Spring 2021

Assisted Network Analysis in Cancer Genomics

Huangdi Yi
Yale University Graduate School of Arts and Sciences, denise.huangdi.yi@gmail.com

Abstract

Assisted Network Analysis in Cancer Genomics

Huangdi Yi

2021

Cancer is a molecular disease. In the past two decades, we have witnessed a surge of high-throughput profiling in cancer research and corresponding development of high-dimensional statistical techniques. In this dissertation, the focus is on gene expression, which has played a uniquely important role in cancer research. Compared to some other types of molecular measurements, for example DNA changes, gene expressions are “closer” to cancer outcomes. In addition, processed gene expression data have good statistical properties, in particular, continuity. In the “early” cancer gene expression data analysis, attention has been on marginal properties such as mean and variance. Genes function in a coordinated way. As such, techniques that take a system perspective have been developed to also take into account the interconnections among genes. Among such techniques, graphical models, with lucid biological interpretations and satisfactory statistical properties, have attracted special attention. Graphical model-based analysis can not only lead to a deeper understanding of genes’ properties but also serve as a basis for other analyses, for example, regression and clustering. Cancer molecular studies usually have limited sizes. In the graphical model-based analysis, the number of parameters to be estimated gets squared. Combined together, they lead to a serious lack of information.

The overarching goal of this dissertation is to conduct more effective graphical model analysis for cancer gene expression studies. One literature review and three methodological projects have been conducted. The overall strategy is to borrow strength from additional information so as to assist gene expression graphical model estimation. In the first chapter, the literature review is conducted. The methods developed in Chapter 2 and Chapter 4 take advantage of information on regulators of gene expressions (such as methylation, copy number variation, microRNA, and others). As they belong to the vertical data integration framework, we first provide a review of such data integration for gene expression data in

Chapter 1. Additional, graphical model-based analysis for gene expression data is reviewed. Research reported in this chapter has led to a paper published in *Briefings in Bioinformatics*. In Chapters 2-4, to accommodate the extreme complexity of information-borrowing for graphical models, three different approaches have been proposed. In Chapter 2, two graphical models, with a gene-expression-only one and a gene-expression-regulator one, are simultaneously considered. A biologically sensible hierarchy between the sparsity structures of these two networks is developed, which is the first of its kind. This hierarchy is then used to link the estimation of the two graphical models. This work has led to a paper published in *Genetic Epidemiology*. In Chapter 3, additional information is mined from published literature, for example, those deposited at PubMed. The consideration is that published studies have been based on many independent experiments and can contain valuable information on genes' interconnections. The challenge is to recognize that such information can be partial or even wrong. A two-step approach, consisting of information-guided and information-incorporated estimations, is developed. This work has led to a paper published in *Biometrics*. In Chapter 4, we slightly shift attention and examine the difference in graphs, which has important implications for understanding cancer development and progression. Our strategy is to link changes in gene expression graphs with those in regulator graphs, which means additional information for estimation. It is noted that to make individual chapters standing-alone, there can be minor overlapping in descriptions.

All methodological developments in this research fit the advanced penalization paradigm, which has been popular for cancer gene expression and other molecular data analysis. This methodological coherence is highly desirable. For the methods described in Chapters 2-4, we have developed new penalized estimations which have lucid interpretations and can directly lead to variable selection (and so sparse and interpretable graphs). We have also developed effective computational algorithms and R codes, which have been made publicly available at Dr. Shuangge Ma's Github software repository. For the methods described in Chapters 2 and 3, statistical properties under ultrahigh dimensional settings and mild regularity conditions have been established, providing the proposed methods a uniquely strong ground. Statistical properties for the method developed in Chapter 4 are relatively straightforward and hence are omitted. For all the proposed methods, we have conducted extensive simulations, comparisons with the most relevant competitors, and data analysis. The practical advantage is fully established.

Overall, this research has delivered a practically sensible information-incorporating strategy for improving graphical model-based analysis for cancer gene expression data, multiple highly competitive methods, R programs that can have broad utilization, and new findings for multiple cancer types.

Assisted Network Analysis in Cancer Genomics

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Huangdi Yi

Dissertation Director: Dr. Shuangge Ma, Department of Biostatistics

June, 2021

Contents

List of Figures

List of Tables

Acknowledgements

The past five years have been such a joyful and extraordinarily rewarding experience for me. I sincerely acknowledge everyone who has offered me help during this journey.

First of all, I give my deepest gratitude to my advisor, Dr. Shuangge Ma, who not only led me into the world of science, but also gave me a lot of valuable advice for my career and life as well. Throughout the years, Dr. Ma has provided the best support I could ever ask for. Whenever I get stuck in a project, he is always available and willing to help. I wouldn't have achieved this without his help and support. I feel so lucky and honored to have Dr. Shuangge Ma as my advisor. His confidence in me and high expectations have been challenging me to improve myself. I am not sure if I will encounter a responsive and supportive supervisor/advisor like him in the future. What I can do is to keep his advice and encouragement in mind and carry on. I would also like to acknowledge my two other committee members, Dr. Hongyu Zhao and Dr. Yuval Kluger. I benefitted a lot from their scientific expertise. I would also like to express my sincere gratitude to Dr. Heping Zhang, Dr. Anita Wang, and Dr. Ching-Ti Liu. They all agreed to be my dissertation readers despite their busy schedules. I am very grateful and look forward to reading their comments.

I would also like to express my sincere gratitude to Dr. Christian Tschudi and Ms. Melanie Elliot. They are like the dad and mom of us, all the Ph.D. students at the Yale School of Public Health. I am so grateful that they are always being there for me, especially when I am finishing up my Ph.D. and have tons of questions about the procedure.

Equally importantly, I would like to extend my thanks to all the members in Dr. Ma's group. Dr. Qingzhao Zhang, Dr. Yifan Sun, Dr. Mengyun Wu, and Dr. Cunjie Lin have made critical contributions to my projects. I could not possibly have done it without their help. Many other lab members, both current and previous, have helped me too. I would like to thank Dr. Yuan Huang, Dr. Xiaoyan Wang, Hao Mei, Yaqing Xu for the casual but helpful discussions and talks.

I would also like to thank Silas Mullins, Wenlan Zang, Weimiao Wu, Wenhao Pan, and Yexin Yang, who are very good friends of mine. My life, especially during the pandemic, would be so boring without these friendships.

Last but not least, I want to thank my mom for her constant support. She has always supported my choices and encouraged me to pursue what I am passionate about. I would thank my cat, Mai, as well. Without him, I can't imagine how I could go through those depressed days in the pandemic. I own my success to my family including my cat. I could not thank you enough.

Chapter 1

Introduction

Genomic data provides valuable insights into cancer biology. Many quantitative genomic analyses have been done, among which network analysis has critical applications, enhancing our understanding of cancer and other complex diseases, and assisting us on the way to personalized medicine. Although the importance of network analysis has been broadly recognized, the results of existing methods are often not sufficiently satisfactory because of the high dimensionality of large-scale networks.

In recent decades, research has been extensively conducted, developing for statistical network models that can more accurately describe how genes are associated with cancer risk, progression, response to treatment, and other outcomes/phenotypes. Many genetic networks have been constructed using various statistical methods, whose findings are valuable and inspiring. However, quite often, the results are far from satisfactory because, compared to the dimensionality of human genome, information from one dataset with hundreds of observations is too limited. With the extreme complexity of cancer, it has been well recognized that a single source/type of data is insufficient, and utilizing additional information from multiple sources/types of data is needed. With the spirit of utilizing additional information, our group has taken a leading role in developing cancer modeling techniques by integrating various types of omics (genetic, epigenetic, genomic, and proteomic) data. In a series of studies, we have built integrated regression models for the prognosis and biomarkers of lung cancer, melanoma, breast cancer, and leukemia, and assisted clustering models. However, this idea has never been tested in network analysis. A critical and practically highly relevant question, which remains unanswered, is "can the usage of additional information (e.g., contained in various types of omics data and prior information) lead to more accurate network analysis."

Our ultimate goal is to build more accurate statistical models for genetic network analysis and differential network analysis by taking advantage of additional information, so as to more effectively identify gene interconnections and network changes. In this dissertation, we significantly expand the network analysis paradigm developed for genetic data and test the feasibility/necessity of using (multiple types of) regulator data and prior knowledge for cancer genetic network modeling. Taking advantage of TCGA data, we also construct network models for multiple types of cancer. This study lays the foundation for developing more advanced network methods and systematically conducting assisted network analysis in cancer genomics.

1.1 Review of vertical data integration for gene expression analysis

Gene expression data has played an essentially important role in many biomedical studies. This has been thoroughly established in a myriad of books, journal articles, and presentations. In gene expression studies, especially those with whole-genome profiling, there is usually "a large number of unknown parameters but a limited sample size" problem, leading to a "lack of information" and low-quality findings such as a lack of reliability and suboptimal modeling/prediction. One solution to this problem is data integration. The existing data integration methods mostly belong to two categories [1]. Under horizontal integration, data from multiple independent studies with comparable designs are integrated [2–5]. Under vertical integration, data on multiple types of omics measurements collected on the same subjects are integrated [6, 7]. Horizontal integration has been reviewed elsewhere [1], and in this chapter, we focus on vertical integration. We note that when data are available on multiple types of omics measurements collected on the same subjects and from multiple independent studies, it is possible to integrate in both ways, for which analysis methods are a "marriage" of those for one-way integration [8–10]. There are also studies that integrate prior information. For example, pathway information from KEGG has been extensively utilized to assist present data analysis [11–13]. Moreover, some studies [14] mine information from published studies deposited at PubMed and use that in model estimation and variable selection. However, they do not involve additionally collected data, and the methods are significantly different. As such they deserve separate reviews.

The surge in vertical data integration studies has been made possible by the growing popularity of multidimensional profiling. A representative example is TCGA (The Cancer Genome Atlas), which is a collective effort organized by the NIH and involves multiple research institutes and universities. In Table 1.1, we present the numbers of measurements on gene expressions as well as their regulators, including point mutations, copy number variations, methylation, and miRNAs, for four representative cancers including breast invasive carcinoma (BRCA), colorectal adenocarcinoma (COADREAD), kidney renal clear cell carcinoma (KIRC), and lung squamous cell carcinoma (LUSC).

Table 1.1: Numbers of measurements on gene expressions and their regulators in four TCGA datasets.

	BRCA	COADREAD	KIRC	LUSC
Gene expression	17,268	17,518	17,243	17,268
Mutation	13,414	15,998	14,054	15,273
Copy number variation	20,871	20,871	21,526	20,871
Methylation	12,328	12,328	1,678	12,328
miRNA	398	299	353	366

Vertical data integration has been motivated by the overlapping as well as independent information contained in gene expressions and their regulators. Gene expressions are regulated by the aforementioned and other regulators, leading to overlapping information. There have been extensive studies on the regulating mechanisms [15–18], although we note that the "gene expressions $\sim$ regulators" modeling is still being explored. With overlapping information, regulators can be used to "verify" findings made with gene expressions, as such, motivating data integration. On the other hand, these regulators, for example methylation, can "interact" with proteins without "passing through" gene expressions. As such, in modeling, regulators can bring additional and useful information not contained in gene expressions, thus bearing the potential of improving model fitting and prediction.

Generically, gene expression data analysis can be classified as marginal and joint [19]. Under marginal analysis, one or a small number of genes are analyzed at a time, whereas under joint analysis, a large number of genes are modeled simultaneously. It can also be classified as unsupervised and supervised. Under unsupervised analysis, no outcome/response data is involved, whereas under supervised analysis, there is an outcome/response of interest. We note that semi-supervised analysis, which is a "combination" of unsupervised and supervised analysis, is also gaining popularity, but will not be reviewed here. For general discussions, we refer to [20, 21]. Below we review data integration methods for marginal and joint analysis as well as unsupervised and supervised analysis separately.

1.1.1 Marginal analysis

Unsupervised analysis

With just a single gene (at a time) and no outcome variable, analysis has been mostly exploratory, for example examining distributional properties (mean, variance, shape, etc.). To the best of our knowledge, there is still no data integration study for this type of analysis. Our own assessment is that there is perhaps no need.

Supervised analysis

Denote Y as the outcome/response of interest, which can be continuous, categorical, or survival (subject to censoring). Denote X as the vector of gene expressions and Z as the vector of regulators. It is noted that the analysis described here and below does not require the collection of all relevant regulators. When there are multiple types of regulators, published studies [22, 23] have recommended combining them and creating a "mega" vector of regulators.

A "standard" marginal analysis proceeds as follows: (a) regress Y on one component of X, and extract the corresponding p-value; (b) conduct (a) for all genes in a parallel manner; and (c) apply the FDR (false discovery rate) or Bonferroni approach to all p-values, and identify significant genes. When regulator data are present, analysis can be revised as follows: (i) for each gene, identify its regulator(s) via analysis or from prior knowledge;

and (ii) confirm findings from the above Step (c) using regulator data. For example, a finding can be more “trustworthy” if the regulator(s) can also be significantly associated with response.

Remarks A potential problem is that the relationship between gene expressions and regulators is “m-to-m”. That is, one gene expression can be regulated by multiple regulators, and one regulator can regulate the expressions of multiple genes. This naturally demands looking at multiple gene expressions/regulators at a time and may lead to invalid marginal analysis results.

1.1.2 Joint analysis

1.1.2.1 Unsupervised analysis

Our limited literature review suggests that most analysis in this category conducts clustering, which can be on samples or genes. The goal of sample clustering is to understand population heterogeneity, identify disease subtypes, etc., whereas the goal of gene clustering is to understand gene functionalities, reduce dimensionality for downstream analysis (e.g., regression), etc. It is also possible to conduct biclustering and cluster both samples and genes. Biclustering with data integration can be potentially realized by combing methods for one-way clustering. We will not review it as studies are still limited.

Clustering samples As illustrated in Figure 1.1, two main strategies have been developed. The first strategy has been developed with the overlapping information in gene expressions and regulators in mind. Under this strategy, three categories of methods have been developed, where the key is to reinforce the same (or similar) clustering by gene expressions and regulators.

The first category contains the late integration methods mainly based on the consensus clustering techniques, such as the assisted weighted normalized cut (AWNCut) approach [23], multi-view genomic data integration (MVDA) approach [24], Bayesian consensus clustering (BayesianCC) [25], integrative context-dependent clustering (Clusternomics) [26], and Bayesian two-way latent structure model (BayesianTWL) [27]. These methods differ in the base clustering techniques, ways for extracting useful gene expression/regulator in-

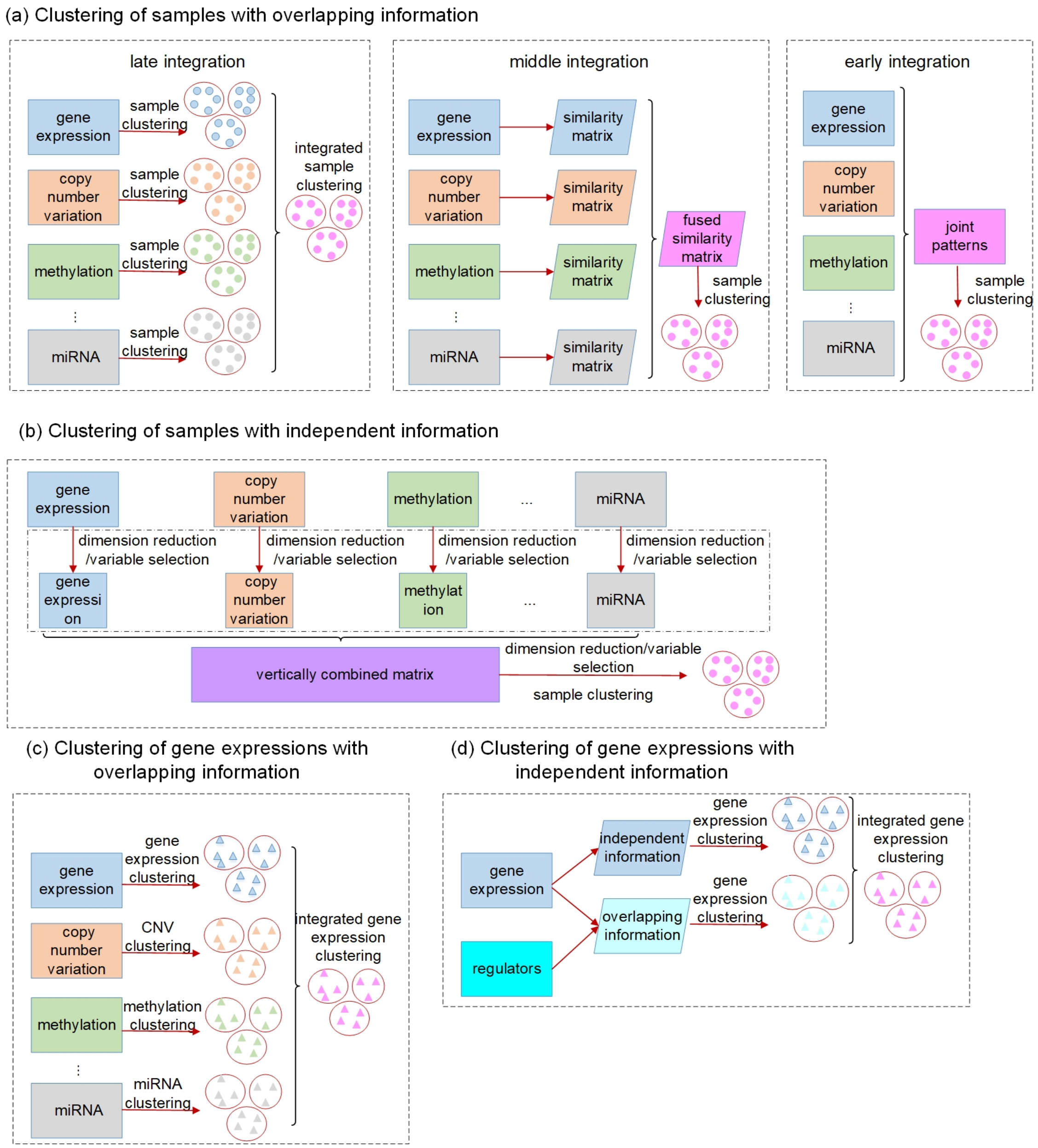


Figure 1.1: Illustration of unsupervised joint vertical integration approaches taking advantage of overlapping and independent information, respectively. CNV stands for copy number variation.

formation, and some other aspects. Here we use the AWNCut as an example to provide some insights into the strategy [23]. Denote n as the number of independent samples. First consider the "standard" NCut analysis. Compute the $n \times n$ adjacency matrices U and V, which measure the "closeness" of any two samples based on gene expressions and regulators, respectively. A simple choice is the inverse of the Euclidean distance. Denote K as the number of sample clusters, and $A_1, \cdots, A_K$ as their index sets. Using gene expression data only, the NCut approach maximizes the objective function:

$$NCut(A_1, \cdots, A_K) = \sum_{k=1}^{K} \frac{cutvol(A_k; U)}{cut(A_k, A_k^c; U)},$$

where A_k^c is the complement of A_k, $cutvol(\cdot)$ measures the within-cluster similarity, and $cut(\cdot)$ measures the across-cluster similarity. With the consideration that not all genes/regulators are equally informative, the AWNCut approach first introduces weights – genes/regulators with higher weights are more informative for clustering. Denote U_w and V_w as the weighted counterparts of U and V, respectively. The AWNCut approach maximizes the objective function:

$$\sum_{k=1}^{K} \left\{ \frac{cutvol(A_k; U_w)}{cut(A_k, A_k^c; U_w)} + \tau \frac{cutvol(A_k; V_w)}{cut(A_k, A_k^c; V_w)} + \lambda \left(\sum_j w_j^X cor\left(X_{A_k,j}, Z_{A_k,\cdot}\right) + \sum_j w_j^Z cor\left(Z_{A_k,j}, X_{A_k,\cdot}\right) \right) \right\},$$

where τ and λ are two data-dependent tuning parameters and can be selected for example using cross validation. w_j^X and w_j^Z are the jth components of the unknown weights for X and Z, respectively. $cor\left(X_{A_k,j}, Z_{A_k,\cdot}\right)$ measures the average correlation between the jth component of X and Z, computed using samples in A_k, and $cor\left(Z_{A_k,j}, X_{A_k,\cdot}\right)$ is defined similarly. It is noted that the clustering structure and weights are optimized simultaneously.

The following observations can be made with this approach and are also applicable to several other consensus clustering methods. First, the key clustering strategy and most important component – the objective function – are built on an existing single-data-type approach (in this case NCut). Second, clusterings are conducted separately using gene expressions and regulators, and consensus is fully reinforced or encouraged. Third, certain mechanisms are needed to remove noises so as to conduct clustering using only informative genes/regulators. With AWNCut, data-dependent weights are imposed, and thresholding can be employed to distinguish signals from noises. With some approaches, regularization has been directly employed for such a purpose.

The second category contains the middle integration methods, which take advantage of similarity based analysis, including the similarity network fusion (SNF) approach [28] and some others [29–31]. In particular, these methods first build similarity matrices of samples using gene expressions and regulators separately, which are often represented as graphs or networks. Fusion techniques, from as simple as average for PINS [29] and NEMO [30] to the more complex eigen-decomposition based for CoALa [31], are applied to these similarity matrices to generate a single combined similarity matrix, which is then partitioned using a conventional clustering method, such as the spectral or k-means clustering. Different from the late integration methods which directly generate cluster memberships for gene expressions and regulators separately, followed by a post hoc integration of these separate clusterings, middle integration conducts integration for similarity matrices in an earlier step.

The third category contains the early integration methods, which first detect joint patterns (overlapping information) across gene expressions and regulators, and then build a single clustering model that accounts for the generated overlapping information. In a sense, the integration is earlier than the aforementioned ones. These methods are mainly based on the joint dimension reduction techniques, among which iCluster [44, 45] is perhaps the most representative. The basic formulation of iCluster is:

$$X = W_X H + \varepsilon_X, \quad Z = W_Z H + \varepsilon_Z,$$

where H is the latent component that connects gene expressions and regulators and induces their dependencies, ε_X and ε_Z are independent "errors" for gene expressions and regulators, respectively, and W_X and W_Z are the coefficient matrices. The objective function is built on the Gaussian distribution assumption with $H \sim N(0, I)$, $\varepsilon_X \sim N(0, \Psi_X)$, and $\varepsilon_Z \sim N\left(0, \Psi_Z\right)$. To accommodate high dimensionality and identify informative genes and regulators, the Lasso penalty is imposed on W_X and W_Z. An EM algorithm is applied for optimization, and cluster memberships are then assigned by applying a standard k-means clustering on the posterior mean $E(H|X, Z)$. Similar to in late integration, regularization is usually employed for sparse estimation. Other examples include iClusterPlus [32], LRAcluster [33], moCluster [34], GST-iCluster [35], iClusterBayes [36], MOFA [37], and others.

Complementary to the first strategy, the second strategy has been developed to take advantage of the independent information in gene expressions and regulators [38–40]. As a representative example, a recent approach DLMI [40] is based on modern deep learning techniques and proceeds as follows: (a) gene expression and regulator data are stacked together and then used as the input of an autoencoder which is an unsupervised, feed-forward, and nonrecurrent neural network (NN); (b) the output of the NN produces new features, which are nonlinear combinations of the original measurements; (c) to make the analysis clinically more relevant, an outcome variable is used for supervised screening and identify marginally important features from Step (b); and (d) the selected features are used to cluster samples with the k-means approach. With this approach, gene expressions and regulators are explicitly pooled in Step (a) to gain more information. This approach is also a good showcase of data integration in the modern deep learning era.

Clustering gene expressions Our limited literature review suggests that, compared to the analysis described in the above subsection, gene expression clustering that integrates regulator data is limited. The graphical presentation is also provided in Figure 1.1.

To take advantage of the overlapping information, we conjecture that it is possible to proceed as follows: (a) for each gene expression, identify its regulators; (b) for a partition of gene expressions, compute the ordinary within-cluster and across-cluster distances; (c) partition regulators based on their associations with gene expressions and the partition in (b). Note that a regulator may belong to multiple clusters. Compute the within-cluster and across-cluster distances; and (d) compute the (weighted) sums of within-cluster and across-cluster distances from (b) and (c), and determine the clustering structure by minimizing the within-cluster distance and maximizing the across-cluster distance. This conjectured approach has been motivated by AWNCut, although we note that it has not been actually executed. And we have not been able to identify a clustering approach motivated by the overlapping information.

To take advantage of the independent information, we consider the ANCut (assisted NCut) approach [41], which is also built on the NCut technique and proceeds as follows. First consider the model:

$$X = \eta\ Z + E,$$

where η is the matrix of unknown regression coefficients, and E is the vector of "random errors" (which may also contain unmeasured or unknown regulating mechanisms). In [41], the estimate of $\hat{\eta}$ is obtained using the elastic net approach, which can accommodate the sparsity of regulations. Denote $\hat{X} = \hat{\eta}Z$ and $\widetilde{X} = X - \hat{X}$. Here a linear regression is adopted to explicitly describe that gene expression data contain information overlapping with regulator data (that is, $\hat{X}$) as well as independent information (that is, $\widetilde{X}$). Denote $\hat{U}$ and $\widetilde{U}$ as the $n \times\ n$ sample adjacency matrices computed using $\hat{X}$ and $\widetilde{X}$, respectively. Denote K as the number of gene clusters, and $A_1, \cdots, A_K$ as their index sets. The ANCut objective function is:

$$\sum_{k=1}^{K} \frac{cutvol(A_k; \hat{U})}{cut(A_k, A_k^c; \hat{U})} + \sum_{k=1}^{K} \frac{cutvol(A_k; \widetilde{U}\ )}{cut(A_k, A_k^c; \widetilde{U}\ )}.$$

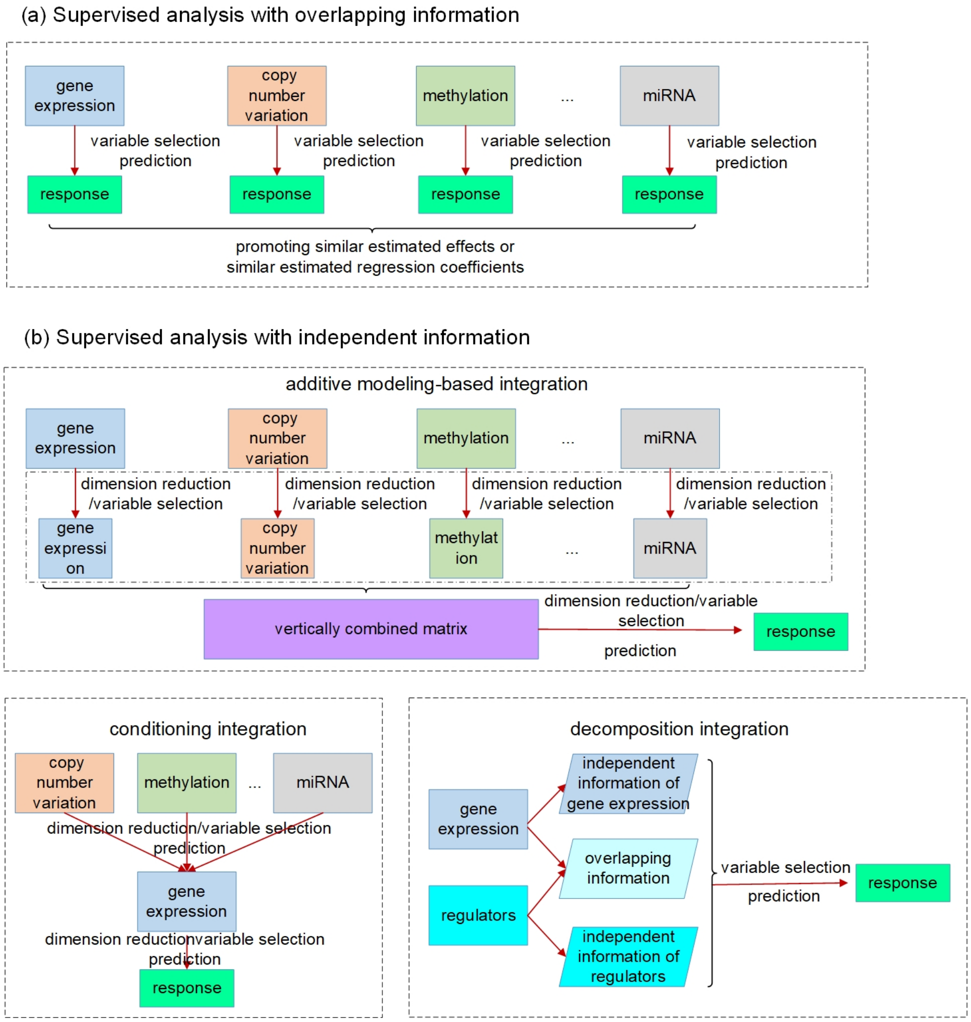
A simplified version, which is suggested as equivalent, has also been developed [41]. The essence of this approach is to first decompose gene expressions into two components and then reinforce that they generate the same clustering results.

Remarks The aforementioned clustering techniques generate disjoint clusters. In the clustering of samples, clustering of gene expressions, and biclustering, fuzzy techniques [42–45] have been developed to allow samples/genes to belong to multiple clusters or not be clustered. Data integration in fuzzy clustering remains limited and may warrant more exploration.

1.1.2.2 Supervised analysis with sparsity

For a specific outcome/response, it is usually true that many or most genes are "noises", demanding certain sparsity in analysis. Sparse results are also more interpretable and more actionable. The strategies of the supervised integration approaches are illustrated in Figure

1.2.


(a) Supervised analysis with overlapping information

(b) Supervised analysis with independent information

gene expression
copy number variation
methylation
...
miRNA
variable selection
prediction
response
promoting similar estimated effects or similar estimated regression coefficients
additive modeling-based integration
dimension reduction /variable selection
gene expression
copy number variation
methylation
miRNA
vertically combined matrix
dimension reduction/variable selection
prediction
response
conditioning integration
copy number variation
methylation
miRNA
dimension reduction/variable selection
prediction
gene expression
dimension reductionvariable selection
prediction
response
decomposition integration
gene expression
regulators
independent information of gene expression
overlapping information
independent information of regulators
variable selection
prediction
response

Figure 1.2: Illustration of supervised joint vertical integration approaches taking advantage of overlapping and independent information, respectively.

Analysis that takes advantage of the overlapping information A well-known representative is collaborative regression (CollRe) [46], which is motivated by the unit-rank canonical correlation analysis. Consider the case with a continuous Y and the model $Y = \beta^\top X + \epsilon$, where β is the vector of unknown regression coefficients and ϵ is the random error. Use subscript i to denote the ith sample. With the Lasso estimation, the objective function is:

$$\sum_{i=1}^{n} (Y_i - \beta^\top X_i)^2 + \lambda|\beta|,$$

where λ is the data-dependent tuning parameter and the l_1 norm is defined as the sum of component-wise absolute values. Following the same strategy, a model can be built using the regulators, and denote the corresponding regression coefficient vector as γ. The collaborative regression approach considers the objective function:

$$\sum_{i=1}^{n} \left(Y_i - \beta^\top X_i\right)^2 + \lambda\left|\beta\right| + \sum_{i=1}^{n} \left(Y_i - \gamma^\top Z_i\right)^2 + \lambda\left|\gamma\right| + \tau \sum_{i=1}^{n} \left(\beta^\top X_i - \gamma^\top Z_i\right)^2,$$

where τ is another data-dependent tuning parameter. This approach explicitly builds two regression models. The key advancement is the last penalty term, which encourages gene expressions and regulators to generate similar estimated effects.

Motivated by the successes of approaches that explicitly model the gene-regulator relationship and possible long-tailed distribution/contamination of the response data, the ARMI (assisted robust marker identification) approach is developed [47]. Specifically, still consider the linear gene expression-regulator model as in Section 3.1.2. In [47], $\hat{\eta}$ is obtained using the Lasso approach. The ARMI approach has objective function:

$$\sum_{i=1}^{n} \left|Y_i - \beta^\top X_i\right| + \lambda\left|\beta\right| + \sum_{i=1}^{n} \left|Y_i - \gamma^\top Z_i\right| + \lambda\left|\gamma\right| + \ \tau \times \left|\beta^\top \hat{\eta} - \gamma\top\right|.$$

Different from collaborative regression, it promotes the similarity of regression coefficients for gene expressions and regulators, as opposed to the estimated effects. In addition, the l_1 loss functions are adopted, which leads to robustness and simplified computation (as all terms are l_1).

Remarks With both collaborative regression and ARMI, the goodness-of-fit functions can be replaced by negative likelihood functions to accommodate other models and data distributions. For example, a followup study [48] extends collaborative regression and develops canonical variate regression (CVR) which can handle multivariate and non-continuous outcomes and allows for multiple-rank modeling. For these two approaches and those described below, the original publications have assumed homogeneity. We conjecture that they can be extended and coupled with the FMR (finite mixture of regression) technique [49, 50] to accommodate heterogeneity. In addition, they have been described with only the additive effects of omics measurements. In practical data analysis, demographic/clinical/environmental variables, which are usually low-dimensional, can be easily incorporated. We conjecture that it is possible to extend the approaches aforementioned and below to accommodate gene-environment interactions [51, 52], although our literature search shows that this has not been pursued.

Analysis that takes advantage of the independent information Conceptually, the most straightforward approach is to pool all omics measurements together and use as input to, for example, penalization estimation and variable selection. As different types of omics data have significantly different dimensionalities and distributional properties, this simple approach barely works in practical data analysis. To tackle this problem, IPF-LASSO proposes using different penalty parameters for different types of predictors [53]. As an "upgrade", the additive modeling approach first applies for example Lasso to each type of omics data separately and identifies a small number of features [54, 55]. The selected features, which have much lower dimensions, are pooled and modeled in an additive manner. The most significant advantage of this approach is simplicity. On the other hand, there is no distinction between gene expressions and regulators.

The conditioning-integration approach has been designed to account for the "order" of omics measurements. That is, compared to regulators, gene expressions are "closer to" outcome/response. This approach proceeds as follows: (a) conduct analysis with gene expression data only, using a "standard" high-dimensional sparse approach, for example Lasso. With this step, the dimensionality of gene expressions is reduced to one; (b) conditional on the one-dimensional gene expression effect, integrate one type of regulator data. This can be achieved using the same approach as in (a); (c) conduct (b) with all types of regulator data (if applicable), and select the type with, for example, the best prediction performance, and integrate; (d) repeat (c) until there is no significant improvement in prediction or all regulator data have been integrated. A significant advantage of this approach is that it does not demand new methodological and computational development. It can also generate a "ranking" of regulator data, facilitating biological interpretations. On the other hand, it does not take full advantage of the regulation relationship.

Overlapping information may be statistically manifested as correlation, which may challenge model estimation. The decomposition-integration approach explicitly exploits the regulation relationship and can effectively eliminate correlation. A representative example is the LRM-SVD approach [22], which proceeds as follows: (a) consider the regulation model $X = \eta\, Z + E$, and denote $\hat{\eta}$ as the estimate of η. In [22], estimation is achieved using Lasso. (b) Conduct sparse SVD (singular value decomposition) with $\hat{\eta}$. Specifically, the first step is conducted by minimizing the objective function:

$$||\hat{\eta} - \lambda \times u^\top v||_2^2 + \tau\left(|u| + |v|\right),$$

where λ is the first singular value, and u and v are singular vectors with the same dimensions as X and Z, respectively. $\hat{\eta}$ is then updated, and the subsequent steps can be conducted in a similar manner. (c) With each sparse SVD, Step (b) leads to rank-one subspaces of X and Z (which are linear combinations of a few components of X and Z, corresponding to the nonzero components of the singular vectors). These rank-one subspaces have been referred to as the "linear regulatory modules (LRMs)" and include co-expressed gene expressions and their coordinated regulators. Denote the collection of such subspaces as X_O. (d) Project X and Z onto X_O, and denote the "residuals" as $\widetilde{X}$ and $\widetilde{Z}$. This is realized using matrix projection operations. (e) Consider the outcome model $Y \sim f(\beta^\top X_O + \alpha^\top \widetilde{X} + \gamma^\top \widetilde{Z})$. In [22], survival data and the accelerated failure time model are considered. Denote $l(\beta, \alpha, \gamma)$ as the lack-of-fit function. The final estimation and variable selection can be achieved by minimizing:

$$l\left(\beta, \alpha, \gamma\right) + \lambda(|\beta| + |\alpha| + |\gamma|).$$

The three decomposed components have lucid interpretations. The LRMs, besides serving as the building blocks for model fitting, can also facilitate understanding biology. In addition, through projection, the three components are statistically independent, facilitating estimation.

1.1.2.3 Supervised analysis without sparsity

The approaches reviewed in Section 3.2 and those alike make the sparsity assumption. In practical data analysis, they usually select only a few gene expressions (and regulators). It has been proposed that there may be many weak signals, which cannot be accommodated by sparse approaches. When biological interpretation is of secondary concern, dense approaches that can accommodate many genes may be advantageous. Studies have suggested that some "black-box" approaches may excel in prediction.

With the additive modeling and conditioning-integration techniques discussed in Section 3.2, dense dimension reduction approaches, such as PCA (principle component analysis), PLS (partial least squares), ICA (independent component analysis), and SIR (slice inverse regression), can be applied as building blocks to accommodate high dimensionality [6]. Examining the decomposition-integration technique suggests that it is designed to be sparse. We have not identified a dense approach that adopts this technique.

In recent studies, deep learning techniques have also been adopted for supervised model building and prediction. Here we note that for data with low-dimensional input and a large number of training samples, the superiority of deep learning in prediction has been well demonstrated. However, the message is less clear with high-dimensional omics data. As a representative, a recent deep learning approach HI-DFNForest [56] proceeds as follows: (a) For gene expression and each type of regulator, data representations are learned separately. This can be achieved using fully connected NNs, although our personal observation is that those with regularization (for example, Lasso) may be more reliable. (b) All the learned representations are integrated into a layer of autoencoder to learn more complex representations. (c) The learned representations from (b) are fed into another NN for the outcome/phenotype. For continuous, categorical, and censored survival outcomes, NNs with various complexity levels have been developed in the recent literature, including MVFA [57], SALMON [58], MDNNMD [59], and others.

Remarks The line between sparse and dense approaches is becoming blurring. Hybrid approaches have been developed, with the hope to "inherit" strengths from both families of approaches. For example, in a study of the gene expression-regulator relationship [60], a sparse canonical correlation analysis approach is developed, which applies the Lasso penalization to correlation analysis. Other examples include the joint and individual variation explained method [61] and penalized co-inertia analysis [62]. In supervised model building, the SPCA (sparse PCA) and SPLS (sparse PLS) techniques have been applied [55,63].

1.1.3 Discussion

Most of the reviewed approaches, for example AWNCut, collaborative regression, conditioning-integration, and many alike in published literature, have roots deep in the existing methods for gene expression only. There are only a few, such as the decomposition-integration approach, that directly take a system perspective. More developments are needed to directly start with the gene expression-regulator system.

Most of the reviewed approaches have been based on penalized variable selection and dimension reduction, which are arguably the most popular high-dimensional techniques. There have also been developments using other techniques, especially including Bayesian, thresholding, and boosting. For example, the iBAG approach [64], which adopts the decomposition-integration strategy, has been developed using the Bayesian technique. With the complexity of omics data, it is unlikely that one technique can beat all. It is of interest to expand the aforementioned studies using alternative techniques and comprehensively compare (for example, consensus clustering using the NCut technique against k-means).

It is indisputable that regulator data contain valuable information. However, in any statistical analysis with a fixed sample size, regulator data contain both signals (which are unknown and need to be identified data-dependently) and noises. Conceptually, if signals overweigh noises, then data integration is worthwhile. However, theoretically, there is still a lack of research on the sufficient (and possibly also necessary) conditions under which data integration is beneficial. We conjecture that this is related to the level of signals, number/ratio of signals, and analysis techniques. There have been a few studies conducting numerical comparisons. For example, in [6], with survival data, the models with gene expression only are compared against those integrating regulators including copy number variation, methylation, and miRNA using C-statistics. Conflicting observations are made across diseases/datasets, further demonstrating the necessity of more statistical investigations on the benefit of data integration.

