result is adapted from Theorem 4.5 in Kuchibhotla and Chakrabortty (2018) by considering Gaussian errors.

Lemma 6 (Theorem 4.1 in Kuchibhotla and Chakrabortty (2018)). *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent random vectors in $\mathbb{R}^p$. Assume each element of $\boldsymbol{X}_i$ is sub-exponential with $\|X_{i,j}\|_{\psi_1} < K_2$, $i \in [n], j \in [p]$. where $\|X_{i,j}\|_{\psi_1} = \sup_{d \geq 1} d^{-1}(E|X_{i,j}|^d)^{1/d}$ denotes the sub-exponential norm. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \boldsymbol{X}^\top \boldsymbol{X}/n$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathbb{E}(\boldsymbol{X}^\top \boldsymbol{X}/n)$. Define*

$$\Upsilon_{n,k} = \max_{j,k} \frac{1}{n} \sum_{i=1}^{n} Var\{X_{i,j} X_{i,k}\}.$$

Then for any $t > 0$, with probability at least $1 - \mathcal{O}(p^{-1})$,

$$\sup_{\|\boldsymbol{v}\|_0 \leq k,\, \|\boldsymbol{v}\|_2 \leq 1} \left| \boldsymbol{v}^\top (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} - \boldsymbol{\Sigma}_{\boldsymbol{X}}) \boldsymbol{v} \right| \precsim k \sqrt{\frac{\Upsilon_{n,k} log\, p}{n}} + K_2^2 \frac{k (log\, n log\, p)^2}{n}.$$

Lemma 7 (Lemma 12 in Loh and Wainwright (2011)). *For any symmetric matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ and if $|\boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}| \leq \delta_1$ for any $\boldsymbol{v} \in \{\boldsymbol{v} : \|\boldsymbol{v}\|_0 \leq 2s \, and \, \|\boldsymbol{v}\|_2 = 1\}$, then*

$$|\boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}| \leq 27 \delta_1 (\|\boldsymbol{v}\|_2^2 + \frac{1}{s} \|\boldsymbol{v}\|_1^2), \; for \; any \quad \boldsymbol{v} \in \mathbb{R}^p.$$

S3 Proofs of Main Results

We begin by recalling the notation used throughout the paper. We denote the true parameters by $\boldsymbol{\beta}_j$, $j \in [p]$, though, in some places and without ambiguities, we use them to denote the corresponding parameters or the arguments in functions. The index set $\{1, \ldots, (p-1)(q+1)\}$ is partitioned into $q+1$ groups, indexed by $(0), (1), \ldots, (q) \subset \{1, \ldots, (p-1)(q+1)\}$. For a group index subset $\mathcal{G} \in \{1, \ldots, q\}$, we let $(\mathcal{G}) = \cup_{h \in \mathcal{G}}(h)$, $(\mathcal{G}^c) = \cup_{h \notin \mathcal{G}}(h)$, and $(\boldsymbol{\beta}_j)_{(\mathcal{G})}$ represent a sub-vector of $\boldsymbol{\beta}_j$ indexed by $(\mathcal{G})$. We define $\|\boldsymbol{\beta}_{j,-0}\|_{1,2} = \sum_{h=1}^{q} \|(\boldsymbol{\beta}_j)_{(h)}\|_2$. We use $\mathcal{S}_j$ to denote the element-wise support set of $\boldsymbol{\beta}_j$, i.e, $\mathcal{S}_j = \{l : (\boldsymbol{\beta}_j)_l \neq 0, l \in [(p-1)(q+1)]\}$, and $\mathcal{G}_j$ to denote the group-wise support set of $\boldsymbol{\beta}_j$, i.e, $\mathcal{G}_j = \{h : (\boldsymbol{\beta}_j)_{(h)} \neq \boldsymbol{0}, h \in [q]\}$. Moreover, we let $s_j = |\mathcal{S}_j|$ and $s_{j,g} = |\mathcal{G}_j|$.

S3.1 Proof of Theorem 1

As $\hat{\boldsymbol{\beta}}_j$ is a minimizer of the objective function (9) and the sparse group penalty function in (9) is convex, Lemma 1 implies that

$$\begin{aligned} & \frac{1}{2n}\|\boldsymbol{z}_j - \boldsymbol{W}\hat{\boldsymbol{\beta}}_j\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}_j\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2} + \frac{1}{2n}\|\boldsymbol{W}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\|_2^2 \\ \leq & \frac{1}{2n}\|\boldsymbol{z}_j - \boldsymbol{W}\boldsymbol{\beta}_j\|_2^2 + \lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}. \end{aligned}$$

Writing $\Delta = \hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j$ and reorganizing terms in the above inequality gives

$$\frac{1}{n}\|\boldsymbol{W}\Delta\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}_j\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2} \leq \frac{1}{n}\langle\boldsymbol{\epsilon}_j, \boldsymbol{W}\Delta\rangle + \lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}.$$

Using the fact that $\|\hat{\boldsymbol{\beta}}_j\|_1 = \|(\hat{\boldsymbol{\beta}}_j)_{\mathcal{S}_j}\|_1 + \|(\hat{\boldsymbol{\beta}}_j)_{\mathcal{S}_j^c}\|_1$, $\|\boldsymbol{\beta}_j\|_1 = \|(\boldsymbol{\beta}_j)_{\mathcal{S}_j}\|_1$, $\|\hat{\boldsymbol{\beta}}_j\|_{1,2} = \|(\hat{\boldsymbol{\beta}}_j)_{(\mathcal{G}_j)}\|_{1,2} + \|(\hat{\boldsymbol{\beta}}_j)_{(\mathcal{G}_j^c)}\|_{1,2}$, $\|\boldsymbol{\beta}_j\|_{1,2} = \|(\boldsymbol{\beta}_j)_{(\mathcal{G}_j)}\|_{1,2}$ and applying the triangle inequalities of $\|\cdot\|_1$ and $\|\cdot\|_{1,2}$, we arrive at

$$\frac{1}{n}\|\boldsymbol{W}\Delta\|_2^2 + \lambda\|\Delta_{\mathcal{S}_j^c}\|_1 + \lambda_g\|\Delta_{(\mathcal{G}_j^c)}\|_{1,2} \leq \frac{1}{n}\langle\boldsymbol{\epsilon}_j, \boldsymbol{W}\Delta\rangle + \lambda\|\Delta_{\mathcal{S}_j}\|_1 + \lambda_g\|\Delta_{(\mathcal{G}_j)}\|_{1,2}. \tag{S2}$$

Defining $\hat{\mathcal{S}}_j = \{l : (\hat{\boldsymbol{\beta}}_j)_l \neq 0, l \in [(p-1)(q+1)]\}$ and letting $\tilde{\mathcal{S}}_j = \mathcal{S}_j \cup \hat{\mathcal{S}}_j$, we obtain

$$\begin{aligned} \langle\boldsymbol{\epsilon}_j, \boldsymbol{W}\Delta\rangle &= \langle\boldsymbol{\epsilon}_j, \boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}\boldsymbol{W}_{\tilde{\mathcal{S}}_j}\Delta_{\tilde{\mathcal{S}}_j}\rangle \\ &= \langle\boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j, \boldsymbol{W}\Delta\rangle \leq \frac{1}{2a_1}\|\boldsymbol{W}\Delta\|_2^2 + \frac{a_1}{2}\|\boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j\|_2^2, \end{aligned} \tag{S3}$$

where the second equality holds with $\tilde{\mathcal{S}}_j = \mathcal{S}_j \cup \hat{\mathcal{S}}_j$ and the last inequality comes from that $2ab \leq ta^2 + b^2/t$ for any $t > 0$. Here, $\boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}$ is the orthogonal projection matrix onto the column space of $\boldsymbol{W}_{\tilde{\mathcal{S}}_j}$. As opposed to the classic techniques that bound $\langle\boldsymbol{\epsilon}, \boldsymbol{W}\Delta\rangle$ with $\|\boldsymbol{W}^\top\boldsymbol{\epsilon}\|_\infty\|\Delta\|_1$ or $\|\boldsymbol{W}^\top\boldsymbol{\epsilon}\|_{\infty,2}\|\Delta\|_{1,2}$ (Bickel et al., 2009; Lounici et al., 2011; Negahban et al., 2012), we bound this term in (S3) with $\|\boldsymbol{W}\Delta\|_2^2/(2a_1) + a_1\|\boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j\|_2^2/2$, which is useful in our proof to more sharply bound the term $\|\boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j\|_2^2$. This indeed is a challenging step as the group lasso penalty term in (9) is not decomposable with respect to $\mathcal{S}_j$ and, hence,

the existing techniques based on decomposable regularizers and null space properties are non-applicable. We provide a new proof, which is divided into three steps.

Step 1: We first bound $\|\boldsymbol{\mathcal{P}}_{\mathcal{S}_j}\boldsymbol{\epsilon}_j\|_2^2$ with some cardinality measures. Given any $\mathcal{J} \subset [(p-1)(q+1)]$ and $\boldsymbol{\gamma} \in \{0,1\}^{(p-1)(q+1)}$ satisfying $\boldsymbol{\gamma}_{\mathcal{J}} = \mathbf{1}$ and $\boldsymbol{\gamma}_{\mathcal{J}^c} = \mathbf{0}$, we write $\mathcal{G}(\mathcal{J}) = \{h : (\boldsymbol{\gamma})_{(h)} \neq \mathbf{0}, h \in [q]\}$. In this step, we aim to show that, given $0 \leq s_{j,g} \leq q+1$ and $0 \leq s_j \leq (p-1)(q+1)$, the following holds

$$\mathbb{P}\left[\sup_{|\mathcal{J}|=s_j, |\mathcal{G}(\mathcal{J})|=s_{j,g}} \|\boldsymbol{\mathcal{P}}_{\mathcal{J}}\boldsymbol{\epsilon}_j\|_2^2 \geq 6\sigma_{\epsilon_j}^2 \left\{s_j \log(ep) + s_{j,g}\log(eq/s_{j,g})\right\} + t\sigma_{\epsilon_j}^2\right] \leq c_1 \exp(-c_2 t),$$

where c_1, c_2 are positive constants.

As the projection matrix $\boldsymbol{\mathcal{P}}_{\mathcal{J}}$ is idempotent, Lemma 2 implies that

$$\|\boldsymbol{\mathcal{P}}_{\mathcal{J}}\boldsymbol{\epsilon}_j\|_2^2/\sigma_{\epsilon_j}^2 \sim \chi_d^2, \quad d < s_j,$$

where d is the rank of $\boldsymbol{\mathcal{P}}_{\mathcal{J}}$. Next, we find the size of $\{\mathcal{J} \subset [(p-1)(q+1)], |\mathcal{J}| = s_j, |\mathcal{G}(\mathcal{J})| = s_{j,g}\}$ by considering two cases (i) $s_{j,g} = s_j$ and (ii) $s_{j,g} < s_j$. These are the only two possible cases since $\mathcal{S}_j$ includes all nonzero elements in $(\boldsymbol{\beta}_j)_{(0)}$ and $(\boldsymbol{\beta}_j)_{(1)}, \ldots, (\boldsymbol{\beta}_j)_{(q)}$.

case (i): $s_{j,g} = s_j$. Here, $\{\mathcal{J} \subset [(p-1)(q+1)], |\mathcal{J}| = s_j, |\mathcal{G}(\mathcal{J})| = s_{j,g}\}$ contains $\binom{q}{s_{j,g}}(p-1)^{s_j}$ elements. It follows from Stirling's approximation that $\log\binom{q}{s_{j,g}} \leq s_{j,g}\log(eq/s_{j,g})$. Therefore, $\log\left\{\binom{q}{s_{j,g}}(p-1)^{s_j}\right\} \leq s_j \log p + s_{j,g}\log(eq/s_{j,g})$.

case (ii): $s_{j,g} < s_j$. The number of elements in $\{\mathcal{J} \subset [(p-1)(q+1)], |\mathcal{J}| = s_j, |\mathcal{G}(\mathcal{J})| = s_{j,g}\}$ is bounded above by $\binom{q}{s_{j,g}}\binom{(p-1)+(p-1)s_{j,g}}{s_j}$. By Stirling's approximation, we have $\log\binom{(p-1)+(p-1)s_{j,g}}{s_j} \leq s_j \log(e(p-1)(s_{j,g}+1)/s_j) \leq s_j \log(ep)$. Therefore, we have $\log\left\{\binom{q}{s_{j,g}}\binom{(p-1)+(p-1)s_{j,g}}{s_j}\right\} \leq s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})$.

