High-Dimensional Inference of Modified Poisson-Type Graphical Models and Robust Sparse CCA, with Applications to Large-Scale Omics Data

by

Rong Zhang

B.Sc. (Hons.) in Statistics, Hong Kong Baptist University, 2013

M.A. in Applied Statistics, University of Pittsburgh, 2015

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Rong Zhang

It was defended on

June 30th, 2020

and approved by

Zhao Ren, Ph.D., Department of Statistics, University of Pittsburgh

Satish Iyengar, Ph.D., Department of Statistics, University of Pittsburgh

Kehui Chen, Ph.D., Department of Statistics, University of Pittsburgh

Wei Chen, Ph.D., Department of Pediatrics, University of Pittsburgh, UPMC Children's Hospital of Pittsburgh

Dissertation Director: Zhao Ren, Ph.D., Department of Statistics, University of Pittsburgh

High-Dimensional Inference of Modified Poisson-Type Graphical Models and Robust Sparse CCA, with Applications to Large-Scale Omics Data

Rong Zhang, PhD

University of Pittsburgh, 2020

Recent advances in high-throughput sequencing have generated different types of high-dimensional omics data. Even though remarkable progress has been made in statistical inference of high-dimensional Gaussian graphical model (GGM) for gene co-expression network analysis and sparse canonical correlation analysis (CCA) for multi-omics study, efficient computation is always a big concern, and methods beyond Gaussian assumption are even largely unknown. To address both computational and methodological challenges, this dissertation covers efficient implementations of statistical inference of high-dimensional GGM (the first part) and novel statistical methods for count-valued RNA-seq data in gene co-expression network analysis (the second part) and heavy-tailed CITE-seq data in multi-omics study (the third part).

In the first part of the dissertation, we develop an extensive and efficient `R` package named `SILGGM` (Statistical Inference of Large-scale Gaussian Graphical Model) that includes four main approaches in statistical inference of high-dimensional GGM. Extensive comparisons illustrate that `SILGGM` can accelerate existing implementations from several to dozens of orders of magnitudes without loss of accuracy. The package is freely available via `CRAN` at `https://cran.r-project.org/package=SILGGM`.

In the second part of the dissertation, we propose a novel two-step procedure in both edge-wise and global statistical inference of three modified Poisson-type graphical models using a cutting-edge generalized low-dimensional projection approach for bias correction. An extensive simulation study illustrates asymptotic normality of edge-wise inference and more accurate inferential results in multiple testing compared to the sole estimation and the inferential method under normal assumption. The application to a novel count-valued RNA-seq data set of childhood atopic asthma in Puerto Ricans demonstrates more biologically meaningful results compared to the sole estimation and the inferential methods based

on Gaussian and nonparanormal graphical models.

In the third part of the dissertation, we propose R-CoLaR, a novel Robust Convex Program with group-Lasso Refinement combining the cutting-edge tail-robust covariance estimation for sparse CCA. Numerical studies and the analysis of the heavy-tailed CITE-seq data of a mucosa-associated lymphoid tissue (MALT) tumor have successfully illustrated the validity and noticeable advantages of R-CoLaR over existing methods of sparse CCA in more accurate estimation and better interpretation of protein-RNA correlation.

Keywords: Gene co-expression network; Multi-omics; High-dimensional statistical inference; Efficient package; Modified Poisson graphical model; RNA-seq; Bias correction; Sparse CCA; Heavy-tailed; Tail-robust covariance estimation.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation presents my research accomplishments during the past five years of study. The research provides new developments of statistical methods and applications in gene co-expression network analysis and multi-omics studies, which will be potentially helpful for biological researchers in facilitating the study of disease mechanisms and new treatment development.

Research was challenging for me initially, but overcoming its difficulties is inseparable to all the help from the people I received during the five years. I would have never been able to finish this dissertation without guidance from my thesis committee members, encouragement from my collaborators and friends, and support from my parents and wife.

I would like to convey my deepest gratitude to my advisor, Dr. Zhao Ren. I really appreciate his careful supervision on each of my research projects and valuable suggestions to improve my dissertation. During the past five years, he has taught me how to be a good researcher by always seeing the big picture of a problem. His deep knowledge in statistical theory and way of thinking a problem have gradually cultivated my insights in statistics, which will be always helpful for my future career.

I would like to thank my committee members, Dr. Satish Iyengar, Dr. Kehui Chen and Dr. Wei Chen for their continuous caring of my research progress. I would like to particularly thank Dr. Wei Chen for giving me an opportunity to learn the beauty of statistics with its application to bioinformatics and genetics. The efficient computational coding and skills that I have learned play important roles in achieving feasible implementations of new methods under high-dimensional settings in my research projects.

I am also grateful to Dr. Juan C. Celedón for sharing the precious RNA-seq gene expression data of childhood atopic asthma in Puerto Ricans to one of my research projects, and Dr. Erick Forno for active collaborations of applying different statistical methods in the study of childhood asthma. These collaborative experiences have made me learn how to act as a good statistician within an interdisciplinary research group.

I really appreciate all the encouragement from Hongyi Xin, Qi Yan, Tao Sun and Soyeon

Kim at Dr. Wei Chen's lab and my fellow students at Department of Statistics. They have made my Ph.D. life enjoyable and memorable.

I want to sincerely thank my father Yeyi Zhang and mother Yaping Hu for their continuous support of my decision to pursue the Ph.D. degree. Finally, I would like to thank my wife Mingyue Fan for her company and support during this period. She has always encouraged me to be a better person.

1.0 Introduction

1.1 Background

High-dimensional statistical estimation and inference is a big field in modern statistical theory and applications. Below are some necessary topics related to the development of this dissertation, including regression, undirected graphical models and canonical correlation analysis (CCA) under high-dimensional settings. For a more comprehensive overview of the development of high-dimensional statistics, see the following well-written textbooks: Bühlmann and van de Geer (2011), Giraud (2014), Hastie et al. (2015).

In many applications of biological research, statistical relationship among the variables has been extensively paid attention to. One typical case is to study the relationship between the variables (or features) and the response, or in other words, to identify the set of variables that influence the response. If there is a response vector $y \in \mathbb{R}^n$ and a covariate matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with n to be the number of observations and p to be the number of covariates under consideration, the classical linear regression model is defined as

$$y = \boldsymbol{X}\beta + \epsilon, \tag{1.1}$$

where $\beta \in \mathbb{R}^p$ is denoted as a vector of regression coefficients, and ϵ_i is denoted as an error term which follows i.i.d. normal distribution $\mathcal{N}(0, \sigma^2)$ for $i = 1, 2, ..., n$. Under the traditional low-dimensional setting with fixed p and increasing sample size n, it is well known that the least squares estimator $\hat{\beta}^{lse}$ can be obtained via minimizing the sum of squared errors, that is

$$\hat{\beta}^{lse} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} ||y - \boldsymbol{X}\beta||_2^2. \tag{1.2}$$

$\hat{\beta}^{lse}$ has many good statistical properties like consistency and efficiency. However, in the era of "big data", there are explosive data sets with p allowed to be larger than n. Under these high-dimensional settings, the traditional approach fails to obtain a consistent estimator of

β, because the inverse of the covariance matrix $\boldsymbol{X}^\intercal\boldsymbol{X}$ is ill-defined. A popular approach that tailors to the high-dimensional linear regression is the Lasso (Tibshirani, 1996), i.e.,

$$\hat{\beta}^{Lasso} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\, \frac{1}{2n}||y-\boldsymbol{X}\beta||_2^2 + \lambda||\beta||_1. \tag{1.3}$$

The extra ℓ_1 penalty on the coefficients of β results in a sparse solution of $\hat{\beta}^{Lasso}$, and the number of zero entries is controlled by the tuning parameter λ. A broad range of literatures have studied theoretical properties of Lasso including the rate of convergence under a variety of norms and the consistency in support recovery under some mild conditions assumed on $\boldsymbol{X}$, see for example Zhao and Yu (2006), Wainwright (2009), Candès and Plan (2009) and Bühlmann and van de Geer (2011). Later on, the scaled Lasso (Sun and Zhang, 2012) was proposed as an alternative approach to deal with the high-dimensional linear regression problems (see also an equivalent approach in Belloni et al. (2011)). The scaled Lasso is to esimate the regression coefficients β and the noise level σ jointly, i.e.

$$(\hat{\beta}^{scaledLasso}, \hat{\sigma}^{scaledLasso}) = \underset{\beta\in\mathbb{R}^p,\sigma>0}{\arg\min}\, \frac{||y-\boldsymbol{X}\beta||_2^2}{2\sigma} + \frac{n\sigma}{2} + \lambda||\beta||_1. \tag{1.4}$$