The reviewed approaches and many in the literature focus on gene expressions and their upstream regulators. In the whole molecular system, there are also proteomic and metabolic measurements. It is possible to further expand the scope of data integration. One possibility is to keep the central role of gene expressions and use downstream data to assist gene expression analysis. For example, multiple studies have used protein-protein interaction information in gene expression data analysis [65, 66]. The second possibility is to consider gene expression as an intermediate step and directly model the whole system. For example, in [67], clustering analysis (MuNCut) is conducted on the "protein-gene expression-regulator" system and identify molecular channels.

Our review has been focused on bulk gene expression data, where, for a specific gene, the measurement is the average of transcription levels within a cell population collected from a biological sample. In the past few years, single cell RNA sequencing (scRNA-seq) is getting increasingly popular. It advances from bulk RNA-seq by measuring mRNA expressions in individual cells and can provide more comprehensive understanding of complex heterogeneous tissues, dynamic biological processes, and other aspects [68]. Parallel single cell sequencing techniques have also been developed for the joint profiling of single cell transcriptome and other molecular layers, such as genome [69], DNA methylation [70], and chromatin accessibility [71], on the same cells, making it potentially possible to conduct data integration at the single cell resolution [72]. Single cell data usually has the count nature and exhibit strong amplification biases, dropouts, and batch effects due to unwanted technical effects, tiny amount of RNA present in a single cell, and other reasons [73], posing tremendous challenges to statistical analysis. The integration approaches reviewed above do not account for these characteristics and cannot be applied to single cell data directly. We conjecture that it is possible to build the single cell counterparts of the review methods. However, significant methodological developments will be needed. The limited existing vertical integration approaches for single cell data include the coupled nonnegative matrix factorization for the clustering of cells [74], multi-omics factor analysis v2 (MOFA+) [75] which is the extension of the unsupervised sample clustering approach MOFA [37], and a few others.

In data integration, higher dimensionality inevitably brings computational challenges. This is multi-faceted. First, it increases data storage and manipulation burden. This can be especially true when, for example, genome-wide SNP data is present. In practical data analysis, pre-processing is usually conducted to significantly reduce dimensionality and hence computational challenges. For example, SNP data can be aggregated to gene-level data [76], or supervised screening can be applied to select the most relevant ones for downstream analysis [6, 22, 55]. This way, the increase in storage and manipulation burden can be moderate. Second, some methods demand the development of new computational algorithms. For example, AWNCut introduces weights, which need to be optimized along with cluster memberships. The decomposition-integration approach LRM-SVD demands a more effective way of conducting sparse SVD. Fortunately, in the reviewed studies, computational algorithms have been developed by "combining" existing techniques. For example, with AWNCut, the simulated annealing technique is repeatedly applied. Deep learning-based integration approaches have taken advantage of the existing algorithms/tools, such as the Keras library [40], TensorFlow [59], and others. Overall, the demand for new computational algorithms has been "affordable". Third, increased dimensionality reduces computational stability. In some studies [23], random-splitting approaches have been applied to evaluate stability. However, there is still a lack of study rigorously quantifying the loss of stability, and whether that can be "compensated" by for example the improvement in prediction.

1.2 Review of network analysis

1.2.1 Significance of genetic network analysis

Omics data have brought valuable insights into quantitative research on many complex diseases [77–79]. Although the associations between different outcomes and individual genes have been widely studied, many individual interconnections studies lack a system prospective. Compared to gene-based research and individual interconnection analysis, genetic network analysis provides an advanced means to illuminate how genes function systematically for complex diseases.

Genetic networks are a representation in which nodes represent genes and edges represent interconnections among genes. Genetic network analysis has attracted a lot of interest, and many methods have been proposed for statistical inference from GE data since the advent of high-throughput sequencing [80–82]. It has critical applications in biological and medical sciences, enhancing our understanding of complex diseases, and assisting us on the way to personalized medicine.

Genetic networks provide important information about gene-gene interconnections such as regulatory associations between regulating genes and their potential targets, which can help solve different biological and biomedical problems. An important application is that genetic networks represent systematically statistical significance of molecular interconnections obtained from high-throughput data. Given a large number of potential gene interconnections among approximately 25,000 genes, the construction of genetic networks largely narrows down the number of connections and pinpoints these critical ones from noisy data. For instance, Butte et al. constructed 202 relevant (sub-)networks from 11,692 genes in 60 cancer cell lines, and some of the network clusters/pathways were found related to different biological functions [83]. Another representative example is the weighted correlation network analysis (WGCNA) [84], based on which many studies have been conducted to construct gene co-expression networks, identify modules, and hub genes. For example, Clark et al. built a genetic network using breast cancer samples from 13 microarray-based GE studies and identified 11 coregulated gene clusters. Most of these transcriptional modules were found to be correlated with tumor grade, survival endpoints for breast cancer, and also its molecular subtypes [85]. Many other findings have also shown that genetic network analysis can facilitate the identification of key biomarkers of cancer and various other complex diseases [86–88].

Not only individual identified genes can be used as biomarkers, but it has been argued that a network itself can also be considered as a biomarker for diagnostic, predictive, or prognostic purposes [89, 90]. This is reasonable especially for complex diseases like cancer, as the characteristics of cancer are represented by interconnected genes with complex "interactions" [91]. For example, using GE data, Yang et al. analyzed the genetic networks of four representative cancer types and showed that prognostic genes in genetic networks have common system-level properties [92]. This study and those alike suggest that we can potentially conduct more accurate prognosis and other analysis if we can account for gene network information in a comprehensive and effective manner. In another relevant study, Dehmer et al. used eigenvalues and entropy-based network measures as biomarkers and demonstrated that they outperform conventional biomarkers using GE data [90]. When more and more established networks from different diseases become available, together with clinical data and drug-dose response information, it will be possible to lead the charge to more personalized medicine [93].

A beneficial concomitant of the increasing availability of genetic networks is the growing possibility of differential network analysis comparing multiple networks from different populations or groups. This will allow us to learn about how interconnections change across various time courses or disease conditions and enrich our biomedical understanding [94]. For example, Islam et al. conducted a computational analysis of published protein interaction networks [95]. In their study, cancer protein interaction networks show a higher level of clustering, or molecular complexes, than the normal ones for all tissues. These networks further predicted some major molecular complexes that might act as the important regulators in cancer progression and potential drug targets.

In conclusion, genetic network analysis is of great importance for solving many different biological and biomedical problems. As the advent of the omics era, when GE and other genetic variants data are becoming increasingly available, it is the appropriate time to develop statistical methods for more advanced genetic network analysis.

1.2.2 Network methods

Compared to the methods mentioned in Section 1.1, genetic network analysis is timely and more informative because it considers gene interconnections in a more systematical way [96]. A distinction is made between undirected networks, where edges link two nodes symmetrically, and directed networks, where edges can be directional [97]. In biomedical studies, undirected networks are often adopted because many types of relationships between two biological entities (e.g., gene co-expression and protein binding) are symmetrical [98]. A network is fully specified by its adjacency matrix, a symmetric matrix whose components encode the network connection strength between nodes. For an unweighted network, each entry in the adjacency matrix is either 1 or 0, representing there is an edge or not. Weighted networks allow the adjacency to take on continuous values between 0 and 1, which are defined by gene similarity.

Statistical methods for genetic network construction can be divided into two families, for identifying unconditional associations and conditional associations. WGCNA and its successors belong to the first category. The connection in a gene co-expression network is often a measure of correlation, mostly commonly Pearson correlation coefficient [99, 100]. This measure describes marginal, linear relationships between genes, i.e., every pair of genes is considered alone, ignoring the presence of all remaining genes. The resulting networks are sometimes very dense, and the natural interpretation of the edges has had only limited success in identifying therapeutic targets [101]. This is partly due to the fact that gene co-expression networks focus only on marginal dependency, and can neither provide a systemic perspective conditional on other genes nor incorporate valuable information from multi-omics data [102, 103].

Studies on network methods for revealing conditional dependence are promising and prosperous. In recent years, there have been many conditional dependency network inference methods [104–106], such as Gaussian graphical models (GGMs) [107], Bayesian networks [108], and Boolean networks [109]. Among them, GGMs are especially attractive because the assumption is intuitive and simple, and they process superior statistical properties. If GE profiles follow a multivariate Gaussian distribution, two genes have a non-zero partial correlation if and only if they are conditionally dependent given other genes, which is, if the corresponding element in the inverse of their covariance matrix, i.e., the precision matrix, is non-zero [107]. GGM may produce a more parsimonious graph than some co-expression networks [110]. This parsimonious graph, in other words, the sparse precision matrix, can be estimated by maximizing a penalized log-likelihood function. For the optimization problem, the desired properties of network can be enforced by restricting the solution space or by constructing an appropriate penalty. Researchers have exploited this flexibility, resulting in diverse literature on analyzing GE data using GGMs [111–113]. However, most GGMs are not comprehensive or informative enough because they cannot incorporate regulator information.

Relationships of genes are often affected by regulator variations, such as CNVs and methylation. When additional information on regulators is available, there are studies on identifying the dependency networks of genes after "removing" the effect of regulators. For example, one may want to infer the gene network incorporating all external variables as well, since the relationships of genes are often affected by external variables (e.g., genetic variations), and gene regulatory relationships may be altered under different conditions such as tissue types. The conditional Gaussian graphical models (cGGMs) have been introduced to achieve this goal. In the studies of cGGMs, multivariate regression has been widely used. Yin and Li proposed a sparse cGGM for studying the conditional dependent relationships among a set of GEs adjusting for possible genetic effects [114]. Yuan and Zhang developed a partial Gaussian graphical Model (pGGM) and showed that it is essentially a regularized conditional maximum likelihood estimator for the regression model [115]. Different from cGGMs, the pGGM approach directly estimates blocks of the full precision matrix via a convex formulation, while the log-likelihood objective function of a cGGM is not convex but biconvex. This significantly simplifies the computational procedure and statistical analysis. For more related studies, we refer to [116–119]. All the aforementioned papers have been focused on either the construction of one network or the construction of several similar networks of the same type, but no comparison has been made between these related but different genetic networks. More specifically, there is a hierarchical structure between the network that is constructed based on GE only and that with both GE and regulators. In particular, the interconnection of two genes can be caused by many reasons, one of which is that these genes are regulated by the same regulators. So when we "remove" the effect of regulators, the interconnection/edge between these genes should disappear if it is only caused by these regulators. In other words, there is a monotone change in edges of the GE network after removing the effects of regulators. However, none of the existing methods has utilized this hierarchical structure. Our development in Chapter 2 will fill this knowledge gap.

Another common limitation of the existing methods is that they fail to take into account existing "prior information". In statistical literature, prior information almost "automatically" leads to Bayesian analysis. Our literature review suggests that Bayesian network analysis remains limited. One example is by Gevaert et al. [120], which developed several Bayesian networks with expert information and used them to predict ectopic pregnancy. Gaussian graphical models have also been estimated using Bayesian techniques [121, 122]. Bayesian techniques for network analysis have a few limitations. They are often difficult to compute; The adopted priors can be somewhat subjective – this is especially true when computation is a major concern; And it is difficult to "customize" priors for different edges – imposing the same prior for all edges may not be sensible as we have extensive knowledge on some edges but little to none on others. Our development in Chapter 3 will fill this knowledge gap.

When there are two or more different conditions/groups (for example, disease stage), differential analysis can be of significant interest. This is also true when the quantity of interest is network. Differential network analysis can especially suffer from "a lack of information" as at least two networks need to be estimated, which increases the number of parameters even more. In Chapter 4, we will continue developing assisted analysis. We will consider the scenario where the goal is to identify the key contributors (genes) to the difference of two (or more) GE networks, when data is also available on regulators. Our strategy is to take advantage of regulator information and GE-regulator relationship to improve the accurate of GE identification. This strategy has been partly motivated by early developments by Dr. Ma's group [23] and also contains further development.

Taken together, although the significance of genetic network analysis has been broadly recognized, the results are far from satisfactory. Part of the reason is that the sample size is relatively small compared to the gene dimensionality. This gap can be partially filled if we can take advantage of abundant information from regulators and existing studies. As such, the purpose of this dissertation is to propose different techniques with different strategies for using additional information to improve genetic network analysis.

1.3 Summary of Significance

GE data provides valuable insights into cancer biology. Many quantitative analyses have been done, among which network analysis has critical applications, enhancing our understanding of disease, and helping us on the way to personalized medicine. Although the importance of network analysis has been broadly recognized, the results of existing methods are not fully satisfactory, mainly because of a "lack of information" caused by the high data dimensionality and small sample size. We see a strong need for more effective methods which can harness "additional information" in regulators and published studies to improve genetic network analysis. In the following chapters, we conduct extensive statistical and numerical studies, which significantly advances the assisted analysis paradigm developed by our group and others to the more complex network analysis. Methodology developments in genetic network analysis is fundamentally meaningful – it lays the foundation for future network analysis and broader data analysis on other cancer types. Taking advantage of TCGA data, this study also has a significant impact on cancer research.

Project 1: Using information on regulators to improve network construction. The aforementioned hierarchical structure between the genetic network based only on GE data and the network that also incorporates regulators of GEs has not yet been taken full advantage of in the estimation of gene networks. Considered that more and more information about regulators is available, we see a great potential of improving network construction if we can develop a novel approach to effectively use the information.

Project 2: Using information from publications to improve network construction. Similar ideas have been brought to regression analyses in [123] and [14], but there is not such a statistical method to comprehensively incorporate prior information in genetic network analysis. Therefore, our second project is to improve the estimation of network structures using additional information from prior knowledge especially by mining published studies.

Project 3: Using information on regulators to improve differential network analysis. Since there are regulations between GEs and regulators, it is meaningful to incorporate regulator information when we look for genes that primarily contribute to the change of networks. Regulator information has been exploited in regression analyses and clustering, but has not been brought to differential network analysis. It is challenging and of interest to adapt and advance this technique in the complex differential network analysis.

Innovation

1. The framework of this dissertation is novel because it contains a series of strategies aiming to take advantage of various information, leading to more accurate and dependable findings in genetic networks, significantly advancing the network analysis paradigm.

2. The first method innovatively uses a penalization model to jointly estimate a gene-expression-only GGM and a gene-expression-regulator GGM, so that it can exploit information on the hierarchical structure in genetic networks. The second method brings forward an innovative way to incorporate prior knowledge in network construction. The third method creatively utilizes the change of regulatory network when estimating the change of genetic network. All methods are biologically well motivated while having a strong statistical ground.

3. The proposed methodological advances adequately tackle the limitations mentioned above in current methods for genetic network analysis. Also, we prove the theoretical properties of the proposed methods so that this study enjoys high statistical rigor.

4. Intensive numerical studies including simulations and real data analyses are conducted. Extensive comparisons are made between our results and existing findings. This dissertation has a high likelihood of success and significant practical impact.

Chapter 2

Project 1: Assisted estimation of gene expression graphical models

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. Here we use GeO-GGM to represent gene-expression-only GGM. Gene expressions are regulated by regulators. GeR-GGMs (gene-expression-regulator 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 TCGA datasets are analyzed, leading to findings different from

the direct competitors. Research reported in this chapter has been published in *Genetic Epidemiology.*

2.1 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 study, 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) [124], Wang et al. (2016) [125], Zhao and Duan (2019) [126], 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 [127, 128] 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 Yuan and Zhang (2014) [115], Ravikumar et al. (2011) [129], and Suzuki (2013) [130] 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 [110] and Bayesian [131]. In this article, we focus on the probabilistic estimation, which may be more popular.

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, GeR-GGMs (gene-expression-regulator GGMs) have been constructed [119], 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. (2014) [119] 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 [118] and other types of biological data, and refer to [116, 132], 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 (that is, 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 [110], which applies Lasso penalization in GGM estimation. Beyond Lasso, other penalization approaches and approaches based on other regularization techniques have also been developed [132, 133]. 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 datasets that share certain similarity and jointly estimate multiple GeO-GGMs (or GeR-GGMs) [134]. There are also studies that borrow information from prior knowledge, for example, functional annotations of genes or published findings [9].

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 [124–126]. 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 [14]. 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 TCGA data. Overall, this study can provide a practical and useful new venue for gene expression network analysis.

2.2 Methods

2.2.1 Strategy

Consider gene expressions G1, G2, and G3, and regulator R (which can be multi-dimensional). In a gene-expression-only 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) co-regulation by R. If G1 and G2 are both regulated by R, then they can be interconnected; (b) co-regulation 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 gene-expression-only graphical model contains *all-causes gene interconnections*, whereas a gene-expression-regulator 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 gene-expression-only graphical model and its gene-expression-regulator counterpart. For a gene-expression-only 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 gene-expression-regulator graphical model, we can potentially improve its identification and estimation of gene connections. The same applies to the gene-expression-regulator 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 co-regulations 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 gene-expression-only network analysis, with no gain of information from regulators but also no loss.

Remarks

Identifying biologically motivated hierarchy to assist data analysis is by no means new. Examples include [135–137], 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 [16], 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:

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

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.

2.2.2 Assisted estimation

Let $Y = (Y_1, \cdots, Y_p)^\top$ denote p gene expressions and $X = (X_1, \cdots, X_q)^\top$ denote q regulators. With multiple types of regulators, their measurements can be stacked together. Consider a dataset $D_1 = \{\mathbf{y}\}_{i=1}^{n_1}$ with n_1 i.i.d. copies of Y and a dataset $D_2 = \{(\mathbf{y}, \mathbf{x})\}_{i=1}^{n_2}$ 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 $Y_i \perp\!\!\!\perp Y_j | Y_{-(i,j)} \Leftrightarrow \widetilde{\Omega}_{ij} = 0$, 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:

$$Q(\widetilde{\Omega}_{YY}, \Omega_{YY}, \Omega_{YX}) = L_1(\widetilde{\Omega}_{YY}) + L_2(\Omega_{YY}, \Omega_{YX}) + P_1(\widetilde{\Omega}_{YY}, \Omega_{YY}) + P_2(\Omega_{YX}), \qquad (2.1)$$

where

$$L_1(\widetilde{\Omega}_{YY}) = -\log\det(\widetilde{\Omega}_{YY}) + \mathrm{tr}(\widetilde{S}_{YY}\widetilde{\Omega}_{YY}),$$

$$L_2(\Omega_{YY}, \Omega_{YX}) = -\log\det(\Omega_{YY}) + \mathrm{tr}(S_{YY}\Omega_{YY}) + 2\mathrm{tr}(S_{YX}^\top \Omega_{YX}) + \mathrm{tr}(S_{XX}\Omega_{YX}^\top \Omega_{YY}^{-1}\Omega_{YX}),$$

$$P_1(\widetilde{\Omega}_{YY}, \Omega_{YY}) = \sum_{i \neq j} \rho(\sqrt{(\widetilde{\Omega}_{YY})_{ij}^2 + (\Omega_{YY})_{ij}^2}; \lambda_1, \gamma) + \sum_{i \neq j} \rho(|(\Omega_{YY})_{ij}|; \lambda_2, \gamma),$$

$$P_2(\Omega_{YX}) = \sum_{i=1}^{p}\sum_{j=1}^{q} \rho(|(\Omega_{YX})_{ij}|; \lambda_2, \gamma).$$

Here $\rho(t; \lambda, \gamma) = \lambda \int_0^{|t|} \left(1 - \frac{x}{\lambda\gamma}\right)_+ dx$ is the MCP (minimax concave penalty [138]), λ_1 and λ_2 are data-dependent tuning parameters, and γ is the regularization parameter. The estimate is defined as the minimizer of (2.1), and a nonzero element corresponds to an interconnection.

Remarks Distributions of regulator data may further deviate from normality. With copy number variation (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 miRNA, 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 non-normality. 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 (Yuan and Zhang 2014), which uses a re-parametrization 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 [139]. 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.

2.2.3 A small example

To gain more intuition into the proposed analysis, we simulate one small dataset 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 2.3. 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 2.1). 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 2.1.

For this specific example, the proposed method has more accurate identification. Specifically, for the gene-expression-regulator network (Ω_{YY}), it identifies 14 true positives and 5 false positives, whereas the alternative separate estimation identifies 11 true positives and 9 false positives. For the gene-expression-only network $(\tilde{\Omega}_{YY})$, the proposed method identifies 16 true positives and 7 false positives, where the alternative identifies 13 true positives and 9 false positives. The alternative identification result violates the hierarchy. Specifically, there are three edges that are identified in the gene-expression-regulator network but not

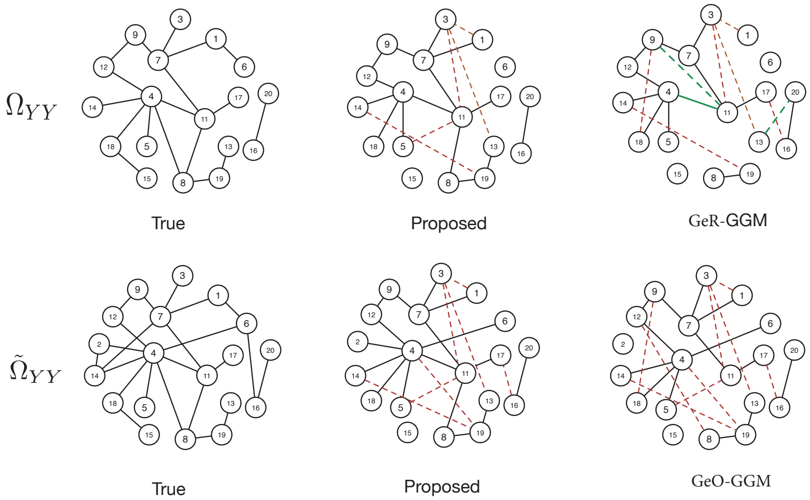


Figure 2.1: Gene expression networks in the small example: true (left), proposed (middle), and alternatives (right). Solid lines: true positives; Dashed lines: false positives; Green lines: identifications that violate the hierarchy.

in the gene-expression-only one. We further examine estimation performance using RMSE (details in Section 2.3). 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 2.3.

2.2.4 Statistical properties

Rigorously establishing statistical properties can provide the proposed approach a stronger ground than those not properly supported. Suppose that gene expressions $(Y_1, \cdots, Y_p)$ are associated with the vertex set $V_1 = \{1, 2, \cdots, p\}$ of the undirected graph $G_1 = (V_1, E_1)$, and that gene expressions plus regulators $(Y_1, \cdots, Y_p, X_1, \cdots, X_q)$ are associated with the vertex set $V_2 = \{1, 2, \cdots, p+q\}$ 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 $\widetilde{\mathcal{A}}_{YY} = \{(i,j)|(\widetilde{\Omega}^*_{YY})_{ij} \neq 0; i, j = 1, \cdots p\}$, $\mathcal{A}_{YY} = \{(i,j)|(\Omega^*_{YY})_{ij} \neq 0; i, j = 1, \cdots p\}$, and $\mathcal{A}_{YX} = \{(i,j)|(\Omega^*_{YX})_{ij} \neq 0; i = 1, \cdots p; j = p, \cdots p+q\}$ 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 $\mathcal{A} = \mathcal{A}_{YY} \cup \mathcal{A}_{YX}$, $\mathcal{A}^c = \{(i,j)|i = 1, \ldots, p; j =$

$1, \ldots, p+q\} \backslash \mathcal{A}$, $\mathcal{A}_1 = \mathcal{A}_{YY} \cup \widetilde{\mathcal{A}}_{YY}$, and $\mathcal{A}_1^c = \{(i,j) | i = 1, \ldots, p; j = 1, \ldots, p\} \backslash \mathcal{A}_1$.

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 $\widetilde{d} := \max_{i=1,\cdots,p} |\{j \in V_1 | (\widetilde{\Omega}^*_{YY})_{ij} \neq 0\}|$ and $d := \max_{i=1,\cdots,p} |\{j \in V_2 | \Omega^*_{ij} \neq 0\}|$.

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 $\kappa_{\widetilde{\Sigma}^*_{YY}} := \|\widetilde{\Sigma}^*_{YY}\|_\infty$. 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 $\widetilde{\Gamma}^* := \widetilde{\Omega}^{*-1}_{YY} \otimes \widetilde{\Omega}^{*-1}_{YY}$, where $\otimes$ denotes the Kronecker product. Consequently, we define $\widetilde{\Gamma}^*_{\mathcal{A}_1\mathcal{A}_1} := \left[\widetilde{\Omega}^{*-1}_{YY} \otimes \widetilde{\Omega}^{*-1}_{YY}\right]_{\mathcal{A}_1\mathcal{A}_1}$, $\kappa_{\widetilde{\Gamma}^*} := \|(\widetilde{\Gamma}^*_{\mathcal{A}_1\mathcal{A}_1})^{-1}\|_\infty$, and $\widetilde{\kappa} := \max_{e \in \mathcal{A}_1^c} \|\widetilde{\Gamma}^*_{e\mathcal{A}_1}(\widetilde{\Gamma}^*_{\mathcal{A}_1\mathcal{A}_1})^{-1}\|_1$. For the gene-expression-regulator graph, we denote its Hessian evaluated at the true values as:

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

Similar to above, we define $\kappa_1 := \max_{e \in \mathcal{A}_1^c} \|H^*_{e\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1}\|_1$, $\kappa_2 := \max_{e \in \widetilde{\mathcal{A}}_{YY} \cup \mathcal{A}^c_{YY}} \|H^*_{e\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1}\|_1$, $\kappa_3 := \max_{e \in \mathcal{A}^c_{YX}} \|H^*_{e\mathcal{A}}(H^*_{\mathcal{A}\mathcal{A}})^{-1}\|_1$, $c_{\Omega^{*-1}_{YY}} := \|\Omega^{*-1}_{YY}\|_\infty$, $c_{\Omega^*_{YX}} := \|\Omega^*_{YX}\|_1$, 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. $\min_{(i,j) \in \mathcal{A}_1} \left(|(\widetilde{\Omega}^*_{YY})_{ij}| + |(\Omega^*_{YY})_{ij}|\right) > \{\gamma + \kappa_{\widetilde{\Gamma}^*}/(\widetilde{\kappa} + 1)\}\lambda_1$.

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

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:* $n_1 > \max\left\{0, C_1 \log(4p^\tau)\widetilde{d}^2 - C_2 \log[4(p \vee q)^\tau]d^2\right\}$, $n_2 > C_2 \log[4(p \vee q)^\tau]d^2$, *where* $C_1 = \left[\max\{\kappa_{\widetilde{\Sigma}^*_{YY}} \kappa_{\widetilde{\Gamma}^*}, \kappa^3_{\widetilde{\Sigma}^*_{YY}} \kappa^2_{\widetilde{\Gamma}^*}\}\right]^2$ *and* $C_2 = c^2_{H^*}\left[\max\{3c_{\Omega^{*-1}_{YY}}, c_{\Omega^*_{YX}}, c_{\Omega^{*-1}_{YY}} c^2_{\Omega^*_{YX}} c^2_X\}\right]^2$. *In addition, the regularization and tuning parameters satisfy* $\lambda_1 > 2(\widetilde{\kappa}+1)c_*\sqrt{\frac{\log(4p^\tau)}{n_1+n_2}}$, *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 $c_* = 40\sqrt{2}\max_{i=1,\cdots,p}(\widetilde{\Omega}^{*-1}_{YY})_{ii}$ *and* $c'_* = \max\left\{40\sqrt{2}\max_i(\Omega^{*-1}_{YY})_{ii}, 2\sqrt{2}c_X\right\}$,

$$\|\widehat{\widetilde{\Omega}}_{YY} - \widetilde{\Omega}^*_{YY}\|_\infty \leq 2c_*\kappa_{\widetilde{\Gamma}^*}\sqrt{\frac{\log(4p^\tau)}{n_1+n_2}}, \tag{2.2}$$

$$\|\widehat{\Omega}_{YY} - \Omega^*_{YY}, \widehat{\Omega}_{YX} - \Omega^*_{YX}\|_\infty \leq 2c'_*c_{H^*}\sqrt{\frac{\log(4(p\vee q)^\tau)}{n_2}}. \tag{2.3}$$

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 (Section 2.6.

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.

2.2.5 Computation

We optimize objective function (2.1) 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: $t = 0$, $\Omega_{YY}^{(t)} = \widehat{\Omega}_{YY}$, $\Omega_{YX}^{(t)} = \widehat{\Omega}_{YX}$, $\widetilde{\Omega}_{YX}^{(t)} = S_{YY}^{-1}$, where $\widehat{\Omega} = \begin{pmatrix} \widehat{\Omega}_{YY} & \widehat{\Omega}_{YX} \\ \widehat{\Omega}_{YX}^\top & \widehat{\Omega}_{XX} \end{pmatrix}$ 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

$$\begin{aligned} M_1((\Omega_{YY})_{ij}) &= \frac{1}{2}\left[(\Omega_{YY})_{ij} - \left((\Omega_{YY}^{(t)})_{ij} - \eta^{(t)} A_{ij}^{(t)}\right)\right]^2 + \eta^{(t)} \rho(\sqrt{(\Omega_{YY})_{ij}^2 + (\widetilde{\Omega}_{YY}^{(t)})_{ij}^2}; \lambda_1, \gamma) \\ &+ \eta^{(t)} \rho(|(\Omega_{YY})_{ij}|; \lambda_2, \gamma). \end{aligned} \tag{2.4}$$

   Here $A^{(t)} = S_{YY} - (\Omega_{YY}^{(t)})^{-1} - (\Omega_{YY}^{(t)})^{-1} \Omega_{YX}^{(t)} S_{XX} (\Omega_{YX}^{(t)})^\top (\Omega_{YY}^{(t)})^{-1}$.

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

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

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

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

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

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

   where

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

   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

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

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

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

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

(2) Determine the step size.

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

$$\begin{aligned} \widetilde{L}_1(\widetilde{\Omega}_{YY}^{\star}) &= L_1(\widetilde{\Omega}_{YY}^{(t)}) + \operatorname{tr}\left((B^{(t)})^{\top}(\widetilde{\Omega}_{YY}^{\star} - \widetilde{\Omega}_{YY}^{(t)})\right) + \frac{1}{2\eta^{(t)}} \parallel \widetilde{\Omega}_{YY}^{\star} - \widetilde{\Omega}_{YY}^{(t)} \parallel_F^2 \\ \widetilde{L}_2(\Omega_{YY}^{\star}, \Omega_{YX}^{\star}) &= L_2(\Omega_{YY}^{(t)}, \Omega_{YX}^{(t)}) + \operatorname{tr}\left((A^{(t)})^{\top}(\Omega_{YY}^{\star} - \Omega_{YY}^{(t)})\right) + \operatorname{tr}\left((C^{(t)})^{\top}(\Omega_{YX}^{\star} - \Omega_{YX}^{(t)})\right) \\ &+ \frac{1}{2\eta^{(t)}}\left[\parallel \Omega_{YY}^{\star} - \Omega_{YY}^{(t)} \parallel_F^2 + \parallel \Omega_{YX}^{\star} - \Omega_{YX}^{(t)} \parallel_F^2\right]. \end{aligned} \tag{2.5}$$

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

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

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

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

$$\parallel \Omega_{YY}^{(t+1)} - \Omega_{YY}^{(t)} \parallel_F + \parallel \widetilde{\Omega}_{YY}^{(t+1)} - \widetilde{\Omega}_{YY}^{(t)} \parallel_F + \parallel \Omega_{YX}^{(t+1)} - \Omega_{YX}^{(t)} \parallel_F \le 10^{-3}$$

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 seconds 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.

2.3 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 ($k = 1, \ldots, 5$), 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 2.1, 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 2.1 in the main text and Tables 2.2-2.6 in the Appendix (Section 2.6), 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.

Table 2.1: Summary statistics on identification and estimation. Homogeneous Ω_{YY}, $\theta = 0.1$, and D_1 and D_2 containing the same 300 samples. In each cell, mean(SD).

	$\tilde{\Omega}_{YY}$				Ω_{YY}				Hierarchy violation	
	Identification			Estimation	Identification			Estimation	Count	Proportion
	Recall	FPR	Fscore		Recall	FPR	Fscore			
$(p, q) = (50, 50)$										
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)		

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), FPR (false positive rate), and F-score (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 2.2, 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 2.1. The rest of the results are presented in Tables 2.2-2.6 in the Appendix (Section 2.6). It is observed that, across all simulation settings, the proposed analysis outperforms the separate estimation. Consider for example the last setting in Table 2.1. 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 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 2.2.1. 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 false positive rates 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 2.2.1, 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 non-normal data.

2.4 Data analysis

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

2.4.1 Cutaneous melanoma (SKCM) data

Following the literature, we focus on the 395 White patients who had non-glabrous skin. Beyond gene expressions, data is also available on copy number variations. 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 copy number variations. We graphically examine the marginal distributions of gene expressions and copy number variations. All distributions are continuous, and the dominating majority are bell-shaped. We also conduct marginal regressions of gene expressions on copy number variations. There are no copy number variations seemingly with complementary effects. As such, the proposed approach can be reasonably applied.

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 2.7 in the Appendix (Section 2.6). The estimated graph structures using the proposed approach are presented in Figure 2.2. Results using the alternative and comparison are presented in Figure 2.4 in the Appendix (Section 2.6). 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 (that is, violation of the hierarchy).

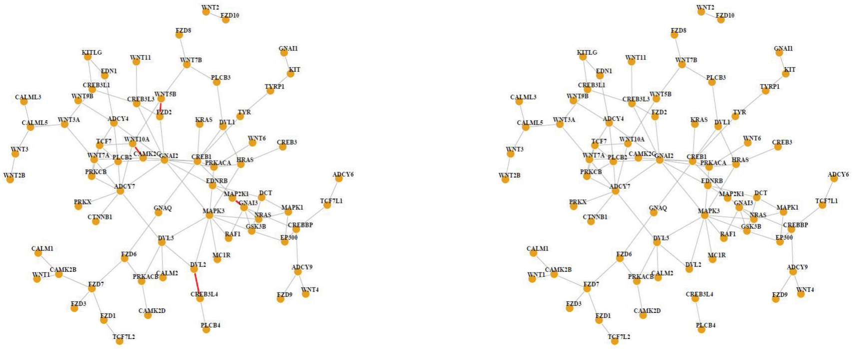


Figure 2.2: Analysis of TCGA SKCM data using the proposed approach: the GeO-GGM (left) and GeR-GGM (right) gene expression networks. Four red edges are identified in the GeO-GGM but not GeR-GGM.

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$^{\mathrm{R141W}}$ 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 dataset 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 dataset, 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.

2.4.2 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 copy number variations 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 copy number variations analyzed. The same exploratory analysis as for the melanoma data is conducted, again suggesting it is reasonable to apply the proposed approach.

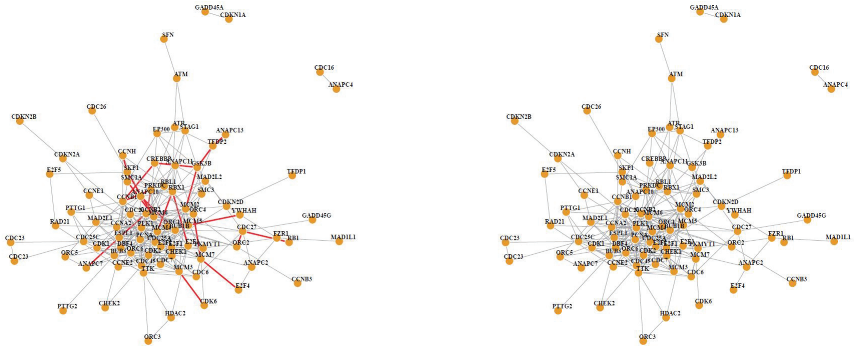


Figure 2.3: Analysis of TCGA lung cancer data using the proposed approach: the GeO-GGM (left) and GeR-GGM (right) gene expression networks. Twenty-one red edges are identified in the GeO-GGM but not GeR-GGM.

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 2.8 in the Appendix (Section 2.6). The estimated graph structures are presented in Figure 2.3 and Figure 2.5 in the Appendix (Section 2.6). 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.

2.5 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 gene-expression-only and gene-expression-regulator 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 [135–137]. 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 non-omics 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 non-normal 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.

2.6 Appendix

2.6.1 Establishment of statistical properties

We first establish some auxiliary lemmas, which will be needed in proving the main theorem. The following additional notations are needed. Let $\widetilde{W} = \widetilde{S}_{YY} - \widetilde{\Sigma}^*_{YY} \in \mathbb{R}^{p\times p}$ denote the "effective noise" in the sample covariance matrix $\widetilde{\Sigma}^*_{YY}$, where $\widetilde{\Sigma}^*_{YY} = \widetilde{\Omega}^{*-1}_{YY}$. The remainder of the difference $\widetilde{\Delta}$ between the estimator $\widehat{\widetilde{\Omega}}_{YY}$ and its true value $\widetilde{\Omega}^*_{YY}$ takes the form $\widetilde{R}(\widetilde{\Delta}) = \widehat{\widetilde{\Omega}}^{-1}_{YY} - \widetilde{\Omega}^{*-1}_{YY} + \widetilde{\Omega}^{*-1}_{YY}\widetilde{\Delta}\widetilde{\Omega}^{*-1}_{YY}$. Similarly, denote $g(\Omega_{YY}, \Omega_{YX}) = (g_1(\Omega_{YY}, \Omega_{YX}), g_2(\Omega_{YY}, \Omega_{YX}))$ as the gradient, $\Delta = \Theta - \Theta^*$ as the difference, and $R(\Delta) = g(\Omega^*_{YY}, \Omega^*_{YX}) - g(\Omega_{YY}, \Omega_{YX}) + H(\Omega^*_{YY}, \Omega^*_{YX})\Delta$ as the remainder in the second part. The following lemma relates $\widetilde{R}(\widetilde{\Delta})$ to $\widetilde{\Delta}$.

Lemma 1. *Suppose that* $\|\widetilde{\Delta}\|_\infty \leq \frac{1}{3\kappa_{\widetilde{\Sigma}^*_{YY}}\widetilde{d}}$ *holds. Then* $\|\widetilde{R}(\widetilde{\Delta})\|_\infty \leq \frac{3}{2}\widetilde{d}\kappa^3_{\widetilde{\Sigma}^*_{YY}}\|\widetilde{\Delta}\|^2_\infty$.

This lemma can be proved by following Lemma 5 of Ravikumar et al. (2011). Similarly, following the proof of Lemma 6 in the same reference, we can establish the following lemma, which provides a control of the deviation $\widetilde{\Delta}$, measured in the element-wise ℓ_∞ norm.

Lemma 2. *Suppose that* $\|\widetilde{W}\|_\infty \leq \frac{1}{2\kappa_{\widetilde{\Gamma}^*}\widetilde{d}}\min\{\frac{1}{3\kappa_{\widetilde{\Sigma}^*_{YY}}}, \frac{1}{3\kappa^3_{\widetilde{\Sigma}^*_{YY}}\kappa_{\widetilde{\Gamma}^*}}\}$ *holds. Then* $\|\widetilde{\Delta}\|_\infty = \|\widehat{\widetilde{\Omega}}_{YY} - \widetilde{\Omega}^*_{YY}\|_\infty \leq 2\kappa_{\widetilde{\Gamma}^*}\|\widetilde{W}\|_\infty$.

For the GeR-GGM, we can establish the following lemmas to control the remainder and difference as well.

Lemma 3. *Suppose that* $\|\Delta\|_\infty \leq \frac{1}{d}\min\{\frac{1}{3c_{\Omega^{*-1}_{YY}}}, \frac{c_{\Omega^*_{YX}}}{2}\}$ *holds. Then* $\|R(\Delta)\|_\infty \leq 206c^4_{\Omega^{*-1}_{YY}}c^2_{\Omega^*_{YX}}c^2_X d\|\Delta\|^2_\infty$.