Combining these two cases, we conclude that $|\{\mathcal{J} \subset [(p-1)(q+1)], |\mathcal{J}| = s_j, |\mathcal{G}(\mathcal{J})| = s_{j,g}\}|$ is bounded above by $s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})$. Applying Lemma 3 and applying the union bound leads to the desired result in Step 1.

Step 2: Using the result from Step 1, we find an upper bound for $\|\boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j\|_2^2$. First, we define

$$r_{s,s_g} = \left(\sup_{|\mathcal{J}|=s, |\mathcal{G}(\mathcal{J})|=s_g} \|\boldsymbol{\mathcal{P}}_{\mathcal{J}}\boldsymbol{\epsilon}_j\|_2^2 - M\sigma_{\epsilon_j}^2 \{s\log(ep) + s_g\log(eq/s_g)\} \right)_+,$$

and $r = \sup_{1\le s\le (p-1)(q+1),\, 0\le s_g\le q} r_{s,s_g}$, where $M > 0$ is a constant to be specified later. Then,

$$\begin{aligned} \|\boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j\|_2^2 &\le M\sigma_{\epsilon_j}^2 \{\tilde{s}_j\log(ep) + \tilde{s}_{j,g}\log(eq/s_{j,g})\} + r \\ &\le M\sigma_{\epsilon_j}^2 \{(s_j + \hat{s}_j)\log(ep) + (s_{j,g} + \hat{s}_{j,g})\log(eq/s_{j,g})\} + r. \end{aligned} \tag{S4}$$

With $M = 9$, the result from Step 1 gives

$$\begin{aligned} \mathbb{P}\{r \ge t\sigma_{\epsilon_j}^2\} &\le \sum_{s=1}^{(p-1)(q+1)} \sum_{s_g=0}^{q} \mathbb{P}\{r_{s,s_g} \ge t\sigma_{\epsilon_j}^2\} \\ &\le \sum_{s=1}^{(p-1)(q+1)} \sum_{s_g=0}^{q} c_1 \exp[-c_2 t - 3c_2\{s\log(ep) + s_g\log(eq/s_g)\}]. \end{aligned}$$

Step 3: This step derives an inequality for $\|\boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j\|_2^2$ by utilizing the computational optimality of $\hat{\boldsymbol{\beta}}_j$. Since the objective function is convex, $\hat{\boldsymbol{\beta}}_j$ is a stationary point of

$$\frac{1}{2n}\|\boldsymbol{z}_j - \boldsymbol{W}\boldsymbol{\beta}_j|_2^2 + \lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}.$$

By the KKT conditions, for any $l \in \hat{\mathcal{S}}_j \cap (0)$, $h \in [q]$, $\hat{\boldsymbol{\beta}}_j$ must satisfy that

$$\lambda\,\mathrm{sign}\{(\hat{\boldsymbol{\beta}}_j)_l\} = \frac{1}{n}\langle \boldsymbol{w}_l, \boldsymbol{z}_j - \boldsymbol{W}\hat{\boldsymbol{\beta}}_j\rangle. \tag{S5}$$

Similarly, for any $l \in \hat{\mathcal{S}}_j \cap (h)$, it must satisfy that

$$\lambda\,\mathrm{sign}\{(\hat{\boldsymbol{\beta}}_j)_l\} + \lambda_g \frac{(\hat{\boldsymbol{\beta}}_j)_l}{\|(\hat{\boldsymbol{\beta}}_j)_{(h)}\|_2^2} = \frac{1}{n}\langle \boldsymbol{w}_l, \boldsymbol{z}_j - \boldsymbol{W}\hat{\boldsymbol{\beta}}_j\rangle. \tag{S6}$$

Squaring both sides of (S5) and (S6) and summing over all $l \in \hat{\mathcal{S}}_j$ gives

$$\lambda^2\hat{s}_j + \lambda_g^2\hat{s}_{j,g} \le \frac{1}{n^2}\|\boldsymbol{W}_{\hat{\mathcal{S}}}^\top(\boldsymbol{z}_j - \boldsymbol{W}\hat{\boldsymbol{\beta}}_j)\|_2^2,$$

due to $\mathrm{sign}\{(\hat{\boldsymbol{\beta}}_j)_l\}(\hat{\boldsymbol{\beta}}_j)_l \ge 0$.

Consider $\boldsymbol{W}_{i\cdot}^\top \boldsymbol{v}$, where $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$, $\boldsymbol{v}_l \in \mathbb{R}^{q+1}$ and $\|\boldsymbol{v}\| = 1$. We have $\boldsymbol{W}_{i\cdot} = (z_1^{(i)}, z_1^{(i)}u_1^{(i)}, \ldots, z_1^{(i)}u_q^{(i)}, \ldots, z_p^{(i)}, z_p^{(i)}u_1^{(i)}, \ldots, z_p^{(i)}u_q^{(i)})$. With a slight abuse of notation, we include the intercept term into $\boldsymbol{u}^{(i)}$ in the subsequent development. Letting $\boldsymbol{V} = [\boldsymbol{v}_1^\top, \ldots, \boldsymbol{v}_p^\top] \in \mathbb{R}^{(q+1)\times p}$, we can reexpress $\boldsymbol{W}_{i\cdot}\boldsymbol{v}$ as $\boldsymbol{W}_{i\cdot}\boldsymbol{v} = \boldsymbol{u}^{(i)\top}\boldsymbol{V}\boldsymbol{z}^{(i)}$, $i \in [n]$. Consequently, by the law of total expectation and Assumption 1, we have

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{v}^\top \frac{\boldsymbol{W}^\top\boldsymbol{W}}{n}\boldsymbol{v}\right) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\boldsymbol{u}^{(i)\top}\boldsymbol{V}\boldsymbol{z}^{(i)}\right)^2 \\
&= \mathbb{E}\left[\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\boldsymbol{u}^{(i)\top}\boldsymbol{V}\boldsymbol{z}^{(i)}\right)^2 \Big| \{\boldsymbol{u}^{(i)}\}_{i\in[n]}\right\}\right] \\
&= \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n \boldsymbol{u}^{(i)\top}\boldsymbol{V}\boldsymbol{\Sigma}(\boldsymbol{u}^{(i)})\boldsymbol{V}^\top\boldsymbol{u}^{(i)}\right\} \\
&\geq \phi_1 \times \mathbb{E}\left\{\mathrm{tr}\left(\boldsymbol{u}^{(1)\top}\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{u}^{(1)}\right)\right\} \\
&\geq \phi_1 \times \lambda_{\min}\left(\mathrm{Cov}(\boldsymbol{u}^{(1)})\right)\mathrm{tr}(\boldsymbol{V}\boldsymbol{V}^\top) = \phi_1/\phi_0,
\end{aligned} \tag{S7}$$

where we have used the fact $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) \geq \lambda_{\min}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B})$ for positive semi-definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and $\mathrm{tr}(\boldsymbol{V}\boldsymbol{V}^\top) = 1$. By the Cauchy–Schwarz inequality, we deduce

$$\max_{j,k}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\{(\boldsymbol{W}_{ij}\boldsymbol{W}_{ik})^2\} = \max_{l_1,l_2,l_3,l_4}\mathbb{E}\left(z_{l_1}^{(1)2}z_{l_2}^{(1)2}u_{l_3}^{(1)2}u_{l_4}^{(1)2}\right) = \mathcal{O}(1), \tag{S8}$$

where we have used the fact that $z_{l_1}^{(i)}$ and $u_{l_2}^{(i)}$ have bounded eighth moments, as they are both sub-Gaussian with a bounded sub-Gaussian norm. It then follows from Lemma 6 and with (S7) and Assumption 3 that, with probability at least $1 - C_0\exp\{-(\log p + \log q)\}$, we have $\|\boldsymbol{W}_{\hat{\mathcal{S}}}\|^2/n \leq M_1$ for some $M_1 > 0$. Since $\hat{\mathcal{S}}_j \in \tilde{\mathcal{S}}$, it holds that

$$\lambda^2\hat{s}_j + {\lambda_g}^2\hat{s}_{j,g} \leq \frac{2M_1}{n}\|\boldsymbol{W}\Delta\|_2^2 + \frac{2M_1}{n}\|\boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j\|_2^2. \tag{S9}$$

Combining (S4) and (S9) and letting

$$\lambda = C\sigma_{\epsilon_j}\sqrt{\log(ep)/n + s_{j,g}\log(eq/s_{j,g})/(ns_j)}, \quad \lambda_g = \sqrt{s_j/s_{j,g}}\lambda,$$

where $C = 3(M_1 a_2)^{1/2}$ for some $a_2 > 0$, we arrive at

$$(1 - \frac{2}{a_2})\|\boldsymbol{\mathcal{P}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j\|_2^2 \leq 9\sigma_{\epsilon_j}^2 \{s_j \log(ep) + s_{j,g}\log(eq/s_{j,g})\} + \frac{2}{a_2}\|\boldsymbol{W}\Delta\|_2^2 + r. \tag{S10}$$

This, together with (S2) and (S3), implies

$$\begin{aligned} & \frac{\|\boldsymbol{W}\Delta\|_2^2}{n} + \lambda\|\Delta_{\mathcal{S}_j^c}\|_1 + \lambda_g\|\Delta_{(\mathcal{G}_j^c)}\|_{1,2} \\ \leq & \frac{1}{2a_1}\frac{\|\boldsymbol{W}\Delta\|_2^2}{n} + \frac{9a_1a_2}{2(a_2-2)}\frac{\sigma_{\epsilon_j}^2\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}}{n} \\ & + \frac{a_1}{a_2-2}\frac{\|\boldsymbol{W}\Delta\|_2^2}{n} + \frac{a_1a_2}{2(a_2-2)n}r + \lambda\|\Delta_{\mathcal{S}_j}\|_1 + \lambda_g\|\Delta_{(\mathcal{G}_j)}\|_{1,2}. \end{aligned} \tag{S11}$$

Next, we have that

$$\frac{\|\Delta_{\mathcal{S}_j}\|_1}{\sqrt{s_j}} + \frac{\|\Delta_{(\mathcal{G}_j)}\|_{1,2}}{\sqrt{s_{j,g}}} \leq \|\Delta_{\mathcal{S}_j}\|_2 + \|\Delta_{(\mathcal{G}_j)}\|_2 \leq 2\frac{\phi_1}{\phi_0}\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2,$$

where the first inequality is due to that $\|\Delta_{\mathcal{S}_j}\|_1 \leq \sqrt{s_j}\|\Delta_{\mathcal{S}_j}\|_2$, $\|\Delta_{(\mathcal{G}_j)}\|_{1,2} \leq \sqrt{s_{j,g}}\|\Delta_{(\mathcal{G}_j)}\|_2$ and the second inequality holds because $\|\Delta_{\mathcal{S}_j}\|_2 + \|\Delta_{(\mathcal{G}_j)}\|_2 < 2\|\Delta\|_2$ trivially with $\Delta_{\mathcal{S}_j}$ and $\Delta_{(\mathcal{G}_j)}$ being the sub-vectors of Δ, and $\lambda_{min}(\boldsymbol{\Sigma}_{\boldsymbol{W}}) \geq \phi_1/\phi_0 > 0$ in (S7). Consequently,

$$\lambda\|\Delta_{\mathcal{S}_j}\|_1 + \lambda_g\|\Delta_{(\mathcal{G}_j)}\|_{1,2} \;\leq\; 2C\frac{\phi_1}{\phi_0}\sqrt{e_j}\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2 \leq a_3 C\frac{\phi_1}{\phi_0}e_j + \frac{1}{a_3}\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2^2,$$

where $e_j = \sigma_{\epsilon_j}^2\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}/n$ and the last inequality comes from that $2ab \leq ta^2 + b^2/t$ for any $t > 0$. Plugging this into (S11), we obtain

$$\begin{aligned} & \left\{1 - \frac{1}{2a_1} - \frac{a_1}{a_2-2}\right\}\frac{\|\boldsymbol{W}\Delta\|_2^2}{n} \\ \leq & \left\{\frac{9a_1a_2}{2(a_2-2)} + Ca_3\frac{\phi_1}{\phi_0}\right\}\frac{\sigma_{\epsilon_j}^2\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}}{n} + \frac{1}{a_3}\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2^2 + \frac{a_1a_2}{2(a_2-2)n}r. \end{aligned} \tag{S12}$$

It remains to bound the distance between $\|\boldsymbol{W}\Delta\|_2^2/n$ and $\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2^2$. To proceed, we first show with probability at least $1 - C'\exp\{-(\log p + \log q)\}$,

$$\sup_{\boldsymbol{v}\in\mathbb{K}_0(2C_{\boldsymbol{\beta}_j}s_j)}\left|\boldsymbol{v}^\top\left(\frac{\boldsymbol{W}^\top\boldsymbol{W}}{n} - \boldsymbol{\Sigma}_{\boldsymbol{W}}\right)\boldsymbol{v}\right| \leq 1/L, \tag{S13}$$

where L is an arbitrarily large constant and $\mathbb{K}_0(2C_{\boldsymbol{\beta}_j}s_j) = \{\boldsymbol{v} : \|\boldsymbol{v}\|_0 \leq 2C_{\boldsymbol{\beta}_j}s_j \text{ and } \|\boldsymbol{v}\|_2 = 1\}$ for some positive constant $C_{\boldsymbol{\beta}_j}$.