A major advantage of the scaled Lasso over the Lasso is the tuning-free property, which enables the usage of a pre-determined universal tuning parameter λ. In the past six years, the rigorous statistical inference of high-dimensional regressions has been developed to make the confidence interval and the hypothesis testing of the estimators available, see for example Zhang and Zhang (2014), van de Geer et al. (2014) and Javanmard and Montanari (2014). The main idea of these techniques is to correct the bias of the Lasso estimator $\hat{\beta}^{Lasso}$ or the scaled Lasso estimator $\hat{\beta}^{scaledLasso}$ to an acceptable level. For example in Zhang and Zhang (2014), the de-biased estimator of coefficient for X_j (the column j of $\boldsymbol{X}$) is defined as

$$\hat{\beta}_j^{debiased} = \hat{\beta}_j^{Lasso} + \frac{v_j^\intercal(y-\boldsymbol{X}\hat{\beta}^{Lasso})}{v_j^\intercal X_j}, \tag{1.5}$$

where $v_j \in \mathbb{R}^n$ is a carefully chosen score vector. One popular choice of v_j is $v_j = X_j - \boldsymbol{X}_{-j}\hat{\gamma}_j$, which resembles the residual of the Lasso regression of X_j on $\boldsymbol{X}_{-j}$ (the covariate matrix without column j). Intuitively, the second term on the right-hand side of (1.5) projects the residual to the direction of v_j so as to correct the bias of $\hat{\beta}_j^{Lasso}$. Theoretically, each $\hat{\beta}_j^{debiased}$ has been proved asymptotically normal under a certain sparsity assumption on β

and a certain assumption on the design matrix $\boldsymbol{X}$.

The second case for the statistical relationship is to study the conditional dependencies among the covariates $X = (X_1, X_2, ..., X_p)$ and visualize them in a network structure. The undirected graphical model (or the Markov random field) is a powerful tool for these specific purposes. Each graph is defined as $G = (V, E)$, which consists of a node set $V = (X_1, X_2, ..., X_p)$ and an edge set $E = \{$pairs of (i, j) if there is an undirected edge between X_i and $X_j\}$. According to Lauritzen (1996), the two nodes X_i and X_j are conditionally independent given the conditions of all the other variables if there is no edge between them; otherwise, they are conditionally independent. The Gaussian graphical model (GGM) is one of the most commonly used models in this field with X following a multivariate normal distribution. For GGM, it is well known that a conditional independence between X_i and X_j is equivalent to a zero entry of $(i, j)^{th}$ element ($\omega_{ij} = 0$) of the precision matrix $\Omega = (\omega_{ij})_{p\times p} = \Sigma^{-1}$, an inverse of the covariance matrix Σ of X (Lauritzen, 1996). In the past decade, the estimation of the high-dimensional GGM has been paid more attention to. In summary, there are two types of approaches: 1. the graphical Lasso (Yuan and Lin, 2007), a log-likelihood based approach with a penalty on the precision matrix; 2. the neighborhood selection (Meinshausen and Bühlmann, 2006), a penalized regression approach for each node X_i on the other nodes X_{-i} which is closely related to the Lasso-type regressions in (1.3) and (1.4). Following these two directions, the rigorous statistical inference of high-dimensional GGM has been developed in the past seven years to allow the edge-wise confidence interval and the p-value of each ω_{ij}, and the multiple testing for all the ω_{ij}'s, see for example Ren et al. (2015), Janková and van de Geer (2015, 2017) and Liu (2013). These approaches of statistical inference essentially rely on the initial estimation to derive a test statistic that follows an asymptotically normal distribution at $\sqrt{n}$ rate.

Moreover, the study of the statistical relationship between two or more types of data has recently played a much more important role in different areas. Particularly in the genomics studies, analsysis on the different types of omics data (or multi-omics data analysis) helps us to better understand a complex biological mechanism. The canonical correlation analysis (CCA) (Hotelling, 1936) is an important technique to study the relationship between two types of data. Essentially, CCA is to find the canonical directions that maximize the

correlation between the two data sets after projecting onto them. Suppose that there are two data sets $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times q}$ with the number of observations to be n and the dimensions to be p and q respectively. In the regime of low-dimensional settings with fixed p, q and increasing n, CCA can be well solved by the singluar value decomposition (SVD) (Hotelling, 1936). But in high-dimensional settings, the conventional SVD approach fails because the inverses of sample covariance matrices $\hat{\boldsymbol{\Sigma}}_x^{-1}$ and $\hat{\boldsymbol{\Sigma}}_y^{-1}$ are ill-defined. Within the past decade, many efforts have been made in developing the new methods of sparse CCA that tailor to the high-dimensional data analysis, see for example Witten and Tibshirani (2009), Chen et al. (2019), Gao et al. (2015), Gao et al. (2017) and Suo et al. (2017).

1.2 Overview of the dissertation

The dissertation is motivated by two types of popular studies in biological research: gene co-expression network analysis and multi-omics study for explanation of complex biological processes and diseases. Gene expression data sets are explosively generated and are usually large-scale with the number of genes (or dimension) p to a thousand or a ten-thousand level. There is an urgent demand for developing an efficient software package for analysis of large amount of high-dimensional data. More importantly, most methods mentioned above are applicable to Gaussian or sub-Gaussian data. However, many omics data sets are non-Gaussian and even have heavy tails, for example, count-valued RNA-seq data and data from the state-of-the-art cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq). Therefore, there is also an urgent demand for developing new methods that tailor to analysis of those non-Gaussian data sets.

In Chapter 2, we focus on the implementation of the statistical inference approaches for high-dimensional Gaussian graphical model (GGM) in large-scale gene co-expression network analysis. Even though the theoretical part of the rigorous statistical inference of high-dimensional GGM has been well developed, there is lack of practical usage for these methods in the real biological network analysis due to no available efficient and user-friendly software packages. In order to narrow the gap between the theory and the practice, we

develop an efficient and extensive `R` package named `SILGGM` (Zhang et al., 2018b) which involves four cutting-edge approaches (Ren et al., 2015; Janková and van de Geer, 2015, 2017; Liu, 2013) that tailor to the edge-wise and the global inference of high-dimensional GGM. The development of `SILGGM` has dramatically accelerated current implementations and allowed the computation of very large-scale settings with even ten thousand of variables available. Furthermore, we compare and validate the accuracy of all the approaches in the very high-dimensional settings with p to a ten-thousand level using `SILGGM`. The different situations to use which method have been discussed based on our comparison studies. In the end, applications of `SILGGM` to a novel single-cell RNA-seq data set with pan T cells for large-scale gene network analysis have shown the advantages of these approaches.

In Chapter 3, we go a step further and pay attention to the method development for the statistical inference of high-dimensional non-Gaussian graphical models, particularly the modified Poisson-type graphical models. Even though GGM is powerful and has been widely used in many applications, it is not suitable for the count-valued data that are discrete and non-negative without any ad hoc transformation. Due to the prevelance of count-valued data in reality (e.g. RNA-seq data), the novel methods for the statistical inference of those non-Gaussian graphical models are necessary. We focus on the three modified Poisson-type graphical models, see Yang et al. (2013) and Inouye et al. (2016). Based on the idea of the debiased approach shown in (1.5) from Zhang and Zhang (2014) and Li et al. (2016), we propose a novel two-step procedure for the edge-wise and the global inference of their conditional dependencies. It should be noted that the statistical inference of those high-dimensional non-Gaussian graphical models are more challenging than the ones of GGM due to the incondordance between conditional dependencies and precision matrices. Besides, we have also provided a computationally efficient implementation to achieve the proposed approach. Extensive numerical studies and applications to a novel RNA-seq gene expression data set of childhood atopic asthma in Puerto Ricans have illustrated the advantages of our proposed method with more accurate and more biologically meaningful inferential results over existing methods under normal and nonparanormal assumptions.

In Chapter 4, we further develop a new method for the estimation of sparse CCA in the non-Gaussian data, particularly the ones with heavy-tailed distributions due to the increasing

popularity of multi-omics study in biological research. There has been several efforts made on the theoretical and methodological development of sparse CCA under high-dimensional settings, see Witten et al. (2009), Chen et al. (2019), Gao et al. (2017) and Suo et al. (2017). One of the examples is the approach of convex optimization with group-Lasso refinement (CoLaR) from Gao et al. (2017). The approach of CoLaR works well under the situations where data has a normal or a roughly normal distribution. However, the estimation errors from the current CoLaR fail to be controlled well if the data has a heavy-tailed distribution (e.g. Student's t, Pareto or Log-normal) due to its non-robust sample covariance estimation. Moreover, we illustrate that the phenomenon of heavy-tailed distributions is common in real data sets even though normalization or an ad-hoc transformation has been made, for example, CITE-seq data which provides abundance of RNA-seq and surface proteins on a same set of cells simultaneously and opens a new door for study on cell-level protein-RNA correlation. Therefore, based on the novel approaches in recent development of tail-robust covariance matrix estimation (Catoni, 2016; Avella-Medina et al., 2018; Ke et al., 2019), we propose a new method called R-CoLaR by robustifying covariance matrix estimation within the current framework of CoLaR. Numerical studies and applications to the CITE-seq data of a mucosa-associated lymphoid tissue (MALT) tumor have shown noticeable advantages of R-CoLaR over existing methods of sparse CCA on heavy-tailed distributions.