Lemma 4. *Suppose that* $\max\{\|g^*_1\|_\infty, \|g^*_2\|_\infty\} \leq \frac{1}{2c_{H^*}d}\min\{\frac{1}{3c_{\Omega^{*-1}_{YY}}}, \frac{c_{\Omega^*_{YX}}}{2}, \frac{1}{412c^4_{\Omega^{*-1}_{YY}}c^2_{\Omega^*_{YX}}c^2_X d\|\Delta\|^2_\infty}\}$ *holds. Then* $\|\Delta\|_\infty \leq 2c_{H^*}\max\{\|g^*_1\|_\infty, \|g^*_2\|_\infty\}$.

The proof of Lemmas 3 and 4 follows that of Lemmas 3 and 4 in Wytock and Kolter (2013).

Proof of Theorem 1. Take the partial derivatives of $L_1(\widetilde{\Omega}_{YY})$, evaluate at $\widehat{\widetilde{\Omega}}_{YY}$, and denote $\widetilde{Z} := \widetilde{S}_{YY} - \widehat{\widetilde{\Omega}}_{YY}^{-1}$. Here $\widetilde{Z}$ is a member of the off-diagonal sub-differential $\partial\|\widehat{\widetilde{\Omega}}_{YY}\|_1$. Similarly, we derive the analytic expressions for the gradient and Hessian of $L_2(\Omega_{YY}, \Omega_{YX})$. Taking the first- and second-order partial derivatives, we obtain

$$g_1(\Omega_{YY}, \Omega_{YX}) = S_{YY} - \Omega_{YY}^{-1} - \Omega_{YY}^{-1}\Omega_{YX}S_{XX}\Omega_{YX}^{\top}\Omega_{YY}^{-1}, \tag{2.6}$$

$$g_2(\Omega_{YY}, \Omega_{YX}) = 2S_{YX} + 2\Omega_{YY}^{-1}\Omega_{YX}S_{XX}, \tag{2.7}$$

$$H(\Omega_{YY}, \Omega_{YX}) = \begin{pmatrix} \Omega_{YY}^{-1} \otimes (\Omega_{YY}^{-1} + 2\Omega_{YY}^{-1}\Omega_{YX}S_{XX}\Omega_{YX}^{\top}\Omega_{YY}^{-1}) & -2\Omega_{YY}^{-1} \otimes S_{XX}\Omega_{YX}^{\top}\Omega_{YY}^{-1} \\ -2\Omega_{YY}^{-1} \otimes \Omega_{YY}^{-1}\Omega_{YX}S_{XX} & 2\Omega_{YY}^{-1} \otimes S_{XX} \end{pmatrix}.$$

Taking a similar strategy as in published literature, we have $Y = -X\Omega_{YX}^{*\top}\Omega_{YY}^{*-1} + E$, where $E \in \mathbb{R}^{n_2 \times p}$, $\mathbb{E}(E_i) = 0$, $\mathrm{Var}(E_i) = \Omega_{YY}^{*-1}$, and E_i's are Gaussian. Define the dual solution $Z := (Z^{YY}, Z^{YX})$ by evaluating (2.6) and (2.7) at the estimator $\hat{\Theta} = (\widehat{\Omega}_{YY}, \widehat{\Omega}_{YX})$, where $Z^{YY} = S_{YY} - \widehat{\Omega}_{YY}^{-1} - \widehat{\Omega}_{YY}^{-1}\widehat{\Omega}_{YX}S_{XX}\widehat{\Omega}_{YX}^{\top}\widehat{\Omega}_{YY}^{-1}$ and $Z^{YX} = 2S_{YX} + 2\widehat{\Omega}_{YY}^{-1}\widehat{\Omega}_{YX}S_{XX}$.

Now we prove that, under the assumed conditions, entries in $\widetilde{Z}$, $|Z^{YY}|$, and $|Z^{YX}|$ satisfy

$$|\widetilde{Z}_{ij}| < \lambda_1 \text{ for all } (i,j) \notin \mathcal{A}_1, \tag{2.8}$$

$$\begin{aligned} &|Z_{ij}^{YY}| < \lambda_1 + \lambda_2 \text{ for } (i,j) \in \mathcal{A}_1^c, \\ &|Z_{ij}^{YY}| < \lambda_2 \text{ for } (i,j) \in \widetilde{\mathcal{A}}_{YY} \cap \mathcal{A}_{YY}^c, \\ &|Z_{ij}^{YX}| < \lambda_2 \text{ for } (i,j) \in \mathcal{A}_{YX}^c. \end{aligned} \tag{2.9}$$

We will first prove inequality (2.8), so as to facilitate the proof of (2) in Theorem 1. We will then prove inequality (2.9).

Some calculations yield

$$\widetilde{Z} = \widetilde{S}_{YY} - \widehat{\widetilde{\Omega}}_{YY}^{-1} = \widetilde{\Omega}_{YY}^{*-1}\widetilde{\Delta}\widetilde{\Omega}_{YY}^{*-1} + \widetilde{W} - \widetilde{R}(\widetilde{\Delta}). \tag{2.10}$$

Then, we vectorize the matrices

$$\mathrm{vec}(\widetilde{\Omega}_{YY}^{*-1}\widetilde{\Delta}\widetilde{\Omega}_{YY}^{*-1}) = (\widetilde{\Omega}_{YY}^{*-1} \otimes \widetilde{\Omega}_{YY}^{*-1})\mathrm{vec}(\widetilde{\Delta}) = \widetilde{\Gamma}^{*}\mathrm{vec}(\widetilde{\Delta}). \tag{2.11}$$

Combining (2.10) and (2.11), and denoting $\overline{\widetilde{\Delta}} = \mathrm{vec}(\widetilde{\Delta})$, $\overline{\widetilde{W}} = \mathrm{vec}(\widetilde{W})$, and $\overline{\widetilde{R}} = \mathrm{vec}(\widetilde{R}(\widetilde{\Delta}))$, we obtain

$$\begin{aligned}\widetilde{Z}_{\mathcal{A}_1^c} &= \widetilde{\Gamma}^*_{\mathcal{A}_1^c\mathcal{A}_1}\overline{\widetilde{\Delta}}_{\mathcal{A}_1} + \overline{\widetilde{W}}_{\mathcal{A}_1^c} - \overline{\widetilde{R}}_{\mathcal{A}_1^c} \\ &= -\widetilde{\Gamma}^*_{\mathcal{A}_1^c\mathcal{A}_1}\left(\widetilde{\Gamma}^*_{\mathcal{A}_1\mathcal{A}_1}\right)^{-1}\left(\overline{\widetilde{W}}_{\mathcal{A}_1} - \overline{\widetilde{R}}_{\mathcal{A}_1}\right) + \overline{\widetilde{W}}_{\mathcal{A}_1^c} - \overline{\widetilde{R}}_{\mathcal{A}_1^c}.\end{aligned}$$

Recalling the definition of $\widetilde{\kappa}$, we have

$$\|\widetilde{Z}_{\mathcal{A}_1^c}\|_\infty \leq \left[\max_{e\in\mathcal{A}_1^c}\|\widetilde{\Gamma}^*_{e\mathcal{A}_1}(\widetilde{\Gamma}^*_{\mathcal{A}_1\mathcal{A}_1})^{-1}\|_1 + 1\right](\|\overline{\widetilde{W}}\|_\infty + \|\overline{\widetilde{R}}\|_\infty) < \lambda_1, \tag{2.12}$$

if

$$\max\left\{\|\widetilde{W}\|_\infty, \|\widetilde{R}(\widetilde{\Delta})\|_\infty\right\} < \frac{\lambda_1}{2(\widetilde{\kappa}+1)}. \tag{2.13}$$

Next we prove that condition (2.13) holds. Consider event $\mathcal{B}_1 = \left\{\|\widetilde{W}\|_\infty \leq c_*\sqrt{\frac{\log(4p^\tau)}{n_1+n_2}}\right\}$ with $c_* = 40\sqrt{2}\max_{i=1,\cdots,p}(\widetilde{\Omega}^{*-1}_{YY})_{ii}$ and $\tau > 2$. Since the samples are independent, by Lemma 1 of Ravikumar et al. (2011), we have

$$\mathbf{P}\left[\left|(\widetilde{S}_{YY})_{ij} - (\widetilde{\Sigma}^*_{YY})_{ij}\right| > \delta_1\right] \leq 4\exp\left\{-\frac{(n_1+n_2)\delta_1^2}{3200\max_{i=1,\cdots,p}(\widetilde{\Omega}^{*-1}_{YY})^2_{ii}}\right\} \tag{2.14}$$

for all $\delta_1 \in \left(0, 25\max_i(\widetilde{\Omega}^{*-1}_{YY})^2_{ii}\right)$. Consequently, it can be obtained that $\mathbf{P}(\mathcal{B}_1) \geq 1 - p^{2-\tau}$ according to Lemma 8 of Ravikumar et al. (2011). Thus, based on the condition on λ_1, half of the condition (2.13) is approved. Accordingly, the following proof is conditional on event $\mathcal{B}_1$.

With the lower bound of sample size $n_1 + n_2 > C_1\log(4p^\tau)\widetilde{d}^2$, we have

$$\|\widetilde{W}\|_\infty \leq c_*\sqrt{\frac{\log(4p^\tau)}{n_1+n_2}} \leq \frac{1}{2\kappa_{\widetilde{\Gamma}^*}\widetilde{d}}\min\left\{\frac{1}{3\kappa_{\widetilde{\Sigma}^*_{YY}}}, \frac{1}{3\kappa^3_{\widetilde{\Sigma}^*_{YY}}\kappa_{\widetilde{\Gamma}^*}}\right\}, \tag{2.15}$$

showing that the assumptions of Lemma 2 are satisfied. Applying this lemma, we conclude that

$$\|\widetilde{\Delta}\|_\infty \leq 2\kappa_{\widetilde{\Gamma}^*}\|\widetilde{W}\|_\infty. \tag{2.16}$$

Consider Lemma 1. We can see that its assumption $\|\widetilde{\Delta}\|_\infty \le \frac{1}{3\kappa_{\widetilde{\Sigma}^*_{YY}}\widetilde{d}}$ holds by applying equations (2.15) and (2.16). Consequently, we have

$$\begin{aligned} \|\widetilde{R}(\widetilde{\Delta})\|_\infty &\le \frac{3}{2}\widetilde{d}\|\widetilde{\Delta}\|_\infty^2 \kappa^3_{\widetilde{\Sigma}^*_{YY}} \le 6\kappa^3_{\widetilde{\Sigma}^*_{YY}} \kappa^2_{\widetilde{\Gamma}^*}\widetilde{d}c_*^2 \cdot \frac{\log(4p^\tau)}{n_1+n_2} \\ &\le c_*\sqrt{\frac{\log(4p^\tau)}{n_1+n_2}} < \frac{\lambda_1}{2(\widetilde{\kappa}+1)}, \end{aligned} \tag{2.17}$$

where the last line follows from the condition on λ_1 and control of sampling noise (2.15). As such, we have proved that condition (2.13) holds, which completes the proof of $\|\widetilde{Z}_{A_1^c}\|_\infty < \lambda_1$. Therefore, we have shown $\|\widetilde{\Delta}\|_\infty = \|\widehat{\widetilde{\Omega}}_{YY} - \widetilde{\Omega}^*_{YY}\|_\infty \le 2\kappa_{\widetilde{\Gamma}^*}c_*\sqrt{\frac{\log(4p^\tau)}{n_1+n_2}}$.

To briefly summarize the proof of Equation (2), we have shown that $\max\left\{\|\widetilde{W}\|_\infty, \|\widetilde{R}(\widetilde{\Delta})\|_\infty\right\} < \frac{\lambda_1}{2(\widetilde{\kappa}+1)}$ holds, allowing us to conclude that $\|\widetilde{Z}_{A_1^c}\|_\infty < \lambda_1$. In this way, $\widetilde{Z}$ is an optimal solution to the corresponding dual problem, and the unimportant entries shrink to zero. Then, we have proved that the ℓ_∞-bound of the difference between $\widehat{\widetilde{\Omega}}_{YY}$ and $\widetilde{\Omega}^*_{YY}$ is bounded by $2\kappa_{\widetilde{\Gamma}^*}c_*\sqrt{\frac{\log(4p^\tau)}{n_1+n_2}}$, as claimed in Theorem 1. Since this is conditioned on event $\mathcal{B}$, the above statements hold with probability $\mathbf{P}(\mathcal{B}_1) \ge 1-p^{2-\tau}$. Moreover, we already have $(\widehat{\widetilde{\Omega}}_{YY})_{\mathcal{A}_1^c} = 0$ by the construction with the MCP penalty.

Now consider the proof of Equation (3). For any Θ, define Δ as the difference between Θ and its true value Θ^*, that is,

$$\Delta := \Theta - \Theta^* = \begin{pmatrix} \Omega_{YY} - \Omega^*_{YY} \\ \Omega_{YX} - \Omega^*_{YX} \end{pmatrix}. \tag{2.18}$$

Recall the definition of the remainder

$$R(\Delta) = g(\Omega^*_{YY}, \Omega^*_{YX}) - g(\Omega_{YY}, \Omega_{YX}) + H(\Omega^*_{YY}, \Omega^*_{YX})\Delta, \tag{2.19}$$

which consists of the residuals of the first order Taylor expansion of the gradient. By the construction of the estimator $(\widehat{\Omega}_{YY}, \widehat{\Omega}_{YX})$, we have $Z_{\mathcal{A}^c} = 0$. Also, both $Z_{\mathcal{A}^c} = 0$ and $Z_{\mathcal{A}}$ satisfy

$$\begin{pmatrix} Z_{\mathcal{A}} \\ Z_{\mathcal{A}^c} \end{pmatrix} = \begin{pmatrix} H^*_{\mathcal{A}\mathcal{A}} & H^*_{\mathcal{A}\mathcal{A}^c} \\ H^*_{\mathcal{A}^c\mathcal{A}} & H^*_{\mathcal{A}^c\mathcal{A}^c} \end{pmatrix} \begin{pmatrix} \Delta_{\mathcal{A}} \\ 0 \end{pmatrix} + \begin{pmatrix} g^*_{\mathcal{A}} \\ g^*_{\mathcal{A}^c} \end{pmatrix} - \begin{pmatrix} R_{\mathcal{A}}(\Delta) \\ R_{\mathcal{A}^c}(\Delta) \end{pmatrix}. \tag{2.20}$$

Therefore, we have $\Delta_{\mathcal{A}} = H_{\mathcal{A}\mathcal{A}}^{*-1}[R_{\mathcal{A}}(\Delta) - g_{\mathcal{A}}^*]$. Plugging this result into $Z_{\mathcal{A}^c}$, we obtain

$$Z_{\mathcal{A}^c} = H_{\mathcal{A}^c\mathcal{A}}^*\Delta_{\mathcal{A}} + g_{\mathcal{A}^c}^* - R_{\mathcal{A}^c}(\Delta) = H_{\mathcal{A}^c\mathcal{A}}^*H_{\mathcal{A}\mathcal{A}}^{*-1}[R_{\mathcal{A}}(\Delta) - g_{\mathcal{A}}^*] + g_{\mathcal{A}^c}^* - R_{\mathcal{A}^c}(\Delta). \quad (2.21)$$

Next, we prove that $Z_{\mathcal{A}^c}$ satisfies (2.9) under mild conditions with a high probability. By (2.20) and (2.21), we have

$$Z_{\mathcal{A}_1^c} = H_{\mathcal{A}_1^c\mathcal{A}}^*H_{\mathcal{A}\mathcal{A}}^{*-1}[R_{\mathcal{A}}(\Delta) - g_{\mathcal{A}}^*] + g_{\mathcal{A}_1^c}^* - R_{\mathcal{A}_1^c}(\Delta). \quad (2.22)$$

From the above equation, we have

$$\|Z_{\mathcal{A}_1^c}\|_\infty \le \left[\max_{e\in\mathcal{A}^c}\|H_{e\mathcal{A}}^*H_{\mathcal{A}\mathcal{A}}^{*-1}\|_1 + 1\right](\|g^*\|_\infty + \|R(\Delta)\|_\infty). \quad (2.23)$$

Similarly, we have

$$\|Z_{\widetilde{\mathcal{A}}_{YY}\cap\mathcal{A}_{YY}^c}\|_\infty \le \left[\max_{e\in\widetilde{\mathcal{A}}_{YY}\cap\mathcal{A}_{YY}^c}\|H_{e\mathcal{A}}^*H_{\mathcal{A}\mathcal{A}}^{*-1}\|_1 + 1\right](\|g^*\|_\infty + \|R(\Delta)\|_\infty), \quad (2.24)$$

$$\|Z_{\mathcal{A}_{YX}^c}\|_\infty \le \left[\max_{e\in\mathcal{A}_{YX}^c}\|H_{e\mathcal{A}}^*H_{\mathcal{A}\mathcal{A}}^{*-1}\|_1 + 1\right](\|g^*\|_\infty + \|R(\Delta)\|_\infty). \quad (2.25)$$

With the definitions of κ_1, κ_2, and κ_3 and condition on the regularization and tuning parameters

$$\max\{\|g^*\|_\infty, \|R(\Delta)\|_\infty\} < \frac{1}{2}\min\left\{\frac{\lambda_1+\lambda_2}{\kappa_1+1}, \frac{\lambda_2}{\kappa_2+1}, \frac{\lambda_2}{\kappa_3+1}\right\}, \quad (2.26)$$

we can derive that $Z_{\mathcal{A}^c}$ satisfies (2.9).

Next, we prove that (2.26) holds with a high probability. Consider event $\mathcal{B}_2 = \left\{\max\{\|g_1^*\|_\infty, \|g_2^*\|_\infty\} \le c'_*\sqrt{\frac{\log(4(p\vee q)^\tau)}{n_2}}\right\}$, where $c'_* = \max\{40\sqrt{2}\max_i(\Omega_{YY}^{*-1})_{ii}, 2\sqrt{2}c_X\}$ and $\tau > 2$.

By the Bonferroni's inequality and Gaussian assumption,

$$\begin{aligned} P\left(\|g_1^*\|_\infty > \delta_2\right) &= P(\|n_2^{-1}E^\top E - \Omega_{YY}^{*-1}\|_\infty > \delta_2) \\ &\le 4p^2\exp\left\{-\frac{n_2\delta_2^2}{3200\max_i(\Omega_{YY}^{*-1})_{ii}^2}\right\} \quad (2.27) \end{aligned}$$

for all $\delta_2 \in \big(0, 40\sqrt{2}\max_i(\Omega^{*-1}_{YY})_{ii}\big)$, and

$$P\left(\|g_2^*\|_\infty > \delta_2\right) = P(\|2n_2^{-1}E^\top X\|_\infty > \delta_2) \leq 2pq\exp\left\{-\frac{n_2\delta_2^2}{8c_X^2}\right\}. \tag{2.28}$$

Let $\delta_2 = c'_*\sqrt{\frac{\log(4(p\vee q)^\tau)}{n_2}}$. We have $P(\mathcal{B}_2) \geq 1 - 2(p\vee q)^{2-\tau}$. Thus, half of (2.26) is established. We then proceed with the proof conditional on $\mathcal{B}_2$. With the bound on sample size $n_2 > C_2\log(4(p\vee q)^\tau)d^2$, we have

$$\|g^*\|_\infty \leq c'_*\sqrt{\frac{\log(4(p\vee q)^\tau)}{n_2}} \leq \frac{1}{2c_{H^*}d}\min\{\frac{1}{3c_{\Omega^{*-1}_{YY}}}, \frac{c_{\Omega^*_{YX}}}{2}, \frac{1}{412c^4_{\Omega^{*-1}_{YY}}c^2_{\Omega^*_{YX}}c_X^2 d\|\Delta\|^2_\infty}\}. \tag{2.29}$$

Thus, the condition for Lemma 4 holds, and $\|\Delta\|_\infty \leq 2c_{H^*}\|g^*\|_\infty$. By Lemma 3,

$$\begin{aligned}\|R(\Delta)\|_\infty &\leq 206c^4_{\Omega^{*-1}_{YY}}c^2_{\Omega^*_{YX}}c_X^2 d\|\Delta\|^2_\infty \leq 824c^4_{\Omega^{*-1}_{YY}}c^2_{\Omega^*_{YX}}c_X^2 dc'^2_*\frac{\log(4(p\vee q)^\tau)}{n_2} \\ &< \frac{1}{2}\min\left\{\frac{\lambda_1+\lambda_2}{\kappa_1+1}, \frac{\lambda_2}{\kappa_2+1}, \frac{\lambda_2}{\kappa_3+1}\right\}. \end{aligned} \tag{2.30}$$

Combining $\|\Delta\|_\infty \leq 2c_{H^*}\|g^*\|_\infty$ and (2.29), together with (2.30), we have established condition (2.26). Therefore, $\|\Delta\|_\infty \leq 2c_{H^*}c'_*\sqrt{\frac{\log(4(p\vee q)^\tau)}{n_2}}$.

To summarize the proof of Equation (3), we have started with the generic primal-dual witness approach, showing that the condition $\max\{\|g^*\|_\infty, \|R(\Delta)\|_\infty\} < \frac{1}{2}\min\left\{\frac{\lambda_1+\lambda_2}{\kappa_1+1}, \frac{\lambda_2}{\kappa_2+1}, \frac{\lambda_2}{\kappa_3+1}\right\}$ holds, which allows us to conclude that the nonzero entries of $\|Z\|_\infty$ are bounded above by the penalties. In this way, Z is an optimal solution to the corresponding dual problem, and the estimates of the zero entries shrink to zero. Then, we have proved that the ℓ_∞-bound of the difference between estimator $\widehat{\Theta}$ and its true value Θ^* is bounded above by $2c_{H^*}c'_*\sqrt{\frac{\log(4(p\vee q)^\tau)}{n_2}}$, as stated in Theorem 1. These above statements hold with probability $\mathbf{P}(\mathcal{B}_2) \geq 1 - 2(p\vee q)^{2-\tau}$.

Together with the proof of Equation (2), we have shown that the results in Theorem 1 hold with probability $1 - 1/p^{\tau-2} - 2/(p\vee q)^{\tau-2}$.

□

2.6.2 Additional numerical results

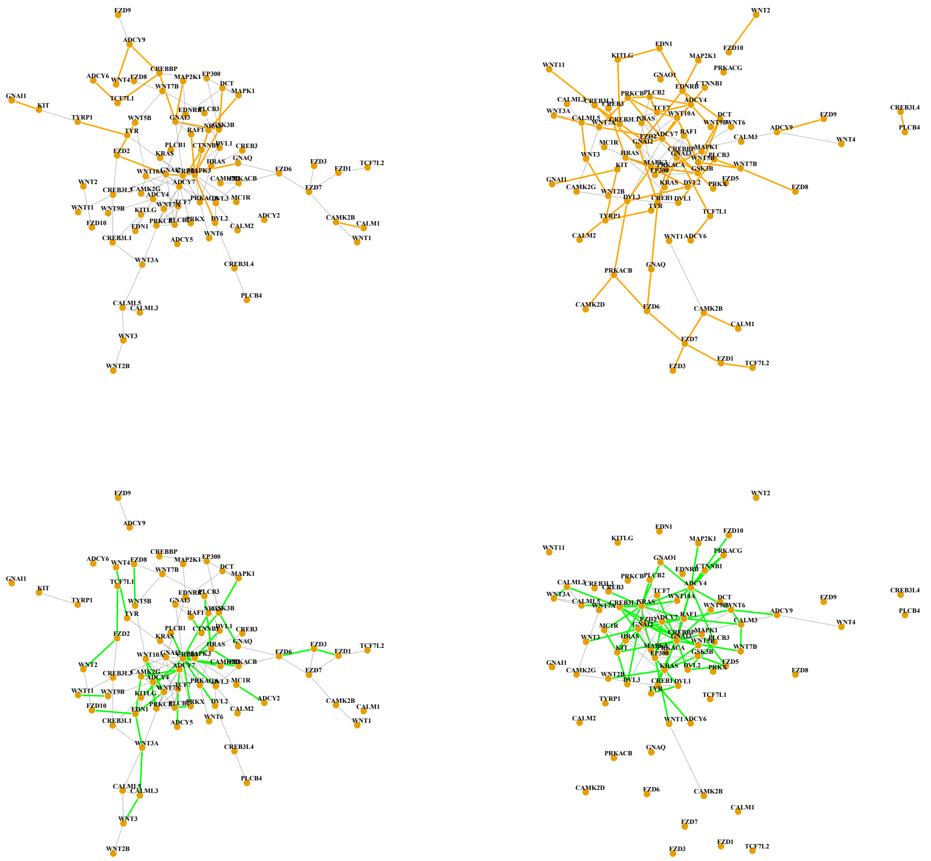


Figure 2.4: Analysis of TCGA SKCM data using the proposed and alternative approaches. Left/right: GeO-GGM/GeR-GGM (only the gene expression network is presented). Upper/lower: proposed/alternative. Grey: edges shared by the two approaches; Orange: edges unique to the proposed approach. Green: edges unique to the alternative approach.

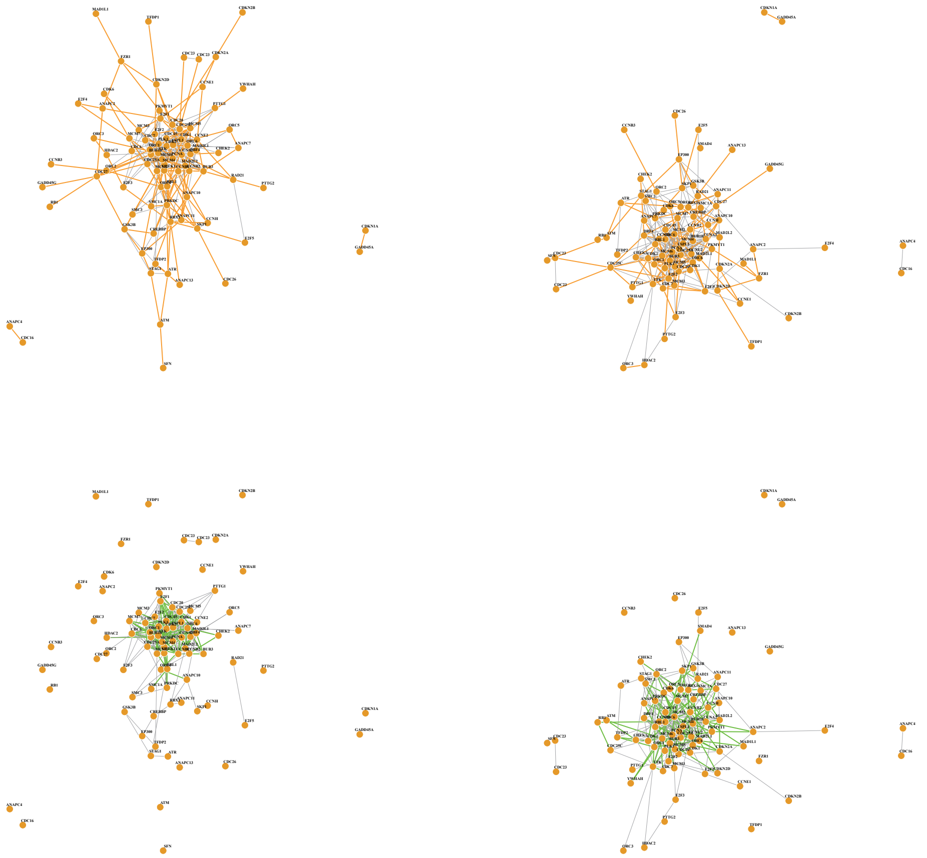


Figure 2.5: Analysis of TCGA lung cancer data using the proposed and alternative approaches. Left/right: GeO-GGM/GeR-GGM (only the gene expression network is presented). Upper/lower: proposed/alternative. Grey: edges shared by the two approaches; Orange: edges unique to the proposed approach. Green: edges unique to the alternative approach.

Table 2.2: Summary statistics on identification and estimation. Homogeneous Ω_{YY}, $\theta = 0.05$, and D_1 and D_2 containing the same 300 samples. In each cell, mean(SD).

	$\tilde{\Omega}_{YY}$				Ω_{YY}				Hierarchy violation	
	Identification			Estimation	Identification			Estimation	Count	Proportion
	Recall	FPR	Fscore		Recall	FPR	Fscore			
$(p, q) = (50, 50)$										
Proposed	0.508 (0.051)	0.027 (0.006)	0.556 (0.04)	19.61 (0.77)	0.605 (0.037)	0.051 (0.007)	0.62 (0.026)	39.38 (1.27)	0 (0)	0 (0)
GeO-GGM	0.456 (0.038)	0.021 (0.01)	0.519 (0.037)	23.01 (1.52)	–	–	–	–	2.46 (0.88)	3.77% (1.07%)
GeR-GGM	–	–	–	–	0.517 (0.036)	0.048 (0.011)	0.558 (0.033)	44.38 (2.49)		
$(p, q) = (50, 100)$										
Proposed	0.515 (0.044)	0.03 (0.005)	0.562 (0.033)	19.65 (0.94)	0.578 (0.039)	0.045 (0.007)	0.597 (0.035)	76.01 (2.72)	0 (0)	0 (0)
GeO-GGM	0.446 (0.044)	0.013 (0.007)	0.538 (0.03)	22.89 (1.48)	–	–	–	–	17.27 (2.35)	33.21% (4%)
GeR-GGM	–	–	–	–	0.501 (0.052)	0.045 (0.012)	0.543 (0.029)	80.01 (4.26)		
$(p, q) = (50, 150)$										
Proposed	0.464 (0.027)	0.048 (0.007)	0.548 (0.024)	19.92 (1.93)	0.56 (0.05)	0.033 (0.008)	0.583 (0.041)	147.7 (3.04)	0 (0)	0 (0)
GeO-GGM	0.436 (0.045)	0.012 (0.01)	0.523 (0.036)	21.83 (1.43)	–	–	–	–	31.4 (4.11)	72.1% (9.8%)
GeR-GGM	–	–	–	–	0.486 (0.038)	0.051 (0.02)	0.533 (0.027)	155.2 (8.06)		
$(p, q) = (100, 50)$										
Proposed	0.483 (0.023)	0.049 (0.004)	0.526 (0.017)	57.92 (1.93)	0.602 (0.02)	0.058 (0.004)	0.581 (0.014)	108.8 (3.04)	0 (0)	0 (0)
GeO-GGM	0.467 (0.024)	0.081 (0.02)	0.464 (0.019)	75.34 (3.25)	–	–	–	–	48.4 (10.5)	7.6% (1.6%)
GeR-GGM	–	–	–	–	0.453 (0.034)	0.05 (0.013)	0.487 (0.016)	145.5 (4.3)		
$(p, q) = (100, 100)$										
Proposed	0.438 (0.016)	0.051 (0.004)	0.493 (0.016)	57.64 (1.68)	0.559 (0.021)	0.052 (0.004)	0.548 (0.019)	214.9 (6.76)	0 (0)	0 (0)
GeO-GGM	0.431 (0.015)	0.071 (0.006)	0.463 (0.012)	74.73 (3.14)	–	–	–	–	179.8 (29.5)	45.3% (8.4%)
GeR-GGM	–	–	–	–	0.436 (0.023)	0.045 (0.014)	0.477 (0.018)	245.6 (9.79)		
$(p, q) = (100, 150)$										
Proposed	0.429 (0.015)	0.014 (0.004)	0.503 (0.012)	56.46 (2.04)	0.522 (0.02)	0.049 (0.005)	0.542 (0.017)	535.3 (22.8)	0 (0)	0 (0)
GeO-GGM	0.421 (0.014)	0.09 (0.008)	0.462 (0.017)	72.66 (3.44)	–	–	–	–	3496 (269.8)	78.6% (5.9%)
GeR-GGM	–	–	–	–	0.4 (0.017)	0.097(0.01)	0.373 (0.016)	693.9 (59.7)		

Table 2.3: Summary statistics on identification and estimation. Block-structured Ω_{YY}, $\theta = 0.1$, and D_1 and D_2 containing the same 300 samples. In each cell, mean(SD).

	$\tilde{\Omega}_{YY}$				Ω_{YY}				Hierarchy violation	
	Identification			Estimation	Identification			Estimation	Count	Proportion
	Recall	FPR	Fscore		Recall	FPR	Fscore			
$(p, q) = (50, 50)$										
Proposed	0.465 (0.065)	0.042 (0.024)	0.477 (0.024)	20.82 (0.814)	0.551 (0.067)	0.057 (0.022)	0.548 (0.026)	40.21 (1.54)	0 (0)	0 (0)
GeO-GGM	0.486 (0.098)	0.088 (0.042)	0.455 (0.025)	28.86 (1.61)	–	–	–	–	3.05 (1.4)	3.9% (1.5%)
GeR-GGM	–	–	–	–	0.458 (0.056)	0.041 (0.022)	0.521 (0.044)	49.85 (2.03)		
$(p, q) = (50, 100)$										
Proposed	0.442 (0.05)	0.05 (0.021)	0.461 (0.022)	20.51 (0.71)	0.483 (0.057)	0.038 (0.02)	0.529 (0.025)	76.99 (2.53)	0 (0)	0 (0)
GeO-GGM	0.469 (0.03)	0.109 (0.051)	0.442 (0.031)	27.807 (1.482)	–	–	–	–	37.13 (8.57)	35.4% (5.2%)
GeR-GGM	–	–	–	–	0.489 (0.043)	0.054 (0.024)	0.506 (0.03)	88.05 (3.4)		
$(p, q) = (50, 150)$										
Proposed	0.432 (0.051)	0.048 (0.029)	0.458 (0.024)	20.02 (1.04)	0.49 (0.044)	0.054 (0.021)	0.509 (0.021)	148.3 (4.3)	0 (0)	0 (0)
GeO-GGM	0.442 (0.03)	0.122 (0.055)	0.45 (0.028)	27.55 (0.987)	–	–	–	–	90.6 (22.8)	53.4% (12%)
GeR-GGM	–	–	–	–	0.394 (0.044)	0.05 (0.023)	0.44 (0.031)	167.1 (7.74)		
$(p, q) = (100, 50)$										
Proposed	0.452 (0.027)	0.083 (0.016)	0.431 (0.017)	62.68 (2.24)	0.474 (0.048)	0.081 (0.02)	0.439 (0.017)	112.1 (3.35)	0 (0)	0 (0)
GeO-GGM	0.41 (0.034)	0.086 (0.018)	0.409 (0.014)	84.48 (3.79)	–	–	–	–	62.6 (15.19)	8.8% (2%)
GeR-GGM	–	–	–	–	0.415 (0.034)	0.051 (0.017)	0.446 (0.026)	142.2 (5.77)		
$(p, q) = (100, 100)$										
Proposed	0.437 (0.034)	0.108 (0.027)	0.421 (0.018)	61.06 (1.79)	0.449 (0.039)	0.079 (0.017)	0.426 (0.015)	216.3 (5.76)	0 (0)	0 (0)
GeO-GGM	0.396 (0.016)	0.079 (0.014)	0.403 (0.015)	83.39 (2.8)	–	–	–	–	947.3 (94.38)	46.8% (5.3%)
GeR-GGM	–	–	–	–	0.398 (0.035)	0.066 (0.022)	0.403 (0.022)	268.9 (11.02)		
$(p, q) = (100, 150)$										
Proposed	0.398 (0.018)	0.088 (0.008)	0.435 (0.018)	59.91 (1.68)	0.453 (0.037)	0.087 (0.018)	0.421 (0.012)	544.1 (19.37)	0 (0)	0 (0)
GeO-GGM	0.469 (0.025)	0.084 (0.012)	0.413 (0.017)	81.55 (2.64)	–	–	–	–	3879 (300.3)	80.4% (9.5%)
GeR-GGM	–	–	–	–	0.482 (0.035)	0.241 (0.034)	0.273 (0.014)	793.3 (52.8)		

Table 2.4: Summary statistics on identification and estimation. Block-structured Ω_{YY}, $\theta = 0.05$, and D_1 and D_2 containing the same 300 samples. In each cell, mean(SD).

	$\tilde{\Omega}_{YY}$				Ω_{YY}				Hierarchy violation	
	Identification			Estimation	Identification			Estimation	Count	Proportion
	Recall	FPR	Fscore		Recall	FPR	Fscore			
$(p,q) = (50,50)$										
Proposed	0.488 (0.047)	0.017 (0.005)	0.565 (0.036)	19.36 (1.11)	0.595 (0.075)	0.015 (0.006)	0.61 (0.031)	39.29 (1.62)	0 (0)	0 (0)
GeO-GGM	0.504 (0.109)	0.031 (0.036)	0.528 (0.025)	22.81 (1.76)	–	–	–	–	6.2 (0.97)	4.9% (0.8%)
GeR-GGM	–	–	–	–	0.479 (0.099)	0.017 (0.023)	0.554 (0.049)	42.65 (1.25)		
$(p,q) = (50,100)$										
Proposed	0.483 (0.052)	0.02 (0.007)	0.553 (0.04)	19.68 (0.981)	0.552 (0.079)	0.016 (0.01)	0.599 (0.034)	75.44 (3.11)	0 (0)	0 (0)
GeO-GGM	0.526 (0.128)	0.045 (0.042)	0.517 (0.031)	23.63 (0.795)	–	–	–	–	21 (8.9)	29.6% (9.74%)
GeR-GGM	–	–	–	–	0.428 (0.078)	0.01 (0.008)	0.544 (0.046)	88.05 (3.4)		
$(p,q) = (50,150)$										
Proposed	0.476 (0.056)	0.022 (0.006)	0.547 (0.041)	19.03 (0.716)	0.5 (0.06)	0.018 (0.009)	0.572 (0.038)	147.7 (4.12)	0 (0)	0 (0)
GeO-GGM	0.476 (0.075)	0.03 (0.03)	0.519 (0.057)	27.55 (0.987)	–	–	–	–	64.8 (5.34)	72.9% (9.9%)
GeR-GGM	–	–	–	–	0.444 (0.036)	0.023 (0.008)	0.509 (0.023)	167.1 (7.74)		
$(p,q) = (100,50)$										
Proposed	0.433 (0.03)	0.033 (0.005)	0.449 (0.016)	59.91 (2.14)	0.479 (0.031)	0.02 (0.005)	0.494 (0.021)	109.3 (3.31)	0 (0)	0 (0)
GeO-GGM	0.4 (0.053)	0.029 (0.008)	0.428 (0.032)	73.92 (3.62)	–	–	–	–	45.6 (12.24)	7.2% (1.9%)
GeR-GGM	–	–	–	–	0.391 (0.031)	0.02 (0.005)	0.454 (0.027)	142.2 (5.77)		
$(p,q) = (100,100)$										
Proposed	0.446 (0.023)	0.032 (0.005)	0.462 (0.02)	59.03 (2.01)	0.443 (0.043)	0.027 (0.01)	0.476 (0.022)	215.1 (6.03)	0 (0)	0 (0)
GeO-GGM	0.408 (0.023)	0.03 (0.011)	0.434 (0.022)	72.14 (3.12)	–	–	–	–	892 (70.19)	44.7% (4.2%)
GeR-GGM	–	–	–	–	0.369 (0.027)	0.021 (0.006)	0.43 (0.014)	244 (7.86)		
$(p,q) = (100,150)$										
Proposed	0.433 (0.031)	0.035 (0.005)	0.454 (0.023)	59.91 (1.68)	0.447 (0.04)	0.026 (0.007)	0.462 (0.021)	544.6 (18.83)	0 (0)	0 (0)
GeO-GGM	0.415 (0.029)	0.037 (0.009)	0.423 (0.021)	73.5 (2.39)	–	–	–	–	3735.9 (298.4)	78.5% (8.8%)
GeR-GGM	–	–	–	–	0.448 (0.029)	0.12 (0.015)	0.275 (0.009)	719.6 (43.01)		

Table 2.5: Summary statistics on identification and estimation. Homogeneous Ω_{YY}, $\theta = 0.1$, and D_1 and D_2 containing non-overlapping samples. In each cell, mean(SD).

