Natural Covariate-adjusted Gaussian Graphical Regression

Ruobin Liu and Guo Yu

Department of Statistics and Applied Probability, University of California, Santa Barbara, Santa Barbara, CA, USA

June 8, 2024

Abstract

Gaussian graphical models (GGMs) are widely used for recovering the conditional independence structure among random variables. Recently, several key advances have been made to exploit an additional set of variables for better estimating the GGMs of the variables of interest. For example, in co-expression quantitative trait locus (eQTL) studies, both the mean expression level of genes as well as their pairwise conditional independence structure may be adjusted by genetic variants local to those genes. Existing methods to estimate covariate-adjusted GGMs either allow only the mean to depend on covariates or suffer from poor scaling assumptions due to the inherent non-convexity of simultaneously estimating the mean and precision matrix. In this paper, we propose a convex formulation that jointly estimates the covariate-adjusted mean and precision matrix by utilizing the natural parametrization of the multivariate Gaussian likelihood. This convexity yields theoretically better performance as the sparsity and dimension of the covariates grow large relative to the number of samples. We verify our theoretical results with numerical simulations and perform a reanalysis of an eQTL study of glioblastoma multiforme (GBM), an aggressive form of brain cancer.

1 Introduction

Graphical models are used to represent the distribution of a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ by relating its conditional independence structure to a graph. This correspondence is particularly salient when $\boldsymbol{X}$ is Gaussian. Letting $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ be the precision matrix of a Gaussian random vector $\boldsymbol{X}$, for $i \neq j$, the components X_i and X_j are conditionally independent given $\boldsymbol{X}_{\{1,\ldots,p\}\setminus\{i,j\}}$ if and only if $\Omega_{ij} = 0$ (Lauritzen, 1996). Models that estimate the conditional independence structure by imposing sparsity on $\boldsymbol{\Omega}$ are known as Gaussian graphical models (GGMs).

There has been considerable research on GGMs and their estimation in the high-dimensional setting (Yuan and Lin, 2007; Friedman et al., 2007; Meinshausen and Bühlmann, 2006). GGMs find major applications in genomics where the network structure of genes is of interest. Such a setting is typically high-dimensional such that the number of genes exceeds the number of samples (Schäfer and Strimmer, 2005). Of particular interest in genomics are co-expression quantitative trait locus (eQTL) studies. In these studies, gene expression data are analyzed alongside information about external genetic markers that are known to confound the expression levels of genes. In order to use GGMs for eQTL studies, one must allow the GGM framework to incorporate covariate information.

Let $\boldsymbol{X} \in \mathbb{R}^p$ be the vector of responses and $\boldsymbol{U} \in \mathbb{R}^q$ the associated vector of covariates. We consider a general model

$$\boldsymbol{X} \mid \boldsymbol{U} = \mathbf{u} \sim N(\boldsymbol{\mu}(\mathbf{u}), \boldsymbol{\Omega}^{-1}(\mathbf{u})) \tag{1}$$

and a specification given by

$$\boldsymbol{\mu}(\mathbf{u}) = \boldsymbol{\Gamma}\mathbf{u}, \quad \boldsymbol{\Omega}(\mathbf{u}) = \mathbf{B}_0. \tag{2}$$

Multivariate regression models (with random design) specify (2) with the goal to improve estimation of $\boldsymbol{\Gamma}$ by also estimating $\mathbf{B}_0$; see Yuan et al. (2007) and references therein for this well-studied setting. By comparison, the focus of the covariate-adjusted GGM framework is to estimate the sparsity pattern of $\boldsymbol{\Omega}(\mathbf{u})$. Several methods exist in the GGM framework to allow the mean of the responses to depend on covariates. For instance, Rothman et al. (2010); Yin and Li (2011); Cai et al. (2013); Chen et al. (2016, 2018) consider GGMs of the form (2) so that the mean of the responses may depend on covariates $\mathbf{u}$ while the graph structure, determined by $\mathbf{B}_0$, does not. These models are termed *conditional Gaussian graphical models* in Yin and Li (2011). Jointly estimating $(\boldsymbol{\Gamma}, \mathbf{B}_0)$ in (2) is challenging. Therefore, covariate-adjusted GGM methods use either alternating (Rothman et al., 2010; Yin and Li, 2011; Chen et al., 2018) or two-stage (Cai et al., 2013; Chen et al., 2016) estimation procedures. Despite these challenges, minimax-optimal error rates in jointly estimating (2) have been established (Chen et al., 2018; Lv et al., 2022).

Although there has been progress in allowing heterogeneous means in GGMs, there is very little work on allowing the graph structure of $\boldsymbol{\Omega}$ to also depend on covariates. Yet this is plausible in the context of eQTL analyses. For example, gene C may mediate the co-expression of genes A and B, but only in the presence of a single-nucleotide polymorphism (SNP) local to gene C. In other words, the expression of genes A and B may be conditionally independent given the rest of the network unless a genetic variant is present near gene C (Fehrmann et al., 2011; Kolberg et al., 2020; Rockman and Kruglyak, 2006). To capture this additional structure, Zhang and Li (2022) extends (2) to allow both the mean and the network structure of the responses to be modified by covariates. Their model specifies (1) by taking

$$\boldsymbol{\mu}(\mathbf{u}) = \boldsymbol{\Gamma}\mathbf{u}, \quad \boldsymbol{\Omega}(\mathbf{u}) = \mathbf{B}_0 + \sum_{h=1}^{q} \mathbf{B}_h u_h \tag{3}$$

where $\mathbf{B}_h, h = 0, 1 \ldots, q$ are sparse, symmetric matrices. Like previous works in covariate-adjusted graphical models, the joint estimation of the parameters $(\boldsymbol{\Gamma}, \mathbf{B}_0, \mathbf{B}_1, \ldots, \mathbf{B}_h)$ in (3) involves a non-jointly-convex objective. Hence Zhang and Li (2022) use a two-stage estimation procedure.

In this work, we contribute a jointly convex formulation of (1) such that both the mean and the graph structure may depend on covariates. Like Zhang and Li (2022), our method uses nodewise regression to estimate the covariate-adjusted graph structure (Meinshausen and Bühlmann, 2006). However, we base our formulation upon a natural parametrization of the multivariate Gaussian likelihood such that each nodewise regression is a convex optimization problem. We recall the Gaussian graphical regression framework of Zhang and Li (2022) and motivate our parametrization in Section 2. In Section 3 we provide the model specification and algorithm to induce a sparse-group structure in the graph. Our theoretical results are discussed in Section 4, namely that the natural parametrization allows for better theoretical scaling of p and q relative to n under the same assumptions as in Zhang and Li (2022). This is demonstrated through extensive simulations in Section 5. Finally we apply our method to perform a reanalysis of data from glioblastoma microforme (GBM) tissue samples in Section 6.

2 Parametrizations for Covariate-adjusted Graphical Models

Let $\boldsymbol{X} \in \mathbb{R}^p$ be a random vector of responses and $\boldsymbol{U} \in \mathbb{R}^q$ the corresponding covariates. In Zhang and Li (2022), the dependence of $\boldsymbol{X}$ on $\boldsymbol{U}$ is given by (1) and (3) so that both the mean and precision matrix are covariate-dependent. Write $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p)^\top$ and let $\Omega_{jk}(\mathbf{u})$ denote the (j,k)-th element of $\boldsymbol{\Omega}(\mathbf{u})$, keeping in mind its dependence on $\mathbf{u}$. Then (3) may be estimated via a neighborhood regression method (Meinshausen and Bühlmann, 2006), which amounts to estimating p separate linear regression models of the form

$$X_j = \mathbf{u}^\top \boldsymbol{\gamma}_j + \sum_{k \neq j}^{p} \beta_{jk0}(X_k - \mathbf{u}^\top \boldsymbol{\gamma}_k) + \sum_{k \neq j}^{p} \sum_{h=1}^{q} \beta_{jkh} u_h (X_k - \mathbf{u}^\top \boldsymbol{\gamma}_k) + \varepsilon_j \tag{4}$$

where $\beta_{jkh} = -[\mathbf{B}_h]_{jk}/\Omega_{jj}$ and $\varepsilon_j \sim N(0, 1/\Omega_{jj})$ for all j, k, and h. This is termed *Gaussian graphical regression* in Zhang and Li (2022).

Note that a least squares criterion based on (4) is not jointly convex owing to the cross term $\beta_{jkh} \times \mathbf{u}^\top \boldsymbol{\gamma}_k$. Hence, Zhang and Li (2022) uses a two-stage estimation method. First, $\boldsymbol{\gamma}_j$ is estimated for all j via the model

$$X_j = \mathbf{u}^\top \boldsymbol{\gamma}_j + \xi_j, \quad \mathbb{E}\, \xi_j = 0, \tag{5}$$

and ℓ_1 penalization is added so that $\hat{\boldsymbol{\gamma}}_j$ is sparse. Second, the observed response vectors $\mathbf{x}_j$ are centered using $\hat{\boldsymbol{\gamma}}_j$ and the coefficients β_{jkh} in (4) are estimated with $\hat{\boldsymbol{\gamma}}_j$ in place of $\boldsymbol{\gamma}_j$ for all $j \in [p]$. As with other two-stage methods, the above procedure incurs an estimation error in the first stage because (5) ignores the dependence on the β_{jkh} terms in (4). Assumptions on the scaling of the ambient dimensions p and q and the sparsity of the coefficient vectors are needed to suppress the model misspecification errors. Next, we will present a formulation of (1) so that the corresponding least squares criterion is convex, obviating the need for a two-stage procedure.

2.1 Convex Formulation

Recall the form of the p-dimensional multivariate Gaussian likelihood expressed in terms of the natural parameters $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Theta}})$ where $\tilde{\boldsymbol{\theta}} = \boldsymbol{\Omega}\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\Theta}} = -\boldsymbol{\Omega}$:

$$L(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Theta}} \mid \mathbf{x}) = \exp\left\{ \tilde{\boldsymbol{\theta}}^\top \mathbf{x} - \frac{1}{2} \mathbf{x}^\top \tilde{\boldsymbol{\Theta}} \mathbf{x} - A(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Theta}}) \right\}.$$

Our formulation is motivated by the fact that the cumulant function A is jointly convex in $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Theta}})$ and therefore so is $-\log L(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Theta}} \mid \mathbf{x})$. Define

$$\boldsymbol{\theta} = \operatorname{diag}(\boldsymbol{\Omega})^{-1} \boldsymbol{\Omega} \boldsymbol{\mu} \quad \text{and} \quad \boldsymbol{\Theta} = -\operatorname{diag}(\boldsymbol{\Omega})^{-1} \boldsymbol{\Omega}, \tag{6}$$

where $\mathrm{diag}(\boldsymbol{\Omega})$ is the $p \times p$ diagonal matrix of $\boldsymbol{\Omega}$. Our model is to incorporate covariates as in (3) to $(\boldsymbol{\theta}, \boldsymbol{\Theta})$ instead of $(\boldsymbol{\mu}, \boldsymbol{\Omega})$, namely

$$\boldsymbol{\theta}(\mathbf{u}) = \boldsymbol{\Gamma}\mathbf{u}, \quad \boldsymbol{\Theta}(\mathbf{u}) = \mathbf{B}_0 + \sum_{h=1}^{q} \mathbf{B}_h u_h. \tag{7}$$

To see that this leads to jointly convex nodewise regression problems, fix $j \in [p]$ and consider the partial regression of component X_j against the covariates $\boldsymbol{U}$ and the remaining components $\boldsymbol{X}_{-j}$ in the general setting (1). For a matrix $\mathbf{M}$ and sets of indices $\mathcal{I}$ and $\mathcal{J}$, denote by $\mathbf{M}_{\mathcal{I},\mathcal{J}}$ the sub-matrix of $\mathbf{M}$ consisting of rows indexed by $\mathcal{I}$ and columns indexed by $\mathcal{J}$. Letting $-j$ indicate all indices excluding j, we have the conditional distribution

$$X_j - \mu_j \mid \boldsymbol{X}_{-j}, \mathbf{U} \sim N\big(\boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}(\boldsymbol{X}_{-j} - \boldsymbol{\mu}_{-j}),\ \boldsymbol{\Sigma}_{jj} - \boldsymbol{\Sigma}_{j,-j}\boldsymbol{\Sigma}_{-j,-j}^{-1}\boldsymbol{\Sigma}_{-j,j}\big), \tag{8}$$

suppressing the dependence of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ on the covariates $\mathbf{u}$ for notational convenience. Defining the error terms $\varepsilon_j \sim N(0, \sigma_{\varepsilon_j}^2)$ where $\sigma_{\varepsilon_j}^2 = 1/\Omega_{jj}$, by the matrix block inversion formula we may write (8) as the linear model

$$X_j = \mu_j - \Omega_{jj}^{-1}\boldsymbol{\Omega}_{j,-j}(\boldsymbol{X}_{-j} - \boldsymbol{\mu}_{-j}) + \varepsilon_j = \theta_j + \boldsymbol{\Theta}_{j,-j}\boldsymbol{X}_{-j} + \varepsilon_j,$$

which is readily seen to be jointly convex in the parameters θ_j and $\boldsymbol{\Theta}_{j,-j}$ as defined in (6). In light of the specification of covariate dependence (7) and with $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)^\top$ and $\beta_{jkh} = [\mathbf{B}_h]_{jk}$, the above leads to the nodewise regression model

$$X_j = \mathbf{u}^\top \boldsymbol{\gamma}_j + \sum_{k \neq j}^{p} \beta_{jk0} X_k + \sum_{k \neq j}^{p} \sum_{h=1}^{q} \beta_{jkh} u_h X_k + \varepsilon_j, \quad \varepsilon_j \sim N(0, \sigma_{\varepsilon_j}^2) \tag{9}$$

which allows for simultaneous estimation of the parameters $\boldsymbol{\gamma}_j$ and β_{jkh} in a jointly convex setting.

Implicit in (6) is that $\Theta_{jj} = -1$ for all $j \in [p]$. This is satisfied in (7) by letting $[\mathbf{B}_0]_{jj} = -1$ and $[\mathbf{B}_h]_{jj} = 0$ for all $h \in [q]$ and $j \in [p]$. By doing so, we assume as in Zhang and Li (2022) that the residual variance $\sigma_{\varepsilon_j}^2$ is not covariate-dependent, allowing us to write $\Omega_{jj} = \Omega_{jj}(\mathbf{u})$ for $j \in [p]$.

3 Estimation

Suppose we collect n independent observations $\{(\mathbf{x}^{(i)}, \mathbf{u}^{(i)})\}_{i=1}^{n}$ of responses and covariates that follow the joint distribution in (1) with (6) and (7). Let $\mathbf{U} = \big[\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)}\big]^\top \in \mathbb{R}^{n \times q}$ be the matrix of covariates and $\mathbf{X} = \big[\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\big]^\top$ the matrix of responses with $\mathbf{x}_j \in \mathbb{R}^n$ denoting the j-th column of $\mathbf{X}$ and $\mathbf{u}_j \in \mathbb{R}^n$ the j-th column of $\mathbf{U}$. For $j \in [p]$, let $\mathbf{W}_{-j}$ be the $n \times (p-1)(q+1)$ matrix of interactions of the remaining responses with the covariates. Concretely,

$$\begin{aligned}
\mathbf{W}_{-j,0} &= [\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times (p-1)}, \\
\mathbf{W}_{-j,h} &= [\mathbf{x}_1 \odot \mathbf{u}_h,\ \ldots,\ \mathbf{x}_{j-1} \odot \mathbf{u}_h,\ \mathbf{x}_{j+1} \odot \mathbf{u}_h,\ \ldots,\ \mathbf{x}_p \odot \mathbf{u}_h] \text{ for } h \in [q], \\
\mathbf{W}_{-j} &= [\mathbf{W}_{-j,0},\ \mathbf{W}_{-j,1},\ \ldots,\ \mathbf{W}_{-j,q}] \in \mathbb{R}^{n \times (p-1)(q+1)},
\end{aligned}$$

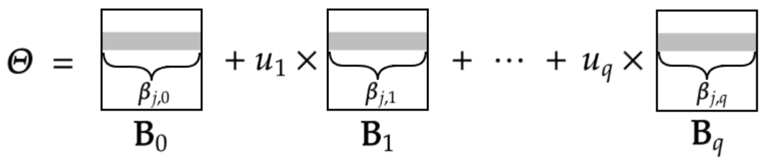

Figure 1: Decomposition of $\mathbf{\Theta}$ into components $\mathbf{B}_h$ according to (7). The block $\boldsymbol{\beta}_{j,h}$ corresponds to the effects of covariate u_h on the partial correlations of response X_j while $\boldsymbol{\beta}_{j,0}$ describes the population effect.

where $\odot$ denotes the elementwise product of two vectors. By writing

$\boldsymbol{\beta}_{j,h} = (\beta_{j1h}, \dots, \beta_{j,j-1,h}, \beta_{j,j+1,h}, \dots, \beta_{jph})^\top \in \mathbb{R}^{p-1}$ and $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{j,0}, \boldsymbol{\beta}_{j,1}, \dots, \boldsymbol{\beta}_{j,q})^\top \in \mathbb{R}^{(p-1)(q+1)}$,

we can view $\boldsymbol{\beta}_j$ as $q+1$ blocks of $(p-1)$-element vectors $\boldsymbol{\beta}_{j,h}$ for $h = \{0, 1, \dots q\}$ so that (9) may be written in the block form

$$X_j = \mathbf{u}^\top \boldsymbol{\gamma}_j + \boldsymbol{X}_{-j}^\top \boldsymbol{\beta}_{j,0} + \sum_{h=1}^{q} (u_h \boldsymbol{X}_{-j})^\top \boldsymbol{\beta}_{j,h} + \varepsilon_j, \quad \varepsilon_j \sim N(0, \sigma_{\varepsilon_j}^2). \tag{10}$$

Figure 1 relates the blocks $\boldsymbol{\beta}_{j,h}$ to $\mathbf{\Theta}$ in (7).

The joint convexity of (10) in $(\boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)$ allows for simultaneous estimation of the penalized problem

$$\underset{\boldsymbol{\gamma}_j, \boldsymbol{\beta}_j}{\text{minimize}} \ \frac{1}{2n} \left\| \mathbf{x}_j - \mathbf{U}\boldsymbol{\gamma}_j - \mathbf{W}_{-j}\boldsymbol{\beta}_j \right\|_2^2 + g_j(\boldsymbol{\gamma}_j, \boldsymbol{\beta}_j) \tag{11}$$

where g_j is an arbitrary convex penalty function. Moreover, by the relationship between $\boldsymbol{\beta}_j$ and rows of $\mathbf{\Theta}$, sparsity-inducing penalties g_j in (11) will give us sparse estimates of $\mathbf{\Theta}$. From (6), we see that $\mathbf{\Theta}$ has the same sparsity pattern as $\mathbf{\Omega}$. Hence we keep the conditional independence interpretation of the sparsity in $\mathbf{\Theta}$ as in $\mathbf{\Omega}$.

3.1 Estimating Covariate-adjusted Graphs

In the nodewise regression approach to GGMs, a symmetrization step is needed to ensure that the estimate of $\mathbf{\Omega}$ is symmetric because each regression is fit independently (Meinshausen and Bühlmann, 2006). In our approach, the nodewise regressions target $\mathbf{\Theta}$, which is not necessarily symmetric due to the different scaling factor for each row in (6). To ensure symmetry in the estimate of $\mathbf{\Omega}$, we perform a similar symmetrization step after performing all nodewise regressions. First, specify a function $H(\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\beta}}_j)$ that estimates the error variance in (10) as $\hat{\sigma}_{\varepsilon_j}^2$. Then for an estimate $\hat{\boldsymbol{\beta}}_j$, set $\tilde{\boldsymbol{\beta}}_j = -\hat{\boldsymbol{\beta}}_j / \hat{\sigma}_{\varepsilon_j}^2$ and define for all $h = 0, 1, \dots, q$ the symmetric matrix

$$[\tilde{\mathbf{B}}_h]_{jk} = [\tilde{\mathbf{B}}_h]_{kj} = \tilde{\beta}_{jkh} \Big[|\tilde{\beta}_{jkh}| < |\tilde{\beta}_{kjh}| \Big] + \tilde{\beta}_{kjh} \Big[|\tilde{\beta}_{jkh}| > |\tilde{\beta}_{kjh}| \Big], \tag{12}$$

where the expression $[P]$ is equal to 1 if P is true and 0 otherwise. The mean vector and precision matrix may then be estimated by (6) via

$$\hat{\mathbf{\Omega}}(\mathbf{u}^{(i)}) = \tilde{\mathbf{B}}_0 + \sum_{h=1}^{q} \tilde{\mathbf{B}}_h u_h^{(i)} \quad \text{and} \quad \hat{\boldsymbol{\mu}}(\mathbf{u}^{(i)}) = (\hat{\mathbf{\Omega}}(\mathbf{u}^{(i)}))^{-1} \operatorname{diag}(\hat{\mathbf{\Omega}}) \hat{\mathbf{\Gamma}} \mathbf{u}^{(i)}. \tag{13}$$

The preceding steps are summarized in Algorithm 1 in Section S1.