In Chapter 5, we provide an overall summary of the three main chapters and a brief discussion of their future works.

2.0 SILGGM: An Extensive R Package for Efficient Statistical Inference in Large-Scale Gene Networks

2.1 Introduction

Gene co-expression network is an undirected graph, where each node represents a gene and each edge between two genes shows a significant co-expression relationship (Stuart et al., 2003). It has been of great biological interests and widely used in exploring underlying mechanisms of complex biological processes since the co-expressed genes are usually functionally related and share a same pathway (Weirauch, 2011; Filteau et al., 2013; Gaiteri et al., 2014; Parikshak et al., 2015). However, it is always a concern whether the inferred gene network structure is trustworthy or not. A partial correlation-based approach to assess the conditional dependence of two genes given the conditions of other genes in a network is a more reliable choice to infer a gene network since the marginal correlation may fail to reflect a true gene-gene relationship without considering other genes' effects. Gaussian graphical model (GGM) is a typical statistical model to interpret gene dependence with the conditions of other genes.

Previous high-throughput sequencing technologies like microarray and bulk RNA-seq have generated many high-dimensional gene expression data sets with a huge number of genes, but these data sets usually have a small number of subjects or samples. Recently, the emergence of the droplet-based single-cell RNA-seq (Macosko et al., 2015; Mazutis et al., 2013) has made the cell-level gene measurements available, and its increasing availability has led to a growing number of even larger gene expression data sets which generally have thousands of subjects and tens of thousands of genes. These high-dimensional settings have imposed bigger statistical and computational challenges in obtaining a reliable gene network.

Due to the assumption of the intrinsically sparse structure of a gene network, two main streams of approaches have been developed in estimating conditional dependence of genes using high-dimensional GGM: (i.) the graphical Lasso, which is a penalized-likelihood approach for precision matrix of GGM (Yuan and Lin, 2007; Friedman et al., 2008; d'Aspremont et al.,

2008) and (ii.) a neighbourhood-based approach with a penalized regression (Meinshausen and Bühlmann, 2006; Yuan, 2010; Sun and Zhang, 2013). Over the recent seven years, more important efforts have been made in rigorous statistical inference of gene-gene conditional dependence with high-dimensional GGM: the bivariate nodewise scaled Lasso (B_NW_SL) (Ren et al., 2015), the de-sparsified nodewise scaled Lasso (D-S_NW_SL) (Janková and van de Geer, 2017), the de-sparsified graphical Lasso (D-S_GL) (Janková and van de Geer, 2015) and the GGM estimation with false discovery rate (FDR) control using scaled Lasso or Lasso (GFC_SL or GFC_L) (Liu, 2013). These approaches have two main advantages over the ones in sole estimation: (i). the obtained estimators of conditional dependence are more precise and asymptotically efficient with each variance equal to the inverse of Fisher information; (ii). the estimators are asymptotically normal under a minimal sparsity condition (e.g. the maximum node degree satisfies $s = o(\sqrt{n}/\log(p))$), so the corresponding confidence intervals or p-values are provided besides point estimators for identifying a more reliable gene network.

There are some existing software packages for gene co-expression network analysis. For example, the popular `R` package `WGCNA` (Langfelder and Horvath, 2008) provides functions to construct a gene co-expression network based on the marginal correlations. In terms of the partial correlation-based approaches particularly for large-scale settings, `glasso` (Friedman et al., 2008) and `huge` (Zhao et al., 2012) are two widely adopted packages for fast estimation of gene-gene conditional dependence based on the high-dimensional GGM. More recent packages include `FastCLIME` (Pang et al., 2014), `flare` (Li et al., 2015) and `XMRF` (Wan et al., 2016). Unlike the marginal correlation-based approaches and high-dimensional GGM estimation, there are in practice few efficient packages or algorithms for the aforementioned approaches of rigorous statistical inference with the partial correlations that are supposed to be more powerful in large-scale gene-gene network analysis. `FastGGM` (Wang et al., 2016) is the recently developed package for an efficient and tuning-free implementation of B_NW_SL and has made the method computationally feasible to tens of thousands of genes. However, some redundant steps in the algorithm can be further improved and the outputs in only a matrix format make the package less friendly to users. Except `FastGGM`, no efficient `R` package has been proposed for the other above related works, and the expensive computation of

naïve implementation also remains a challenge for these approaches.

To enhance the influence of these cutting-edge statistical inference works in practical usage and address the computational challenge in high-dimensional settings even with large sample sizes, we develop a more comprehensive package called `SILGGM` (**S**tatistical **I**nference of **L**arge-scale **G**aussian **G**raphical **M**odel) that includes B_NW_SL, D-S_NW_SL, D-S_GL and GFC_SL or GFC_L. `SILGGM` has significantly increased the efficiency of each approach using fast algorithms, the `Rcpp` library (Eddelbuettel et al., 2011) and some additional optimizations. It also provides a consistent framework of statistically efficient inference on both individual gene pair and all gene pairs by extending the implementation of B_NW_SL, D-S_NW_SL and D-S_GL to global inference with FDR control under the framework of GFC_SL or GFC_L. Compared to `FastGGM`, `SILGGM` has several advantages. First, some steps in inner product calculations are optimized in the core algorithm of `SILGGM`, so B_NW_SL is performed even faster than its implementation in `FastGGM`. Second, `SILGGM` can accommodate users' different research purposes with a new functionality of global inference for FDR control and with more flexible choices of methods. Third, based on users' preference, the outputs in `SILGGM` can also be saved in a table format that is able to be further used directly in multiple platforms for network visualization like `Cytoscape` (Shannon et al., 2003), `BisoGenet` (Martin et al., 2010) and `BiNA` (Gerasch et al., 2014). Overall, the package `SILGGM` is an extensive and user-friendly tool that aims to facilitate large-scale gene network analysis with rigorous statistical inference and to show more trustworthy statistical results in a biological sense.

2.2 Design and implementation

In GGM, a set of p-dimensional random variables $X = (X_1, X_2, ..., X_p)^\intercal$ follows a multivariate normal distribution with mean μ (assuming $\mu = 0$ without loss of generality) and covariance matrix Σ. The conditional dependence between each pair of variables is reflected in a precision matrix $\Omega = (\omega_{ij})_{p\times p} = \Sigma^{-1}$, the inverse of Σ. For instance, if X_i and X_j are conditionally dependent, then equivalently, the corresponding element in Ω is $\omega_{ij} \neq 0$

(Lauritzen, 1996). In the gene network analysis, we regard X_i as the i^{th} gene. Therefore, the inference between gene i and j is equivalent to the inference of an individual ω_{ij}, and the global inference of whole-scale gene pairs is based on a multiple testing procedure with all ω_{ij}'s.

2.2.1 Software architecture

We focus on the high-dimensional settings with p (the number of genes) allowed to be far larger than n (the number of subjects). The `SILGGM` package has one main function `SILGGM()` with various arguments and its workflow is described in Figure 2.1.