	$\tilde{\Omega}_{YY}$				Ω_{YY}				Hierarchy violation	
	Identification			Estimation	Identification			Estimation	Count	Proportion
	Recall	FPR	Fscore		Recall	FPR	Fscore			
$(p, q) = (50, 50)$										
Proposed	0.478 (0.04)	0.01 (0.004)	0.62 (0.036)	15.76 (0.59)	0.602 (0.04)	0.016 (0.004)	0.7 (0.031)	39.29 (1.62)	0 (0)	0 (0)
GeO-GGM	0.482 (0.03)	0.042 (0.008)	0.572 (0.029)	18.82 (0.96)	–	–	–	–	5.2 (3.9)	7.9% (4.8%)
GeR-GGM	–	–	–	–	0.58 (0.063)	0.033 (0.016)	0.638 (0.039)	46.76 (2.3)		
$(p, q) = (50, 100)$										
Proposed	0.482 (0.03)	0.014 (0.004)	0.615 (0.03)	15.31 (0.52)	0.571 (0.035)	0.014 (0.004)	0.68 (0.023)	75.78 (2.58)	0 (0)	0 (0)
GeO-GGM	0.504 (0.044)	0.055 (0.008)	0.581 (0.021)	17.77 (0.8)	–	–	–	–	21 (8.9)	29.6% (9.74%)
GeR-GGM	–	–	–	–	0.542 (0.043)	0.051 (0.017)	0.578 (0.031)	86.77 (4.8)		
$(p, q) = (50, 150)$										
Proposed	0.491 (0.024)	0.034 (0.015)	0.616 (0.022)	15.1 (0.491)	0.501 (0.043)	0.009 (0.003)	0.638 (0.037)	150.3 (5.2)	0 (0)	0 (0)
GeO-GGM	0.472 (0.026)	0.057 (0.01)	0.575 (0.028)	17.88 (1.142)	–	–	–	–	97 (24.88)	57.3% (11.5%)
GeR-GGM	–	–	–	–	0.432 (0.037)	0.059 (0.012)	0.463 (0.039)	173 (6.88)		
$(p, q) = (100, 50)$										
Proposed	0.568 (0.02)	0.044 (0.002)	0.6 (0.014)	36.71 (0.81)	0.61 (0.021)	0.022 (0.002)	0.68 (0.015)	109.8 (3.45)	0 (0)	0 (0)
GeO-GGM	0.468 (0.031)	0.033 (0.006)	0.559 (0.029)	46.49 (0.91)	–	–	–	–	150.9 (20.66)	26.4 % (2.8%)
GeR-GGM	–	–	–	–	0.506 (0.033)	0.033 (0.007)	0.568 (0.023)	135.2 (3.5)		
$(p, q) = (100, 100)$										
Proposed	0.53 (0.017)	0.022 (0.002)	0.592 (0.02)	35.97 (0.62)	0.605 (0.023)	0.026 (0.003)	0.666 (0.018)	215.1 (6.03)	0 (0)	0 (0)
GeO-GGM	0.453 (0.028)	0.028 (0.005)	0.548 (0.026)	46.47 (1.62)	–	–	–	–	912.4 (90.44)	68.4% (6.6%)
GeR-GGM	–	–	–	–	0.418 (0.027)	0.044 (0.013)	0.476 (0.024)	263 (9.3)		
$(p, q) = (100, 150)$										
Proposed	0.553 (0.031)	0.042 (0.009)	0.572 (0.013)	35.72 (0.78)	0.532 (0.02)	0.016 (0.002)	0.637 (0.016)	545.9 (25.16)	0 (0)	0 (0)
GeO-GGM	0.446 (0.043)	0.042 (0.022)	0.534 (0.023)	45.8 (1.6)	–	–	–	–	4162.3 (338.3)	87.3% (7.54%)
GeR-GGM	–	–	–	–	0.428 (0.017)	0.157 (0.021)	0.294 (0.022)	728.7 (55.09)		

Table 2.6: Summary statistics on identification and estimation. Homogeneous Ω_{YY}, $\theta = 0.1$, and D_1 and D_2 containing non-overlapping samples. In each cell, mean(SD).

	$\tilde{\Omega}_{YY}$				Ω_{YY}				Hierarchy violation	
	Identification			Estimation	Identification			Estimation	Count	Proportion
	Recall	FPR	Fscore		Recall	FPR	Fscore			
$(p,q) = (50,50)$										
Proposed	0.512 (0.02)	0.058 (0.009)	0.576 (0.021)	15.07 (0.54)	0.558 (0.052)	0.033 (0.008)	0.623 (0.03)	39.62 (1.48)	0 (0)	0 (0)
GeO-GGM	0.428 (0.024)	0.032 (0.008)	0.524 (0.026)	17.46 (0.68)	–	–	–	–	4.1 (2.6)	9% (5.3%)
GeR-GGM	–	–	–	–	0.558 (0.052)	0.039 (0.017)	0.597 (0.03)	46.87 (2.33)		
$(p,q) = (50,100)$										
Proposed	0.506 (0.031)	0.052 (0.029)	0.571 (0.023)	15.69 (0.41)	0.552 (0.029)	0.039 (0.006)	0.592 (0.027)	75.722 (3.09)	0 (0)	0 (0)
GeO-GGM	0.419 (0.052)	0.028 (0.009)	0.517 (0.023)	17.46 (0.68)	–	–	–	–	51.8 (10.58)	46.1% (9.7%)
GeR-GGM	–	–	–	–	0.494 (0.063)	0.034 (0.008)	0.545 (0.024)	82.92 (2.03)		
$(p,q) = (50,150)$										
Proposed	0.498 (0.041)	0.058 (0.017)	0.562 (0.019)	15.31 (0.89)	0.455 (0.018)	0.045 (0.003)	0.508 (0.014)	109.02 (3.25)	0 (0)	0 (0)
GeO-GGM	0.428 (0.024)	0.033 (0.006)	0.521 (0.025)	17 (1.48)	–	–	–	–	96.4 (19.57)	27.8% (6.5%)
GeR-GGM	–	–	–	–	0.447 (0.051)	0.061 (0.019)	0.462 (0.03)	166.9 (5.63)		
$(p,q) = (100,50)$										
Proposed	0.526 (0.04)	0.054 (0.02)	0.559 (0.019)	38.03 (0.45)	0.6 (0.037)	0.04 (0.006)	0.626 (0.032)	148.3 (4.3)	0 (0)	0 (0)
GeO-GGM	0.449 (0.016)	0.043 (0.004)	0.505 (0.014)	44.87 (0.51)	–	–	–	–	288.8 (22.12)	36.7% (2.7%)
GeR-GGM	–	–	–	–	0.485 (0.024)	0.044 (0.007)	0.528 (0.025)	136.6 (3.7)		
$(p,q) = (100,100)$										
Proposed	0.482 (0.035)	0.042 (0.002)	0.556 (0.019)	37.7 (0.73)	0.478 (0.022)	0.047 (0.004)	0.511 (0.016)	214.6 (5.54)	0 (0)	0 (0)
GeO-GGM	0.451 (0.042)	0.067 (0.024)	0.492 (0.016)	44.18 (1.61)	–	–	–	–	1004.1 (89.42)	68% (5.6%)
GeR-GGM	–	–	–	–	0.404 (0.042)	0.053 (0.011)	0.432 (0.029)	233.1 (9.9)		
$(p,q) = (100,150)$										
Proposed	0.51 (0.012)	0.075 (0.007)	0.546 (0.013)	36.84 (0.76)	0.527 (0.022)	0.061 (0.006)	0.515 (0.019)	546.1 (19.86)	0 (0)	0 (0)
GeO-GGM	0.426 (0.018)	0.062 (0.006)	0.484 (0.012)	45.3 (1.55)	–	–	–	–	4196.9 (243.9)	83.1% (4%)
GeR-GGM	–	–	–	–	0.402 (0.02)	0.165 (0.026)	0.292 (0.017)	747.2 (56.43)		

Table 2.7: Analysis of TCGA SKCM data: numbers of overlapping. In each cell, gene-expression-only/gene-expression-regulator analysis.

	Proposed	GeO-GGM	GeR-GGM
Proposed	101/97	80/–	–/29
GeO-GGM		119/–	–/–
GeR-GGM			–/99

Table 2.8: Analysis of TCGA lung cancer data: numbers of overlapping. In each cell, gene-expression-only/gene-expression-regulator analysis.

	Proposed	GeO-GGM	GeR-GGM
Proposed	284/263	150/–	–/146
GeO-GGM		278/–	–/–
GeR-GGM			–/258

Chapter 3

Project 2: Information-incorporated Gaussian graphical model for gene expression data

Abstract

In the analysis of gene expression data, network approaches take a system perspective and have played an irreplaceably important role. Gaussian graphical models (GGM) have been popular in the network analysis of gene expression data. They investigate the conditional dependence between genes and "transform" the problem of estimating network structures into a sparse estimation of precision matrices. When there is a moderate to large number of genes, the number of parameters to be estimated may overwhelm the limited sample size, leading to unreliable estimation and selection. In this article, we propose incorporating information from previous studies (for example, those deposited at PubMed) to assist estimating the network structure in the present data. It is recognized that such information can be partial, biased, or even wrong. A penalization-based estimation approach is developed, shown to have consistency properties, and realized using an effective computational algorithm. Simulation demonstrates its competitive performance under various

information accuracy scenarios. The analysis of TCGA lung cancer prognostic genes leads to network structures different from the alternatives. Research reported in this chapter has been published in *Biometrics*.

3.1 Introduction

In the analysis of gene expression data, network approaches have played important roles. They take a system perspective and examine the interconnections among genes as well as their individual properties. There have been quite a few network analysis approaches [140], among which Gaussian graphical model (GGM) has been popular because of its lucid interpretations, satisfactory statistical properties, and computational advantages. GGM assumes the multivariate normal distribution, under which the conditional independence of two nodes is equivalent to a zero value of the corresponding element in the precision matrix. As such, determining the network structure amounts to a sparse estimation of the precision matrix, for which penalization and other techniques have been adopted. For GGM estimation, we refer to reviews including [141,142]. Relevant works also include [107, 143, 144], and references therein. It is recognized that practical gene expression data may have distributions deviated from normal. Nevertheless, with proper data processing, the GGM technique has been extensively applied to gene expression data and led to interesting findings.

With p genes, the number of conditional dependence to be estimated is $p(p-1)/2$ and may easily exceed the sample size, leading to unreliable estimation and selection. Multiple remedies have been developed, including jointly analyzing multiple independent datasets to increase power (which demands the availability of data from studies other than the present one), jointly analyzing gene expressions and their regulators (which demands multidimensional profiling and the availability of data on regulators), imposing structural constraints (which demands additional information or assumptions on network structure), among others.

Our goal is to improve gene expression GGM analysis via borrowing additional information, which is similar to that of some existing studies [114, 118]. Different from these studies, we consider additional information contained in publications. To make the idea clearer, we conduct a simple search of "gene TP53, gene MKI67, lung cancer" in PubMed and get 23 hits. In comparison, the search of "gene TP53, gene CD68, lung cancer" leads to zero hit. That is, at least 23 publications have simultaneously reported genes TP53 and MKI67 as well as lung cancer, compared to none for genes TP53 and CD68. Published articles represent a large number of past studies and contain valuable information. On the other hand, we also note that information retrieved from a simple search may be crude and not fully trustworthy. For example, although the article by [145] also comes up in the search, the context does not seem to be related to lung cancer, and there is a lack of direct suggestion on the interconnection between genes TP53 and MKI67. Further, it is also unclear whether the interconnections between genes TP53 and MKI67 reported in these publications are conditional or unconditional. Nevertheless, even with just this simple search, it may be "safe" to conclude that genes TP53 and MKI67 are more likely to be interconnected compared to genes TP53 and CD68.

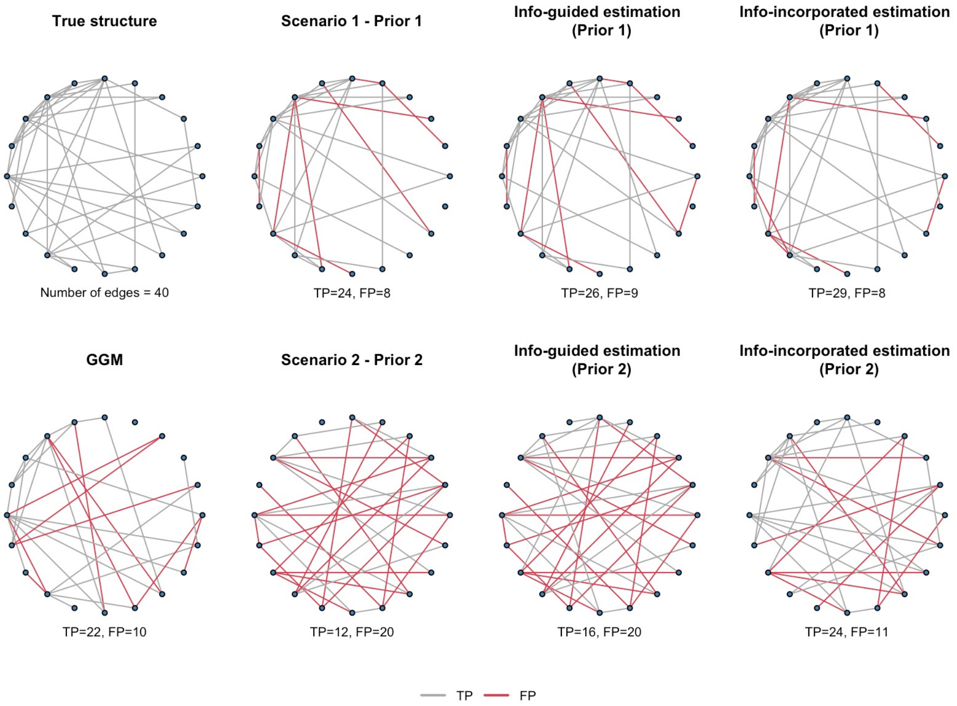


Figure 3.1: Small example. The first column: upper – true structure, lower – GGM; The second to fourth columns: prior information, information-guided analysis, and information-incorporated analysis. Upper and lower: two scenarios of prior information.

A small example We use a small example to provide some insights into the proposed approach. In Figure 3.1, we present a graph with 40 edges. Directly applying the GGM leads to 22 TPs and 10 FPs, where TP/FP stand for true/false positive. We first consider a set of high-quality information with 24 TPs and 8 FPs. The information-guided estimation (defined below in Step II) retains all those TPs and FPs and also add 2 TPs and 1 FP. The proposed information-incorporated estimation (defined below in Step III) further refines and leads to 29 TPs and 8 FPs, improving over the GGM. We also consider an alternative scenario with low-quality information (12 TPs and 20 FPs). In this case, the proposed approach can data-dependently and effectively "disregard" incorrect information and leads to 24 TPs and 11 FPs – an overall performance comparable to the GGM. Although small, this example can already suggest that the proposed analysis has the potential for improving performance by incorporating additional information.

This article is built on the existing GGM studies for gene expression data. To improve the often-unsatisfactory performance caused by a lack of information, our proposal is to incorporate additional information. Our goal is to develop an approach that not only has satisfactory numerical performance but also is theoretically well-grounded. Compared to that from other sources, additional information retrieved from publications can be advantageous in multiple ways. For example, it is built on a large number of (mostly independent) published studies and, in a sense, can be more reliable than that from a few additional datasets. It is more cost-effective and does not involve collecting additional data. In existing works, information contained in published literature has been utilized in multiple ways. Some draw conclusions based solely on such information [146,147]. They differ significantly from our analysis by not having a present dataset to be analyzed. Some other works conduct qualitative and quantitative comparisons between the present data analysis results with those in the literature, to verify the present findings. They also differ significantly from our analysis by not considering the additional information in the estimation procedure. Our analysis is more aligned with that of [123] and [14], both of which use information in publications to assist the present penalized regression analysis. On the other hand, it differs from these two studies by conducting GGM-based analysis, which has significantly different data structures, analysis goals, and objective function. For GGM, there have been studies that improve the present analysis by shrinking the estimate towards a known target [148, 149]. They demand very detailed information (e.g., not only whether two nodes are connected but also the strength of the edge), making it not feasible to accommodate a large number of published findings. There are also information-incorporated studies in other domains, for example Bayesian [150,151], whose strategies are fundamentally different from the proposed. Overall, with the popularity of GGM for gene expression data, often-unsatisfactory performance when directly applying GGM, and differences from the existing information-incorporated analyses, this study is warranted beyond the existing works.

3.2 Methods

Let $X = (X^{(1)}, \cdots, X^{(p)})$ denote the $p-$dimensional gene expression measurements with a multivariate normal distribution $\mathcal{N}_p(\mu, \Sigma)$. Discussions on alternative distributions are provided below. Denote the precision matrix as $\Theta = \Sigma^{-1}$ and the corresponding graph as $\mathcal{G} = (V, E)$, where $V = \{1, \cdots, p\}$ and the edges $E = (e_{ij})_{1\leq i<j\leq p}$ describe the conditional independence relationships among $X^{(1)}, \cdots, X^{(p)}$. The edge between $X^{(i)}$ and $X^{(j)}$ is absent if and only if $X^{(i)}$ and $X^{(j)}$ are independent conditional on the other variables, which corresponds to $\theta_{ij} = 0$.

With n iid samples $\{X_1, \cdots, X_n\}$, up to a constant, the negative log-likelihood function is

$$L(\Theta; S) = -\log|\Theta| + \mathrm{tr}(S\Theta),$$

where $S = (s_{ij})$ is the sample covariance matrix, and $|\Theta|$ is the determinant of Θ. To regularize estimation and generate interpretable networks, penalized estimation has been developed. Denote $p_{\lambda_n}(\cdot)$ as the penalty function, where λ_n is a tuning parameter. The objective function is

$$L_{\lambda_n}(\Theta; S) = L(\Theta; S) + \sum_{i\neq j} p_{\lambda_n}(|\theta_{ij}|).$$

To incorporate information contained in published literature to improve performance of the penalized GGM, our approach consists of the following main steps:

Step I: Information retrieval In our data analysis, we use PubMatrix [152], a publicly available text mining tool, to mine PubMed. It conducts a search of the co-occurrence of two lists of keywords, which in our case are two identical lists of genes. If desirable, the context of analysis, for example "lung cancer", can be further added. It delivers the number of publications in PubMed that include a specific pair of genes. A simple demonstration of the submit and result pages is provided in Figure 3.4 in Appendix, Section 3.6.2. Denote $\mathcal{E}^p$ as the index set of retrieved gene connections. That is, if $(i, j) \in \mathcal{E}^p$, then there are suggestions that genes i and j are interconnected. $\mathcal{E}^p$ is symmetric: if $(i, j) \in \mathcal{E}^p$, $(j, i) \in \mathcal{E}^p$.

We acknowledge that not all publications are included in PubMed, and, as mentioned in Section 3.1, the retrieved information may not be fully relevant. It can also be partial or even wrong. With more recent and sophisticated text mining tools [153], it may be possible to refine the mining and remove some irrelevant "findings". Text mining is a moving field, and more sophisticated tools can sometimes be harder to implement. In addition, text mining usually cannot identify incorrect information contained in literature. As such, it may not be possible to obtain fully accurate information. Information can also be manually scrutinized. However, it is not practical with a large number of genes and relevant publications. Luckily, as discussed below and shown in simulation, the proposed approach does not demand the full accuracy of retrieved information and can "robustly" accommodate partial and incorrect information.

Step II: Information-guided analysis Consider the penalized objective function:

$$L_{\gamma_n, \mathcal{E}^p}(\Theta; S) = L(\Theta; S) + \sum_{(i,j) \notin \mathcal{E}^p} p_{\gamma_n}(|\theta_{ij}|), \tag{3.1}$$

where notations have similar implications as above. In this analysis, the retrieved information is *fully trusted.* That is, if two genes have been suggested as interconnected in the literature, we automatically include the corresponding edge via not imposing penalty. Penalization is imposed on other parameters to search for additional signals. Denote the estimator from (3.1) as $\hat{\Theta}^p_{\gamma_n}$. Compute $\hat{\Sigma}^p = (\hat{\Theta}^p_{\gamma_n})^{-1}$. For both (3.1) and this inversion, if needed, a small ridge penalty can be imposed to stabilize estimation. $\hat{\Sigma}^p$ is the artificial covariance matrix when the network sparsity structure estimation is guided by the retrieved information.

It is noted that, in (3.1), we only account for whether there is any evidence but not the amount of evidence. This is due to concerns on the potential "research bias" of published studies, "selection bias" of PubMed, crudeness of our text mining, and other factors.

Step III: Information-incorporated analysis Consider the penalized objective function:

$$L_{\lambda_n,\eta}(\Theta; S, \hat{\Sigma}^p) \quad = \quad L(\Theta; S) + \eta L(\Theta; \hat{\Sigma}^p) + \sum_{i \neq j} p_{\lambda_n}(|\theta_{ij}|),$$

where $\eta \geq 0$ is a tuning parameter, and the other notations have the same implications as above. It can also be rewritten as:

$$\begin{aligned} L_{\lambda_n,\eta}(\Theta; S, \hat{\Sigma}^p) &= -(1+\eta)\log|\Theta| + \operatorname{tr}\left\{(S + \eta\hat{\Sigma}^p)\Theta\right\} + \sum_{i \neq j} p_{\lambda_n}(|\theta_{ij}|), \\ &= (1+\eta) L_{\frac{\lambda_n}{1+\eta}}(\Theta; \tilde{S}_\eta), \qquad (3.2) \end{aligned}$$

where $\tilde{S}_\eta = (S + \eta\hat{\Sigma}^p)/(1+\eta)$ is the weighted sum of the observed sample covariance matrix and that obtained in Step II. Denote $\hat{\Theta}_{\lambda_n,\eta} = \arg\min L_{\lambda_n,\eta}(\Theta; S, \hat{\Sigma}^p)$ as the final information-incorporated GGM estimate.

In this step of analysis, the penalty has the same implication as in a standard GGM. The goodness-of-fit has two components: one from the observed data and the other from the information-guided analysis. η is introduced to data-dependently balance them. Intuitively, when the additional information has higher quality, a larger η is preferred and can lead to more utilization of such information. On the other hand, when the quality of the additional information is poor, a small η value can lead to analysis basically relying on the observed data. As such, the proposed approach has the potential to incorporating additional information while flexibly allowing it to be not fully accurate.

Remarks We acknowledge that the proposed way of incorporating information is not sufficiently refined. For example, it is possible that multiple publications report the same two genes, but their conclusions contradict. That is, there is uncertainty in the available information, which is also related to the irreproducibility of findings. The present text mining cannot identify such conflicting/uncertain information. As such, it is not accommodated in the proposed analysis. This is the price paid for mining a large number of publications and gene pairs. If an information uncertainty measure is available for each gene pair, we conjecture that it is possible to revise Step II: instead of automatically including those suggested in the literature, weighted penalization can be applied to make those with less conflicting/uncertain information more easily selected.

3.2.1 Statistical properties

Denote $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ as the minimum and maximum eigenvalues of a symmetric matrix A. For a matrix B, define the Frobenius norm as $\|B\|_F = \text{tr}^{1/2}(B^T B)$. Denote $\mathcal{A} = \{(i,j) : \theta^0_{ij} \neq 0\}$ and $\mathcal{A}^c = \{(i,j) : \theta^0_{ij} = 0\}$. Further denote $\mathcal{A}^- = \{(i,j) : i \neq j, \theta^0_{ij} \neq 0\}$, and $s = |\mathcal{A}^-|$ as the size of $\mathcal{A}^-$. Let $a_n = \max_{(i,j)\in\mathcal{A}^-} p'_{\lambda_n}(|\theta^0_{ij}|), b_n = \max_{(i,j)\in\mathcal{A}^-} p''_{\lambda_n}(|\theta^0_{ij}|)$. Assume the following conditions.

(C1) There are constants ϕ_1 and ϕ_2 such that $0 < \phi_1 < \lambda_{\min}(\Sigma_0) \leq \lambda_{\max}(\Sigma_0) < \phi_2 < \infty$.

(C2) $a_n = O\left(\frac{1}{1+\eta}\sqrt{\frac{(p+s)\log p}{(s+1)n}}\right)$, $b_n = o(1)$, and $\min_{(i,j)\in\mathcal{A}} |\theta^0_{ij}|/\lambda_n \to \infty$ as $n \to \infty$.

(C3) $p_{\lambda_n}(\cdot)$ is singular at the origin, with $\lim_{t\downarrow 0} p_{\lambda_n}(t)/(\lambda_n t) = k > 0$.

(C4) There are constants C and D such that, when $\theta_1, \theta_2 > C\lambda_n$, $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \leq D|\theta_1 - \theta_2|$.

(C5) $\hat{\Sigma}^p = (\hat{\sigma}^p_{ij})$ satisfies that $\max_{i,j} |\hat{\sigma}^p_{ij} - \sigma^0_{ij}| \leq \frac{C_0}{\eta}\sqrt{\frac{\log p}{n}}$ with probability tending to 1, where C_0 is a large positive constant.

Conditions (C1)-(C4) have been commonly assumed in the literature [113]. Multiple penalties satisfy the above assumptions, and the minimax concave penalty (MCP) is adopted in our numerical study. (C5) concerns with the information-guided estimator. With linear regression, a similar condition has been assumed in [123]. It shows the connection between how reliable the information-guided estimator is and how much we should "trust" the additional information. More specifically, more reliable additional information leads to more reliable information-guided estimation, which in turn results in a larger η value. Lastly, a larger η leads to a more accurate information-incorporated estimator. When $\eta = 0$, Condition (C5) is automatically satisfied, and Theorem 1 reduces to the results in [113].

Theorem 2. *Suppose that Conditions (C1)-(C5) hold. If* $\frac{(p+s)\log p}{n(1+\eta)^2} = o(1)$ *and* $\frac{(p+s)\log p}{n} = O(\lambda_n^2)$*, then there exists a local minimizer* $\hat{\Theta}_{\lambda_n,\eta}$ *of* $L_{\lambda_n,\eta}(\Theta; S, \hat{\Sigma}^p)$ *such that*

$$\|\hat{\Theta}_{\lambda_n,\eta} - \Theta_0\|_F^2 = O_p\left(\frac{(p+s)\log p}{n(1+\eta)^2}\right),$$

and with probability tending to 1,

$$sign(\hat{\theta}_{ij}^{\lambda_n,\eta}) = sign(\theta_{ij}^0).$$

Remarks and proof are presented in the Supporting Information S1. With the GGM framework, the result and proof differ significantly from those for linear regression [123]. With the additional information, they also differ considerably from those for an "ordinary" GGM.

Alternatively, the objective function can be written as:

$$L_{\lambda_n,\tau}(\Theta; S, \hat{\Sigma}^p) = (1-\tau)L(\Theta; S) + \tau L(\Theta; \hat{\Sigma}^p) + \sum_{i \neq j} p_{\lambda_n}(|\theta_{ij}|),$$

where $\tau \in [0, 1]$, and a larger value of τ corresponds to more reliable information. Then $L_{\lambda_n,\tau}(\Theta; S, \hat{\Sigma}^p)$ can be rewritten as

$$\begin{aligned} L_{\lambda_n,\tau}(\Theta; S, \hat{\Sigma}^p) &= -\log|\Theta| + (1-\tau)\mathrm{tr}(S\Theta) + \tau\mathrm{tr}(\hat{\Sigma}^p\Theta) + \sum_{i \neq j} p_{\lambda_n}(|\theta_{ij}|) \\ &= -\log|\Theta| + \mathrm{tr}\left[\{(1-\tau)S + \tau\hat{\Sigma}^p\}\Theta\right] + \sum_{i \neq j} p_{\lambda_n}(|\theta_{ij}|) \\ &= L_{\lambda_n}(\Theta; \tilde{S}_\tau), \end{aligned} \tag{3.3}$$

where $\tilde{S}_\tau = (1-\tau)S + \tau\hat{\Sigma}^p$. Let $\hat{\Theta}_{\lambda_n,\tau} = \arg\min L_{\lambda_n,\tau}(\Theta; S, \hat{\Theta}^p)$. We modify the assumed conditions as:

(C2') $a_n = O\left((1-\tau)\sqrt{\frac{(p+s)\log p}{(s+1)n}}\right)$, $b_n = o(1)$, and $\min_{(i,j)\in\mathcal{A}} |\theta_{ij}^0|/\lambda_n \to \infty$ as $n \to \infty$.

(C5') $\max_{i,j} |\hat{\sigma}_{ij}^p - \sigma_{ij}^0| \leq \frac{1-\tau}{\tau} C_0 \sqrt{\frac{\log p}{n}}$ with probability tending to 1, where C_0 is a large positive constant.

Then, under Conditions (C1), (C2'), (C3), (C4) and (C5'), if $(1-\tau)^2\frac{(p+s)\log p}{n} = o(1)$ and $\frac{(p+s)\log p}{n} = O(\lambda_n^2)$, we can show that:

$$\|\hat{\Theta}_{\lambda_n,\tau} - \Theta_0\|_F^2 = O_p\left((1-\tau)^2\frac{(p+s)\log p}{n}\right),$$

and $\hat{\theta}_{ij}^{\lambda_n,\tau} = 0$ for all $(i,j) \in \mathcal{A}^c$, with probability tending to 1.

3.2.2 Computation

A significant advantage of the proposed approach is that it does not demand new computational development. Specifically, for computing the information-guided estimator, the existing algorithms [154] can be applied by setting tunings as zero for components of the precision matrix corresponding to edges in $\mathcal{E}^p$. For computing the information-incorporated estimator, with the rewritten function (3.2) (or (3.3)), the existing algorithms can be directly applied. Convergence, computational complexity, and other results follow the literature [155, 156]. For selecting γ_n, λ_n, and η, we conduct a three-dimensional grid search and adopt cross validation-type techniques (more details below). In addition, in simulation, we also consider the ROC (Receiver Operating Characteristic) and other techniques, which can "reduce" the impact of tuning parameter selection. R programs for implementing the proposed approach and a demonstrating example are available at `www.github.com/shuanggema`.

3.3 Simulation

For the structure of the precision matrix, we consider four popular choices, namely the Erdos-Renyi, scale-free, nearest-neighbor, and banded (positive and negative) structures. Briefly, we generate the Erdos-Renyi network with two sub-networks that have probabilities 0.05 and 0.07 for drawing an edge between two arbitrary nodes. The scale-free network is generated using the popular Barabasi-Albert algorithm. It starts with an initial connected network with a small number of nodes. New nodes are added to the network one at a time, and each new node is connected to a certain number of existing nodes with a probability that is proportional to the number of edges that the existing nodes already have. The nearest-neighbor network is generated by modifying the data generating mechanism described in [157]. Specifically, we generate p points randomly on a unit square, calculate all pairwise distances, and find the k nearest neighbors of each point. The nearest-neighbor network is obtained by linking any two points that are among the k-nearest neighbors of each other. k controls the sparsity level of the network, and we set $k = 4$ in our simulation. With the Banded(+) network, the precision matrix has a block-diagonal structure with 7 blocks (for $p = 50$) and 13 blocks (for $p = 100$). Within each block, the band has width ranging from two to four (on each side, and diagonal not included). All nonzero off-diagonal elements are positive. With the Banded(-) network, the precision matrix is similar to that with the Banded(+), except that the nonzero elements for adjacent nodes are negative. The average numbers of total edges are presented in the simulation tables.

For the Erdos-Renyi, scale-free, and nearest-neighbor networks, the precision matrices are determined based on the corresponding network structures. In particular, elements not corresponding to edges are set as zero. For elements corresponding to edges, we generate their values randomly from a uniform distribution with support $[-0.4, -0.1] \cup [0.1, 0.4]$ – this setting is referred to as "strong signal". In addition, we also consider the "weak signal" setting, where elements are equal to 80% of those under the strong signal setting. To ensure positive definiteness, we set $\theta_{ii} = \sum_{j \neq i} \theta_{ij} + 0.1$. For the banded networks, the precision matrices are directly generated. Finally, the covariance matrix $\Sigma = \Theta^{-1}$. We consider $p = 50$ and 100, with corresponding sample sizes 100 and 300.

As discussed above, additional information may not be fully correct. To examine this aspect, we consider four scenarios. Under Scenario 1, information is 100% correct. That is, it contains all TPs and no FPs. Under Scenarios 2-4, respectively, about 70%, 50%, and 30% of the information is correct. More detailed information on the numbers of TPs and FPs in the additional information are provided in Table 3.2 (Section 3.6).

To better gauge performance of the proposed approach, we consider the following alternatives. The first is the benchmark, which couples the GGM with MCP penalization (referred to as "GGM"). This approach is based on the observed data only. The second is the information-guided analysis, under which the additional information is fully trusted.

The third is generalized gLasso [148,149], which is recent and has competitive performance. It demands a target precision matrix, which is generated as follows. With the true precision matrix, we compute the standard deviation (sd) of all nonzero elements. The target matrix has elements that correspond to the additional information being nonzero (and the rest being zero). For these nonzero elements, we add random "perturbations" generated from N(0, sd) or N(0, sd/2) to the true values. The two settings are referred to as L and S, standing for large and small perturbations, respectively. In this process, the symmetry and positive definiteness need to be preserved. It is noted that, as the target sensibly differs from the true, and also with the complexity of GGM, a larger perturbation not necessarily corresponds to worse performance. We also note that there exist more remotely related alternatives. Comparing with the above three approaches can the most directly establish the merit of the proposed approach.

When evaluating the proposed and alternative approaches, of the most interest is edge identification. For each generated dataset (training), we simulate an independent testing dataset under the same settings. Estimates are generated using the training data, and optimal tunings are selected based on the likelihood computed using the "testing data + training data estimates". The TP and FP rates are calculated under the optimal tunings. This procedure closely mimics and is computationally simpler than cross validation, and has been adopted in multiple studies. Under each setting, 100 replicates are simulated. Results for $p = 50$ and 100 are summarized in Table 3.1 and Table 3.3 (Section 3.6.2), respectively. Across the whole spectrum of simulation, the proposed analysis is observed to have competitive performance. Consider for example Table 3.1, the ER network, weak signal, and additional information Scenario 1. The average (TP, FP) values are (41.8, 11.7) for GGM, (139.8, 0.0) for the information-guided analysis, (149.7, 0.4) for the information-incorporated analysis, (107.7, 57.9) for generalized gLasso-L, and (105.1, 56.8) for generalized gLasso-S. As another example, consider Table 3.1, the Banded(-) network, strong signal, and additional information Scenario 4. The (TP, FP) values are (175.2, 45.7) for GGM, (151.6, 50.9) for the information-guided analysis, (178.9, 32.2) for the information-incorporated analysis, (174.5, 179.0) for generalized gLasso-L, and (175.7, 189.6) for generalized gLasso-S. In general, performance of the proposed approach deteriorates as the quality of additional information

Table 3.1: Simulation results with selected optimal tunings: mean (sd) TPs and FPs for $p = 50$. (ER: Erdos-Renyi; SF: scale-free; NN: nearest-neighbor; Banded(+): positive banded; and Banded(-): negative banded. L/S with generalized gLasso: large/small perturbations added.)

		ER		SF		NN		Banded(+)		Banded(-)	
	Total edges	152		200		184		244		244	
		TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
Weak signal											
	GGM	41.8(12.6)	11.7(8.0)	23.5(8.2)	10.1(7.6)	51.6(11.6)	21.9(14.2)	96.1(11.6)	30.8(15.0)	107.7(15.1)	37.4(22.8)
Scenario 1	Info-guided	139.8(4.2)	0.0(0.0)	178.3(6.2)	0.0(0.0)	168.7(4.6)	0.0(0.0)	229.9(3.9)	0.0(0.0)	230.1(5.0)	0.0(0.0)
	Info-incorporated	149.7(2.1)	0.4(0.8)	188.9(4.8)	0.7(1.4)	180.9(2.1)	1.1(1.7)	240.2(4.2)	3.6(3.9)	240.8(3.2)	3.1(4.1)
	Generalized gLasso – L	107.7(5.8)	57.9(9.5)	131.4(7.9)	162.4(79.1)	129.9(6.8)	192.4(66.4)	177.5(7.5)	79.1(14.5)	185.3(6.7)	75.1(10.3)
	Generalized gLasso – S	105.1(6.5)	56.8(9.0)	120.7(6.5)	170.4(77.7)	127.7(6.8)	217.6(35.5)	171.4(8.2)	70.9(21.1)	169.9(8.0)	68.7(14.5)
Scenario 2	Info-guided	129.3(4.3)	3.9(2.3)	128.5(5.3)	30.2(3.7)	132.9(4.9)	25.7(4.9)	153.0(12.9)	52.9(14.2)	150.5(12.1)	54.9(16.1)
	Info-incorporated	138.3(2.7)	4.1(2.4)	135.3(4.9)	28.0(3.5)	137.9(6.2)	19.5(6.2)	175.1(16.5)	36.1(15.5)	167.5(17.6)	33.4(14.5)
	Generalized gLasso – L	106.7(5.6)	64.3(10.2)	117.2(10.5)	198.9(71.6)	124.9(6.5)	241.1(34.0)	156.9(9.8)	153.1(33.7)	153.6(10.6)	157.0(42.2)
	Generalized gLasso – S	103.0(6.1)	63.1(10.6)	103.7(11.8)	146.4(76.4)	119.6(8.4)	191.4(76.8)	144.7(7.6)	108.5(18.1)	139.6(6.2)	112.8(12.8)
Scenario 3	Info-guided	104.2(6.0)	27.7(5.0)	95.9(5.6)	49.0(6.8)	101.1(8.3)	51.2(10.6)	139.7(13.9)	81.4(24.5)	153.1(16.4)	91.6(23.0)
	Info-incorporated	107.3(7.7)	19.1(7.9)	96.4(20.8)	36.4(16.4)	92.9(10.0)	16.6(8.2)	142.0(13.6)	35.3(12.0)	153.3(15.0)	40.2(13.9)
	Generalized gLasso – L	96.1(6.1)	85.4(9.9)	95.7(15.1)	192.0(76.5)	110.3(13.4)	218.5(73.8)	147.8(13.2)	189.8(58.2)	155.9(12.3)	188.2(53.5)
	Generalized gLasso – S	92.7(5.8)	77.8(11.2)	90.2(12.4)	147.9(78.2)	108.9(14.5)	166.5(87.2)	129.2(6.8)	116.0(20.5)	146.3(8.6)	102.5(18.0)
Scenario 4	Info-guided	63.5(7.6)	55.3(11.2)	50.3(5.4)	91.2(9.2)	71.3(16.2)	100.5(16.7)	117.9(13.9)	114.0(23.4)	123.0(17.1)	114.5(22.4)
	Info-incorporated	65.7(8.1)	14.6(6.8)	39.0(11.0)	13.1(8.8)	65.2(9.7)	19.3(11.1)	124.7(18.0)	38.1(16.7)	130.7(18.2)	39.1(17.8)
	Generalized gLasso – L	70.1(4.8)	105.7(10.8)	69.9(16.1)	187.1(68.7)	104.1(17.1)	269.9(63.0)	128.3(15.3)	214.3(49.1)	135.7(18.0)	220.7(52.9)
	Generalized gLasso – S	70.8(5.7)	102.3(12.3)	65.4(14.8)	153.2(60.7)	79.7(12.9)	113.7(52.4)	112.1(11.3)	129.7(23.1)	124.9(8.7)	133.8(12.8)
Strong signal											
	GGM	73.5(13.5)	18.3(11.8)	63.3(10.6)	28.9(12.6)	85.5(11.2)	26.3(10.8)	158.3(33.0)	41.3(25.8)	175.2(21.7)	45.7(20.1)
Scenario 1	Info-guided	146.9(2.8)	0.0(0.0)	188.0(5.5)	0.0(0.0)	177.4(4.0)	0.0(0.0)	239.7(2.7)	0.1(0.4)	239.0(2.9)	0.0(0.0)
	Info-incorporated	150.3(2.4)	1.2(1.4)	193.8(3.5)	2.9(2.6)	182.0(3.2)	3.3(2.8)	227.9(15.2)	8.1(9.5)	229.1(12.3)	4.4(5.7)
	Generalized gLasso – L	130.3(5.5)	88.4(39.5)	158.4(7.5)	201.6(74.6)	146.3(8.0)	119.3(76.8)	237.9(3.3)	84.1(55.6)	237.3(3.4)	176.1(20.6)
	Generalized gLasso – S	114.1(6.7)	65.9(13.2)	151.2(8.7)	197.4(63.1)	140.0(9.1)	165.5(81.4)	237.8(2.8)	56.8(46.5)	236.1(3.4)	109.4(49.3)
Scenario 2	Info-guided	136.9(3.8)	4.1(2.5)	140.4(5.7)	30.1(4.8)	143.8(4.1)	24.9(4.4)	178.5(18.0)	32.9(16.4)	175.9(20.6)	27.8(11.7)
	Info-incorporated	140.3(3.3)	3.9(3.0)	143.5(6.4)	27.7(7.6)	145.3(6.3)	15.6(7.3)	187.1(21.0)	23.6(14.5)	191.9(17.0)	22.3(13.9)
	Generalized gLasso – L	130.9(5.5)	88.2(9.9)	140.3(9.2)	251.1(56.3)	139.7(11.1)	206.3(74.5)	203.9(6.3)	208.2(31.7)	201.3(8.7)	190.8(29.9)
	Generalized gLasso – S	117.1(6.5)	75.6(9.3)	125.3(12.5)	204.5(69.7)	132.9(13.5)	178.8(81.0)	191.7(13.7)	170.7(58.0)	203.6(5.3)	207.7(19.8)
Scenario 3	Info-guided	119.5(6.2)	25.9(8.3)	113.6(8.3)	55.9(12.8)	117.1(11.5)	52.9(13.6)	168.1(27.5)	45.2(18.2)	176.3(18.6)	37.3(16.5)
	Info-incorporated	119.5(5.4)	12.7(5.7)	106.8(17.3)	29.0(14.5)	114.5(10.2)	18.1(8.3)	180.8(23.1)	32.7(18.2)	187.4(14.6)	26.7(23.4)
	Generalized gLasso – L	122.7(6.4)	134.7(52.5)	141.7(8.5)	298.0(40.2)	134.5(12.5)	227.3(73.1)	191.7(10.8)	190.2(47.5)	205.3(5.4)	233.0(20.8)
	Generalized gLasso – S	113.3(5.1)	99.6(10.4)	128.3(11.1)	230.2(62.3)	130.4(13.1)	205.9(77.3)	194.3(9.3)	191.9(47.0)	190.3(8.6)	123.5(32.4)
Scenario 4	Info-guided	84.1(6.7)	52.6(7.7)	85.3(14.0)	113.4(20.6)	105.0(14.0)	105.5(17.7)	140.7(26.1)	47.4(18.2)	151.6(28.8)	50.9(37.7)
	Info-incorporated	89.5(9.6)	15.8(7.1)	77.3(9.7)	28.0(10.9)	94.4(10.2)	19.9(7.4)	169.6(29.6)	31.3(16.9)	178.9(20.7)	32.2(15.3)
	Generalized gLasso – L	99.0(7.8)	134.7(12.6)	124.4(8.5)	324.3(23.4)	133.9(9.0)	303.2(45.0)	186.1(6.5)	259.5(29.1)	174.5(15.6)	179.0(46.1)
	Generalized gLasso – S	96.8(7.8)	127.0(10.0)	121.1(11.7)	298.3(42.1)	110.9(15.8)	177.1(62.5)	168.7(20.5)	212.1(53.3)	175.7(16.1)	189.6(43.4)

deteriorates from Scenario 1 to 4, which is as expected. Performance can vary significantly across network structures. As has been noted in the literature, the considered networks have significantly different properties. It is unclear how such differences affect performance under the proposed approach. We note that there is a lack of such research in the literature. We expect it to be highly challenging and postpone to future research. It is also observed that, under a handful of settings, generalized gLasso identifies a few more TPs, at the price of many more FPs.