Equation (12) is the "and-rule" to symmetrize the matrices $\tilde{\mathbf{B}}_h$ for $h = 0, 1, \ldots, q$; $[\tilde{\mathbf{B}}_h]_{jk}$ is nonzero if both $\tilde{\beta}_{jkh}$ and $\tilde{\beta}_{kjh}$ are nonzero. A less conservative estimate would be given by the "or-rule", namely

$$[\tilde{\mathbf{B}}_h]_{jk} = [\tilde{\mathbf{B}}_h]_{kj} = \tilde{\beta}_{jkh}\Big[|\tilde{\beta}_{jkh}| \geq |\tilde{\beta}_{kjh}|\Big] + \tilde{\beta}_{kjh}\Big[|\tilde{\beta}_{jkh}| \leq |\tilde{\beta}_{kjh}|\Big],$$

so that $[\tilde{\mathbf{B}}_h]_{jk}$ is nonzero if either $\tilde{\beta}_{jkh}$ or $\tilde{\beta}_{kjh}$ is nonzero. We elect to use the more conservative rule and note that both approaches are asymptotically equivalent (Meinshausen and Bühlmann, 2006) and that the and-rule has been considered before in covariate-adjusted graphical models (Cai et al., 2013; Zhang and Li, 2022).

3.2 Sparse-group Structure

With the application to eQTL studies in mind, we will focus on a particular sparsity-inducing penalty g_j in (14). We wish to identify *elementwise* sparsity within a group h, amounting to sparsity in the coefficient vector $\boldsymbol{\beta}_{j,h}$. The interpretation is that a covariate u_h affects the conditional independence of X_j and some, but not all other responses X_k. At the same, we wish to identify *groupwise* sparsity, where $\boldsymbol{\beta}_{j,h} = \mathbf{0}$ for certain groups h. This means that u_h has no effect on the conditional independence of X_j and the other responses.

We consider the following convex problem:

$$\underset{\boldsymbol{\gamma}_j, \boldsymbol{\beta}_j}{\text{minimize}} \; \frac{1}{2n}\big\|\mathbf{x}_j - \mathbf{U}\boldsymbol{\gamma}_j - \mathbf{W}_{-j}\boldsymbol{\beta}_j\big\|_2^2 + \eta\|\boldsymbol{\gamma}_j\|_1 + \lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2} \tag{14}$$

where $\eta \geq 0$, $\lambda \geq 0$, and $\lambda_g \geq 0$ are tuning parameters and $\|\boldsymbol{\beta}_{j,-0}\|_{1,2} = \sum_{h=1}^{q}\|\boldsymbol{\beta}_{j,h}\|_2$ is the group lasso penalty. Together, the penalty on $\boldsymbol{\beta}_j$ is the sparse-group lasso penalty developed in Simon et al. (2013). The group lasso penalty is not applied to the coefficients governing the population graph. Following the lasso literature (Reid et al., 2016; Yu and Bien, 2019), a straightforward estimate of $\sigma^2_{\varepsilon_j}$ is given by

$$\hat{\sigma}^2_{\varepsilon_j} = \frac{\|\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}}_j - \mathbf{W}_{-j}\hat{\boldsymbol{\beta}}_j\|_2^2}{n - \hat{s}_{\beta_j} - \hat{s}_{\gamma_j}},$$

where $\hat{\boldsymbol{\gamma}}_j$ and $\hat{\boldsymbol{\beta}}_j$ are solutions of (14) and $\hat{s}_{\gamma_j}$ and $\hat{s}_{\beta_j}$ are the number of nonzero elements of $\hat{\boldsymbol{\gamma}}_j$ and $\hat{\boldsymbol{\beta}}_j$.

3.3 Implementation

We use block coordinate descent to solve (14) where the blocks are given by $\{\boldsymbol{\gamma}_j, \boldsymbol{\beta}_{j,0}, \ldots, \boldsymbol{\beta}_{j,q}\}$. Since (14) is convex and separable in these blocks, this approach is guaranteed to converge to the optimal solution (Tseng, 2001). The block update steps are given in Section S1.1 with the full algorithm presented in Algorithm 2. In our implementation, we use k-fold cross-validation to select the tuning parameter triplet $(\eta, \lambda, \lambda_g)$. To do so, we set a parameter $\lambda_0 > 0$ and mixture parameters $\alpha_g, \alpha_s \in [0, 1]$ so that the penalty in (14) may be written as

$$\alpha_g\lambda_0\|\boldsymbol{\gamma}_j\|_1 + (1-\alpha_g)\alpha_s\lambda_0\|\boldsymbol{\beta}_j\|_1 + (1-\alpha_g)(1-\alpha_s)\lambda_0\|\boldsymbol{\beta}_{j,-0}\|_{1,2}. \tag{15}$$

Then for fixed α_g and α_s, we run the method on a path of λ_0, taking advantage of warm restarts. Tuning all three parameters may not be necessary in practice as the performance is fairly robust to fixed choices of α_s; see Figure 3.

4 Theoretical Properties

In this section we will analyze the estimation error and support recovery of (14). For two sequences of real numbers a_n and b_n, we write $a_n \precsim b_n$ if $a_n = O(b_n)$, i.e. there exists constants $C > 0$ and $N > 0$ so that $a_n < Cb_n$ for all $n \geq N$. If $a_n \precsim b_n$ and $b_n \precsim a_n$, we write $a_n \asymp b_n$. We write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. With the understanding that j is fixed, we will suppress the subscript j when referring to $\boldsymbol{\gamma}_j$, $\boldsymbol{\beta}_j$ and $\boldsymbol{\varepsilon}_j$ for notational convenience.

Let $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ be the solution to (14) and let $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ be the true parameters. Denote by S_γ and S_β the support sets of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ respectively. Further let $S_{\beta,g}$ index the active groups of $\boldsymbol{\beta}$, that is $S_{\beta,g} = \{h : \boldsymbol{\beta}_{j,h} \neq \mathbf{0},\ h \in [q]\}$. Denote by s_γ, s_β and $s_{\beta,g}$ the cardinalities of these sets. For a square matrix $\mathbf{M}$, denote by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ its minimum and maximum eigenvalue, respectively.

We require the following assumptions on the model (1) with specification (7).

Assumption 1. *The covariates $\{\mathbf{u}^{(i)}\}_{i=1}^n$ are i.i.d. mean zero random vectors with covariance matrix satisfying*

$$\phi_0 \leq \lambda_{\min}(\mathrm{Cov}(\mathbf{u}^{(i)})) \leq \lambda_{\max}(\mathrm{Cov}(\mathbf{u}^{(i)})) \leq \phi_1$$

for some constants $0 < \phi_0 \leq \phi_1 < \infty$. Furthermore, there exists a constant $M > 0$ such that $|\mathbf{u}_j^{(i)}| \leq M$ for all $j \in [q]$ and $i \in [n]$.

Assumption 1 is the same as Assumptions 1 and 5 in Zhang and Li (2022) without requiring a bound on $\|\beta\|_1$. The boundedness of the covariates is not restrictive in the eQTL setting since SNPs are often binary coded.

Denote by $[\mathbf{U}, \mathbf{W}_{-j}]$ the $n \times (p(q+1)-1)$ matrix that is the concatenation of the covariate matrix $\mathbf{U}$ with the interaction matrix $\mathbf{W}_{-j}$. Define the following Gram matrix of covariates, responses, and interactions:

$$\boldsymbol{\Sigma}_{\mathbf{UW}} = \mathbb{E}\left(\frac{[\mathbf{U}, \mathbf{W}_{-j}]^\top [\mathbf{U}, \mathbf{W}_{-j}]}{n}\right).$$

Our next assumption bounds the eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{UW}}$, which is needed to characterize the joint distribution of $\mathbf{u}^{(i)}$, $\mathbf{x}^{(i)}$, and their interactions.

Assumption 2. *We assume there exist positive constants m_0, M_0 such that*

$$m_0 \leq \lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{UW}}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{UW}}) \leq M_0.$$

Elementwise boundedness in Assumption 1 implies that $\mathbf{u}^{(i)}$ is elementwise sub-Gaussian. Our theoretical analysis further requires that our design matrix $[\mathbf{U}, \mathbf{W}_{-j}]$ is elementwise sub-Gaussian. Assumption 3 ensures this since each entry of $\mathbf{W}_{-j}$ is the product of a sub-Gaussian and a bounded random variable.

Assumption 3. *The marginal distribution of $\boldsymbol{X}$ is elementwise sub-Gaussian with bounded sub-Gaussian norm.*

The following assumption controls how the sparsity of the true parameters may scale with the sample size n.

Assumption 4. *Let p and q be the number of responses and covariates, respectively. The true sparsities s_γ and s_β satisfy*

1. $(s_\gamma + s_\beta)\sqrt{\log(pq)} = o(\sqrt{n})$,
2. $(s_\gamma + s_\beta)\log(pq) = o(n/\log n)$.

In practice we consider estimates $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ of varying sparsities and select from this pool of candidate models via cross-validation. Define $\hat{s}_\gamma^{\max}$ and $\hat{s}_\beta^{\max}$ to be the maximum sparsity of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ out of all candidate models, chosen so that $s_\gamma < s_\gamma^{\max}$ and $s_\beta < s_\beta^{\max}$. Our first theorem describes the ℓ_2 estimation error of the nodewise solution.

Theorem 1. *Under Assumptions 1-4, with $(\hat{s}_\gamma^{\max} + \hat{s}_\beta^{\max})\log(pq) = O(\sqrt{n})$, and with*

$$\begin{aligned}
\eta &= C\frac{\sigma_\varepsilon}{\sqrt{n}}\left(\log(eq/s_\gamma) + \frac{s_\beta}{s_\gamma}\log(ep) + \frac{s_{\beta,g}}{s_\gamma}\log(eq/s_{\beta,g})\right)^{1/2},\\
\lambda &= C\frac{\sigma_\varepsilon}{\sqrt{n}}\left(\frac{s_\gamma}{s_\beta}\log(eq/s_\gamma) + \log(ep) + \frac{s_{\beta,g}}{s_\beta}\log(eq/s_{\beta,g})\right)^{1/2},\\
\lambda_g &= C\frac{\sigma_\varepsilon}{\sqrt{n}}\left(\frac{s_\gamma}{s_{\beta,g}}\log(eq/s_\gamma) + \frac{s_\beta}{s_{\beta,g}}\log(ep) + \log(eq/s_{\beta,g})\right)^{1/2},
\end{aligned} \tag{16}$$

we have

$$\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \precsim \frac{\sigma_\varepsilon^2}{n}(s_\gamma \log(eq/s_\gamma) + s_\beta \log(ep) + s_{\beta,g}\log(eq/s_{\beta,g})).$$

with probability at least $1 - C_1\exp\{-C_2(s_\gamma\log(eq/s_\gamma) + s_\beta\log(ep) + s_{\beta,g}\log(eq/s_{\beta,g}))\}$ *where C, C_1, and C_2 are positive constants.*

The bound on the sparsity of candidate models helps to bound the errors in Theorem 1 and is also assumed in Zhang and Li (2022).

Our second result proves that our method achieves support recovery with high probability under some additional conditions. Denote the (k,ℓ)-th entry of $\boldsymbol{\Sigma}_{\mathbf{UW}}$ by $\boldsymbol{\Sigma}_{\mathbf{UW}}(k,\ell)$.

Assumption 5. *Define*

$$\tau_j = 1 + \max\left(\sqrt{\frac{s_\gamma}{s_\beta}} + \sqrt{\frac{s_\gamma}{s_{\beta,g}}}, \sqrt{\frac{s_\beta}{s_\gamma}} + \sqrt{\frac{s_\beta}{s_{\beta,g}}}, \sqrt{\frac{s_{\beta,g}}{s_\gamma}} + \sqrt{\frac{s_{\beta,g}}{s_\beta}}\right).$$

We assume that there exists a constant $c_0 > 2/m_0$ such that

$$\max_{k\neq\ell}|\boldsymbol{\Sigma}_{\mathbf{UW}}(k,\ell)| \leq \frac{1}{c_0(1+16\tau_j)(s_\beta + s_\gamma)}.$$

Theorem 2. *Suppose Assumptions 1-5 hold. Additionally, suppose* $\log p \asymp \log q$ *and that*

$$n \geq A_1(s_\gamma \log(eq/s_\gamma) + s_\beta \log(ep) + s_{\beta,g} \log(eq/s_{\beta,g}))$$

for some constant $A_1 > 0$*. Then with* η, λ, *and* λ_g *defined as in Theorem 1, we have*

$$\max\Big\{\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_\infty, \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty\Big\} \leq \frac{3}{m_0}(\eta + \lambda + \lambda_g)\left(1 + \frac{6(1+8\tau_j)^2}{(1+16\tau_j)(c_0 m_0 - 2)}\right)$$

with probability at least $1 - C_3 \exp(-C_4 \log p)$ *for some positive constants* C_3, C_4*.*

Theorem 2 implies that our method achieves support recovery with high probability if $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ satisfy a minimal signal strength condition.

Corollary 1. *Define*

$$\hat{S}_\gamma = \left\{k : |\hat{\gamma}_k| > \frac{3}{m_0}(\eta + \lambda + \lambda_g)\left(1 + \frac{6(1+8\tau_j)^2}{(1+16\tau_j)(c_0 m_0 - 2)}\right)\right\},$$
$$\hat{S}_\beta = \left\{k : |\hat{\beta}_k| > \frac{3}{m_0}(\eta + \lambda + \lambda_g)\left(1 + \frac{6(1+8\tau_j)^2}{(1+16\tau_j)(c_0 m_0 - 2)}\right)\right\}$$

If it holds that

$$\min\left\{\min_{\ell \in S_\gamma}|\gamma_\ell|, \min_{k \in S_\beta}|\beta_k|\right\} \geq 2 \cdot \frac{3}{m_0}(\eta + \lambda + \lambda_g)\left(1 + \frac{6(1+8\tau_j)^2}{(1+16\tau_j)(c_0 m_0 - 2)}\right),$$

we have that $\mathbb{P}(\hat{S}_\gamma = S_\gamma$ *and* $\hat{S}_\beta = S_\beta) \geq 1 - C_3 \exp(-C_4 \log p)$*.*

Zhang and Li (2022) provides two theorems related to the estimation error of $\hat{\boldsymbol{\beta}}$; an *oracle* rate where the true $\boldsymbol{\Gamma}$ in (3) is known and a rate where $\hat{\boldsymbol{\Gamma}}$ is estimated via independent regressions in stage 1 of the procedure. It is shown that the oracle rate, given by

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \precsim \frac{\sigma_{\varepsilon_j}^2}{n}(s_\beta \log(ep) + s_{\beta,g} \log(eq/s_{\beta,g})),$$

can be achieved by the two-stage rate under the assumptions $\log(pq) = O(n^{1/6})$, $s_\beta = o(n^{1/6})$, and $s_\gamma = o(n^{1/3})$. By comparison, Theorem 1 shows that our method achieves the oracle rate under milder assumptions; more explicitly, Assumption 4 is satisfied when $\log(pq) = O(n^{1/3})$ and $s_\gamma + s_\beta = o(n^{1/3})$. The milder scaling assumptions in $\log(pq)$ and s_β in our formulation stems from the crucial fact that (14) is jointly convex in $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, allowing for their simultaneous estimation. By comparison, a more stringent scaling assumption is needed to control the estimation error of $\boldsymbol{\gamma}_j$ incurred in stage 1 of Zhang and Li (2022). These comments apply as well to the support recovery results in Theorem 2.

5 Simulation study

In this section, we compare our method *Natural Covariate-adjusted Gaussian Graphical Regression* (`ncagr`) with the Gaussian graphical model regression method developed in Zhang and Li (2022) (`RegGMM`) under extensive simulations. We are interested in the behavior of the two methods under data generated according to both models.

For each simulation setting, we generate a $\mathbf{\Gamma} \in \mathbb{R}^{p \times q}$ sparse matrix by selecting uniformly $s_g = 250$ of its elements to be 0.5. We generate the population graph $\mathbf{B}_0$ using a preferential attachment algorithm (Barabási and Albert, 1999) with power 1.1. We select $q_e = 5$ out of q covariates to have nonzero effects and generate the covariate-adjusted components $\mathbf{B}_h$ for $h \in [q_e]$ as Erdős-Renyi graphs (Erdős and Rényi, 1959) with edge probability $v_e = 0.01$; the remaining $\mathbf{B}_h$ are identically zero. Our choice of random graph models are such that the covariate-adjusted graphs $\mathbf{B}_h$ are more sparse than the population graph $\mathbf{B}_0$. See Clauset et al. (2009) for a discussion of random graph models. After determining the graph structure, we generate entries in $\mathbf{B}_h$ by sampling uniformly from $[-0.5, -0.35] \cup [0.35, 0.5]$ and then scaling each row by the row sum to ensure diagonal dominance

To generate a covariate vector $\mathbf{u}^{(i)} \in \mathbb{R}^q$, we select each entry from $\mathrm{Unif}(0, 1)$ and then standardize each covariate across the n observations to have mean zero and unit variance. With each set of $\mathbf{\Gamma}$ and $\mathbf{B}_h$, we generate a dataset under the original model (3) by setting $\boldsymbol{\mu}^{(i)} = \mathbf{\Gamma}\mathbf{u}^{(i)}$ and $\mathbf{\Omega}^{(i)} = \mathbf{B}_0 + \sum_{h=1}^{q} \mathbf{B}_h u_h^{(i)}$ and a second dataset under (7), termed the "natural model", via $\mathbf{\Theta}^{(i)} = \mathbf{B}_0 + \sum_{h=1}^{q} \mathbf{B}_h u_h^{(i)}$, $\mathbf{\Omega}^{(i)} = -\mathbf{\Theta}^{(i)}$, and $\boldsymbol{\mu}^{(i)} = \mathbf{\Sigma}^{(i)}\mathbf{\Gamma}\mathbf{u}^{(i)}$. Finally, we sample a vector of responses $\mathbf{x}^{(i)} \sim N_p(\boldsymbol{\mu}^{(i)}, \mathbf{\Sigma}^{(i)})$ for $i \in [n]$. In both models, we set $\mathrm{diag}(\mathbf{\Omega}^{(i)}) = 1$ so that $\sigma^2_{\varepsilon_j} = 1$.

For each of the settings $p = 25, q = 50$; $p = 50, q = 50$; and $p = 25, q = 100$, we generate 100 independent data sets of $n = 200$ and $n = 400$ samples and run both `RegGMM` and `ncagr` on each data set. For a fair comparison, both methods have two tuning parameters selected via 5-fold cross-validation; on a path of 100 λ parameters and 10 mixture parameters. With the notation in (15), this means cross-validating over 100 values of λ_0 and $\alpha_s = 0, 0.1, \ldots, 1$, while $\alpha_g = 0.1$ is fixed. In Figure 3, we show that the performance is fairly robust to choices of α_s, suggesting that fixing α_s is reasonable.

We report the mean and standard error of the following metrics after applying (12): TPR_β, the true positive rate of nonzero entries of $\tilde{\mathbf{B}}_h$; FPR_β, the false positive rate; and the estimation error of $\tilde{\boldsymbol{\beta}}_j$ given by $\boldsymbol{\beta}_{\text{err}} = \sum_{j=1}^{p} \|\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_2$. The results are shown in Table 1. We see `ncagr` outperforms `RegGMM` in the natural model and both `RegGMM` and `ncagr` show improved performance on data generated under the natural model compared to that generated under the original model. We also give the in-sample prediction error of $\hat{\mathbf{\Omega}}$ and $\hat{\boldsymbol{\mu}}$ from the simulations as in (13); see Table S1.

Next, we investigate through simulations the scaling performance in the sample size n. We fixed $p = 25, q = 100$ and considered a varying sample size of $n = 100, 110, \ldots, 400$, generating 20 independent data sets of size n at each value. In Figure 2, we plot the TPR_β and $\boldsymbol{\beta}_{\text{err}}$ of both `RegGMM` and `ncagr`. We see that `ncagr` outperforms `RegGMM` in terms of both support recovery and squared error. This comparison verifies the theoretical results of Section 4, that laxer scaling assumptions in n and $\log(pq)$ are needed to guarantee good performance.