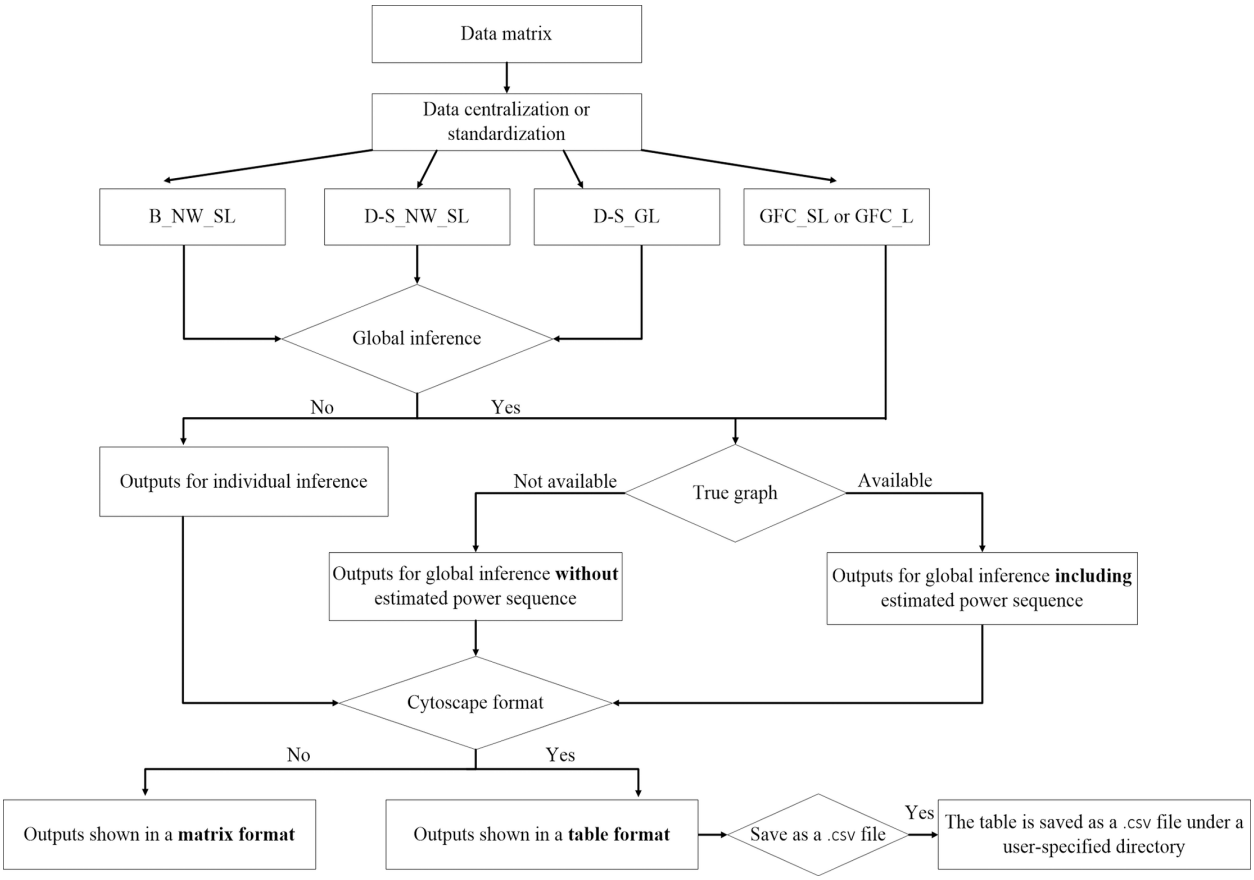

Figure 2.1: The workflow of the `SILGGM` package.

The setup of the `SILGGM()` function is very simple. It only takes an n by p gene expression data matrix as an input. The gene names can be specified in each column by users. Without loss of generality, the data matrix is further centralized by subtracting its mean or standardized by subtracting the mean and adjusting the variance to one before the formal statistical inference, but the final results are returned in an original scale.

The `method` argument in the function `SILGGM()` supports four approaches in rigorous statistical inference: B_NW_SL, D-S_NW_SL, D-S_GL, GFC_SL or GFC_L. In the original four papers, the first three methods are developed for inference of each individual ω_{ij}, while the last one is proposed particularly for simultaneous inference of all ω_{ij}'s. All of the four methods (see more details in Appendix A.1) can be summarized into two steps. The first step involves a Lasso-type regularization approach. The graphical Lasso is performed in D-S_GL, while $O(p)$ or $O(sp)$ runs of nodewise Lasso-type regressions are conducted among the other three methods. The second step is to obtain $(p^2-p)/2$ test statistics: (i.) the estimators $\hat{\omega}_{ij}$'s for B_NW_SL; (ii.) the de-sparsified estimators $\check{\omega}_{ij}$'s for D-S_NW_SL and D-S_GL; (iii.) the de-sparsified newly-constructed test statistics $\hat{T}_{ij}$'s for GFC_SL or GFC_L, each of which is asymptotically efficient and normal at $\sqrt{n}$ rate under a minimal sparseness condition.

As it can be seen, GFC_SL or GFC_L essentially relies on asymptotically normal test statistics for testing on ω_{ij}'s, so the implementations of the other three methods can also be extended to global inference under its FDR framework (Liu, 2013) that has been rigorously proved to be valid in high-dimensional settings. The `global` argument in the function determines whether or not to perform global inference in the other three methods. Since global inference needs FDR control, an α-level sequence with $\alpha = 0.05, 0.1$ is pre-specified by the `alpha` argument in the function, and it can be customized by users with different values.

Outputs are shown with the different types of inference. For individual inference of gene i and j, `SILGGM` not only provides the estimator $\hat{\omega}_{ij}$ or $\check{\omega}_{ij}$, but also obtains the associated confidence interval, z-score and p-value. Each output of gene i and j is encoded in the $(i,j)^{th}$ element of a p by p symmetric matrix with diagonal elements equal to 0. For global inference with a pre-specified α-level sequence, `SILGGM` further returns the estimated FDR sequence based on $\hat{T}_{ij}$'s or z-scores of $\hat{\omega}_{ij}$'s or $\check{\omega}_{ij}$'s, the corresponding threshold sequence for absolute

values of test statistics and a series of decisions for conditional dependence between each gene pair (a list of p by p adjacency matrices with each off-diagonal element value of 1 = conditionally dependent or 0 = conditionally independent). If the true structure of a gene network is available (e.g. a simulation study or a real study with sufficient prior knowledge), `SILGGM` also includes the estimated power sequence with respect to the estimated FDR sequence. Users can input the true structure in a matrix format via the `true_graph` argument in the `SILGGM()` function.

In addition to present the above outputs from both individual and global inference in a matrix format, the function `SILGGM()` provides the `cytoscape_format` argument as an alternative to show them in a table format that can be saved as a `.csv` file by using the `csv_save` argument to a directory specified by the directory argument. The `.csv` file is compatible with multiple popular platforms for network visualization. In order to show the validity of this alternative, we have applied `SILGGM` to to a public gene expression microarray data set on the lymphoblastoid cells of $n = 258$ asthmatic children (Brazma et al., 2003; Liang et al., 2013) and $p = 1953$ genes with the largest inter-sample variance by using the method GFC_SL with an FDR control at the level of 0.05. Figure 2.2(A) gives a table in the `.csv` file with the 20 most significant gene pairs based on a rank of the absolute values of test statistics $\hat{T}_{ij}$'s with the hub gene CLK1 that has been proved to be susceptible to asthma (Verheyen et al., 2004). The first two columns ("gene 1" and "gene 2") show the names of each non-overlapped gene pair. The following column "test_statistic" indicates the test statistic $\hat{T}_{ij}$ of gene i and j. At the end, the column "global_decision_0.05" shows the decision for conditional dependence between each gene pair under global inference with FDR control at the 0.05 level. All the gene pairs are conditionally dependent in this example. Furthermore, we import the `.csv` file to `Cytoscape (version 3.4.0)` and obtain the corresponding network visualization shown in Figure 2.2(B).

2.2.2 Features of efficient implementations

Computational efficiency is a prominent advantage of `SILGGM`. The core algorithms in the package are developed with the `Rcpp` library (Eddelbuettel et al., 2011) which highly

(A)

gene1	gene2	test_statistic	global_decision_0.05
CLK1	C1ORF63	7.068319578	1
CLK1	DNAJB1	6.370053216	1
CLK1	RGS1	5.519126715	1
CLK1	RNF146	5.291264048	1
CLK1	POLQ	-5.191657692	1
CLK1	PIAS1	-5.131982503	1
CLK1	DNAJB9	4.986321414	1
CLK1	EGR1	-4.808012448	1
CLK1	ARMCX3	4.588409301	1
CLK1	DNAJB4	4.583513721	1
CLK1	HMMR	-4.55370168	1
CLK1	MIR21	-4.552792912	1
CLK1	TRA2A	4.481474871	1
CLK1	PPP4R2	4.441497616	1
CLK1	FAM76B	4.393816269	1
CLK1	C1ORF199	4.390922346	1
CLK1	PDE4B	-4.373685294	1
CLK1	SLC16A6	-4.318410531	1
CLK1	PIK3C2A	-4.136358875	1
CLK1	IER3	-4.11946634	1

(B)