For the GGM and information-guided analysis, which have performance closer to that of the proposed approach, we also conduct additional evaluations. Specifically, we consider a sequence of tunings, evaluate identification accuracy at each tuning point, and summarize the overall performance using pAUC (partial Area Under Curve) under the ROC framework. Compared to AUC, pAUC can better describe performance when the FP rate is controlled below a reasonable level. In addition, we also consider the number of TPs when a fixed number of edges are identified. With p =50 and 100, respectively, we consider 300 and 600 identified edges, and refer to the numbers of TPs as Top300 and Top600, respectively. It is noted that multiple tunings may lead to the numbers of identified edges equal to 300 (or 600). For all approaches, we choose the one with the best performance. This measure may slightly favor the proposed approach, which has more tunings. For $p = 50$, the pAUC and Top300 values are summarized in Tables 3.4 and 3.6 (Section 3.6.2), respectively. The corresponding results for $p = 100$ are summarized in Tables 3.5 and 3.7 (Section 3.6.2), respectively. It is observed that the information-incorporated analysis has performance either being or close to the best. Consider for example the setting with $p = 50$, weak signal, and ER network structure. GGM has pAUC*100 equal to 53.3. For the (information-guided, information-incorporated) dual, the pAUC*100 values are (100, 96.8), (96.4, 92.9), (83.8, 79.9), and (58.9, 61.8) under Scenario 1-4, respectively.

To further explore the proposed approach, with $p = 50$, we summarize the selected τ values in Table 3.8 (Section 3.6.2). It is observed that, as the quality of additional information deteriorates, the value of τ decreases, suggesting less information incorporation. Under most settings, there is more information incorporation with weak signals. Significant differences across network structures are again observed.

3.4 Data analysis

TCGA (The Cancer Genome Atlas) is one of the largest and most comprehensive cancer projects jointly organized by the NCI and NHGRI. For over thirty cancer types, it has published comprehensive molecular and other types of data. We analyze TCGA data because of its high quality, easy accessibility, and high scientific impact. In particular, we analyze the gene expression data on LUAD (lung adenocarcinoma) and LUSC (lung squamous cell carcinoma), two subtypes of lung cancer. In TCGA, gene expressions were measured using the Illumina Hiseq2000 RNA Sequencing Version 2 analysis platform and processed and normalized using the RSEM software. More detailed information is available in the literature [158]. We examine data graphically and observe that the processed data mostly have unimodal and bell-shaped distributions, although some deviations from normality are observed. One remedy is to replace the simple correlation with robust correlations that do not rely on the normality assumption [159]. Then the proposed information-incorporated approach can be directly applied. However, the alternative correlations may not be as easily interpretable. In addition, quite a few published studies have used the simple correlation, conducted network analysis, and generated useful findings. As such, we choose to use the simple correlation. In principle, it is possible to conduct whole-genome analysis. However, with limited sample sizes, the findings may be unreliable. In addition, only a small number of genes are "interesting" in the context of lung cancer. As such, we take a "candidate gene" approach. In particular, the 61 gene panel developed and validated in [160] is adopted. This panel has been shown as having important biomedical implications, for example, for lung cancer prognosis. Matching this panel with the gene names in TCGA leads to 50 genes for analysis. Compared to the sample sizes (517 for LUAD and 501 for LUSC), the number of parameters to be estimated is large. The correlation heatmaps are shown in Figures 3.5 and 3.6 (Appendix, Section 3.6.2), where we observe different patterns.

With PubMatrix, for a given gene pair, the number of PubMed publications ranges from 1 to 1,486. More information is provided in Figure 3.2 (which contains the numbers in the log scale) and Figure 3.7 (Appendix, Section 3.6.2, which shows the histogram of the numbers).

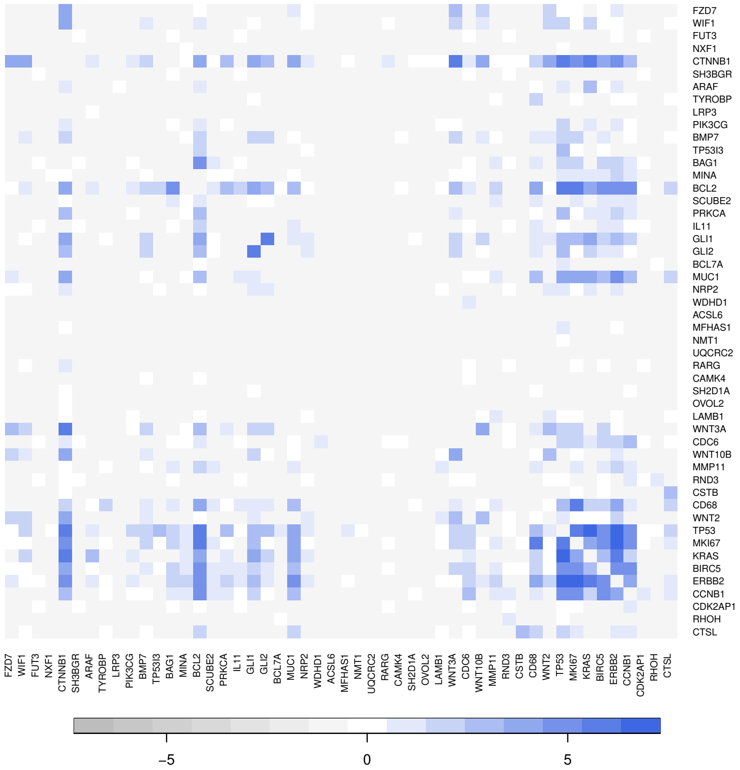
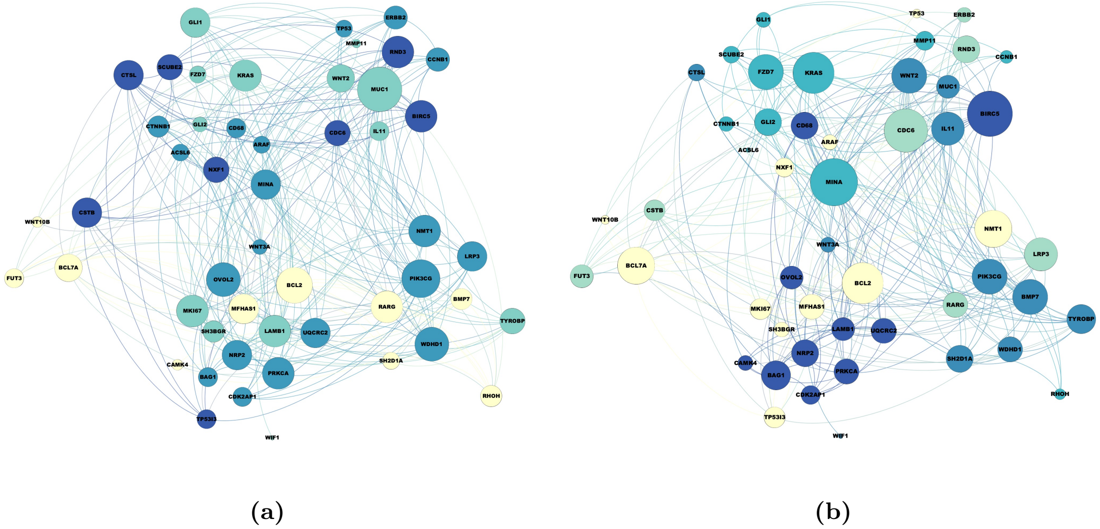


Figure 3.2: Data analysis: number of relevant publications (in log scale) for any gene pair.

The generalized gLasso approach demands a known precision matrix as target, which is not available. As such, it is not applied. For the GGM, information-guided, and information-incorporated approaches, tuning parameters are selected using 4-fold cross validation. The estimated precision matrices are provided at `https://github.com/DeniseYi`. The network structures are graphically presented in Figure 3.3 for the information-incorporated approach and Figure 3.8 (Section 3.6.2) for the alternatives. Briefly, with the LUAD data, they identify 530 (proposed), 526 (GGM), and 554 (information-guided) edges. The proposed approach has 486 and 474 overlapping edges with the two alternatives. With the LUSC data, they identify 534 (proposed), 520 (GGM), and 592 (information-guided) edges. The proposed approach has 486 and 480 overlapping edges with the two alternatives. With the proposed approach, the τ values are 0.57 (LUAD) and 0.64 (LUSC), suggesting that there is considerable information incorporation. The identified network structures are significantly different, with p-values < 0.001 using a permutation test [161]. As noted in the literature, with a large number of edges, it is not feasible to examine biological implications of the findings. In addition, in biological literature, research on genes' *conditional interconnections* remains limited. As such, we do not pursue biological implications of the differences in findings.



(a) **(b)**

Figure 3.3: Gene networks constructed using the proposed approach: (a) LUAD, (b) LUSC.

To gain further insights into the analysis results, we conduct a random splitting-based evaluation. Specifically, data is randomly split into a training and a testing set with sizes 3 : 1. With the training set, the proposed and alternative approaches are applied. Then the predicted likelihood is computed using the testing set. This process is repeated 100 times. In addition, for each edge identified using the whole dataset (without splitting), we compute its probability of being identified in the random splits. Such a probability has been referred to as the OOI (observed occurrence index) in the literature and reflects the stability of finding, with a higher value indicating higher stability. The average predicted negative log-likelihood values are 69.4 (proposed), 73.1 (GGM), and 73.6 (information-guided) for the LUAD data, and 65.7 (proposed), 67.5 (GGM), and 67.5 (information-guided) for the LUSC data. For all the edges identified using the whole dataset, their average OOI values are 0.836 (proposed), 0.806 (GGM), and 0.822 (information-guided) for the LUAD data, and 0.839 (proposed), 0.814 (GGM), and 0.830 (information-guided) for the LUSC data.

3.5 Discussion

In the network analysis of gene expression and other molecular data, the "lack of information" problem is likely to persist in the foreseeable future. We have developed a way of improving gene expression network construction via incorporating additional information contained in published articles. Carefully examining the proposed procedure suggests that it can also accommodate some other sources of information on network edges (for example, as contained in protein-protein interactions) and other types of molecular data. In terms of methodology, it complements the existing literature and differs from the ordinary GGM, generalized gLasso, regression-based works, and Bayesian approaches. The consistency results have provided a strong basis for the proposed approach and may also shed light into other network analysis methods. Simulation has shown that the proposed approach has competitive performance even when the additional information is only partially correct. With two lung cancer datasets, findings different from the alternative approaches have been made.

This study can be extended in multiple ways. It will be of interest to couple the proposed information-incorporated strategy with network constructions that accommodate non-normal data [159]. Estimation and selection can also be realized using other regularization techniques. More refined text mining tools can be adopted to generate more reliable information, which may further improve performance. It will also be of interest to associate an uncertainty measure with the extracted information, which can describe the conflict of published findings, and incorporate such a measure in estimation. It has been suggested that incorrect information does not happen independently. Although this is not difficult to comprehend, modeling and incorporating it in analysis demands significant future research. We also defer bioinformatics examinations of the findings to future works.

3.6 Appendix

3.6.1 Additional details on Theorem 1

Remarks Under the Frobenius norm, we establish the $O(\sqrt{\frac{(p+s)\log p}{n(1+\eta)^2}})$ convergence rate. The term $(p+s)/n$ is the optimal rate for the total squared errors with $p+s$ nonzero elements. The logarithmic factor $\log p$ is the price paid for high dimensionality. The term $1/(1+\eta)^2$ with $\eta \geq 0$ is the gain by incorporating additional information. When $\eta = 0$, our result reduces to that in [113]. It requires $(p+s)\log p/\{n(1+\eta)^2\} = o(1)$, which means $p < n$ if η is bounded. As shown in some publications, if we change the Frobenius norm to other norms, then the convergence rate result may change, with a possibility of accommodating $n < p$. Significant additional developments will be needed to establish consistency and convergence in other norms for the $n < p$ case. We conjecture that rates like $O(\sqrt{\frac{\log p}{n}})$ under the elementwise maximum-norm as in [129] may be possible. We note that the Frobenius norm and accompanying convergence rate results are common in the literature, and postpone investigation on other norms to future research.

Proof Define the operator norm of a matrix B as $\|B\| = \lambda_{\max}^{1/2}(B^T B)$. We first prove that there exists a local minimizer $\hat{\Theta}_{\lambda_n,\eta}$ satisfying:

$$\|\hat{\Theta}_{\lambda_n,\eta} - \Theta_0\|_F^2 = O_p\left(\frac{(p+s)\log p}{n(1+\eta)^2}\right).$$

Then we show that $\hat{\Theta}_{\lambda_n,\eta}$ also enjoys $\hat{\theta}_{ij}^{\lambda_n,\eta} = 0$ for all $(i,j) \in \mathcal{A}^c$. The proof can be achieved via the following steps.

<u>Step 1</u>. Denote $\Delta = \Theta - \Theta_0 = (\delta_{ij})$ and

$$\begin{aligned} Q(\Delta) &= L_{\lambda_n,\eta}(\Theta_0 + \Delta; S, \hat{\Sigma}^p) - L_{\lambda_n,\eta}(\Theta_0; S, \hat{\Sigma}^p) \\ &= -(1+\eta)\left(\log|\Theta_0 + \Delta| - \log|\Theta_0|\right) + \operatorname{tr}\left\{(S + \eta\hat{\Sigma}^p)\Delta\right\} \\ &\quad + \sum_{i \neq j}\left\{p_{\lambda_n}(|\theta_{ij}^0 + \delta_{ij}|) - p_{\lambda_n}(|\theta_{ij}^0|)\right\}. \end{aligned}$$

Applying Taylor's expansion to $f(t) = \log|\Theta + t\Delta|$ with the integral remainder, we have

$$\begin{aligned}&\log|\Theta_0 + \Delta| - \log|\Theta_0| \\ &\quad = \mathrm{tr}(\Sigma_0\Delta) - vec(\Delta)^T \left\{\int_0^1 (1-v)(\Theta_0 + v\Delta)^{-1} \otimes (\Theta_0 + v\Delta)^{-1} dv\right\} vec(\Delta).\end{aligned}$$

Then, we can rewrite $Q(\Delta)$ as

$$\begin{aligned} Q(\Delta) &= \mathrm{tr}\left\{(S - \Sigma_0)\Delta + \eta(\hat{\Sigma}^p - \Sigma_0)\Delta\right\} \\ &\quad +(1+\eta)\cdot vec(\Delta)^T \left\{\int_0^1 (1-v)(\Theta_0 + v\Delta)^{-1} \otimes (\Theta_0 + v\Delta)^{-1} dv\right\} vec(\Delta) \\ &\quad + \sum_{(i,j)\in\mathcal{A}^c} \left\{p_{\lambda_n}(|\theta_{ij}^0 + \delta_{ij}|) - p_{\lambda_n}(|\theta_{ij}^0|)\right\} + \sum_{(i,j)\in\mathcal{A}^-} \left\{p_{\lambda_n}(|\theta_{ij}^0 + \delta_{ij}|) - p_{\lambda_n}(|\theta_{ij}^0|)\right\} \\ &=: I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{3.4}$$

Consider the set $\mathcal{N} = \left\{\Delta : \Delta = \Delta^T, \|\Delta\|_F = Kr_n\right\}$ with K being a large constant and

$$r_n = \sqrt{\frac{(p+s)\log p}{n(1+\eta)^2}}.$$

For $\Delta \in \mathcal{N}$, we have

$$\begin{aligned} |I_1| &= \left|\mathrm{tr}\left\{(S - \Sigma_0)\Delta + \eta(\hat{\Sigma}^p - \Sigma_0)\Delta\right\}\right| \\ &\leq \left|\sum_{(i,j)\in\mathcal{A}} \left\{(s_{ij} - \sigma_{ij}^0)\delta_{ij} + \eta(\hat{\sigma}_{ij}^p - \sigma_{ij}^0)\delta_{ij}\right\}\right| + \left|\sum_{(i,j)\in\mathcal{A}^c} \left\{(s_{ij} - \sigma_{ij}^0)\delta_{ij} + \eta(\hat{\sigma}_{ij}^p - \sigma_{ij}^0)\delta_{ij}\right\}\right| \\ &=: I_{11} + I_{12}. \end{aligned} \tag{3.5}$$

Note that

$$I_{11} \leq \sqrt{p+s}\left(\max_{(i,j)\in\mathcal{A}} |s_{ij} - \sigma_{ij}^0| + \eta \max_{(i,j)\in\mathcal{A}} |\hat{\sigma}_{ij}^p - \sigma_{ij}^0|\right)\|\Delta\|_F.$$

By Lemma A.3 of [162], with probability tending to 1,

$$\max_{i,j} |s_{ij} - \sigma_{ij}^0| \leq C_1\sqrt{\frac{\log p}{n}},$$

where C_1 is a large constant. Together with Condition (C5), we have

$$I_{11} \leq \sqrt{p+s}(C_0 + C_1)\sqrt{\frac{\log p}{n}} K r_n = (1+\eta)(C_0 + C_1) K r_n^2, \tag{3.6}$$

with probability tending to 1. Furthermore,

$$I_{12} \leq \sum_{(i,j)\in\mathcal{A}^c} \left| \left\{ (s_{ij} - \sigma_{ij}^0) + \eta(\hat{\sigma}_{ij}^p - \sigma_{ij}^0) \right\} \delta_{ij} \right| \leq (C_0 + C_1)\sqrt{\frac{\log p}{n}} \sum_{(i,j)\in\mathcal{A}^c} |\delta_{ij}|,$$

with probability tending to 1. For $(i,j) \in \mathcal{A}^c$,, $\theta_{ij}^0 = 0$. Thus, by Condition (C3), we can find a constant $k^* > 0$ such that

$$I_3 = \sum_{(i,j)\in\mathcal{A}^c} p_{\lambda_n}(|\delta_{ij}|) \geq \lambda_n k^* \sum_{(i,j)\in\mathcal{A}^c} |\delta_{ij}|.$$

With the above arguments, we have, with probability tending to 1,

$$I_3 - I_{12} \geq \left\{ \lambda_n k^* - (C_0 + C_1)\sqrt{\frac{\log p}{n}} \right\} \sum_{(i,j)\in\mathcal{A}^c} |\delta_{ij}|.$$

With the assumption that $\frac{(p+s)\log p}{n} = O(\lambda_n^2)$, we can see from the above that

$$I_3 - I_{12} \geq 0, \tag{3.7}$$

with probability tending to 1. Using Condition (C1) and result (18) in [162], we can get that

$$\begin{aligned} I_2 &\geq (1+\eta)\|vec(\Delta)\|^2 \int_0^1 (1-v)\lambda_{\min}^2(\Theta_0 + v\Delta)^{-1} dv \\ &\geq (1+\eta)\frac{1}{2}\|vec(\Delta)\|^2 \min_{0\leq v\leq 1} \lambda_{\min}^2(\Theta_0 + v\Delta)^{-1} \\ &\geq (1+\eta)\frac{1}{2}\|vec(\Delta)\|^2 \left(\|\Theta_0\| + \|\Delta\|\right)^{-2} \\ &\geq (1+\eta)\frac{K^2 r_n^2}{4\phi_2^2}. \end{aligned} \tag{3.8}$$

For I_4, using Taylor's expansion, we obtain

$$|I_4| \leq \sum_{(i,j)\in\mathcal{A}^-} \left\{ p'_{\lambda_n}(|\theta^0_{ij}|)|\delta_{ij}| + p''_{\lambda_n}(|\tilde{\theta}_{ij}|)\frac{\delta^2_{ij}}{2} \right\},$$

where $\tilde{\theta}_{ij}$ is on the line segment jointing zero and θ^0_{ij}. By Condition (C4) and the Cauchy-Schwartz inequality, we can conclude

$$|I_4| \leq \sqrt{s}a_n\|\Delta\|_F + b_n\|\Delta\|^2_F \leq C_2 K r^2_n + o(K^2 r^2_n), \tag{3.9}$$

where C_2 is a positive constant, and the second inequality follows from Condition (C2).

Combining (3.4)-(3.9), we have that $Q(\Delta) > 0$ with probability tending to 1, when K is sufficiently large. Following arguments similar to [163], we can prove that

$$\|\hat{\Delta}\|^2_F = \|\hat{\Theta}_{\lambda_n,\eta} - \Theta_0\|^2_F = O_p\left(r^2_n\right).$$

Step 2. For $(i,j) \in \mathcal{A}^c$, the derivative of $L_{\lambda_n,\eta}(\Theta; S, \hat{\Sigma}^p)$ with respect to θ_{ij} is

$$\frac{\partial L_{\lambda_n,\eta}(\Theta; S, \hat{\Sigma}^p)}{\partial\theta_{ij}} = 2\left\{(s_{ij} - \sigma_{ij}) + \eta(\hat{\sigma}^p_{ij} - \sigma_{ij}) + p'_{\lambda_n}(|\theta_{ij}|)\mathrm{sign}(\theta_{ij})\right\}. \tag{3.10}$$

For $\hat{\Theta}_{\lambda_n,\eta}$, a minimizer of $L_{\lambda_n,\eta}(\Theta; S, \hat{\Sigma}^p)$, it suffices to show that for all $\hat{\theta}^{\lambda_n,\eta}_{ij}$ with $(i,j) \in \mathcal{A}^c$, the sign of $\frac{\partial L_{\lambda_n,\tau}(\Theta;S,\hat{\Sigma}^p)}{\partial\theta_{ij}}\Big|_{\theta_{ij}=\hat{\theta}^{\lambda_n,\eta}_{ij}}$ depends only on $\mathrm{sign}(\hat{\theta}^{\lambda_n,\eta}_{ij})$ with probability tending to 1, and the optimum is at zero, so that $\hat{\theta}^{\lambda_n,\eta}_{ij} = 0$ for all $(i,j) \in \mathcal{A}^c$ with probability tending to 1.

Firstly, we have that $\|\hat{\Theta}_{\lambda_n,\eta} - \Theta_0\|_F = O_p(r_n)$ with $r_n \to 0$. Following the arguments in the proof of Theorem 2 in [113], we have, with probability tending to 1,

$$\max_{i,j}|s_{ij} - \sigma_{ij}| \leq \max_{i,j}|s_{ij} - \sigma^0_{ij}| + \max_{i,j}|\sigma^0_{ij} - \sigma_{ij}| \leq C_1\sqrt{\frac{\log p}{n}} + C_3 K r_n,$$

and

$$\max_{i,j}\eta|\hat{\sigma}^p_{ij} - \sigma_{ij}| \leq \max_{i,j}\eta|\hat{\sigma}^p_{ij} - \sigma^0_{ij}| + \max_{i,j}\eta|\sigma^0_{ij} - \sigma_{ij}| \leq C_0\sqrt{\frac{\log p}{n}} + \eta C_3 K r_n,$$

where C_3 is a positive constant. Therefore, we have

$$\max_{i,j} |(s_{ij} - \sigma_{ij}) + \eta(\hat{\sigma}_{ij}^{p} - \sigma_{ij})| \leq (C_0 + C_1)\sqrt{\frac{\log p}{n}} + (1+\eta)C_3 K r_n \leq C_4 \sqrt{\frac{(p+s)\log p}{n}},$$

with probability tending to 1, where C_4 is a positive constant.

Secondly, for any $\hat{\theta}_{ij}^{\lambda_n,\eta}$ in a small neighborhood of 0 and some positive constant C_5, we have

$$p'_{\lambda_n}(|\hat{\theta}_{ij}^{\lambda_n,\eta}|) \geq C_5 \lambda_n,$$

by Conditions (C3) and (C4). Then by setting $\lambda_n > \frac{C_4}{C_5}\sqrt{\frac{(p+s)\log p}{n}}$, we have that $p'_{\lambda_n}(|\theta_{ij}|)\text{sign}(\theta_{ij})$ dominates the other part, which yields that the sign of $\frac{\partial L_{\lambda_n,\tau}(\Theta; S, \hat{\Sigma}^p)}{\partial \theta_{ij}}\Big|_{\theta_{ij}=\hat{\theta}_{ij}^{\lambda_n,\eta}}$ equals $\text{sign}(\hat{\theta}_{ij}^{\lambda_n,\eta})$ with probability tending to 1. The theorem is proved. □

3.6.2 Additional numerical results

Table 3.2: Simulation settings: numbers of TPs and FPs in the additional information. (ER: Erdos-Renyi; SF: scale-free; NN: nearest-neighbor; and banded structure.)

		ER	SF	NN	Banded
$p = 50$					
Scenario 1	TP	152	200	184	244
	FP	0	0	0	0
Scenario 2	TP	140	140	140	140
	FP	12	60	44	104
Scenario 3	TP	100	100	100	100
	FP	52	100	84	144
Scenario 4	TP	40	40	40	40
	FP	112	160	144	204
$p = 100$					
Scenario 1	TP	556	600	630	516
	FP	0	0	0	0
Scenario 2	TP	490	490	490	490
	FP	66	110	140	26
Scenario 3	TP	350	350	350	350
	FP	206	250	280	166
Scenario 4	TP	140	140	140	140
	FP	416	460	490	376

Table 3.3: Simulation results with selected optimal tunings: mean (sd) TPs and FPs for $p = 100$. (ER: Erdos-Renyi; SF: scale-free; NN: nearest-neighbor; Banded(+): positive banded; and Banded(-): negative banded. L/S with generalized gLasso: large/small perturbations added.)

		ER		SF		NN		Banded(+)		Banded(-)	
	Total edges	556		600		630		516		516	
		TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
Weak signal											
	GGM	146.7(11.0)	44.7(10.1)	114.8(18.4)	54.2(17.7)	230.7(39.4)	46.1(95.1)	285.2(7.5)	19.3(6.2)	303.1(8.0)	17.6(5.3)
Scenario 1	Info-guided	510.3(8.0)	0.0(0.0)	494.9(16.7)	0.0(0.0)	592.8(7.4)	0.0(0.0)	493.7(5.7)	0.0(0.0)	493.7(6.0)	0.0(0.0)
	Info-incorporated	526.3(29.8)	0.5(1.3)	278.1(46.8)	1.8(2.4)	577.2(69.5)	3.7(5.7)	504.2(4.6)	0.2(0.6)	506.3(3.9)	0.3(0.7)
	Generalized gLasso – L	307.1(13.5)	17.9(6.3)	380.3(12.1)	44.4(9.4)	351.1(13.8)	5.9(2.9)	394.9(8.6)	68.1(9.7)	401.5(10.7)	63.2(10.7)
	Generalized gLasso – S	282.4(11.7)	17.3(6.7)	272.3(11.4)	34.7(7.5)	314.1(15.0)	6.2(3.3)	365.3(8.1)	64.3(10.0)	373.0(9.3)	61.5(11.5)
Scenario 2	Info-guided	454.9(7.8)	33.5(8.8)	409.1(16.0)	44.9(10.1)	500.3(14.8)	87.8(15.0)	474.5(6.0)	4.9(2.2)	475.5(5.9)	4.3(2.0)
	Info-incorporated	472.0(39.8)	24.9(17.5)	234.8(9.5)	11.0(4.7)	516.9(23.1)	64.6(19.5)	487.3(5.0)	0.9(1.4)	490.5(4.4)	0.7(1.2)
	Generalized gLasso – L	292.1(10.7)	53.3(6.7)	313.4(10.8)	104.1(9.3)	289.5(15.3)	76.6(4.2)	382.7(8.6)	77.0(11.1)	393.6(10.1)	72.7(11.9)
	Generalized gLasso – S	266.5(13.1)	32.3(7.1)	263.2(11.4)	62.1(7.4)	300.5(14.4)	36.1(3.7)	362.9(9.2)	70.0(11.1)	367.1(10.0)	66.9(12.4)
Scenario 3	Info-guided	350.5(19.7)	113.9(18.3)	320.2(27.9)	116.9(30.9)	428.3(10.2)	174.3(15.9)	416.0(28.8)	40.3(26.8)	413.9(22.6)	30.2(20.0)
	Info-incorporated	344.0(47.0)	59.2(31.2)	185.1(10.1)	21.9(7.6)	425.4(62.6)	95.7(58.2)	441.3(7.5)	4.4(3.0)	443.4(5.4)	4.5(2.6)
	Generalized gLasso – L	225.9(11.3)	126.8(6.5)	266.5(10.0)	197.1(10.7)	263.7(13.7)	177.3(5.6)	377.0(8.9)	127.8(14.6)	379.3(8.6)	125.7(16.0)
	Generalized gLasso – S	218.7(13.1)	107.5(5.7)	212.9(9.4)	117.4(8.0)	263.1(13.5)	88.9(3.3)	347.1(8.4)	102.9(11.3)	363.7(8.6)	90.3(13.0)
Scenario 4	Info-guided	234.5(9.8)	234.1(19.3)	209.1(22.4)	244.6(25.1)	299.3(14.0)	274.9(23.3)	417.8(5.7)	123.1(12.7)	422.8(6.8)	119.4(16.1)
	Info-incorporated	191.9(29.3)	53.2(25.8)	122.6(20.9)	30.5(16.0)	314.2(41.0)	100.1(46.6)	349.7(6.6)	6.7(3.7)	362.1(7.7)	6.2(4.0)
	Generalized gLasso – L	151.9(9.5)	266.8(10.6)	171.4(9.3)	281.5(12.2)	204.5(14.7)	283.3(4.3)	342.1(9.2)	192.1(16.1)	352.1(6.5)	198.3(21.9)
	Generalized gLasso – S	141.0(9.3)	156.4(8.9)	137.7(8.2)	177.2(9.1)	203.1(15.0)	183.3(4.7)	330.9(9.3)	131.9(15.3)	345.0(5.8)	129.2(18.6)
Strong signal											
	GGM	254.6(12.4)	51.7(9.9)	188.0(10.0)	68.4(11.1)	226.8(15.4)	29.9(7.2)	301.4(18.9)	5.1(4.0)	327.1(21.4)	5.7(3.5)
Scenario 1	Info-guided	531.6(5.2)	0.0(0.0)	538.8(9.1)	0.0(0.0)	592.3(8.3)	0.0(0.0)	513.4(2.0)	0.0(0.0)	513.8(2.1)	0.0(0.0)
	Info-incorporated	544.3(4.6)	17.0(11.7)	410.9(94.7)	3.0(3.2)	542.0(79.8)	3.7(6.7)	514.3(1.8)	2.9(2.7)	514.7(2.1)	1.9(1.9)
	Generalized gLasso – L	385.8(11.2)	43.7(10.1)	429.7(12.2)	58.3(13.2)	340.4(12.1)	6.7(4.2)	508.1(2.9)	21.5(7.0)	509.4(4.2)	12.8(4.1)
	Generalized gLasso – S	306.8(14.4)	26.9(8.6)	329.1(10.6)	46.7(11.4)	330.5(14.2)	6.9(3.8)	509.5(3.3)	19.8(5.4)	511.7(3.4)	9.9(5.3)
Scenario 2	Info-guided	478.3(5.2)	29.7(5.6)	447.2(7.5)	42.5(7.6)	499.9(11.8)	88.9(14.0)	495.1(11.2)	2.1(2.3)	498.5(5.7)	1.2(1.3)
	Info-incorporated	502.9(23.5)	41.9(13.2)	340.2(62.3)	17.1(13.4)	515.0(27.9)	58.8(20.4)	485.9(10.3)	2.1(1.9)	490.6(7.2)	3.8(3.3)
	Generalized gLasso – L	367.0(13.4)	90.3(9.3)	351.3(7.1)	131.3(12.0)	298.9(13.2)	80.0(3.9)	499.0(5.3)	24.1(5.9)	497.3(4.9)	17.7(4.9)
	Generalized gLasso – S	306.5(11.7)	58.3(8.8)	292.3(9.0)	75.8(11.0)	284.7(14.8)	45.6(4.3)	498.3(5.6)	14.5(5.8)	499.5(4.5)	19.9(4.9)
Scenario 3	Info-guided	417.8(9.1)	132.5(11.3)	401.7(15.4)	150.2(22.4)	425.6(11.5)	172.3(12.9)	463.5(21.2)	17.3(8.1)	465.5(24.0)	14.7(6.3)
	Info-incorporated	364.7(31.0)	33.3(31.0)	269.7(28.3)	29.5(20.9)	428.1(59.5)	100.2(48.8)	434.9(18.4)	4.5(3.5)	445.6(11.5)	5.1(3.7)
	Generalized gLasso – L	314.9(11.9)	158.2(10.2)	308.5(9.3)	212.9(15.0)	282.5(14.0)	158.0(4.7)	453.7(5.4)	50.7(12.1)	460.2(5.4)	42.1(8.2)
	Generalized gLasso – S	270.8(14.5)	97.9(8.7)	252.7(9.0)	128.9(11.4)	260.5(14.1)	79.1(4.4)	452.7(5.4)	49.1(10.0)	456.2(6.9)	43.5(8.5)
Scenario 4	Info-guided	314.9(11.3)	230.3(13.6)	272.7(8.5)	253.8(15.3)	301.5(11.1)	272.0(16.6)	423.7(31.8)	35.3(16.9)	445.6(5.3)	35.9(7.5)
	Info-incorporated	294.9(27.7)	67.7(27.6)	204.8(20.2)	49.6(21.2)	322.8(45.4)	103.9(46.9)	364.8(14.5)	6.0(3.2)	377.7(18.3)	5.9(3.3)
	Generalized gLasso – L	265.2(10.8)	305.8(12.5)	228.2(9.9)	331.7(16.0)	199.7(11.2)	288.9(6.9)	404.4(6.1)	77.6(12.2)	417.1(4.7)	78.7(9.7)
	Generalized gLasso – S	223.2(12.4)	183.9(11.7)	194.1(10.1)	215.9(12.5)	201.1(12.6)	152.5(5.0)	406.9(5.6)	85.5(10.6)	412.3(5.0)	76.3(9.0)

Table 3.4: Simulation results of pAUC: mean×100 (sd×100) for $p = 50$. (ER: Erdos-Renyi; SF: scale-free; NN: nearest-neighbor; Banded(+): positive banded; and Banded(-): negative banded.)

		ER	SF	NN	Banded(+)	Banded(-)
Weak signal						
	GGM	53.3(2.4)	40.6(2.3)	64.6(2.9)	66.0(2.4)	69.2(2.7)
Scenario 1	Info-guided	100(0.0)	100(0.0)	100(0.0)	100(0.0)	100(0.0)
	Info-incorporated	96.8(1.7)	90.2(1.9)	98.0(0.7)	97.5(3.2)	97.9(0.3)
Scenario 2	Info-guided	96.4(0.6)	79.7(1.1)	89.6(1.6)	83.9(1.0)	84.1(0.8)
	Info-incorporated	92.9(1.4)	70.8(2.2)	87.1(3.2)	85.9(2.4)	86.8(2.3)
Scenario 3	Info-guided	83.8(1.5)	66.9(2.0)	78.8(2.3)	75.3(2.1)	77.7(1.8)
	Info-incorporated	79.9(1.6)	58.2(3.4)	77.0(3.3)	78.9(2.3)	82.2(2.2)
Scenario 4	Info-guided	58.9(1.7)	44.4(2.2)	62.7(2.1)	60.4(1.9)	62.1(2.2)
	Info-incorporated	61.8(2.7)	45.0(2.6)	68.1(3.3)	71.6(2.2)	73.6(3.0)
Strong signal						
	GGM	68.1(3.0)	52.2(2.5)	75.8(2.0)	72.0(2.2)	74.5(2.3)
Scenario 1	Info-guided	100(0.0)	100(0.0)	100(0.0)	100(0.0)	100(0.0)
	Info-incorporated	98.1(1.2)	92.8(1.2)	99.2(0.8)	96.9(1.7)	97.9(1.1)
Scenario 2	Info-guided	97.4(0.7)	82.7(1.1)	91.9(1.3)	85.6(1.7)	85.9(1.0
	Info-incorporated	96.0(1.1)	76.0(1.4)	91.0(2.0)	80.0(2.7)	83.2(2.7)
Scenario 3	Info-guided	88.4(1.1)	72.4(1.3)	83.9(1.8)	77.3(1.8)	79.7(1.6)
	Info-incorporated	88.2(2.0)	67.1(2.8)	84.6(1.9)	78.2(3.1)	79.4(4.2)
Scenario 4	Info-guided	67.3(2.1)	52.8(2.2)	70.9(2.8)	66.5(3.2)	68.2(2.3)
	Info-incorporated	73.5(2.7)	56.1(2.6)	78.7(2.1)	77.1(2.5)	77.7(2.1)

Table 3.5: Simulation results of pAUC: mean×100 (sd×100) for $p = 100$. (ER: Erdos-Renyi; SF: scale-free; NN: nearest-neighbor; Banded(+): positive banded; and Banded(-): negative banded.)

