Received: 20 August 2020 | Revised: 16 December 2020 | Accepted: 31 December 2020

DOI: 10.1002/gepi.22377

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

Assisted estimation of gene expression graphical models

Huangdi Yi[1] | **Qingzhao Zhang**[2] | **Yifan Sun**[3] | **Shuangge Ma**[1,2]

[1]Department of Biostatistics, Yale University, New Haven, Connecticut, USA

[2]Department of Statistics, Key Laboratory of Econometrics, Ministry of Education, School of Economics, The Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, China

[3]Center of Applied Statistics, School of Statistics, Renmin University of China, Beijing, China

Correspondence
Shuangge Ma, Department of Biostatistics, Yale University, New Haven, CT 06520, USA.
Email: shuangge.ma@yale.edu

Yifan Sun, Center of Applied Statistics, School of Statistics, Renmin University of China, Beijing, China.
Email: sunyifan1984@163.com

Funding information
National Natural Science Foundation of China, Grant/Award Number: 11971404; Higher Education Discipline Innovation Project, Grant/Award Number: B13028; National Science Foundation, Grant/Award Number: 1916251; National Institutes of Health, Grant/Award Numbers: CA216017, CA241699, CA121974, CA196530; Fund for building world-class universities (disciplines) of Renmin University of China; National Bureau of Statistics of China, Grant/Award Number: 2019LZ11

Abstract

In the study of gene expression data, network analysis has played a uniquely important role. To accommodate the high dimensionality and low sample size and generate interpretable results, regularized estimation is usually conducted in the construction of gene expression Gaussian Graphical Models (GGM). Here we use GeO-GGM to represent gene-expression-only GGM. Gene expressions are regulated by regulators. gene-expression-regulator GGMs (GeR-GGMs), which accommodate gene expressions as well as their regulators, have been constructed accordingly. In practical data analysis, with a "lack of information" caused by the large number of model parameters, limited sample size, and weak signals, the construction of both GeO-GGMs and GeR-GGMs is often unsatisfactory. In this article, we recognize that with the regulation between gene expressions and regulators, the sparsity structures of a GeO-GGM and its GeR-GGM counterpart can satisfy a hierarchy. Accordingly, we propose a joint estimation which reinforces the hierarchical structure and use the construction of a GeO-GGM to assist that of its GeR-GGM counterpart and vice versa. Consistency properties are rigorously established, and an effective computational algorithm is developed. In simulation, the assisted construction outperforms the separation construction of GeO-GGM and GeR-GGM. Two The Cancer Genome Atlas data sets are analyzed, leading to findings different from the direct competitors.

KEYWORDS
assisted estimation, gene expressions, graphical models, hierarchy

1 | INTRODUCTION

In biomedical research, gene expression data have been routinely generated. A long array of analysis has been conducted, among which network analysis has played a uniquely important role. Network analysis can not only lead to a deeper understanding of how genes affect each other but also serve as the basis of other important analyses, for example, regression and clustering. There are two main families of gene expression network construction: unconditional and conditional. In an unconditional construction, when quantifying whether two gene expressions are connected, information in other genes is not accounted for. In contrast, a conditional construction quantifies whether two gene expressions are connected *conditional on* the rest of the genes. In a sense,

with a system perspective, conditional construction can be more informative and more comprehensive. Statistically, it is more challenging as the analysis of each gene interconnection involves a large number of parameters.

In this article, we consider Gaussian Graphical Model (GGM), which is possibly the most popular conditional network construction approach. It has been extensively applied to the analysis of gene expression data and led to biologically useful findings. Representative examples include Dobra et al. (2004), Wang et al. (2016), Zhao and Duan (2019), and others. We acknowledge that the GGM approach is not ideal in the sense that it makes the multivariate normal distribution assumption, whereas practical gene expression data may have distributions deviating from normal. In the literature, there have been several works (Liu et al., 2012; Xue & Zou, 2012) relaxing this assumption, and we note that *the proposed technique can be directly coupled with these works*. However, these alternatives are not as lucidly interpretable as the GGM. In addition, when gene expression data are properly processed (possibly with transformations), our data examination suggests that usually the distributions are bell-shaped and unimodal. Considering the lucid interpretation and satisfactory performance observed in published data analysis, we choose the GGM for gene expression data while cautioning that exploratory analysis should be conducted in practice (to examine deviation from normality) before applying the proposed approach. We refer to X. T. Yuan and Zhang (2014), Ravikumar et al. (2011), and Suzuki (2013) for methodological developments, statistical properties, computational algorithms, and applications of GGMs under high-dimensional settings. There are multiple ways for estimating GGMs, in particular including probabilistic (Friedman et al., 2008) and Bayesian (Williams, 2018). In this article, we focus on the probabilistic estimation, which may be more popular. GGM is related to Bayesian networks (Jensen, 1996). In particular, both study conditional dependence. However, they are significantly different as Bayesian networks are directed, while GGM is not. In addition, Bayesian networks usually deal with categorically distributed nodes, whereas GGM assumes the continuous Gaussian distribution.

The levels of gene expressions are not "rootless" but instead highly regulated by regulators including copy number variations (CNVs), methylation, microRNAs, and others. In the past few years, we have witnessed a surge of multidimensional profiling studies, which collect measurements on gene expressions as well as their regulators on the same subjects. Such studies make it possible to jointly analyze gene expressions and their regulators, more informatively describing the whole molecular picture. In the context of network analysis, gene-expression-regulator GGMs (GeR-GGMs) have been constructed (Chiquet et al., 2017), under which the analysis of interconnection for two gene expressions is conditional on *the other gene expressions as well as regulators*. We refer to Chiquet et al. (2017) and other published studies for the rational and merit of GeR-GGM analysis. To differentiate the two types of analysis, we use GeO-GGM to represent a gene-expression-only GGM analysis. We note that such techniques are also applicable to other types of molecular data (Chun et al., 2013) and other types of biological data, and refer to Li et al. (2012), M. Yuan (2010), and others for additional relevant discussions.

Gene expression data analysis is challenged by the "high dimensional variables, small sample size" problem, which gets more serious in network analysis where the number of unknown parameters gets squared—this is especially true in GeR-GGM constructions. To accommodate the high dimensionality and generate sparse networks that match the underlying biology (i.e., a specific gene is only connected to a few other genes), regularized estimation has been extensively conducted. Among the existing approaches, the most famous is perhaps graphical Lasso (Friedman et al., 2008), which applies Lasso penalization in GGM estimation. Beyond Lasso, other penalization approaches and approaches based on other regularization techniques have also been developed (Witten & Tibshirani, 2009; M. Yuan, 2010). Despite satisfactory theoretical properties of the graphical Lasso and other regularized estimation approaches, in practical data analysis, numerical results are still often unsatisfactory, which can be attributable to a "lack of information" caused by the large number of unknown parameters, small sample size, and weak signals. To overcome this problem, various "information borrowing" techniques have been developed. For example, the horizontal data integration techniques pool multiple independent data sets that share certain similarity and jointly estimate multiple GeO-GGMs (or GeR-GGMs) (Cai et al., 2016). There are also studies that borrow information from prior knowledge, for example, functional annotations of genes or published findings (Mihaylov et al., 2019).

Our goal is to conduct more effective GGM analysis of gene expression data, when regulator data is available for at least some subjects (more detailed data setting described below). The gene expression networks generated by our analysis have the same implications and can be utilized in the same manner as in the literature (Dobra et al., 2004; Wang et al., 2016; Zhao & Duan, 2019). This study has been motivated by the importance of graphical models in the analysis of gene expression data, still not fully satisfactory performance of the existing analysis, and hence demand for new and more effective network

construction. It has been made possible by the growing popularity of multidimensional profiling. Significantly different from the existing studies, a new analysis strategy is proposed to borrow information across a GeO-GGM and its corresponding GeR-GGM, so that the estimation of the GeO-GGM can assist the estimation of the GeR-GGM, and vice versa. Loosely speaking, this strategy shares some similar spirit with the vertical data integration (Wang et al., 2019). This study may advance from the existing literature in the following aspects. The first is to propose a biologically sensible hierarchy between the GeO-GGM and GeR-GGM, which motivates our methodological development and has not been accounted for in the literature. Second, although the proposed penalized estimation shares some similarity with published studies, its application to the present context is new and novel. Third, statistical and numerical properties are rigorously established, providing the proposed method a stronger ground than some of the existing studies that are limited to numerical developments. Last but equally important, our study can provide new insights into gene interconnections for cutaneous melanoma and lung cancer and showcase how to extract more information from the The Cancer Genome Atlas (TCGA) data. Overall, this study can provide a practical and useful new venue for gene expression network analysis.

2 | METHODS

2.1 | Strategy

Consider gene expressions G1, G2, and G3, and regulator R (which can be multidimensional). In a GeO network analysis, the goal is to quantify, for example, (G1, G2)|G3, that is, the interconnection between G1 and G2 conditional on G3. This interconnection can be caused by multiple factors: (a) coregulation by R. If G1 and G2 are both regulated by R, then they can be interconnected; (b) coregulation by regulators other than R. Most if not all profiling studies are "incomplete," in the sense that not all regulators are measured; (c) direct effects such as gene interference; and (d) mechanisms yet to be identified. In the analysis of (G1, G2)|G3, G1 and G2 are interconnected if any of the above exists. In the analysis that accommodates regulators, the goal is to quantify (G1, G2)|(G3, R), that is, the interconnection between G1 and G2 caused by (b)–(d), after removing (accounting for) (a), and conditional on G3.