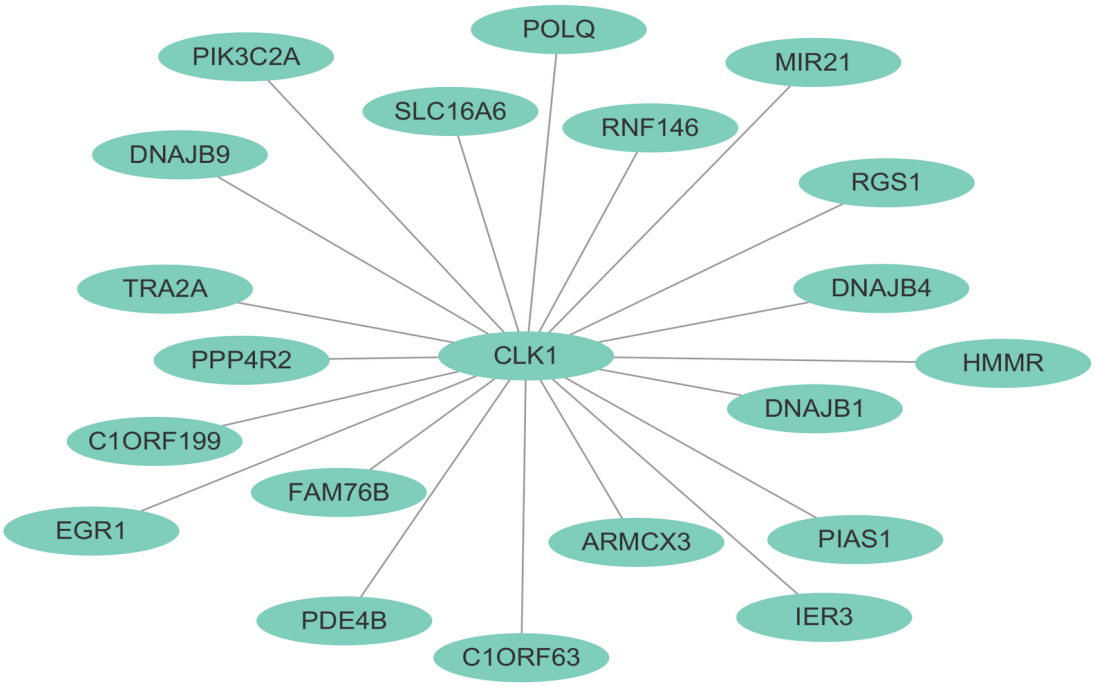

Figure 2.2: An example of table-format outputs and the corresponding network visualization. (A) A table in the `.csv` file generated by the `SILGGM` package using the method GFC_SL. (B) The corresponding network visualization.

speeds up the loop operation and makes the implementation of `C++` code available in `R`. In addition to the fast programming language, there are many other key features of efficient implementations making `SILGGM` feasible in high-dimensional settings. We outline the details according to the two summarized steps of all the approaches as below.

In the first step, based on the same optimization in `FastGGM` (Wang et al., 2016), we pre-calculate and save the covariance matrix to avoid its repetitive calculation before solving each Lasso-type problem. Then, we apply the cyclical coordinate descent algorithm with covariance update (Friedman et al., 2010) that has been shown much faster than other competing methods like the LARS procedure (Efron et al., 2004) in solving Lasso-type problems. To further increase the efficiency, some tuning-free schemes (e.g. the scaled Lasso with tuning parameter $\lambda = \sqrt{2\log(p/\sqrt{n})/n}$, the graphical Lasso with a certain $\lambda = \sqrt{\log(p)/n}$ suggested in Janková and van de Geer (2015) are applied to avoid inefficient tuning selection. Our coding with the scaled Lasso is more efficient than directly using the package `scalreg` (Sun and Zhang, 2013) which is built on the `lars` package (Efron et al., 2004). To conduct the graphical Lasso in D-S_GL, we use the package `glasso (version 1.8)` due to the great improvement in its efficiency by the screening procedures (Witten et al., 2011). In addition, for GFC_L which requires tuning selection for FDR control, we apply the "warm start" optimizations (Friedman et al., 2010) to boost the procedure.

In the second step, we facilitate inner product operations to derive each de-sparsified test statistic. To be more specific, we consider the sparsity of Lasso-type estimators from the first step and make inner product calculations only on the non-zero elements. For D-S_NW_SL or D-S_GL, to obtain $\check{\omega}_{ij}$ needs an inner product which requires a total number of operations $O(p^3)$ with naïve calculation (see (A.7) and (A.10) in Appendix A.1). By considering the sparsity, the total number of operations can be reduced to $O(sp^2)$, and s is usually much smaller than p in high-dimensional settings.

In addition to the aforementioned optimizations, we optimize the inner product operations between the whole data matrix and regression coefficients by considering the sparsity of estimated coefficients when solving each scaled Lasso problem. The idea behind the optimization is same as the one used in the second step, and it reduces the redundant steps and enables B_NW_SL to perform even faster in `SILGGM` than in `FastGGM`.

2.3 Results

We illustrate the efficiency of `SILGGM` through simulation studies and real data analysis. In simulation, we considered four popular graph structures for gene network studies: Band graph, Hub graph, Erdős - Rényi (E-R) random graph and Scale-free random graph, as shown in Figure 2.3 that was generated by the `R` package `huge`. We not only evaluated time efficiency of `SILGGM`, but also made an extensive validation testing the estimation accuracy of `SILGGM` for both individual and global inference particularly in the very high-dimensional scenarios. More detailed results about individual and global inference are presented in Appendices A.3 and A.4 respectively. The real data analysis of `SILGGM` is based on a novel single-cell RNA-seq study on the gene expression of Pan T cells.

2.3.1 Performance benchmark in simulation

To the best of our knowledge, the online `MATLAB` code (see "models.txt" and "GFC-lasso.txt" from `http://math.sjtu.edu.cn/faculty/weidongl/Publication/code.rar`) is the only publicly available implementation of GFC_L prior to the development of `SILGGM`. In order to compare its time performance with GFC_L implementation in `SILGGM`, we set $n = 100$ and simulated three types of graph settings: Band, Hub and E-R (see the details of the graph settings in Appendix A.2), same as those in Liu (2013) with $p = 50, 100, 200$. Total timings (in seconds) over 100 replications on a single CPU were recorded for GFC_L with FDR control at the 0.1 and the 0.2 levels using `SILGGM` and the `MATLAB` code, as shown in Table 2.1. GFC_L implemented with `SILGGM` is generally around 60 times faster among all the scenarios and can be up to 70 times in some cases. The above simulations were conducted on a PC with Intel Core i5-3230M CPU @ 2.60GHz. The significant speed improvement in GFC_L implementation is mainly due to the incorporation of `Rcpp` library and the optimization of redundant steps of FDR calculation in tuning selection for FDR control.

Then, we evaluated the timing performance of B_NW_SL using `SILGGM` compared to the current package `FastGGM`. As shown in Table 2.2, the E-R graph settings (see Ap-