Finally, we show that the performance is fairly robust to choices of α_s in (15), suggesting that fixing α_s is reasonable. We ran `ncagr` over the path $\alpha_s = 0.05, 0.10, \ldots, 0.95, 1.00$ and performed 5-fold cross-validation over 100 values of λ_0 and $\alpha_g = 0.1, 0.2, \ldots, 1$. This was repeated for 20

Model	n	(p, q)	Method	TPR_β	FPR_β	$\boldsymbol{\beta}_{\text{err}}$
natural	200	(25, 50)	`ncagr`	**0.991** (0.015)	0.006 (0.001)	**3.767** (0.370)
			`RegGMM`	0.798 (0.059)	**0.004** (0.001)	5.109 (0.392)
		(50, 50)	`ncagr`	**0.666** (0.043)	**0.003** (0.000)	**11.081** (0.410)
			`RegGMM`	0.392 (0.047)	0.004 (0.001)	13.169 (0.342)
		(25, 100)	`ncagr`	**0.998** (0.007)	0.007 (0.001)	7.163 (0.900)
			`RegGMM`	0.812 (0.065)	**0.001** (0.000)	**4.742** (0.496)
	400	(25, 50)	`ncagr`	**0.999** (0.004)	0.004 (0.001)	**2.235** (0.172)
			`RegGMM`	0.921 (0.040)	**0.003** (0.001)	4.248 (0.265)
		(50, 50)	`ncagr`	**0.879** (0.028)	**0.002** (0.000)	**8.239** (0.294)
			`RegGMM`	0.686 (0.050)	0.009 (0.002)	10.774 (0.351)
		(25, 100)	`ncagr`	**1.000** (0.000)	0.004 (0.001)	**2.546** (0.208)
			`RegGMM`	0.925 (0.039)	**0.001** (0.000)	3.859 (0.294)
original	200	(25, 50)	`ncagr`	0.510 (0.075)	0.010 (0.002)	9.178 (0.242)
			`RegGMM`	**0.636** (0.073)	**0.005** (0.002)	**6.610** (0.352)
		(50, 50)	`ncagr`	0.149 (0.029)	**0.003** (0.000)	15.278 (0.779)
			`RegGMM`	**0.195** (0.040)	**0.003** (0.001)	**14.565** (0.334)
		(25, 100)	`ncagr`	0.569 (0.077)	0.010 (0.001)	9.930 (1.155)
			`RegGMM`	**0.581** (0.076)	**0.002** (0.001)	**7.032** (0.360)
	400	(25, 50)	`ncagr`	0.762 (0.064)	0.010 (0.001)	8.129 (0.217)
			`RegGMM`	**0.876** (0.053)	**0.007** (0.002)	**4.735** (0.300)
		(50, 50)	`ncagr`	0.270 (0.039)	**0.002** (0.000)	14.512 (0.178)
			`RegGMM`	**0.464** (0.049)	0.007 (0.001)	**12.590** (0.312)
		(25, 100)	`ncagr`	0.772 (0.055)	0.009 (0.001)	8.623 (0.155)
			`RegGMM`	**0.858** (0.051)	**0.003** (0.001)	**5.054** (0.288)

Table 1: Mean and standard error of performance metrics over 100 data sets. "Natural" indicates data sets generated according to the natural specification (7) whereas "original" indicates those generated according to (3).

replications over three simulation settings. The averaged results are shown in Figure 3.

6 GBM eQTL Reanalysis

Glioblastoma multiforme (GBM) is the most malignant type of brain cancer and patient prognosis is typically very poor. Although there has been research on the genetic signaling pathways involved in the proliferation of GBM, it remains largely incurable; see Hanif et al. (2017) for a survey. It is important to understand the conditional independence structure of genes involved in GBM in order to discover new drug therapies (Kwiatkowska et al., 2013). Our estimated graphs describe the conditional independence of co-expressions in a gene network; hence, we refer to estimated networks and effects of SNPs on this network.

We reanalyze a GBM eQTL data set that was reported in Zhang and Li (2022). The data set contains microarray and SNP profling data of $n = 401$ GBM patients from the REMBRANDT trial (GSE108476) We use the expression levels of $p = 73$ genes known to belong to the human glioma

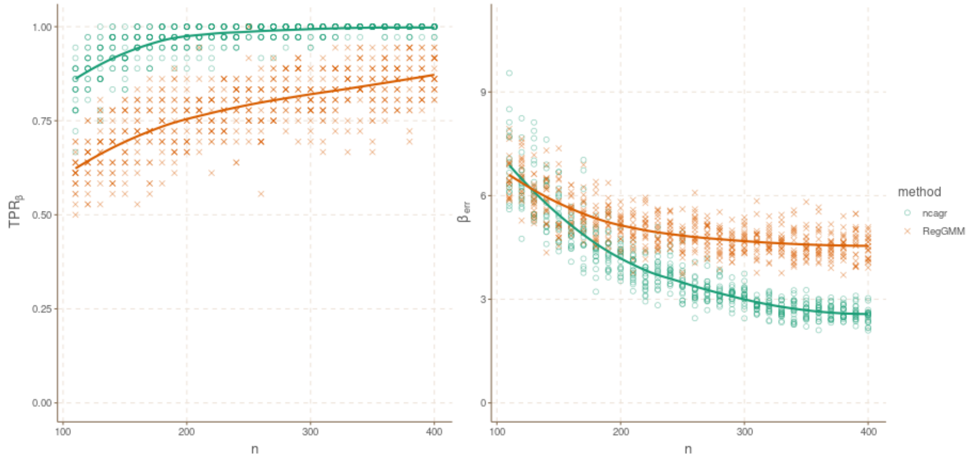

Figure 2: Performance of RegGMM and ncagr as the sample size varies. For each n, we show the TPR$_\beta$ (*left*) and $\boldsymbol{\beta}_{\text{err}}$ (*right*) of twenty replications with $p = 25$, $q = 100$. We see that ncagr outperforms RegGMM under the natural setting under both metrics.

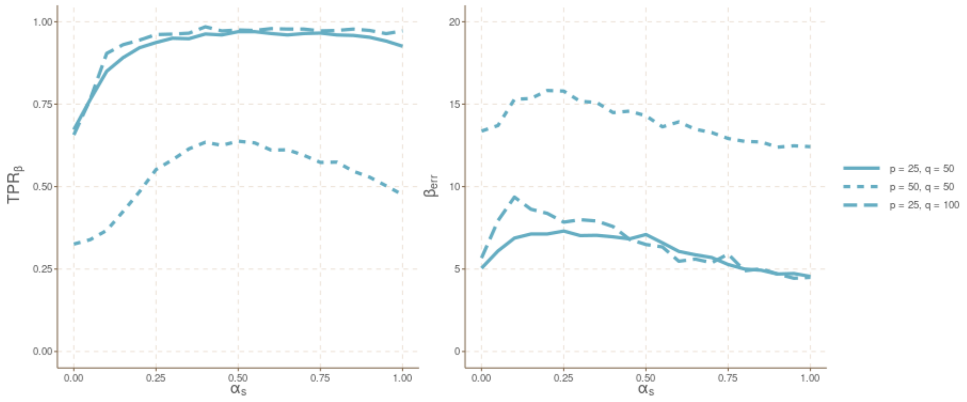

Figure 3: Robustness to choice of the mixing parameter α_s. The average TPR$_\beta$ (*left*) and $\boldsymbol{\beta}_{\text{err}}$ (*right*) over twenty replications for three settings are shown, where α_s is fixed and α_g and λ_0 are selected via 5-fold cross-validation.

pathway according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000); the genes and pathways are detailed in Table S3. We also consider $q = 118$ SNPs that are local to these 73 genes. The SNPs are binary-coded, with 0 indicating homozygous major alleles at that locus and 1 otherwise. Our data set is slightly different from the one used in Zhang and Li (2022); we have a larger cohort size (compared to $n = 178$ in Zhang and Li 2022) and do not include age and sex as covariates.

We ran `ncagr` on the data set using 5-fold cross-validation over $\alpha_g = 0.1, 0.2, \ldots, 0.9$, $\alpha_s = 0.1, 0.2, \ldots, 0.9$, and 100 values of λ_0. Our method identified 56 SNPs that potentially modify expressions in the network. However, many of the identified edges in these networks have small weights. Since cross-validation tends to select dense models, we elect to set to zero those entries in $\tilde{\mathbf{B}}_h$ below a certain threshold in order to achieve a more interpretable result. From Figure S1 (*left*), 0.005 appears to be a reasonable threshold. After thresholding, 10 SNPs are estimated to have effects on the network; see Table S4.

We also ran `RegGMM` on the data set for comparison, using 5-fold cross-validation for a path of 100 λ parameters and 10 mixture parameters. `RegGMM` identified 16 SNPs with nonzero effects on the network. However, similarly to `ncagr`, many of the estimated edges have very small weights. We choose to threshold these edges as well. From Figure S1 (*right*), 0.005 is also a reasonable threshold for these edges.

Figure 4 shows the estimated population (covariate-independent) network from `ncagr` and `RegGMM`. It can be seen that many estimated edges overlap and that `ncagr` estimates a somewhat denser network. For example, SHC4 and CALML4 are highly connected nodes in both population networks.

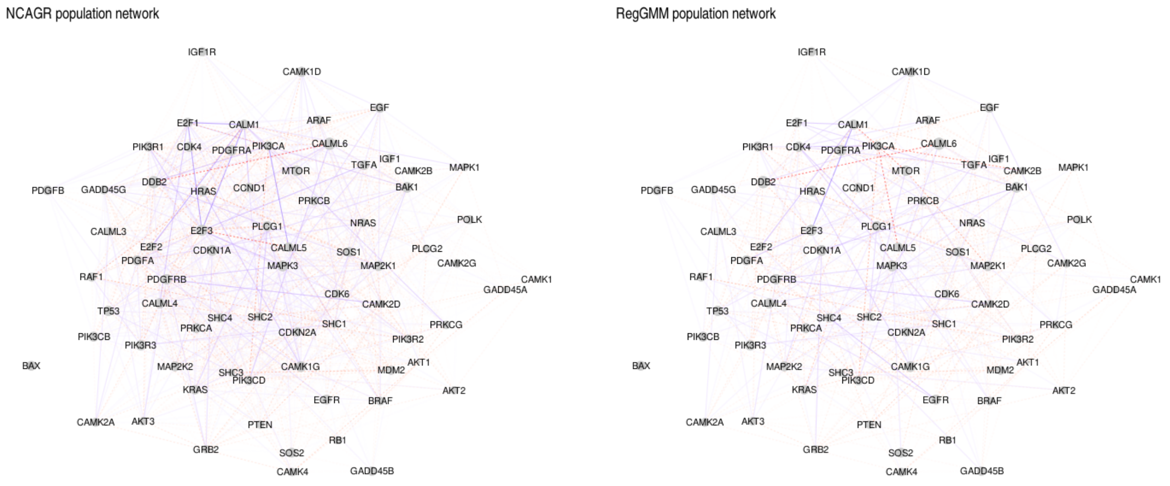

Figure 4: Population network from GBM eQTL data estimated by `ncagr` (*left*) and `RegGMM` (*right*). The graph structure is determined by the estimates of $\mathbf{B}_0$ in (3) and (7). Solid blue lines and dashed red lines indicate positive and negative edge weights, respectively.

Four SNPs are identified by `RegGMM` to affect the network of co-expressions. Out of the four, `rs1267622` is also identified by `ncagr` to have a nonzero effect. It is interesting to look at the estimated effect of `rs1267622`, a variant of the BRAF gene, by the two methods, shown in Figure 5.

The results from both methods suggest that this SNP may modify the co-expressions of PDGFRA

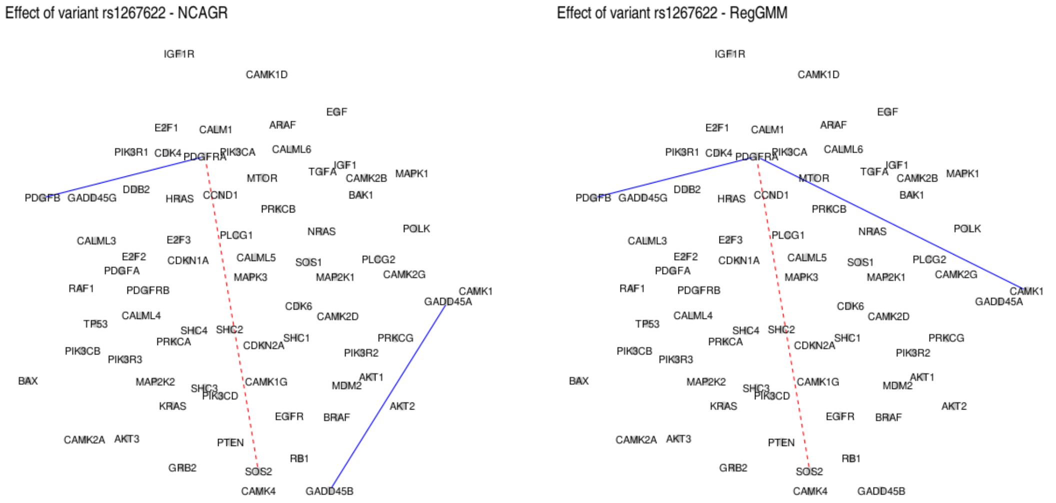

Figure 5: Estimated covariate networks for the SNP `rs1267622`, a variant local to the BRAF gene. The graph structure corresponds to the sparsity pattern of the matrix $\mathbf{B}_h$ corresponding to this SNP in (3) and (7). Solid blue lines and dashed red lines indicate positive and negative effects of this SNP on the partial correlation of co-expressions. An effect on the co-expression of GADD45A and GADD45B is detected by `ncagr` while `RegGMM` detects an effect between PDGFRA and SOS2. The remaining two edges, between PDGRFA and PDGFB and PDGRFA and SOS2, are detected by both methods.

and PDGFB as well as PDGFRA and SOS2. This is plausible since all three of these genes lie in the Ras-Raf-MEK-ERK pathway along with BRAF (Table S3). There is evidence that variants on this pathway are associated with the proliferation of certain cancers (Gonzalez-Hormazabal et al., 2019). Beyond this, the two methods differ; `ncagr` estimates this variant to modify the co-expressions of GADD45A and GADD45B while `RegGMM` suggests that the co-expressions of PDGFRA and CAMK1 are modified.

We now focus on the connections between four genes in particular, namely PIK3CA, CALML5, E2F1, and E2F3. PIK3CA is one of the most highly mutated oncogenes in a variety of cancers and resides in the PI3K/Akt/mTOR signaling pathway (Samuels and Velculescu, 2004). CALML5 is a calcium-binding protein that is part of the calcium (Ca^{+2}) signaling pathway, which is known to have diverse roles in explaining GBM biology and is a topic of active research (Azab et al., 2020; Cheng et al., 2021). E2F1 and E2F3 are oncogenic transcription factors. The over-expression of E2F3 is known to be vital in the development of various types of cancers including GBM (Zhang et al., 2019; Wu et al., 2021; Feng et al., 2018).

The estimated connections among these four genes in the population network are shown in Figure 6. While `RegGMM` and `ncagr` detect mostly the same edges in the population, only `ncagr` finds that these co-expressions are modified by the presence of SNPs. These effects are summarized in Figure 6 (*right*). For instance, it is estimated that a variant local to the CALML4 gene may mediate the induction of the PI3K/Akt/mTOR pathway by expressions in E2F1. PIK3CA and E2F1 are part of interconnected signaling pathways implicated in cancer progression; E2F1 is involved in regulating the expression of PIK3CA and downstream signaling components (Ladu et al., 2008). Hence, this discovery may give a clue for the role of CALML4 on regulating this pathway.

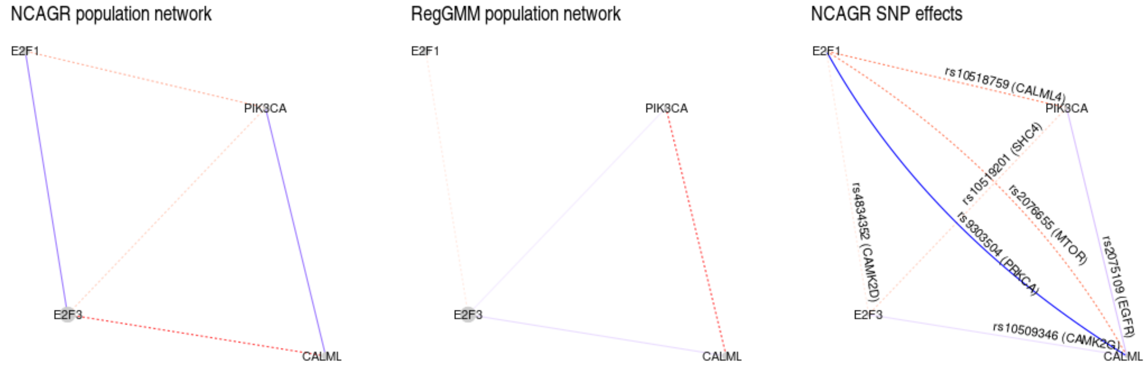

Figure 6: Estimated population network from `ncagr` (*left*) and `RegGMM` (*middle*) highlighting four select genes. An edge indicates the partial correlation between the expressions of two genes conditional on the rest of the network (not shown). *Right*: the estimated effect of SNPs on co-expressions among the selected genes according to `ncagr`. The gene to which the SNP is local is given in parentheses. Solid blue lines and dashed red lines indicate positive and negative edges, respectively. `RegGMM` did not detected any effects of SNPs on these co-expressions.

7 Discussion

In this work, we contributed to the covariate-adjusted graphical model literature by developing a framework allowing for jointly convex optimization of the mean and precision matrix. Our theoretical work implies that the convex formulation allows for more relaxed scaling assumptions in the sparsities of $\boldsymbol{\beta}_j$ in relation to the sample size n and this is confirmed by our simulation results.

There are a few directions for future work in our framework. First, our method relies on tuning the triplet $(\eta, \lambda, \lambda_g)$ as in 3.3. This will be too computationally intensive for very high-dimensional data sets, yet fixing a parameter such as α_s may not be desired. Therefore, adapting tuning-free methods such as the square-root lasso (Belloni et al., 2011) to our method would be an important contribution. A direction for theoretical work would be to establish the minimax rate of (14). Although our estimator achieves the same rate as in Zhang and Li (2022), the optimality of this rate has yet to be established.

Our R package `ncagr` is available on GitHub[1].

[1] `https://github.com/roobnloo/ncagr`

Acknowledgements

We are grateful to Emma Zhang for generously making the RegGMM software and GBM eQTL data set available to us.

Use was made of computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 2308708) at UC Santa Barbara.

References

Azab, M. A., Alomari, A., and Azzam, A. Y. (2020). Featuring how calcium channels and calmodulin affect glioblastoma behavior. a review article. *Cancer Treatment and Research Communications* **25,** 100255.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286,** 509–512.

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.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98,** 791–806.

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., Beroukhim, R., Bernard, B., Wu, C.-J., Genovese, G., Shmulevich, I., Barnholtz-Sloan, J., Zou, L., Vegesna, R., Shukla, S. A., Ciriello, G., Yung, W., and Zhang (2013). The somatic genomic landscape of glioblastoma. *Cell* **155,** 462–477.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100,** 139–156.

Chen, J., Xu, P., Wang, L., Ma, J., and Gu, Q. (2018). Covariate adjusted precision matrix estimation via nonconvex optimization. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 922–931. PMLR.

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, Q., Tang, A., Wang, Z., Fang, N., Zhang, Z., Zhang, L., Li, C., and Zeng, Y. (2021). Cald1 modulates gliomas progression via facilitating tumor angiogenesis. *Cancers* **13,**.

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

Erdős, P. and Rényi, A. (1959). On random graphs. i. *Publicationes Mathematicae Debrecen* **6,** 290–297.

Fehrmann, R. S. N., Jansen, R. C., Veldink, J. H., Westra, H.-J., Arends, D., Bonder, M. J., Fu, J., Deelen, P., Groen, H. J. M., Smolonska, A., Weersma, R. K., Hofstra, R. M. W., Buurman, W. A., Rensen, S., Wolfs, M. G. M., Platteel, M., Zhernakova, A., Elbers, C. C., Festen, E. M., Trynka, G., Hofker, M. H., Saris, C. G. J., Ophoff, R. A., van den Berg, L. H., van Heel, D. A., Wijmenga, C., te Meerman, G. J., and Franke, L. (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.

Feng, Z., Peng, C., Li, D., Zhang, D., Li, X., Cui, F., Chen, Y., and He, Q. (2018). E2f3 promotes cancer growth and is overexpressed through copy number variation in human melanoma. *OncoTargets and Therapy* pages 5303–5313.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9,** 432–441.

Gonzalez-Hormazabal, P., Musleh, M., Bustamante, M., Stambuk, J., Pisano, R., Valladares, H., Lanzarini, E., Chiong, H., Rojas, J., Suazo, J., Castro, V. G., Jara, L., and Berger, Z. (2019). Polymorphisms in ras/raf/mek/erk pathway are associated with gastric cancer. *Genes* **10,**.