		ER	SF	NN	Banded(+)	Banded(-)
Weak signal						
	GGM	59.5(1.1)	43.4(1.5)	72.0(1.2)	75.7(1.5)	78.1(1.9)
Scenario 1	Info-guided	100(0.0)	100(0.0)	100(0.0)	100(0.0)	100(0.0)
	Info-incorporated	93.1(2.4)	83.9(1.7)	96.3(0.9)	95.4(2.6)	96.0(2.9)
Scenario 2	Info-guided	93.5(0.7)	87.1(0.4)	88.1(0.8)	98.9(0.4)	98.9(2.4)
	Info-incorporated	87.1(0.9	74.1(1.9)	89.8(0.9)	95.7(3.8)	96.2(0.8)
Scenario 3	Info-guided	78.0(0.7)	73.6(0.6)	78.8(1.5)	94.7(0.6)	94.5(0.4)
	Info-incorporated	77.5(1.0)	65.1(1.7)	85.4(1.0)	94.1(0.9)	94.6(1.2)
Scenario 4	Info-guided	54.0(0.7)	47.9(0.6)	59.7(4.4)	78.7(1.4)	80.0(1.1)
	Info-incorporated	65.0(1.2)	47.2(1.9)	77.4(0.8)	89.3(2.1)	89.8(1.7)
Strong signal						
	GGM	71.8(2.1)	53.9(1.9)	72.4(1.2)	87.0(1.1)	88.6(0.8)
Scenario 1	Info-guided	100(0.0)	100(0.0)	100(0.0)	100(0.0)	100(0.0)
	Info-incorporated	95.7(3.3)	87.6(1.0)	96.2(0.5)	99.1(0.3)	99.0(0.3)
Scenario 2	Info-guided	95.0(0.4)	88.3(0.5)	88.0(0.4)	99.6(6.3)	99.7(7.3)
	Info-incorporated	90.6(0.7)	78.5(0.8)	89.8(0.8)	95.8(3.4)	95.5(4.0)
Scenario 3	Info-guided	82.6(0.9)	77.5(0.9)	79.3(1.7)	95.4(0.5)	95.3(0.4)
	Info-incorporated	82.4(1.3)	71.2(1.4)	85.5(0.9)	90.2(2.3)	91.2(1.9)
Scenario 4	Info-guided	64.0(1.6)	53.5(0.8)	60.1(1.8)	84.9(1.6)	85.7(0.9)
	Info-incorporated	75.9(1.8)	57.0(1.8)	77.5(1.2)	88.8(1.3)	89.9(1.0)

Table 3.6: Simulation results of Top300: mean (sd) TPs for $p = 50$. (ER: Erdos-Renyi; SF: scale-free; NN: nearest-neighbor; Banded(+): positive banded; and Banded(-): negative banded.)

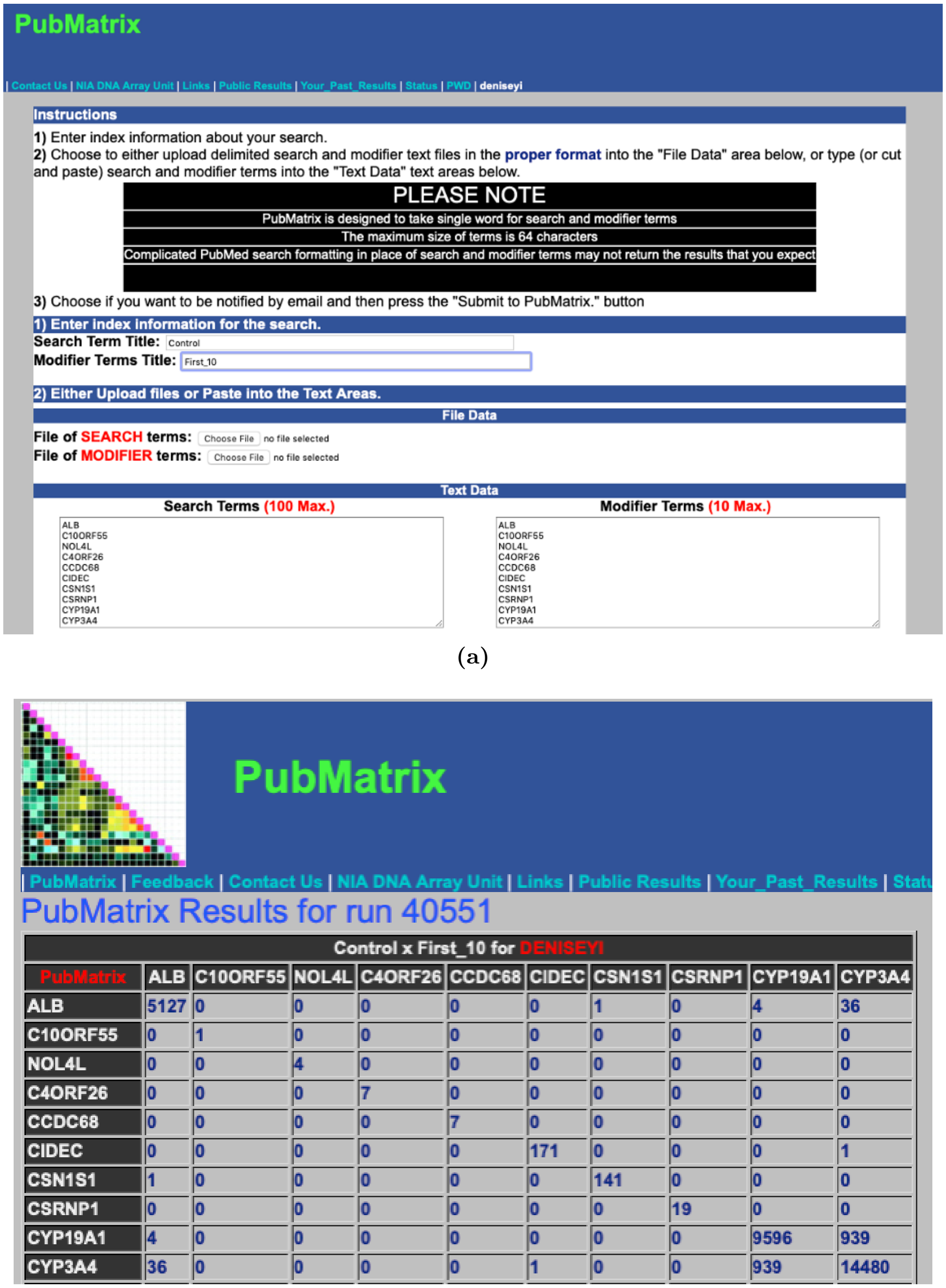
		ER	SF	NN	Banded(+)	Banded(-)
	Total edges	152	200	184	244	244
Weak signal						
	GGM	58.0(8.9)	60.0(8.9)	90.0(11.9)	121.0(10.4)	125.0(7.4)
Scenario 1	Info-guided	152.0(0.0)	200.0(0.0)	184.0(0.0)	244.0(0.0)	244.0(0.0)
	Info-incorporated	149.6(2.2)	142.1(7.3)	179.5(5.1)	241.5(2.2)	241.7(1.9)
Scenario 2	Info-guided	144.0(2.0)	146.9(2.7)	156.6(4.2)	178.6(5.8)	175.5(4.2)
	Info-incorporated	141.3(2.8)	118.6(8.9)	155.5(5.7)	188.7(8.7)	186.6(7.0)
Scenario 3	Info-guided	118.5(4.3)	113.5(3.3)	132.5(6.9)	152.4(4.1)	155.9(4.7)
	Info-incorporated	115.5(5.1)	106.4(3.7)	128.3(5.5)	159.3(6.7)	163.3(6.5)
Scenario 4	Info-guided	71.4(5.6)	66.3(5.0)	89.1(12.2	103.8(5.5)	103.5(4.9)
	Info-incorporated	83.5(8.5)	75.4(7.4)	106.3(7.3)	140.3(7.3)	146.3(7.1)
Strong signal						
	GGM	58.0(8.9)	60.0(8.9)	90.0(11.9)	121.0(10.4)	125.0(7.4)
Scenario 1	Info-guided	152.0(0.0)	200.0(0.0)	184.0(0.0)	244.0(0.0)	244.0(0.0)
	Info-incorporated	145.6(4.1)	161.3(7.2)	183.1(1.5)	238.0(15.8)	240.0(14.5)
Scenario 2	Info-guided	146.2(2.2)	154.5(3.8)	158.0(5.5)	177.7(5.1)	178.3(3.8)
	Info-incorporated	141.9(3.3)	134.5(6.1)	164.4(4.1)	161.7(16.4)	170.4(14.4)
Scenario 3	Info-guided	125.7(3.7)	127.5(5.8)	132.5(6.8)	153.3(4.6)	158.0(2.7)
	Info-incorporated	126.0(4.7)	119.9(5.0)	144.9(6.2)	151.6(12.1)	155.5(11.1)
Scenario 4	Info-guided	84.3(5.7)	85.6(9.1)	100.7(6.2)	112.5(3.5)	115.3(4.6)
	Info-incorporated	100.6(5.5)	98.1(5.2)	130.1(5.6)	144.8(10.6)	144.2(10.0)

Table 3.7: Simulation results of Top600: mean (sd) TPs for $p = 100$. (ER: Erdos-Renyi; SF: scale-free; NN: nearest-neighbor; Banded(+): positive banded; and Banded(-): negative banded.)

		ER	SF	NN	Banded(+)	Banded(-)
	Total edges	556	600	630	516	516
Weak signal						
	GGM	177.8(12.4)	145.3(9.7)	241.7(10.5)	289.0(12.1)	307.9(9.3)
Scenario 1	Info-guided	556.0(0.0)	600.0(0.0)	600.0(0.0)	516.0(0.0)	516.0(0.0)
	Info-incorporated	354.5(12.0)	366.9(47.7)	443.6(49.0)	516.0(13.8)	516.0(19.7)
Scenario 2	Info-guided	495.6(1.4)	491.3(1.8)	493.6(5.3)	516.0(1.5)	516.0(2.1)
	Info-incorporated	334.8(12.4)	421.3(30.6)	411.9(42.6)	491.9(47.2)	483.4(51.9)
Scenario 3	Info-guided	360.9(2.1)	368.5(4.6)	338.1(3.7)	470.0(6.3)	470.5(6.1)
	Info-incorporated	404.7(35.2)	312.1(6.6)	381.1(9.8)	456.9(11.7)	350.8(24.9)
Scenario 4	Info-guided	169.8(4.4)	171.0(5.6)	160.5(6.8)	288.7(3.7)	289.9(7.1)
	Info-incorporated	219.3(12.1)	173.0(10.6)	175.8(26.0)	370.9(23.8)	376.8(22.6)
Strong signal						
	GGM	288.0(13.2)	215.7(10.8)	245.2(14.5)	345.3(25.4)	353.8(21.8)
Scenario 1	Info-guided	556.0(0.0)	600.0(0.0)	600.0(0.0)	516.0(0.0)	516.0(0.0)
	Info-incorporated	442.5(10.7)	352.3(11.5)	438.9(45.5)	516.0(5.2)	516.0(6.2)
Scenario 2	Info-guided	501.3(2.6)	498.7(2.8)	492.9(5.8)	516.0(1.3)	516.0(1.8)
	Info-incorporated	421.9(11.0)	323.5(11.7)	426.2(47.8)	516.0(14.0)	516.0(20.1)
Scenario 3	Info-guided	381.1(5.1)	393.7(5.9)	342.8(21.4)	482.2(5.8)	480.3(6.3)
	Info-incorporated	370.9(9.6)	294.1(16.3)	382.4(12.1)	462.7(11.6)	400.7(6.1)
Scenario 4	Info-guided	209.0(6.4)	212.5(6.4)	160.6(6.4)	284.2(3.7)	291.5(3.7)
	Info-incorporated	313.6(10.3)	242.1(10.4)	169.5(8.6)	380.1(18.7)	383.6(17.1)

Table 3.8: Simulation: average τ value for $p = 50$. (ER: Erdos-Renyi; SF: scale-free; NN: nearest-neighbor; Banded(+): positive banded; and Banded(-): negative banded.)

	ER	SF	NN	Banded(+)	Banded(-)
Weak signal					
Scenario 1	0.98	0.97	0.97	0.95	0.96
Scenario 2	0.92	0.94	0.91	0.80	0.78
Scenario 3	0.80	0.76	0.56	0.76	0.70
Scenario 4	0.51	0.50	0.46	0.72	0.68
Strong signal					
Scenario 1	0.94	0.92	0.89	0.65	0.71
Scenario 2	0.88	0.81	0.74	0.50	0.54
Scenario 3	0.62	0.54	0.56	0.47	0.45
Scenario 4	0.50	0.53	0.50	0.38	0.41


PubMatrix
| Contact Us | NIA DNA Array Unit | Links | Public Results | Your_Past_Results | Status | PWD | deniseyi
Choose File
Choose File
| PubMatrix | Feedback | Contact Us | NIA DNA Array Unit | Links | Public Results | Your_Past_Results | Statu

Instructions

1) Enter index information about your search.
2) Choose to either upload delimited search and modifier text files in the **proper format** into the "File Data" area below, or type (or cut and paste) search and modifier terms into the "Text Data" text areas below.

PLEASE NOTE

PubMatrix is designed to take single word for search and modifier terms

The maximum size of terms is 64 characters

Complicated PubMed search formatting in place of search and modifier terms may not return the results that you expect

3) Choose if you want to be notified by email and then press the "Submit to PubMatrix." button

1) Enter index information for the search.

Search Term Title: Control
Modifier Terms Title: First_10

2) Either Upload files or Paste into the Text Areas.

File Data

File of SEARCH terms: no file selected
File of MODIFIER terms: no file selected

Text Data

Search Terms (100 Max.)

ALB
C10ORF55
NOL4L
C4ORF26
CCDC68
CIDEC
CSN1S1
CSRNP1
CYP19A1
CYP3A4

Modifier Terms (10 Max.)

ALB
C10ORF55
NOL4L
C4ORF26
CCDC68
CIDEC
CSN1S1
CSRNP1
CYP19A1
CYP3A4

(a)

PubMatrix Results for run 40551

Control x First_10 for DENISEYI

PubMatrix	ALB	C10ORF55	NOL4L	C4ORF26	CCDC68	CIDEC	CSN1S1	CSRNP1	CYP19A1	CYP3A4
ALB	5127	0	0	0	0	0	1	0	4	36
C10ORF55	0	1	0	0	0	0	0	0	0	0
NOL4L	0	0	4	0	0	0	0	0	0	0
C4ORF26	0	0	0	7	0	0	0	0	0	0
CCDC68	0	0	0	0	7	0	0	0	0	0
CIDEC	0	0	0	0	0	171	0	0	0	1
CSN1S1	1	0	0	0	0	0	141	0	0	0
CSRNP1	0	0	0	0	0	0	0	19	0	0
CYP19A1	4	0	0	0	0	0	0	0	9596	939
CYP3A4	36	0	0	0	0	1	0	0	939	14480

(b)

Figure 3.4: Sample PubMatrix (a) submit and (b) result pages.

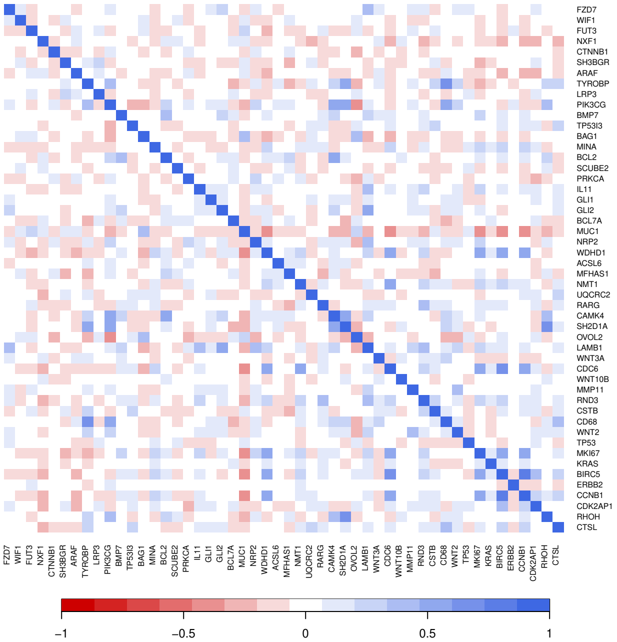


Figure 3.5: Analysis of LUAD data: heatmap of correlation.

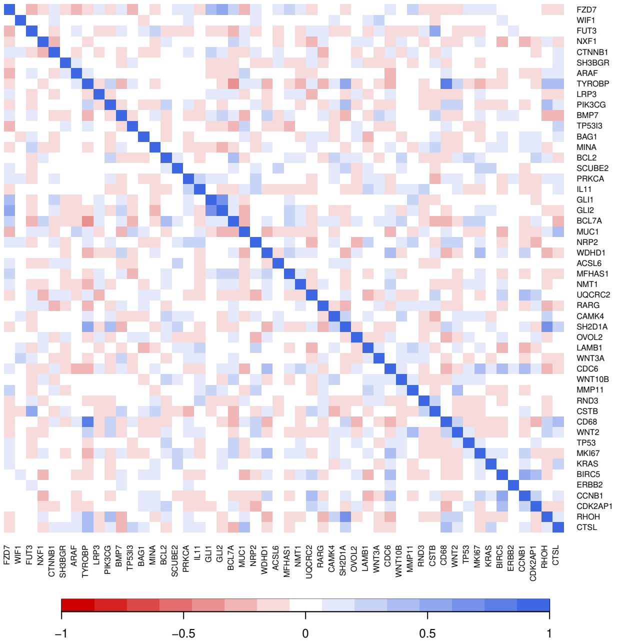


Figure 3.6: Analysis of LUSC data: heatmap of correlation.

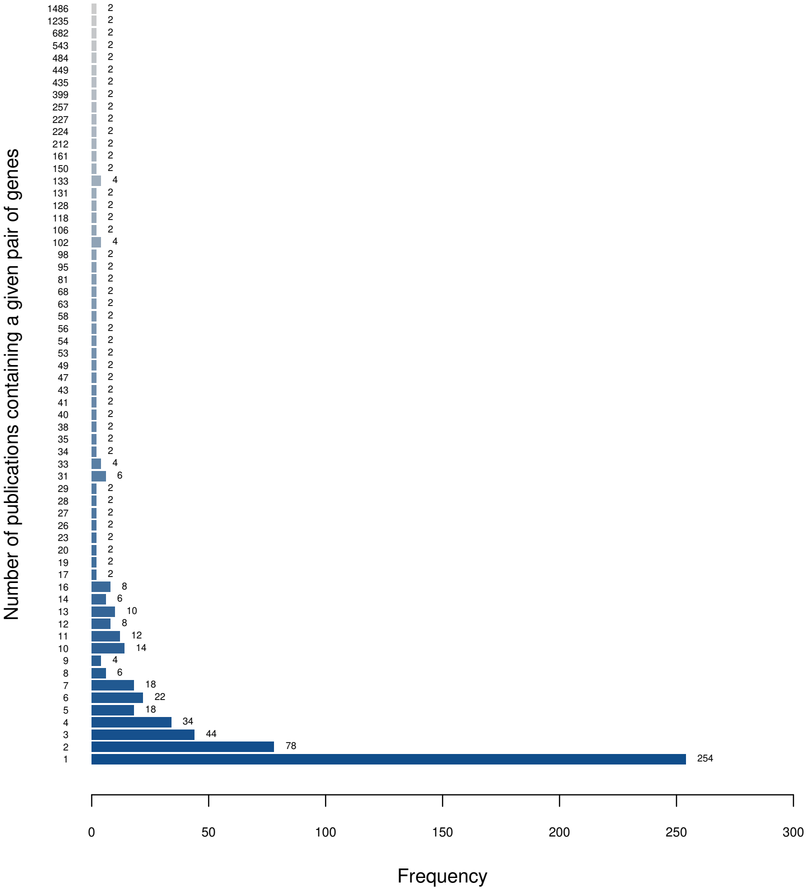


Figure 3.7: Data analysis: distribution of the number of publications including a given pair of genes.

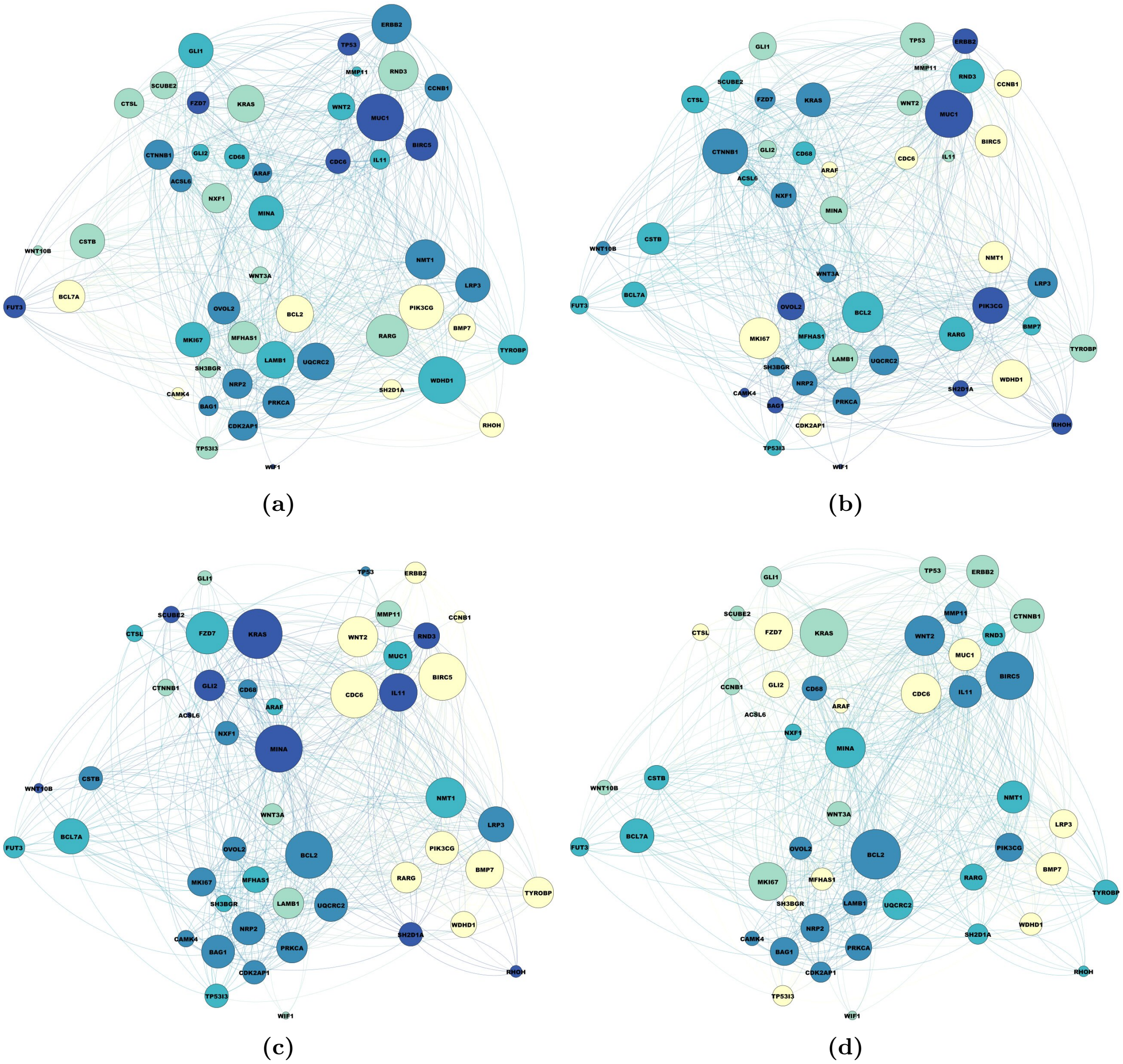


Figure 3.8: Gene networks constructed using the alternatives. (a) LUAD with GGM. (b) LUAD with information-guided. (c) LUSC with GGM. (d) LUSC with information-guided.

Chapter 4

Project 3: Assisted differential network analysis for gene expression data

Abstract

When there are two or more conditions/groups (for example, cancer and normal, deceased and alive, and different stages/subtypes), differential analysis targets at identifying key differences and has important implications. In network differential analysis, spectral clustering and other techniques can identify key contributors and reveal important biological mechanisms that lead to the differences. Network differential analysis involves the estimation of at least two networks and can be more challenging with the significantly increased number of parameters. In this chapter, we further develop the assisted analysis strategy, take advantage of multidimensional profiling data, and propose incorporating regulator information to improve the identification of key genes (that lead to differences in GE networks). An effective computational algorithm is developed. Comprehensive simulation is conducted, showing that the proposed approach can outperform benchmarks in terms of identification accuracy. The analysis of TCGA lung adenocarcinoma (LUAD) data leads to findings with sensible interpretations and different from the alternatives. Overall, this study can significantly expand the scope of differential network analysis and assisted analysis. A manuscript

based on this chapter will be submitted for publication soon.

4.1 Introduction

Gene expression data has been playing a uniquely important role in cancer research. The analysis of gene expression data has led to a deeper elucidation of cancer etiology as well as actionable targets for the development of treatment/prevention strategies. An important type of analysis is to compare gene expression properties under different conditions, which can be cancer and normal, deceased and alive, different subtypes, and others. For practical examples of such analysis, we refer to [164–167], and others.

When comparing gene expression properties between conditions, the simplest way is to compare (normalized) means (medians, etc.), which leads to the commonly conducted differential gene identification analysis. It has been recognized that the first moment (of gene expression distribution) does not contain all relevant information. Accordingly, variance (second moment) based analysis has been conducted, motivated by the genetic principle that higher variations indicate less stable gene expressions, which may increase disease susceptibility and severity. Further advancing from such *marginal analysis* – which analyzes one gene at a time, *network-based analysis* has been conducted. Such analysis takes a system perspective and describes properties of not only individual gene expressions but also their interconnections.

Gene expression network analysis can be mainly classified into two categories: unconditional analysis and conditional analysis. In unconditional analysis, the goal is to quantify whether any two gene expressions are independent while "ignoring" other genes. A representative example of unconditional analysis is the WGCNA (Weighted Gene Co-Expression Network Analysis) pioneered by Peter Langfelder and Steve Horvath [84]. As demonstrated in the WGCNA and other analyses, the variance-covariance matrix is often the simplest starting point of unconditional analysis. Unconditional gene expression networks can be both directional and undirectional, both weighted and unweighted, and both sparse and dense [97]. In comparison, in conditional analysis, the goal is to quantify whether two gene expressions are independent conditional on the rest of the genes. The "simplest" and most extensively conducted conditional analysis is perhaps the Gaussian Graphical Model (GGM), under which it is assumed that gene expressions have joint normal distributions. Under this specific assumption, determining conditional independence is equivalent to determining whether the precision matrix (which is the inverse of the variance-covariance matrix) has the corresponding element being zero. When the normality assumption is too stringent, approaches have been developed to relax the normality assumption, replace Pearson's correlation (in the variance-covariance matrix) with robust, for example Kendall's tau, correlations, and then proceed in the same way as under the GGM. In both unconditional and conditional analysis, when sparsity is desirable (which is usually the case for gene expression analysis), regularization can be applied. It is noted that this may be routinely needed in conditional analysis, which demands the joint estimation of a large number of parameters. For example, for GGM, the graphical Lasso approach, which applies Lasso penalization to GGM estimation, has been popular.

Consider comparing gene expressions between two conditions. Here we note that the discussions and proposed approach are also applicable to the comparison of more than two conditions. The first step, very naturally, is to determine whether the gene expression networks under the two conditions are significantly different. This naturally poses a hypothesis testing problem. A "straightforward" approach is to first take the difference between two networks (variance-covariance matrices under unconditional analysis, precision matrices under conditional analysis, etc.), and then take a certain norm of this difference. Norms considered in the literature include the Frobenius norm, ℓ_∞ norm, and others [168, 169]. Some studies have derived the asymptotic distributions of such norms, which is often a very challenging problem [170]. An alternative solution is to apply, for example, permutation-based techniques [171, 172].

For many "simple" problems, for example the comparison between subtypes or between normal and cancer, significant differences are apparently expected – this has been confirmed by many published analyses. In this case, the natural next step is to identify which genes lead to the differences. This corresponds to differential gene analysis [173, 174], which has been established as having important implications. For both unconditional and conditional analysis, with the difference of networks, a simple approach is to examine which genes correspond to the large elements. A statistically more rigorous approach is via spectral clustering [140, 175]. With the difference of networks, spectral clustering amounts to conducting SVD (singular value decomposition). In our analysis, sparsity is assumed, under which it is postulated that only a small number of genes contribute to the difference. Then regularization is needed along with SVD to differentiate "signals" from "noises". A "straightforward" choice is the SSVD (sparse SVD) technique, with which some components of the singular vectors can be estimated as exactly zero [176]. When the first sparse singular vector is estimated, the nonzero components correspond to the first *gene expression cluster* that causes the difference [175]. If desirable, the SSVD procedure can be continued to identify the subsequent gene expression clusters that also contribute to the difference. This analysis pipeline has been developed in the literature [175] and shown to have satisfactory performance. Here we note that this analysis can be conducted in the same manner for both unconditional and conditional networks.

As well recognized in the literature, network analysis is challenged by the high dimensionality of parameters and limited sample size, which may lead to unsatisfactory estimation and identification [177]. This can be especially true in the identification of difference, where at least two networks need to be estimated. Gene expressions are heavily regulated. We note that here we take a loose definition and generically refer to molecular mechanisms that can affect gene expression levels, including but not limited to copy number variation and other DNA mutations, methylation, and microRNA, as "regulators". Under other (possibly simpler) contexts, assisted analysis has been developed to take advantage of regulator information and assist the analysis of gene expressions. One example is collaborative regression [46]. Here for a low-dimensional outcome, a regression model is built using gene expressions only, and a separate model is built using regulators only. With the gene-expression-regulator relationship, this approach promotes that the two models lead to similar estimates for the outcome variable. This approach may be limited by not explicitly accounting for the regulation relationship. To tackle this problem, the ARMI approach is developed which includes the addition gene-expression-regulator modeling step [47]. In addition, it also has built-in robustness to accommodate long-tailed outcome distributions. Assisted analysis has also been conducted in clustering and other contexts. In a very recent study (Chapter 2), assisted analysis has also been conducted on gene expression networks. The goal of that study is to more accurately estimate the conditional gene-expression-only and gene-expression-regulator networks, through linking them via a hierarchy. It is noted that this approach has been developed for conditional networks only. The aforementioned studies have provided extensive evidence on the effectiveness of gene expression analysis assisted by regulators.

In this study, we consider the comparison of gene expression networks between two (or more) conditions. As mentioned above, multiple comparison scenarios can be accommodated. For each subject (under all conditions), it is assumed that both gene expression and regulator measurements are available. Here it is noted that, as in the published assisted analysis, the proposed analysis does not demand all collected regulators are relevant, or all relevant regulators are collected – as such, it can be sufficiently flexible. Our analysis goal, as in some published studies, is to identify the subset of gene expressions that contribute to the difference of gene expression networks, built on the regularized spectral clustering technique. This study may advance from the existing literature in the following important aspects. First, it advances from the differential analysis based on mean (median) and variance by examining the interconnections among genes. Second, it advances from the existing difference-in-networks analysis by taking advantage of the information in regulators. It also advances from the existing spectral clustering analysis by simultaneously analyzing gene expressions and regulators. Third, its data settings are fundamentally different from those in Chapter 2. In particular, in Chapter 2, the two networks have different formulating components: one without regulators, and the other with regulators. In contrast, in the present analysis, regulator data is available for all subjects under both conditions. This study also advances by conducting both conditional and unconditional network analysis. Fourth, this study also advances from the regression-based assisted analysis by conducting more complex network analysis.

4.2 Methods

4.2.1 Strategy

The first challenge of differential network analysis is to quantify network changes. A very recent effort that tackles a related task uses the Generalized Hamming Distance (GHD) to quantify the differences between two networks, and then adopts an iterative technique to identify the set of genes that contribute most to the change [168]. Let Y_1, Y_2 be the vectors of GE variables from two different groups or stages, and X_1, X_2 be the vectors for corresponding regulator variables. Denote the GE networks constructed using Y_1 and Y_2 as $G_1 \in \mathbb{R}^{p\times p}$ and $G_2 \in \mathbb{R}^{p\times p}$, respectively, and the corresponding regulator networks constructed using X_1 and X_2 as $R_1 \in \mathbb{R}^{q\times q}$ and $R_2 \in \mathbb{R}^{q\times q}$. It is noted that here there is a slight abuse of notation. The ”networks” describe the interconnections among variables. Following the idea of GHD, a natural alternative measure of the GE network difference is to form a matrix $G_{\text{diff}} \in \mathbb{R}^{p\times p}$ with elements $(G_{1,ij} - G_{2,ij})$, where $G_{1,ij}$ and $G_{2,ij}$ are the (i, j)th elements in G_1 and G_2, respectively. Similarly, we also define/compute the network difference for regulators, which is denoted as $R_{\text{diff}} \in \mathbb{R}^{q\times q}$.

A “classic” method to detect the key contributors to network changes is sparse singular value decomposition (SSVD), which has already been widely used in clustering and identifying interpretable row-column associations with high-dimensional data matrices [176]. In our case, the benchmark analysis is to apply SSVD to G_{diff}, the network changes of GEs. The singular value decomposition (SVD) of G_{diff} can be written as $G_{\text{diff}} = VDW^\top = \sum_{k=1}^{p} s_k^G v_k w_k^\top$, where $V = (v_1, \cdots, v_p)$ and $W = (w_1, \cdots, w_p)$ are two matrices of orthonormal singular vectors, and $D = \text{diag}(s_1^G, \cdots, s_p^G)$ is a diagonal matrix with positive singular values $s_1^G \geq \cdots \geq s_p^G$ on its diagonal. SVD decomposes G_{diff} into a summation of rank-one matrices $s_k^G v_k w_k^\top$. With the ordered singular values, the first term, $G_{\text{diff}} \approx G_{\text{diff}}^{(1)} \equiv s_1^G v_1 w_1^\top$ provides the best rank-one approximation to G_{diff}. By using regularization, SSVD seeks a sparse low-rank matrix approximation. It requires that the vector v_k is sparse. Spectrum analysis theory stipulates that genes identified in $G_{\text{diff}}^{(1)}$ represents the key contributors to the network differences. This is easy to comprehend with $s_1^G v_1 w_1^\top$ providing the best rank-one approximation and containing the most information of G_{diff} . It is noted that regulator information is not accommodated in the conventional SSVD analysis. Now consider the analysis of regulator data. And similarly, we examine the differences in regulator networks. Similarly, the rank-one approximation to R_{diff} can be written as $R_{\text{diff}} = U\tilde{D}Z^{\top} \approx R_{\text{diff}}^{(1)} \equiv s_1^R u_1 z_1^{\top}$, where $U = (u_1, \cdots, u_q)$ and $Z = (z_1, \cdots, z_p)$ are two matrices of orthonormal singular vectors, and $\tilde{D} = \text{diag}(s_1^R, \cdots, s_q^R)$ is a diagonal matrix with positive singular values $s_1^R \geq \cdots \geq s_q^R$ on its diagonal.

The proposed assisted differential network analysis is motivated by the work of Li et al. [23] and Lee et al. [176], whose strategies are to reinforce "concordance" between GE- and regulator-based clustering analysis. As in the published SSVD and spectral clustering analyses, we first focus on extracting the first layers of the GE and regulator matrices; the subsequent layers can be extracted sequentially from the residual matrices after removing the preceding layers. The strategy of assisted analysis has been developed in early studies by Dr. Ma's group [47]. And it has been found that, with the assistance of information contained in regulators, assisted analysis can cost-effectively improve identification and estimation over that limited to GE data only. Here, we adopt this strategy in our differential network analysis with the intention to improve the identification of key contributors to network changes. With a little abuse of notations, we propose the assisted differential network analysis objective function as:

$$\begin{aligned} Q(\boldsymbol{v}, \boldsymbol{w}, s^G, \boldsymbol{u}, \boldsymbol{z}, s^R) =& \|G_{\text{diff}} - s^G \boldsymbol{v}\boldsymbol{w}^{\top}\|_F^2 + \|R_{\text{diff}} - s^R \boldsymbol{u}\boldsymbol{z}^{\top}\|_F^2 \\ & + P_{\boldsymbol{v}} + P_{\boldsymbol{w}} + P_{\boldsymbol{u}} + P_{\boldsymbol{z}} - P_{\text{similarity}} \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} & P_{\boldsymbol{v}} = \rho\left(|\boldsymbol{v}|; \lambda_1, a\right), \quad P_{\boldsymbol{w}} = \rho\left(|\boldsymbol{w}|; \lambda_1, a\right), \\ & P_{\boldsymbol{u}} = \rho\left(|\boldsymbol{u}|; \lambda_2, a\right), \quad P_{\boldsymbol{z}} = \rho\left(|\boldsymbol{z}|; \lambda_2, a\right), \\ & P_{\text{similarity}} = \lambda_3 I(\boldsymbol{v} \neq 0) \cdot [cor(Y_1, X_1) + cor(Y_2, X_2)] \cdot I(\boldsymbol{u} \neq 0). \end{aligned}$$

where $\boldsymbol{v}$ and $\boldsymbol{w}$ are the first orthogonal singular vectors of G_{diff}; s^G is the first singular value of G_{diff}; $\boldsymbol{u}$ and $\boldsymbol{z}$ are the first orthogonal singular vectors of R_{diff}; s^R is the first singular value of R_{diff}; $\rho(|\cdot|;\lambda,a)=\lambda\int_0^{|\cdot|}\left(1-\frac{x}{\lambda a}\right)_+ dx$ is the MCP (minimax concave penalty); λ_1 is a tuning parameter controlling the penalty on singular values of GEs; λ_2 is a tuning parameter controlling the penalty on singular values of regulators; λ_3 is a tuning parameter controlling the promotion of correlation of GEs and regulators; and a is the regularization parameter.

Rationale The proposed method has been motivated by the following considerations. Similar to other assisted analyses, it involves the joint analysis of GEs and their regulators. However, the detailed strategy differs significantly from that in Chapter 2 and others. More specifically, in (4.1), if λ_3 in $P_{similarity}$ reduces to 0, then the analysis simplifies to two differential network analyses, with one on GEs and the other on regulators. More specifically, the SSVD-based analyses can identify important GEs (that contribute to the differences in GE networks) and regulators (that contributes to the differences in regulator networks). The key advancement is the introduction of $P_{similarity}$, which connects the two analyses. This spirit is somewhat similar to that in Chapter 2, however, the strategy is significantly different. $P_{similarity}$ *encourages the set of important GEs and that of important regulators to be correlated.* The underlying assumption is that a set of important regulators cause significant differences in the regulator networks; and they regulate a set of important GEs that cause significant differences in the GE networks. It is noted that using correlations to describe GE-regulator relationships may be too simplified. However, it has been adopted in [23] and others and shown as effective. In principle, objective function (4.1) itself is sufficient for numerical and theoretical investigation purposes. To simplify computation, in the following section, an approximation is introduced, which does not change the key properties of the proposed approach but can facilitate the adoption of existing computational techniques.