Given (S8), it follows from Lemma 6 and Assumption 3 that with probability at least $1 - C'\exp\{-(\log p + \log q)\}$

$$\left|\boldsymbol{v}^\top \left(\frac{\boldsymbol{W}^\top \boldsymbol{W}}{n} - \boldsymbol{\Sigma}_{\boldsymbol{W}}\right)\boldsymbol{v}\right| = o(1).$$

Combing this with the result in Lemma 7, we have, with probability at least $1-C'\exp\{-(\log p+\log q)\}$,

$$\left|\Delta^\top \left(\frac{\boldsymbol{W}^\top \boldsymbol{W}}{n} - \boldsymbol{\Sigma}_{\boldsymbol{W}}\right)\Delta\right| \leq \frac{1}{L'}\left(\|\Delta\|_2^2 + \frac{1}{s_j}\|\Delta\|_1^2\right), \tag{S14}$$

where L' is an arbitrarily large positive constant. Plugging (S14) into (S12) and choosing proper constants a_1, a_2 and a_3 (e.g., $a_1 = 2$, $a_2 = 6$, $a_3 = 6$), we have

$$\frac{1}{2}\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2^2 \precsim \frac{\sigma_{\epsilon_j}^2\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}}{n} + \frac{1}{L'}\left(\|\Delta\|_2^2 + \frac{1}{s_j}\|\Delta\|_1^2\right) + \frac{\sigma_{\epsilon_j}^2}{n}, \tag{S15}$$

with probability at least $1 - c_1\exp[-c_2'\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}]$, due to that

$$\mathbb{P}\left[r \geq M_0\sigma_{\epsilon_j}^2\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}\right] \leq c_1\exp[-c_2'\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}],$$

for a large positive constant M_0. Next, taking $a_1 = 2 - \sqrt{2}$ and $a_2 = 6$ in (S11) and using the expressions for λ, λ_g, we have with probability at least $1 - c_1\exp[-c_2'\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}]$ that

$$\frac{\|\Delta_{\mathcal{S}_j^c}\|_1}{\sqrt{s_j}} + \frac{\|\Delta_{(\mathcal{G}_j^c)}\|_{1,2}}{\sqrt{s_{j,g}}} \leq \sigma_{\epsilon_j}\sqrt{\frac{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})}{n}} + \frac{\|\Delta_{\mathcal{S}_j}\|_1}{\sqrt{s_j}} + \frac{\|\Delta_{(\mathcal{G}_j)}\|_{1,2}}{\sqrt{s_{j,g}}}. \tag{S16}$$

Adding $\|\Delta_{\mathcal{S}_j}\|_1/\sqrt{s_j}$ to both sides of (S16), we get

$$\frac{\|\Delta\|_1}{\sqrt{s_j}} \leq \sqrt{e_j} + 3\|\Delta\|_2. \tag{S17}$$

Plugging (S17) into (S15) and with $\lambda_{min}(\boldsymbol{\Sigma}_{\boldsymbol{W}}) \geq \phi_1/\phi_0 > 0$ in (S7), we have

$$\|\Delta\|_2^2 \precsim \frac{\sigma_{\epsilon_j}^2}{n}\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\} + \frac{\sigma_{\epsilon_j}^2}{n},$$

with probability at least $1 - C_1 \exp[-C_2\{s_j \log(ep) + s_{j,g}\log(eq/s_{j,g})\}$, for some positive constants C_1, C_2.

□

S3.2 Proof of Theorem 2

We first establish the element-wise error bound of $\hat{\boldsymbol{\beta}}_j$ in (14), which is to be accomplished in three steps.

Step 1: With $\boldsymbol{\Psi} = \boldsymbol{W}^\top \boldsymbol{W}/n$, this step shows that with probability at least $1-2\exp\{-c_3(\log p + \log q)\}$,

$$\|\boldsymbol{\Psi}\Delta\|_\infty \leq \frac{3\eta_j \lambda}{2} \tag{S18}$$

for a constant $c_3 > 0$. We also show with probability at least $1 - 2\exp\{-c_3' \log p\}$ that

$$\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 4\eta_j \|\Delta_{\mathcal{S}_j}\|_1. \tag{S19}$$

The KKT conditions state that $\boldsymbol{\theta}$, an optimizer of (9), satisfies

$$\begin{cases} \big(\boldsymbol{W}^\top(\boldsymbol{z}_j - \boldsymbol{W}\boldsymbol{\theta})/n\big)_l = \text{sign}(\theta_l)\lambda & \text{if } \theta_l \neq 0, l \in (0) \\ \big(\boldsymbol{W}^\top(\boldsymbol{z}_j - \boldsymbol{W}\boldsymbol{\theta})/n\big)_l = \text{sign}(\theta_l)\lambda + \lambda_g \frac{\theta_l}{\|\boldsymbol{\theta}_{(h)}\|_2} & \text{if } \theta_l \neq 0, l \in (h) \\ \big|\big(\boldsymbol{W}^\top(\boldsymbol{z}_j - \boldsymbol{W}\boldsymbol{\theta})/n\big)_l\big| < \eta_j\lambda & \text{if } \theta_l = 0. \end{cases}$$

Thus, $\hat{\boldsymbol{\beta}}_j$ must satisfy that

$$\left\| \frac{1}{n}\boldsymbol{W}^\top(\boldsymbol{z}_j - \boldsymbol{W}\hat{\boldsymbol{\beta}}_j) \right\|_\infty \leq \eta_j \lambda.$$

If we can show that with high probability

$$\frac{1}{n}\left\|\boldsymbol{W}^\top \boldsymbol{\epsilon}_j\right\|_\infty \leq \frac{\eta_j \lambda}{2}, \tag{S20}$$

we conclude that (S18) holds with at least the same probability.

To prove (S20), we first define $V_l = \boldsymbol{w}_l^\top \boldsymbol{\epsilon}_j/n$, $l \in [(p-1)(q+1)]$, where $\boldsymbol{w}_l$ is the lth column of $\boldsymbol{W}$. As $\boldsymbol{u}_h^{(i)}$ is bounded as assumed in Assumption 1, $w_l^{(i)}$ is sub-Gaussian. Hence,

V_l is a sum of independent sub-exponential random variables and $\mathrm{Var}(w_l^{(i)}) \leq M^2\phi_2$ for some $M > 0$. Applying Lemma 4 gives that

$$\mathbb{P}\left(|V_l| > \frac{\eta_j\lambda}{2}\right) \leq 2\exp\left(-\frac{c\eta_j^2\lambda^2 n}{4M^2\phi_2\sigma_{\epsilon_j}^2}\right) \leq 2\exp\{-C_0'(\log p + \log q)\}, \tag{S21}$$

where $C_0' = cC^2/(4M^2\phi_2)$, and C is as defined in (12), and the last inequality is due to

$$\eta_j\lambda \geq C\sigma_{\epsilon_j}\sqrt{\frac{\log p}{n}} + C\sigma_{\epsilon_j}\sqrt{\frac{\log q}{n}}.$$

With constant C in (12) chosen sufficiently large, we have $C_0' > 1$. Applying the union bound inequality gives

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\left\|\boldsymbol{W}^\top\boldsymbol{\epsilon}_j\right\|_\infty \geq \frac{\eta_j\lambda}{2}\right) &\leq \mathbb{P}\left(\max_l |V_l| \geq \frac{\eta_j\lambda}{2}\right)\\ &\leq 2\exp\{-(C_0'-1)(\log p + \log q)\}.\end{aligned}$$

We have finished showing (S18) of Step 1 by taking $c_3 = C_0' - 1$.

Next, we show that $\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 3\eta_j\|\Delta_{\mathcal{S}_j}\|_1$ with probability greater than $1-\exp\{-c_3'(\log p)\}$. The definition of $\hat{\boldsymbol{\beta}}_j$ implies that

$$\frac{1}{2n}\|\boldsymbol{z}_j - \boldsymbol{W}\hat{\boldsymbol{\beta}}_j\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}_j\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2} \leq \frac{1}{2n}\|\boldsymbol{\epsilon}_j\|_2^2 + \lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}.$$

Developing the left hand side of the above inequality gives

$$\lambda\|\hat{\boldsymbol{\beta}}_j\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2} \leq \lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2} + \frac{1}{n}\Delta^\top\boldsymbol{W}^\top\boldsymbol{\epsilon}_j. \tag{S22}$$

Recall that $V_l = \boldsymbol{w}_l^\top\boldsymbol{\epsilon}_j/n$, $l \in [(p-1)(q+1)]$, where $\boldsymbol{w}_l$ is the lth column of $\boldsymbol{W}$. Using similar arguments as in (S20) and again applying Lemma 4 gives that

$$\mathbb{P}\left(|V_l| > \frac{\lambda}{2}\right) \leq 2\exp\left(-\frac{c\lambda^2 n}{4M^2\phi_2\sigma_{\epsilon_j}^2}\right) \leq 2\exp\{-C_0''(\log p)\}, \tag{S23}$$

where $C_0'' = cC^2/(4M^2\phi_2)$, C is as defined in (12), and the last inequality is due to $\lambda^2 n \geq C^2\sigma_{\epsilon_j}^2\log p$. Applying the union bound inequality then gives

$$\mathbb{P}\left(\frac{1}{n}\left\|\boldsymbol{W}^\top\boldsymbol{\epsilon}_j\right\|_\infty \geq \frac{\lambda}{2}\right) \leq \mathbb{P}\left(\max_l |V_l| \geq \frac{\lambda}{2}\right) \leq 2\exp\{\log q - (C_0''-1)(\log p)\}.$$

As $\log p \asymp \log q$, when C is chosen sufficiently large, we have, for some $c_3' > 0$, $\exp\{\log q - (C_0'' - 1)(\log p)\} \leq \exp\{-c_3' \log p\}$. Defining event $\mathcal{A}$ as

$$\mathcal{A} = \left\{ \frac{1}{n} \left\| \boldsymbol{W}^\top \boldsymbol{\epsilon}_j \right\|_\infty \leq \frac{\lambda}{2} \right\}, \tag{S24}$$

we have $\mathbb{P}(\mathcal{A}^c) \leq 2\exp\{-c_3' \log p\}$. Given $\mathcal{A}$, (S22) leads to

$$2\|\hat{\boldsymbol{\beta}}_j\|_1 + 2\sqrt{\frac{s_j}{s_{j,g}}}\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2} \leq 2\|\boldsymbol{\beta}_j\|_1 + 2\sqrt{\frac{s_j}{s_{j,g}}}\|\boldsymbol{\beta}_{j,-0}\|_{1,2} + \|\Delta\|_1,$$

as $\lambda_g = \sqrt{s_j/s_{j,g}}\lambda$ specified in (12).