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.

Hanif, F., Muzaffar, K., Perveen, k., Malhi, S., and Simjee, S. (2017). Glioblastoma multiforme: A review of its epidemiology and pathogenesis through clinical presentation and treatment. *Asian Pacific Journal of Cancer Prevention* **18,**.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28,** 27–30.

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,**.

Kuchibhotla, A. and Chakrabortty, A. (2022). Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *Information and Inference A Journal of the IMA* **11,** 1389–1456.

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

Ladu, S., Calvisi, D. F., Conner, E. A., Farina, M., Factor, V. M., and Thorgeirsson, S. S. (2008). E2f1 inhibits c-myc-driven apoptosis via pik3ca/akt/mtor and cox-2 in a mouse model of human liver cancer. *Gastroenterology* **135,** 1322–1332.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics* **28,** 1302–1338.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford University Press.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics* **40,** 1637 – 1664.

Lv, X., Cui, W., and Liu, Y. (2022). A sharp analysis of covariate adjusted precision matrix estimation via alternating projected gradient descent. *IEEE Signal Processing Letters* **29,** 877–881.

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.

Network, C. G. A. R. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica* **26,** 35–67.

Rockman, M. and Kruglyak, L. (2006). Genetics of global gene expression. *Nature reviews. Genetics* **7,** 862–72.

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.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4,**.

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

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

Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices.

Wu, L., Li, J., Xu, Y., Lou, X., Sun, M., and Wang, S. (2021). Expression and prognostic value of e2f3 transcription factor in non-small cell lung cancer. *Oncology letters* **21,** 411.

Yin, J. and Li, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics* **5,** 2630–2650.

Yu, G. and Bien, J. (2019). Estimating the error variance in a high-dimensional linear model. *Biometrika* **106,** 533–546.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **69,** 329–346.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94,** 19–35.

Zhang, G., Dong, Z., Prager, B. C., Kim, L. J., Wu, Q., Gimple, R. C., Wang, X., Bao, S., Hamerlik, P., and Rich, J. N. (2019). Chromatin remodeler hells maintains glioma stem cells through e2f3 and myc. *JCI Insight* **4,**.

Zhang, J. and Li, Y. (2022). High-dimensional gaussian graphical regression models with covariates. *Journal of the American Statistical Association* **0,** 1–13.

Supplemental Materials
Natural Covariate-adjusted Graphical Regression

S1 Computational Details

```
Algorithm 1: Natural Covariate-adjusted Graphical Regression
Input  : $\mathbf{U} \in \mathbb{R}^{n\times q}$ matrix of covariates.
         $\mathbf{X} \in \mathbb{R}^{n\times p}$ matrix of responses.
         $g_j : \mathbb{R}^q \times \mathbb{R}^{(p-1)(q+1)} \to \mathbb{R}$ convex penalty functions for $j \in [p]$.
         $H(\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\beta}}_j)$ a function to estimate the error variance $\sigma^2_{\varepsilon_j}$.
Output : $\hat{\boldsymbol{\Gamma}} \in \mathbb{R}^{p\times q}$, $\tilde{\mathbf{B}}_0, \tilde{\mathbf{B}}_1, \dots \tilde{\mathbf{B}}_q \in \mathbb{R}^{p\times p}$, and $\hat{\sigma}^2_{\varepsilon_j}$ for $j \in [p]$.

for $j = 1, 2, \dots, p$ do
    Define
        $\mathbf{W}_{-j,0} = [\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_p] \in \mathbb{R}^{n\times(p-1)}$,
        $\mathbf{W}_{-j,h} = [\mathbf{x}_1 \odot \mathbf{u}_h,\ \dots,\ \mathbf{x}_{j-1} \odot \mathbf{u}_h,\ \mathbf{x}_{j+1} \odot \mathbf{u}_i,\ \dots,\ \mathbf{x}_p \odot \mathbf{u}_h]$ for $h \in [q]$,
        $\mathbf{W}_{-j} = [\mathbf{W}_{-j,0},\ \mathbf{W}_{-j,1},\ \dots,\ \mathbf{W}_{-j,q}] \in \mathbb{R}^{n\times(p-1)(q+1)}$.
    Solve
        $(\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\beta}}_j) = \underset{\boldsymbol{\gamma}_j, \boldsymbol{\beta}_j}{\arg\min}\ \frac{1}{2n}\left\|\mathbf{x}_j - \mathbf{U}\boldsymbol{\gamma}_j - \mathbf{W}_{-j}\boldsymbol{\beta}_j\right\|_2^2 + g_j(\boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)$.    (S1)
    Set $\hat{\sigma}^2_{\varepsilon_j} = H(\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\beta}}_j)$.
    Set $\tilde{\beta}_{jkh} = -\hat{\beta}_{jkh}/\hat{\sigma}^2_{\varepsilon_j}$ for $k \neq j$, $h = 0, 1, \dots, q$.
end
Set $\hat{\boldsymbol{\Gamma}} = [\hat{\boldsymbol{\gamma}}_1, \dots, \hat{\boldsymbol{\gamma}}_p]^\top$.
Set $[\tilde{\mathbf{B}}_0]_{jj} = -1$ and $[\tilde{\mathbf{B}}_h]_{jj} = 0$ for $j \in [p]$, $h \in [q]$.
for $h = 0, 1, \dots, q$ do
    Apply the and-rule:
        $[\tilde{\mathbf{B}}_h]_{jk} = [\tilde{\mathbf{B}}_h]_{kj} = \tilde{\beta}_{jkh}\mathbf{1}_{\{|\tilde{\beta}_{jkh}|<|\tilde{\beta}_{kjh}|\}} + \tilde{\beta}_{kjh}\mathbf{1}_{\{|\tilde{\beta}_{jkh}|>|\tilde{\beta}_{kjh}|\}}$.
end
return $\hat{\boldsymbol{\Gamma}}, \tilde{\mathbf{B}}_0, \tilde{\mathbf{B}}_1, \dots, \tilde{\mathbf{B}}_q, \hat{\sigma}^2_{\varepsilon_1}, \dots, \hat{\sigma}^2_{\varepsilon_p}$.
```

S1.1 Algorithm for Block Coordinate Descent

In this section, we fully specify the algorithm to solve each nodewise regression problem (14). Each block update is an update on either a standard lasso problem ($\boldsymbol{\gamma}_j$ and $\boldsymbol{\beta}_{j,0}$) or a sparse group lasso problem ($\boldsymbol{\beta}_{j,h}$ for $h \in [q]$). Since (14) is convex and block-separable, this algorithm is guaranteed to converge to the optimal solution (Tseng, 2001). With the understanding that $j \in [p]$ is fixed, define the residual vectors

$$\begin{aligned}
\mathbf{r} &= \mathbf{x}_j - \mathbf{U}\boldsymbol{\gamma}_j - \sum_{\ell=0}^{q} \mathbf{W}_{-j,\ell}\boldsymbol{\beta}_{j,\ell}, \\
\mathbf{r}_\gamma &= \mathbf{x}_j - \sum_{\ell=0}^{q} \mathbf{W}_{-j,\ell}\boldsymbol{\beta}_{j,\ell}, \\
\mathbf{r}_{\beta,h} &= \mathbf{x}_j - \mathbf{U}\boldsymbol{\gamma}_j - \sum_{\ell\neq h}^{q} \mathbf{W}_{-j,\ell}\boldsymbol{\beta}_{j,\ell}, \;\; h \in \{0, 1, \dots, q\}.
\end{aligned}$$

Then we can represent the subproblems corresponding to $\boldsymbol{\gamma}_j$, $\boldsymbol{\beta}_{j,0}$, $\boldsymbol{\beta}_{j,1}, \dots, \boldsymbol{\beta}_{j,q}$ as follows:

$$\underset{\boldsymbol{\gamma}_j}{\text{minimize}} \frac{1}{2n}\left\|\mathbf{r}_\gamma - \mathbf{U}\boldsymbol{\gamma}_j\right\|_2^2 + \eta\|\boldsymbol{\gamma}_j\|_1, \tag{S2a}$$

$$\underset{\boldsymbol{\beta}_{j,0}}{\text{minimize}} \frac{1}{2n}\left\|\mathbf{r}_{\beta,0} - \mathbf{W}_{-j,0}\boldsymbol{\beta}_{j,0}\right\|_2^2 + \lambda\|\boldsymbol{\beta}_{j,0}\|_1, \tag{S2b}$$

$$\underset{\boldsymbol{\beta}_{j,h}}{\text{minimize}} \frac{1}{2n}\left\|\mathbf{r}_{\beta,h} - \mathbf{W}_{-j,h}\boldsymbol{\beta}_{j,h}\right\|_2^2 + \lambda\|\boldsymbol{\beta}_{j,h}\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,h}\|_2, \quad h \in [q]. \tag{S2c}$$

One iteration of our block coordinate descent algorithm performs a single update step for (S2a), (S2b), and (S2c) in turn, updating the estimates and the residual vectors along the way.

Since (S2a) and (S2b) are standard lasso problems, we update each component using the well-known expressions

$$\begin{aligned}
\hat{\gamma}_j &\leftarrow \mathcal{S}_\eta\left(\hat{\gamma}_j + \frac{1}{n}\mathbf{U}^\top \mathbf{r}_{\gamma,j}\right), \\
\hat{\boldsymbol{\beta}}_{j,0} &\leftarrow \mathcal{S}_\lambda\left(\hat{\boldsymbol{\beta}}_{j,0} + \frac{1}{n}\mathbf{W}_{-j,0}^\top \mathbf{r}_{\beta,0}\right),
\end{aligned}$$

where $\mathcal{S}_c(\mathbf{x}) = \operatorname{sign}(\mathbf{x})(\mathbf{x} - c)_+$ is the soft-thresholding operator applied elementwise to a vector $\mathbf{x}$. The specification (S2c) is the sparse-group lasso problem and our approach follows the implementation given in Simon et al. (2013). A brief review follows. For a fixed $h \in [q]$, we wish to update $\hat{\boldsymbol{\beta}}_{j,h}$. First we check the subgradient condition

$$\|\mathcal{S}_\lambda(\mathbf{W}_{-j,h}^\top \mathbf{r}_{\beta,h})\|_2 \le \lambda_g;$$

if this is satisfied, we may set $\hat{\boldsymbol{\beta}}_{j,h} \leftarrow \mathbf{0}$ and are done with the update. Otherwise, define the loss function

$$\ell(\boldsymbol{\beta}_{j,h}) = \frac{1}{2n}\left\|\mathbf{r}_{\beta,h} - \mathbf{W}_{-j,h}\boldsymbol{\beta}_{j,h}\right\|_2^2$$

and the gradient step with step size t

$$U(\boldsymbol{\beta}_{j,h}, t) = \left(1 - \frac{t(1-\alpha)\lambda}{\left\|\mathcal{S}_{t\lambda}\left(\boldsymbol{\beta}_{j,h} - t\nabla\ell(\boldsymbol{\beta}_{j,h})\right)\right\|_2}\right)_+ \mathcal{S}_{t\lambda}\left(\boldsymbol{\beta}_{j,h} - t\nabla\ell(\boldsymbol{\beta}_{j,h})\right).$$

Then initialize counter $k = 1$, $\boldsymbol{\theta}_k = \hat{\boldsymbol{\beta}}_{j,h}$, step size $t = 1$, and execute the following until convergence:

1. Update the gradient
$$\mathbf{g} = \nabla \ell(\hat{\boldsymbol{\beta}}_{j,h}),$$

2. Optimize the step size via $t \leftarrow 0.8t$ until the following condition holds:
$$\ell(U(\hat{\boldsymbol{\beta}}_{j,h}, t)) \leq \ell(\hat{\boldsymbol{\beta}}_{j,h}) + \mathbf{g}^\top \Delta_{(k,t)} + \frac{1}{2t} \left\|\Delta_{(k,t)}\right\|_2^2$$
where $\Delta_{(k,t)} = U(\hat{\boldsymbol{\beta}}_{j,h}, t) - \hat{\boldsymbol{\beta}}_{j,h}$,

3. Update the state variable
$$\boldsymbol{\theta}_{k+1} = U(\hat{\boldsymbol{\beta}}_{j,h}, t),$$

4. Perform a Nesterov acceleration step
$$\hat{\boldsymbol{\beta}}_{j,h} \leftarrow \boldsymbol{\theta}_k + \frac{k}{k+3}(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k),$$

5. Set $k \leftarrow k + 1$.

These steps come from a Majorization-Minimization algorithm combined with Nesterov acceleration and step size selection. In practice, not much time is spent in the inner loop where the step size t is selected. We refer the reader to Simon et al. (2013) for a detailed explanation of these steps. The combined steps are presented in Algorithm 2.

Algorithm 2: Blockwise descent for NCAGR nodewise regression

```
Input      : $\mathbf{U} \in \mathbb{R}^{n \times q}$ matrix of covariates.
             $\mathbf{x}_j \in \mathbb{R}^n$ vector of response $j$.
             $[\mathbf{W}_{-j,0},\ \mathbf{W}_{-j,1},\ \ldots,\ \mathbf{W}_{-j,q}] \in \mathbb{R}^{n \times (p-1)(q+1)}$ interaction matrices.
Output     : $\hat{\boldsymbol{\gamma}}_j \in \mathbb{R}^q$, $\hat{\boldsymbol{\beta}}_j \in \mathbb{R}^{(p-1)(q+1)}$
Initialize : $\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\beta}}_j$
Compute the residual vector: $\mathbf{r} \leftarrow \mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}}_j - \sum_{h=0}^{q} \mathbf{W}_{-j,h}\hat{\boldsymbol{\beta}}_{j,h}$ repeat
    /* Block update for $\hat{\boldsymbol{\gamma}}_j$ */
    $\mathbf{r}_\gamma \leftarrow \mathbf{r} + \mathbf{U}\hat{\boldsymbol{\gamma}}_j$
    $\hat{\boldsymbol{\gamma}}_j \leftarrow \mathcal{S}_\eta(\hat{\boldsymbol{\gamma}}_j + \mathbf{U}^\top \mathbf{r}_\gamma / n)$
    $\mathbf{r} \leftarrow \mathbf{r}_\gamma - \mathbf{U}\hat{\boldsymbol{\gamma}}_j$
    /* Block update for $\hat{\boldsymbol{\beta}}_{j,0}$ */
    $\mathbf{r}_{\beta,0} \leftarrow \mathbf{r} + \mathbf{W}_{-j,0}\hat{\boldsymbol{\beta}}_{j,0}$
    $\hat{\boldsymbol{\beta}}_{j,0} \leftarrow \mathcal{S}_\lambda\left(\hat{\boldsymbol{\beta}}_{j,0} + \frac{1}{n}\mathbf{W}_{-j,0}^\top \mathbf{r}_{\beta,0}\right)$
    $\mathbf{r} \leftarrow \mathbf{r}_{\beta,0} - \mathbf{W}_{-j,0}\hat{\boldsymbol{\beta}}_{j,0}$

    for $h = 1, 2, \ldots, q$ do
        /* Block update for $\hat{\boldsymbol{\beta}}_{j,h}$ */
        if $\|\mathcal{S}_\lambda(\mathbf{W}_{-j,h}^\top \mathbf{r}_{\beta,h})\|_2 \leq \lambda_g$ then
            $\hat{\boldsymbol{\beta}}_{j,h} \leftarrow \mathbf{0}$
        else
            Initialize $k \leftarrow 1$; $\boldsymbol{\theta}_k \leftarrow \hat{\boldsymbol{\beta}}_{j,h}$; $t \leftarrow 1$
            repeat
                $\mathbf{g} \leftarrow \nabla \ell(\hat{\boldsymbol{\beta}}_{j,h})$
                repeat
                    $t \leftarrow 0.8t$
                    $\Delta_{(k,t)} \leftarrow U(\hat{\boldsymbol{\beta}}_{j,h}, t) - \hat{\boldsymbol{\beta}}_{j,h}$
                until $\ell(U(\hat{\boldsymbol{\beta}}_{j,h})) \leq \ell(\hat{\boldsymbol{\beta}}_{j,h}) + \mathbf{g}^\top \Delta_{(k,t)} + \frac{1}{2t}\|\Delta_{(k,t)}\|_2^2$;
                $\boldsymbol{\theta}_{k+1} \leftarrow U(\hat{\boldsymbol{\beta}}_{j,h}, t)$
                $\hat{\boldsymbol{\beta}}_{j,h} \leftarrow \boldsymbol{\theta}_k + \frac{k}{k+3}(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k)$
                $k \leftarrow k + 1$
            until convergence;
        end
    end
until convergence;
return $\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\beta}}_j$
```

S2 Additional Simulation Results

In addition to the metrics reported in Table 1, we computed the errors in the predicted precision matrix $\hat{\boldsymbol{\Omega}}$ and mean $\hat{\boldsymbol{\mu}}$ via (13) for each data set:

- $\boldsymbol{\Omega}_{\text{err}} = \sum_{i=1}^{n} \|\hat{\boldsymbol{\Omega}}_i - \boldsymbol{\Omega}_i\|_{F,\text{off}}^2 / n$, where $\|\cdot\|_{F,\text{off}}$ is the Frobenius norm of off-diagonal entries,
- $\boldsymbol{\mu}_{\text{err}} = \sum_{i=1}^{n} \|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|_2^2 / n$.

The means and standard errors for all simulation settings are shown in Table S1. As these are prediction rather than estimation errors, the results are not guaranteed to follow the theory developed in Section 4.

Model	n	(p, q)	Method	$\boldsymbol{\Omega}_{\text{TPR}}$	$\boldsymbol{\Omega}_{\text{FPR}}$	$\boldsymbol{\Omega}_{\text{err}}$	$\boldsymbol{\mu}_{\text{err}}$
natural	200	(25, 50)	`ncagr`	**0.991** (0.015)	0.265 (0.031)	**0.661** (0.155)	15.988 (1.694)
			`RegGMM`	0.880 (0.045)	**0.149** (0.044)	1.217 (0.176)	**14.607** (1.097)
		(50, 50)	`ncagr`	**0.724** (0.046)	**0.140** (0.019)	**2.612** (0.211)	14.719 (0.822)
			`RegGMM`	0.575 (0.048)	0.177 (0.035)	3.609 (0.175)	**12.670** (0.556)
		(25, 100)	`ncagr`	**0.999** (0.006)	0.474 (0.036)	2.565 (0.950)	36.043 (4.104)
			`RegGMM`	0.883 (0.047)	**0.104** (0.034)	**1.072** (0.200)	**17.579** (1.148)
	400	(25, 50)	`ncagr`	**0.999** (0.004)	0.187 (0.029)	**0.230** (0.033)	**8.487** (0.643)
			`RegGMM`	0.955 (0.028)	**0.112** (0.042)	0.841 (0.096)	12.698 (1.089)
		(50, 50)	`ncagr`	**0.898** (0.028)	**0.114** (0.016)	**1.482** (0.099)	**8.619** (0.375)
			`RegGMM`	0.810 (0.044)	0.312 (0.045)	2.465 (0.151)	10.085 (0.398)
		(25, 100)	`ncagr`	**1.000** (0.000)	0.355 (0.036)	**0.290** (0.047)	**11.191** (0.840)
			`RegGMM`	0.951 (0.028)	**0.082** (0.034)	0.719 (0.104)	14.885 (1.129)
original	200	(25, 50)	`ncagr`	0.695 (0.088)	0.384 (0.045)	3.471 (0.187)	6.705 (0.454)
			`RegGMM`	**0.799** (0.062)	**0.205** (0.054)	**1.928** (0.191)	**4.184** (0.280)
		(50, 50)	`ncagr`	0.246 (0.044)	**0.117** (0.017)	5.003 (3.152)	7.327 (0.417)
			`RegGMM`	**0.410** (0.045)	0.123 (0.036)	**4.352** (0.189)	**4.962** (0.272)
		(25, 100)	`ncagr`	**0.830** (0.073)	0.625 (0.041)	4.643 (5.044)	10.300 (0.607)
			`RegGMM`	0.762 (0.061)	**0.182** (0.053)	**2.165** (0.204)	**5.374** (0.337)
	400	(25, 50)	`ncagr`	0.861 (0.059)	0.397 (0.038)	2.728 (0.139)	3.376 (0.238)
			`RegGMM`	**0.940** (0.035)	**0.252** (0.058)	**1.010** (0.121)	**2.096** (0.157)
		(50, 50)	`ncagr`	0.351 (0.046)	**0.107** (0.015)	4.271 (0.104)	3.682 (0.197)
			`RegGMM`	**0.646** (0.048)	0.238 (0.040)	**3.295** (0.152)	**2.483** (0.143)
		(25, 100)	`ncagr`	0.899 (0.046)	0.584 (0.044)	3.048 (0.115)	4.472 (0.228)
			`RegGMM`	**0.926** (0.031)	**0.249** (0.060)	**1.145** (0.118)	**2.665** (0.141)

Table S1: Additional prediction metrics to supplement Table 1.