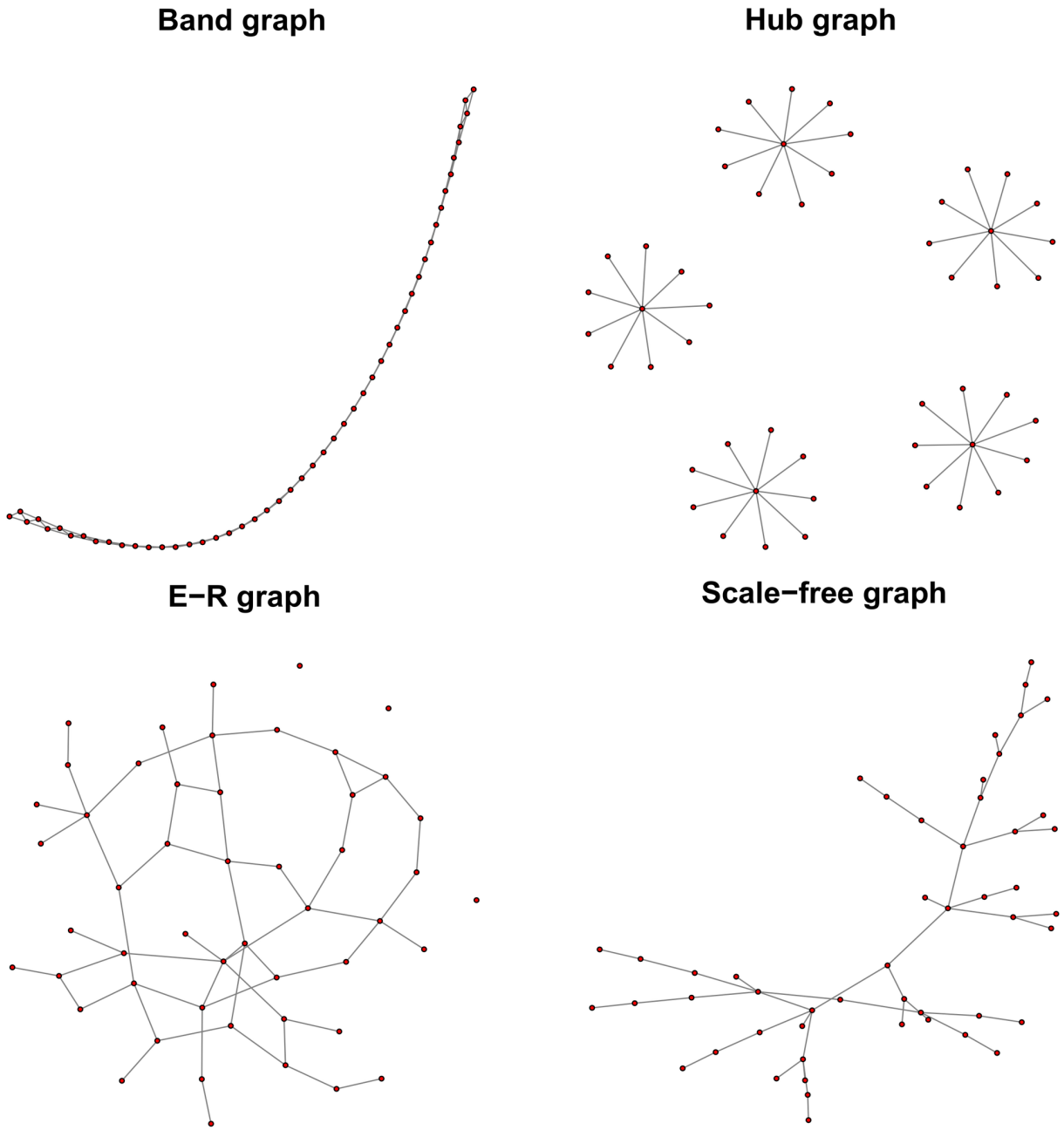

Figure 2.3: Four possible graph structures in simulation studies.

Table 2.1: Total timings (in seconds) of GFC_L (`SILGGM`) and GFC_L (`MATLAB`).

	GFC_L (`SILGGM`)			GFC_L (`MATLAB`)		
p	50	100	200	50	100	200
Band	22.1	77.8	370.9	1396.2	4730.5	20039.3
Hub	22.8	84.1	363.7	1556.8	5227.7	19747.2
E-R	20.2	89.6	377.3	1495.7	5627.0	21482.2

pendix A.2) same as those in Wang et al. (2016) were simulated with $n = 400, 800$ and $p = 800, 1000, 2000, 5000, 10000$ to make sure that the expected node degree of each graph, which is the value of π (the probability of $\omega_{ij} \neq 0$ for $i \neq j$) times p, is around 4 or 5. The first column of Table 2.2 also gives the estimated average node degree of each case. We carried out the experiments on a Linux server with Intel Xeon CPU E5-2695 v2 @ 2.40GHz. To be as fair as possible, we performed B_NW_SL without global inference in `SILGGM`, so the outputs are same as the ones from `FastGGM`. Timings (in seconds) for one run on a single CPU with both `SILGGM` and `FastGGM` are reported in Table 2.2 using the same simulated data set from each graph setting. As it can be seen, B_NW_SL is implemented even faster in `SILGGM` among all of the scenarios, and the computational cost of each scenario is reduced by 20% to 56%.

In addition to the time evaluation, we validated the accuracy of estimation results from all the approaches for both individual and global inference in the very large-scale settings with relatively small sample sizes ($n = 800$, $p = 5000$ and $n = 800$, $p = 10000$).

We at first assessed the performance of individual inference of each $(i, j)^{th}$ gene pair ($H_0 : \omega_{ij} = 0$ vs. $H_1 : \omega_{ij} \neq 0$) in terms of the estimation for an entire graph. The empirical Type I error (the probability of falsely rejecting H_0 when there is actually a known zero partial correlation between gene i and j) under a pre-specified level of 0.05 for p-values and the corresponding Type II error (the probability of failing to reject H_1 when there is

Table 2.2: Total timings (in seconds) of B_NW_SL (`SILGGM`) and B_NW_SL (`FastGGM`).

Average node degree	π	p	n	B_NW_SL (`SILGGM`)	B_NW_SL (`FastGGM`)
4.045	0.005	800	400	24.9	36.6
4.994	0.005	1000	800	72.8	145.2
4.970	0.0025	2000	800	411.7	938.5
5.0264	0.001	5000	800	5772.6	8663.7
5.0498	0.0005	10000	800	40080.6	49650.2

actually a known non-zero partial correlation between gene i and j) were measured for Band graph (same as that described in Liu (2013)), E-R graph (same as that described in Liu (2013)) and Scale-free graph (see Appendix A.3). The good performance of the empirical Type I and Type II error rates has shown the validity of all the approaches in the package `SILGGM` for individual inference even in the very high-dimensional scenarios (see more detailed results in Appendix A.3). To make a further comparison for individual inference, we also evaluated the average empirical coverage probabilities for the 95% confidence intervals of the ω_{ij}'s for the "non-zero partial correlation" set (a set of all pairs with non-zero ω_{ij}'s) and the "zero partial correlation" set (a set of all pairs with zero ω_{ij}'s) respectively. Since GFC_SL or GFC_L provides no confidence intervals, we included the other three approaches here. According to the results from the three graph settings, the overall performance of the confidence intervals among B_NW_SL, D-S_NW_SL and D-S_GL are good in terms of the entire graph structure. But in terms of the confidence intervals for the non-zero partial correlations, B_NW_SL and D-S_NW_SL outperform D-S_GL. Moreover, the performance of B_NW_SL is more stable than that of D-S_NW_SL in the different settings (see more details in Appendix A.3). Therefore, for individual inference of a gene pair which further requires the information of a confidence interval, B_NW_SL is a more desirable choice compared to the other approaches, but D-S_NW_SL can be an alternative to save time for the very high-

dimensional cases.

Then, we evaluated the performance of global inference of all gene pairs for the overall partial correlation recovery in the very large-scale settings based on the same three graph settings (Band, E-R and Scale-free) used in individual inference. Unlike individual inference, global inference generally requires a multiple testing procedure for tests on all $H_0 : \omega_{ij} = 0$ vs. $H_1 : \omega_{ij} \neq 0$ with $1 \leq i < j \leq p$ simultaneously. Therefore, to make global inference of all gene pairs in a large graph, we always recommend controlling FDR to avoid the inflation of false positives. The testing results from the three graph settings indicate that the FDRs of all the methods are effectively controlled below the desired level for both $p = 5000$ and $p = 10000$. The corresponding power values (the proportions of the correctly identified elements among the known non-zero partial correlations) and the Matthews correlation coefficients (MCCs) also demonstrate comparably good performance of all the methods (see Appendix A.4 for more details). Overall speaking, the good performance of FDR, power and MCC has shown the validity of all the approaches in correctly identifying the zero and the non-zero partial correlations in a global sense even for the very high-dimensional scenarios.

2.3.2 Gene network analysis in a droplet-based single-cell data set with pan T cells

We also applied the `SILGGM` package to a novel public single-cell RNA-seq data set with pan T cells isolated from peripheral blood mononuclear cells of a healthy human donor (`https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/t_3k`). The data set generated by the latest `CellRanger` pipeline (Zheng et al., 2017) includes $n = 3555$ cells. After filtering out the unexpressed genes, we considered $p = 2000$ genes with the largest inter-sample variance.

Since the genes in the data set are measured with the unique molecular identifier (UMI) counts (Islam et al., 2014), we need to transform the count values before the use of `SILGGM`. According to Jia et al. (2017), it is reasonable to take a $\log 2$(UMI counts $+1$) transformation and to perform a nonparanormal transformation (Liu et al., 2009) using the function `huge.npn()` in the package `huge` on the continuized data to make it Gaussian because the

transformation procedure preserves the underlying network structure. Then, we performed each approach in `SILGGM` under global inference with FDR control at the 0.01 level. As comparison studies, we also applied the graphical Lasso (GLasso) using the package `huge`, the marginal correlation-based approach with the Pearson's correlation (PearsonCorr) and the maximum likelihood estimation (MLE) of the partial correlation by directly inverting sample covariance matrix to the same transformed data set. GLasso was run with the default parameters except using the rotational information criterion (Zhao et al., 2012; Lysen, 2009) for tuning selection. Since GLasso only provides point estimates, a non-zero partial-correlation estimate here implies a conditional dependence between the gene pair. For PearsonCorr and MLE, we used the same thresholding procedure among the other approaches in `SILGGM` to control FDR at the 0.01 level based on the z-scores of the Fisher z-transformation of Pearson's correlation and the z-scores of MLEs on all the gene pairs.