Adding $\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_1$ and $\sqrt{\frac{s_j}{s_{j,g}}}\|\hat{\boldsymbol{\beta}}_{j,-0} - \boldsymbol{\beta}_{j,-0}\|_{1,2}$ to both sides and noting $\|\Delta\|_1 = \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_1$, we obtain

$$\begin{aligned} & \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_1 + 2\|\hat{\boldsymbol{\beta}}_j\|_1 + 2\sqrt{\frac{s_j}{s_{j,g}}}\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2} + \sqrt{\frac{s_j}{s_{j,g}}}\|\hat{\boldsymbol{\beta}}_{j,-0} - \boldsymbol{\beta}_{j,-0}\|_{1,2} \\ \leq \; & 2\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_1 + 2\|\boldsymbol{\beta}_j\|_1 + 2\sqrt{\frac{s_j}{s_{j,g}}}\|\boldsymbol{\beta}_{j,-0}\|_{1,2} + 2\sqrt{\frac{s_j}{s_{j,g}}}\|\hat{\boldsymbol{\beta}}_{j,-0} - \boldsymbol{\beta}_{j,-0}\|_{1,2}, \end{aligned}$$

which leads to

$$\begin{aligned} & \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_1 + \sqrt{\frac{s_j}{s_{j,g}}}\|\hat{\boldsymbol{\beta}}_{j,-0} - \boldsymbol{\beta}_{j,-0}\|_{1,2} \\ & \leq 2(\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_1 + \|\boldsymbol{\beta}_j\|_1 - \|\hat{\boldsymbol{\beta}}_j\|_1) + 2\sqrt{\frac{s_j}{s_{j,g}}}(\|\boldsymbol{\beta}_{j,-0} - \boldsymbol{\beta}_{j,-0}\|_{1,2} + \|\boldsymbol{\beta}_{j,-0}\|_{1,2} - \|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2}). \end{aligned}$$

Since $\|(\hat{\boldsymbol{\beta}}_j)_l - (\boldsymbol{\beta}_j)_l\|_1 + \|(\boldsymbol{\beta}_j)_l\|_1 - \|(\hat{\boldsymbol{\beta}}_j)_l\|_1 = 0$ for $l \in \mathcal{S}_j^c$ and $\|(\hat{\boldsymbol{\beta}}_j)_{(h)} - (\boldsymbol{\beta}_j)_{(h)}\|_{1,2} + \|(\boldsymbol{\beta}_j)_{(h)}\|_{1,2} - \|(\hat{\boldsymbol{\beta}}_j)_{(h)}\|_{1,2} = 0$ for $h \in \mathcal{G}_j^c$, using the triangular inequality yields that

$$\|\Delta_{\mathcal{S}_j^c}\|_1 \leq \|\Delta\|_1 \leq 4\|\Delta_{\mathcal{S}_j}\|_1 + 4\sqrt{\frac{s_j}{s_{j,g}}}\|\Delta_{\mathcal{G}_j}\|_{1,2}.$$

Therefore, conditional on event $\mathcal{A}$ in (S24), we have that

$$\|\Delta_{\mathcal{S}_j^c}\|_1 + \sqrt{\frac{s_j}{s_{j,g}}}\|\Delta_{\mathcal{G}_j^c}\|_{1,2} \leq 4\|\Delta_{\mathcal{S}_j}\|_1 + 4\sqrt{\frac{s_j}{s_{j,g}}}\|\Delta_{\mathcal{G}_j}\|_{1,2},$$

which further implies $\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 4\eta_j\|\Delta_{\mathcal{S}_j}\|_1$ because $\|\Delta_{\mathcal{G}_j}\|_{1,2} \leq \|\Delta_{\mathcal{S}_j}\|_1$.

Step 2: The step bounds the diagonal and off-diagonal elements of $\mathbf{\Psi}$. We first bound the diagonal elements, i.e., $\mathbf{\Psi}_{ll} = \|\boldsymbol{w}_l\|_2^2/n$, where $w_l^{(i)}$ is sub-Gaussian as $u_l^{(i)}$ is bounded. Under Assumptions 1 and 2, we have $\mathrm{Var}(w_l^{(i)}) \leq M^2\phi_2$ and by (S7), we have $\mathbf{\Sigma}_{\boldsymbol{W}}(l,l) \geq \phi_1/\phi_0$. Using the concentration inequality for sub-exponential random variables in Lemma 4, we have

$$\mathbb{P}\left(|\mathbf{\Psi}_{ll} - \mathbf{\Sigma}_{\boldsymbol{W}}(l,l)| > \mathbf{\Sigma}_{\boldsymbol{W}}(l,l)/2\right) \leq 2\exp(-c_4 n),$$

for some positive constant c_4. Immediately,

$$\mathbb{P}(\mathbf{\Psi}_{ll} \notin [\phi_0/(2\phi_1), 2M^2\phi_2]) \leq 2\exp(-c_4 n) \tag{S25}$$

because

$$\mathbb{P}(\mathbf{\Psi}_{ll} \notin [\phi_0/(2\phi_1), 2M^2\phi_2]) \leq \mathbb{P}\left(|\mathbf{\Psi}_{ll} - \mathbf{\Sigma}_{\boldsymbol{W}}(l,l)| > \mathbf{\Sigma}_{\boldsymbol{W}}(l,l)/2\right).$$

Similarly, for the off diagonal elements, i.e., $\mathbf{\Psi}_{kl} = \boldsymbol{w}_k^\top \boldsymbol{w}_l/n$, by noting that $\|\mathbf{\Sigma}(\boldsymbol{u}^{(i)})\|_{\max} \leq \phi_2$, we have

$$\begin{aligned} &\mathbb{P}\left\{\mathbf{\Psi}_{kl} \notin [-\frac{1}{c_0(1+8\eta_j)s_j}, \frac{3}{c_0(1+8\eta_j)s_j}]\right\} \\ \leq\ &\mathbb{P}(|\mathbf{\Psi}_{kl} - \mathbf{\Sigma}_{\boldsymbol{W}}(k,l)| \geq 2\mathbf{\Sigma}_{\boldsymbol{W}}(k,l)) \leq 2\exp(-c_5 n), \end{aligned} \tag{S26}$$

for a positive constant c_5.

Step 3: This step establishes that, conditional on event $\mathcal{A}$ and that

$$\mathbf{\Psi}_{ll} \in [\phi_0/(2\phi_1), 2M^2\phi_2], \quad \mathbf{\Psi}_{kl} \in [-\frac{1}{c_0(1+8\eta_j)s_j}, \frac{3}{c_0(1+8\eta_j)s_j}], \tag{S27}$$

we have

$$\min_{\|\boldsymbol{v}_{\mathcal{S}_j^c}\|_1 \leq 3\eta_j\|\boldsymbol{v}_{\mathcal{S}_j}\|_1} \frac{\|\boldsymbol{W}\boldsymbol{v}\|_2}{\sqrt{n}\|\boldsymbol{v}_{\mathcal{S}_j}\|_2} \geq \sqrt{\frac{\phi_0}{2\phi_1} - \frac{1}{c_0}} > 0.$$

First, we have that

$$\begin{aligned} \frac{\|\boldsymbol{W}\boldsymbol{v}_{\mathcal{S}_j}\|_2^2}{n\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} &= \frac{\boldsymbol{v}_{\mathcal{S}_j}^\top \mathrm{diag}(\mathbf{\Psi})\boldsymbol{v}_{\mathcal{S}_j}}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} + \frac{\boldsymbol{v}_{\mathcal{S}_j}^\top(\mathbf{\Psi} - \mathrm{diag}(\mathbf{\Psi}))\boldsymbol{v}_{\mathcal{S}_j}}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} \\ &\geq \frac{\phi_0}{2\phi_1} - \frac{1}{c_0(1+8\eta_j)s_j}\frac{\|\boldsymbol{v}_{\mathcal{S}_j}\|_1^2}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2}, \end{aligned}$$

where $\text{diag}(\boldsymbol{\Psi})$ is a diagonal matrix with the diagonal elements being identical to those of $\boldsymbol{\Psi}$, and the last inequality follows from (S27). Furthermore,

$$\begin{aligned}\frac{\|\boldsymbol{W}\boldsymbol{v}\|_2^2}{n\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} &\geq \frac{\|\boldsymbol{W}\boldsymbol{v}_{\mathcal{S}_j}\|_2^2}{n\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} + 2\frac{\boldsymbol{v}_{\mathcal{S}_j}^\top \boldsymbol{\Psi}\boldsymbol{v}_{\mathcal{S}_j^c}}{n\|\boldsymbol{v}_{\mathcal{S}_j}\|_2} \\ &\geq \frac{\phi_0}{2\phi_1} - \frac{1}{c_0(1+8\eta_j)s_j}\frac{\|\boldsymbol{v}_{\mathcal{S}_j}\|_1^2}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} - \frac{2}{c_0(1+8\eta_j)s_j}\frac{\|\boldsymbol{v}_{\mathcal{S}_j}\|_1\|\boldsymbol{v}_{\mathcal{S}_j^c}\|_1}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} \\ &\geq \frac{\phi_0}{2\phi_1} - \frac{1+8\eta_j}{c_0(1+8\eta_j)s_j}\frac{\|\boldsymbol{v}_{\mathcal{S}_j}\|_1^2}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} \geq \frac{\phi_0}{2\phi_1} - \frac{1}{c_0} > 0,\end{aligned}$$

where we have used the results that $\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 4\eta_j\|\Delta_{\mathcal{S}_j}\|_1$ on event $\mathcal{A}$ and the fact that $\|\boldsymbol{v}_{\mathcal{S}_j}\|_1 \leq \sqrt{s_j}\|\boldsymbol{v}_{\mathcal{S}_j}\|_2$. Thus, we have finished Step 3.

Lastly, based on the results from Steps 1-3, we find the ℓ_∞ bound of the error of $\hat{\boldsymbol{\beta}}_j$. For $l \in [(p-1)(q+1)]$, it is true that

$$\Big(\boldsymbol{\Psi}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\Big)_l = \boldsymbol{\Psi}_{ll}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)_l + \sum_{k\neq l}\boldsymbol{\Psi}_{kl}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)_k.$$

Given (S27), we have

$$\left|\Big(\boldsymbol{\Psi}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\Big)_l - \boldsymbol{\Psi}_{ll}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)_l\right| \leq \frac{3}{c_0(1+8\eta_j)s_j}\sum_{k\neq l}|(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)_k|,$$

and also

$$\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_\infty \leq 2\phi_1\|\boldsymbol{\Psi}\Delta\|_\infty + \frac{6\phi_1}{c_0\phi_0(1+8\eta_j)s_j}\|\Delta\|_1. \tag{S28}$$

With $\Delta = \hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j$, and conditioning on $\|\boldsymbol{\Psi}\Delta\|_\infty \leq \frac{3\eta_j\lambda}{2}$ and $\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 4\eta_j\|\Delta_{\mathcal{S}_j}\|_1$ from Step 1, we have that

$$\frac{\|\boldsymbol{W}\Delta\|_2^2}{n} \leq \|\boldsymbol{\Psi}\Delta\|_\infty\|\Delta\|_1 \leq \frac{3\eta_j\lambda}{2}(1+4\eta_j)\sqrt{s_j}\|\Delta_{\mathcal{S}_j}\|_2, \tag{S29}$$

while Step 3 also gives that $\|\boldsymbol{W}\Delta\|_2^2/n \geq \{\phi_0/(2\phi_1) - 1/c_0\}\|\Delta_{\mathcal{S}_j}\|_2^2$. Combining the above two inequalities yields that $\{\phi_0/(2\phi_1) - 1/c_0\}\|\Delta_{\mathcal{S}_j}\|_2^2 \leq \frac{3\eta_j\lambda}{2}(1+4\eta_j)\sqrt{s_j}\|\Delta_{\mathcal{S}_j}\|_2$, and, therefore,

$$\|\Delta_{\mathcal{S}_j}\|_2 \leq 3\eta_j\lambda(1+4\eta_j)\frac{c_0\phi_1}{c_0\phi_0 - 2\phi_1}\sqrt{s_j}.$$

With $\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 4\eta_j\|\Delta_{\mathcal{S}_j}\|_1$, it follows that $\|\Delta\|_1 \leq (1+4\eta_j)\|\Delta_{\mathcal{S}_j}\|_1 \leq (1+4\eta_j)\sqrt{s_j}\|\Delta_{\mathcal{S}_j}\|_2$, and, therefore,

$$\|\Delta\|_1 \leq 3\eta_j\lambda(1+4\eta_j)^2\frac{c_0\phi_1}{c_0\phi_0-2\phi_1}s_j.$$

Plugging this into (S28), we get the desired result that with probability at least $1-C_1'\exp\{-C_2'\log p\}$,

$$\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\|_\infty \leq \left(3\phi_1\eta_j+\frac{18\phi_1^2(1+4\eta_j)^2\eta_j}{\phi_0(c_0\phi_0-2\phi_1)(1+8\eta_j)}\right)\lambda,$$

for some positive constants C_1', C_2'.