4.2.2 Computation

As discussed above, when $\lambda_3 = 0$, objective function (4.1) simplifies to two SSVDs, for which there are effective algorithms. To take advantage of such algorithms, our strategy is to approximate the newly added penalty − which involves indicator functions and is not differentiable – and make it differentiable. With the approximation, the newly added penalty can be combined with the goodness-of-fit measures. Specifically, we consider the approximation:

$$I(v_j \neq 0) \approx \left(1 - \exp\left(-\frac{v_j^2}{\tau}\right)\right) \approx \left[1 - \exp\left(-\frac{\tilde{v}_j^2}{\tau}\right)\right] + \exp\left(-\frac{\tilde{v}_j^2}{\tau}\right) \cdot \frac{2\tilde{v}_j}{\tau} \cdot (v_j - \tilde{v}_j),$$

where $\tilde{v}_j$ is a point not "far away" from the last round of estimation. It is noted that similar approximations have been adopted in the literature, and that other approximations to the indicator function may work equally well. Then, the approximation of $I(\boldsymbol{v} \neq 0)$ in matrix form is:

$$I(\boldsymbol{v} \neq 0) \approx C_1(\tilde{\boldsymbol{v}}) + C_2(\tilde{\boldsymbol{v}})(\boldsymbol{v} - \tilde{\boldsymbol{v}}),$$

where $C_1(\tilde{\boldsymbol{v}})$ is a $p \times 1$ vector, and $C_2(\tilde{\boldsymbol{v}})$ is a diagonal matrix, both of which are constant depending on $\tilde{\boldsymbol{v}}$. Similarly, we can approximate $I(\boldsymbol{u} \neq 0)$ at $\tilde{\boldsymbol{u}}$ as $I(\boldsymbol{u} \neq 0) \approx C_3(\tilde{\boldsymbol{u}}) + C_4(\tilde{\boldsymbol{u}})(\boldsymbol{u} - \tilde{\boldsymbol{u}})$.

As in "ordinary" coordinate descent computations, we alternately minimize objective function 4.1 with respect to $\boldsymbol{v}$, $\boldsymbol{w}$, $\boldsymbol{u}$, and $\boldsymbol{z}$, after plugging in the approximations. To simplify notation, we denote $T = [cor(Y_1, X_1) + cor(Y_2, X_2)]$. The algorithm is summarized below.

Algorithm

Step 1. Initialization. Apply the standard SVD to G_{diff} and R_{diff}, respectively. Let $\{\tilde{s}^G, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}}; \tilde{s}^R, \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{z}}\}$ denote the first SVD triplets. It is noted that when dimensionality is high, SSVD can be adopted to stabilize estimation and distinguish signals from noises.

Step 2. Update:

(a) Set $\boldsymbol{v}_{\text{temp}} = \frac{1}{2} sign\,[2G_{\text{diff}}\tilde{\boldsymbol{w}} + \lambda_3 C_2(\tilde{\boldsymbol{v}}) T I(\tilde{\boldsymbol{u}} \neq 0)] \cdot [|2G_{\text{diff}}\tilde{\boldsymbol{w}} + \lambda_3 C_2(\tilde{\boldsymbol{v}}) T I(\tilde{\boldsymbol{u}} \neq 0)| - \dot{\rho}(|\tilde{\boldsymbol{v}}|; \lambda_1, a)]_+$. Let $s^G = \sqrt{||\boldsymbol{v}_{\text{temp}}||_F \cdot ||\tilde{\boldsymbol{w}}||_F}$ and $\boldsymbol{v} = \boldsymbol{v}_{\text{temp}}/s^G$.

(b) Set $\boldsymbol{w}_{\text{temp}} = \frac{1}{2} sign\,[G_{\text{diff}}\boldsymbol{v}] \cdot [2G_{\text{diff}}\boldsymbol{v} - \dot{\rho}(|\tilde{\boldsymbol{w}}|; \lambda_1, a)]_+$. Let $s^G = \sqrt{||\boldsymbol{v}||_F \cdot ||\boldsymbol{w}_{\text{temp}}||_F}$ and $\boldsymbol{w} = \boldsymbol{w}_{\text{temp}}/s^G$.

(c) Set $\boldsymbol{u}_{\text{temp}} = \frac{1}{2} sign\left[2R_{\text{diff}}\tilde{\boldsymbol{z}} + \lambda_3 C_4(\tilde{\boldsymbol{u}})^\top T^\top I(\boldsymbol{v} \neq 0)\right] \cdot \left[|2R_{\text{diff}}\tilde{\boldsymbol{z}} + \lambda_3 C_4(\tilde{\boldsymbol{u}})^\top T^\top I(\boldsymbol{v} \neq 0)| - \dot{\rho}(|\tilde{\boldsymbol{u}}|; \lambda_2, a)\right]_+$. Let $s^R = \sqrt{||\boldsymbol{u}_{\text{temp}}||_F \cdot ||\tilde{\boldsymbol{z}}||_F}$ and $\boldsymbol{u} = \boldsymbol{u}_{\text{temp}}/s^R$.

(d) Set $\boldsymbol{z}_{\text{temp}} = \frac{1}{2} sign\left[R_{\text{diff}}\boldsymbol{u}\right] \cdot \left[2R_{\text{diff}}\boldsymbol{u} - \dot{\rho}(|\tilde{\boldsymbol{z}}|; \lambda_2, a)\right]_+$. Let $s^R = \sqrt{||\boldsymbol{u}||_F \cdot ||\boldsymbol{z}_{\text{temp}}||_F}$ and $\boldsymbol{z} = \boldsymbol{z}_{\text{temp}}/s^R$.

Step 3. Set $\tilde{\boldsymbol{v}} = \boldsymbol{v}$, $\tilde{\boldsymbol{w}} = \boldsymbol{w}$, $\tilde{\boldsymbol{u}} = \boldsymbol{u}$, and $\tilde{\boldsymbol{z}} = \boldsymbol{z}$. Repeat Step 2 until convergence.

In data analysis, we conclude convergence when the difference between the estimates from two consecutive steps is smaller than a prespecified cutoff. Convergence properties can be established following those for SSVD, which is omitted here. In all of our simulation and data analysis, convergence is achieved within 20 iterations. The proposed approach involves three tuning parameters λ_1, λ_2, and λ_3. λ_1, λ_2 controls sparsity, as in "regular" SSVD; and λ_3 controls the level of correlation between important GEs and important regulators. In numerical analysis, we conduct a three-dimensional grid search. In simulation study, considering that different approaches (the proposed and alternatives) have different numbers of tuning parameters, we also consider a ROC (Receiver Operating Characteristic) based approach for evaluation, which can "eliminate" the impact of tuning parameter selection. To facilitate data analysis, we have developed R programs implementing the proposed approach and made them publicly available at `https://github.com/DeniseYi`.

4.3 Simulation

For modeling the relationship between GEs and regulators, following [19], we consider

$$Y = XB + W, \tag{4.2}$$

where X is the $n \times q$ data matrix of regulators; Y is the $n \times p$ data matrix of GEs; B is the $q \times p$ matrix of unknown regression coefficients and represents the "transition" from regulators to GEs; and W is an $n \times p$ matrix and accommodates both "random errors" as well as regulation mechanisms not measured. The expression level of a specific gene is only affected by a small number of regulators (that is, B is sparse). However, the set of regulators and strengths of their effects are unknown in a real-world problem. For the structure of the covariance matrix of the regulators, Σ_X, we consider two different scenarios: A1) a block diagonal structure with block size ten and each block is in the Erdos-Renyi structure.

Briefly, we generate the Erdos-Renyi network that has probability 0.05 for drawing an edge between two arbitrary nodes. Regulator changes between two different groups or stages are constructed as the change of one block matrix. A2) on the basis of a), we change some blocks to diagonal sub-matrices. Regulator changes are constructed as the change of some grouped regulators in one block plus the change of several isolated regulators.

For the regression coefficient matrix, B, we consider four different structures. B1) Strong effect block diagonal: A block diagonal structure with all elements in the blocks generated from a uniform distribution $U(0.9, 1)$. The dimensions of the blocks are matched with those of the regulator covariance matrix. B2) Strong effect "milky way": on the basis of structure B1), a small portion (2%) of the off-block-diagonal elements are randomly generated from the same uniform distribution $U(0.9, 1)$. Their positions are randomly simulated. B3) Weak effect block diagonal: Different from B1), all elements in the blocks are generated from a uniform distribution $U(0.27, 0.3)$. B4) Weak effect "milky way": on the basis of structure B3), a small portion (2%) of the off-block-diagonal elements are randomly generated from the same uniform distribution $U(0.27, 0.3)$.

For the structure of the covariance matrix of the noise, Σ_W,we consider both independently errors and correlated errors. C1) Independent errors are generated from diagonal matrix with diagonal elements from $N(1, 0.1)$. C2) The covariance matrix of the correlated errors has the same block structure as the GEs. Each block is generated from $MVN(0, \Sigma_{p_i}(\rho))$ – a multivariate normal distribution with mean zero and covariance $\Sigma_{p_i}(\rho)) = \rho^{|i-j|}$, where p_i is the size of the block i and $\rho = 0.3$ in simulation. The data matrix of GEs are simulated from the outcome generating model (4.2). Set $n = 200$ and $(p, q) = (50, 100)$. Here we note that, although smaller than n, the values of p and q are reasonable. Even though whole-genome studies may have a much higher dimensionality, to improve analysis reliability, it is a common practice to focus on a smaller set of genes, which can be screened biologically or statistically. It is also noted that even with moderate p and q, the number of parameters involved is still much larger than n.

Simulation is conducted to assess the performance of assisted SSVD in different scenarios. In addition, as a reference, we consider the following alternatives.

Alt.1 **SVD.** Consider the SVD estimate. Similarly, we obtain the rank-one approximation to G_{diff} and R_{diff}, respectively. Then find the first group of significant genes/regulators to the difference. This is the benchmark approach, involves GEs or regulators only.

Alt.2 **IRLBA.** Consider the SSVD estimate using the augmented implicitly restarted Lanczos bidiagonalization approach [178]. This approach seeks the rank-one approximation to the difference matrices, but with the requirement that the singular vectors are sparse. We directly use the `'irlba'` R package [179]. It conducts SSVD to GEs and regulators separately.

Alt.3 **BSSVD.** Consider the biclustering SSVD estimate using Lasso penalty developed in [176]. It is similar to Alt.2, but is optimized by a different algorithm. Again, we consider both G_{diff} and R_{diff}, but apply SSVD to them separately.

The proposed assisted SSVD and alternative approaches all involve tuning parameters. For Alt.1, we can apply a series of cutoffs to obtain the most significant genes/regulators, and thus can be viewed as a tuning. For Alt.2, there is a tuning parameter controlling the number of non-zero elements in the singular vector. Focusing on specific tuning parameter values may not generate a comprehensive picture. To solve this problem, we adopt the ROC (Receiver Operating Characteristic) approach, which considers a set of tuning parameter values, evaluates identification at each value, and uses the ROC-based measures for evaluation. This evaluation approach has been extensively adopted in the literature. In our simulation, the AUC (area under the ROC curve) is adopted as the overall identification accuracy measure.

AUCs are computed based on 100 replicates. In each scenario, we compare estimations of four matrices including $\widehat{\Sigma}_X$, $\widehat{\Sigma}_Y$, $\widehat{\Omega}_X$, and $\widehat{\Omega}_Y$ among different approaches. Results under strong association between X and Y are shown in Table 4.1, and those under weak association are shown in Table 4.2. It is observed that the proposed assisted SSVD approach has competitive performance across the whole spectrum of simulation. In general, the proposed approach has the best performance, followed by two SSVD approaches. The SVD approach has the least satisfactory performance. It is noticed that under strong association when the covariance matrix of X is set as block-diagonal, the estimation $\widehat{\Omega}_X$

usually has the largest AUCs across all approaches; whereas when the covariance matrix of X is set to contain isolated variables, the estimation $\widehat{\Sigma}_Y$ usually has the best AUCs across all approaches (See Table 4.1). For example in the first row of Table 4.1, when the covariance matrix of X is block-diagonal, the coefficient matrix is block-diagonal, and the error terms are independent, the AUC value of $\widehat{\Omega}_X$ of the proposed method is 0.670 (sd = 0.130), it is the largest among all AUCs in this scenario. The three other alternatives have the AUCs 0.623 (sd = 0.150), 0.658 (sd = 0.154), and 0.658 (sd = 0.150). In the fifth row of Table 4.1, when the covariance matrix of X contains isolated variables, the coefficient matrix is block-diagonal, and the error terms are independent, the AUC value of $\widehat{\Sigma}_Y$ of the proposed method is 0.812 (sd = 0.171). The three other alternatives have the AUCs 0.737 (sd = 0.210), 0.801 (sd = 0.266), and 0.805 (sd = 0.267). The proposed approach has the largest AUC with the smallest sd.

This pattern does not exist under weak association between X and Y. Under weak association, the estimation $\widehat{\Sigma}_X$ has the most satisfactory AUCs across different approaches in general (See Table 4.2). For example in the first row of Table 4.2, when the covariance matrix of X is block-diagonal, the coefficient matrix is block-diagonal, and the error terms are independent, the AUC value of $\widehat{\Sigma}_X$ of the proposed method is 0.703 (sd = 0.152). It is the largest among all AUCs in this scenario, followed by Alt.3: 0.621 (sd = 0.175), Alt.2: 0.593 (sd = 0.181), and Alt.1: 0.554 (sd = 0.177). As expected, because the proposed approach jointly analyze GEs and regulators and borrow information with accounting for the regulation relationship, they have superior performance. As shown in Tables 4.3 and 4.4 in Appendix, we consider $(p, q) = (50, 50)$ and $n = 200$ under different scenarios as above. It is also observed that the proposed approach has competitive performance across the whole spectrum of simulation.

4.4 Data Analysis

TCGA (The Cancer Genome Atlas) is one of the largest and most comprehensive cancer projects jointly organized by the NCI and NHGRI. For over thirty cancer types, it has published comprehensive molecular and other types of data. We analyze TCGA data because

Table 4.1: Simulation results of AUC: mean×100 (sd×100) for $p = 50$, $q = 100$, $n = 200$ (Strong association between X and Y).

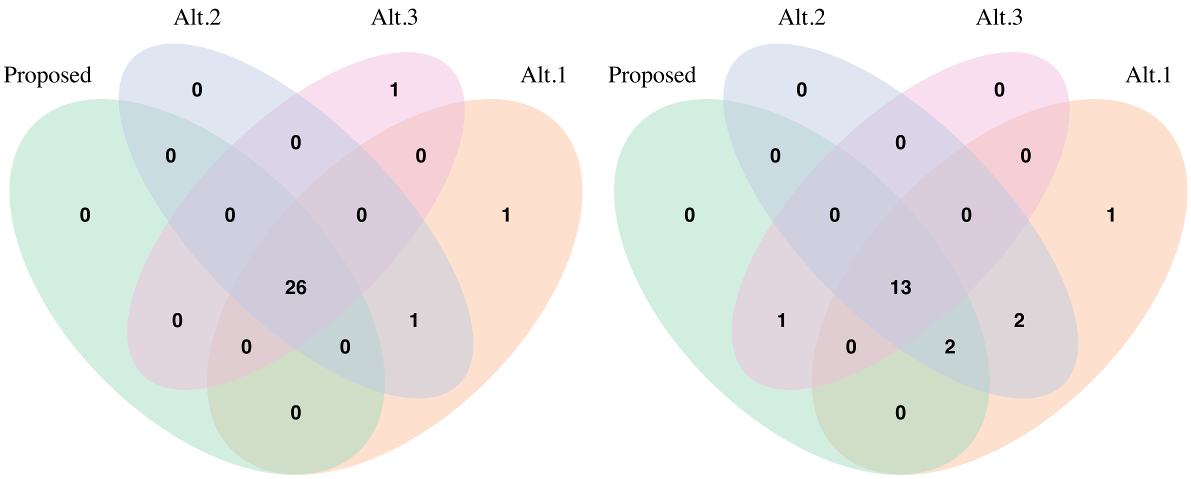
Scenario	Approach	X	B	W	$\widehat{\Sigma}_X$	$\widehat{\Sigma}_Y$	$\widehat{\Omega}_X$	$\widehat{\Omega}_Y$
A1)B1)C1)	Proposed	Block-diagonal	Block-diagonal	Independent	64.3 (13.0)	48.5 (18.5)	67.0 (15.0)	60.5 (15.1)
	Alt.1	Block-diagonal	Block-diagonal	Independent	58.8 (14.0)	40.1 (25.7)	62.3 (15.0)	42.9 (13.5)
	Alt.2	Block-diagonal	Block-diagonal	Independent	62.8 (14.7)	43.7 (28.2)	65.8 (15.4)	49.2 (15.2)
	Alt.3	Block-diagonal	Block-diagonal	Independent	62.8 (14.6)	46.6 (29.0)	65.8 (15.0)	49.6 (13.4)
A2)B1)C1)	Proposed	Containing isolate	Block-diagonal	Independent	76.4 (7.8)	81.2 (17.1)	45.5 (11.0)	57.9 (15.0)
	Alt.1	Containing isolate	Block-diagonal	Independent	67.9 (8.2)	73.7 (21.0)	41.5 (10.1)	39.3 (12.4)
	Alt.2	Containing isolate	Block-diagonal	Independent	72.3 (8.5)	80.1 (26.6)	45.2 (10.5)	49.3 (15.3)
	Alt.3	Containing isolate	Block-diagonal	Independent	74.8 (8.1)	80.5 (26.7)	46.3 (9.9)	47.9 (15.4)
A1)B2)C1)	Proposed	Block-diagonal	Milky-way	Independent	61.8 (17.7)	52.1 (12.5)	69.1 (13.5)	50.4 (9.8)
	Alt.1	Block-diagonal	Milky-way	Independent	57.3 (17.1)	48.9 (17.7)	61.7 (14.2)	35.1 (9.7)
	Alt.2	Block-diagonal	Milky-way	Independent	61.2 (18.0)	54.3 (18.7)	66.8 (14.5)	40.9 (10.7)
	Alt.3	Block-diagonal	Milky-way	Independent	63.4 (17.8)	53.1 (17.1)	68.5 (13.7)	39.7 (10.7)
A2)B2)C1)	Proposed	Containing isolate	Milky-way	Independent	75.0 (7.6)	80.6 (17.0)	46.2 (9.9)	54.5 (13.9)
	Alt.1	Containing isolate	Milky-way	Independent	69.3 (8.7)	67.6 (25.3)	41.7 (9.4)	35.4 (12.9)
	Alt.2	Containing isolate	Milky-way	Independent	73.4 (9.4)	76.7 (28.7)	45.4 (9.6)	44.1 (15.3)
	Alt.3	Containing isolate	Milky-way	Independent	76.0 (8.6)	80.5 (25.3)	45.9 (9.5)	45.9 (12.5)
A1)B1)C2)	Proposed	Block-diagonal	Block-diagonal	Banded	60.4 (14.3)	52.8 (20.7)	70.0 (12.1)	56.1 (14.5)
	Alt.1	Block-diagonal	Block-diagonal	Banded	55.5 (15.5)	48.7 (23.7)	61.6 (13.8)	40.8 (14.8)
	Alt.2	Block-diagonal	Block-diagonal	Banded	59.0 (16.3)	53.3 (26.9)	65.4 (14.4)	47.2 (17.0)
	Alt.3	Block-diagonal	Block-diagonal	Banded	60.0 (15.3)	52.3 (27.5)	66.2 (12.4)	47.4 (16.7)
A2)B1)C2)	Proposed	Containing isolate	Block-diagonal	Banded	78.3 (8.3)	86.5 (14.4)	44.8 (9.6)	65.2 (18.4)
	Alt.1	Containing isolate	Block-diagonal	Banded	71.6 (8.3)	74.6 (21.4)	40.3 (9.6)	45.4 (16.1)
	Alt.2	Containing isolate	Block-diagonal	Banded	76.4 (8.7)	83.6 (26.9)	44.5 (10.1)	54.6 (18.7)
	Alt.3	Containing isolate	Block-diagonal	Banded	77.6 (7.7)	86.3 (24.0)	45.4 (10.1)	55.1 (18.3)
A1)B2)C2)	Proposed	Block-diagonal	Milky-way	Banded	60.2 (16.4)	52.3 (14.4)	65.8 (13.2)	47.1 (12.8)
	Alt.1	Block-diagonal	Milky-way	Banded	55.6 (17.1)	48.0 (18.4)	62.1 (12.8)	31.1 (8.7)
	Alt.2	Block-diagonal	Milky-way	Banded	58.9 (17.5)	54.3 (17.9)	65.9 (13.3)	35.7 (10.0)
	Alt.3	Block-diagonal	Milky-way	Banded	61.2 (17.6)	51.6 (17.2)	67.4 (13.0)	37.6 (10.3)
A2)B2)C2)	Proposed	Containing isolate	Milky-way	Banded	76.4 (7.4)	78.0 (13.5)	44.3 (9.3)	59.1 (13.2)
	Alt.1	Containing isolate	Milky-way	Banded	70.2 (7.7)	69 (19)	40.1 (8.9)	43.6 (14.1)
	Alt.2	Containing isolate	Milky-way	Banded	74.7 (7.9)	74.2 (23.1)	43.4 (9.3)	51.8 (15.8)
	Alt.3	Containing isolate	Milky-way	Banded	76.2 (7.2)	76.0 (23.4)	45.4 (9.7)	52.0 (15.7)

Table 4.2: Simulation results of AUC: mean×100 (sd×100) for $p = 50$, $q = 100$, $n = 200$ (Weak association between X and Y).

Scenario	Approach	X	B	W	$\widehat{\Sigma}_X$	$\widehat{\Sigma}_Y$	$\widehat{\Omega}_X$	$\widehat{\Omega}_Y$
A1)B3)C1)	Proposed	Block-diagonal	Block-diagonal	Independent	70.3 (15.2)	46.7 (18.1)	68.2 (13.3)	65.7 (13.8)
	Alt.1	Block-diagonal	Block-diagonal	Independent	55.4 (17.7)	37.9 (20.2)	62.0 (14.6)	46.1 (12.3)
	Alt.2	Block-diagonal	Block-diagonal	Independent	59.3 (18.1)	43.6 (19.1)	66.0 (14.5)	54.3 (14.4)
	Alt.3	Block-diagonal	Block-diagonal	Independent	62.1 (17.5)	43.9 (20.0)	67.8 (12.4)	53.3 (15.4)
A2)B3)C1)	Proposed	Containing isolate	Block-diagonal	Independent	83.6 (6.1)	82.3 (19.3)	45.7 (10.9)	59.3 (11.9)
	Alt.1	Containing isolate	Block-diagonal	Independent	71.3 (7.7)	62.0 (23.2)	41.5 (10.5)	39.9 (10.0)
	Alt.2	Containing isolate	Block-diagonal	Independent	75.9 (8.2)	70.5 (27.6)	45.5 (10.8)	48.3 (11.7)
	Alt.3	Containing isolate	Block-diagonal	Independent	77.7 (6.8)	70.0 (26.7)	45.9 (10.6)	51.3 (11.0)
A1)B4)C1)	Proposed	Block-diagonal	Milky-way	Independent	72.0 (12.7)	57.9 (16.4)	65.2 (14.2)	59.2 (10.1)
	Alt.1	Block-diagonal	Milky-way	Independent	55.6 (14.2)	50.8 (17.1)	60.8 (14.6)	41.4 (10.7)
	Alt.2	Block-diagonal	Milky-way	Independent	58.4 (15.0)	56.0 (17.6)	64.6 (15.1)	47.0 (10.4)
	Alt.3	Block-diagonal	Milky-way	Independent	59.7 (16.0)	52.4 (15.6)	66.1 (15.6)	48.1 (10.6)
A2)B4)C1)	Proposed	Containing isolate	Milky-way	Independent	82.0 (7.7)	83.0 (21.1)	45.5 (10.7)	56.2 (14.1)
	Alt.1	Containing isolate	Milky-way	Independent	71.5 (9.1)	65.4 (26.0)	41.5 (11.6)	36.3 (14.4)
	Alt.2	Containing isolate	Milky-way	Independent	76.6 (9.0)	75.4 (28.9)	44.9 (12.0)	44.1 (16.0)
	Alt.3	Containing isolate	Milky-way	Independent	78.0 (8.3)	76.0 (28.4)	45.0 (10.9)	46.3 (15.4)
A1)B3)C2)	Proposed	Block-diagonal	Block-diagonal	Banded	69.4 (14.7)	49.8 (18.9)	69.6 (13.3)	57.5 (15.0)
	Alt.1	Block-diagonal	Block-diagonal	Banded	54.9 (15.5)	43.0 (21.6)	64.7 (13.0)	44.2 (14.0)
	Alt.2	Block-diagonal	Block-diagonal	Banded	58.9 (16.9)	48.6 (21.8)	68.9 (14.2)	51.3 (16.4)
	Alt.3	Block-diagonal	Block-diagonal	Banded	60.4 (15.3)	48.1 (22.9)	68.9 (14.1)	50.4 (14.9)
A2)B3)C2)	Proposed	Containing isolate	Block-diagonal	Banded	82.8 (6.4)	84.5 (19.4)	44.8 (10.6)	55.4 (18.4)
	Alt.1	Containing isolate	Block-diagonal	Banded	69.8 (7.5)	65.2 (24.1)	40.3 (10.8)	37.9 (16.5)
	Alt.2	Containing isolate	Block-diagonal	Banded	73.9 (7.9)	73.1 (28.0)	44.0 (10.8)	46.7 (20.0)
	Alt.3	Containing isolate	Block-diagonal	Banded	74.8 (7.4)	75.6 (25.4)	45.6 (11.1)	46.8 (19.1)
A1)B4)C2)	Proposed	Block-diagonal	Milky-way	Banded	73.3 (13.8)	55.9 (15.0)	68.2 (12.1)	55.6 (11.5)
	Alt.1	Block-diagonal	Milky-way	Banded	58.0 (14.8)	48.2 (15.0)	61.9 (13.3)	40.3 (10.9)
	Alt.2	Block-diagonal	Milky-way	Banded	62.4 (15.4)	52.3 (14.9)	66.6 (13.3)	45.7 (11.8)
	Alt.3	Block-diagonal	Milky-way	Banded	63.7 (16.6)	54.5 (15.9)	68.4 (12.1)	46.4 (10.8)
A2)B4)C2)	Proposed	Containing isolate	Milky-way	Banded	80.0 (6.8)	74.1 (17.6)	46.5 (9.8)	60.5 (14.2)
	Alt.1	Containing isolate	Milky-way	Banded	70.3 (7.8)	57.6 (20.4)	42.0 (10.1)	44.2 (14.2)
	Alt.2	Containing isolate	Milky-way	Banded	74.4 (8.2)	64.4 (22.8)	45.6 (10.3)	52.2 (14.6)
	Alt.3	Containing isolate	Milky-way	Banded	75.9 (8.0)	64.8 (22.8)	46.7 (10.2)	51.7 (14.4)

of its high quality, easy accessibility, and high scientific impact. In particular, we analyze the TCGA data on LUAD (lung adenocarcinoma), a subtype of lung cancer. Data on the gene expressions and copy number variations of 512 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 224 gene expressions and 228 copy number variations analyzed. Preparation has been done to obtain the difference networks of GEs and CNV. First, samples have been dichotomized based on the pathologic tumor stage. Specifically, Stage I, Stage IA, and Stage IB are in one group. The remaining stages are considered as the other group. This dichotomy is biologically sensible. Second, we have conducted marginal screening between Stage and GEs and obtained 57 relevant genes. Also, we have added 23 least relevant genes to the subset genes we used, to mimic the "noise" in a real-world problem. We have considered GEs for $p = 80$ as well as their corresponding CNV. Both the covariance matrices in the unconditional analysis framework and the precision matrices in the conditional analysis framework have been constructed, followed by the difference networks.

Data is analyzed using the proposed and alternative approaches. Tuning parameters are selected using a BIC-type criterion. As in the previous analysis, we focus on results for gene expressions. Genes identified using the proposed and alternative approaches and their estimates for the unconditional and conditional differential network analyses are shown in Tables 4.5 and 4.6 in Appendix, respectively. Summary comparison results are provided in Figure 4.1. As we see, the proposed method and the alternatives have similar results. In particular, for the differential analysis based on the covariance matrix in the unconditional framework, the proposed method identifies 26 change contributors, among which, all genes are identified by all approaches. For the differential analysis based on the precision matrix in the conditional framework, the proposed method identifies 16 change contributors, among which, 13 genes are identified by all approaches, 2 genes are uniquely identified by Alt.1 and Alt.2, and 1 gene is identified by Alt.3 only. For the genes identified by the proposed unconditional differential network analysis, we present the correlation heatmaps for the groups in Figure 4.2 in Appendix. Simply eyeballing the plots suggests significant differences, which can provide support to the proposed analysis.



(a)

(b)

Figure 4.1: Venn diagrams of differential analysis using the proposed method and the alternatives. (a) based on the covariance matrix of GEs. (b) based on the precision matrix of GEs.

It is found that the proposed analysis can identify biologically sensible change contributor genes. For example, genes CHUK, MET, PIK3CA, and ELK1 have been observed in multiple studies. The loss of CHUK mRNA expression in lung adenomas has been confirmed by eRT-PCR analysis of CHUK exons 6 and 7; and two models were established showing that CHUK is a major NSCLC tumor suppressor [180]. Paik et al. have found responses to MET inhibitors in patients with stage IV lung adenocarcinomas harboring MET mutations causing exon 14 skipping [181]. Yamamoto et al. analyzed PIK3CA mutations in exons 9 and 20 in lung cancer cell lines and tumors, and identified PIK3CA mutations among all the major histologic subtypes [182]. Sheng et al. have found ELK1-induced upregulation of HOXA10-AS improved LUAD progression through increasing Wnt/β-catenin signaling [183]. Differential network analysis in the conditional interconnection framework has also identified biologically sensible change contributor genes including VEGFC and RASSF1. Evidence has been provided in the literature that VEGFC/Flt-4-mediated invasion and metastasis of lung cancer cells were found to require upregulation of the neural cell adhesion molecule contactin-1 through activation of the Src-p38 MAPK-C/EBP-dependent pathway [184]. The RASSF1 gene is located in the chromosomal segment of 3p21.3. The high allelic loss in a variety of cancers suggested a crucial role of this region in tumorigenesis. Re-expression of RASSF1A reduced the growth of human cancer cells supporting a role for RASSF1 as a tumor suppressor gene. RASSF1A inactivation and K-ras activation are mutually exclusive events in the development of certain carcinomas [185].

4.5 Discussion

In this chapter, we have somewhat switched gear and conducted network differential analysis. The strategy is consistent with one of those reviewed in Chapter 1. That is, the identified "important" GEs should be connected with the important regulators. As discussed in Chapter 1, this has a strong biological ground and is related to that in [23], that in the LRM study, and others. The "interconnections" in the identified GEs and regulators can significantly facilitate interpretation. Building on the spectral clustering technique, we have developed an approach that has lucid interpretations and a formulation that is methodologically consistent with [23] and other penalized assisted estimations. An effective computational algorithm has been developed. Simulation and data analysis have shown competitive performance of the proposed approach.

As in some other assisted analyses, the proposed approach does not demand the collection of all relevant regulators. However, it is easy to comprehend that, if the collected regulators are not informative, promoting the correlations between important GEs and noises may negatively impact performance. To simplify notation, in methodological development, we have considered two groups. The situation gets complicated when there are multiple groups. Say there are three ordered groups/conditions I, II, and III. One possibility will be to conduct pairwise analysis using the proposed approach. Another possibility is that, considering the order of the three groups, analysis is conducted on group I-II and also group II-III. And then, considering certain similarity between the two sets of analysis is further promoted. This may demand more complex formulation and computation, however, no fundamental change to the proposed strategy. We defer this to future research.

In this chapter, we have focused on methodological and computational development.

Theoretical developments on SSVD have been conducted in the literature. It is conjectured that consistency properties (on estimation and variable selection) can follow from that for SSVD and the proof in [186]. We omit the proof here.

4.6 Appendix

Table 4.3: Simulation results of AUC: mean×100 (sd×100) for $p = 50$, $q = 50$, $n = 200$ (Strong association between X and Y).

Scenario	Approach	X	B	W	$\widehat{\Sigma}_X$	$\widehat{\Sigma}_Y$	$\widehat{\Omega}_X$	$\widehat{\Omega}_Y$
A1)B1)C1)	Proposed	Block-diagonal	Block-diagonal	Independent	75.7 (9.9)	83.5 (27.2)	80.6 (12.6)	62.1 (11.5)
	Alt.1	Block-diagonal	Block-diagonal	Independent	67.4 (10.5)	81.8 (26.8)	67.7 (13.9)	44.7 (11.1)
	Alt.2	Block-diagonal	Block-diagonal	Independent	70.3 (11.1)	84.9 (30.8)	73.2 (13.8)	47.5 (11.8)
	Alt.3	Block-diagonal	Block-diagonal	Independent	70.8 (9.2)	84.3 (29.7)	76.5 (13.3)	48.1 (11.4)
A2)B1)C1)	Proposed	Containing isolate	Block-diagonal	Independent	76.6 (8.9)	66.9 (26.9)	49.4 (8.7)	54.6 (11.0)
	Alt.1	Containing isolate	Block-diagonal	Independent	68.4 (12.3)	55.9 (28.4)	41.9 (8.4)	48.0 (11.8)
	Alt.2	Containing isolate	Block-diagonal	Independent	73.3 (13.3)	58.3 (31.1)	45.4 (9.2)	51.3 (12.4)
	Alt.3	Containing isolate	Block-diagonal	Independent	73.9 (11.1)	57.4 (29.5)	45.2 (9.9)	51.5 (11.3)
A1)B2)C1)	Proposed	Block-diagonal	Milky-way	Independent	86.9 (8.1)	83.3 (23.0)	81.7 (13.7)	67.2 (12.4)
	Alt.1	Block-diagonal	Milky-way	Independent	68.6 (11.2)	78.7 (23.3)	69.1 (15.2)	48.3 (13.3)
	Alt.2	Block-diagonal	Milky-way	Independent	70.5 (12.8)	82.8 (27.4)	73.1 (16.0)	52.1 (13.7)
	Alt.3	Block-diagonal	Milky-way	Independent	71.2 (11.4)	83.1 (25.3)	75.3 (16.3)	52.7 (12.4)
A2)B2)C1)	Proposed	Containing isolate	Milky-way	Independent	82.0 (11.2)	57.1 (24.7)	45.1 (11.1)	53.2 (9.7)
	Alt.1	Containing isolate	Milky-way	Independent	70.3 (12.7)	52.0 (24.3)	39.4 (10.7)	44.9 (9.6)
	Alt.2	Containing isolate	Milky-way	Independent	74.3 (13.2)	55.5 (26.6)	42.4 (11.1)	48.9 (10.7)
	Alt.3	Containing isolate	Milky-way	Independent	73.5 (12.7)	54.4 (24.9)	43.6 (11.2)	46.9 (8.8)
A1)B1)C2)	Proposed	Block-diagonal	Block-diagonal	Banded	74.8 (10.3)	88.1 (21.7)	84.0 (11.4)	61.1 (13.4)
	Alt.1	Block-diagonal	Block-diagonal	Banded	67.0 (11.3)	85.2 (25.1)	72.6 (12.6)	44.9 (11.3)
	Alt.2	Block-diagonal	Block-diagonal	Banded	69.1 (11.6)	88.7 (29.5)	77.1 (13.3)	49.3 (12.3)
	Alt.3	Block-diagonal	Block-diagonal	Banded	70.2 (10.4)	90.5 (22.4)	80.1 (13.4)	49.1 (10.7)
A2)B1)C2)	Proposed	Containing isolate	Block-diagonal	Banded	74.7 (9.7)	67.0 (23.7)	50.1 (8.1)	58.2 (9.5)
	Alt.1	Containing isolate	Block-diagonal	Banded	68.6 (10.9)	55.1 (24.3)	40.6 (7.9)	50.4 (10.2)
	Alt.2	Containing isolate	Block-diagonal	Banded	72.1 (12.6)	54.9 (25.6)	43.7 (8.0)	54.2 (11.1)
	Alt.3	Containing isolate	Block-diagonal	Banded	73.1 (11.1)	58.4 (27.2)	44.8 (7.7)	53.7 (9.7)
A1)B2)C2)	Proposed	Block-diagonal	Milky-way	Banded	86.0 (9.0)	78.4 (31.1)	81.1 (10.1)	65.6 (11.7)
	Alt.1	Block-diagonal	Milky-way	Banded	67.7 (12.4)	70.9 (31.6)	66.9 (12.2)	48.0 (12.4)
	Alt.2	Block-diagonal	Milky-way	Banded	68.9 (13.8)	75.4 (33.9)	71.4 (12.3)	51.5 (11.7)
	Alt.3	Block-diagonal	Milky-way	Banded	70.8 (12.9)	73.2 (30.6)	75.8 (11.8)	52.0 (10.2)
A2)B2)C2)	Proposed	Containing isolate	Milky-way	Banded	83.1 (9.4)	50.7 (23.8)	48.7 (12.2)	56.7 (13.7)
	Alt.1	Containing isolate	Milky-way	Banded	69.9 (12.6)	45.3 (21.8)	43.4 (12.6)	49.7 (13.9)
	Alt.2	Containing isolate	Milky-way	Banded	74.8 (13.0)	47.4 (24.6)	46.8 (12.7)	53.1 (14.7)
	Alt.3	Containing isolate	Milky-way	Banded	74.7 (11.5)	46.8 (24.0)	46.9 (11.3)	52.3 (13.8)

Table 4.4: Simulation results of AUC: mean×100 (sd×100) for $p = 50$, $q = 50$, $n = 200$ (Weak association between X and Y).