A GeO graphical model contains *all-causes gene interconnections*, whereas a GeR graphical model contains *only gene interconnections not explained by the analyzed regulators*. Motivated by this consideration, we proposed the hierarchy:

the edge set in the gene-expression-regulator graphical model is a subset of that in the gene-expression-only graphical model.

This hierarchy connects a GeO graphical model and its GeR counterpart. For a GeO graphical model, this hierarchy amounts to additional information. That is, if we can effectively take advantage of this hierarchy and "borrow strength" from its corresponding GeR graphical model, we can potentially improve its identification and estimation of gene connections. The same applies to the GeR graphical model. It is noted that this specific biologically sensible hierarchy has not been considered in the literature and can provide a way of information borrowing significantly different from the existing ones.

The above discussions are applicable to the scenario with gene coregulations by regulators not measured. As such, the proposed analysis does not demand the collection of all regulators. It also does not demand the collected regulators all being informative. In the worst-case scenario, R only contains unrelated noises. Then the proposed analysis will basically reduce to a GeO network analysis, with no gain of information from regulators but also no loss.

2.1.1 | Remarks

Identifying biologically motivated hierarchy to assist data analysis is by no means new. Examples include Schadt et al. (2005), Yazdani et al. (2020), Zhu et al. (2012), and a few others. In a sense, they provide support to our general strategy of improving estimation/selection with the assistance of the hierarchy. Our literature review suggests that our study fundamentally differs from the existing hierarchies/approaches in one or more of the following aspects. First, the aforementioned and some other hierarchy-incorporating studies address problems other than conditional network analysis using the GGM technique. Second, although some of the existing studies also deal with high-dimensional data, they conduct the analysis of a small number of variables at a time and hence does not demand regularized estimation/selection. Third, hierarchy is not reinforced with penalization, which is one of the state-of-the-art high-dimensional techniques. Fourth, as shown below, the joint analysis of high-dimensional variables and penalized estimation demand challenging methodological, computational, as well as theoretical developments, which are not present in the literature.

There are also other ways of jointly analyzing gene expression and regulator data related to the network analysis paradigm. For example, in Wu et al. (2018), the

associations between gene expressions and their regulators are analyzed, taking into account the interconnections among genes/regulators. However, these studies do not focus on the construction of gene networks, and there is no counterpart of the proposed hierarchy.

Strictly speaking, it is possible to design settings under which the proposed hierarchy fails. With a slight abuse of notation, we use $G1, G2, R1$ and $R2$ to also denote the variables representing gene expressions and regulators. Considering the linear regression models for generating gene expressions:

$$G1 = R1 + R2 + \epsilon_1, \quad G2 = R1 - R2 + \epsilon_2,$$

where $R1, R2$ are independent and $N(0, 1)$ distributed, and ϵ_1, ϵ_2 are random errors. Here $G1$ and $G2$ are independent. However, conditional on $R1$, they are not. Our preliminary exploration suggests that *it is possible to design more complicated settings, for example, involving more genes and regulators, however, they share the same spirit.* Failure of the hierarchy demands regulators with completely complementary effects *and* that only one part of such regulators is measured. When $R1$ and $R2$ are two different types of regulators, our extensive literature search suggests that, to date, regulators with such complementary effects have not been identified. When $R1$ and $R2$ are the same type, studies have found regulators with strongly negatively correlated effects—but they are correlated, not independent. Under the worst-case scenario that independent and complementary $R1, R2$ do exist, a closer examination of our methodology and theoretical development suggests that, because of the existence of the interconnection conditional on the regulators (in the GeR-GGM), the interconnection in the GeO-GGM will be identified. Thus, there will be a false positive discovery. However, with the estimation consistency results described below, the estimate of the edge will converge to zero. More discussions are provided below.

2.2 | Assisted estimation

Let $Y = (Y_1, ..., Y_p)^\top$ denote p gene expressions and $X = (X_1, ..., X_q)^\top$ denote q regulators. With multiple types of regulators, their measurements can be stacked together. Consider a data set $D_1 = \{\mathbf{y}\}_{i=1}^{n_1}$ with n_1 i.i.d. copies of Y and a data set $D_2 = \{(\mathbf{y}, \mathbf{x})\}_{i=1}^{n_2}$ with n_2 i.i.d. copies of (Y, X). The GeO-GGM and GeR-GGM analysis will be conducted on D_1 and D_2, respectively. Our strategy is to simultaneously estimate the GeO-GGM and GeR-GGM, borrow information across each other via the hierarchy, and improve performance for both. The proposed analysis can flexibly accommodate multiple scenarios. The first scenario is where the same samples have both gene expression and regulator measurements. In this case, D_1 contains only gene expression measurements, while D_2 contains both gene expression and regulator measurements on the same samples. This scenario is considered in our simulation and data analysis. Under the second scenario, D_1 and D_2 are generated by different studies, and there is no overlapping subject. This scenario is also considered in our simulation. Under the third scenario, in a single study, some samples have only gene expression measurements, while others have both gene expression and regulator measurements.

Under the GGM framework, it is assumed that Y and $Y|X$ are Gaussian distributed. The graph structures are fully determined by the precision matrices. Specifically, first consider the GeO-GGM. Denote $\widetilde{\Sigma}_{YY}$ and $\widetilde{\Omega}_{YY}$ as the covariance and precision matrices of Y, respectively. Then $Y_i \perp\!\!\!\perp Y_j \mid Y_{-(i,j)} \Leftrightarrow \widetilde{\Omega}_{ij} = 0$, where $\widetilde{\Omega}_{ij}$ is the (i, j)th element of $\widetilde{\Omega}$ and $Y_{-(i,j)}$ is Y with the ith and jth elements removed. Further consider the GeR-GGM. Denote the precision matrix of (Y, X) as $\Omega = \begin{pmatrix} \Omega_{YY} & \Omega_{YX} \\ \Omega_{YX}^\top & \Omega_{XX} \end{pmatrix}$. Then $(\Omega_{YY})_{ij} = 0$ is equivalent to $Y_i \perp\!\!\!\perp Y_j \mid Y_{-(i,j)}, X$, where $(\Omega_{YY})_{ij}$ is the (i, j)th entry of Ω_{YY}.

We adopt penalization, a state-of-the-art high-dimensional technique, for the estimation and identification of graph structures. To reinforce the hierarchy and realize information borrowing, we propose jointly estimating the GeO-GGM and GeR-GGM. Denote $\widetilde{S}_{YY}$ as the empirical covariance matrix calculated using $D_1 \cup D_2$, S_{YY} as the empirical covariance matrix calculated using D_2, S_{YX} as the empirical correlation matrix calculated using D_2, and S_{XX} as the empirical correlation matrix calculated using D_2. We propose the objective function:

$$Q(\widetilde{\Omega}_{YY}, \Omega_{YY}, \Omega_{YX}) = L_1(\widetilde{\Omega}_{YY}) + L_2(\Omega_{YY}, \Omega_{YX}) + P_1(\widetilde{\Omega}_{YY}, \Omega_{YY}) + P_2(\Omega_{YX}), \quad (1)$$

where

$$\begin{aligned} L_1(\widetilde{\Omega}_{YY}) &= -\operatorname{logdet}(\widetilde{\Omega}_{YY}) + \operatorname{tr}(\widetilde{S}_{YY}\widetilde{\Omega}_{YY}), L_2(\Omega_{YY}, \Omega_{YX}) \\ &= -\operatorname{logdet}(\Omega_{YY}) + \operatorname{tr}(S_{YY}\Omega_{YY}) \\ &\quad + 2\operatorname{tr}(S_{YX}^\top\Omega_{YX}) + \operatorname{tr}(S_{XX}\Omega_{YX}^\top\Omega_{YY}^{-1}\Omega_{YX}), P_1 \\ &\quad (\widetilde{\Omega}_{YY}, \Omega_{YY}) \\ &= \sum_{i\neq j}\rho\left(\sqrt{(\widetilde{\Omega}_{YY})_{ij}^2 + (\Omega_{YY})_{ij}^2}; \lambda_1, \gamma\right) \\ &\quad + \sum_{i\neq j}\rho(|(\Omega_{YY})_{ij}|; \lambda_2, \gamma), P_2(\Omega_{YX}) \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q}\rho(|(\Omega_{YX})_{ij}|; \lambda_2, \gamma). \end{aligned}$$

Here $\rho(t;\lambda,\gamma)=\lambda\int_0^{|t|}\left(1-\frac{x}{\lambda\gamma}\right)_+dx$ is the minimax concave penalty (MCP; Zhang, 2010), λ_1 and λ_2 are data-dependent tuning parameters, and γ is the regularization parameter. The estimate is defined as the minimizer of (1), and a nonzero element corresponds to an interconnection.

Remarks: Distributions of regulator data may further deviate from normality. With CNV (which is analyzed in this study), although the raw measurements are discrete, with proper processing as in TCGA, data distributions are continuous and mostly bell-shaped. As such, it can be reasonable to analyze under the GGM framework. With continuously distributed regulators such as methylation and microRNA, marginal transformations can be applied to get closer to normality. With for example SNP, gene-level data aggregation and transformation may lead to distributions closer to continuous and normal. However, if not, we propose following the literature and replacing the simple correlation with robust, for example, rank-based, correlations to accommodate nonnormality. Then the proposed approach can be applied.