S3.3 Proof of Theorem 3

We start the proof by noting that the two events in (19) hold with the specified probability by applying Lemma 5, which is applicable because

$$\sup_{i\in[n]}\mathrm{Var}(z_j^{(i)})=\mathcal{O}(1),\ \frac{1}{n}\sum_{i=1}^n\mathrm{Var}(z_j^{(i)})=\mathcal{O}(1),$$

due to $\{\boldsymbol{\Sigma}(\boldsymbol{u}^{(i)})\}_{jj}\leq\phi_2$, $j\in[p]$ as assumed in Assumption 2. In what follows, we show (20), conditional on that the two events, as specified in (19), hold.

When $\boldsymbol{\Gamma}$ is unknown, compared to the oracle regression equation $\boldsymbol{z}_j=\boldsymbol{W}\boldsymbol{\beta}_j+\boldsymbol{\epsilon}_j$, we only have access to the noisy equation

$$\hat{\boldsymbol{z}}_j=\hat{\boldsymbol{W}}\boldsymbol{\beta}_j+\boldsymbol{E}_j,$$

where $\boldsymbol{E}_j=\boldsymbol{\epsilon}_j+(\hat{\boldsymbol{z}}_j-\boldsymbol{z}_j)+(\boldsymbol{W}-\hat{\boldsymbol{W}})\boldsymbol{\beta}_j$. As $\hat{\boldsymbol{\beta}}_j$ is a minimizer of the convex objective function (18), Lemma 1 implies

$$\begin{aligned}&\frac{1}{2n}\|\hat{\boldsymbol{z}}_j-\hat{\boldsymbol{W}}\hat{\boldsymbol{\beta}}_j\|_2^2+\lambda\|\hat{\boldsymbol{\beta}}_j\|_1+\lambda_g\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2}+\frac{1}{2n}\|\hat{\boldsymbol{W}}(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)\|_2^2\\ \leq\ &\frac{1}{2n}\|\hat{\boldsymbol{z}}_j-\hat{\boldsymbol{W}}\boldsymbol{\beta}_j\|_2^2+\lambda\|\boldsymbol{\beta}_j\|_1+\lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}.\end{aligned}$$

With $\Delta=\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j$, reorganizing terms in the above inequality leads to

$$\frac{1}{n}\|\hat{\boldsymbol{W}}\Delta\|_2^2+\lambda\|\hat{\boldsymbol{\beta}}_j\|_1+\lambda_g\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2}\leq\frac{1}{n}\langle\boldsymbol{E}_j,\hat{\boldsymbol{W}}\Delta\rangle+\lambda\|\boldsymbol{\beta}_j\|_1+\lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}.$$

Next, with $\Delta_{\boldsymbol{E}_j} = (\hat{\boldsymbol{z}}_j - \boldsymbol{z}_j) + (\boldsymbol{W} - \hat{\boldsymbol{W}})\boldsymbol{\beta}_j$, we have that

$$\begin{aligned}\frac{1}{n}\langle \boldsymbol{E}_j, \hat{\boldsymbol{W}}\Delta\rangle &= \frac{1}{n}\langle \boldsymbol{\epsilon}_j, \hat{\boldsymbol{W}}\Delta\rangle + \frac{1}{n}\langle \Delta_{\boldsymbol{E}_j}, \hat{\boldsymbol{W}}\Delta\rangle \\ &\leq \frac{1}{n}\langle \boldsymbol{\epsilon}_j, \hat{\boldsymbol{W}}\Delta\rangle + \frac{1}{2n}\|\Delta_{\boldsymbol{E}_j}\|_2^2 + \frac{1}{2n}\|\hat{\boldsymbol{W}}\Delta\|_2^2.\end{aligned}$$

Using similar arguments as in (S2), we obtain

$$\begin{aligned}&\frac{1}{2n}\|\hat{\boldsymbol{W}}\Delta\|_2^2 + \lambda\|\Delta_{\mathcal{S}_j^c}\|_1 + \lambda_g\|\Delta_{(\mathcal{G}_j^c)}\|_{1,2} \\ &\leq \frac{1}{n}\langle \boldsymbol{\epsilon}_j, \hat{\boldsymbol{W}}\Delta\rangle + \frac{1}{2n}\|\Delta_{\boldsymbol{E}_j}\|_2^2 + \lambda\|\Delta_{\mathcal{S}_j}\|_1 + \lambda_g\|\Delta_{(\mathcal{G}_j)}\|_{1,2}.\end{aligned} \tag{S30}$$

Consider the two stochastic terms $\langle \boldsymbol{\epsilon}_j, \hat{\boldsymbol{W}}\Delta\rangle$ and $\|\Delta_{\boldsymbol{E}_j}\|_2^2$. For the latter, recall that our proof is conditional on the events of (19). Then, with probability at least $1 - \exp(\log p + \log q - \tau_1 \log q)$,

$$\frac{1}{n}\|\Delta_{\boldsymbol{E}_j}\|_2^2 \leq \frac{1}{n}\|\hat{\boldsymbol{z}}_j - \boldsymbol{z}_j\|_2^2 + \frac{\max_j \|\boldsymbol{W}_{\cdot j} - \hat{\boldsymbol{W}}_{\cdot j}\|_2^2}{n} \cdot \|\boldsymbol{\beta}_j\|_1^2 \precsim \sigma_{\epsilon_j}^2 \frac{t\log q}{n}, \tag{S31}$$

where the last inequality is true due to Assumption 5 and (19).

Next, we bound $\langle \boldsymbol{\epsilon}_j, \hat{\boldsymbol{W}}\Delta\rangle$. Defining $\hat{\mathcal{S}}_j = \left\{l : (\hat{\boldsymbol{\beta}}_j)_l \neq 0, l \in [(p-1)(q+1)]\right\}$ and letting $\tilde{\mathcal{S}}_j = \mathcal{S}_j \cup \hat{\mathcal{S}}_j$, we have that

$$\begin{aligned}\langle \boldsymbol{\epsilon}_j, \hat{\boldsymbol{W}}\Delta\rangle &= \langle \boldsymbol{\epsilon}_j, \hat{\boldsymbol{\mathcal{P}}}_{\tilde{\mathcal{S}}_j}\hat{\boldsymbol{W}}_{\tilde{\mathcal{S}}_j}\Delta_{\tilde{\mathcal{S}}_j}\rangle \\ &= \langle \hat{\boldsymbol{\mathcal{P}}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j, \hat{\boldsymbol{W}}\Delta\rangle \leq \frac{1}{2a_1}\|\hat{\boldsymbol{W}}\Delta\|_2^2 + \frac{a_1}{2}\|\hat{\boldsymbol{\mathcal{P}}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j\|_2^2,\end{aligned} \tag{S32}$$

where $\hat{\boldsymbol{\mathcal{P}}}_{\tilde{\mathcal{S}}_j}$ is the orthogonal projection matrix onto the column space of $\hat{\boldsymbol{W}}_{\tilde{\mathcal{S}}_j}$. Using the same argument as in (S4), we have

$$\|\hat{\boldsymbol{\mathcal{P}}}_{\tilde{\mathcal{S}}_j}\boldsymbol{\epsilon}_j\|_2^2 < M\sigma_{\epsilon_j}^2\left\{(s_j + \hat{s}_j)\log(ep) + (s_{j,g} + \hat{s}_{j,g})\log(eq/s_{j,g})\right\} + \hat{r}, \tag{S33}$$

where

$$\hat{r} = \sup_{\substack{1\leq s\leq (p-1)(q+1)\\ 0\leq s_g\leq q}}\left(\sup_{|\mathcal{J}|=s, |\mathcal{G}(\mathcal{J})|=s_g}\|\hat{\boldsymbol{\mathcal{P}}}_{\mathcal{J}}\boldsymbol{\epsilon}_j\|_2^2 - M\sigma_{\epsilon_j}^2\left\{s\log(ep) + s_g\log(eq/s_g)\right\}\right)_+.$$

Setting $M = 9$, by Step 1 in Section S3.1, we have

$$\mathbb{P}\{\hat{r} \geq t\sigma_{\epsilon_j}^2\} < \sum_{s=1}^{(p-1)(q+1)} \sum_{s_g=0}^{q} c_1 \exp(-c_2 t) \exp[-3c_2 \{s\log(ep) + s_g \log(eq/s_g)\}].$$

We move to bound $\|\hat{\boldsymbol{\mathcal{P}}}_{\tilde{\mathcal{S}}_j} \boldsymbol{\epsilon}_j\|_2^2$ by using the computational optimality of $\hat{\boldsymbol{\beta}}_j$. As in (S5) and (S6), it follows that

$$\lambda^2 \hat{s}_j + {\lambda_g}^2 \hat{s}_{j,g} \leq \frac{1}{n^2} \|\hat{\boldsymbol{W}}_{\hat{\mathcal{S}}_j} (\hat{\boldsymbol{z}}_j - \hat{\boldsymbol{W}} \hat{\boldsymbol{\beta}}_j)\|_2^2. \tag{S34}$$

We also have that $\|\hat{\boldsymbol{W}}_{\hat{\mathcal{S}}_j} - \boldsymbol{W}_{\hat{\mathcal{S}}_j}\|/\sqrt{n} \precsim \sqrt{\hat{s}_j t \log q/n}$ conditional on the events of (19). As $\hat{s}_j < s_\lambda = \mathcal{O}(n^{1/2})$, $t = o(n^{1/3})$and $\log q = \mathcal{O}(n^{1/6})$, we have $\|\hat{\boldsymbol{W}}_{\hat{\mathcal{S}}_j} - \boldsymbol{W}_{\hat{\mathcal{S}}_j}\|/\sqrt{n} = o(1)$. Together with the result $\|\boldsymbol{W}_{\hat{\mathcal{S}}_j}\|^2/n \leq M_1$ from Step 3 in Section S3.1, we have that $\|\hat{\boldsymbol{W}}_{\hat{\mathcal{S}}_j}\|^2 \leq M_3$ for some $M_3 > 0$. It then follows from the Cauchy-Schwarz inequality that

$$\begin{aligned} \lambda^2 \hat{s}_j + {\lambda_g}^2 \hat{s}_{j,g} \quad &\leq \quad \frac{3M_3}{n} \|\hat{\boldsymbol{W}} \Delta\|_2^2 \\ &+ \quad \frac{3M_3}{n} \|\hat{\boldsymbol{\mathcal{P}}}_{\tilde{\mathcal{S}}_j} \boldsymbol{\epsilon}_j\|_2^2 + \frac{3M_3}{n} \|\Delta_{\boldsymbol{E}_j}\|_2^2. \end{aligned} \tag{S35}$$

Set $\lambda = C\sigma_{\epsilon_j} \sqrt{\{\log(ep)/n + s_{j,g}\log(eq/s_{j,g})/(ns_j)\}}$ and $\lambda_g = \sqrt{s_j/s_{j,g}}\lambda$, where $C = 3(a_2 M_3)^{1/2}$ for some $a_2 > 0$. Combining (S33) and (S35), we have that

$$\begin{aligned} (1 - \frac{3}{a_2}) \|\hat{\boldsymbol{\mathcal{P}}}_{\tilde{\mathcal{S}}_j} \boldsymbol{\epsilon}_j\|_2^2 \quad &\leq \quad 9\sigma_{\epsilon_j}^2 \{s_j \log(ep) + s_{j,g}\log(eq/s_{j,g})\} \\ &+ \quad \frac{3}{a_2} \|\hat{\boldsymbol{W}} \Delta\|_2^2 + \frac{3}{a_2} \|\Delta_{\boldsymbol{E}_j}\|_2^2 + \hat{r}. \end{aligned} \tag{S36}$$