S3 Additional Results from GBM eQTL Analysis

S3.1 Thresholding

Both `ncagr` and `RegGMM` estimate graphs containing edges with very low weights. Thus we only interpret those edges with weights above a certain threshold. Figure S1 shows there is a meaningful separation between edges with weights above and below 0.005 for the result of both models.

Model	n	(p, q)	Method	$\boldsymbol{\beta}^{0}_{\text{TPR}}$	$\boldsymbol{\beta}^{0}_{\text{FPR}}$	$\boldsymbol{\beta}^{-0}_{\text{TPR}}$	$\boldsymbol{\beta}^{-0}_{\text{FPR}}$
natural	200	(25, 50)	ncagr	**0.989** (0.020)	0.119 (0.019)	**0.995** (0.018)	0.004 (0.001)
			RegGMM	0.872 (0.066)	**0.048** (0.018)	0.662 (0.127)	**0.003** (0.001)
		(50, 50)	ncagr	**0.749** (0.064)	0.058 (0.010)	**0.611** (0.059)	**0.002** (0.000)
			RegGMM	0.516 (0.083)	**0.044** (0.013)	0.310 (0.066)	0.004 (0.001)
		(25, 100)	ncagr	**0.999** (0.007)	0.205 (0.026)	**0.996** (0.017)	0.005 (0.001)
			RegGMM	0.870 (0.069)	**0.030** (0.014)	0.705 (0.123)	**0.001** (0.000)
	400	(25, 50)	ncagr	**1.000** (0.004)	0.107 (0.022)	**0.999** (0.008)	**0.002** (0.001)
			RegGMM	0.951 (0.038)	**0.028** (0.011)	0.865 (0.087)	**0.002** (0.001)
		(50, 50)	ncagr	**0.882** (0.036)	0.048 (0.009)	**0.877** (0.036)	**0.002** (0.000)
			RegGMM	0.760 (0.068)	**0.047** (0.011)	0.636 (0.071)	0.008 (0.002)
		(25, 100)	ncagr	**1.000** (0.000)	0.182 (0.024)	1.000 (0.000)	0.003 (0.001)
			RegGMM	0.939 (0.046)	**0.017** (0.010)	0.900 (0.072)	**0.001** (0.000)
original	200	(25, 50)	ncagr	0.662 (0.098)	0.200 (0.033)	0.229 (0.115)	0.006 (0.001)
			RegGMM	**0.796** (0.090)	**0.077** (0.022)	**0.341** (0.141)	**0.004** (0.002)
		(50, 50)	ncagr	0.303 (0.063)	0.059 (0.011)	0.047 (0.025)	**0.001** (0.000)
			RegGMM	**0.342** (0.087)	**0.038** (0.014)	**0.097** (0.038)	0.002 (0.001)
		(25, 100)	ncagr	0.759 (0.090)	0.335 (0.036)	0.218 (0.124)	0.007 (0.001)
			RegGMM	**0.777** (0.092)	**0.082** (0.024)	**0.219** (0.110)	**0.002** (0.001)
	400	(25, 50)	ncagr	0.897 (0.061)	0.251 (0.033)	0.512 (0.146)	**0.005** (0.001)
			RegGMM	**0.942** (0.043)	**0.045** (0.013)	**0.754** (0.117)	0.006 (0.002)
		(50, 50)	ncagr	0.480 (0.067)	0.061 (0.010)	0.131 (0.040)	**0.001** (0.000)
			RegGMM	**0.651** (0.076)	**0.049** (0.012)	**0.340** (0.075)	0.006 (0.001)
		(25, 100)	ncagr	0.923 (0.054)	0.356 (0.046)	0.493 (0.128)	0.005 (0.001)
			RegGMM	**0.939** (0.046)	**0.043** (0.015)	**0.708** (0.119)	**0.003** (0.001)

Table S2: TPR and FPR of population network and covariate networks considered separately.

S3.2 Gene Signaling Pathways and Estimated SNP Effects

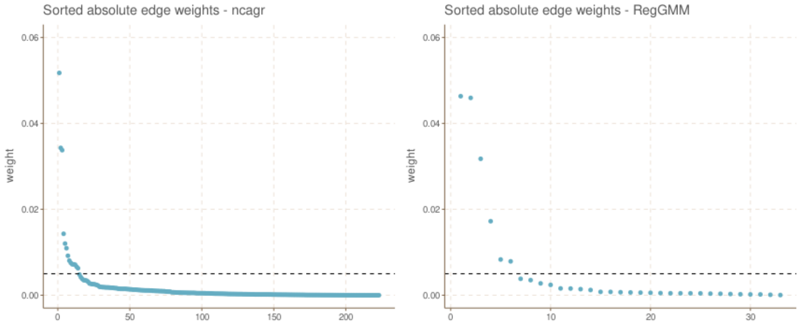

Figure S1: Absolute value of edges in covariate-adjusted networks after running `ncagr` on the GBM data set and the chosen threshold level of 0.005 (dashed line).

name	genes	references
PI3K/Akt/mTOR signaling pathway	PIK3CA, PIK3CB, PIK3CD, PIK3R3, PTEN, AKT1, AKT2, AKT3 MTOR, IGF1, PRKCA	Network (2008)
Ras-Raf-MEK-ERK signaling pathway	EGF, EGFR, GRB2, SOS1, SOS2, IGF1 SHC1, SHC2, SHC3, SHC4 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 (2008)

Table S3: Gene signalling pathways related to GBM.

S4 Proofs

Our proof strategies for Theorems 1 and 2 follow that of Zhang and Li (2022), with modifications made to accommodate our concatenated design matrix $[\mathbf{U}, \mathbf{W}]$. First we state a few lemmas.

Lemma 1 (Bellec et al. (2018) Lemma 1)**.** *Let $g : \mathbb{R}^d \to \mathbb{R}$ be any convex function and let*

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \big\{ \|\mathbf{y} - \mathbf{H}\boldsymbol{\beta}\|_2^2 + g(\boldsymbol{\beta}) \big\}$$

where $\mathbf{H} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Then for all $\boldsymbol{\beta} \in \mathbb{R}^d$,

$$\frac{1}{2n}\|\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}\|_2^2 + g(\hat{\boldsymbol{\beta}}) + \frac{1}{2n}\|\mathbf{H}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2 \le \frac{1}{2n}\|\mathbf{y} - \mathbf{H}\boldsymbol{\beta}\|_2^2 + g(\boldsymbol{\beta}).$$

Lemma 2 (Graybill and Marsaglia (1957) Theorem F)**.** *Let $\boldsymbol{\varepsilon} \sim N_p(\mathbf{0}, \sigma^2 \mathbf{I}_p)$ and let A be a $p \times p$ idempotent matrix with rank $r \le p$. Then $\boldsymbol{\varepsilon}^\top A \boldsymbol{\varepsilon} / \sigma^2 \sim \chi_r^2$.*

SNP	co-expressed genes
rs1267622	(PDGFRA, PDGFB), (PDGRFA, SOS2), (GADD45A, GADD45B)
rs10519201	(E2F3, PIK3CA), (E2F3, BAX)
rs10488141	(SOS2, E2F1), (SOS2, GADD45B)
rs2076655	(E2F1, CALML5)
rs10509346	(E2F3, CALML5)
rs10518759	(E2F1, PIK3CA)
rs10512510	(E2F3, BRAF)
rs9303504	(E2F1, CALML5)
rs4834352	(E2F1, E2F3)
rs2075109	(PIK3CA, CALML5)

Table S4: Estimated SNP effects on gene co-expressions according to `ncagr`.

Lemma 3 (Laurent and Massart (2000) Lemma 1). *Suppose that $U \sim \chi^2_r$. For any $x > 0$ it holds that*

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

Lemma 4 (Vershynin (2011) Proposition 5.16). *Let $X_1, \ldots, X_n$ be independent, mean zero sub-exponential random variables. Let $v_1 = \max_i \|X_i\|_{\psi_1}$ where $\|\cdot\|_{\psi_1}$ is the sub-exponential norm. Then there exists a constant c such that for any $t > 0$ we have*

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

Lemma 5 comes from Theorem 4.1 in Kuchibhotla and Chakrabortty (2022) applied to marginally sub-Gaussian random vectors.

Lemma 5 (Kuchibhotla and Chakrabortty (2022) Theorem 4.1). *Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be independent random vectors in $\mathbb{R}^p$. Assume each element of $\mathbf{Z}_i$ is sub-Gaussian with bounded sub-Gaussian norm for all $i \in [n]$. Let $\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}} = \mathbf{Z}^\top \mathbf{Z}/n$ and $\boldsymbol{\Sigma}_{\mathbf{Z}} = \mathbb{E}[\mathbf{Z}^\top \mathbf{Z}/n]$. Define*

$$\Upsilon_n = \max_{j,k} \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(Z_{ij} Z_{ik}).$$

Then we have

$$\sup_{\|\mathbf{v}\|_0 \leq k, \|\mathbf{v}\|_2 \leq 1} \left|\mathbf{v}^\top (\hat{\boldsymbol{\Sigma}}_{\mathbf{Z}} - \boldsymbol{\Sigma}_{\mathbf{Z}})\mathbf{v}\right| \lesssim k\sqrt{\frac{\Upsilon_n \log p}{n}} + \frac{k \log n \log p}{n}$$

with probability at least $1 - O(1/p)$.

We may apply Lemma 5 because by Assumption 3, $[\mathbf{U}, \mathbf{W}_{-j}]$ is elementwise sub-Gaussian, each entry being the product of a sub-Gaussian and a bounded random variable. Furthermore, we have by the Cauchy-Schwarz inequality that

$$\max_{\ell,k} \frac{1}{n}\sum_{i=1}^n \mathrm{Var}([\mathbf{U}, \mathbf{W}_{-j}]_{i\ell}[\mathbf{U}, \mathbf{W}_{-j}]_{ik}) \leq \max_{\ell_1,\ell_2,\ell_3,\ell_4} \mathbb{E}\Big(X_{\ell_1}^{(1)^2} X_{\ell_2}^{(1)^2} U_{\ell_3}^{(1)^2} U_{\ell_4}^{(1)^2}\Big) = O(1)$$

since the entries of $\mathbf{X}^{(1)}$ and $\mathbf{U}^{(1)}$ have bounded moments. Thus $\Upsilon_n = O(1)$ in our setting.

Lemma 6 (Loh and Wainwright (2012) Lemma 12)**.** *Let* $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ *be a symmetric matrix such that* $|\mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v}| \leq \delta_1$ *for all* $\mathbf{v} \in \mathbb{R}^p$ *with* $\|\mathbf{v}\|_2 = 1$ *and* $\|\mathbf{v}\|_0 \leq 2s$*. It holds for all* $\mathbf{v} \in \mathbb{R}^p$ *that*

$$|\mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v}| \leq 27\delta_1 \left(\|\mathbf{v}\|_2^2 + \frac{1}{s}\|\mathbf{v}\|_1^2 \right).$$

Define the matrices:

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{UW}} = \frac{1}{n}[\mathbf{U}, \mathbf{W}_{-j}]^\top [\mathbf{U}, \mathbf{W}_{-j}], \quad \boldsymbol{\Sigma}_{\mathbf{UW}} = \mathbb{E}(\hat{\boldsymbol{\Sigma}}_{\mathbf{U},\mathbf{W}}).$$

Lemma 7. *For a set of indices* $S \subset [p(q+1)-1]$*, denote by* $[\mathbf{U}, \mathbf{W}_{-j}]_S$ *the submatrix of* $[\mathbf{U}, \mathbf{W}_{-j}]$ *with columns indexed by* S*. Let* $\|\cdot\|_{\mathrm{op}}$ *denote the matrix operator norm. Under Assumptions 1-4, there exist constants* M_{uw} *and* C_0 *such that with probability at least* $1 - C_0 \exp(-\log(pq))$ *we have*

$$\frac{1}{n}\|[\mathbf{U}, \mathbf{W}_{-j}]_S\|_{\mathrm{op}}^2 \leq M_{uw}$$

for all S *satisfying* $|S| \leq \hat{s}_\gamma^{\max} + \hat{s}_\beta^{\max}$*, provided that* $(\hat{s}_\gamma^{\max} + \hat{s}_\beta^{\max})\log(pq) = O(\sqrt{n})$*, as assumed in Theorem 1.*

Proof. Letting $k = |S|$, it suffices to show that $\sup_{\|\mathbf{v}\|_0 \leq k,\, \|\mathbf{v}\|_2 \leq 1} \mathbf{v}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{UW}} \mathbf{v}$ is bounded. We may write

$$\sup_{\|\mathbf{v}\|_0 \leq k,\, \|\mathbf{v}\|_2 \leq 1} \mathbf{v}^\top \hat{\boldsymbol{\Sigma}}_{\mathbf{UW}} \mathbf{v} = \sup_{\|\mathbf{v}\|_0 \leq k,\, \|\mathbf{v}\|_2 \leq 1} \left\{ \mathbf{v}^\top \left(\hat{\boldsymbol{\Sigma}}_{\mathbf{UW}} - \boldsymbol{\Sigma}_{\mathbf{UW}} \right) \mathbf{v} + \mathbf{v}^\top \boldsymbol{\Sigma}_{\mathbf{UW}} \mathbf{v} \right\}.$$

By Assumption 2, the second term is bounded. For the first term, by Lemma 5 we have

$$\sup_{\|\mathbf{v}\|_0 \leq k,\, \|\mathbf{v}\|_2 \leq 1} \left\{ \mathbf{v}^\top \left(\hat{\boldsymbol{\Sigma}}_{\mathbf{UW}} - \boldsymbol{\Sigma}_{\mathbf{UW}} \right) \mathbf{v} \right\} \precsim \left(\frac{k^2 \log(pq)}{n} \right)^{1/2} + \frac{k \log(pq)}{n / \log n}$$

with probability at least $1 - C_0 \exp(-\log(pq))$ and we see that the right-hand side is bounded by Assumption 4. □

S4.1 Proof of Theorem 1

For ease of notation, we will drop the dependence of $\boldsymbol{\gamma}_j$, $\boldsymbol{\beta}_j$, $\boldsymbol{\varepsilon}_j$ and $\mathbf{W}_{-j}$ on j. Let $S_\beta, S_\gamma, \hat{S}_\beta, \hat{S}_\gamma$ be the support sets of $\boldsymbol{\beta}, \boldsymbol{\gamma}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$, respectively. Let $S_{\beta,g} = \{h \in [q] : \boldsymbol{\beta}_h \neq \mathbf{0}\}$ index the blocks $\boldsymbol{\beta}_h$ of $\boldsymbol{\beta}$ that are not identically zero and let $\hat{S}_{\beta,g}$ be the corresponding block indices for $\hat{\boldsymbol{\beta}}$. For any vector $\mathbf{v}$ and set of block indices S, let $\mathbf{v}_{(S)}$ denote the sub-vector containing blocks in S. Let $s_\beta, s_\gamma, s_{\beta,g}, \hat{s}_\beta, \hat{s}_\gamma, \hat{s}_{\beta,g}$ be the number of elements in $S_\beta, S_\gamma, S_{\beta,g}, \hat{S}_\beta, \hat{S}_\gamma, \hat{S}_{\beta,g}$, respectively.

Our proof occurs in three steps.

Step 1

In this step we bound the error $\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2/n$ by the stochastic term $\langle \boldsymbol{\varepsilon}, \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta} \rangle / n$, which is then bounded by a projection of $\boldsymbol{\varepsilon}$ onto the columns of $[\mathbf{U}, \mathbf{W}]$.

Since our penalty function

$$g(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \eta\|\boldsymbol{\gamma}\|_1 + \lambda\|\boldsymbol{\beta}\|_1 + \lambda_g\|\boldsymbol{\beta}_{-0}\|_{1,2}$$

is convex, by Lemma 1 we have

$$\frac{1}{2n}\|\mathbf{x} - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}}\|_2^2 + g(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) + \frac{1}{2n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 \le \frac{1}{2n}\|\mathbf{x} - \mathbf{U}\boldsymbol{\gamma} - \mathbf{W}\boldsymbol{\beta}\|_2^2 + g(\boldsymbol{\gamma}, \boldsymbol{\beta})$$

where $\boldsymbol{\nu} = \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$ and $\boldsymbol{\Delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. Now since $\boldsymbol{\varepsilon} = \mathbf{x} - \mathbf{U}\boldsymbol{\gamma} - \mathbf{W}\boldsymbol{\beta}$ we may write

$$\begin{aligned}\frac{1}{2n}\|\mathbf{x} - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}}\|_2^2 &= \frac{1}{2n}\|\boldsymbol{\varepsilon} - \mathbf{U}\boldsymbol{\nu} - \mathbf{W}\boldsymbol{\Delta}\|_2^2 \\ &= \frac{1}{2n}\|\boldsymbol{\varepsilon}\|_2^2 - \frac{1}{n}\langle \boldsymbol{\varepsilon}, \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\rangle + \frac{1}{2n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2.\end{aligned}$$

Plugging this into the previous expression and substituting the penalty expression then yields

$$\begin{aligned}&\frac{1}{n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 + \eta\|\hat{\boldsymbol{\gamma}}\|_1 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{-0}\|_{1,2} \\ &\quad \le \frac{1}{n}\langle \boldsymbol{\varepsilon}, \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\rangle + \eta\|\boldsymbol{\gamma}\|_1 + \lambda\|\boldsymbol{\beta}\|_1 + \lambda_g\|\boldsymbol{\beta}_{-0}\|_{1,2}.\end{aligned}$$

Notice that $\|\boldsymbol{\Delta}_{S_\beta^c}\|_1 = \|\hat{\boldsymbol{\beta}}_{S_\beta^c}\|_1$, $\|\hat{\boldsymbol{\beta}}\|_1 = \|\hat{\boldsymbol{\beta}}_{S_\beta}\|_1 + \|\hat{\boldsymbol{\beta}}_{S_\beta^c}\|_1$, and $\|\boldsymbol{\beta}\|_1 = \|\boldsymbol{\beta}_{S_\beta}\|_1$. Hence we can express the above as

$$\begin{aligned}&\frac{1}{n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 + \eta\|\hat{\boldsymbol{\gamma}}\|_1 + \underbrace{\lambda\|\hat{\boldsymbol{\beta}}_{S_\beta}\|_1 + \lambda\|\boldsymbol{\Delta}_{S_\beta^c}\|_1}_{\lambda\|\hat{\boldsymbol{\beta}}\|_1} + \lambda_g\|\hat{\boldsymbol{\beta}}_{-0}\|_{1,2} \\ &\quad \le \frac{1}{n}\langle \boldsymbol{\varepsilon}, \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\rangle + \eta\|\boldsymbol{\gamma}\|_1 + \underbrace{\lambda\|\boldsymbol{\beta}_{S_\beta}\|_1}_{\lambda\|\boldsymbol{\beta}\|_1} + \lambda_g\|\boldsymbol{\beta}_{-0}\|_{1,2}.\end{aligned}$$

Thus we have

$$\begin{aligned}&\frac{1}{n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 + \eta\|\hat{\boldsymbol{\gamma}}\|_1 + \lambda\|\boldsymbol{\Delta}_{S_\beta^c}\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{-0}\|_{1,2} \\ &\quad \le \frac{1}{n}\langle \boldsymbol{\varepsilon}, \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\rangle + \eta\|\boldsymbol{\gamma}\|_1 + \lambda(\|\boldsymbol{\beta}_{S_\beta}\|_1 - \|\hat{\boldsymbol{\beta}}_{S_\beta}\|_1) + \lambda_g\|\boldsymbol{\beta}_{-0}\|_{1,2} \\ &\quad \le \frac{1}{n}\langle \boldsymbol{\varepsilon}, \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\rangle + \eta\|\boldsymbol{\gamma}\|_1 + \lambda\|\boldsymbol{\Delta}_{S_\beta}\|_1 + \lambda_g\|\boldsymbol{\beta}_{-0}\|_{1,2}\end{aligned}$$

using the triangle inequality for the ℓ_1 norm. The same development holds for $\eta\|\hat{\boldsymbol{\gamma}}\|_1$. Finally, notice that

$$\|\boldsymbol{\Delta}_{(S_{\beta,g}^c)}\|_{1,2} = \|\hat{\boldsymbol{\beta}}_{(S_{\beta,g}^c)}\|_{1,2}, \; \|\hat{\boldsymbol{\beta}}_{-0}\|_{1,2} = \|\hat{\boldsymbol{\beta}}_{(S_{\beta,g})}\|_{1,2} + \|\hat{\boldsymbol{\beta}}_{(S_{\beta,g}^c)}\|_{1,2}, \text{ and } \|\boldsymbol{\beta}_{-0}\|_{1,2} = \|\boldsymbol{\beta}_{(S_{\beta,g})}\|_{1,2}.$$

Hence the previous development holds for the $\lambda_g\|\hat{\boldsymbol{\beta}}_{-0}\|_{1,2}$ term by the triangle inequality of $\|\cdot\|_{1,2}$. All in all we have

$$\begin{aligned}&\frac{1}{n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 + \eta\|\boldsymbol{\nu}_{S_\gamma^c}\|_1 + \lambda\|\boldsymbol{\Delta}_{S_\beta^c}\|_1 + \lambda_g\|\boldsymbol{\Delta}_{(S_{\beta,g}^c)}\|_{1,2} \\ &\quad \le \frac{1}{n}\langle \boldsymbol{\varepsilon}, \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\rangle + \eta\|\boldsymbol{\nu}_{S_\gamma}\|_1 + \lambda\|\boldsymbol{\Delta}_{S_\beta}\|_1 + \lambda_g\|\boldsymbol{\Delta}_{(S_{\beta,g})}\|_{1,2}.\end{aligned} \tag{S1}$$

Now let $\mathcal{I}$ and $\mathcal{J}$ be arbitrary index sets of the columns of $\mathbf{U}$ and $\mathbf{W}$ respectively. Denote by $\mathcal{P}_{\mathcal{I},\mathcal{J}}$ the orthogonal projection onto the columns of $[\mathbf{U},\mathbf{W}]$ indexed by $(\mathcal{I},\mathcal{J})$. Let $\mathcal{I}_0 = S_\gamma \cup \hat{S}_\gamma$ and $\mathcal{J}_0 = S_\beta \cup \hat{S}_\beta$ denote the unions of the true and estimated support sets of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. We seek to bound the stochastic term

$$\begin{aligned}\langle \boldsymbol{\varepsilon}, \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\rangle &= \langle \mathcal{P}_{\mathcal{I}_0,\mathcal{J}_0}(\boldsymbol{\varepsilon}), \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\rangle \\ &\leq \|\mathcal{P}_{\mathcal{I}_0,\mathcal{J}_0}(\boldsymbol{\varepsilon})\|_2 \|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2 \leq \frac{1}{2a_1}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 + \frac{a_1}{2}\|\mathcal{P}_{\mathcal{I}_0,\mathcal{J}_0}(\boldsymbol{\varepsilon})\|_2^2.\end{aligned} \tag{S2}$$

The last inequality follows from the fact that $2xy \leq ax^2 + y^2/a$ holds for any constant $a > 0$ and real numbers x and y.