Methodologically advancing from many of the existing studies, the proposed approach jointly estimates the GeO-GGM and GeR-GGM. We note that this differs from the joint analysis of multiple GeO-GGMs. There are two lack-of-fit functions. L_1 is standard for the GeO-GGM. In the GeR-GGM estimation, the interconnections among regulators are not of interest. As such, we adopt a partial GGM approach (X. T. Yuan & Zhang, 2014), which uses a reparametrization and effectively avoids the Ω_{XX} term in L_2. This is computationally advantageous especially when the dimension of the regulators is high. In addition, this avoids making additional assumptions on the interconnections among regulators. It is noted that, when needed, the full GeR-GGM lack-of-fit function can be adopted. As described above, the proposed approach can accommodate the scenario where some samples are used for the construction of both L_1 and L_2. However, as can be seen from the theoretical development below, there are no correlation or "double dipping" problems.

The proposed penalties have two components. The first, P_1, is a sparse group penalty built on MCP. It generates sparse estimates (graphs) and, equally importantly, reinforces the hierarchy. Specifically, if the estimate of $(\Omega_{YY})_{ij}$ is nonzero, the estimate of $(\widetilde{\Omega}_{YY})_{ij}$ is guaranteed to be nonzero (Huang et al., 2012). This way, estimates in the GeO-GGM and those in the GeR-GGM affect each other. For estimating one network, estimates of the other network provide additional information through the hierarchy, realizing information borrowing.

The second component, P_2, is a "standard" sparse penalty. Ω_{YX}, which describes the conditional interconnections between gene expressions and regulators, is also expected to be sparse. As such, P_2 is imposed to generate sparsity and accommodate the high dimensionality. We note that the above discussions are valid as long as a "GeO-GGM + GeR-GGM" estimation problem is sensibly formulated. In particular, all the three different D_1+D_2 data scenarios described above can be accommodated.

Consider the scenario that the hierarchy is actually violated, that is, the true value of $(\Omega_{YY})_{ij}$ is nonzero but that of $(\widetilde{\Omega}_{YY})_{ij}$ is zero. In this case, the proposed approach will generate nonzero estimates for both, leading to a false-discovery with respect to $(\widetilde{\Omega}_{YY})_{ij}$. With the estimation consistency established below, the estimate for the zero entry will be very small. In practical data analysis, small estimates in $(\widetilde{\Omega}_{YY})$ can raise alarm, with which one needs to more carefully examine data to identify potential violation of the hierarchy. If found, separate estimation of the GeO-GGM and GeR-GGM will be needed.

2.3 | A small example

To gain more intuition into the proposed analysis, we simulate one small data set with $p=20$, $q=20$, and $n=n_1=100$. Ω_{YY} has a homogeneous structure with $\theta=0.1$. More details on the simulation settings are provided in Section 3. The true data generating model has a total of 36 nonzero off-diagonal entries in Ω_{YY} and 46 nonzero off-diagonal entries in $\tilde{\Omega}_{YY}$ (left panels of Figure 1). Beyond the proposed method, we also consider the alternative that separately estimates the GeO-GGM and GeR-GGM using the MCP technique, to explicitly demonstrate the benefit of joint estimation. The estimated network structures are also shown in Figure 1.

For this specific example, the proposed method has more accurate identification. Specifically, for the GeR network (Ω_{YY}), it identifies 14 true-positives and five false-positives, whereas the alternative separate estimation identifies 11 true-positives and nine false-positives. For the GeO network ($\tilde{\Omega}_{YY}$), the proposed method identifies 16 true positives and seven false-positives, where the alternative identifies 13 true positives and nine false-positives. The alternative identification result violates the hierarchy. Specifically, there are three edges that are identified in the GeR network but not in the GeO one. We further examine estimation performance using RMSE (details in Section 3). The RMSE values of Ω_{YY} are 14.48 (proposed) and 15.61 (alternative), and those of $\tilde{\Omega}_{YY}$ are 7.65 (proposed) and 8.04 (alternative). More definitive results based on larger scale simulations are presented in Section 3.

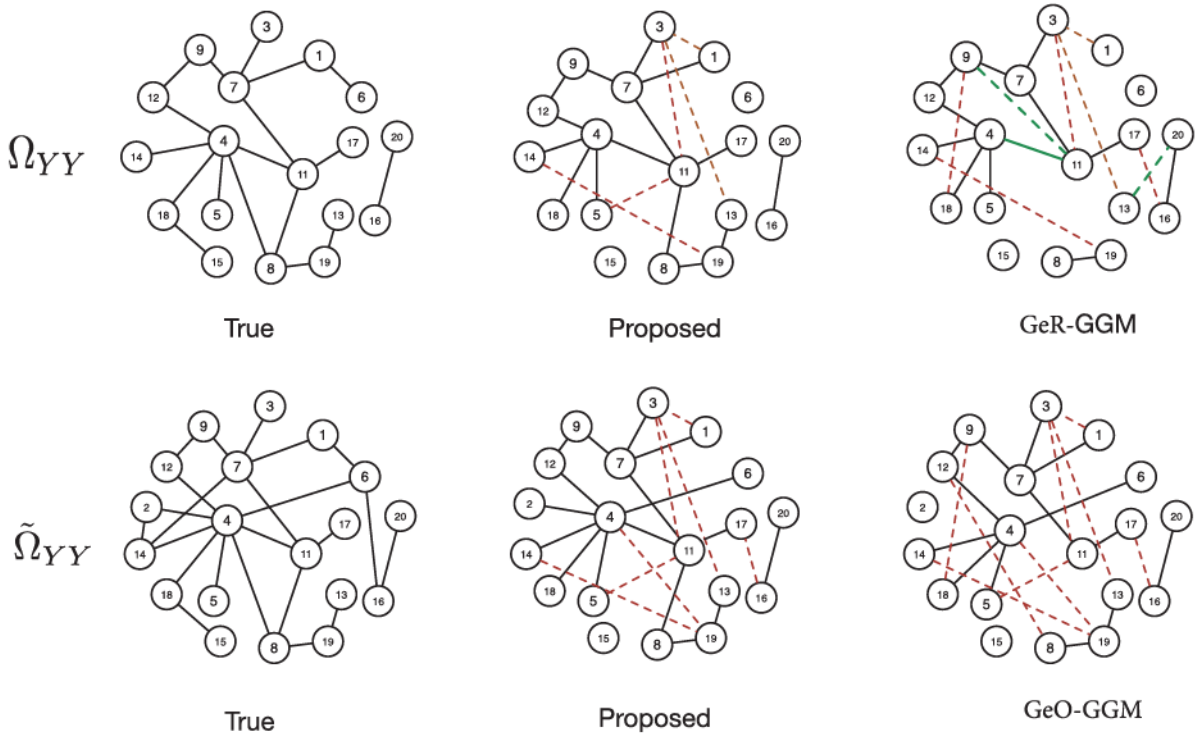

FIGURE 1 Gene expression networks in the small example: true (left), proposed (middle), and alternatives (right). Solid lines: true-positives; dashed lines: false-positives; green lines: identifications that violate the hierarchy

2.4 | Statistical properties

Rigorously establishing statistical properties can provide the proposed approach a stronger ground than those not properly supported. Suppose that gene expressions $(Y_1, ..., Y_p)$ are associated with the vertex set $V_1 = \{1, 2, ..., p\}$ of the undirected graph $G_1 = (V_1, E_1)$, and that gene expressions plus regulators $(Y_1, ..., Y_p, X_1, ..., X_q)$ are associated with the vertex set $V_2 = \{1, 2, ..., p + q\}$ of the undirected graph $G_2 = (V_2, E_2)$. Here E_1 and E_2 are the sets of edges. We first define the following support sets and their complements. Let $\widetilde{\mathcal{A}}_{YY} = \{(i,j)|(\widetilde{\Omega}^*_{YY})_{ij} \neq 0; i,j = 1, ...p\}$, $\mathcal{A}_{YY} = \{(i,j)|(\Omega^*_{YY})_{ij} \neq 0; i,j = 1, ...p\}$, and $\mathcal{A}_{YX} = \{(i,j)|(\Omega^*_{YX})_{ij} \neq 0; i = 1, ...p; j = p, ...p + q\}$ be the sets of indices of all nonzero elements in $\widetilde{\Omega}^*_{YY}$, Ω^*_{YY}, and Ω^*_{YX}, respectively. Here and below, values with superscript "*" denote the true values. Further denote $\mathcal{A} = \mathcal{A}_{YY} \cup \mathcal{A}_{YX}$, $\mathcal{A}^c = \{(i,j)|i = 1, ..., p; j = 1, ..., p + q\}\backslash\mathcal{A}$, $\mathcal{A}_1 = \mathcal{A}_{YY} \cup \widetilde{\mathcal{A}}_{YY}$, and $\mathcal{A}_1^c = \{(i,j)|i = 1, ..., p; j = 1, ..., p\}\backslash\mathcal{A}_1$.

Define the following estimates:

$$\widehat{\widetilde{\Omega}}_{YY} = \arg\min_{\widetilde{\Omega}_{YY}\succ 0, (\widetilde{\Omega}_{YY})_{\mathcal{A}_1^c}=0} L_1(\widetilde{\Omega}_{YY}),$$

$$\widehat{\Theta} = \arg\min_{\Omega_{YY}\succ 0, \Theta_{\mathcal{A}^c}=0} L_2(\Theta),$$

where $\Theta = (\Omega_{YY}, \Omega_{YX})$. We also denote the maximum degrees of the two graphs as $\widetilde{d} := \max_{i=1,...,p} |\{j \in V_1 | (\widetilde{\Omega}^*_{YY})_{ij} \neq 0\}|$ and $d := \max_{i=1,...,p} |\{j \in V_2 | \Omega^*_{ij} \neq 0\}|$.