Plugging the above inequality and (S32) into (S30), we have

$$\begin{aligned} &\frac{\|\hat{\boldsymbol{W}} \Delta\|_2^2}{2n} + \lambda \|\Delta_{\mathcal{S}_j^c}\|_1 + \lambda_g \|\Delta_{(\mathcal{G}_j^c)}\|_{1,2} \\ \leq \quad &\frac{1}{2a_1} \frac{\|\hat{\boldsymbol{W}} \Delta\|_2^2}{n} + \frac{9a_1 a_2}{2(a_2 - 3)} \frac{\sigma_{\epsilon_j}^2 \{s_j \log(ep) + s_{j,g}\log(eq/s_{j,g})\}}{n} \\ &+ \frac{3a_1}{2(a_2 - 3)} \frac{\|\hat{\boldsymbol{W}} \Delta\|_2^2}{n} + \frac{a_1 a_2}{2(a_2 - 3)n} \hat{r} + \frac{1}{2n} \|\Delta_{\boldsymbol{E}_j}\|_2^2 \\ &+ \frac{a_1 a_2}{2(a_2 - 3)n} \|\Delta_{\boldsymbol{E}_j}\|_2^2 + \lambda \|\Delta_{\mathcal{S}_j}\|_1 + \lambda_g \|\Delta_{(\mathcal{G}_j)}\|_{1,2}. \end{aligned} \tag{S37}$$

As in Section S3.1, we have that

$$\frac{\|\Delta_{\mathcal{S}_j}\|_1}{\sqrt{s_j}} + \frac{\|\Delta_{(\mathcal{G}_j)}\|_{1,2}}{\sqrt{s_{j,g}}} \leq \|\Delta_{\mathcal{S}_j}\|_2 + \|\Delta_{(\mathcal{G}_j)}\|_2 \leq 2\frac{\phi_1}{\phi_0}\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2.$$

Consequently,

$$\lambda\|\Delta_{\mathcal{S}_j}\|_1 + \lambda_g\|\Delta_{(\mathcal{G}_j)}\|_{1,2} \leq 2C\frac{\phi_1}{\phi_0}\sqrt{e'_j}\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2. \tag{S38}$$

where $e'_j = \sigma^2_{\epsilon_j}\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}/n$. Combining (S37) and (S38), we have

$$\begin{aligned}
&\left\{\frac{1}{2} - \frac{1}{2a_1} - \frac{3a_1}{2(a_2-3)}\right\}\frac{\|\hat{\boldsymbol{W}}\Delta\|_2^2}{n} \\
\leq\ &\left\{\frac{9a_1a_2}{2(a_2-3)} + Ca_3\frac{\phi_1}{\phi_0}\right\}\frac{\sigma^2_{\epsilon_j}\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}}{n} \\
&+\frac{1}{a_3}\frac{\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2^2}{n} + \frac{a_1a_2}{2(a_2-2)n}\hat{r} + C''\left\{\frac{1}{2n} + \frac{a_1a_2}{2(a_2-3)n}\right\}.
\end{aligned}$$

Therefore, by choosing proper constants a_1, a_2 and a_3 (e.g., $a_1 = 4$, $a_2 = 51$, $a_3 = 4$), we have, with probability at least $1 - c_1\exp[-c'_2\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}]$,

$$\frac{\|\hat{\boldsymbol{W}}\Delta\|_2^2}{n} - \frac{\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2^2}{2n} \precsim \frac{\sigma^2_{\epsilon_j}\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}}{n}, \tag{S39}$$

where we have used the fact that

$$\begin{aligned}
&\mathbb{P}\left[\hat{r} \geq M_0\sigma^2_{\epsilon_j}\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}\right] \\
&\leq c_1\exp[-c'_2\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\}],
\end{aligned}$$

for a large constant M_0.

We then bound the difference between $\|\hat{\boldsymbol{W}}\Delta\|_2^2/n$ and $\|\boldsymbol{\Sigma}_{\boldsymbol{W}}^{1/2}\Delta\|_2^2/n$. To do this, we first show that, with probability at least $1 - c_6\exp[c_7\{\log p - (\tau_1 - 1)\log q\}]$,

$$\sup_{\boldsymbol{v}\in\mathbb{K}_0(2C_{\boldsymbol{\beta}_j}s_j)}\left|\boldsymbol{v}^\top\left(\frac{\hat{\boldsymbol{W}}^\top\hat{\boldsymbol{W}}}{n} - \boldsymbol{\Sigma}_{\boldsymbol{W}}\right)\boldsymbol{v}\right| \leq 1/L, \tag{S40}$$

where L is a large constant and $\mathbb{K}_0(2C_{\boldsymbol{\beta}_j}s_j) = \{\boldsymbol{v} : \|\boldsymbol{v}\|_0 \leq 2C_{\boldsymbol{\beta}_j}s_j \text{ and } \|\boldsymbol{v}\|_2 = 1\}$ for some positive constant $C_{\boldsymbol{\beta}_j}$. Notice that

$$\left|\boldsymbol{v}^\top\left(\frac{\hat{\boldsymbol{W}}^\top\hat{\boldsymbol{W}}}{n} - \frac{\boldsymbol{W}^\top\boldsymbol{W}}{n}\right)\boldsymbol{v}\right| \leq \frac{2|\boldsymbol{v}^\top\boldsymbol{W}^\top(\hat{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{v}|}{n} + \frac{\|(\hat{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{v}\|_2^2}{n},$$

where we have used that $\|\boldsymbol{v}\|_1 \leq \sqrt{2C_{\boldsymbol{\beta}_j} s_j}\|\boldsymbol{v}\|_2$. For the second term on the right hand-side, we have with probability at least $1 - 3\exp\{\log p - (\tau_1 - 1)\log q\}$,

$$\frac{\|(\hat{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{v}\|_2}{\sqrt{n}} \leq \frac{\max_j \|\boldsymbol{W}_{\cdot j} - \hat{\boldsymbol{W}}_{\cdot j}\|_2}{\sqrt{n}} \cdot \|\boldsymbol{v}\|_1 \precsim \sqrt{\frac{2C_{\boldsymbol{\beta}_j} s_j t \log q}{n}} = o(1),$$

and, moreover,

$$\frac{|\boldsymbol{v}^\top \boldsymbol{W}^\top (\hat{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{v}|}{n} \leq \frac{\|(\hat{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{v}\|_2}{\sqrt{n}} \times \frac{\|\boldsymbol{W}\boldsymbol{v}\|_2}{\sqrt{n}}. \tag{S41}$$

Next, using a similar argument as in Section S3.1, it follows from (S8) and Lemma 6 that with probability at least $1 - \mathcal{O}\{(pq)^{-1}\}$

$$\left| \boldsymbol{v}^\top \left(\frac{\boldsymbol{W}^\top \boldsymbol{W}}{n} - \boldsymbol{\Sigma}_{\boldsymbol{W}} \right) \boldsymbol{v} \right| = o(1).$$

Using the same argument as in (S7) and by Assumption 5, we can show $\|\boldsymbol{\Sigma}_{\boldsymbol{W}}\|$ is upper bounded by $\phi_0' \phi_1$. Thus, we have that $\|\boldsymbol{W}\boldsymbol{v}\|_2/\sqrt{n} = \mathcal{O}(1)$ with probability at least $1 - \mathcal{O}\{(pq)^{-1}\}$. Putting this together with (S41), we have shown (S40).

Next, conditioning on (S40) and using the result in Lemma 7, we have

$$\left| \Delta^\top \left(\frac{\hat{\boldsymbol{W}}^\top \hat{\boldsymbol{W}}}{n} - \boldsymbol{\Sigma}_{\boldsymbol{W}} \right) \Delta \right| \leq \frac{1}{L'} \left(\|\Delta\|_2^2 + \frac{1}{s_j} \|\Delta\|_1^2 \right). \tag{S42}$$

Plugging this into (S39), we have

$$\begin{aligned} \frac{\|\boldsymbol{W}\Delta\|_2^2}{2n} &\precsim \frac{\sigma_{\epsilon_j}^2 \{s_j \log(ep) + s_{j,g} \log(eq/s_{j,g})\}}{n} \\ &+ \frac{1}{L'} \left(\|\Delta\|_2^2 + \frac{1}{s_j} \|\Delta\|_1^2 \right) + \frac{\sigma_{\epsilon_j}^2}{n}. \end{aligned} \tag{S43}$$

By choosing an appropriate a_1 given a_2 in (S37), we have with probability at least $1 - c_1 \exp[-c_2'\{s_j \log(ep) + s_{j,g} \log(eq/s_{j,g})\}]$,

$$\frac{\|\Delta_{\mathcal{S}_j^c}\|_1}{\sqrt{s_j}} + \frac{\|\Delta_{(\mathcal{G}_j^c)}\|_{1,2}}{\sqrt{s_{j,g}}} \leq \sqrt{e_j'} + \frac{\|\Delta_{\mathcal{S}_j}\|_1}{\sqrt{s_j}} + \frac{\|\Delta_{(\mathcal{G}_j)}\|_{1,2}}{\sqrt{s_{j,g}}}. \tag{S44}$$

Adding $\|\Delta_{\mathcal{S}_j}\|_1/\sqrt{s_j}$ to both sides of (S44), we get

$$\frac{\|\Delta\|_1}{\sqrt{s_j}} \leq \sqrt{e_j'} + 3\|\Delta\|_2 \tag{S45}$$

Plugging (S45) into (S43) and by (S7), we have

$$\|\Delta\|_2^2 \precsim \frac{\sigma_{\epsilon_j}^2}{n}\{s_j \log(ep) + s_{j,g}\log(eq/s_{j,g})\} + \frac{\sigma_{\epsilon_j}^2}{n},$$

with probability at least $1 - C_3\exp[C_4\{\log p - (\tau_1 - 1)\log q\}]$, for some positive constants C_3 and C_4.

□

S3.4 Proof of Theorem 4

We first establish the ℓ_∞ norm bound of $\hat{\boldsymbol{\beta}}_j$ in (21) with three steps.

Step 1: In this step, we show that, with probability at least $1-c_8\exp[c_9\{\log p-(\tau_1-1)\log q)\}]$ for some $c_8, c_9 > 0$,

$$\left\|\hat{\boldsymbol{\Psi}}\Delta\right\|_\infty \leq \frac{3\eta_j\lambda}{2}. \tag{S46}$$

And it also holds with probability at least $1 - c_8'\exp[c_9'\{\log p - (\tau_1 - 1)\log q)\}]$ that, for some $c_8', c_9' > 0$,

$$\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 4\eta_j\|\Delta_{\mathcal{S}_j}\|_1. \tag{S47}$$

Under the KKT conditions, if $\boldsymbol{\theta}$ is an optimum of (18), then

$$\begin{cases} \left(\hat{\boldsymbol{W}}^\top(\hat{\boldsymbol{z}}_j - \hat{\boldsymbol{W}}\boldsymbol{\theta})/n\right)_l = \text{sign}(\theta_l)\lambda & \text{if } \theta_l \neq 0, l \in (0) \\ \left(\hat{\boldsymbol{W}}^\top(\hat{\boldsymbol{z}}_j - \hat{\boldsymbol{W}}\boldsymbol{\theta})/n\right)_l = \text{sign}(\theta_l)\lambda + \lambda_g\frac{\theta_l}{\|\boldsymbol{\theta}_{(h)}\|_2} & \text{if } \theta_l \neq 0,\ l \in (h) \\ \left|\left(\hat{\boldsymbol{W}}^\top(\hat{\boldsymbol{z}}_j - \hat{\boldsymbol{W}}\boldsymbol{\theta})/n\right)_l\right| < \eta_j\lambda & \text{if } \theta_l = 0. \end{cases}$$

As such, any solution $\hat{\boldsymbol{\beta}}_j$ satisfies that

$$\left\|\frac{1}{n}\hat{\boldsymbol{W}}^\top(\hat{\boldsymbol{z}}_j - \hat{\boldsymbol{W}}\hat{\boldsymbol{\beta}}_j)\right\|_\infty \leq \eta_j\lambda.$$

If we can show that with high probability

$$\frac{1}{n}\left\|\hat{\boldsymbol{W}}^\top\boldsymbol{E}_j\right\|_\infty \leq \frac{\eta_j\lambda}{2}, \tag{S48}$$

then we can reach the desired conclusion that with high probability

$$\left\|\hat{\boldsymbol{\Psi}}\Delta\right\|_\infty \leq \frac{3\eta_j\lambda}{2}.$$

Now we consider the inequality in (S48), and note that

$$\frac{1}{n}\left\|\hat{\boldsymbol{W}}^\top \boldsymbol{E}_j\right\|_\infty \leq \underbrace{\frac{1}{n}\|\boldsymbol{W}^\top\boldsymbol{\epsilon}_j\|_\infty}_{\text{I}} + \underbrace{\frac{1}{n}\|(\hat{\boldsymbol{W}}-\boldsymbol{W})^\top\boldsymbol{\epsilon}_j\|_\infty}_{\text{II}} + \underbrace{\frac{1}{n}\|\hat{\boldsymbol{W}}^\top\Delta_{\boldsymbol{E}_j}\|_\infty}_{\text{III}}.$$

Consider term (I). Define $V_l = \boldsymbol{w}_l^\top\boldsymbol{\epsilon}_j/n$, $j \in [(p-1)(q+1)]$. Using a similar argument as in (S21), we have

$$\mathbb{P}\left(|V_l| > \frac{\eta_j\lambda}{2}\right) \leq 2\exp\left(-\frac{c\eta_j^2\lambda^2 n}{4M^2\phi_2\sigma_{\epsilon_j}^2}\right) \leq 2\exp\{-C_0'(\log p + \log q)\},$$

where $C_0' = cC/(4M^2\phi_2) > 1$. Using the union bound inequality, we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\left\|\boldsymbol{W}^\top\boldsymbol{\epsilon}_j\right\|_\infty \geq \frac{\eta_j\lambda}{2}\right) \leq\ & \mathbb{P}\left(\max_l |V_l| \geq \frac{\eta_j\lambda}{2}\right)\\ \leq\ & 2\exp\{-(C_0'-1)(\log p+\log q)\}.\end{aligned}$$

For term (II), it is true that

$$\frac{1}{n}\|(\hat{\boldsymbol{W}}-\boldsymbol{W})^\top\boldsymbol{\epsilon}_j\|_\infty = \frac{1}{n}\max_l|\langle\hat{\boldsymbol{w}}_l-\boldsymbol{w}_l,\boldsymbol{\epsilon}_j\rangle| \leq \frac{\max_l\|\hat{\boldsymbol{w}}_l-\boldsymbol{w}_l\|_2}{\sqrt{n}}\cdot\frac{\|\boldsymbol{\epsilon}_j\|_2}{\sqrt{n}}.$$

Using Lemma 4, we have that for any large constant $M_4 > 0$,

$$\mathbb{P}\left(\frac{\|\boldsymbol{\epsilon}_j\|_2}{\sqrt{n}} > M_4^{1/2}\sigma_{\epsilon_j}\right) \leq \mathbb{P}\left(\left|\frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n} - \sigma_{\epsilon_j}^2\right| > (M_4-1)\sigma_{\epsilon_j}^2\right) \leq 2\exp(-b_0\log q).$$

We further have $\mathbb{P}(\|\hat{\boldsymbol{w}}_l-\boldsymbol{w}_l\|_2/\sqrt{n}) \precsim t^{1/2}\lambda_1$ with probability at least $1-3\exp(-\tau_1\log q)$ and $t^{1/2}\lambda_1 = o(\eta_j\lambda)$, which is true as $t = o(\sqrt{n/\log q})$. Applying the union bound, we have

$$\mathbb{P}\left(\frac{1}{n}\|(\hat{\boldsymbol{W}}-\boldsymbol{W})^\top\boldsymbol{\epsilon}_j\|_\infty \geq \frac{\eta_j\lambda}{2}\right) \leq b_0'\exp\{\log p+\log q-\tau_1\log q)\}. \tag{S49}$$

Moving to term (III), notice that there exists a large constant $M_4' > 0$ such that

$$\begin{aligned}\frac{1}{n}\|\hat{\boldsymbol{W}}^\top\Delta_{\boldsymbol{E}_j}\|_\infty &= \|(\boldsymbol{W}+\hat{\boldsymbol{W}}-\boldsymbol{W})^\top\Delta_{\boldsymbol{E}_j}/n\|_\infty\\ &\leq \frac{1}{n}\|\boldsymbol{W}^\top\Delta_{\boldsymbol{E}_j}\|_\infty + \frac{\max_l\|\hat{\boldsymbol{w}}_l-\boldsymbol{w}_l\|_2}{\sqrt{n}}\cdot\frac{\|\Delta_{\boldsymbol{E}_j}\|_2}{\sqrt{n}}\\ &\leq \frac{1}{n}\|\boldsymbol{W}^\top\Delta_{\boldsymbol{E}_j}\|_\infty + M_4' t\lambda_1^2,\end{aligned}$$

with probability at least $b_0'' \exp\{\log p + \log q - \tau_1 \log q)\}$, where the last inequality follows from (19) and (S31). Since $t^{1/2}\lambda_1 = o(\eta_j \lambda)$, it suffices to bound the term of $\|\boldsymbol{W}^\top \Delta_{\boldsymbol{E}_j}\|/n$. To this end, we have

$$\frac{1}{n}\|\boldsymbol{W}^\top \Delta_{\boldsymbol{E}_j}\|_\infty < \frac{\max_l \|\boldsymbol{w}_l\|_2}{\sqrt{n}} \cdot \frac{\|\Delta_{\boldsymbol{E}_j}\|_2}{\sqrt{n}}.$$

As $\boldsymbol{w}_l$ has independent sub-Gaussian entries with a bounded sub-Gaussian norm, using a similar argument as for term (II) we have

$$\mathbb{P}\left(\frac{1}{n}\|\boldsymbol{W}^\top \Delta_{\boldsymbol{E}_j}\|_\infty \geq \frac{\eta_j \lambda}{2}\right) \leq b_0''' \exp\{\log p + \log q - \tau_1 \log q)\}. \tag{S50}$$

Combining terms (I)-(III), we have finished showing (S46) of Step 1.

Next, we prove that $\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 4\eta_j \|\Delta_{\mathcal{S}_j}\|_1$. By the definition of $\hat{\boldsymbol{\beta}}_j$, we have

$$\frac{1}{2n}\|\hat{\boldsymbol{z}}_j - \hat{\boldsymbol{W}}\hat{\boldsymbol{\beta}}_j\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}_j\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2} \leq \frac{1}{2n}\|\boldsymbol{E}_j\|_2^2 + \lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}.$$

Developing the left hand side of the above inequality, we have

$$\lambda\|\hat{\boldsymbol{\beta}}_j\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2} \leq \lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2} + \frac{1}{n}\Delta^\top \hat{\boldsymbol{W}}^\top \boldsymbol{E}_j.$$

Define an event

$$\mathcal{A}_1 = \left\{\frac{1}{n}\|\boldsymbol{W}^\top \Delta_{\boldsymbol{E}_j}\|_\infty \leq \frac{\lambda}{2}\right\}. \tag{S51}$$

We again write

$$\frac{1}{n}\left\|\hat{\boldsymbol{W}}^\top \boldsymbol{E}_j\right\|_\infty \leq \underbrace{\frac{1}{n}\|\boldsymbol{W}^\top \boldsymbol{\epsilon}_j\|_\infty}_{\text{I}} + \underbrace{\frac{1}{n}\|(\hat{\boldsymbol{W}} - \boldsymbol{W})^\top \boldsymbol{\epsilon}_j\|_\infty}_{\text{II}} + \underbrace{\frac{1}{n}\|\hat{\boldsymbol{W}}^\top \Delta_{\boldsymbol{E}_j}\|_\infty}_{\text{III}}.$$

For term (I), first define $V_l = \boldsymbol{w}_l^\top \boldsymbol{\epsilon}_j/n$, $j \in [(p-1)(q+1)]$. Using a similar argument as in (S23), we have

$$\mathbb{P}\left(\frac{1}{n}\left\|\boldsymbol{W}^\top \boldsymbol{\epsilon}_j\right\|_\infty \geq \frac{\lambda}{2}\right) \leq \mathbb{P}\left(\max_l |V_l| \geq \frac{\lambda}{2}\right) \leq \exp\{-c_3' \log p\};$$

for term (II), using the same argument as in (S49) and by noting $t^{1/2}\lambda_1 = o(\lambda)$, which is true as $t = o(\sqrt{n/\log q})$, we have

$$\mathbb{P}\left(\frac{1}{n}\|(\hat{\boldsymbol{W}} - \boldsymbol{W})^\top \boldsymbol{\epsilon}_j\|_\infty \geq \frac{\lambda}{2}\right) \leq b_0' \exp\{\log p + \log q - \tau_1 \log q)\};$$

for term (III), using the same argument as in (S50) and again noting $t^{1/2}\lambda_1 = o(\lambda)$, we have

$$\mathbb{P}\left(\frac{1}{n}\|\boldsymbol{W}^\top \Delta_{\boldsymbol{E}_j}\|_\infty \geq \frac{\lambda}{2}\right) \leq b_0''' \exp\{\log p + \log q - \tau_1 \log q)\}.$$

Combining terms (I)-(III), we have $\mathbb{P}(\mathcal{A}_1^c) \leq c_8' \exp[c_9'\{\log p - (\tau_1 - 1)\log q)\}]$. Conditioning on $\mathcal{A}_1$ in (S51), we have that

$$2\|\hat{\boldsymbol{\beta}}_j\|_1 + 2\sqrt{\frac{s_j}{s_{j,g}}}\|\hat{\boldsymbol{\beta}}_{j,-0}\|_{1,2} \leq 2\|\boldsymbol{\beta}_j\|_1 + 2\sqrt{\frac{s_j}{s_{j,g}}}\|\boldsymbol{\beta}_{j,-0}\|_{1,2} + \|\Delta\|_1.$$

The rest of the arguments is similar to Step 1 in the proof of Theorem 2, which leads us to the desired result of $\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 4\eta_j \|\Delta_{\mathcal{S}_j}\|_1$ given $\mathcal{A}_1$.

Step 2: In this step, we bound the diagonal and off-diagonal elements of $\hat{\boldsymbol{\Psi}}$. First, we consider the diagonal elements of $\hat{\boldsymbol{\Psi}}$, i.e., $\hat{\boldsymbol{\Psi}}_{ll}$'s. Note that

$$\frac{\|\boldsymbol{w}_l\|_2}{\sqrt{n}} - \frac{\|\hat{\boldsymbol{w}}_l - \boldsymbol{w}_l\|_2}{\sqrt{n}} \leq \frac{\|\hat{\boldsymbol{w}}_l\|_2}{\sqrt{n}} \leq \frac{\|\boldsymbol{w}_l\|_2}{\sqrt{n}} + \frac{\|\hat{\boldsymbol{w}}_l - \boldsymbol{w}_l\|_2}{\sqrt{n}}.$$

Since $\|\hat{\boldsymbol{w}}_l - \boldsymbol{w}_l\|_2/\sqrt{n} = \mathcal{O}(t^{1/2}\lambda_1) = o(1)$ and $\mathbb{P}(\boldsymbol{\Psi}_{ll} \notin [\phi_0/(2\phi_1), 2M^2\phi_2]) \leq 2\exp(-c_4 n)$ from (S25), we have that $\phi_0/(3\phi_1) \leq \hat{\boldsymbol{\Psi}}_{ll} \leq 3M^2\phi_2$ with probability at least $1 - 2\exp(-c_4' n)$. Next, we consider the off-diagonal elements. It holds for $k \neq l$ that

$$\|\boldsymbol{w}_k^\top \boldsymbol{w}_l - \hat{\boldsymbol{w}}_k^\top \hat{\boldsymbol{w}}_l\|_2 \leq \|\boldsymbol{w}_k^\top (\boldsymbol{w}_l - \hat{\boldsymbol{w}}_l)\|_2 + \|\hat{\boldsymbol{w}}_l^\top (\hat{\boldsymbol{w}}_k - \boldsymbol{w}_k)\|_2.$$

Since $\|\hat{\boldsymbol{w}}_l - \boldsymbol{w}_l\|_2/\sqrt{n} = \mathcal{O}(t^{1/2}\lambda_1) = o(1)$ and (S26) shows

$$\mathbb{P}\left\{\boldsymbol{\Psi}_{kl} \notin [-\frac{1}{c_0(1+8\eta_j)s_j}, \frac{3}{c_0(1+8\eta_j)s_j}]\right\} \leq 2\exp(-c_5 n),$$

the following holds with probability at least $1-\exp(-c_5' n)$,

$$\max_{k\neq l} \hat{\mathbf{\Psi}}_{kl} \in [-\frac{2}{c_0(1+8\eta_j)s_j}, \frac{4}{c_0(1+8\eta_j)s_j}].$$