Following Zhang and Li (2022), we first bound the term $\|\mathcal{P}_{\mathcal{I}_0,\mathcal{J}_0}(\boldsymbol{\varepsilon})\|_2^2$ with a counting argument. For fixed s'_γ, s'_β, and $s'_{\beta,g}$, we will bound the cardinality of the set

$$\mathcal{H}(s'_\gamma, s'_\beta, s'_{\beta,g}) = \{(\mathcal{I},\mathcal{J}) \subset [q] \times [(p-1)(q+1)] : |\mathcal{I}| = s'_\gamma,\ |\mathcal{J}| = s'_\beta,\ |g(\mathcal{J})| = s'_{\beta,g}\}$$

where $g(\mathcal{J})$ is the number of nonzero groups of $\boldsymbol{\beta}_{\mathcal{J}}$. For ease of notation, we will write $\mathcal{H}$ while keeping in mind its dependence on s'_γ, s'_β, and $s'_{\beta,g}$. We will show that

$$\log|\mathcal{H}| \leq s'_\gamma \log\frac{eq}{s'_\gamma} + s'_{\beta,g}\log\frac{eq}{s'_{\beta,g}} + s'_\beta \log(ep)$$

by considering the two cases $s'_\beta = s'_{\beta,g}$ and $s'_\beta > s'_{\beta,g}$ separately. These are the only cases since we cannot have more nonzero groups than nonzero elements.

1. $s'_\beta = s'_{\beta,g}$: In this case, we have $|\mathcal{H}| \leq \binom{q}{s'_\gamma}\binom{q}{s'_{\beta,g}}(p-1)^{s'_\beta}$. Hence

$$\begin{aligned}\log|\mathcal{H}| &\leq \log\binom{q}{s'_\gamma} + \log\binom{q}{s'_{\beta,g}} + s'_\beta\log(p-1) \\ &\leq s'_\gamma \log\frac{eq}{s'_\gamma} + s'_{\beta,g}\log\frac{eq}{s'_{\beta,g}} + s'_\beta\log(ep)\end{aligned}$$

   where we use $\log\binom{n}{k} \leq k\log(en/k)$ which follows from Stirling's approximation.

2. $s'_\beta > s'_{\beta,g}$: In this case, the cardinality is bounded by

$$|\mathcal{H}| \leq \binom{q}{s'_\gamma}\binom{q}{s'_{\beta,g}}\binom{(p-1)(s'_{\beta,g}+1)}{s'_\beta}.$$

   Since by Stirling's approximation

$$\log\binom{(p-1)(s'_{\beta,g}+1)}{s'_\beta} \leq s'_\beta \log\frac{e(p-1)(s'_{\beta,g}+1)}{s'_\beta} \leq s'_\beta\log(ep),$$

   we have

$$\log|\mathcal{H}| \leq s'_\gamma \log\frac{eq}{s'_\gamma} + s'_{\beta,g}\log\frac{eq}{s'_{\beta,g}} + s'_\beta\log(ep)$$

   as desired.

Define k_0 to be the exponential of the right hand side of the above inequality, so that $|\mathcal{H}| \leq k_0$. For any $(\mathcal{I}, \mathcal{J}) \in \mathcal{H}$, since $\mathcal{P}_{\mathcal{I},\mathcal{J}}$ is idempotent, Lemma 2 implies

$$\|\mathcal{P}_{\mathcal{I},\mathcal{J}}(\boldsymbol{\varepsilon})\|_2^2/\sigma_\varepsilon^2 \sim \chi_d^2$$

where $d \leq |\mathcal{I}| + |\mathcal{J}| = s'_\gamma + s'_\beta$ is the rank of $\mathcal{P}_{\mathcal{I},\mathcal{J}}$. By Lemma 3 we have for arbitrary $t' > 0$ that

$$\mathbb{P}\Big(\|\mathcal{P}_{\mathcal{I},\mathcal{J}}(\boldsymbol{\varepsilon})\|_2^2 \geq \sigma_\varepsilon^2(2\sqrt{dt'} + d + 2t')\Big) \leq e^{-t'}.$$

Since $2\sqrt{dt'} \leq d + t'$ and $d \leq s'_\gamma + s'_\beta \leq \log k_0$, we have

$$\mathbb{P}\big(\|\mathcal{P}_{\mathcal{I},\mathcal{J}}(\boldsymbol{\varepsilon})\|_2^2 \geq \sigma_\varepsilon^2(2\log k_0 + 3t')\big) \leq e^{-t'}.$$

Taking the supremum over $\mathcal{H}$ and applying the union bound yields

$$\mathbb{P}\left(\sup_{(\mathcal{I},\mathcal{J})\in\mathcal{H}} \|\mathcal{P}_{\mathcal{I},\mathcal{J}}(\boldsymbol{\varepsilon})\|_2^2 \geq \sigma_\varepsilon^2(2\log k_0 + 3t')\right) \leq |\mathcal{H}|e^{-t'}.$$

Now set $t' = t/3 + \log k_0$ for $t > 0$. Then $|\mathcal{H}|e^{-t'} = |\mathcal{H}|/k_0 \cdot e^{-t/3} \leq e^{-t/3}$ since $|\mathcal{H}| \leq k_0$. Substituting these expressions into the previous bound (and recalling that we have fixed s'_γ, s'_β, and $s'_{\beta,g}$) yields

$$\mathbb{P}\left(\sup_{(\mathcal{I},\mathcal{J})\in\mathcal{H}(s'_\gamma,s'_\beta,s'_{\beta,g})} \|\mathcal{P}_{\mathcal{I},\mathcal{J}}(\boldsymbol{\varepsilon})\|_2^2 \geq 5\sigma_\varepsilon^2\left[s'_\gamma \log\frac{eq}{s'_\gamma} + s'_{\beta,g}\log\frac{eq}{s'_{\beta,g}} + s'_\beta\log(ep)\right] + \sigma_\varepsilon^2 t\right) \leq e^{-t/3}, \tag{S3}$$

recalling that

$$\log k_0 = s'_\gamma \log\frac{eq}{s'_\gamma} + s'_{\beta,g}\log\frac{eq}{s'_{\beta,g}} + s'_\beta\log(ep).$$

This gives a concentration bound of $\|\mathcal{P}_{\mathcal{I},\mathcal{J}}(\boldsymbol{\varepsilon})\|_2^2$ over all possible subsets $\mathcal{I}$ of columns of $\mathbf{U}$ and $\mathcal{J}$ of columns of $\mathbf{W}$ satisfying $|\mathcal{I}| = s'_\gamma$, $|\mathcal{J}| = s'_\beta$, and $|g(\mathcal{J})| = s'_{\beta,g}$. Recalling that $\mathcal{I}_0$ and $\mathcal{J}_0$ are the support sets of $\boldsymbol{\nu}$ and $\boldsymbol{\Delta}$, we can now bound $\|\mathcal{P}_{\mathcal{I}_0,\mathcal{J}_0}(\boldsymbol{\varepsilon})\|_2^2$. Define

$$r(s'_\gamma, s'_\beta, s'_{\beta,g}) = \left(\sup_{(\mathcal{I},\mathcal{J})\in\mathcal{H}(s'_\gamma,s'_\beta,s'_{\beta,g})} \|\mathcal{P}_{\mathcal{I},\mathcal{J}}(\boldsymbol{\varepsilon})\|_2^2 - 5\sigma_\varepsilon^2\left\{s'_\gamma\log\frac{eq}{s'_\gamma} + s'_{\beta,g}\log\frac{eq}{s'_{\beta,g}} + s'_\beta\log(ep)\right\}\right)_+$$

and

$$r = \sup_{s'_\gamma,s'_\beta,s'_{\beta,g}} r(s'_\gamma, s'_\beta, s'_{\beta,g}).$$

It is clear that

$$|\mathcal{I}_0| \leq s_\gamma + \hat{s}_\gamma,\ |\mathcal{J}_0| \leq s_\beta + \hat{s}_\beta,\ \text{and } |g(\mathcal{J}_0)| \leq s_{\beta,g} + \hat{s}_{\beta,g}.$$

Thus we have

$$\|\mathcal{P}_{\mathcal{I}_0,\mathcal{J}_0}(\boldsymbol{\varepsilon})\|_2^2 \leq 5\sigma_\varepsilon^2\left\{(s_\gamma + \hat{s}_\gamma)\log\frac{eq}{s_\gamma} + (s_\beta + \hat{s}_\beta)\log(ep) + (s_{\beta,g} + \hat{s}_{\beta,g})\log\frac{eq}{s_{\beta,g}}\right\} + r. \tag{S4}$$

We also have the following concentration inequality on r for $t > 0$, which comes from the definition of $r(s'_\gamma, s'_\beta, s'_{\beta,g})$ along with (S3):

$$\mathbb{P}(r \geq t\sigma_\varepsilon^2) \leq \sum_{s'_\gamma, s'_\beta, s'_{\beta,g}} \mathbb{P}(r(s'_\gamma, s'_\beta, s'_{\beta,g}) \geq t\sigma_\varepsilon^2) \leq \sum_{s'_\gamma, s'_\beta, s'_{\beta,g}} e^{-t/3} \leq q^3 p\, c_1 e^{-t/3},$$

where $c_1 > 0$ is a constant and the sum is taken over $s'_\gamma \in [q]$, $s'_\beta \in [(p-1)(q+1)]$, $s'_{\beta,g} \in [q]$. Then for a sufficiently large constant $\tilde{M}$ and some $c_2 > 0$, by letting

$$t = \tilde{M}(s_\gamma \log(eq/s_\gamma) + s_\beta \log(ep) + s_{\beta,g} \log(eq/s_{\beta,g}))$$

we have

$$\begin{aligned} &\mathbb{P}\Big\{r \geq \tilde{M}\sigma_\varepsilon^2(s_\gamma \log(eq/s_\gamma) + s_\beta \log(ep) + s_{\beta,g} \log(eq/s_{\beta,g}))\Big\} \\ &\qquad \leq c_1 \exp\{-c_2(s_\gamma \log(eq/s_\gamma) + s_\beta \log(ep) + s_{\beta,g} \log(eq/s_{\beta,g}))\}. \end{aligned} \tag{S5}$$

Step 2

We now use the KKT optimality conditions to bound $\|\mathcal{P}_{\mathcal{I}_0,\mathcal{J}_0}(\boldsymbol{\varepsilon})\|_2^2$ in terms of $\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2$. This takes care of the stochastic term after plugging back into (S1) and (S2). To ease the notation when describing the conditions, define the vectors

$$\mathbf{r}_\gamma = \frac{1}{n}\mathbf{U}^\top(\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}}), \quad \mathbf{r}_\beta = \frac{1}{n}\mathbf{W}^\top(\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}})$$

Let $\hat{\boldsymbol{\beta}}_h$ be the h-th block of $\hat{\boldsymbol{\beta}}$. By the KKT conditions, we know that an optimizer $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ of (14) must satisfy

$$(\mathbf{r}_\gamma)_\ell = \eta \text{ sign}(\hat{\gamma}_\ell) \qquad \text{for } \hat{\gamma}_\ell \neq 0 \tag{S6}$$

$$(\mathbf{r}_\beta)_\ell = \lambda \text{ sign}((\hat{\boldsymbol{\beta}}_0)_\ell) \qquad \text{for } (\hat{\boldsymbol{\beta}}_0)_\ell \neq 0 \tag{S7}$$

$$(\mathbf{r}_\beta)_\ell = \lambda \text{ sign}((\hat{\boldsymbol{\beta}}_h)_\ell) + \lambda_g \frac{(\hat{\boldsymbol{\beta}}_h)_\ell}{\|\hat{\boldsymbol{\beta}}_h\|_2} \qquad \text{for } (\hat{\boldsymbol{\beta}}_h)_\ell \neq 0,\ h \in [q]. \tag{S8}$$

Squaring both sides of (S6) and summing over ℓ gives

$$\eta^2 \hat{s}_\gamma = \frac{1}{n^2}\|\mathbf{U}_{\hat{S}_\gamma}^\top(\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}})\|_2^2$$

and doing the same for (S7) and (S8) gives

$$\lambda^2 \hat{s}_\beta + \lambda_g^2 \hat{s}_{\beta,g} \leq \frac{1}{n^2}\|\mathbf{W}_{\hat{S}_\beta}^\top(\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}})\|_2^2$$

since in (S8) the cross term $\text{sign}((\hat{\boldsymbol{\beta}}_{j,h})_\ell) \times (\hat{\boldsymbol{\beta}}_{j,h})_\ell$ is nonnegative. We have shown that

$$\begin{aligned} \eta^2 \hat{s}_\gamma + \lambda^2 \hat{s}_\beta + \lambda_g^2 \hat{s}_{\beta,g} &\leq \frac{1}{n^2}\Big\|[\mathbf{U}_{\hat{S}_\gamma}, \mathbf{W}_{\hat{S}_\beta}]^\top(\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}})\Big\|_2^2 \\ &= \frac{1}{n^2}\Big\|[\mathbf{U}_{\hat{S}_\gamma}, \mathbf{W}_{\hat{S}_\beta}]^\top(\boldsymbol{\varepsilon} - \mathbf{U}\boldsymbol{\nu} - \mathbf{W}\boldsymbol{\Delta})\Big\|_2^2. \end{aligned} \tag{S9}$$

Using that $\|\mathbf{a}-\mathbf{b}\|_2^2 \le 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$ for any vectors $\mathbf{a}$ and $\mathbf{b}$ along with Lemma 7, we have with probability at least $1 - C_0 \exp(-\log(pq))$ that

$$\begin{aligned} \eta^2 \hat{s}_\gamma + \lambda^2 \hat{s}_\beta + \lambda_g^2 \hat{s}_{\beta,g} &\le \frac{2}{n^2}\left\| [\mathbf{U}_{\hat{S}_\gamma}, \mathbf{W}_{\hat{S}_\beta}]^\top \boldsymbol{\varepsilon} \right\|_2^2 + \frac{2}{n^2}\left\| [\mathbf{U}_{\hat{S}_\gamma}, \mathbf{W}_{\hat{S}_\beta}]^\top (\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}) \right\|_2^2 \\ &\le \frac{2}{n} M_{uw} \|\mathcal{P}_{\mathcal{I}_0,\mathcal{J}_0}(\boldsymbol{\varepsilon})\|_2^2 + \frac{2}{n} M_{uw} \|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2, \end{aligned} \tag{S10}$$

where the last inequality uses that $\hat{S}_\gamma \subset \mathcal{I}_0$ and $\hat{S}_\beta \subset \mathcal{J}_0$.

Now let

$$\begin{aligned} \eta &= C\frac{\sigma_\varepsilon}{\sqrt{n}}\left(\log(eq/s_\gamma) + \frac{s_\beta}{s_\gamma}\log(ep) + \frac{s_{\beta,g}}{s_\gamma}\log(eq/s_{\beta,g}) \right)^{1/2}, \\ \lambda &= C\frac{\sigma_\varepsilon}{\sqrt{n}}\left(\frac{s_\gamma}{s_\beta}\log(eq/s_\gamma) + \log(ep) + \frac{s_{\beta,g}}{s_\beta}\log(eq/s_{\beta,g}) \right)^{1/2}, \\ \lambda_g &= C\frac{\sigma_\varepsilon}{\sqrt{n}}\left(\frac{s_\gamma}{s_{\beta,g}}\log(eq/s_\gamma) + \frac{s_\beta}{s_{\beta,g}}\log(ep) + \log(eq/s_{\beta,g}) \right)^{1/2}. \end{aligned}$$

where $C = \sqrt{5M_{uw}a_2}$ for some $a_2 > 2$.

Combining (S10) and (S4) gives

$$\begin{aligned} &\left(1 - \frac{2}{a_2}\right)\|\mathcal{P}_{\mathcal{I}_0,\mathcal{J}_0}(\boldsymbol{\varepsilon})\|_2^2 \\ &\quad\le 5\sigma_\varepsilon^2(s_\gamma \log(eq/s_\gamma) + s_\beta \log(ep) + s_{\beta,g}\log(eq/s_{\beta,g})) + \frac{2}{a_2}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 + r \end{aligned} \tag{S11}$$

It is then straightforward to multiply both sides by $a_1 a_2/(2(a_2-2))$ and plug into (S1) and (S2) to get

$$\begin{aligned} &\frac{1}{n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 + \eta\|\boldsymbol{\nu}_{S_\gamma^c}\|_1 + \lambda\|\boldsymbol{\Delta}_{S_\beta^c}\|_1 + \lambda_g\|\boldsymbol{\Delta}_{S_{\beta,g}^c}\|_{1,2} \\ &\quad\le \frac{1}{2a_1}\cdot\frac{1}{n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 + \frac{5a_1a_2}{2(a_2-2)}\cdot E_j + \frac{a_1}{a_2-2}\cdot\frac{1}{n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 \\ &\qquad + \frac{a_1a_2}{2(a_2-2)}\cdot\frac{r}{n} + \eta\|\boldsymbol{\nu}_{S_\gamma}\|_1 + \lambda\|\boldsymbol{\Delta}_{S_\beta}\|_1 + \lambda_g\|\boldsymbol{\Delta}_{S_{\beta,g}}\|_{1,2} \end{aligned} \tag{S12}$$

where $E_j = (\sigma_\varepsilon^2/n)\cdot(s_\gamma\log(eq/s_\gamma) + s_\beta\log(ep) + s_{\beta,g}\log(eq/s_{\beta,g}))$.