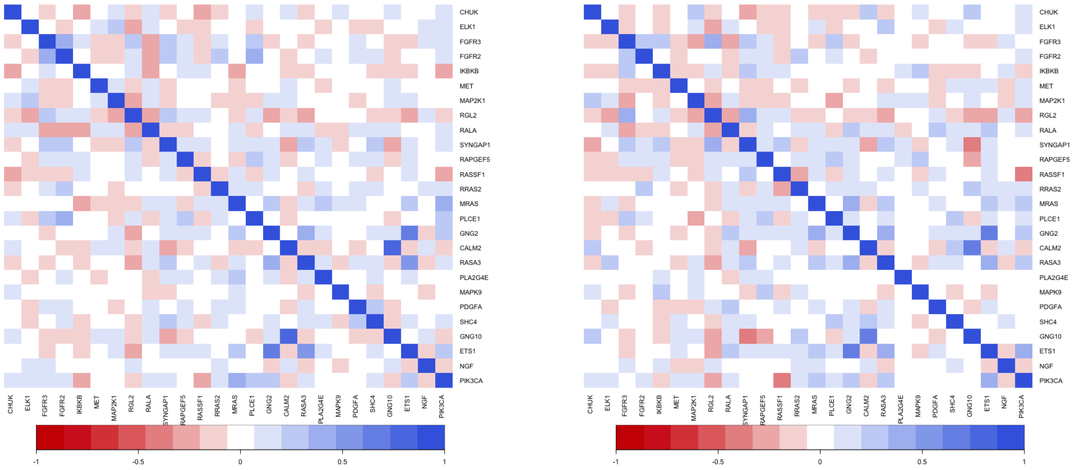
Scenario	Approach	X	B	W	$\widehat{\Sigma}_X$	$\widehat{\Sigma}_Y$	$\widehat{\Omega}_X$	$\widehat{\Omega}_Y$
A1)B3)C1)	Proposed	Block-diagonal	Block-diagonal	Independent	80.3 (7.7)	94.4 (15.5)	77.1 (15.5)	63.7 (10.4)
	Alt.1	Block-diagonal	Block-diagonal	Independent	65.9 (9.9)	82.8 (20.0)	68.4 (15.7)	47.8 (9.0)
	Alt.2	Block-diagonal	Block-diagonal	Independent	69.1 (10.4)	86.2 (22.6)	72.6 (16.6)	51.0 (9.2)
	Alt.3	Block-diagonal	Block-diagonal	Independent	68.1 (10.4)	84.1 (21.8)	75.0 (16.3)	49.6 (8.7)
A2)B3)C1)	Proposed	Containing isolate	Block-diagonal	Independent	76.1 (9.6)	55.9 (12.0)	47.6 (11.8)	50.8 (8.3)
	Alt.1	Containing isolate	Block-diagonal	Independent	69.6 (11.4)	55.0 (21.1)	42.6 (10.6)	43.6 (7.2)
	Alt.2	Containing isolate	Block-diagonal	Independent	73.1 (12.4)	57.6 (21.8)	45.6 (11.2)	46.7 (7.6)
	Alt.3	Containing isolate	Block-diagonal	Independent	74.5 (10.7)	57.7 (22.1)	45.8 (11.7)	46.5 (8.1)
A1)B4)C1)	Proposed	Block-diagonal	Milky-way	Independent	84.5 (10.2)	83 (25)	79.1 (11.6)	66.7 (11.2)
	Alt.1	Block-diagonal	Milky-way	Independent	64.7 (10.8)	73.4 (26.8)	69.5 (12.6)	48.3 (11.8)
	Alt.2	Block-diagonal	Milky-way	Independent	67.2 (11.5)	77.8 (28.3)	73.7 (13.0)	52.3 (12.6)
	Alt.3	Block-diagonal	Milky-way	Independent	69.0 (13.5)	77.7 (27.3)	76.3 (13.0)	51.6 (11.0)
A2)B4)C1)	Proposed	Containing isolate	Milky-way	Independent	81.8 (7.4)	52.4 (15.4)	49.0 (10.4)	55.2 (7.6)
	Alt.1	Containing isolate	Milky-way	Independent	68.2 (13.4)	50.6 (17.7)	43.1 (9.4)	49.0 (7.9)
	Alt.2	Containing isolate	Milky-way	Independent	72.7 (14.3)	52.4 (17.9)	46.4 (9.6)	52.0 (7.6)
	Alt.3	Containing isolate	Milky-way	Independent	73.7 (9.9)	49.0 (15.0)	47.2 (9.3)	50.3 (7.5)
A1)B3)C2)	Proposed	Block-diagonal	Block-diagonal	Banded	84.2 (8.4)	95.4 (13.0)	77.8 (13.2)	61.4 (10.4)
	Alt.1	Block-diagonal	Block-diagonal	Banded	66.5 (10.7)	85.6 (20.0)	66.9 (14.0)	45.3 (10.0)
	Alt.2	Block-diagonal	Block-diagonal	Banded	68.4 (11.4)	90.0 (21.4)	71.7 (15.3)	48.8 (10.4)
	Alt.3	Block-diagonal	Block-diagonal	Banded	67.8 (12.7)	87.8 (21.6)	74.2 (16.0)	49.2 (10.2)
A2)B3)C2)	Proposed	Containing isolate	Block-diagonal	Banded	73.2 (11.9)	63.8 (16.4)	48.9 (10.7)	57.6 (9.6)
	Alt.1	Containing isolate	Block-diagonal	Banded	66.1 (12.6)	58.7 (18.4)	44.5 (9.9)	48.4 (10.4)
	Alt.2	Containing isolate	Block-diagonal	Banded	69.7 (14.0)	59.9 (19.1)	47.7 (10.6)	50.8 (11.1)
	Alt.3	Containing isolate	Block-diagonal	Banded	70.6 (13.1)	62.3 (20.3)	47.4 (11.1)	52.2 (9.4)
A1)B4)C2)	Proposed	Block-diagonal	Milky-way	Banded	85.9 (7.5)	90.1 (20.4)	77.7 (15.6)	64.9 (10.4)
	Alt.1	Block-diagonal	Milky-way	Banded	68.1 (8.9)	78.8 (25.2)	67.5 (15.4)	48.5 (10.9)
	Alt.2	Block-diagonal	Milky-way	Banded	70.2 (9.7)	85.6 (27.1)	71.8 (16.3)	51.7 (11.8)
	Alt.3	Block-diagonal	Milky-way	Banded	70.2 (8.8)	82.3 (25.3)	74.3 (17.2)	52.7 (11.1)
A2)B4)C2)	Proposed	Containing isolate	Milky-way	Banded	79.3 (11.4)	55.3 (15.2)	45.2 (9.3)	54.3 (9.8)
	Alt.1	Containing isolate	Milky-way	Banded	66.9 (11.7)	51.7 (17.1)	40.6 (9.0)	48.4 (9.7)
	Alt.2	Containing isolate	Milky-way	Banded	70.9 (12.8)	53.5 (17.5)	44.0 (9.6)	51.2 (10.1)
	Alt.3	Containing isolate	Milky-way	Banded	71.0 (12.7)	52.5 (17.0)	44.6 (8.9)	50.6 (9.8)

Table 4.5: Unconditional differential network analysis: genes identified using different approaches and estimates.

	Proposed	Alt.1	Alt.2	Alt.3
CHUK	-0.0105	-0.0139	-0.0023	-0.0026
ELK1	0.0124	0.0144	0.0028	0.0030
FGFR3	0.1645	0.0492	0.0382	0.0348
FGFR2	0.1270	0.0388	0.0277	0.0269
IKBKB	0.0245	0.0175	0.0060	0.0055
MET	0.0376	0.0195	0.0080	0.0082
MAP2K1	-0.0061	-0.0136	-0.0020	-0.0017
RGL2	0.0315	0.0195	0.0080	0.0070
RALA	-0.0914	-0.0311	-0.0198	-0.0195
SHC1	0	0	0	0.0002
ZAP70	0	-0.0116	0	0
SYNGAP1	0.0109	0.0149	0.0034	0.0027
RAPGEF5	-0.0016	-0.0118	-0.0001	-0.0007
RASSF1	-0.0052	-0.0134	-0.0018	-0.0015
RRAS2	4.7801	0.9925	0.9977	0.9980
MRAS	0.0591	0.0244	0.0130	0.0127
PLA2G2D	0	-0.0127	-0.0011	0
PLCE1	0.1089	0.0356	0.0244	0.0232
GNG2	-0.0020	-0.0131	-0.0015	-0.0008
CALM2	-0.0125	-0.0163	-0.0047	-0.0031
RASA3	-0.0097	-0.0149	-0.0033	-0.0024
PLA2G4E	0.0362	0.0193	0.0078	0.0080
MAPK9	-0.0277	-0.0182	-0.0067	-0.0062
PDGFA	0.1216	0.0376	0.0264	0.0258
SHC4	0.0280	0.0177	0.0061	0.0062
GNG10	-0.0267	-0.0188	-0.0073	-0.0060
ETS1	-0.0054	-0.0139	-0.0023	-0.0015
NGF	0.0051	0.0130	0.0014	0.0015
PIK3CA	0.0089	0.0147	0.0031	0.0023

Table 4.6: Conditional differential network analysis: genes identified using different approaches and estimates.

	Proposed	Alt.1	Alt.2	Alt.3
GNB3	-0.2295	-0.1049	-0.0481	-0.0510
GNG7	0	-0.0801	-0.0030	0
PIK3CD	-0.0830	-0.1088	-0.0426	-0.0225
PLCG2	0	0.0777	0.0081	0
PRKCB	0.1364	0.1139	0.0565	0.0437
MAPK10	0.0899	0.1154	0.0503	0.0160
VEGFC	0.3392	0.1163	0.0545	0.0603
ZAP70	2.8513	0.6122	0.6990	0.7002
SYNGAP1	-0.0313	0	0	-0.0017
RASGRP2	0.0681	0.1220	0.0592	0.0145
RRAS2	-0.0241	-0.0787	-0.0110	0
PLA2G2D	-0.3057	-0.1336	-0.0836	-0.0875
PAK7	-1.6813	-0.3232	-0.3174	-0.3874
RASAL3	-2.3769	-0.5322	-0.6173	-0.5813
RASSF5	0.0162	0.0729	0.0073	0
RASA3	-0.2541	-0.1370	-0.0738	-0.0685
FGF22	-0.0628	-0.0812	-0.0137	-0.0096
REL	0	-0.0724	0	0
NGF	-0.1395	-0.0872	-0.0202	-0.0129



(a) **(b)**

Figure 4.2: Heatmaps of correlation for the genes identified by the proposed assisted differential network analysis. (a) Group 1. (b) Group 2.

Chapter 5

Conclusion

In this dissertation, we have first conducted comprehensive literature review. This effort has helped us better understand the ground for our methodological developments. Equally importantly, with the publication in Briefings in Bioinformatics, it may also informative for researchers interested in gene expression-centric vertical data integration. In Chapters 2-4, we have conducted three somewhat "independent" methodological developments. This "independence" can be partly seen from our separate and parallel publications. On the other hand, the three methods also have strong interconnections. Specifically, they have addressed assisting gene expression network analysis using complementary information, which may suggest the possibility of "integrating" such methods into a "mega" one and more effectively and comprehensively use additional information. In addition, they have all been built on the effective penalization technique. Penalization has been the favorable choice in GGM and other network analysis and has demonstrated superior statistical and numerical performance. On the other hand, it is recognized that there are many other regularized estimation and variable selection techniques, including thresholding, boosting, Bayesian, and others. It is conjectured that the proposed analysis strategies can be coupled with these techniques. Numerically, new computational algorithms will need to be developed, and new simulation and data analysis will need to be conducted and evaluated. Theoretically, they may pose more challenges. Our limited literature review suggests that, with the fast and extensive developments in the past two decades, the techniques for establishing estimation and variable selection properties with penalization methods are relatively mature – however, this is not true with other regularization methods. In our data analyses, we have focused on the TCGA data. A quick examination suggests that there is no hurdle applying the proposed methods to other data sources. The advantages of TCGA data (and hence our reasoning for choosing such data) have been discussed in this dissertation and extensively in the literature.

This dissertation has opened the door for much more extensive developments. Methodologically, as mentioned above, it is of interest to couple the proposed assisted strategies with other regularization methods. Also, as mentioned in Chapter 1, the proposed analyses, loosely speaking, belong to the vertical data integration paradigm. For improving network analysis, horizontal data integration addresses from a different perspective, and has also been highly successful. A more comprehensive (and practical) scenario includes multiple independent gene expression datasets, and within each dataset, regulator data and/or prior information are present. The analysis of such data will demand effectively combining the proposed methods with the horizontal integration ones. Simply quickly thinking of this analysis can already suggest significant challenges. In particular, different datasets may measure different types/sets of regulators (for example, one dataset has methylation measurements, while another dataset only has microRNA measurements). In this case, the methods developed in Chapters 2 and 4 will need to be significantly revised and advanced. In Chapter 3, the information extracted using PubMatrix has been “rough”. It is of interest to rerun analysis once more refined text mining is conducted. In some published studies, especially when the field/topic is narrow, manual information curation has been done. Such information can still be partial, but less likely to be wrong. Our approach can still be applied, however, with the new characteristics of information, it may not be optimal. With such prior information, a new method may be demanded. A closer examination of information suggests that different gene pairs differ not only in the amount of information (number of publications) but also the level of certainty. That is, the interconnections between some gene pairs have been repeatedly established using not only analytic but also functional approaches. In comparison, for some other gene pairs, there have been a large number of studies (and hence a high amount of information), but the conclusions remain not definitive, that is, the level of certainty is limited. Built on the proposed approach, new methodological development will be needed to accommodate the above and other more complex scenarios.

Under all three projects, we have conducted careful data analysis and comparison. As discussed in the three chapters, we have a reasonable level of confidence in our data analysis results. However, as in many other biostatistical studies, such findings are not meant to be final/decisive. The unconditional interconnections among genes can be established with high confidence using functional experiments, although we do note that with a huge number of gene pairs, this will be a long process. However, to the best of our knowledge, although the conditional dependence among genes is statistically clearly and well defined, it is unclear how that can be verified in functional studies. With the extensive network and other conditional dependence analysis, we see a strong need for designing and conducting such functional validation. However, this is far beyond this dissertation.

Overall, this dissertation has significantly advanced network analysis for gene expression data and the assisted analysis strategy. The proposed methods can enjoy broad applicability, and their routine applications will be significantly facilitated with the development and publication of software. Our methodological developments can also enrich the family of penalized techniques, and our theoretical developments can provide further insight into high dimensional estimation theories. Our data analysis results have provided additional insights for the biology of multiple important cancers. Also, as partly described above, this dissertation has paved the road for extensive future developments.

Bibliography

[1] S. Richardson, G. C. Tseng, and W. Sun. Statistical methods in integrative genomics. *Annual review of statistics and its application*, 3:181–209, 2016.

[2] Q. Zhao, X. Shi, J. Huang, J. Liu, Y. Li, and S. Ma. Integrative analysis of '-omics' data using penalty functions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):99–108, 2015.

[3] Y. Huang, Q. Zhang, S. Zhang, J. Huang, and S. Ma. Promoting similarity of sparsity structures in integrative analysis with penalization. *Journal of the American Statistical Association*, 112(517):342–350, 2017.

[4] K. Fang, X. Fan, Q. Zhang, and S. Ma. Integrative sparse principal component analysis. *Journal of Multivariate Analysis*, 166:1–16, 2018.

[5] X. Fan, K. Fang, S. Ma, and Q. Zhang. Integrating approximate single factor graphical models. *Statistics in Medicine*, 39(2):146–155, 2020.

[6] Q. Zhao, X. Shi, Y. Xie, J. Huang, B. Shia, and S. Ma. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from tcga. *Briefings in bioinformatics*, 16(2):291–303, 2015.

[7] K. J. Karczewski and M. P. Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299, 2018.

[8] D. Lin, J. Zhang, J. Li, H. He, H.-W. Deng, and Y.-P. Wang. Integrative analysis of multiple diverse omics datasets by sparse group multitask regression. *Frontiers in cell and developmental biology*, 2:62, 2014.

[9] I. Mihaylov, M. Kańduła, M. Krachunov, and D. Vassilev. A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. *Biology direct*, 14(1):22, 2019.

[10] J. Y. Park and E. F. Lock. Integrative factorization of bidimensionally linked matrices. *Biometrics*, 76(1):61–74, 2020.

[11] G. Michailidis. Statistical challenges in biological networks. *Journal of Computational and Graphical Statistics*, 21(4):840–855, 2012.

[12] C. B. Peterson, F. C. Stingo, and M. Vannucci. Joint bayesian variable and graph selection for regression models with network-structured predictors. *Statistics in medicine*, 35(7):1017–1031, 2016.

[13] B. Gao, X. Liu, H. Li, and Y. Cui. Integrative analysis of genetical genomics data incorporating network structures. *Biometrics*, 75(4):1063–1075, 2019.

[14] X. Wang, Y. Xu, and S. Ma. Identifying gene-environment interactions incorporating prior information. *Statistics in medicine*, 38(9):1620–1633, 2019.

[15] X. Shi, Q. Zhao, J. Huang, Y. Xie, and S. Ma. Deciphering the associations between gene expression and copy number alteration using a sparse double laplacian shrinkage approach. *Bioinformatics*, 31(24):3977–3983, 2015.

[16] C. Wu, Q. Zhang, Y. Jiang, and S. Ma. Robust network-based analysis of the associations between (epi) genetic measurements. *Journal of multivariate analysis*, 168:119–130, 2018.

[17] L. Cantini, C. Isella, C. Petti, G. Picco, S. Chiola, E. Ficarra, M. Caselle, and E. Medico. Microrna–mrna interactions underlying colorectal cancer molecular subtypes. *Nature communications*, 6(1):1–12, 2015.

[18] Y. Wang, J. M. Franks, M. L. Whitfield, and C. Cheng. Biomethyl: an r package for biological interpretation of dna methylation data. *Bioinformatics*, 35(19):3635–3641, 2019.

[19] X. Shi, H. Yi, and S. Ma. Measures for the degree of overlap of gene signatures and applications to tcga. *Briefings in bioinformatics*, 16(5):735–744, 2015.

[20] S. Ma and J. Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9(5):392–403, 2008.

[21] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5):971–989, 2015.

[22] R. Zhu, Q. Zhao, H. Zhao, and S. Ma. Integrating multidimensional omics data for cancer outcome. *Biostatistics*, 17(4):605–618, 2016.

[23] Y. Li, R. Bie, S. J. Teran Hidalgo, Y. Qin, M. Wu, and S. Ma. Assisted gene expression-based clustering with awncut. *Statistics in medicine*, 37(29):4386–4403, 2018.

[24] A. Serra, M. Fratello, V. Fortino, G. Raiconi, R. Tagliaferri, and D. Greco. Mvda: a multi-view genomic data integration methodology. *BMC bioinformatics*, 16(1):261, 2015.

[25] E. F. Lock and D. B. Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616, 2013.

[26] E. Gabasova, J. Reid, and L. Wernisch. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLoS computational biology*, 13(10):e1005781, 2017.

[27] D. M. Swanson, T. Lien, H. Bergholtz, T. Sørlie, and A. Frigessi. A bayesian two-way latent structure model for genomic data integration reveals few pan-genomic cluster subtypes in a breast cancer cohort. *Bioinformatics*, 35(23):4886–4897, 2019.

[28] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333, 2014.

[29] T. Nguyen, R. Tagett, D. Diaz, and S. Draghici. A novel approach for data integration and disease subtyping. *Genome research*, 27(12):2025–2039, 2017.

[30] N. Rappoport and R. Shamir. Nemo: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 35(18):3348–3356, 2019.

[31] A. Khan and P. Maji. Approximate graph laplacians for multimodal data clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[32] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

[33] D. Wu, D. Wang, M. Q. Zhang, and J. Gu. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC genomics*, 16(1):1022, 2015.

[34] C. Meng, D. Helm, M. Frejno, and B. Kuster. mocluster: identifying joint patterns across multiple omics data sets. *Journal of Proteome Research*, 15(3):755–765, 2016.

[35] S. Kim, S. Oesterreich, S. Kim, Y. Park, and G. C. Tseng. Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. *Biostatistics*, 18(1):165–179, 2017.

[36] Q. Mo, R. Shen, C. Guo, M. Vannucci, K. S. Chan, and S. G. Hilsenbeck. A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, 19(1):71–86, 2018.

[37] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124, 2018.

[38] S. Kim, J. D. Herazo-Maya, D. D. Kang, B. M. Juan-Guardela, J. Tedrow, F. J. Martinez, F. C. Sciurba, G. C. Tseng, and N. Kaminski. Integrative phenotyping

framework (ipf): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC genomics*, 16(1):924, 2015.

[39] Z. Huo and G. Tseng. Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. *The annals of applied statistics*, 11(2):1011, 2017.

[40] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6):1248–1259, 2018.

[41] S. J. T. Hidalgo, M. Wu, and S. Ma. Assisted clustering of gene expression data using ancut. *BMC genomics*, 18(1):623, 2017.

[42] D. Dembele and P. Kastner. Fuzzy c-means method for clustering microarray data. *bioinformatics*, 19(8):973–980, 2003.

[43] I. A. Maraziotis. A semi-supervised fuzzy clustering algorithm applied to gene expression data. *Pattern Recognition*, 45(1):637–648, 2012.

[44] S. J. Teran Hidalgo, T. Zhu, M. Wu, and S. Ma. Overlapping clustering of gene expression data using penalized weighted normalized cut. *Genetic epidemiology*, 42(8):796–811, 2018.

[45] L.-C. Chen, P. S. Yu, and V. S. Tseng. Wf-msb: A weighted fuzzy-based biclustering method for gene expression data. *International journal of data mining and bioinformatics*, 5(1):89–109, 2011.

[46] S. M. Gross and R. Tibshirani. Collaborative regression. *Biostatistics*, 16(2):326–338, 2015.

[47] H. Chai, X. Shi, Q. Zhang, Q. Zhao, Y. Huang, and S. Ma. Analysis of cancer gene expression data with an assisted robust marker identification approach. *Genetic epidemiology*, 41(8):779–789, 2017.

[48] C. Luo, J. Liu, D. K. Dey, and K. Chen. Canonical variate regression. *Biostatistics*, 17(3):468–483, 2016.

[49] G. McLachlan and D. Peel. Finite mixture models.,(john wiley & sons: New york.). *Wiley Series in Probability and Statistics*, 2000.

[50] M. Liu, Q. Zhang, K. Fang, and S. Ma. Structured analysis of the high-dimensional fmr model. *Computational Statistics & Data Analysis*, 144:106883, 2020.

[51] D. J. Hunter. Gene–environment interactions in human diseases. *Nature Reviews Genetics*, 6(4):287–298, 2005.

[52] M. Wu and S. Ma. Robust genetic interaction analysis. *Briefings in Bioinformatics*, 20(2):624–637, 2018.

[53] A.-L. Boulesteix, R. D. Bin, X. Jiang, and M. Fuchs. IPF-LASSO: Integrative l1-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine*, 2017:1–14, 2017.

[54] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine, and C. Sander. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS ONE*, 6(11):e24709, 2011.

[55] Y. Jiang, X. Shi, Q. Zhao, M. Krauthammer, B. E. G. Rothberg, and S. Ma. Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. *Genomics*, 107(6):223–230, 2016.

[56] J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood, and H. Dawood. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. *BMC Bioinformatics*, 20(1), 2019.

[57] T. Ma and A. Zhang. Multi-view factorization AutoEncoder with network constraints for multi-omic integrative analysis. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018.

[58] Z. Huang, X. Zhan, S. Xiang, T. S. Johnson, B. Helm, C. Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han, and K. Huang. SALMON: Survival analysis learning with multi-omics neural networks on breast cancer. *Frontiers in Genetics*, 10, 2019.

[59] D. Sun, M. Wang, and A. Li. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(3):841–850, 2019.

[60] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[61] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, 2013.

[62] E. J. Min, S. E. Safo, and Q. Long. Penalized co-inertia analysis with applications to -omics data. *Bioinformatics*, 35(6):1018–1025, 2018.

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

[64] W. Wang, V. Baladandayuthapani, J. S. Morris, B. M. Broom, G. Manyam, and K.-A. Do. ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159, 2013.

[65] C. Wu, J. Zhu, and X. Zhang. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC bioinformatics*, 13(1):182, 2012.

[66] M. H. Kabir, R. Patrick, J. W. Ho, and M. D. O'Connor. Identification of active signaling pathways by integrating gene expression and protein interaction data. *BMC systems biology*, 12(9):77–87, 2018.

[67] S. J. T. Hidalgo and S. Ma. Clustering multilayer omics data using muncut. *BMC genomics*, 19(1):198, 2018.

[68] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):1–17, 2018.

[69] I. C. Macaulay, W. Haerty, P. Kumar, Y. I. Li, T. X. Hu, M. J. Teng, M. Goolam, N. Saurat, P. Coupland, L. M. Shirley, et al. G&t-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods*, 12(6):519–522, 2015.

[70] C. Angermueller, S. J. Clark, H. J. Lee, I. C. Macaulay, M. J. Teng, T. X. Hu, F. Krueger, S. A. Smallwood, C. P. Ponting, T. Voet, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods*, 13(3):229–232, 2016.

[71] J. Cao, D. A. Cusanovich, V. Ramani, D. Aghamirzaie, H. A. Pliner, A. J. Hill, R. M. Daza, J. L. McFaline-Figueroa, J. S. Packer, L. Christiansen, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, 2018.

[72] T. Stuart and R. Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, 2019.

[73] R. Petegrosso, Z. Li, and R. Kuang. Machine learning and statistical methods for clustering single-cell rna-sequencing data. *Briefings in bioinformatics*, 21(4):1209–1223, 2020.

[74] Z. Duren, X. Chen, M. Zamanighomi, W. Zeng, A. T. Satpathy, H. Y. Chang, Y. Wang, and W. H. Wong. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences*, 115(30):7723–7728, 2018.

[75] R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, and O. Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):1–17, 2020.

[76] H.-C. Yang, P.-L. Wang, C.-W. Lin, C.-H. Chen, and C.-H. Chen. Integrative analysis of single nucleotide polymorphisms and gene expression efficiently distinguishes samples from closely related ethnic populations. *BMC genomics*, 13(1):346, 2012.

[77] C. Lengauer, K. W. Kinzler, and B. Vogelstein. Genetic instabilities in human cancers. *Nature*, 396(6712):643, 1998.

[78] W. M. Linehan, M. M. Walther, and B. Zbar. The genetic basis of cancer of the kidney. *The Journal of urology*, 170(6):2163–2172, 2003.

[79] J. E. Dancey, P. L. Bedard, N. Onetto, and T. J. Hudson. The genetic basis for cancer treatment decisions. *Cell*, 148(3):409–420, 2012.

[80] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(1):S7, 2006.

[81] P. E. Meyer, F. Lafitte, and G. Bontempi. minet: Ar/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*, 9(1):461, 2008.

[82] F. Emmert-Streib, G. Glazko, R. De Matos Simoes, et al. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Frontiers in genetics*, 3:8, 2012.

[83] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97(22):12182–12186, 2000.

[84] P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[85] C. Clarke, S. F. Madden, P. Doolan, S. T. Aherne, H. Joyce, L. O'driscoll, W. M. Gallagher, B. T. Hennessy, M. Moriarty, J. Crown, et al. Correlating transcriptional

networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis*, 34(10):2300–2308, 2013.

[86] N. Wisniewski, M. Cadeiras, G. Bondar, R. Cheng, K. Shahzad, D. Onat, F. Latif, Y. Korin, E. Reed, R. Fakhro, et al. Weighted gene coexpression network analysis (wgcna) modeling of multiorgan dysfunction syndrome after mechanical circulatory support therapy. *The Journal of Heart and Lung Transplantation*, 32(4):S223, 2013.

[87] H. Tian, D. Guan, and J. Li. Identifying osteosarcoma metastasis associated genes by weighted gene co-expression network analysis (wgcna). *Medicine*, 97(24), 2018.

[88] Q. Yang, R. Wang, B. Wei, C. Peng, L. Wang, G. Hu, D. Kong, and C. Du. Candidate biomarkers and molecular mechanism investigation for glioblastoma multiforme utilizing wgcna. *BioMed research international*, 2018, 2018.

[89] R. Ben-Hamo and S. Efroni. Gene expression and network-based analysis reveals a novel role for hsa-mir-9 and drug control over the p38 network in glioblastoma multiforme progression. *Genome medicine*, 3(11):77, 2011.

[90] M. Dehmer, L. A. Mueller, and F. Emmert-Streib. Quantitative network measures as biomarkers for classifying prostate cancer disease states: a systems approach to diagnostic biomarkers. *PloS one*, 8(11):e77602, 2013.

[91] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[92] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature communications*, 5:3231, 2014.

[93] I. S. Chan and G. S. Ginsburg. Personalized medicine: progress and promise. *Annual review of genomics and human genetics*, 12:217–244, 2011.

[94] T. Ideker and N. J. Krogan. Differential network biology. *Molecular systems biology*, 8(1):565, 2012.

[95] M. F. Islam, M. M. Hoque, R. S. Banik, S. Roy, S. S. Sumi, F. N. Hassan, M. T. S. Tomal, A. Ullah, and K. T. Rahman. Comparative analysis of differential network modularity in tissue specific normal and cancer protein interaction networks. *Journal of clinical bioinformatics*, 3(1):19, 2013.

[96] F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in cell and developmental biology*, 2:38, 2014.

[97] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, NJ, 1996.

[98] Y. V. Sun. Integration of biological networks and pathways with genetic association studies. *Human genetics*, 131(10):1677–1686, 2012.

[99] Y. Rahmatallah, F. Emmert-Streib, and G. Glazko. Gene sets net correlations analysis (gsnca): a multivariate differential coexpression test for gene sets. *Bioinformatics*, 30(3):360–368, 2013.

[100] A. T. McKenzie, I. Katsyv, W.-M. Song, M. Wang, and B. Zhang. Dgca: a comprehensive r package for differential gene correlation analysis. *BMC systems biology*, 10(1):106, 2016.

[101] S. L. Carter, C. M. Brechbühler, M. Griffin, and A. T. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, 2004.

[102] R. De Smet and K. Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10):717, 2010.

[103] T. Grimes, S. S. Potter, and S. Datta. Integrating gene regulatory pathways into differential network analysis of gene expression data. *Scientific reports*, 9(1):5479, 2019.

[104] K. Baba, R. Shibata, and M. Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

[105] M. Dehmer and F. Emmert-Streib. *Analysis of complex networks: from biology to linguistics*. John Wiley & Sons, 2009.

[106] T. V. Tatarinova and Y. Nikolsky. *Biological Networks and Pathway Analysis*. Springer, 2017.

[107] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[108] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[109] R.-S. Wang, A. Saadatpour, and R. Albert. Boolean modeling in systems biology: an overview of methodology and applications. *Physical biology*, 9(5):055001, 2012.

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

[111] A. Wille, P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, et al. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome biology*, 5(11):R92, 2004.

[112] S. Ma, Q. Gong, and H. J. Bohnert. An arabidopsis gene network based on the graphical gaussian model. *Genome research*, 17(11):1614–1625, 2007.

[113] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254–4278, 2009.

[114] J. Yin and H. Li. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied statistics*, 5(4):2630, 2011.

[115] X.-T. Yuan and T. Zhang. Partial gaussian graphical model estimation. *IEEE Transactions on Information Theory*, 60(3):1673–1687, 2014.

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

[117] T. T. Cai, H. Li, W. Liu, and J. Xie. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1):139–156, 2013.

[118] H. Chun, M. Chen, B. Li, and H. Zhao. Joint conditional gaussian graphical models with multiple sources of genomic data. *Frontiers in genetics*, 4:294, 2013.

[119] J. Chiquet, T. Mary-Huard, and S. Robin. Structured regularization for conditional gaussian graphical models. *Statistics and Computing*, 27(3):789–804, 2017.

[120] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor, and G. Condous. Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Human reproduction*, 21(7):1824–1831, 2006.

[121] C. Peterson, F. C. Stingo, and M. Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

[122] A. Mohammadi, E. C. Wit, et al. Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.

[123] Y. Jiang, Y. He, and H. Zhang. Variable selection with prior information for generalized linear models via the prior lasso method. *Journal of the American Statistical Association*, 111(513):355–376, 2016.

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

[125] T. Wang, Z. Ren, Y. Ding, Z. Fang, Z. Sun, M. L. MacDonald, R. A. Sweet, J. Wang, and W. Chen. Fastggm: an efficient algorithm for the inference of gaussian graphical model in biological networks. *PLoS computational biology*, 12(2):e1004755, 2016.

[126] H. Zhao and Z.-H. Duan. Cancer genetic network inference using gaussian graphical models. *Bioinformatics and biology insights*, 13:1177932219839402, 2019.

[127] H. Liu, F. Han, and C.-h. Zhang. Transelliptical graphical models. *Advances in neural information processing systems*, pp. 800–808, 2012.

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

[129] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

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

[131] D. R. Williams. Bayesian estimation for gaussian graphical models: Structure learning, predictability, and network comparisons. *PsyArXiv*, 2018.

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

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

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

[135] E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, 2005.

[136] A. Yazdani, R. Mendez-Giraldez, A. Yazdani, M. R. Kosorok, and P. Roussos. Differential gene regulatory pattern in the human brain from schizophrenia using transcriptomic-causal network. *BMC bioinformatics*, 21(1):1–19, 2020.

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

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

[139] J. Huang, P. Breheny, and S. Ma. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4), 2012.

[140] S. van Dam, U. Vosa, A. van der Graaf, L. Franke, and J. P. de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592, 2018.

[141] J. Fan, Y. Liao, and H. Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 2016.

[142] M. Drton and M. H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

[143] Z. Ren, T. Sun, C.-H. Zhang, H. H. Zhou, et al. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

[144] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, 5(1):21, 2011.

[145] R. Yamaguchi, M. Tanaka, H. Otsuka, M. Yamaguchi, Y. Kaneko, T. Fukushima, H. Terasaki, S. Isobe, O. Nakashima, and H. Yano. Neuroendocrine small cell carcinoma of the breast: report of a case. *Medical molecular morphology*, 42(1):58–61, 2009.

[146] J. Hur, A. Özgür, Z. Xiang, and Y. He. Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining. *Journal of biomedical semantics*, 3(1):18, 2012.

[147] S. Zhao, C. Su, Z. Lu, and F. Wang. Recent advances in biomedical literature mining. *Briefings in Bioinformatics*, 2020.

[148] W. N. van Wieringen. The generalized ridge estimator of the inverse covariance matrix. *Journal of Computational and Graphical Statistics*, 28(4):932–942, 2019.

[149] W. N. van Wieringen, K. A. Stam, C. F. Peeters, and M. A. van de Wiel. Updating of the gaussian graphical model through targeted penalized estimation. *Journal of Multivariate Analysis*, 178:104621, 2020.

[150] A. Deeter, M. Dalman, J. Haddad, and Z.-H. Duan. Inferring gene and protein interactions using pubmed citations and consensus bayesian networks. *PloS one*, 12(10):e0186004, 2017.

[151] M. C. de Souza and C. H. Higa. Reverse engineering of gene regulatory networks combining dynamic bayesian networks and prior biological knowledge. In *International Conference on Computational Science and Its Applications*, volume pp. 323–336. Springer, 2018.

[152] K. G. Becker, D. A. Hosack, G. Dennis, R. A. Lempicki, T. J. Bright, C. Cheadle, and J. Engel. Pubmatrix: a tool for multiplex literature mining. *BMC bioinformatics*, 4(1):61, 2003.

[153] C. Zhang and M. Q. Zhang. Biomedical literature mining. In *Bioinformatics: A Concept-Based Introduction*, volume pp. 115–127. Springer, 2009.

[154] K. Mohan, M. Chung, S. Han, D. Witten, S.-I. Lee, and M. Fazel. Structured learning of gaussian graphical models. In *Advances in neural information processing systems*, volume pp. 620–628, 2012.

[155] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[156] O. Banerjee, L. E. Ghaoui, A. d'Aspremont, and G. Natsoulis. Convex optimization techniques for fitting sparse gaussian graphical models. In *Proceedings of the 23rd international conference on Machine learning*, volume pp. 89–96, 2006.

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

[158] K. Tomczak, P. Czerwińska, and M. Wiznerowicz. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A):A68, 2015.

[159] H. Liu, F. Han, M. Yuan, J. Lafferty, L. Wasserman, et al. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

[160] D. J. Raz, M. R. Ray, J. Y. Kim, B. He, M. Taron, M. Skrzypski, M. Segal, D. R. Gandara, R. Rosell, and D. M. Jablons. A multigene assay is prognostic of survival in patients with early-stage lung adenocarcinoma. *Clinical Cancer Research*, 14(17):5565–5570, 2008.

[161] C. D. van Borkulo, L. Boschloo, J. Kossakowski, P. Tio, R. A. Schoevers, D. Borsboom, and L. J. Waldorp. Comparing network structures on three aspects: A permutation test. *Manuscript submitted for publication*, 10, 2017.

[162] A. J. Rothman, P. J. Bickel, E. Levina, J. Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[163] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[164] S. D. Zhao, T. T. Cai, and H. Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.

[165] X.-F. Zhang, L. Ou-Yang, X.-M. Zhao, and H. Yan. Differential network analysis from cross-platform gene expression data. *Scientific reports*, 6(1):1–12, 2016.

[166] H. Yuan, R. Xi, C. Chen, and M. Deng. Differential network analysis via lasso penalized d-trace loss. *Biometrika*, 104(4):755–770, 2017.

[167] S. Zhao, S. Ottinger, S. Peck, C. Mac Donald, and A. Shojaie. Network differential connectivity analysis. *arXiv preprint arXiv:1909.13464*, 2019.

[168] D. Ruan, A. Young, and G. Montana. Differential analysis of biological networks. *BMC bioinformatics*, 16(1):327, 2015.

[169] F. Liu, D. Choi, L. Xie, and K. Roeder. Global spectral clustering in dynamic networks. *Proceedings of the National Academy of Sciences*, 115(5):927–932, 2018.

[170] V. Nikiforov. Beyond graph energy: Norms of graphs and matrices. *Linear Algebra and its Applications*, 506:82–138, 2016.

[171] X. Cui and G. A. Churchill. Statistical tests for differential expression in cdna microarray experiments. *Genome biology*, 4(4):1–10, 2003.

[172] W. T. Barry, A. B. Nobel, and F. A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 2005.

[173] S. B. Cho, J. Kim, and J. H. Kim. Identifying set-wise differential co-expression in gene expression microarray data. *BMC bioinformatics*, 10(1):1–13, 2009.

[174] V. C. Jardim, C. C. Moreno, and A. Fujita. Computational tools for comparing gene coexpression networks. In *Networks in Systems Biology*, volume pp. 19–30. Springer, 2020.

[175] W. Yu, S. Zhao, Y. Wang, B. N. Zhao, W. Zhao, and X. Zhou. Identification of cancer prognosis-associated functional modules using differential co-expression networks. *Oncotarget*, 8(68):112928, 2017.

[176] M. Lee, H. Shen, J. Z. Huang, and J. Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095, 2010.

[177] J. Liu, C. Wang, M. Danilevsky, and J. Han. Large-scale spectral clustering on graphs. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer, 2013.

[178] J. Baglama and L. Reichel. Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing*, 27(1):19–42, 2005.

[179] J. Baglama, L. Reichel, B. Lewis, and M. B. Lewis. Package 'irlba'. 2019.

[180] E. Chavdoula, D. M. Habiel, E. Roupakia, G. S. Markopoulos, E. Vasilaki, A. Kokkalis, A. P. Polyzos, H. Boleti, D. Thanos, A. Klinakis, et al. Chuk/ikk-α loss in lung epithelial cells enhances nsclc growth associated with hif up-regulation. *Life science alliance*, 2(6), 2019.

[181] P. K. Paik, A. Drilon, P.-D. Fan, H. Yu, N. Rekhtman, M. S. Ginsberg, L. Borsu, N. Schultz, M. F. Berger, C. M. Rudin, et al. Response to met inhibitors in patients with stage iv lung adenocarcinomas harboring met mutations causing exon 14 skipping. *Cancer discovery*, 5(8):842–849, 2015.

[182] H. Yamamoto, H. Shigematsu, M. Nomura, W. W. Lockwood, M. Sato, N. Okumura, J. Soh, M. Suzuki, I. I. Wistuba, K. M. Fong, et al. Pik3ca mutations and copy number gains in human lung cancers. *Cancer research*, 68(17):6913–6921, 2008.

[183] K. Sheng, J. Lu, and H. Zhao. Elk1-induced upregulation of lncrna hoxa10-as promotes lung adenocarcinoma progression by increasing wnt/β-catenin signaling. *Biochemical and biophysical research communications*, 501(3):612–618, 2018.

[184] J.-L. Su, P.-C. Yang, J.-Y. Shih, C.-Y. Yang, L.-H. Wei, C.-Y. Hsieh, C.-H. Chou, Y.-M. Jeng, M.-Y. Wang, K.-J. Chang, et al. The vegf-c/flt-4 axis promotes invasion and metastasis of cancer cells. *Cancer cell*, 9(3):209–223, 2006.

[185] R. Dammann, U. Schagdarsurengin, M. Strunnikova, M. Rastetter, C. Seidel, L. Liu, S. Tommasi, and G. Pfeifer. Epigenetic inactivation of the ras-association domain family 1 (rassf1a) gene and its function in human carcinogenesis. *Histology and histopathology*, 2003.

[186] M. Wu, Q. Zhang, and S. Ma. Structured gene-environment interaction analysis. *Biometrics*, 76(1):23–35, 2020.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA