Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

High-Dimensional Gaussian Graphical Regression Models with Covariates

Jingfei Zhang & Yi Li

To cite this article: Jingfei Zhang & Yi Li (2022): High-Dimensional Gaussian Graphical Regression Models with Covariates, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2034632

To link to this article: https://doi.org/10.1080/01621459.2022.2034632

View supplementary material

Published online: 14 Mar 2022.

Submit your article to this journal

Article views: 1400

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2022.2034632

Taylor & Francis
Taylor & Francis Group

Check for updates

High-Dimensional Gaussian Graphical Regression Models with Covariates

Jingfei Zhang[a] and Yi Li[b]

[a]Department of Management Science, University of Miami, Coral Gables, FL; [b]Department of Biostatistics, University of Michigan, Ann Arbor, MI

ABSTRACT

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

ARTICLE HISTORY
Received March 2021
Accepted January 2022

KEYWORDS
Co-expression QTL; Gaussian graphical model with covariates; Nonasymptotic convergence rate; Subject-specific Gaussian graphical model; Sparse group lasso

1. Introduction

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

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

Though the literature on graphical models has been steadily growing (e.g., Meinshausen and Bühlmann 2006; Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008; Peng et al. 2009; Fan, Feng, and Wu 2009; Xie et al. 2020), relatively few frameworks permit subject-specific graphical model estimation with theoretical justifications. Several works (Rothman, Levina, and Zhu 2010; Yin and Li 2011; Li, Chun, and Zhao 2012; Lee and Liu 2012; Cai et al. 2012; Lin et al. 2016; Chen et al. 2016) considered covariate-dependent Gaussian graphical models, wherein the mean of the nodes depends on covariates, while the network structure is constant across all of the subjects. Guo et al. (2011) and Danaher, Wang, and Witten (2014) jointly estimated several group-specific Gaussian graphical models, where the graph structure is allowed to vary with discrete covariates; Liu et al. (2010) proposed a graph-valued regression, which partitions the covariate space into several subspaces and fits separate Gaussian graphical models for each subspace using graphical lasso. As noted by Cheng et al. (2014), it may be difficult to interpret the relationship between the covariates and the graphical models, as even the adjacent covariate subspaces may differ much. Kolar, Parikh, and Xing (2010) considered a nonparametric approach for conditional covariance estimation with continuous covariates. Cheng et al. (2014) considered a conditional Ising model for binary data where the log-odds is modeled as a linear function of external covari-

CONTACT Jingfei Zhang ezhang@bus.miami.edu Department of Management Science, University of Miami, Coral Gables, FL. Yi Li yili@umich.edu Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.
© 2022 American Statistical Association

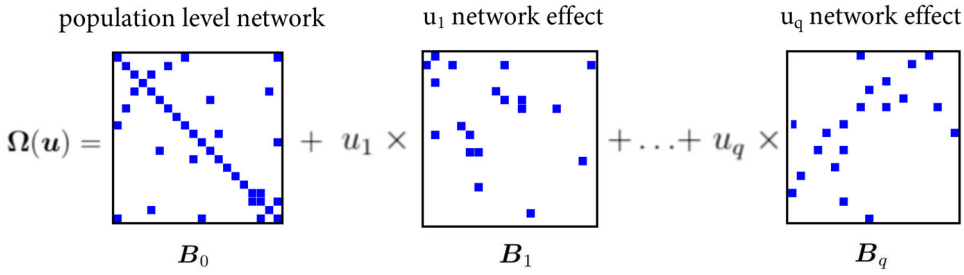

Figure 1. An illustration of the subject-specific Gaussian graphical model.

ates. Ni, Stingo, and Baladandayuthapani (2019) considered a conditional DAG model that allows the graph structure to vary with a finite number of discrete or continuous covariates, and assumed a known hierarchical ordering of the nodes. Such preknowledge may not always be available in practical settings.

We propose a Gaussian graphical regression model that allows the network structure to vary with external covariates (discrete and continuous) of high dimensions. Specifically, both the mean and the precision matrix are modeled as functions of covariates, enabling estimation of subject-specific graphical models; see Figure 1. To facilitate estimation, we show that our proposed model can be formulated as a sequence of linear regression models that include the interactions between response variables (e.g., gene expressions) and external covariates (e.g., genetic variants); Section 2.2. Our model accommodates the setting where both response variables and external covariates are high dimensional, which is frequently encountered in genetic studies, and includes the existing conditional mean Gaussian graphical model (e.g., Yin and Li 2011) as a special case. To estimate coefficients in the covariate-dependent precision matrix, we impose a sparse group lasso penalty that encourages effective covariates to be sparse and their effects on edges to be sparse as well.

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

Our work contributes to both methodology and theory. As to *methodology*, we propose a flexible subject-specific graphical model that depends on a large number of external covariates. We employ a combined sparsity structure that encourages effective covariates and the effect of effective covariates on the network to be simultaneously sparse. With respect to *theory*, we carry out a thorough investigation of the simultaneously sparse estimator, by deriving tight nonasymptotic estimation error bounds and establishing variable selection consistency. Our work addresses the theoretical challenges arising from regressing both the means and the precision matrices on external covariates. Moreover, as the simultaneously sparse regularizer is nondecomposable, the existing techniques using decomposable regularizers and null space properties (Negahban et al. 2012) are not applicable; see Section 4. Thus, our techniques may advance high-dimensional regression with simultaneously sparse structures. Finally, though motivated by a biological application, our method provides a general regression framework of associating networks with external covariates and is broadly applicable to other scientific fields that involve networks.

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

2. Graphical Regression Models

2.1. Notation and Preamble

We start with some notation. Given a vector $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we use $\|\boldsymbol{x}\|_0$, $\|\boldsymbol{x}\|_1$, $\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_\infty$ to denote the ℓ_0, ℓ_1, ℓ_2 and ℓ_∞ norms, respectively, and use $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle$ to denote the inner product of $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$. We write $[d] = \{1, 2, \ldots, d\}$. Given an index set $\mathcal{S} \in [d]$, we use $\boldsymbol{x}_\mathcal{S} \in \mathbb{R}^{|\mathcal{S}|}$ to denote the sub-vector of $\boldsymbol{x}$ corresponding to index $\mathcal{S}$. For a matrix $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$, we let $\|\boldsymbol{X}\|$ and $\|\boldsymbol{X}\|_{\max} = \max_{ij} X_{ij}$ denote the spectral norm and element-wise max norm, respectively. Given $\mathcal{S} \in [d_2]$, we use $\boldsymbol{X}_\mathcal{S} \in \mathbb{R}^{d_1 \times |\mathcal{S}|}$ to denote the sub-matrix with columns indexed in $\mathcal{S}$. We use $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ to denote the smallest and largest eigenvalues of a matrix, respectively. For two positive sequences a_n and b_n, write $a_n \precsim b_n$ or $a_n = \mathcal{O}(b_n)$ if there exist $c > 0$ and $N > 0$ such that $a_n < cb_n$ for all $n > N$, and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$; write $a_n \asymp b_n$ if $a_n \precsim b_n$ and $b_n \precsim a_n$.

Suppose $\boldsymbol{X} = (X_1, \ldots, X_p) \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. Denote the precision matrix $\boldsymbol{\Sigma}^{-1}$ by $(\sigma^{ij})_{p \times p}$. Under a Gaussian distribution,

$\sigma^{ij} \neq 0$ is equivalent to X_i and X_j being conditionally dependent given all other X variables (Lauritzen 1996). Let $\boldsymbol{X}_{-j} = \{X_k : k \in [p], k \neq j\}$. Meinshausen and Bühlmann (2006) and Peng et al. (2009) related $(\sigma^{ij})_{p\times p}$ to the coefficients in this linear regression model:

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

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

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

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

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

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

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

2.2. High-Dimensional Gaussian Graphical Regression with Covariates

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

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

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

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

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

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

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

where $\beta_{jkh} = -[\boldsymbol{B}_h]_{jk}/\sigma^{jj}$ and $\text{var}(\epsilon_j) = 1/\sigma^{jj}$, for all j, k and h. Model (5) provides a regression framework for estimating the mean and precision parameters in (4), by adding to (1) or (3) the interactions between $\boldsymbol{X}_{-j}$ and $\boldsymbol{u}$. Correspondingly, the partial correlation between X_j and X_k, conditional on all other X variables, is modeled as a function of $\boldsymbol{u}$, forming the basis of Gaussian graphical regression. The diagonal elements of $\boldsymbol{\Omega}(\boldsymbol{u})$ (i.e., σ^{jj}'s) are connected to the residual variances in (5), that is, $\text{var}(\epsilon_j) = 1/\sigma^{jj}$. From this perspective, assuming σ^{jj} to be free of $\boldsymbol{u}$ may be viewed as assuming the residual variance of Z_j, after removing effects of $\boldsymbol{u}$, $\boldsymbol{Z}_{-j}$ and the interactions between $\boldsymbol{u}$ and $\boldsymbol{Z}_{-j}$, to not dependent on $\boldsymbol{u}$, which is plausible in the context of regression. However, as (3) is a regression-type representation of the precision matrix, caution must be exercised when comparing the residual terms in (3) to the error terms in a standard regression problem; see more discussions in Section 8. Obviously, model (5) includes models (1) and (3) as special cases with $\beta_{jkh} = 0$ for all j, k and h.

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

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

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

$$\boldsymbol{\beta}_j = \big(\underbrace{\beta_{j10}, \ldots, \beta_{jp0}}_{\substack{\text{group 0} \\ \boldsymbol{b}_{j0}\text{: population level} \\ \text{edges of node } j}}, \underbrace{\beta_{j11}, \ldots, \beta_{jp1}}_{\substack{\text{group 1} \\ \boldsymbol{b}_{j1}: u_1\text{'s effect on} \\ \text{edges of node } j}}, \ldots, \underbrace{\beta_{j1q}, \ldots, \beta_{jpq}}_{\substack{\text{group q} \\ \boldsymbol{b}_{jq}: u_q\text{'s effect on} \\ \text{edges of node } j}} \big)^\top. \tag{7}$$

When both p and q are large, to ensure the estimability of $\boldsymbol{\beta}_j$, we impose on it simultaneous group sparsity and element-wise sparsity. With groups illustrated in (7), we assume $\boldsymbol{\beta}_j$ is *group sparse*, stipulating that effective covariates are sparse, that is, only a few covariates may impact edges and those impactful covariates are termed effective covariates. We further assume $\boldsymbol{\beta}_j$ is *element-wise sparse*. That is, effective covariates may influence only a few edges. These simultaneous sparsity assumptions are well supported by genetic studies (van der Wijst et al. 2018a). We exclude $\boldsymbol{b}_{j0}$ from the group sparsity constraint (but not the element-wise sparsity constraint), as it determines the population level regulatory network. With covariate $\boldsymbol{u}$ and sparsity on

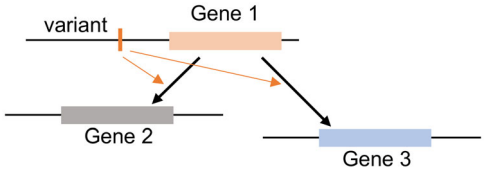

Figure 2. An illustration of gene co-expressions: the genetic variant is a trans-eQTL modulating co-expressions of pairs (1,2) and (1,3), where Gene 1 is an upstream gene and Genes 2 and 3 are downstream genes.

$\boldsymbol{\beta}_j$'s, it is possible that (X_j, X_k) and (X_k, X_s) are conditionally dependent, while (X_j, X_s) are conditionally independent. This type of structures is biological plausible. Consider as an example our motivating data application in co-expression QTL identification. It is possible for genetic variants located near a gene (say, Gene 1), called the trans-acting expression quantitative trait loci (trans-eQTLs) (Fehrmann et al. 2011), to alter how intensively Gene 1 may regulate (e.g., activate, inhibit) two downstream Genes 2 and 3, while not altering the coexpression between Genes 2 and 3. In fact, the two downstream Genes 2 and 3 can be independent conditional on the rest of the gene network, regardless of what the upstream trans-eQTLs might be (Brynedal et al. 2017; Gong et al. 2018; Kolberg et al. 2020); see Figure 2 for an illustration.

Model (6) can be viewed as an interaction model. Our later development does not abide by the common hierarchical principle for the inclusion of interactions, that is, an interaction is allowed only if the main effects are present (Hao, Feng, and Zhang 2018; She, Wang, and Jiang 2018). This is because gene co-expressions may occur only for certain genetic variations (Wang et al. 2013; van der Wijst et al. 2018a), in which case, β_{jkh} (i.e., effect of u_h on edge (j,k)) can be nonzero while β_{jk0} is zero (i.e., population level edge (j,k)). Section 8 discusses modifications of our proposal if hierarchy is to be enforced.