Motivated by Jia et al. (2017), we applied the power law (Clauset et al., 2009; Adamic et al., 2001) to evaluate the performance of the overall network structure inferred by the different approaches. The power law illustrates the relationship $p(m) \propto m^{-\lambda}$ for some positive λ. Here, m refers to the node degree, and $p(m)$ denotes the probability of the m-degree nodes. Many studies have indicated that biological networks are scale-free, and the node degrees possess a power-law distribution (Barabási and Albert, 1999; Barabasi and Oltvai, 2004; Almaas and Barabási, 2006; Lima-Mendez and van Helden, 2009). The $\log 2 - \log 2$ plots of degree distribution of inferred networks are shown in Figure 2.4, where the blue curves are fitted by the `R` function `lowess()`. All the approaches in our package `SILGGM` fit the power-law relationship well, but Glasso, PearsonCorr and MLE do not. Even if $n > p$ in this data set, the values of n and p share the same order such that MLE becomes unstable and increases bias of estimation. Thus, all the inferred network structures by `SILGGM` are biologically meaningful and much more reliable. Furthermore, we can see that the performance of the other three methods based on the nodewise Lasso-type regressions in `SILGGM` is even better than that of D-S_GL since the plot of D-S_GL shows some noise in the tail.

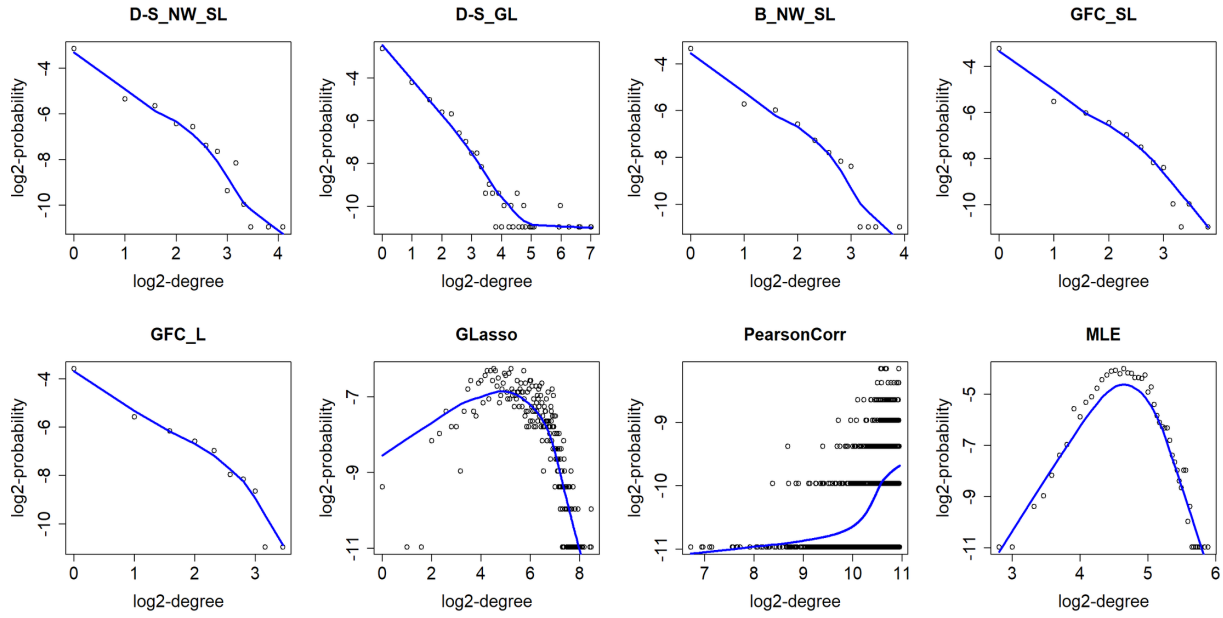

Figure 2.4: The $\log 2 - \log 2$ plots of degree distribution of inferred networks by the different approaches.

2.4 Conclusion and discussion

The source code of the package and a complete reference manual including dependencies, usage of all package functions and associated examples are freely available via `CRAN` at `https://cran.r-project.org/package=SILGGM`. The details of package installation are described in Appendix A.5.

The package `SILGGM` is computationally efficient compared to the `MATLAB` implementation of GFC_L and the `R` package `FastGGM`. Since `R` is a publicly free platform and has been more widely used in biological research compared to `MATLAB` which is a piece of commercially licensed software and has less accessibility to biologists, the `R` platform-based `SILGGM` will play a more important role in accelerating the biological gene network studies. `SILGGM` is also statistically efficient with both individual and global inference due to the theoretical justification of the four approaches and the validation of estimation accuracy in simulation

studies. The analytical results from the single-cell data with Pan T cells further reflect the statistical efficiency of `SILGGM` since inferred gene networks are more reliable. Moreover, the comprehensiveness of `SILGGM` allows users to have more flexible choices of methods depending on the specific purpose of their study. Due to its computational feasibility, analytical reliability in results and methodological comprehensiveness, `SILGGM` can become a valuable and powerful tool to a wide range of biological researchers for high-dimensional or even whole genome-wide co-expression network analysis.

In practice, users have flexible options on the approaches provided by `SILGGM` with respect to the specific purpose of their study. In a whole genome-wide study which is based on global inference of all gene pairs, GFC_L is the one we recommend when n is small (e.g. $n = 100$) because the tuning selection in GFC_L is beneficial for FDR control. When n becomes larger (e.g. $n = 800$) but may be still relatively small to p, users can choose any of the four approaches due to their similar performance among the different settings (see Appendix A.4 for more details). If the study purpose is to evaluate a small set of genes such as certain gene pathways that contribute to an important biological mechanism or a particular gene such as a hub gene that is closely related to a specific disease, among the inference results from all gene pairs, we recommend users choosing B_NW_SL since it provides confidence intervals in addition to p-values and its performance of confidence intervals is always good in the different settings (see Appendix A.3 for detailed comparisons). In a very large-scale setting with p increased to a ten thousand level, D-S_NW_SL is an alternative to save running time. Alternatively, if only the information of p-values is needed, we also recommend GFC_SL or GFC_L.

Besides high-dimensional microarray and bulk RNA-seq data, we intend to promote the application of `SILGGM` to single-cell RNA-seq data with both large n and p. The data sets from single-cell RNA-seq have substantial advantages over the ones from population-level microarray or bulk RNA-seq for us to explore the structure of a gene co-expression network due to larger sample sizes (Macosko et al., 2015) and inherent cell-to-cell variability. According to Wills et al. (2013), the gene network from a single-cell study is able to further reveal potential functionally-related gene pairs which are masked from the bulk sequencing.

3.0 Inference of Large Modified Poisson-Type Graphical Models: Application to RNA-Seq Data in Childhood Atopic Asthma Studies

3.1 Introduction

Recent developments of high-throughput sequencing technologies have generated unprecedented amounts of RNA-seq data for transcriptomics. Network studies of conditional dependency among genes provide new insights to understand a complex biological process or disease.

Gaussian graphical model (GGM) has been widely used in characterizing the conditional relationships among genes in a biological network. However, discrete omics data sets from the next generation sequencing technology are common because the count values are usually used to quantify the genetic or genomic information. One typical example is the bulk RNA-seq data which summarizes the expression of each gene using the number of counts mapped to it. Another example is the droplet-based single-cell RNA-seq data which quantifies the cell-level gene expression with unique molecular identifiers (UMIs) (Islam et al., 2014), a direct counting of transcript copies. Therefore, the use of GGM on those non-Gaussian discrete-type data requires a continuous transformation, for example, using the fragments per kilo base of transcript per milliona (FPKM) or a log transformation on the count values. Converting count values into continuous values tends to alter their biological meanings with the straightforward interpretation and sometimes can be inappropriate (Zwiener et al., 2014). Poisson distribution, however, is a popular choice and has been shown more reasonable than using FPKM in modeling the count data (Anders and Huber, 2010). To describe the conditional dependency among genes from count-valued omics data, Besag (1974) proposed a natural extension of the univariate Poisson model to a multivariate case, and Yang et al. (2015) further extended this to a general graphical model setting called the Poisson graphical model (PGM). Moreover, three modified Poisson-type graphical models: the truncated PGM (TPGM), the sub-linear PGM (SPGM) (Yang et al., 2013) and the square-root PGM (SqrtPGM) (Inouye et al., 2016), were proposed to overcome the major drawback of PGM

for count data modeling (see Section 3.2 for more details).