Consider the ℓ_1 and ℓ_∞ norms. Specifically, for a matrix $A \in \mathbb{R}^{l\times m}$, $\| A \|_1 = \max_{1\le j\le m} \sum_{i=1}^{l} | A_{ij} |$, and $\| A \|_\infty = \max_{1\le i\le l} \sum_{j=1}^{m} | A_{ij} |$. Denote $\kappa_{\Sigma^*_{YY}} := \| \Sigma^*_{YY} \|_\infty$. With results on matrix derivatives, it can be shown that the Hessian of $\log\det(\widetilde{\Omega}_{YY})$, evaluated at $\widetilde{\Omega}^*_{YY}$, takes the form $\tilde{\Gamma}^* := \widetilde{\Omega}^{*-1}_{YY} \otimes \widetilde{\Omega}^{*-1}_{YY}$, where $\otimes$ denotes the Kronecker product. Consequently, we define $\tilde{\Gamma}^*_{\mathcal{A}_1\mathcal{A}_1} := [\widetilde{\Omega}^{*-1}_{YY} \otimes \widetilde{\Omega}^{*-1}_{YY}]_{\mathcal{A}_1\mathcal{A}_1}$, $\kappa_{\tilde{\Gamma}^*} := \| (\tilde{\Gamma}^*_{\mathcal{A}_1\mathcal{A}_1})^{-1} \|_\infty$, and $\tilde{\kappa} := \max_{e\in\mathcal{A}_1^c} \| \tilde{\Gamma}^*_{e\mathcal{A}_1} (\tilde{\Gamma}^*_{\mathcal{A}_1\mathcal{A}_1})^{-1} \|_1$. For the GeR graph, we denote its Hessian evaluated at the true values as:

$$H^* := H(\Omega^*_{YY}, \Omega^*_{YX}) = \begin{pmatrix} \Omega^{*-1}_{YY} \otimes (\Omega^{*-1}_{YY} + 2\Omega^{*-1}_{YY}\Omega^*_{YX}S_{XX}\Omega^{*T}_{YX}\Omega^{*-1}_{YY}) & -2\Omega^{*-1}_{YY} \otimes S_{XX}\Omega^{*T}_{YX}\Omega^{*-1}_{YY} \\ -2\Omega^{*-1}_{YY} \otimes \Omega^{*-1}_{YY}\Omega^*_{YX}S_{XX} & 2\Omega^{*-1}_{YY} \otimes S_{XX} \end{pmatrix}.$$

Similar to above, we define $\kappa_1 := \max_{e\in\mathcal{A}_1^c} \| H^*_{e\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1} \|_1$, $\kappa_2 := \max_{e\in\widetilde{\mathcal{A}}_{YY}\cup\mathcal{A}_{YX}} \| H^*_{e\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1} \|_1$, $\kappa_3 := \max_{e\in\mathcal{A}^c_{YX}} \| H^*_{e\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1} \|_1$, $c_{\Omega^{*-1}_{YY}} := \| \Omega^{*-1}_{YY} \|_\infty$, $c_{\Omega^*_{YX}} := \| \Omega^*_{YX} \|_1$, and $c_{H^*} := \| \Omega^{*-1}_{\mathcal{A}\mathcal{A}} \|_\infty$.

The following conditions, which pertain the model, sample size, and edge signals, are assumed.

They are comparable to those in the existing GGM studies.

Condition 1. $\min_{(i,j)\in\mathcal{A}_1}\left(|(\widetilde{\Omega}^*_{YY})_{ij}| + |(\Omega^*_{YY})_{ij}|\right) > \{\gamma + \kappa_{\widetilde{\Gamma}^*}/(\tilde{\kappa}+1)\}\lambda_1$.

Condition 2. $\min_{(i,j)\in\mathcal{A}\setminus\mathcal{A}_1}\left(|(\Omega^*_{YY})_{ij}|, |(\Omega^*_{YX})_{ij}|\right) > c_{H^*}\min\left\{\frac{\lambda_1+\lambda_2}{\kappa_1+1}, \frac{\lambda_2}{\kappa_2+1}, \frac{\lambda_2}{\kappa_3+1}\right\} + (\lambda_1 \vee \lambda_2)\gamma$.

Condition 3. $\max_j \| X_j \|_2/\sqrt{n_2} \le c_X$, where c_X is a constant.

Under these conditions, we can establish the following consistency properties.

Theorem 1. *Suppose that the sample sizes satisfy:* $n_1 > \max\{0, C_1\log(4p^\tau)\tilde{d}^2 - C_2\log[4(p\vee q)^\tau]d^2\}$, $n_2 > C_2\log[4(p\vee q)^\tau]d^2$, *where* $C_1 = [\max\{\kappa_{\Sigma^*_{YY}}\kappa_{\widetilde{\Gamma}^*}, \kappa^3_{\Sigma^*_{YY}}\kappa^2_{\widetilde{\Gamma}^*}\}^2]$ *and* $C_2 = c^2_{H^*}[\max\{3c_{\Omega^{*-1}_{YY}}, c_{\Omega^*_{YX}}, c_{\Omega^{*-1}_{YY}}c^2_{\Omega^*_{YX}}c^2_X\}^2]$. *In addition, the regularization and tuning parameters satisfy* $\lambda_1 > 2(\tilde{\kappa}+1)c_*\sqrt{\frac{\log(4p^\tau)}{n_1+n_2}}$, *and* $\min\left\{\frac{\lambda_1+\lambda_2}{\kappa_1+1}, \frac{\lambda_2}{\kappa_2+1}, \frac{\lambda_2}{\kappa_3+1}\right\} > 2c'_*\sqrt{\frac{\log(4(p\vee q)^\tau)}{n_2}}$. *For some* $\tau > 2$ *and probability at least* $1 - 1/p^{\tau-2} - 2/(p\vee q)^{\tau-2}$:

(I) the estimates have nonzero entries that are the same as those of the true values;

(II) with $c_* = 40\sqrt{2}\max_{i=1,\ldots,p}(\widetilde{\Omega}^{*-1}_{YY})_{ii}$ *and* $c'_* = \max\{40\sqrt{2}\max_i(\Omega^{*-1}_{YY})_{ii}, 2\sqrt{2}c_X\}$,

$$\| \widetilde{\widehat{\Omega}}_{YY} - \widetilde{\Omega}^*_{YY} \|_\infty \le 2c_*\kappa_{\widetilde{\Gamma}^*}\sqrt{\frac{\log(4p^\tau)}{n_1+n_2}}, \tag{2}$$

$$\| \widehat{\Omega}_{YY} - \Omega^*_{YY}, \widehat{\Omega}_{YX} - \Omega^*_{YX} \|_\infty \le 2c'_*c_{H^*}\sqrt{\frac{\log(4(p\vee q)^\tau)}{n_2}}. \tag{3}$$

These results have the following theoretical implications in an asymptotic sense. Under mild conditions, result (I) establishes that the important and unimportant edges can be correctly distinguished. Result (II) further establishes that, asymptotically, the estimates can be very close to the true values. As such, the proposed method is theoretically guaranteed to recover the true GeO-GGM and GeR-GGM structures. Such a theoretical rigor is not presented in many of the existing studies. With the two sets of estimates, complexity of graph models, and differences in the imposed penalties, the proof differs significantly from the literature and is highly nontrivial. It can also shed insights for other network analysis studies. Details are presented in the Appendix.

As for most theoretical studies, there is a "gap" between theoretical conclusions and practical applications. For example, the consistency is in an asymptotic sense with sample sizes go to infinity, while with any practical data, sample size is finite.

2.5 | Computation

We optimize objective function (1) using the proximal gradient decent (PGD) technique. The proposed algorithm adopts the backtracking line search to determine the step size. Specifically, it proceeds as follows:

1. Initialize: $t = 0$, $\Omega^{(t)}_{YY} = \widehat{\Omega}_{YY}$, $\Omega^{(t)}_{YX} = \widehat{\Omega}_{YX}$, $\widetilde{\Omega}^{(t)}_{YX} = S^{-1}_{YY}$, where $\widehat{\Omega} = \begin{pmatrix} \widehat{\Omega}_{YY} & \widehat{\Omega}_{YX} \\ \widehat{\Omega}^\top_{YX} & \widehat{\Omega}_{XX} \end{pmatrix}$ is calculated from data. $\eta^{(0)} = 1$.
2. Update:
   (1) Calculate
      a. For each (i,j)th off-diagonal element, minimize $M_1((\Omega_{YY})_{ij})$ with respect to $(\Omega_{YY})_{ij}$, where

$$M_1((\Omega_{YY})_{ij}) = \frac{1}{2}[(\Omega_{YY})_{ij} - ((\Omega^{(t)}_{YY})_{ij} - \eta^{(t)}A^{(t)}_{ij})]^2 + \eta^{(t)}\rho\left(\sqrt{(\Omega_{YY})^2_{ij} + \left(\widetilde{\Omega}^{(t)}_{YY}\right)^2_{ij}}; \lambda_1, \gamma\right) + \eta^{(t)}\rho(|(\Omega_{YY})_{ij}|; \lambda_2, \gamma). \tag{4}$$

Here $A^{(t)} = S_{YY} - \left(\Omega^{(t)}_{YY}\right)^{-1} - \left(\Omega^{(t)}_{YY}\right)^{-1}\Omega^{(t)}_{YX}S_{XX}\left(\Omega^{(t)}_{YX}\right)^\top\left(\Omega^{(t)}_{YY}\right)^{-1}$.

      b. For each (i,j)th off-diagonal element, minimize $M_2((\widetilde{\Omega}_{YY})_{ij})$ with respect to $(\widetilde{\Omega}_{YY})_{ij}$, where

$$M_2((\widetilde{\Omega}_{YY})_{ij}) = \frac{1}{2}\left[(\widetilde{\Omega}_{YY})_{ij} - \left(\left(\widetilde{\Omega}^{(t)}_{YY}\right)_{ij} - \eta^{(t)}B^{(t)}_{ij}\right)\right]^2 + \eta^{(t)}\rho\left(\sqrt{(\Omega^{(t)}_{YY})^2_{ij} + (\widetilde{\Omega}_{YY})^2_{ij}}; \lambda_1, \gamma\right).$$

Here $B^{(t)} = S_{YY} - (\widetilde{\Omega}^{(t)}_{YY})^{-1}$.