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

3. Estimation

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

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

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

Denote the squared error loss function by

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

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

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

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

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

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

Symmetrization can also be achieved via

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

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

Two parameters λ and λ_g in (9) require tuning; in our procedure, they are jointly selected via L-fold cross-validation. As in Simon et al. (2013) and Cai, Zhang, and Zhou (2019), we rewrite $\lambda = \alpha\lambda_0$ and $\lambda_g = (1-\alpha)\lambda_0$, where α reflects the weight of the lasso penalty relative to the group lasso penalty and λ_0 reflects the total amount of regularization. We assess a set of values for $\alpha \in [0,1]$, with $\alpha = 0$ and 1 corresponding to lasso and group lasso, respectively; for each α, a sequence of λ_0 values are considered to obtain the whole regularization path (Vincent and Hansen 2014). Finally, we choose the combination of (α, λ_0) that minimizes the cross-validation error. In our implementations, we consider $\alpha \in \{0, 0.1, 0.2, \ldots, 0.9, 1\}$ and $L = 5$, and note that the result is fairly robust to the choices of α (see Section 6).

4. Theoretical Properties

In this section, we derive the nonasymptotic ℓ_2 convergence rate of the sparse group lasso estimator from (9) and establish variable selection consistency. Our theoretical investigation is challenged by several unique aspects of the model. First, as the design matrix $\boldsymbol{W}_{-j} \in \mathbb{R}^{n\times(p-1)(q+1)}$ includes high-dimensional interaction terms between $\boldsymbol{z}^{(i)}$ and $\boldsymbol{u}^{(i)}$, and the variance of $\boldsymbol{z}^{(i)}$ is a function of $\boldsymbol{u}^{(i)}$, characterizing the joint distribution of each row in $\boldsymbol{W}_{-j}$ is difficult and requires a delicate treatment.

Second, as the combined penalty term $\lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}$ is not decomposable, the classic techniques for decomposable regularizers and null space properties (Negahban et al. 2012) are not applicable. Standard treatments of the stochastic term (Bickel, Ritov, and Tsybakov 2009; Lounici et al. 2011; Negahban et al. 2012) such as $\langle \boldsymbol{\epsilon}, \boldsymbol{W}_{-j}\Delta\rangle \le \|\boldsymbol{W}_{-j}^\top\boldsymbol{\epsilon}\|_\infty\|\Delta\|_1$, where $\Delta \in \mathbb{R}^{(p-1)(q+1)}$, can only yield an ℓ_2 convergence rate comparable to that from the lasso or the group lasso. Utilizing the statistical properties and the computational optimality of the sparse group lasso estimator in (9), we derive two interrelated bounds on the stochastic term. The first bound characterizes the cardinality measure of the covariate space, while the second one utilizes the Karush–Kuhn–Tucker condition and properties of the combined regularizer. Combining these bounds, we give a sharp upper bound on the stochastic term, and show our proposed estimator possesses an improved ℓ_2 error bounds compared to the lasso and the group lasso when the true coefficients are simultaneously sparse; see Section S3.1, supplementary materials.

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

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

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

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

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

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

Let s_λ denote the number of nonzero entries in a candidate model such that $s_j < s_\lambda \le n$. Given an s_λ satisfying the conditions in Theorem 1, we choose $\lambda_{\max}$ and $\lambda_{\min}$ to be the upper and lower limits of λ_0 for each α, respectively, corresponding to an empty model with no variables selected and a sparse model with s_λ variables selected.

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

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

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

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

where C, C_1, and C_2 are positive constants.

Theorem 1 shows that our proposed estimator enjoys an improved ℓ_2 error bound over both the lasso and the group lasso under simultaneous sparsity. Specifically, given that the dimension of $\boldsymbol{\beta}_j$ is $(p-1)(q+1)$ and $s_{j,g} \le s_j$, applying the regular lasso regularizer $\lambda\|\boldsymbol{\beta}_j\|_1$ alone would yield an error bound of $(s_j/n)\log(pq)$ (Negahban et al. 2012), which is slower than that in (13) when $\log p/\log q = o(1)$ and $s_{j,g}/s_j = o(1)$, corresponding to group sparsity. Moreover, when $p > n + 1$, estimating with the group lasso regularizer $\lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}$ alone, which excludes $(\boldsymbol{\beta}_j)_{(0)}$, is not feasible, because the dimension of the latter (i.e., $p-1$) exceeds n. If we utilize a group lasso regularizer $\lambda_g\|\boldsymbol{\beta}_j\|_{1,2}$ that includes $(\boldsymbol{\beta}_j)_{(0)}$, the estimator would have an ℓ_2 error bound of $(s_{j,g}/n)\log q + (s_{j,g}/n)p$ (Lounici et al. 2011), which is slower than that in (13) when $\log q/p = o(1)$ and $s_j/s_{j,g} = o(p/\log p)$, corresponding to within-group sparsity. While the optimality of these error bounds warrants further investigation, the combined regularizer $\lambda\|\boldsymbol{\beta}_j\|_1 + \lambda_g\|\boldsymbol{\beta}_{j,-0}\|_{1,2}$ may improve upon both the regular lasso and group lasso regularizers, when the true underlying coefficients are both element-wise and group sparse. In Theorem 1, the condition $s_\lambda(\log p + \log q) = \mathcal{O}(\sqrt{n})$ upper bounds the size of candidate models, which in turn helps to bound $\langle \boldsymbol{\epsilon}, \boldsymbol{W}\Delta\rangle$. The parameter s_λ can be set to $c\sqrt{n}/\max\{\log p, \log q\})$ for some $c > 0$; by Assumption 3, it follows that $s_j = o(s_\lambda)$.

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

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

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

the covariance matrix $\mathbf{\Sigma}_{\boldsymbol{W}}$ satisfies that

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

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

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

Specifically, by the law of total probability, we write

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

Hence, Assumption 4 holds when the correlations among $\{Z_j\}_{j\in[p]}$ and among $\{U_h\}_{h\in[q]}$ are not too large. A trivial sufficient condition is $\mathrm{cov}(\boldsymbol{u}^{(1)}) = \boldsymbol{I}$. Furthermore, if $s_j = \mathcal{O}(1)$, Assumption 4 is satisfied when $\max_{l_1\neq l_2}[\mathrm{cov}(\boldsymbol{z}^{(1)})]_{l_1,l_2}\times\max_{l_3\neq l_4}[\mathrm{cov}(\boldsymbol{u}^{(1)})]_{l_3,l_4}$ is less than some positive constant. Similar correlation conditions include the neighborhood stability condition (Meinshausen and Bühlmann 2006) and the irrepresentability condition (Zhao and Yu 2006); see Van De Geer and Bühlmann (2009) for a discussion of these relationships.