Step 3: We show that conditional on $\mathcal{A}_1$ and that

$$\max_{l} \hat{\mathbf{\Psi}}_{ll} \in [\phi_0/(3\phi_1), 3M^2\phi_2], \quad \max_{k\neq l} \hat{\mathbf{\Psi}}_{kl} \in [-\frac{2}{c_0(1+8\eta_j)s_j}, \frac{4}{c_0(1+8\eta_j)s_j}], \tag{S52}$$

it holds that

$$\min_{\|\boldsymbol{v}_{\mathcal{S}_j^c}\|_1 \leq 3\eta_j \|\boldsymbol{v}_{\mathcal{S}_j}\|_1} \frac{\|\hat{\boldsymbol{W}}\boldsymbol{v}\|_2}{\sqrt{n}\|\boldsymbol{v}_{\mathcal{S}_j}\|_2} \geq \sqrt{\frac{\phi_0}{3\phi_1} - \frac{2}{c_0}} > 0.$$

First, given (S52), we have that

$$\begin{aligned}\frac{\|\hat{\boldsymbol{W}}\boldsymbol{v}_{\mathcal{S}_j}\|_2^2}{n\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} &= \frac{\boldsymbol{v}_{\mathcal{S}_j}^\top \mathrm{diag}(\hat{\mathbf{\Psi}})\boldsymbol{v}_{\mathcal{S}_j}}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} + \frac{\boldsymbol{v}_{\mathcal{S}_j}^\top(\hat{\mathbf{\Psi}} - \mathrm{diag}(\hat{\mathbf{\Psi}}))\boldsymbol{v}_{\mathcal{S}_j}}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} \\ &\geq \frac{\phi_0}{3\phi_1} - \frac{2}{c_0(1+8\eta_j)s_j}\frac{\|\boldsymbol{v}_{\mathcal{S}_j}\|_1^2}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2}.\end{aligned}$$

Furthermore, given (S52) and $\mathcal{A}_1$, we have that

$$\begin{aligned}\frac{\|\hat{\boldsymbol{W}}\boldsymbol{v}\|_2^2}{n\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} &\geq \frac{\|\hat{\boldsymbol{W}}\boldsymbol{v}_{\mathcal{S}_j}\|_2^2}{n\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} + 2\frac{v_{\mathcal{S}_j}^\top \hat{\mathbf{\Psi}}\boldsymbol{v}_{\mathcal{S}_j^c}}{n\|\boldsymbol{v}_{\mathcal{S}_j}\|_2} \\ &\geq \frac{\phi_0}{3\phi_1} - \frac{2}{c_0(1+8\eta_j)s_j}\frac{\|\boldsymbol{v}_{\mathcal{S}_j}\|_1^2}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} - \frac{4}{c_0(1+8\eta_j)s_j}\frac{\|\boldsymbol{v}_{\mathcal{S}_j}\|_1\|\boldsymbol{v}_{\mathcal{S}_j^c}\|_1}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} \\ &\geq \frac{\phi_0}{3\phi_1} - \frac{2(1+8\eta_j)}{c_0(1+8\eta_j)s_j}\frac{\|\boldsymbol{v}_{\mathcal{S}_j}\|_1^2}{\|\boldsymbol{v}_{\mathcal{S}_j}\|_2^2} \geq \frac{\phi_0}{3\phi_1} - \frac{2}{c_0} > 0,\end{aligned}$$

where we have used the results that $\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 4\eta_j\|\Delta_{\mathcal{S}_j}\|_1$ and the fact that $\|\boldsymbol{v}_{\mathcal{S}_j}\|_1 \leq \sqrt{s_j}\|\boldsymbol{v}_{\mathcal{S}_j}\|_2$.

Lastly, with results from Steps 1-3, we find the ℓ_∞ bound of $\boldsymbol{\beta}_j$. For $l \in [(p-1)(q+1)]$, it is true that

$$\left(\hat{\mathbf{\Psi}}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\right)_l = \hat{\mathbf{\Psi}}_{ll}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)_l + \sum_{k\neq l}\hat{\mathbf{\Psi}}_{kl}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)_k$$

Given (S52) from Step 2, we have

$$\left|\left(\hat{\mathbf{\Psi}}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\right)_l - \hat{\mathbf{\Psi}}_{ll}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)_l\right| \leq \frac{4}{c_0(1+8\eta_j)s_j}\sum_{k\neq l}|(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)_k|,$$

SNP	co-expressed genes
rs10492975	(CALML5,PIK3R2), (CALML5,CAMK1)
rs723211	(HRAS,CALML5), (CALML5,CAMK1G)
rs1347069	(SHC4,CDKN2A)
rs473698	(PRKCG,CAMK1)
rs4118334	(SHC2,CAMK1)
rs882664	(PRKCA,CAMK1)
rs1267622	(SHC3,RAF1)

Table S6: Identified co-expression QTLs and the corresponding co-expressed genes.

and also

$$\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_\infty \leq 3\phi_1 \left\|\hat{\boldsymbol{\Psi}}\Delta\right\|_\infty + \frac{12\phi_1}{c_0\phi_0(1+8\eta_j)s_j}\|\Delta\|_1. \tag{S53}$$

With $\Delta = \hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j$ and given $\left\|\hat{\boldsymbol{\Psi}}\Delta\right\|_\infty \leq \frac{3\eta_j\lambda}{2}$ and $\|\Delta_{\mathcal{S}_j^c}\|_1 \leq 4\eta_j\|\Delta_{\mathcal{S}_j}\|_1$ from Step 1, we have

$$\frac{\|\hat{\boldsymbol{W}}\Delta\|_2^2}{n} \leq \|\hat{\boldsymbol{\Psi}}\Delta\|_\infty\|\Delta\|_1 \leq \frac{3\eta_j\lambda}{2}(4\eta_j+1)\sqrt{s_j}\|\Delta_{\mathcal{S}_j}\|_1.$$

We also have from Step 3 that $\|\hat{\boldsymbol{W}}\Delta\|_2^2/n \geq \{\phi_0/(3\phi_1) - 2/c_0\}\|\Delta_{\mathcal{S}_j}\|_2^2$ given (S52) and $\mathcal{A}_1$. Combining these two inequalities and by noting $\|\Delta\|_1 \leq (1+4\eta_j)\sqrt{s_j}\|\Delta_{\mathcal{S}_j}\|_2$, we have that

$$\|\Delta\|_1 \leq \frac{3\eta_j\lambda}{2}(1+4\eta_j)^2\frac{3c_0\phi_1}{c_0\phi_0 - 6\phi_1}s_j.$$

Plugging this into (S53), we obtain that

$$\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_\infty \leq \frac{9}{2}\left\{\phi_1\eta_j + \frac{12\phi_1^2(1+4\eta_j)^2}{\phi_0(c_0\phi_0 - 6\phi_1)(1+8\eta_j)}\right\}\lambda.$$

with probability at least $1 - C_5\exp[C_6\{\log p - (\tau_1 - 1)\log q\}]$, for some positive constants C_5 and C_6.

S4 Additional results of data analysis

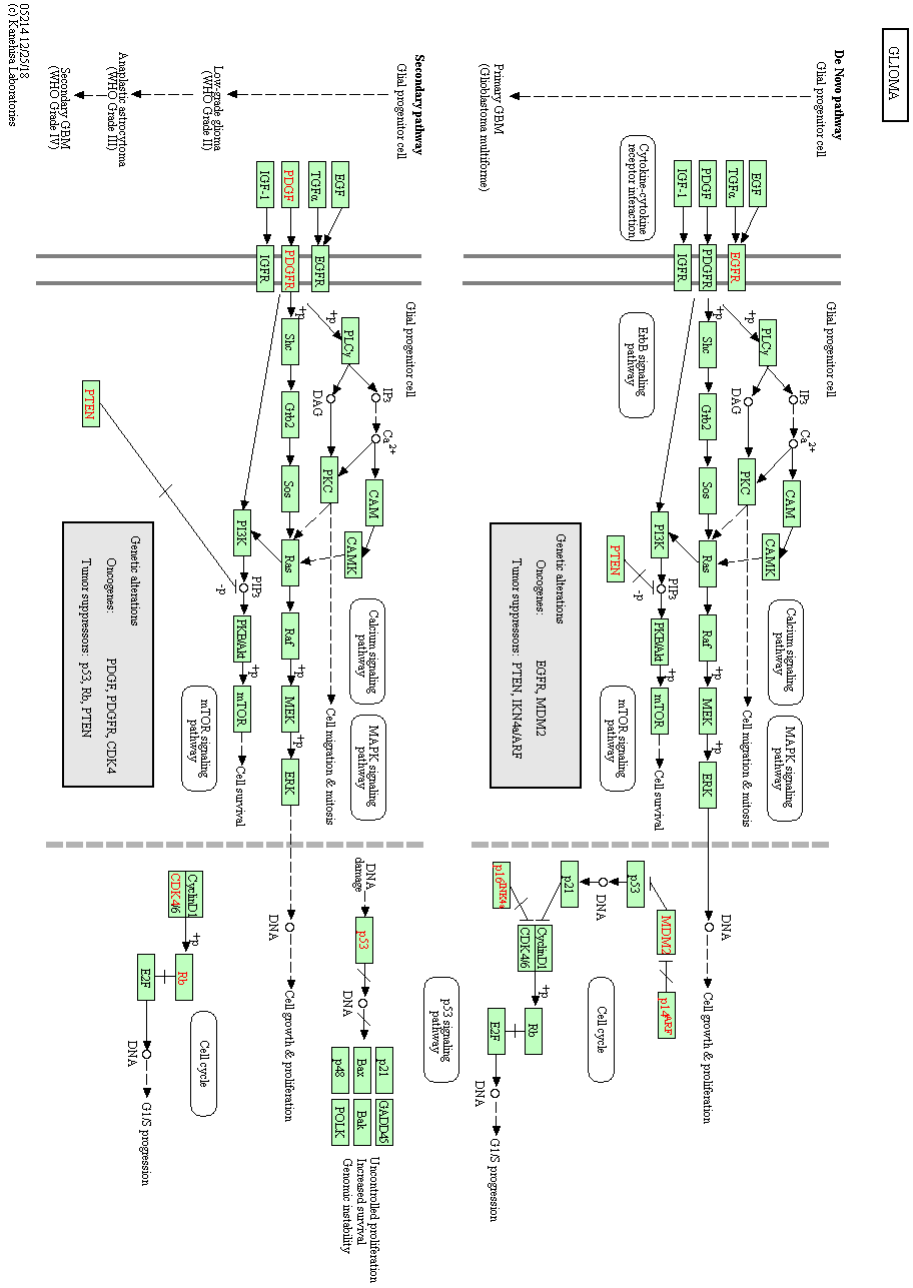

Figure S1: The KEGG human glioma pathway. This figure is downloaded from https://www.genome.jp/kegg/ (Kanehisa and Goto, 2000).

Additional references

Balasubramanian, K., Fan, J., and Yang, Z. (2018), Tensor methods for additive index models under discordance and heterogeneity, *arXiv preprint arXiv:1807.06693*.

Bellec, P. C., Dalalyan, A. S., Grappin, E., and Paris, Q. (2018), On the prediction loss of the lasso in the partially labeled setting, *Electronic Journal of Statistics*, 12, 3443–3472.

Graybill, F. A. and Marsaglia, G. (1957), Idempotent matrices and quadratic forms in the general linear hypothesis, *The Annals of Mathematical Statistics*, 28, 678–686.

Kuchibhotla, A. K. and Chakrabortty, A. (2018), Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression, *arXiv preprint arXiv:1804.02605*.

Laurent, B. and Massart, P. (2000), Adaptive estimation of a quadratic functional by model selection, *Annals of Statistics*, 1302–1338.

Loh, P.-L. and Wainwright, M. J. (2011), High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity, in *Advances in Neural Information Processing Systems*, pp. 2726–2734.

Vershynin, R. (2010), Introduction to the non-asymptotic analysis of random matrices, *arXiv preprint arXiv:1011.3027*.