With $\gamma > \eta^{(t)}$, the solutions are

$$(\Omega^{\star}_{YY})_{ij}=\begin{cases}\dfrac{R^{(t)}_{ij}}{1-\eta^{(t)}/\gamma}\left(1-\dfrac{\lambda_1\eta^{(t)}}{\sqrt{\left(R^{(t)}_{ij}\right)^2+\left(\left(\widetilde{\Omega}^{(t)}_{YY}\right)_{ij}-\eta^{(t)}B^{(t)}_{ij}\right)^2}}\right)_+ & \text{if } \sqrt{\left(R^{(t)}_{ij}\right)^2+\left(\left(\widetilde{\Omega}^{(t)}_{YY}\right)_{ij}-\eta^{(t)}B^{(t)}_{ij}\right)^2}\le\gamma\lambda_1\\ R^{(t)}_{ij} & \text{if } \sqrt{\left(R^{(t)}_{ij}\right)^2+\left(\left(\widetilde{\Omega}^{(t)}_{YY}\right)_{ij}-\eta^{(t)}B^{(t)}_{ij}\right)^2}>\gamma\lambda_1\end{cases}$$

$$\left(\widetilde{\Omega}^{\star}_{YY}\right)_{ij}=\begin{cases}\dfrac{\left(\widetilde{\Omega}^{(t)}_{YY}\right)_{ij}-\eta^{(t)}B^{(t)}_{ij}}{1-\eta^{(t)}/\gamma}\left(1-\dfrac{\lambda_1\eta^{(t)}}{\sqrt{\left(R^{(t)}_{ij}\right)^2+\left(\left(\widetilde{\Omega}^{(t)}_{YY}\right)_{ij}-\eta^{(t)}B^{(t)}_{ij}\right)^2}}\right)_+ & \text{if } \sqrt{\left(R^{(t)}_{ij}\right)^2+\left(\left(\widetilde{\Omega}^{(t)}_{YY}\right)_{ij}-\eta^{(t)}B^{(t)}_{ij}\right)^2}\le\gamma\lambda_1\\ \left(\widetilde{\Omega}^{(t)}_{YY}\right)_{ij}-\eta^{(t)}B^{(t)}_{ij} & \text{if } \sqrt{\left(R^{(t)}_{ij}\right)^2+\left(\left(\widetilde{\Omega}^{(t)}_{YY}\right)_{ij}-\eta^{(t)}B^{(t)}_{ij}\right)^2}>\gamma\lambda_1\end{cases}$$

where

$$R^{(t)}_{ij}=\begin{cases}\dfrac{S\left(\left(\Omega^{(t)}_{YY}\right)_{ij}-\eta^{(t)}A^{(t)}_{ij},\lambda_2\eta^{(t)}\right)}{1-\eta^{(t)}/\gamma} & \text{if } \left|\left(\Omega^{(t)}_{YY}\right)_{ij}-\eta^{(t)}A^{(t)}_{ij}\right|\le\gamma\lambda_2\\ \left(\Omega^{(t)}_{YY}\right)_{ij}-\eta^{(t)}A^{(t)}_{ij} & \text{if } \left|\left(\Omega^{(t)}_{YY}\right)_{ij}-\eta^{(t)}A^{(t)}_{ij}\right|>\gamma\lambda_2\end{cases}.$$

Here $S(z,\lambda)=(1-\frac{\lambda}{|z|})_+z$.

c. For each (i,j)th element, minimize $M_3((\Omega_{YX})_{ij})$ with respect to $(\Omega_{YX})_{ij}$, where

$$M_3((\Omega_{YX})_{ij})=\frac{1}{2}\left[(\Omega_{YX})_{ij}-\left(\left(\Omega^{(t)}_{YX}\right)_{ij}-\eta C^{(t)}_{ij}\right)\right]^2+\eta^{(t)}\rho(|(\Omega_{YX})_{ij}|;\lambda_2,\gamma).$$

Here $C^{(t)}=2[(\Omega^{\star}_{YY})^{-1}\Omega^{(t)}_{YX}S_{XX}+S_{YX}]$.

With $\gamma>\eta$, the solution is

$$\left(\Omega^{\star}_{YX}\right)_{ij}=\begin{cases}\dfrac{S((\Omega^{(t)}_{YX})_{ij}-\eta^{(t)}C^{(t)}_{ij},\lambda_2\eta^{(t)})}{1-\eta^{(t)}/\gamma} & \text{if } |(\Omega^{(t)}_{YX})_{ij}-\eta^{(t)}C^{(t)}_{ij}|\le\gamma\lambda_2\\ (\Omega^{(t)}_{YX})_{ij}-\eta^{(t)}C^{(t)}_{ij} & \text{if } |(\Omega^{(t)}_{YX})_{ij}-\eta^{(t)}C^{(t)}_{ij}|>\gamma\lambda_2\end{cases}.$$

(2) Determine the step size.

Calculate the quadratic approximations of $L_1(\widetilde{\Omega}^{\star}_{YY})$ and $L_2(\Omega^{\star}_{YY},\Omega^{\star}_{YX})$:

$$\begin{aligned}\widetilde{L}_1(\widetilde{\Omega}^{\star}_{YY})&=L_1\left(\widetilde{\Omega}^{(t)}_{YY}\right)+\operatorname{tr}\left((B^{(t)})^{\top}\left(\widetilde{\Omega}^{\star}_{YY}-\widetilde{\Omega}^{(t)}_{YY}\right)\right)\\&\quad+\frac{1}{2\eta^{(t)}}\left\|\widetilde{\Omega}^{\star}_{YY}-\widetilde{\Omega}^{(t)}_{YY}\right\|_F^2+\widetilde{L}_2(\Omega^{\star}_{YY},\Omega^{\star}_{YX})\\&=L_2(\Omega^{(t)}_{YY},\Omega^{(t)}_{YX})+\operatorname{tr}\left((A^{(t)})^{\top}(\Omega^{\star}_{YY}-\Omega^{(t)}_{YY})\right)\\&\quad+\operatorname{tr}((C^{(t)})^{\top}(\Omega^{\star}_{YX}-\Omega^{(t)}_{YX}))\\&\quad+\frac{1}{2\eta^{(t)}}[\|\Omega^{\star}_{YY}-\Omega^{(t)}_{YY}\|_F^2\\&\quad+\|\Omega^{\star}_{YX}-\Omega^{(t)}_{YX}\|_F^2].\end{aligned}\tag{5}$$

If $L_1(\widetilde{\Omega}^{\star}_{YY})+L_2(\Omega^{\star}_{YY},\Omega^{\star}_{YX})>\widetilde{L}_1(\widetilde{\Omega}^{\star}_{YY})+\widetilde{L}_2(\Omega^{\star}_{YY},\Omega^{\star}_{YX}),\eta^{(t)}\leftarrow 0.5\eta^{(t)}$, and return to Step 1; else continue.

(3) Update the estimates of Ω_{YY}, $\widetilde{\Omega}_{YY}$, and Ω_{YX} as

$$\left(\Omega_{YY}^{(t+1)}\right)_{ij} \leftarrow \begin{cases} (\Omega_{YY}^{\star})_{ij} & i \neq j \\ (\Omega_{YY}^{(t)})_{ij} & i = j \end{cases}, \quad \left(\widetilde{\Omega}_{YY}^{(t+1)}\right)_{ij}$$
$$\leftarrow \begin{cases} (\widetilde{\Omega}_{YY}^{\star})_{ij} & i \neq j \\ \left(\widetilde{\Omega}_{YY}^{(t)}\right)_{ij} & i = j \end{cases}, \quad \left(\Omega_{YX}^{(t+1)}\right)_{ij}$$
$$\leftarrow \left(\Omega_{YX}^{\star}\right)_{ij}.$$

3. Repeat Step 2 until convergence. In numerical study, we use

$$\| \Omega_{YY}^{(t+1)} - \Omega_{YY}^{(t)} \|_F + \left\| \widetilde{\Omega}_{YY}^{(t+1)} - \widetilde{\Omega}_{YY}^{(t)} \right\|_F + \| \Omega_{YX}^{(t+1)} - \Omega_{YX}^{(t)} \|_F \leq 10^{-3}$$

as the convergence criterion, where $\| A \|_F \equiv \sqrt{\sum_{i=1}^{l}\sum_{j=1}^{m} |a_{ij}|^2}$ for matrix $A \in \mathbb{R}^{l\times m}$.