We also have that

$$\begin{aligned} \frac{\|\boldsymbol{\nu}_{S_\gamma}\|_1}{\sqrt{s_\gamma}} + \frac{\|\boldsymbol{\Delta}_{S_\beta}\|_1}{\sqrt{s_\beta}} + \frac{\|\boldsymbol{\Delta}_{S_{\beta,g}}\|_{1,2}}{\sqrt{s_{\beta,g}}} &\le \|\boldsymbol{\nu}_{S_\gamma}\|_2 + \|\boldsymbol{\Delta}_{S_\beta}\|_2 + \|\boldsymbol{\Delta}_{S_{\beta,g}}\|_2 \\ &\le 2\left\|\begin{pmatrix}\boldsymbol{\nu} \\ \boldsymbol{\Delta}\end{pmatrix}\right\|_2 \le 2m_0^{-1/2}\left\|\boldsymbol{\Sigma}_{\mathbf{UW}}^{1/2}\begin{pmatrix}\boldsymbol{\nu} \\ \boldsymbol{\Delta}\end{pmatrix}\right\|_2 \end{aligned}$$

where the last inequality follows from Assumption 2. With our chosen values of η, λ, and λ_g, we have after multiplying through by $C\sqrt{E_j}$ that

$$\begin{aligned}\eta\|\boldsymbol{\nu}_{S_\gamma}\|_1 + \lambda\|\boldsymbol{\Delta}_{S_\beta}\|_1 + \lambda_g\|\boldsymbol{\Delta}_{S_{\beta,g}}\|_{1,2} &\le 2Cm_0^{-1/2}\sqrt{E_j}\left\|\boldsymbol{\Sigma}_{\mathbf{UW}}^{1/2}\begin{pmatrix}\boldsymbol{\nu}\\ \boldsymbol{\Delta}\end{pmatrix}\right\|_2 \\ &\le a_3\frac{C^2}{m_0}E_j + \frac{1}{a_3}\left\|\boldsymbol{\Sigma}_{\mathbf{UW}}^{1/2}\begin{pmatrix}\boldsymbol{\nu}\\ \boldsymbol{\Delta}\end{pmatrix}\right\|_2^2\end{aligned}$$

for any $a_3 > 0$. Plugging this into (S12) gives

$$\begin{aligned}&\left(1 - \frac{1}{2a_1} - \frac{a_1}{a_2 - 2}\right)\frac{1}{n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2 \\ &\quad\le \left(\frac{5a_1a_2}{2(a_2-2)} + a_3\frac{C^2}{m_0}\right)E_j + \frac{1}{a_3}\left\|\boldsymbol{\Sigma}_{\mathbf{UW}}^{1/2}\begin{pmatrix}\boldsymbol{\nu}\\ \boldsymbol{\Delta}\end{pmatrix}\right\|_2^2 + \frac{a_1a_2}{2(a_2-2)}\cdot\frac{r}{n}.\end{aligned}\tag{S13}$$

Step 3

We now bound the difference between $\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2/n$ and $\left\|\boldsymbol{\Sigma}_{\mathbf{UW}}^{1/2}\begin{pmatrix}\boldsymbol{\nu}\\ \boldsymbol{\Delta}\end{pmatrix}\right\|_2^2$. Let $L > 0$ be an arbitrarily large constant. By Assumption 4 and Lemma 5, we have with probability at least $1 - C'\exp(-\log(pq))$ that

$$\sup_{\|\mathbf{v}\|_0 \le 2(s_\gamma + s_\beta),\, \|\mathbf{v}\|_2 = 1}\left|\mathbf{v}^\top\left(\frac{[\mathbf{U},\mathbf{W}]^\top[\mathbf{U},\mathbf{W}]}{n} - \boldsymbol{\Sigma}_{\mathbf{UW}}\right)\mathbf{v}\right| \le \frac{1}{27L}$$

for sufficiently large n. By Lemma 6, it holds with probability at least $1 - C'\exp(-\log(pq))$ that

$$\begin{aligned}&\left|(\boldsymbol{\nu}^\top,\, \boldsymbol{\Delta}^\top)\left(\frac{[\mathbf{U},\mathbf{W}]^\top[\mathbf{U},\mathbf{W}]}{n} - \boldsymbol{\Sigma}_{\mathbf{UW}}\right)\begin{pmatrix}\boldsymbol{\nu}\\ \boldsymbol{\Delta}\end{pmatrix}\right| \\ &\quad\le \frac{1}{L}\left(\|\boldsymbol{\nu}\|_2^2 + \|\boldsymbol{\Delta}\|_2^2 + \frac{1}{s_\gamma + s_\beta}(\|\boldsymbol{\nu}\|_1^2 + \|\boldsymbol{\Delta}\|_1^2)\right) \\ &\quad\le \frac{1}{L}\left(\|\boldsymbol{\nu}\|_2^2 + \|\boldsymbol{\Delta}\|_2^2 + \frac{1}{s_\gamma}\|\boldsymbol{\nu}\|_1^2 + \frac{1}{s_\beta}\|\boldsymbol{\Delta}\|_1^2\right)\end{aligned}\tag{S14}$$

for sufficiently large n. Plugging this into (S13) gives

$$\begin{aligned}&\left(1 - \frac{1}{2a_1} - \frac{a_1}{a_2-2} - \frac{1}{a_3}\right)\left\|\boldsymbol{\Sigma}_{\mathbf{UW}}^{1/2}\begin{pmatrix}\boldsymbol{\nu}\\ \boldsymbol{\Delta}\end{pmatrix}\right\|_2^2 \\ &\quad\le \left(\frac{5a_1a_2}{2(a_2-2)} + a_3\frac{C^2}{m_0}\right)E_j + \frac{a_1a_2}{2(a_2-2)}\cdot\frac{r}{n} \\ &\qquad + \left(1 - \frac{1}{2a_1} - \frac{a_1}{a_2-2}\right)\left(\left\|\boldsymbol{\Sigma}_{\mathbf{UW}}^{1/2}\begin{pmatrix}\boldsymbol{\nu}\\ \boldsymbol{\Delta}\end{pmatrix}\right\|_2^2 - \frac{1}{n}\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}\|_2^2\right) \\ &\quad\le \left(\frac{5a_1a_2}{2(a_2-2)} + a_3\frac{C^2}{m_0}\right)E_j + \frac{a_1a_2}{2(a_2-2)}\cdot\frac{r}{n} \\ &\qquad + \left(1 - \frac{1}{2a_1} - \frac{a_1}{a_2-2}\right)\cdot\frac{1}{L}\left(\|\boldsymbol{\nu}\|_2^2 + \|\boldsymbol{\Delta}\|_2^2 + \frac{1}{s_\gamma}\|\boldsymbol{\nu}\|_1^2 + \frac{1}{s_\beta}\|\boldsymbol{\Delta}\|_1^2\right).\end{aligned}$$

Recalling that the above holds for any $a_1 > 0$, $a_2 > 2$, and $a_3 > 0$, choose $a_1 = 2$, $a_2 = 6$, and $a_3 = 6$ to get

$$\frac{1}{2}\left\|\mathbf{\Sigma}_{\mathbf{UW}}^{1/2}\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_2^2 \precsim E_j + \frac{1}{L}\left(\|\boldsymbol{\nu}\|_2^2 + \|\mathbf{\Delta}\|_2^2 + \frac{1}{s_\gamma}\|\boldsymbol{\nu}\|_1^2 + \frac{1}{s_\beta}\|\mathbf{\Delta}\|_1^2\right) + \frac{r}{n}.$$

By the concentration bound on r in (S5), we in fact have

$$\frac{1}{2}\left\|\mathbf{\Sigma}_{\mathbf{UW}}^{1/2}\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_2^2 \precsim E_j + \frac{1}{L}\left(\|\boldsymbol{\nu}\|_2^2 + \|\mathbf{\Delta}\|_2^2 + \frac{1}{s_\gamma}\|\boldsymbol{\nu}\|_1^2 + \frac{1}{s_\beta}\|\mathbf{\Delta}\|_1^2\right) \tag{S15}$$

with probability at least $1 - c_1 \exp\{-c_2(s_\gamma \log(eq/s_\gamma) + s_\beta \log(ep) + s_{\beta,g} \log(eq/s_{\beta,g}))\}$. Then by Assumption 2 we have with the same high probability

$$\frac{m_0}{2}(\|\boldsymbol{\nu}\|_2^2 + \|\mathbf{\Delta}\|_2^2) \precsim E_j + \frac{1}{L}\left(\|\boldsymbol{\nu}\|_2^2 + \|\mathbf{\Delta}\|_2^2 + \frac{1}{s_\gamma}\|\boldsymbol{\nu}\|_1^2 + \frac{1}{s_\beta}\|\mathbf{\Delta}\|_1^2\right). \tag{S16}$$

Next, taking $a_1 = 2 - \sqrt{2}$, $a_2 = 6$ in (S12) cancels out the $\|\mathbf{U}\boldsymbol{\nu} + \mathbf{W}\mathbf{\Delta}\|_2^2/n$ terms. Appealing again to (S5) yields

$$\frac{\|\boldsymbol{\nu}_{S_\gamma^c}\|_1}{\sqrt{s_\gamma}} + \frac{\|\mathbf{\Delta}_{S_\beta^c}\|_1}{\sqrt{s_\beta}} + \frac{\|\mathbf{\Delta}_{(S_{\beta,g}^c)}\|_{1,2}}{\sqrt{s_{\beta,g}}} \leq \sqrt{E_j} + \frac{\|\boldsymbol{\nu}_{S_\gamma}\|_1}{\sqrt{s_\gamma}} + \frac{\|\mathbf{\Delta}_{S_\beta}\|_1}{\sqrt{s_\beta}} + \frac{\|\mathbf{\Delta}_{(S_{\beta,g})}\|_{1,2}}{\sqrt{s_{\beta,g}}}.$$

with probability at least $1 - c_1 \exp\{-c_2(s_\gamma \log(eq/s_\gamma) + s_\beta \log(ep) + s_{\beta,g} \log(eq/s_{\beta,g}))\}$. Adding $\|\boldsymbol{\nu}_{S_\gamma}\|_1/\sqrt{s_\gamma} + \|\mathbf{\Delta}_{S_\beta}\|_1/\sqrt{s_\beta}$ to both sides and using $\|\boldsymbol{\nu}_{S_\gamma}\|_1 \leq \sqrt{s_\gamma}\|\boldsymbol{\nu}\|_2$, $\|\mathbf{\Delta}_{S_\beta}\|_1 \leq \sqrt{s_\beta}\|\mathbf{\Delta}\|_2$, and $\|\mathbf{\Delta}_{(S_{\beta,g})}\|_{1,2} \leq \sqrt{s_{\beta,g}}\|\mathbf{\Delta}\|_2$ yields

$$\frac{\|\boldsymbol{\nu}\|_1}{\sqrt{s_\gamma}} + \frac{\|\mathbf{\Delta}\|_1}{\sqrt{s_\beta}} \leq \sqrt{E_j} + 2\|\boldsymbol{\nu}\|_2 + 3\|\mathbf{\Delta}\|_2$$

and after squaring both sides we have

$$\frac{\|\boldsymbol{\nu}\|_1^2}{s_\gamma} + \frac{\|\mathbf{\Delta}\|_1^2}{s_\beta} \leq k_0 E_j + k_1\|\boldsymbol{\nu}\|_2^2 + k_2\|\mathbf{\Delta}\|_2^2 \tag{S17}$$

for some absolute constants $k_0, k_1, k_2 > 0$. Plugging (S17) into the right-hand side of (S16) yields

$$\frac{m_0}{2}(\|\boldsymbol{\nu}\|_2^2 + \|\mathbf{\Delta}\|_2^2) \precsim \left(1 + \frac{k_0}{L}\right)E_j + \frac{k_1 + 1}{L}\|\boldsymbol{\nu}\|_2^2 + \frac{k_2 + 1}{L}\|\mathbf{\Delta}\|_2^2$$

Finally, plugging in the expression for E_j and recalling that L is arbitrarily large yields

$$\|\boldsymbol{\nu}\|_2^2 + \|\mathbf{\Delta}\|_2^2 \precsim \frac{\sigma_\varepsilon^2}{n}(s_\gamma \log(eq/s_\gamma) + s_\beta \log(ep) + s_{\beta,g} \log(eq/s_{\beta,g}))$$

with probability at least $1 - C_1 \exp\{-C_2(s_\gamma \log(eq/s_\gamma) + s_\beta \log(ep) + s_{\beta,g} \log(eq/s_{\beta,g}))\}$ for some positive constants C_1, C_2 as desired. □

S4.2 Proof of Theorem 2

The proof of Theorem 2 occurs in three steps.

Step 1

Recall that $\hat{\mathbf{\Sigma}}_{\mathbf{UW}} = [\mathbf{U}, \mathbf{W}]^\top [\mathbf{U}, \mathbf{W}]/n$. We wish to show that with high probability

$$\left\| \hat{\mathbf{\Sigma}}_{\mathbf{UW}} \begin{pmatrix} \boldsymbol{\nu} \\ \mathbf{\Delta} \end{pmatrix} \right\|_\infty \leq \frac{3}{2}(\eta + \lambda + \lambda_g). \tag{S18}$$

To ease the notation for this step, define the vectors

$$\mathbf{r}_\gamma = \frac{1}{n}\mathbf{U}^\top(\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}}), \quad \mathbf{r}_\beta = \frac{1}{n}\mathbf{W}^\top(\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}})$$

Let $\hat{\boldsymbol{\beta}}_h$ indicate the h-th block of $\hat{\boldsymbol{\beta}}$. By the KKT conditions, we know that an optimizer $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ of (14) must satisfy

$$\begin{cases} (\mathbf{r}_\gamma)_\ell = \eta \ \mathrm{sign}(\hat{\gamma}_\ell) & \hat{\gamma}_\ell \neq 0 \\ |(\mathbf{r}_\gamma)_\ell| \leq \eta & \hat{\gamma}_\ell = 0 \\ (\mathbf{r}_\beta)_\ell = \lambda \ \mathrm{sign}((\hat{\boldsymbol{\beta}}_0)_\ell) & (\hat{\boldsymbol{\beta}}_0)_\ell \neq 0 \\ (\mathbf{r}_\beta)_\ell = \lambda \ \mathrm{sign}((\hat{\boldsymbol{\beta}}_h)_\ell) + \lambda_g \frac{(\hat{\boldsymbol{\beta}}_h)_\ell}{\|\hat{\boldsymbol{\beta}}_h\|_2} & (\hat{\boldsymbol{\beta}}_h)_\ell \neq 0, \ h \in [q] \\ |(\mathbf{r}_\beta)_\ell| \leq \lambda + \lambda_g & (\hat{\boldsymbol{\beta}}_h)_\ell = 0, \ h \in [q] \end{cases}$$

so we have for all $\ell \in [p(q+1) - 1]$

$$\left| \left([\mathbf{U}, \mathbf{W}]^\top (\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}})/n \right)_\ell \right| \leq \eta + \lambda + \lambda_g.$$

Recalling that

$$\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}} = \mathbf{x}_j - \mathbf{U}\boldsymbol{\nu} - \mathbf{W}\mathbf{\Delta} - \mathbf{U}\boldsymbol{\gamma} - \mathbf{W}\boldsymbol{\beta} = \boldsymbol{\varepsilon} - \mathbf{U}\boldsymbol{\nu} - \mathbf{W}\mathbf{\Delta},$$

it follows by the triangle inequality that

$$\begin{aligned} \left\| \hat{\mathbf{\Sigma}}_{\mathbf{UW}} \begin{pmatrix} \boldsymbol{\nu} \\ \mathbf{\Delta} \end{pmatrix} \right\|_\infty - \left\| \frac{1}{n}[\mathbf{U}, \mathbf{W}]^\top \boldsymbol{\varepsilon} \right\|_\infty &\leq \left\| \frac{1}{n}[\mathbf{U}, \mathbf{W}]^\top \boldsymbol{\varepsilon} - \frac{1}{n}[\mathbf{U}, \mathbf{W}]^\top [\mathbf{U}, \mathbf{W}] \begin{pmatrix} \boldsymbol{\nu} \\ \mathbf{\Delta} \end{pmatrix} \right\|_\infty \\ &= \left\| \frac{1}{n}[\mathbf{U}, \mathbf{W}]^\top (\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}}) \right\|_\infty \\ &\leq \eta + \lambda + \lambda_g. \end{aligned}$$

Thus to show (S18) it suffices to show that with high probability

$$\left\| \frac{1}{n}[\mathbf{U}, \mathbf{W}]^\top \boldsymbol{\varepsilon} \right\|_\infty \leq \frac{\eta + \lambda + \lambda_g}{2}. \tag{S19}$$

To see this, let M_Σ denote the bound on the sub-Gaussian norm of elements of $\boldsymbol{X}$ as assumed in Assumption 3. Recall that each element of $\boldsymbol{U}$ is bounded by a constant M by Assumption 1.

Then the product of elements U_ℓ of $\boldsymbol{U}$ and X_ℓ of $\boldsymbol{X}$ with the error ε is sub-exponential with norm satisfying

$$\begin{aligned}
&\max_\ell \|U_\ell\, \varepsilon\|_{\psi_1} \leq \max_\ell \|U_\ell\|_{\psi_2} \|\varepsilon\|_{\psi_2} \leq M\sigma_\varepsilon \\
&\max_\ell \|X_\ell\, \varepsilon\|_{\psi_1} \leq \max_\ell \|U_\ell\|_{\psi_2} \max_\ell \|X_\ell\|_{\psi_2} \|\varepsilon\|_{\psi_2} \leq MM_\Sigma\sigma_\varepsilon.
\end{aligned}$$

Define the vector $\mathbf{v} = [\mathbf{U}, \mathbf{W}]^\top \boldsymbol{\varepsilon}/n$, where each element v_i is a sum of sub-exponential functions with bounded norm. By Lemma 4 we have

$$\mathbb{P}\left\{ |v_i| > \frac{1}{2}(\eta + \lambda + \lambda_g) \right\} \leq 2\exp\left(-c \min\left(\frac{n(\eta + \lambda + \lambda_g)^2}{4M^2 M_\Sigma^2 \sigma_\varepsilon^2}, \frac{n(\eta + \lambda + \lambda_g)}{2MM_\Sigma\sigma_\varepsilon} \right) \right).$$

where $c > 0$ is a constant coming from Lemma 4. Since by (16)

$$\eta + \lambda + \lambda_g \geq C\sigma_\varepsilon \sqrt{\frac{\log p}{n}} \iff \frac{(\eta + \lambda + \lambda_g)^2 n}{C^2\sigma_\varepsilon^2} \geq \log p,$$

we have

$$\frac{(\eta + \lambda + \lambda_g)^2 n}{C^2\sigma_\varepsilon^2} \cdot C_0'' \geq C_0'' \log p$$

where $C_0'' = C^2/(4M^2 M_\Sigma^2)$. Similarly, the inequality

$$\frac{(\eta + \lambda + \lambda_g)n}{C\sigma_\varepsilon} \geq \sqrt{n \log p}$$

implies

$$\frac{(\eta + \lambda + \lambda_g)n}{C\sigma_\varepsilon} \cdot \sqrt{C_0''} \geq \sqrt{C_0''} \cdot \sqrt{n \log p} \geq \tilde{A}^{-1} \sqrt{C_0''} \log p$$

provided that $\log p \leq \tilde{A}n$ for some constant $\tilde{A} > 0$, which we assumed in the hypothesis of Theorem 2. The above shows that we can pick either argument of the minimum in Lemma 4 to develop and that

$$\mathbb{P}\left\{ |v_i| > \frac{1}{2}(\eta + \lambda + \lambda_g) \right\} \leq 2\exp(-C_0' \log p).$$

for some constant C_0 that increases with C, recalling that C comes from (16). Then by the union bound we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n} \|[\mathbf{U}, \mathbf{W}]^\top \boldsymbol{\varepsilon}\|_\infty \geq \frac{\eta + \lambda + \lambda_g}{2} \right) &= \mathbb{P}\left(\max_i |v_i| \geq \frac{\eta + \lambda + \lambda_g}{2} \right) \\
&\leq (p(q+1)) \cdot 2\exp(-C_0' \log p) \\
&\leq 2\exp(-C_0' \log p + \log p + \log(q+1)) \\
&\leq 2\exp(-c_3 \log p)
\end{aligned}$$

for some $c_3 > 0$. In the last step we used $\log p \asymp \log q$. This shows (S19) holds with high probability.

Next we will show with high probability that

$$\|\boldsymbol{\nu}_{S_\gamma^c}\|_1 + \|\boldsymbol{\Delta}_{S_\beta^c}\|_1 \leq 8\tau_j(\|\boldsymbol{\nu}_{S_\gamma}\|_1 + \|\boldsymbol{\Delta}_{S_\beta}\|_1). \tag{S20}$$

Since $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ is optimal, we have that

$$\frac{1}{2n}\|\mathbf{x}_j - \mathbf{U}\hat{\boldsymbol{\gamma}} - \mathbf{W}\hat{\boldsymbol{\beta}}\|_2^2 + \eta\|\hat{\boldsymbol{\gamma}}\|_1 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{-0}\|_{1,2} \le \frac{1}{2n}\|\boldsymbol{\varepsilon}_j\|_2^2 + \eta\|\boldsymbol{\gamma}\|_1 + \lambda\|\boldsymbol{\beta}\|_1 + \lambda_g\|\boldsymbol{\beta}_{-0}\|_{1,2}.$$

Rearranging the above leads to

$$\eta\|\hat{\boldsymbol{\gamma}}\|_1 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{-0}\|_{1,2} \le \eta\|\boldsymbol{\gamma}\|_1 + \lambda\|\boldsymbol{\beta}\|_1 + \lambda_g\|\boldsymbol{\beta}_{-0}\|_{1,2} + \frac{1}{n}\langle \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}, \boldsymbol{\varepsilon}\rangle. \tag{S21}$$

By the same argument leading to (S19), the events

$$\mathcal{A}_1 = \left\{\frac{1}{n}\|\mathbf{U}^\top\boldsymbol{\varepsilon}\|_\infty \le \frac{\eta}{2}\right\} \quad \text{and} \quad \mathcal{A}_2 = \left\{\frac{1}{n}\|\mathbf{W}^\top\boldsymbol{\varepsilon}\|_\infty \le \frac{\lambda}{2}\right\}$$

hold with probability at least $1 - 2\exp(-c_4 \log p)$ and $1 - 2\exp(-c_5 \log p)$ respectively for some constants $c_4 > 0$ and $c_5 > 0$.