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

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

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

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

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

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

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

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

where $\mathcal{P}_{\hat{\mathcal{S}}_j}$ is the projection matrix onto the column space of $\boldsymbol{W}_{\hat{\mathcal{S}}_j}$. The estimator in (16) can alternatively be written as $\hat{\sigma}_{\epsilon_j}^2 = \frac{1}{n-\hat{s}_j}(1-\gamma_n^2)\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$, where $\gamma_n^2 = \boldsymbol{\epsilon}^\top\mathcal{P}_{\hat{\mathcal{S}}_j}\boldsymbol{\epsilon}/\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$ represents the fraction of bias in $\hat{\sigma}_{\epsilon_j}^2$. Under conditions in Theorem 2 and using a result in (S10), we get $\gamma_n^2 \asymp \sigma_{\epsilon_j}^2\left\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\right\}/n$. Therefore, $\hat{\sigma}_{\epsilon_j}^2$ is consistent, provided that $\sigma_{\epsilon_j}^2\left\{s_j\log(ep) + s_{j,g}\log(eq/s_{j,g})\right\}/n \to 0$.

5. Estimation with Unknown $\mathbf{\Gamma}$: A Two-Step Procedure

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

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

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

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

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

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

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

Step 1 poses a regular lasso penalty on $\mathbf{\Gamma}$, as commonly done in the covariate-adjusted Gaussian graphical model literature (Rothman, Levina, and Zhu 2010; Cai et al. 2012; Yin and Li 2013; Chen et al. 2016). Note that each regression in Step 1 is of dimension q. When q is large, it may be necessary to consider a sparse group penalty (as in Step 2) that encourages $\mathbf{\Gamma}$ to be both element-wise and group sparse.

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

The theoretical development for the two-step procedure is challenging. Step 1 involves a regularized regression $\boldsymbol{x}_j = \boldsymbol{H}\boldsymbol{\gamma}_j + \boldsymbol{z}_j$ with heteroscedastic errors as the variance of $z_j^{(i)}$ is a function of $\boldsymbol{u}^{(i)}$. In Step 2, both the response vector $\hat{\boldsymbol{z}}_j$ and the design

matrix $\hat{\boldsymbol{W}}$ inherit approximation errors from $\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}$, which further complicates the analysis of the sparse group lasso estimator. We show that $\hat{\boldsymbol{\beta}}_j$ in (18) can achieve the same convergence rate as that in Theorem 1 (i.e., the noiseless case), and thus, enjoys the oracle property (as if $\boldsymbol{\Gamma}$ were known).

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

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

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

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

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

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

for some positive constants C_3, C_4.

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

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

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

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

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

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

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

6. Simulations

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

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

and a group lasso estimator

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

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

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

The population level network is assumed to follow a scale-free network model, with the degrees of nodes generated from a power-law distribution (Clauset, Shalizi, and Newman 2009) with parameter 2.5. We randomly select q_e out of q covariates to have nonzero effects, and the graphs for these q_e covariates follow an Erdos–Renyi model with edge probability v_e; see the graph structures in Figure 3. We set $\sigma^{jj} = 1$ for $j \in [p]$. The initial nonzero coefficients β_{jkh} are generated from Uniform$([-0.5, -0.35] \cup [0.35, 0.5])$. For each j, we rescale $\{\beta_{jkh}\}_{k\neq j\in[p],h\in\{0\}\cup[q]}$ by dividing each entry by $\sum_{k\neq j\in[p],h\in\{0\}\cup[q]}|\beta_{jkh}|$. After rescaling, for each j, k and h, we use the average of β_{jkh} and β_{kjh} to fill the entries at jkh and kjh. This process results in symmetry with diagonal dominance and, thus, ensures the positive definiteness of the precision matrices. We set $s_{\boldsymbol{\Gamma}} = 125$, $q_e = 5$, $v_e = 0.01$, and consider $n = 200, 400$, $p = 25, 50$ and $q = 50, 100$, with 1224 to 4949 parameters to estimate.

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

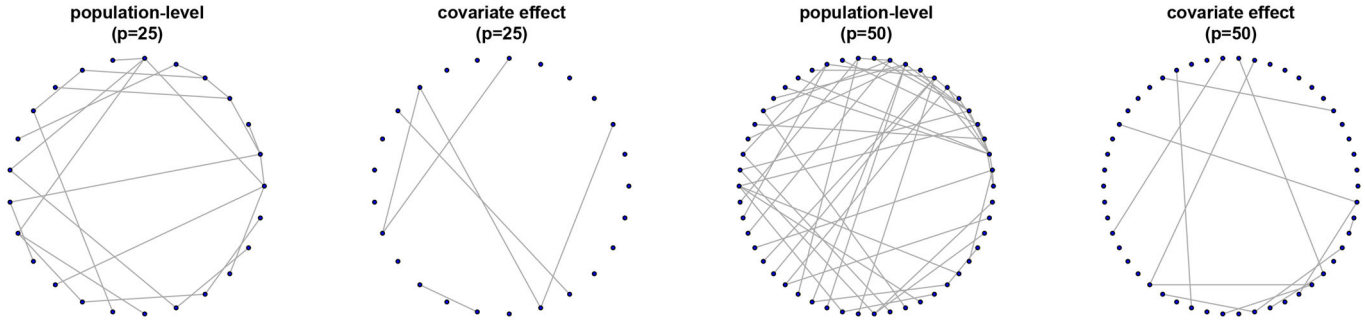

Figure 3. Graphs corresponding to the population-level effects and covariate effects with $p = 25$ or 50. When illustrating covariate effects, we randomly pick only one of the q_e effective covariates.

Table 1. Estimation accuracy of $\boldsymbol{\beta}_j$'s in simulations with varying sample size n, network size p and covariate dimension q.

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

NOTE: The three methods are RegGMM, the lasso estimator in (23) and the group lasso estimator in (24). Marked in boldface are those achieving the best evaluation criteria in each setting.

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

Table 1 reports the average criteria for estimating $\boldsymbol{\beta}_j$'s, with standard errors in the parentheses, over 200 data replications. It shows that the proposed RegGMM outperforms the competing methods in both estimation accuracy and selection accuracy for different sample sizes n, network sizes p and covariate dimensions q. This is consistent with our theoretical findings. Moreover, the estimation errors of RegGMM decrease as n increases, or as p and q decrease, confirming the theoretical results in Theorem 3. For additional analyses, we evaluate some higher dimensional cases by increasing p, q to 300 or 400, and present the results in Section S1.3, supplementary materials. In Section S1.4, we compare several benchmark solutions, including the standard Gaussian graphical model estimated using the neighborhood selection method (Meinshausen and Bühlmann 2006) and the graphical lasso estimation method (Friedman, Hastie, and Tibshirani 2008), the conditional mean Gaussian graphical model (Cai et al. 2012) and the stratified Gaussian graphical model (Danaher, Wang, and Witten 2014).

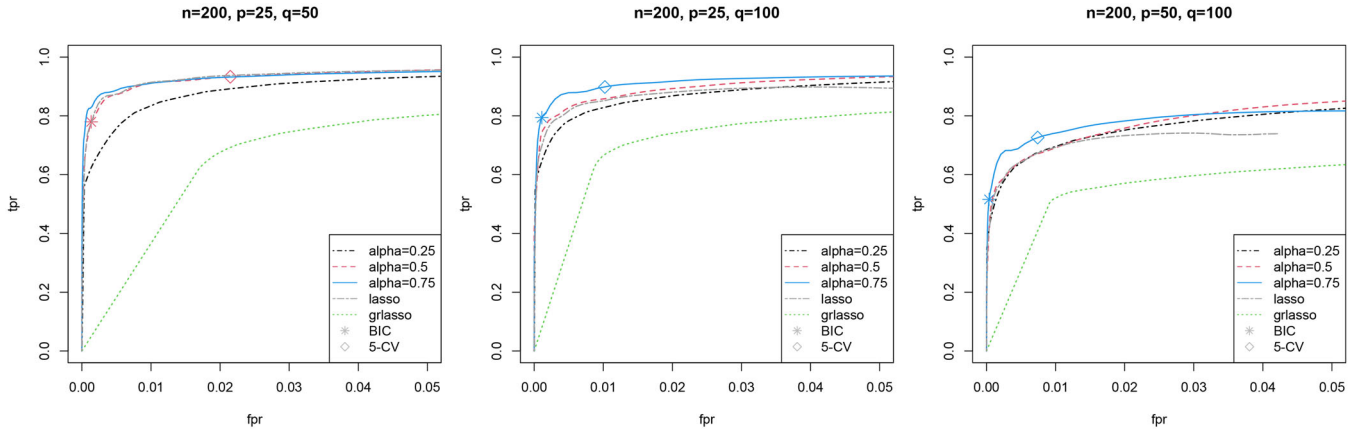

Figure 4. The ROC curves for RegGMM under $\alpha = 0.25, 0.5, 0.75$, lasso and the group lasso (grlasso). For RegGMM, the values selected by BIC and 5-fold cross-validation are marked on the curves.

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

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

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

Finally, we investigate the computation cost that may occur during the tuning process. Table 2 shows the 5-fold cross-validation computation time for one node (or gene) and all p nodes for a given α. The simulations were run on an iMac with a 3.6 GHz Intel Core i9 processor. As the number of parameters is $\mathcal{O}(p^2 q)$, the total computing cost is expected to be roughly quadratic in p and linear in q, as seen in Table 2. Our method enables a parallel implementation over the p node-wise regressions and the working values of α, in which case the computing time on each core is, for example, 16 sec when $p = 50$ and $q = 100$.

7. Co-expression QTL Analysis

Our application focuses on glioblastoma multiforme (GBM), the most aggressive and fatal subtype of brain cancer (Bleeker, Molenaar, and Leenstra 2012), as featured in the REMBRANDT trial (GSE108476) with a subcohort of $n = 178$ GBM patients. Since existing therapies remain largely ineffective (Bleeker, Molenaar, and Leenstra 2012), it is imperative to explore more effective treatment, such as new gene therapies (Kwiatkowska et al. 2013). Understanding the molecular underpinning of the disease is the key. In the study, all of these 178 patients had undergone microarray and single nucleotide polymorphism (SNP) chip profiling, with both gene expression and SNP data available for analysis. Specifically, the extracted RNA from each tumor sample was processed using microarrays with 23,521 genes assayed on each array. Genomic DNA from each sample was hybridized to SNP chips, which covers 116,204 SNP loci with a mean intermarker distance of 23.6kb. The raw data were preprocessed and normalized using standard pipelines; see Gusev et al. (2018) for more details.

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

We have evaluated several benchmark methods in Section S1.4 of the supplementary materials; however, these methods are not designed to and cannot detect eQTL variants. Therefore, we have elected to apply the proposed two-step procedure in Section 5 to this dataset. It is common in penalized regressions to standardize predictors to ensure they be on the same scale (Tibshirani 1997). For example, the covariates in the model are standardized to have mean zero and variance one. The scheme does not alter interpretations of the model; see discussions in Section S1.5, supplementary materials. Tuning parameters in both steps of the estimation procedure are selected via 5-fold cross-validation, and post-processing, as in (10), generates the