In all of our numerical analysis, convergence is satisfactorily achieved. The proposed algorithm is computationally affordable. With fixed tunings, the analysis of one simulated data (details described below) takes about 30 s on a regular laptop. The proposed approach involves the MCP regularization parameter γ. As in published studies, we examine a few values and find that $\gamma = 6$ leads to the best performance for our numerical examples. λ_1 and λ_2 are obtained using V-fold cross validation.

3 | SIMULATION

The precision matrix Ω can be decomposed into four submatrices: Ω_{YY}, Ω_{YX}, $\Omega_{YX}^{\top}$, and Ω_{XX}, which are generated as follows. Each entry of Ω_{YX} is generated independently, and equals 1 with probability θ and 0 with probability $1-\theta$. For Ω_{YY}, we consider the following structures: (a) a homogeneous structure, under which each off-diagonal entry of Ω_{YY} is independently drawn from a Bernoulli distribution with a success probability of θ. The diagonal elements of Ω_{YY} are zero; (b) a block structure, under which Ω_{YY} equals $\begin{bmatrix} \mathbf{A}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{A}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{A}_5 \end{bmatrix}$. For each block $A_k (k = 1, ..., 5)$, the diagonal elements are zero, and the off-diagonal elements are independently drawn from a Bernoulli distribution with a success probability of θ. All elements of Ω_{XX} are set as 0.5. To ensure the positive-definiteness of Ω, we add a diagonal matrix $\sigma\mathbf{I}$, and σ is set as 10. $\tilde{\Omega}_{YY}$ that follows this data generation is sparse. For example, for the setting described in Table 1, about 13.0% of its elements are nonzero. In addition, this data generation leads to graphs that satisfy the hierarchy. We generate i.i.d. observations from $N(0, \Sigma)$ with $\Sigma = \Omega^{-1}$. As shown in Table 1 in the main text and tables 2–6 in the appendix, we consider $\theta = 0.1$ and 0.05. For the (p, q) dual, we consider (50, 50), (50, 100), (50, 150), (100, 50), (100, 100), and (100, 150). To demonstrate the broad applicability of the proposed approach, we consider two different scenarios for D_1 and D_2. More specifically, we first consider the first scenario described in the "Assisted estimation" section, where D_1 and D_2 contain the same subjects and $n_1 = n_2 = 300$. Here the subjects are "analyzed twice," first with Y only and then with both Y and X. Then we consider the second scenario, where D_1 and D_2 contain no overlapping subjects. Here $n_1 = 200$ and $n_2 = 300$. Under all simulation settings, the numbers of unknown parameters are much larger than the sample sizes.

In our analysis, of the most interest is the estimation and identification of sparsity structure for the precision matrices $\tilde{\Omega}_{YY}$ and Ω_{YY}. Three measures are adopted to measure identification accuracy, including recall (which measures the true-positive rate), false-positive rate (FPR), and Fscore (which is the harmonic mean of precision and recall). Estimation accuracy is measured using the Frobenius norm of the difference between the estimated and true precision matrices. The proposed approach has been motivated by the hierarchy. As such, we also evaluate the count and proportion of the hierarchy being violated (meaning $(\widetilde{\Omega}_{YY})_{ij} = 0$ but $(\Omega_{YY})_{ij} \neq 0$).

For comparison, we consider the separate estimation of GeO-GGM and GeR-GGM, for which we adopt the MCP penalization. For the estimation of GeR-GGM, following the reasonings described in Section 2, the partial GGM technique is adopted. Although there are potentially other approaches for estimating the graphs, comparing with the separate estimation can the most directly establish the merit of the proposed joint estimation. For the separate estimation, the same regularization parameter is adopted, and the tuning parameters are also chosen using V-fold cross validation.

Under each setting, 200 replicates are simulated. Summary statistics for the setting with a homogeneous Ω_{YY}, D_1 and D_2 containing the same 300 subjects, and $\theta = 0.1$ are presented in Table 1. The rest of the results are presented in tables 2–6 in the appendix. It is observed that, across all simulation settings, the proposed analysis outperforms the separate estimation. Consider for example the last setting in Table 1. For the estimation of $\tilde{\Omega}_{YY}$, the proposed approach has (recall, FPR, Fscore) = (0.421, 0.024, 0.471), compared to (0.406, 0.038, 0.429) of the

TABLE 1 Summary statistics on identification and estimation

	$\tilde{\Omega}_{YY}$				Ω_{YY}				Hierarchy violation	
	Identification				Identification					
	Recall	FPR	Fscore	Estimation	Recall	FPR	Fscore	Estimation	Count	Proportion
$(p, q) = (50, 50)$										
Proposed	0.454 (0.048)	0.033 (0.008)	0.532 (0.04)	20.49 (0.92)	0.552 (0.053)	0.021 (0.008)	0.62 (0.047)	39.87 (1.44)	0 (0)	0 (0)
GeO-GGM	0.454 (0.044)	0.043 (0.009)	0.511 (0.039)	27.72 (1.39)	–	–	–	–	17.78 (1.4)	17.73% (1.8%)
GeR-GGM	–	–	–	–	0.475 (0.067)	0.022 (0.012)	0.544 (0.04)	51.35 (2.06)		
$(p, q) = (50, 100)$										
Proposed	0.466 (0.038)	0.043 (0.007)	0.54 (0.036)	20.95 (0.90)	0.525 (0.041)	0.025 (0.007)	0.584 (0.037)	76.85 (2.49)	0 (0)	0 (0)
GeO-GGM	0.443 (0.03)	0.064 (0.009)	0.494 (0.031)	27.807 (1.48)	–	–	–	–	29.32 (3.75)	26.75% (5.72%)
GeR-GGM	–	–	–	–	0.447 (0.042)	0.022 (0.009)	0.524 (0.028)	90.85 (3.88)		
$(p, q) = (50, 150)$										
Proposed	0.464 (0.027)	0.048 (0.007)	0.542 (0.044)	21.07 (1.04)	0.542 (0.046)	0.027 (0.005)	0.566(0.04)	150.3 (4.87)	0 (0)	0 (0)
GeO-GGM	0.491 (0.03)	0.104 (0.038)	0.51 (0.034)	27.95 (1.04)	–	–	–	–	95.6 (13.75)	47.6% (6.8%)
GeR-GGM	–	–	–	–	0.41 (0.045)	0.01 (0.023)	0.518 (0.018)	166.9 (5.55)		
$(p, q) = (100, 50)$										
Proposed	0.468 (0.024)	0.028 (0.003)	0.496 (0.021)	61.99 (2.37)	0.474 (0.024)	0.028 (0.002)	0.499 (0.022)	114.8 (3.35)	0 (0)	0 (0)
GeO-GGM	0.41 (0.033)	0.031 (0.007)	0.439 (0.014)	84.32 (4.01)	–	–	–	–	198.2 (21.71)	27.4% (5.4%)
GeR-GGM	–	–	–	–	0.373 (0.02)	0.021 (0.004)	0.441 (0.019)	151.5 (5.04)		
$(p, q) = (100, 100)$										
Proposed	0.462 (0.023)	0.029 (0.003)	0.48 (0.02)	60.21 (1.37)	0.489 (0.024)	0.029 (0.003)	0.492 (0.02)	220 (5.85)	0 (0)	0 (0)
GeO-GGM	0.397 (0.025)	0.034 (0.01)	0.422 (0.019)	80.94 (4.41)	–	–	–	–	814.2 (85.11)	47% (6.3%)
GeR-GGM	–	–	–	–	0.355 (0.022)	0.021 (0.009)	0.426 (0.015)	269.6 (10.67)		
$(p, q) = (100, 150)$										
Proposed	0.421 (0.028)	0.024 (0.002)	0.471 (0.025)	60.23 (1.85)	0.478 (0.031)	0.024 (0.002)	0.502 (0.025)	538.6 (18.35)	0 (0)	0 (0)
GeO-GGM	0.406 (0.025)	0.038 (0.009)	0.429 (0.022)	77.79 (3.07)	–	–	–	–	3620 (365)	79.21% (5.8%)
GeR-GGM	–	–	–	–	0.45 (0.028)	0.112(0.032)	0.287 (0.007)	696.9 (40.4)		

Note: Homogeneous Ω_{YY}, $\theta = 0.1$, and D_1 and D_2 containing the same 300 samples. In each cell, mean (*SD*).

GeO-GGM. In the evaluation of estimation accuracy, the proposed approach has the Frobenius norm of the difference between the estimated and true precision matrices equal to 60.23, compared to 77.79 of the GeO-GGM. For the estimation of Ω_{YY}, the proposed approach has (recall, FPR, Fscore) = (0.478, 0.024, 0.502), compared to (0.45, 0.112, 0.287) of the GeR-GGM. In the evaluation of estimation accuracy, the Frobenius norms are 538.6 (proposed) and 696.9 (GeR-GGM), respectively. With the separate estimation, 79.2% of the hierarchy are violated. Similar findings are made with the other settings. We have also simulated data with similar structures but different parameter values and made similar observations.