Conditional on $\mathcal{A}_1$ and $\mathcal{A}_2$, we have by Hölder's inequality

$$\begin{aligned}\frac{1}{n}\langle \mathbf{U}\boldsymbol{\nu} + \mathbf{W}\boldsymbol{\Delta}, \boldsymbol{\varepsilon}\rangle &= \frac{1}{n}\langle \mathbf{U}\boldsymbol{\nu}, \boldsymbol{\varepsilon}\rangle + \frac{1}{n}\langle \mathbf{W}\boldsymbol{\Delta}, \boldsymbol{\varepsilon}\rangle \\ &\le \frac{1}{n}\|\mathbf{U}^\top\boldsymbol{\varepsilon}\|_\infty\|\boldsymbol{\nu}\|_1 + \frac{1}{n}\|\mathbf{W}^\top\boldsymbol{\varepsilon}\|_\infty\|\boldsymbol{\Delta}\|_1 \\ &\le \frac{\eta}{2}\|\boldsymbol{\nu}\|_1 + \frac{\lambda}{2}\|\boldsymbol{\Delta}\|_1\end{aligned}$$

so that (S21) becomes

$$\eta\|\hat{\boldsymbol{\gamma}}\|_1 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 + \lambda_g\|\hat{\boldsymbol{\beta}}_{-0}\|_{1,2} \le \eta\|\boldsymbol{\gamma}\|_1 + \lambda\|\boldsymbol{\beta}\|_1 + \lambda_g\|\boldsymbol{\beta}_{-0}\|_{1,2} + \frac{\eta}{2}\|\boldsymbol{\nu}\|_1 + \frac{\lambda}{2}\|\boldsymbol{\Delta}\|_1.$$

Now multiply through by 2, bring the LHS over to the RHS, and add $\eta\|\boldsymbol{\nu}\|_1 + \lambda\|\boldsymbol{\Delta}\|_1$ to both sides to find

$$\begin{aligned}\eta\|\boldsymbol{\nu}\|_1 + \lambda\|\boldsymbol{\Delta}\|_1 \le\ & 2\eta(\|\boldsymbol{\gamma}\|_1 - \|\hat{\boldsymbol{\gamma}}\|_1 + \|\boldsymbol{\nu}\|_1) \\ &+ 2\lambda\Big(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}\|_1 + \|\boldsymbol{\Delta}\|_1\Big) \\ &+ 2\lambda_g\Big(\|\boldsymbol{\beta}_{-0}\|_{1,2} - \|\hat{\boldsymbol{\beta}}_{-0}\|_{1,2} + \|\boldsymbol{\Delta}_{-0}\|_{1,2}\Big),\end{aligned}$$

where we also added $2\lambda_g\|\boldsymbol{\Delta}_{-0}\|_{1,2}$ to the RHS. Then using the triangle inequality and the fact that the terms in parentheses vanish outside of the respective support sets S_γ, S_β, and $S_{\beta,g}$, we have

$$\eta\|\boldsymbol{\nu}\|_1 + \lambda\|\boldsymbol{\Delta}\|_1 \le 4\eta\|\boldsymbol{\nu}_{S_\gamma}\|_1 + 4\lambda\|\boldsymbol{\Delta}_{S_\beta}\|_1 + 4\lambda_g\|\boldsymbol{\Delta}_{(S_{\beta,g})}\|_{1,2}. \tag{S22}$$

Now recall in Assumption 5 we defined

$$\tau_j = 1 + \max\left(\sqrt{\frac{s_\gamma}{s_\beta}} + \sqrt{\frac{s_\gamma}{s_{\beta,g}}}, \sqrt{\frac{s_\beta}{s_\gamma}} + \sqrt{\frac{s_\beta}{s_{\beta,g}}}, \sqrt{\frac{s_{\beta,g}}{s_\gamma}} + \sqrt{\frac{s_{\beta,g}}{s_\beta}}\right)$$

and that $\eta\sqrt{s_\gamma} = \lambda\sqrt{s_\beta} = \lambda_g\sqrt{s_{\beta,g}}$ in (16). This yields

$$\frac{1}{\tau_j} = \min\left(\frac{\eta}{\eta + \lambda + \lambda_g}, \frac{\lambda}{\eta + \lambda + \lambda_g}, \frac{\lambda_g}{\eta + \lambda + \lambda_g}\right).$$

Divide (S22) by $\eta + \lambda + \lambda_g$ to find

$$\|\boldsymbol{\nu}_{S_\gamma^c}\|_1 + \|\boldsymbol{\Delta}_{S_\beta^c}\|_1 \le \|\boldsymbol{\nu}\|_1 + \|\boldsymbol{\Delta}\|_1 \le 4\tau_j(\|\boldsymbol{\nu}_{S_\gamma}\|_1 + 2\|\boldsymbol{\Delta}_{S_\beta}\|_1) \le 8\tau_j(\|\boldsymbol{\nu}_{S_\gamma}\|_1 + \|\boldsymbol{\Delta}_{S_\beta}\|_1)$$

as desired.

Step 2

In this step, we bound the diagonal difference $|\hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell,\ell) - \mathbf{\Sigma}_{\mathbf{UW}}(\ell,\ell)|$. Notice that $\hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell,\ell)$ is a sum of sub-exponential random variables with sub-exponential norm bounded by $M^2M_\Sigma^2$, where M is from Assumption 1 and M_Σ from Assumption 3. We have by Lemma 4

$$\begin{aligned}\mathbb{P}\Big(|\hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell,\ell) - \mathbf{\Sigma}_{\mathbf{UW}}(\ell,\ell)| > \frac{1}{2}\mathbf{\Sigma}_{\mathbf{UW}}(\ell,\ell)\Big) &\le 2\exp\Big(-cn\min\Big(\frac{\mathbf{\Sigma}_{\mathbf{UW}}(\ell,\ell)^2}{4M^2M_\Sigma^2}, \frac{\mathbf{\Sigma}_{\mathbf{UW}}(\ell,\ell)}{2MM_\Sigma}\Big)\Big)\\ &\le 2\exp\Big(-cn\min\Big(\frac{m_0^2}{4M^2M_\Sigma^2}, \frac{m_0}{2MM_\Sigma}\Big)\Big)\\ &\le 2\exp(-c_6 n)\end{aligned}$$

where we used $m_0 \le \mathbf{\Sigma}_{\mathbf{UW}}(\ell,\ell)$ by Assumption 2. Since

$$|\hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell,\ell) - \mathbf{\Sigma}_{\mathbf{UW}}(\ell,\ell)| \le \frac{1}{2}\mathbf{\Sigma}_{\mathbf{UW}}(\ell,\ell)$$

implies

$$\frac{1}{2}\mathbf{\Sigma}_{\mathbf{UW}}(\ell,\ell) \le \hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell,\ell) \le 2\mathbf{\Sigma}_{\mathbf{UW}}(\ell,\ell)$$

which by Assumption 2 implies

$$\frac{m_0}{2} \le \hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell,\ell) \le 2M_0,$$

it follows from the above concentration bound that

$$\mathbb{P}\Big(\frac{m_0}{2} \le \hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell,\ell) \le 2M_0\Big) \ge 1 - 2\exp(-c_6 n). \tag{S23}$$

Next we bound the off-diagonal terms. Again by Lemma 4

$$\begin{aligned}&\mathbb{P}\Big(|\hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell,k) - \mathbf{\Sigma}_{\mathbf{UW}}(\ell,k)| \ge 2\mathbf{\Sigma}_{\mathbf{UW}}(\ell,k)\Big)\\ &\quad\le 2\exp\Big(-cn\min\Big(\frac{4\mathbf{\Sigma}_{\mathbf{UW}}(\ell,k)^2}{M^2M_\Sigma^2}, \frac{2\mathbf{\Sigma}_{\mathbf{UW}}(\ell,k)}{MM_\Sigma}\Big)\Big)\\ &\quad\le 2\exp\Big(-cn\min\Big(\frac{4M_0^2}{M^2M_\Sigma^2}, \frac{4M_0}{MM_\Sigma}\Big)\Big)\\ &\quad\le 2\exp(-c_7 n).\end{aligned}$$

By Assumption 5, we have

$$\begin{aligned}&-2\mathbf{\Sigma}_{\mathbf{UW}}(k,\ell) \le \hat{\mathbf{\Sigma}}_{\mathbf{UW}}(k,\ell) - \mathbf{\Sigma}_{\mathbf{UW}}(k,\ell) \le 2\mathbf{\Sigma}_{\mathbf{UW}}(k,\ell)\\ &\implies -\mathbf{\Sigma}_{\mathbf{UW}}(k,\ell) \le \hat{\mathbf{\Sigma}}_{\mathbf{UW}}(k,\ell) \le 3\mathbf{\Sigma}_{\mathbf{UW}}(k,\ell)\\ &\implies -\frac{1}{c_0(1+16\tau_j)(s_\beta+s_\gamma)} \le \hat{\mathbf{\Sigma}}_{\mathbf{UW}}(k,\ell) \le \frac{3}{c_0(1+16\tau_j)(s_\beta+s_\gamma)}\end{aligned}$$

Combining this with the above concentration bound yields

$$\mathbb{P}\Big(-\frac{1}{c_0(1+16\tau_j)(s_\beta+s_\gamma)} \le \mathbf{\Sigma}_{\mathbf{UW}}(k,\ell) \le \frac{3}{c_0(1+16\tau_j)(s_\beta+s_\gamma)}\Big) \ge 1-2\exp(-c_7 n). \tag{S24}$$

Step 3

Define the index set $\tilde{S} = S_\gamma \cup \{q + i \mid i \in S_\beta\}$. With slight abuse of notation, for a vector $\mathbf{v} \in \mathbb{R}^{p(q+1)-1}$ let $\mathbf{v}_{\tilde{S}} \in \mathbb{R}^{p(q+1)-1}$ be the vector that equals $\mathbf{v}$ on the set $\tilde{S}$ and is zero on $\tilde{S}^c$, so that it has $s_\gamma + s_\beta$ nonzero elements. Define $\mathbf{v}_{\tilde{S}^c}$ analogously.

We will show that conditional on $\mathcal{A}_1$, $\mathcal{A}_2$, and the bounds

$$\begin{gathered} \frac{m_0}{2} \leq \hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell, \ell) \leq 2M_0, \\ -\frac{1}{c_0(1+16\tau_j)(s_\beta + s_\gamma)} \leq \hat{\mathbf{\Sigma}}_{\mathbf{UW}}(k, \ell) \leq \frac{3}{c_0(1+16\tau_j)(s_\beta + s_\gamma)}, \end{gathered} \tag{S25}$$

we have

$$\inf_{\mathbf{v}\in\mathcal{V}} \frac{\|[\mathbf{U}, \mathbf{W}]\mathbf{v}\|_2}{\sqrt{n}\|\mathbf{v}_{\tilde{S}}\|_2} \geq \sqrt{\frac{m_0}{2} - \frac{1}{c_0}} > 0.$$

where $\mathcal{V} = \{\mathbf{v} \in \mathbb{R}^{p(q+1)-1} \mid \|\mathbf{v}_{\tilde{S}^c}\|_1 \leq 8\tau_j \|\mathbf{v}_{\tilde{S}}\|_1\}$

First we have that

$$\begin{aligned} \frac{\|[\mathbf{U}, \mathbf{W}]\mathbf{v}_{\tilde{S}}\|_2^2}{n\|\mathbf{v}_{\tilde{S}}\|_2^2} &= \frac{\mathbf{v}_{\tilde{S}}^\top \operatorname{diag}(\hat{\mathbf{\Sigma}}_{\mathbf{UW}})\mathbf{v}_{\tilde{S}}}{\|\mathbf{v}_{\tilde{S}}\|_2^2} + \frac{\mathbf{v}_{\tilde{S}}^\top (\hat{\mathbf{\Sigma}}_{\mathbf{UW}} - \operatorname{diag}(\hat{\mathbf{\Sigma}}_{\mathbf{UW}}))\mathbf{v}_{\tilde{S}}}{\|\mathbf{v}_{\tilde{S}}\|_2^2} \\ &\geq \frac{m_0}{2} - \frac{1}{c_0(1+16\tau_j)(s_\beta + s_\gamma)} \frac{\|\mathbf{v}_{\tilde{S}}\|_1^2}{\|\mathbf{v}_{\tilde{S}}\|_2^2}. \end{aligned}$$

Then by $\|\mathbf{a} + \mathbf{b}\|_2^2 = \|\mathbf{a}\|_2^2 + 2\mathbf{a}^\top\mathbf{b} + \|\mathbf{b}\|_2^2$ we have

$$\begin{aligned} \frac{\|[\mathbf{U}, \mathbf{W}]\mathbf{v}\|_2^2}{n\|\mathbf{v}_{\tilde{S}}\|_2^2} &\geq \frac{\|[\mathbf{U}, \mathbf{W}]\mathbf{v}_{\tilde{S}}\|_2^2}{n\|\mathbf{v}_{\tilde{S}}\|_2^2} + 2\frac{\mathbf{v}_{\tilde{S}}^\top \hat{\mathbf{\Sigma}}_{\mathbf{UW}}\mathbf{v}_{\tilde{S}^c}}{n\|\mathbf{v}_{\tilde{S}}\|_2^2} \\ &\geq \frac{m_0}{2} - \frac{1}{c_0(1+16\tau_j)(s_\beta + s_\gamma)} \frac{\|\mathbf{v}_{\tilde{S}}\|_1^2}{\|\mathbf{v}_{\tilde{S}}\|_2^2} - \frac{2}{c_0(1+16\tau_j)(s_\beta + s_\gamma)} \frac{\|\mathbf{v}_{\tilde{S}}\|_1 \|\mathbf{v}_{\tilde{S}^c}\|_1}{\|\mathbf{v}_{\tilde{S}}\|_2^2} \\ &\geq \frac{m_0}{2} - \frac{1}{c_0(1+16\tau_j)(s_\beta + s_\gamma)} \frac{\|\mathbf{v}_{\tilde{S}}\|_1^2}{\|\mathbf{v}_{\tilde{S}}\|_2^2} - \frac{16\tau_j}{c_0(1+16\tau_j)(s_\beta + s_\gamma)} \frac{\|\mathbf{v}_{\tilde{S}}\|_1^2}{\|\mathbf{v}_{\tilde{S}}\|_2^2} \\ &\geq \frac{m_0}{2} - \frac{1+16\tau_j}{c_0(1+16\tau_j)(s_\beta + s_\gamma)} \frac{\|\mathbf{v}_{\tilde{S}}\|_1^2}{\|\mathbf{v}_{\tilde{S}}\|_2^2} \\ &\geq \frac{m_0}{2} - \frac{(1+16\tau_j)(s_\beta + s_\gamma)}{c_0(1+16\tau_j)(s_\beta + s_\gamma)} = \frac{m_0}{2} - \frac{1}{c_0} > 0. \end{aligned}$$

In the last step we used that $c_0 > 2/m_0$ in Assumption 5. We have shown that

$$\frac{1}{n}\left\|[\mathbf{U}, \mathbf{W}]\begin{pmatrix}\boldsymbol{\nu} \\ \mathbf{\Delta}\end{pmatrix}\right\|_2^2 \geq \left(\frac{m_0}{2} - \frac{1}{c_0}\right)\left\|\begin{pmatrix}\boldsymbol{\nu} \\ \mathbf{\Delta}\end{pmatrix}_{\tilde{S}}\right\|_2^2 \tag{S26}$$

Final step

It is true that for $\ell \in [p(q+1)-1]$ we have

$$\left(\hat{\mathbf{\Sigma}}_{\mathbf{UW}}\begin{pmatrix}\boldsymbol{\nu} \\ \mathbf{\Delta}\end{pmatrix}\right)_\ell = \hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell, \ell)\begin{pmatrix}\boldsymbol{\nu} \\ \mathbf{\Delta}\end{pmatrix}_\ell + \sum_{k\neq\ell} \hat{\mathbf{\Sigma}}_{\mathbf{UW}}(k, \ell)\begin{pmatrix}\boldsymbol{\nu} \\ \mathbf{\Delta}\end{pmatrix}_k.$$

Then by (S25) and the triangle inequality we have

$$\begin{aligned}\left|\hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell,\ell)\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_\ell\right| - \left|\left(\hat{\mathbf{\Sigma}}_{\mathbf{UW}}\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right)_\ell\right| &\le \left|\hat{\mathbf{\Sigma}}_{\mathbf{UW}}(\ell,\ell)\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_\ell - \left(\hat{\mathbf{\Sigma}}_{\mathbf{UW}}\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right)_\ell\right|\\ &\le \left|\sum_{k\neq\ell}\hat{\mathbf{\Sigma}}_{\mathbf{UW}}(k,\ell)\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_k\right|\\ &\le \frac{3}{c_0(1+16\tau_j)(s_\gamma+s_\beta)}\sum_{k\neq\ell}\left|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_k\right|.\end{aligned}$$

Rearranging terms and applying (S25) yields

$$\frac{m_0}{2}\left|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_\ell\right| \le \left|\left(\hat{\mathbf{\Sigma}}_{\mathbf{UW}}\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right)_\ell\right| + \frac{3}{c_0(1+16\tau_j)(s_\gamma+s_\beta)}\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_1.$$

Since this holds for all $\ell \in [p(q+1)-1]$, we have shown that

$$\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_\infty \le \frac{2}{m_0}\left\|\hat{\mathbf{\Sigma}}_{\mathbf{UW}}\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_\infty + \frac{6}{c_0 m_0(1+16\tau_j)(s_\gamma+s_\beta)}\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_1. \tag{S27}$$

Combining (S18) and (S20) we have

$$\begin{aligned}\frac{1}{n}\left\|[\mathbf{U},\mathbf{W}]\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_2^2 &\le \left\|\hat{\mathbf{\Sigma}}_{\mathbf{UW}}\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_\infty\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_1\\ &\le \frac{3}{2}(\eta+\lambda+\lambda_g)\cdot(1+8\tau_j)\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_{\tilde{S}}\right\|_1\\ &\le \frac{3}{2}(\eta+\lambda+\lambda_g)\cdot(1+8\tau_j)\sqrt{s_\gamma+s_\beta}\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_{\tilde{S}}\right\|_2\end{aligned}$$

and combining this with (S26) gives us

$$\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_{\tilde{S}}\right\|_2^2 \le \frac{3}{2}(\eta+\lambda+\lambda_g)\cdot(1+8\tau_j)\left(\frac{2c_0}{c_0m_0-2}\right)\sqrt{s_\gamma+s_\beta}\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_{\tilde{S}}\right\|_2$$

and therefore

$$\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_{\tilde{S}}\right\|_2 \le 3(\eta+\lambda+\lambda_g)(1+8\tau_j)\left(\frac{c_0}{c_0m_0-2}\right)\sqrt{s_\gamma+s_\beta}.$$

On the other hand, by (S20) we have

$$\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_1 \le (1+8\tau_j)\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_{\tilde{S}}\right\|_1 \le (1+8\tau_j)\sqrt{s_\beta+s_\gamma}\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}_{\tilde{S}}\right\|_2$$

so combining this with the above yields

$$\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \mathbf{\Delta}\end{pmatrix}\right\|_1 \le 3(\eta+\lambda+\lambda_g)(1+8\tau_j)^2\left(\frac{c_0}{c_0m_0-2}\right)(s_\gamma+s_\beta)$$

Finally, plugging this into (S27) yields

$$\begin{aligned}\left\|\begin{pmatrix}\boldsymbol{\nu}\\ \boldsymbol{\Delta}\end{pmatrix}\right\|_\infty &\leq \frac{3}{m_0}(\eta+\lambda+\lambda_g)+\frac{18(\eta+\lambda+\lambda_g)(1+8\tau_j)^2}{c_0m_0(1+16\tau_j)}\left(\frac{c_0}{c_0m_0-2}\right)\\ &= \frac{3}{m_0}(\eta+\lambda+\lambda_g)\left(1+\frac{6(1+8\tau_j)^2}{(1+16\tau_j)(c_0m_0-2)}\right)\end{aligned}$$

as desired. □