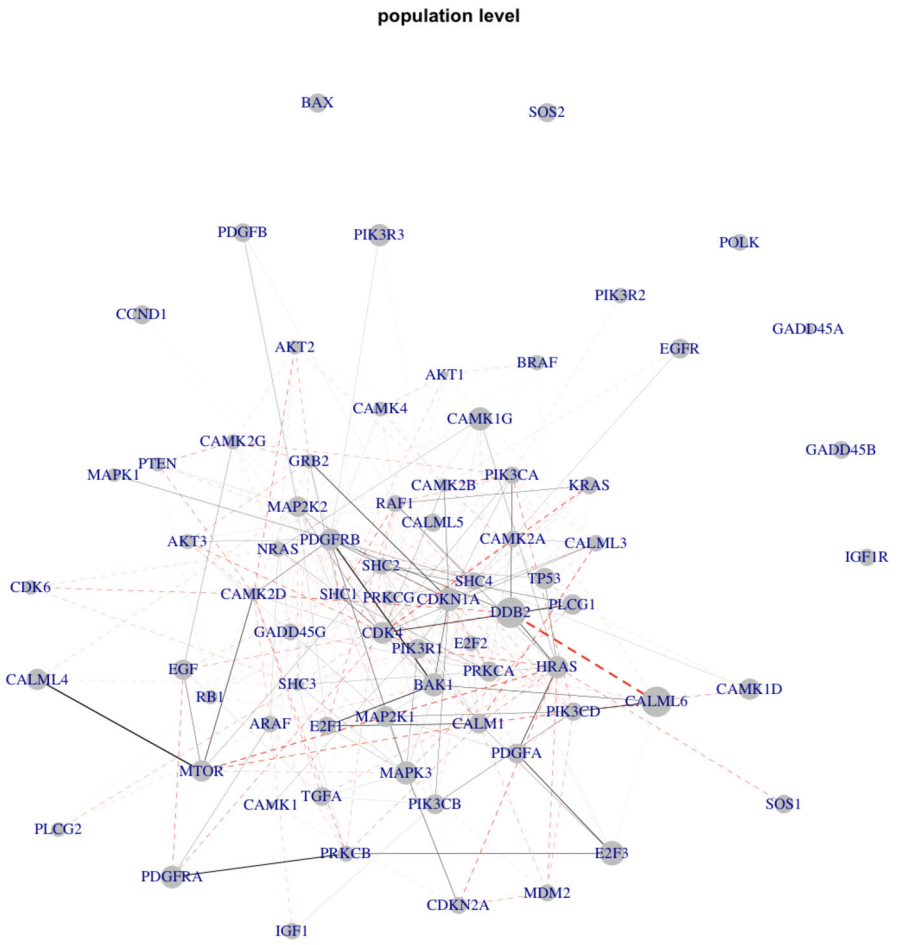

Figure 5. The graph corresponding to the population-level effect. The node sizes are proportional to mean expression levels and the edge weights are proportional to $\hat{B}_0$. Edges with positive (negative) effects on partial correlations are shown in red dashed (black solid) lines.

Table 3. Pathways and genes involves in each pathway.

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

final estimates. Out of the 120 covariates considered, nine SNPs are estimated to have nonzero effects on the network.

We first examine the population level network. Most of the well-connected genes in Figure 5 are known to be associated with cancer. For example, PIK3CA is a protein coding gene and is one of the most highly mutated oncogenes identified in human cancers (Samuels and Velculescu 2004); mutations in the PIK3CA gene are found in many types of cancer, including cancer of the brain, breast, ovary, lung, colon, and stomach (Samuels and Velculescu 2004). The PIK3CA gene is a part of the PI3K/AKT/MTOR signaling pathway, which is one of the core pathways in human GBM and other types of cancer (The Cancer Genome Atlas Research Network 2008). TP53 is also a highly connected gene in the estimated network. This gene encodes a tumor suppressor protein containing transcriptional activation, and is the most frequently mutated gene in human cancer; the P53 signaling pathway is also one of the core pathways in human GBM and other types of cancer (The Cancer

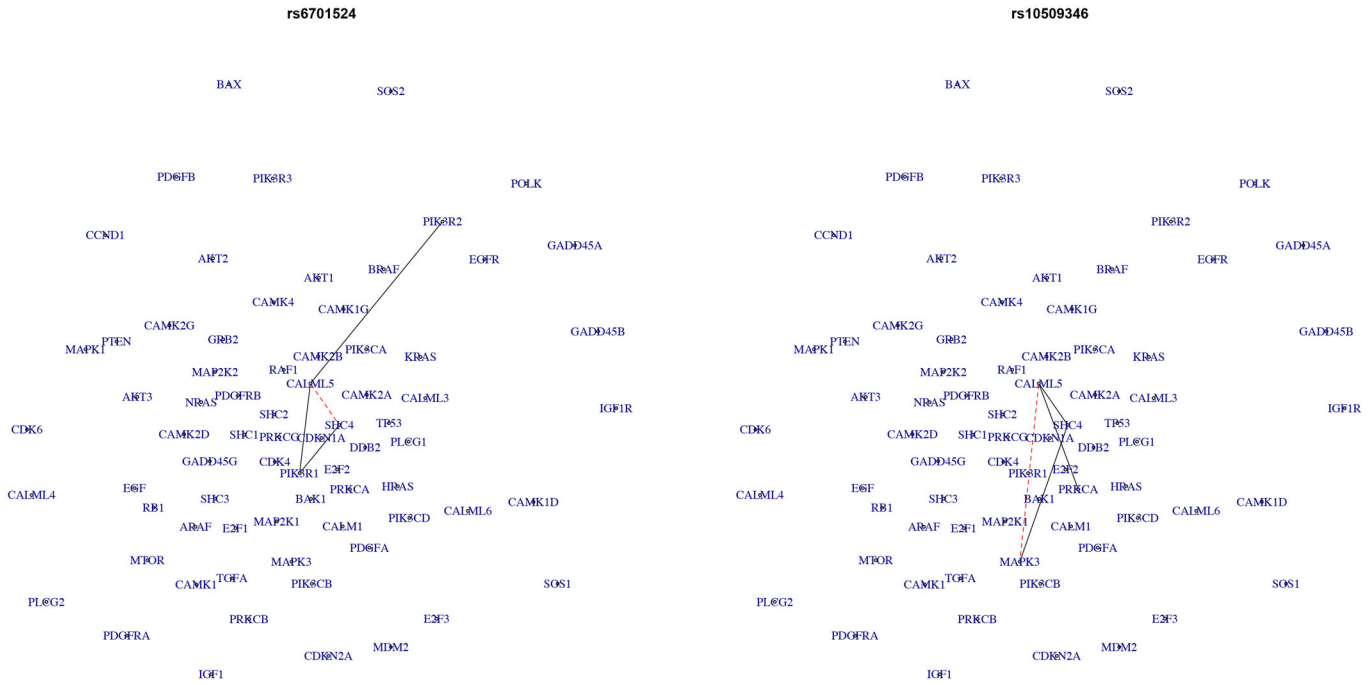

Figure 6. Graphs depending on each covariate (i.e., different SNPs). Edges that have positive (negative) effects on partial correlations are shown in red dashed (black solid) lines.

Genome Atlas Research Network 2008). In Figure 5, we can identify several core pathways in human GBM including the PI3K/ AKT/MTOR, Ras-Raf-MEK-ERK, calcium and p53 signaling pathways; see Table 3 for genes included in each pathway. These findings are in agreement to the existing literature on GBM genes and pathways (The Cancer Genome Atlas Research Network 2008; Brennan et al. 2013; Maklad, Sharma, and Azimi 2019).

We next examine the covariate effects on the network. Identified are nine co-expression QTLs, namely, `rs6701524`, `rs10509346`, `rs10492975`, `rs723211`, `rs1347069`, `rs473698`, `rs4118334`, `rs882664`, and `rs1267622`. The network effects of `rs6701524` are shown in Figure 6 (left panel). This SNP, residing in MTOR, is found to affect CALML5's co-expression with PIK3R1, and also with PIK3R2 and SHC4. This is an interesting finding as PI3K/MTOR is a key pathway in GBM development and progression, and inhibition of PI3K/MTOR signaling was found effective in increasing survival with GBM tumor (Batsios et al. 2019). This co-expression QTL can potentially regulate the co-expressions of CALML5, PIK3R1, PIK3R2, MTOR, and play an important role in activating the PI3K/MTOR pathway.

Shown in Figure 6 (right panel) are the network effects of `rs10509346`, a variant of CAMK2G. The figure indicates that this SNP affects the co-expressions of CAMK2G with genes in the Ras-Raf-MEK-ERK pathway. This agrees to the finding that the Ras-Raf-MEK-ERK pathway is modulated by Ca^{+2} and calmodulin (Agell et al. 2002). Moreover, based on our analysis, `rs10492975` regulates the co-expressions of CALML5, PIK3R2 and CAMK1; `rs723211` is associated with the co-expressions of CALML5 and other genes; `rs1347069` influences the co-expressions of SHC4 and CDKN2A; `rs473698` may modify the co-expressions of PRKCG and CAMK1; `rs4118334` modulates the co-expressions of SHC2 and CAMK1; `rs882664` influences the co-expressions of PRKCA and CAMK1; `rs1267622` may alter the co-expressions of SHC3 and RAF1. See details in Table S6. As co-expression QTL identification has sparked recent interest, these findings warrant more in-depth investigation.

8. Discussion

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

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

the hierarchy between main effects and interaction terms by reorganizing $\boldsymbol{\beta}_j$ as

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

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

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

$$\begin{aligned} Z_j \times g(\boldsymbol{v}_j^\top \boldsymbol{u}) &= \sum_{k\neq j}^{p} \theta_{jk0} Z_k + \sum_{k\neq j}^{p}\sum_{h=1}^{q} \theta_{jkh} u_h Z_k + \tilde{\epsilon}_j, \\ \text{var}(\tilde{\epsilon}_j) &= g(\boldsymbol{v}_j^\top \boldsymbol{u}) \end{aligned} \tag{26}$$

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

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

where $\mathcal{D}$ denotes the observed data, $\boldsymbol{z}_j^{(i)}$ collects the samples of the jth variable and $\boldsymbol{w}_i$ is the ith row of the design matrix $\boldsymbol{W}_{-j}$; see their definitions in Section 3. Through $\ell_j(\boldsymbol{v}_j, \boldsymbol{\theta}_j|\mathcal{D})$, $\boldsymbol{v}_j$ and $\boldsymbol{\theta}_j$ can be estimated with sparse penalties. This new iterative estimation procedure is more computationally demanding and requires a new theoretical analysis. We leave its full investigation as future research.

Supplementary Materials

The supplementary materials collect all technical proofs and additional numerical details and results.

Acknowledgments

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

Funding

The work is partially supported by grants from NIH and NSF.

References

Agell, N., Bachs, O., Rocamora, N., and Villalonga, P. (2002), "Modulation of the Ras/Raf/MEK/ERK Pathway by Ca2+, and Calmodulin," *Cellular Signalling*, 14, 649–654. [11]

Batsios, G., Viswanath, P., Subramani, E., Najac, C., Gillespie, A. M., Santos, R. D., Molloy, A. R., Pieper, R. O., and Ronen, S. M. (2019), "PI3K/mTOR Inhibition of IDH1 Mutant Glioma Leads to Reduced 2HG Production that is Associated with Increased Survival," *Scientific Reports*, 9, 1–15. [11]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [5]

Bleeker, F. E., Molenaar, R. J., and Leenstra, S. (2012), "Recent Advances in the Molecular Understanding of Glioblastoma," *Journal of Neuro-oncology*, 108, 11–27. [9]

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. K., Zhang, W., Sougnez, C., Mikkelsen, T., Aldape, K., Bigner, D. D., Van Meir, E. G., Prados, M., Sloan, A., Black, K. L., Eschbacher, J., Finocchiaro, G., Friedman, W., Andrews, D. W., Guha, A., Iacocca, M., O'Neill, B. P., Foltz, G., Myers, J., Weisenberger, D. J., Penny, R., Kucherlapati, R., Perou, C. M., Neil Hayes, D., Gibbs, R., Marra, M., Mills, G. B., Lander, E., Spellman, P., Wilson, R., Sander, C., Weinstein, J., Meyerson, M., Gabriel, S., Laird, P. W., Haussler, D., Getz, G., Chin, L., and TCGA Research Network. (2013), "The Somatic Genomic Landscape of Glioblastoma," *Cell*, 155, 462–477. [10,11]

Brynedal, B., Choi, J., Raj, T., Bjornson, R., Stranger, B. E., Neale, B. M., Voight, B. F., and Cotsapas, C. (2017), "Large-Scale Trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation," *The American Journal of Human Genetics*, 100, 581–591. [4]

Cai, T. T., Li, H., Liu, W., and Xie, J. (2012), "Covariate-Adjusted Precision Matrix Estimation with an Application in Genetical Genomics," *Biometrika*, 100, 139–156. [1,3,6,7,8]

Cai, T. T., Zhang, A., and Zhou, Y. (2019), "Sparse Group Lasso: Optimal Sample Complexity, Convergence Rate, and Statistical Inference," arXiv preprint arXiv:1909.09851. [2,4,5]

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. [1,3,5,6,7]

Cheng, J., Levina, E., Wang, P., and Zhu, J. (2014), "A Sparse Ising Model with Covariates," *Biometrics*, 70, 943–953. [1,4]

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009), "Power-law Distributions in Empirical Data," *SIAM Review*, 51, 661–703. [7]

Danaher, P., Wang, P., and Witten, D. M. (2014), "The Joint Graphical Lasso for Inverse Covariance Estimation Across Multiple Classes," *Journal of the Royal Statistical Society*, Series B, 76, 373–397. [1,6,8]

Fan, J., Feng, Y., and Wu, Y. (2009), "Network Exploration via the Adaptive LASSO and SCAD Penalties," *The Annals of Applied Statistics*, 3, 521–541. [1]

Fehrmann, R. S., Jansen, R. C., Veldink, J. H., Westra, H.-J., Arends, D., Bonder, M. J., Fu, J., Deelen, P., Groen, H. J., Smolonska, A., 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. [4]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 9, 432–441. [1,8]

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., Feng, J., Liu, R., Diao, L., Guo, A.-Y., Miao, X., and Han, L. (2018), "PancanQTL: Systematic

Identification of Cis-eQTLs and Trans-eQTLs in 33 Cancer Types," *Nucleic Acids Research*, 46, D971–D976. [4]

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011), "Joint Estimation of Multiple Graphical Models," *Biometrika*, 98, 1–15. [1]

Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J.-C., Fine, H., and Madhavan, S. (2018), "The REMBRANDT Study, a Large Collection of Genomic Data from Brain Cancer Patients," *Scientific Data*, 5, 180158. [9]

Hao, N., Feng, Y., and Zhang, H. H. (2018), "Model Selection for High-Dimensional Quadratic Regression via Regularization," *Journal of the American Statistical Association*, 113, 615–625. [4,5]

Kanehisa, M., and Goto, S. (2000), "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, 28, 27–30. [9]

Kolar, M., Parikh, A. P., and Xing, E. P. (2010), "On Sparse Nonparametric Conditional Covariance Selection," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 559–566. [1]

Kolberg, L., Kerimov, N., Peterson, H., and Alasoo, K. (2020), "Co-expression Analysis Reveals Interpretable Gene Modules Controlled by Trans-Acting Genetic Variants," *Elife*, 9, e58705. [4]

Kwiatkowska, A., Nandhu, M. S., Behera, P., Chiocca, E. A., and Viapiano, M. S. (2013), "Strategies in Gene Therapy for Glioblastoma," *Cancers*, 5, 1271–1305. [9]

Lauritzen, S. L. (1996), *Graphical Models* (Vol. 17), Oxford: Clarendon Press. [3]

Lee, W., and Liu, Y. (2012), "Simultaneous Multiple Response Regression and Inverse Covariance Matrix Estimation via Penalized Gaussian Maximum Likelihood," *Journal of Multivariate Analysis*, 111, 241–255. [1,3]

Li, B., Chun, H., and Zhao, H. (2012), "Sparse Estimation of Conditional Graphical Models with Application to Gene Networks," *Journal of the American Statistical Association*, 107, 152–167. [1,3]

Li, B., and Solea, E. (2018), "A Nonparametric Graphical Model for Functional Data with Application to Brain Networks Based on fMRI," *Journal of the American Statistical Association*, 113, 1637–1655. [1]

Li, Y., Nan, B., and Zhu, J. (2015), "Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure," *Biometrics*, 71, 354–363. [4]

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

Liu, H., Chen, X., Wasserman, L., and Lafferty, J. D. (2010), "Graph-Valued Regression," in *Advances in Neural Information Processing Systems*, pp. 1423–1431. [1]

Lounici, K., Pontil, M., Van De Geer, S., and Tsybakov, A. B. (2011), "Oracle Inequalities and Optimal Inference Under Group Sparsity," *The Annals of Statistics*, 39, 2164–2204. [5]

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004), "Genomic Analysis of Regulatory Network Dynamics Reveals Large Topological Changes," *Nature*, 431, 308–312. [1]

Maklad, A., Sharma, A., and Azimi, I. (2019), "Calcium Signaling in Brain Cancers: Roles and Therapeutic Targeting," *Cancers*, 11, 145. [10,11]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462. [1,3,4,6,8]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers," *Statistical Science*, 27, 538–557. [2,5]

Ni, Y., Stingo, F. C., and Baladandayuthapani, V. (2019), "Bayesian Graphical Regression," *Journal of the American Statistical Association*, 114, 184–197. [2]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of the American Statistical Association*, 104, 735–746. [1,3]

Rothman, A. J., Levina, E., and Zhu, J. (2010), "Sparse Multivariate Regression with Covariance Estimation," *Journal of Computational and Graphical Statistics*, 19, 947–962. [1,3,6]

Samuels, Y., and Velculescu, V. E. (2004), "Oncogenic Mutations of PIK3CA in Human Cancers," *Cell Cycle*, 3, 1221–1224. [10]

She, Y., Wang, Z., and Jiang, H. (2018), "Group Regularized Estimation Under Structural Hierarchy," *Journal of the American Statistical Association*, 113, 445–454. [4,12]

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245. [4,6]

Tibshirani, R. (1997), "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, 16, 385–395. [9]

The Cancer Genome Atlas Research Network. (2008), "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways," *Nature*, 455, 1061–1068. [10,11]

Van De Geer, S. A., and Bühlmann, P. (2009), "On the Conditions Used to Prove Oracle Results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392. [6]

van der Wijst, M. G., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., and Franke, L. (2018a), "Single-Cell RNA Sequencing Identifies Celltype-Specific Cis-eQTLs and Co-expression QTLs," *Nature Genetics*, 50, 493–497. [1,3,4]

van der Wijst, M. G., de Vries, D. H., Brugge, H., Westra, H.-J., and Franke, L. (2018b), "An Integrative Approach for Building Personalized Gene Regulatory Networks for Precision Medicine," *Genome Medicine*, 10, 1–15. [1]

Vincent, M., and Hansen, N. R. (2014), "Sparse Group Lasso and High Dimensional Multinomial Classification," *Computational Statistics & Data Analysis*, 71, 771–786. [4,6]

Wang, L., Zheng, W., Zhao, H., and Deng, M. (2013), "Statistical Analysis Reveals Co-expression Patterns of Many Pairs of Genes in Yeast are Jointly Regulated by Interacting Loci," *PLoS Genetics*, 9, e1003414. [1,4]

Wang, Y., Joseph, S. J., Liu, X., Kelley, M., and Rekaya, R. (2012), "SNPxGE2: A Database for Human SNP–Coexpression Associations," *Bioinformatics*, 28, 403–410. [1]

Wu, Y., and Wang, L. (2020), "A Survey of Tuning Parameter Selection for High-Dimensional Regression," *Annual Review of Statistics and Its Application*, 7, 209–226. [3]

Xie, S., Li, X., McColgan, P., Scahill, R. I., Zeng, D., and Wang, Y. (2020), "Identifying Disease-Associated Biomarker Network Features Through Conditional Graphical Model," *Biometrics*, 76, 995–1006. [1]

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. [1,2,3]

——— (2013), "Adjusting for High-Dimensional Covariates in Sparse Precision Matrix Estimation by ℓ_1-Penalization," *Journal of Multivariate Analysis*, 116, 365–381. [6]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [5]

——— (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [1]

Zhang, A., and Xia, D. (2018), "Tensor SVD: Statistical and Computational Limits," *IEEE Transactions on Information Theory*, 64, 7311–7338. [11]

Zhang, J., Sun, W. W., and Li, L. (2019), "Mixed-Effect Time-Varying Network Model and Application in Brain Connectivity Analysis," *Journal of the American Statistical Association*, 115, 2022–2036. [1]

Zhang, T. (2009), "Some Sharp Performance Bounds for Least Squares Regression with l1 Regularization," *The Annals of Statistics*, 37, 2109–2144. [6]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [6]