Remarks: As an experiment, we simulate data with some of the nonzero elements violating the hierarchy, using the strategy described in Section 2.1.1. We observe that, for those satisfying the hierarchy, estimation, and identification results are similar to those above. For those violating the hierarchy, estimation errors are slightly inflated, and higher FPRs are observed, as expected. The overall performance is reasonable. Here we also note that, when all or the majority of the nonzero elements violate the hierarchy, the proposed approach is expected to perform unsatisfactorily. However, as this is biologically insensible as discussed in Section 2.1.1, we do not further examine this scenario. In the second experiment, we dichotomize the simulated X at the medians and create 0/1 data. The proposed approach can still be applied. However, the numerical results are much less satisfactory. As discussed above, modifications are recommended with nonnormal data.

4 | DATA ANALYSIS

We download TCGA data on two cancers from the cBioPortal (http://www.cbioportal.org/).

4.1 | Skin cutaneous melanoma (SKCM) data

Following the literature, we focus on the 395 White patients who had nonglabrous skin. Beyond gene expressions, data is also available on CNVs. Our goal is to construct the GeO-GGM and GeR-GGM analysis (with a focus on gene expressions in the latter analysis). Although in principle the proposed analysis can be conducted at a larger scale, with considerations on the limited sample size and large number of parameters, we conduct pathway-specific analysis. Specifically, we download the KEGG pathway database "c2.cp.kegg.v6.2.symbols.gmt" from the Broad Institute. This database contains information on 186 pathways, and we select the "KEGG-MELANOGENESIS" pathway, which has a top relevance for melanoma, to conduct analysis. By matching with the pathway information, we obtain 87 gene expressions and 101 CNVs. We graphically examine the marginal distributions of gene expressions and CNVs. All distributions are continuous, and the dominating majority are bell-shaped. We also conduct marginal regressions of gene expressions on CNVs. There are no CNVs seemingly with complementary effects. As such, the proposed approach can be reasonably applied.

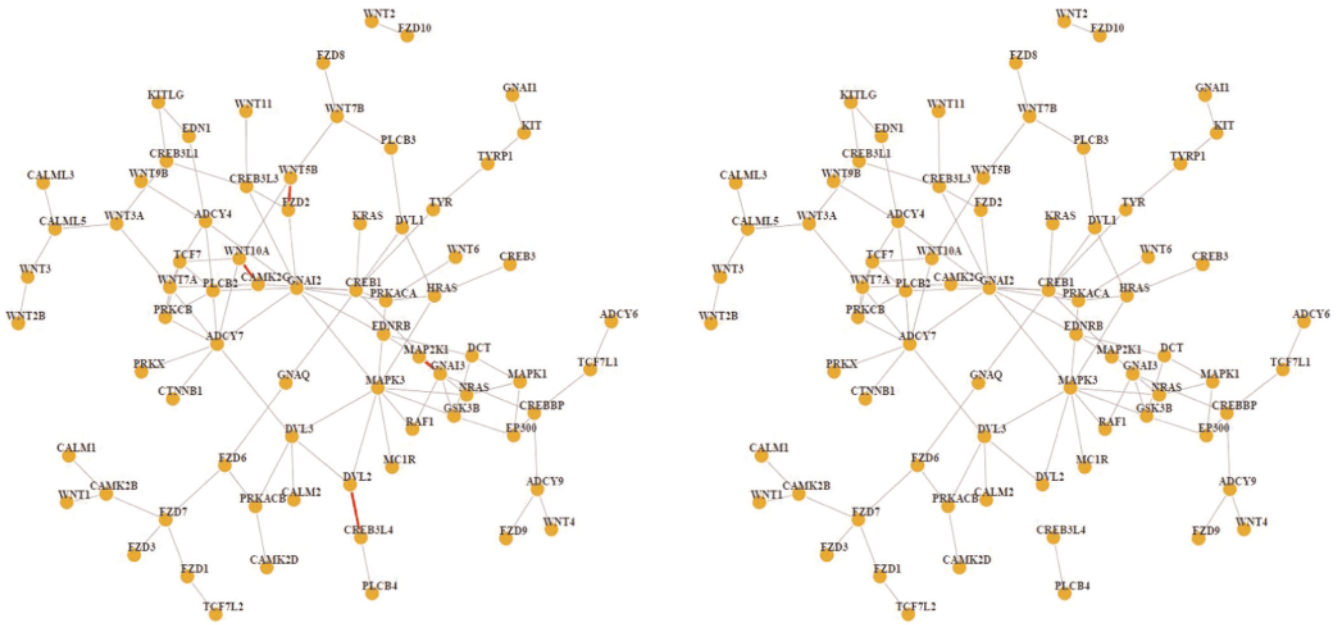

FIGURE 2 Analysis of TCGA SKCM data using the proposed approach: the GeO-GGM (left) and GeR-GGM (right) gene expression networks. Four red edges are identified in the GeO-GGM but not GeR-GGM. GeO-GGM; GeR-GGM; SKCM; TCGA

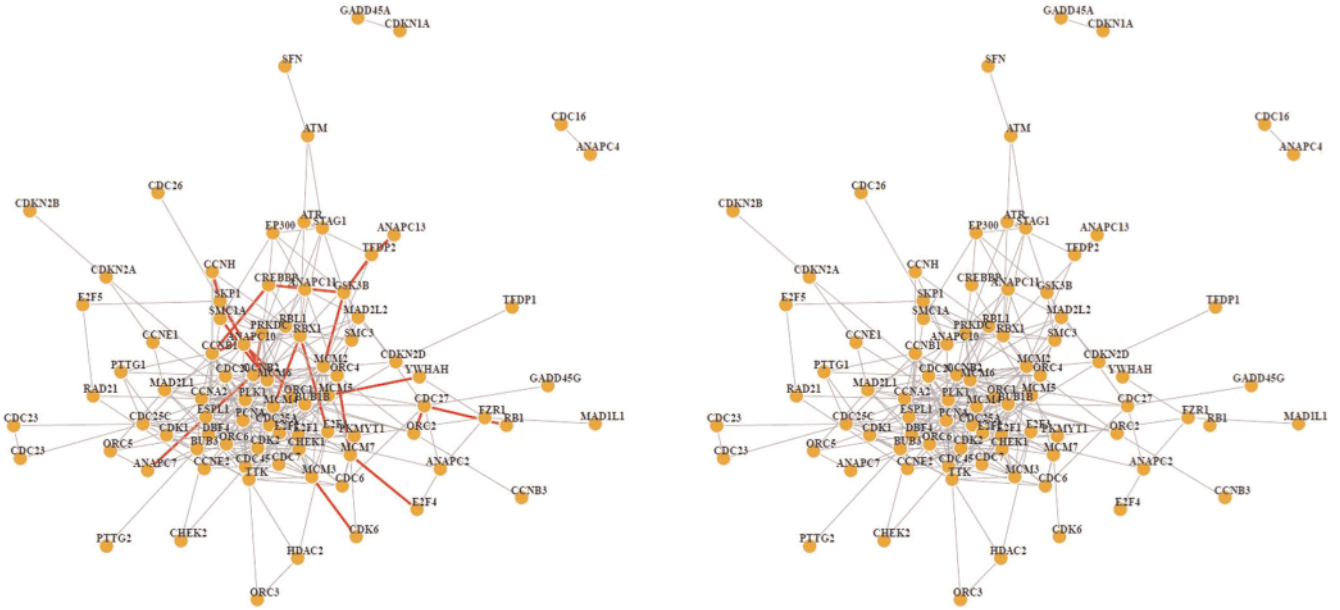

FIGURE 3 Analysis of TCGA lung cancer data using the proposed approach: the GeO-GGM (left) and GeR-GGM (right) gene expression networks. Twenty-one red edges are identified in the GeO-GGM but not GeR-GGM

We apply the proposed approach and alternative separate estimation. Tuning and regularization parameters are selected in the same manner as in simulation. Summary comparison result is presented in table 7 in the appendix. The estimated graph structures using the proposed approach are presented in Figure 2. Results using the alternative and comparison are presented in figure 4 in the appendix. For gene expressions, the proposed approach identifies 101 edges in the GeO-GGM and 97 edges in the GeR-GGM, and the hierarchy is satisfied. For the gene expression edges in the GeR-GGM with moderate to large estimates, we examine the corresponding GeO-GGM estimates and do not observe very small values, showing no alarm of hierarchy violation. For gene expressions, the separate estimation identifies 119 edges in the GeO-GGM and 99 edges in the GeR-GGM, and the edge sets differ significantly from those of the proposed approach. It identifies 76 edges in the GeR-GGM that are not identified in the GeO-GGM (i.e., violation of the hierarchy).

In network analysis, a large number of edges are estimated. In addition, the conditional connections among genes are still not fully understood. Our examination of published gene expression network studies does not suggest a well-established way of evaluating the identification results. To gain some insights, we conduct literature search and find that some gene interconnections identified by the proposed but not alternative analysis may have important biological implications. For example, genes FZD7 and CAMK2B both also belong to the Proteoglycans in cancer pathway and have been suggested as having coordinated functions. Profiling analysis has suggested that the oncogenic roles of CREB3L1/3 fusions in sclerosing epithelioid fibrosarcoma induction might be very similar. Studies have suggested the coordinated down-regulations of Calm1 and Camk2b in the $cTnT^{R141W}$ transgenic model. Genes CREBBP and GNAI3 both also belong to the molecular mechanisms of cancer pathway and have related functions. Genes CREBBP and TCF7L1 both have been identified in the pathways in cancer, which play a key role in multiple cancers. Genes GNAI3 and MAP2K1 are both associated with multiple cancer types for specific populations. Gene FZD2 is highly correlated with gene GNAI2 in the Wnt pathway. Such results, although not meant to be conclusive, can provide some support to the proposed analysis.

We further adopt a random splitting-based approach for evaluation. Specifically, the data set is randomly split into a training and a testing set with sizes 4:1. We apply the proposed and alternative approaches to the training set, and then evaluate the negative log-likelihood functions L_1 and L_2 on the testing set. This process is repeated 100 times. The average L_1 values are 82.3 (proposed) and 87.5 (alternative), and the average L_2 values are 503.7 (proposed) and 727.2 (alternative), respectively. With this random splitting approach, we are also able to evaluate the stability of identification. For the edges identified using the whole data set, we compute their probabilities of being identified in the random splits. Such probabilities have been referred to as the Observed Occurrence Index (OOI), with higher values indicating more stable estimation. For

gene expression edges, the average OOI values are 0.89 (proposed) and 0.81 (alternative) for the GeO-GGM, and 0.80 (proposed) and 0.71 (alternative) for the GeR-GGM, respectively. Overall, the proposed approach has improved estimation/prediction and stability.

4.2 | Lung cancer data

We follow the literature and focus on patients who had no neoadjuvant therapy before tumor sample collection. Data on the gene expressions and CNVs of 519 samples are available for analysis. As above, we also conduct the analysis of one KEGG pathway. Specifically, the "KEGG-CELL-CYCLE-PATHWAY," which contains genes playing important roles in cell cycle and lung cancer prognosis, is analyzed. There are a total of 102 gene expressions and 101 CNVs analyzed. The same exploratory analysis as for the melanoma data is conducted, again suggesting it is reasonable to apply the proposed approach.

Data is analyzed using the proposed and alternative approaches. As in the previous analysis, we focus on results for gene expressions. Summary comparison results are provided in Table 8 in the Appendix. The estimated graph structures are presented in figures 3 and 5 in the appendix. The proposed approach identifies 285 edges in the GeO-GGM and 263 edges in the GeR-GGM, and the hierarchy is satisfied. The separate estimation identifies 278 (GeO-GGM) and 258 (GeR-GGM) edges, with a total of 148 edges violating the hierarchy. Examining the estimates also does not raise any alarm on possible hierarchy violation. It is found that the proposed analysis can identify biologically sensible gene interconnections missed by the alternative. For example, the coordination of genes CCNH and CCNB1 has been observed in multiple studies. Genes CDC6 and CHEK1 have been suggested as coordinated. The interconnection between CCNE2 and E2F1 has been shown to play a vital role in aberrant coronary vascular smooth muscle cell proliferation. The random splitting approach as described above is applied for evaluation. The proposed approach has average L_1 and L_2 values 90.7 and 158.1, respectively, which are lower than their alternative counterparts 94.9 and 169.6. In the stability evaluation, the OOI values of the proposed approach are 0.88 (GeO-GGM) and 0.88 (GeR-GGM), compared to 0.79 (GeO-GGM) and 0.76 (GeR-GGM) of the separate estimation.

5 | DISCUSSION

In this article, we have developed a new approach that well fits the GGM framework for gene expression data but can have improved estimation/identification performance. Although loosely speaking there have been other works on information borrowing in gene network analysis, the proposed strategy of borrowing information between GeO and GeR networks is new and novel. A new hierarchy in the sparsity structures of the two networks, which is biologically sensible, has been proposed. It differs from the hierarchies identified for other omics problems (Schadt et al., 2005; Yazdani et al., 2020; Zhu et al., 2012). Along with the high dimensionality in a single model/estimation, it has led to a penalized estimation significantly different from those in the literature. Extensive and highly nontrivial methodological, theoretical, and computational developments have been conducted. The proposed analysis can flexibly accommodate multiple scenarios. Overall, this study can expand the GGM analysis paradigm and provide a practical and effective way of estimating gene expression networks.

The proposed analysis demands multidimensional profiling data, which is getting increasingly routine. It does not have strict requirements on the type and "quality" of collected regulators. In particular, it does not demand the collection of all factors that may affect gene expressions. As such, it can enjoy broad applicability. Graphical models have also been constructed for omics data other than gene expression and nonomics data. As long as there are underlying determinants for the variables of main interest, the proposed analysis can be applied. It will be of interest to systematically examine graphical modeling for nonnormal data using the proposed technique. However, literature indicates that a significant amount of separate investigation may be needed. We postpone it to future research. It may be of theoretical interest to study scenarios with regulators having completely complementary effects. However, without much practical value, it is not pursued. Although the sound biological implications and improved prediction/stability can support the validity of our data analysis to a certain extent, it is of interest but beyond our scope to independently validate the findings.

ACKNOWLEDGMENTS

We thank the editor and reviewers for their careful review and insightful comments. This study was supported by Bureau of Statistics of China (2019LZ11), Fund for building world-class universities (disciplines) of Renmin University of China, National Natural Science Foundation of China (11971404), Basic Scientific Project 71988101 of National Science Foundation of China, 111 Project (B13028), NIH (CA216017, CA241699, CA121974, CA196530), and NSF (1916251).

DATA AVAILABILITY STATEMENT

The TCGA data sets that support the findings of this study are available from the Cancer Genome Atlas Program. Data are generated by The TCGA Research Network at https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

ORCID

Yifan Sun http://orcid.org/0000-0002-4235-331X
Shuangge Ma http://orcid.org/0000-0001-9001-4999

REFERENCES

Cai, T. T, Li, H., Liu, W., & Xie, J. (2016). Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, *26*(2), 445–464.

Cancer Genome Atlas Program. (2020). National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

Chiquet, J., Mary-Huard, T., & Robin, S. (2017). Structured regularization for conditional Gaussian graphical models. *Statistics and Computing*, *27*, 789–804.

Chun, H., Chen, M., Li, B., & Zhao, H. (2013). Joint conditional Gaussian graphical models with multiple sources of genomic data. *Frontiers in Genetics*, *4*, 294.

Dobra, A., Hans, C., Jones, B., Nevins, J. R, Yao, G., & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, *90*(1), 196–212.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Huang, J., Breheny, P., & Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *27*(4), 481–499.

Jensen, F. V. (1996). *An introduction to bayesian networks*. Springer.

Li, B., Chun, H., & Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, *107*(497), 152–167.

Liu, H., Han, F., & Zhang, C. H. (2012). Transelliptical graphical models. *Advances in Neural Information Processing Systems* (pp. 800–808).

Mihaylov, I., Kandula, M., Krachunov, M., & Vassilev, D. (2019). A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. *Biology Direct*, *14*(1), 22.

Ravikumar, P., Wainwright, M. J, Raskutti, G., & Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, *5*, 935–980.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K, Monks, S., Reitman, M., Zhang, C., Lum, P. Y, Leonardson, A., Thieringer, R., Metzger, J. M, Yang, L., Castle, J., Zhu, H., Kash, S. F, Drake, T. A., ... Lusis, A. J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, *37*(7), 710–717.

Suzuki, T. (2013). Dual averaging and proximal gradient descent for online alternating direction multiplier method. *International Conference on Machine Learning* (pp. 392–400).

Wang, S., Shi, X., Wu, M., & Ma, S. (2019). Horizontal and vertical integrative analysis methods for mental disorders omics data. *Scientific Reports*, *9*(1), 1–12.

Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M. L, Sweet, R. A., Wang, J., & Chen, W. (2016). FastGGM: An efficient algorithm for the inference of gaussian graphical model in biological networks. *PLOS Computational Biology*, *12*(2), e1004755.

Williams, D. (2018). *Bayesian estimation for gaussian graphical models: structure learning, predictability, and network comparisons*. https://psyarxiv.com/x8dpr/

Witten, D., & Tibshirani, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *Journal of the Royal Statistical Society: Series B*, *71*, 615–636.

Wu, C., Zhang, Q., Jiang, Y., & Ma, S. (2018). Robust network-based analysis of the associations between (epi)genetic measurements. *Journal of Multivariate Analysis*, *168*, 119–130.

Wytock, M., & Kolter, Z. (2013). Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. *International Conference on Machine Learning* (pp. 1265–1273).

Xue, L., & Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, *40*(5), 2541–2571.

Yazdani, A., Mendez-Giraldez, R., Yazdani, A., Kosorok, M. R, & Roussos, P. (2020). Differential gene regulatory pattern in the human brain from schizophrenia using transcriptomic-causal network. *BMC Bioinformatics*, *21*(1), 469.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, *11*, 2261–2286.

Yuan, X. T, & Zhang, T. (2014). Partial gaussian graphical model estimation. *IEEE Transactions on Information Theory*, *60*(3), 1673–1687.

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

Zhao, H., & Duan, Z. H. (2019). Cancer genetic network inference using gaussian graphical models. *Bioinformatics and Biology Insights*, *13*.

Zhu, J., Sova, P., Xu, Q., Dombek, K. M, Xu, E. Y, Vu, H., Tu, Z., Brem, R. B, Bumgarner, R. E, & Schadt, E. E. (2012). Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLOS Biology*, *10*(4), e1001301.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Yi H, Zhang Q, Sun Y, Ma S. Assisted estimation of gene expression graphical models. *Genetic Epidemiology*. 2021;1–14. https://doi.org/10.1002/gepi.22377