On the other hand, omics data sets are usually large-scale with the number of genes p allowed to be far larger than the sample size n. To provide reliable estimation for pairwise conditional dependency with its confidence interval and p-value under such settings, statistical inference of high-dimensional GGM has been well developed within the recent seven years, see Liu (2013), Ren et al. (2015), Janková and van de Geer (2015, 2017). Inference of large non-Gaussian graphical models has recently started being paid attention to, see Li et al. (2016) and Cai et al. (2019) for Ising graphical model (IGM). Unfortunately, all current methods based on the three aforementioned high-dimensional modified Poisson-type graphical models only involve estimation, and a unified framework for their statistical inference is still largely unknown.

In this chapter, we intend to propose a new inferential procedure that particularly tailors to the analysis of non-negative, discrete and high-dimensional transcriptomic data based on the modified Poisson-type graphical models. Our motivation comes from the novel RNA-seq gene expression data from the study of the Epigenetic Variation and Childhood Asthma in Puerto Ricans (EVA-PR) aged 9-20 years (Forno et al., 2019). To our knowledge, it is the first study of atopic asthma in nasal epithelium of a large sample of Hispanic children. Further details of the data are deferred to Appendix B.1.

Atopic asthma is one of the most prevalent diseases affecting all ages, but efficient methods for its accurate diagnosis are still under development. Clinicians have recently considered using nasal epithelial samples which are much easier to extract and more disease-relevant to replace white blood cell samples in study of the pathogenesis of atopic asthma. According to Forno et al. (2019), studies in nasal epithelial samples provide promising results in identifying epigenetic variants of childhood atopic asthma in Puerto Ricans. Besides, Pandey et al. (2018) has illustrated differentially expressed genes from transcriptomic profiles that are more closely related to the mechanism of asthma using adult nasal epithelial samples. However, conditional dependence among genes underlying atopic asthma from nasal epithelium is largely unknown, a knowledge of which will no doubt facilitate its accurate diagnosis and the development of its precision medicine.

Inspired by the cutting-edge low-dimensional projection estimator (LDPE) approach in

inference of high-dimensional linear regression (Zhang and Zhang, 2014) and the recent developments in statistical inference of large IGM, we have developed a novel two-step procedure in inference of pairwise conditional dependency from large modified Poisson-type graphical models. The first step involves ℓ_1-penalized node-wise regressions, and the second step is based on a likelihood-based non-linear projection which relies on the graph structure itself and is intrinsically different from the essentially linear projection approach considered in van de Geer et al. (2014) for generalized linear models. For further details, please refer to Section 3.3. From the computational perspective, our method only requires $O(p)$ ℓ_1-penalized regressions due to the novelty of our second step, and is computationally less intensive than the composite likelihood approach and the score matching method proposed in Wang and Kolar (2016) and Yu et al. (2016) respectively targeting on exponential family graphical model inference.

In Section 3.2, we briefly review the properties of three typical modified Poisson-type graphical models. We formally propose a general framework of our procedure with its application to the three modified Poisson-type models in Section 3.3. Section 3.4 includes implementations with selection of tuning parameters. Then, we demonstrate the validity and advantages of our procedure through simulations in Section 3.5 and a real application to the motivating RNA-seq gene expression data of childhood allergic asthma in Section 3.6. We finally conclude with discussion in Section 3.7.

3.2 The modified Poisson-type graphical models

Let $X = (X_1, X_2, ..., X_p)^\intercal$ be a sequence of genes with each $X_i \in \{0, 1, 2, ...\}$ for $i = 1, 2, ..., p$. An undirected Poisson graph $G = (V, E)$ associated with X consists of the node set $V = \{X_1, X_2, ..., X_p\}$ and the edge set $E = \{\text{pairs of } (i, j) \text{ if there is an undirected edge between } X_i \text{ and } X_j\}$. X_i and X_j are conditionally dependent given all the other genes $\{X_r, r \neq i, j\}$ if and only if there is an edge between the two nodes. More formally speaking, the joint distribution of PGM (Yang et al., 2015) is defined as $\mathbb{P}_{\psi,\mathbf{\Theta}}(X) = \exp(\sum_{1\leq i<j\leq p} \theta_{ij} X_i X_j + \sum_{i=1}^{p}(\psi_i X_i - \log(X_i!)) - A(\psi, \mathbf{\Theta}))$, where $A(\psi, \mathbf{\Theta})$ is the log-normalization constant. The

parameter θ_{ij} represents the pairwise strength between nodes X_i and X_j and is encoded in a parameter set $\mathbf{\Theta}$. It is easy to see that X_i and X_j are conditionally independent if and only if $\theta_{ij} = 0$. Therefore, if the two nodes are connected in a graph, we set $\theta_{ij} \neq 0$; otherwise, $\theta_{ij} = 0$. However, PGM can only model negative pairwise dependency (or $\theta_{ij} \leq 0$) if $A(\psi, \mathbf{\Theta}) < +\infty$ is achieved. This fact is due to $x^2/\log(x!) \to +\infty$ as $x \to +\infty$, which can be shown by the Stirling's approximation. To overcome the major constraint of PGM, three modified Poisson-type graphical models are proposed in the literature to allow for both positive and negative dependencies between pairwise nodes.

3.2.1 TPGM

Since the domain of PGM is $\{0, 1, ...\}^p$, the quadratic terms dominate the distribution when count values are very large, which leads to negative dependency. Therefore, a natural remedy is to truncate the domain of each node to a finite level so as to capture both positive and negative dependencies. We can make a reasonable assumption that each node X_i is bounded by a finite number D_i with $i = 1, 2, ..., p$. The joint distribution of TPGM (Yang et al., 2013) is defined as

$$\mathbb{P}_{\psi,\mathbf{\Theta}}(X) \propto \exp\Big(\sum_{i=1}^{p} \psi_i X_i + \sum_{1 \leq i < j \leq p} \theta_{ij} X_i X_j - \sum_{i=1}^{p} \log(X_i!)\Big), \tag{3.1}$$

which has the same format as PGM but with a different log-normalization constant due to the domain $X_i \in \{0, 1, ..., D_i\}$ for $i = 1, 2, ..., p$. We mention that the Ising graphical model (IGM) studied in Ravikumar et al. (2010), Li et al. (2016) and Cai et al. (2019) is a special case of TPGM when $D_i = 1$ for all $i = 1, 2, ..., p$.

3.2.2 SPGM

Unlike TPGM, Yang et al. (2013) also proposed sub-linear PGM (SPGM), an alternative to modify the original PGM without a change on the domain of each node. Specifically, by replacing the linear statistic of each node in $\psi_i X_i$ and $\theta_{ij} X_i X_j$ in (3.1) with a newly-constructed statistic that increases even slower than a linear term, both positive and negative dependencies are allowed in the modified distribution without a domination of quadratic

terms when the value of each node goes to $+\infty$. Therefore, a modified statistic for each node X_i with $i = 1, 2, ..., p$ is defined as

$$S(X_i) = \begin{cases} X_i & \text{if} \quad X_i \leq D_{i0} \\ -\frac{1}{2(D_{i1}-D_{i0})}X_i^2 + \frac{D_{i1}}{D_{i1}-D_{i0}}X_i - \frac{D_{i0}^2}{2(D_{i1}-D_{i0})} & \text{if} \quad D_{i0} < X_i \leq D_{i1} \\ \frac{D_{i0}+D_{i1}}{2} & \text{if} \quad X_i \geq D_{i1}, \end{cases}$$

where D_{i0} and D_{i1} are pre-defined thresholds. The joint distribution of SPGM is thus defined as

$$\mathbb{P}_{\psi,\mathbf{\Theta}}(X) \propto \exp(\sum_{i=1}^{p} \psi_i S(X_i) + \sum_{1\leq i<j\leq p} \theta_{ij} S(X_i)S(X_j) - \sum_{i=1}^{p} \log(X_i!)).$$

SPGM will be close to the original PGM as the upper threshold $D_{i1} \to +\infty$. In particular, SPGM still has a relatively thick tail, which is approachable to the Poisson case.

3.2.3 SqrtPGM

In addition to the aforementioned two models, Inouye et al. (2016) proposed a new class of parametric graphical model called Square Root Graphical Model that allows both positive and negative dependencies. In the Poisson case, SqrtPGM essentially uses square root to replace the linear statistic of each node in $\psi_i X_i$ and $\theta_{ij} X_i X_j$ in (3.1), so the interaction terms become linear to avoid the problem that the quadratic terms dominate the distribution when the value of each node goes to $+\infty$. The joint distribution of SqrtPGM is thus defined as

$$\mathbb{P}_{\psi,\mathbf{\Theta}}(X) \propto \exp(\sum_{i=1}^{p} \psi_i \sqrt{X_i} + \sum_{1\leq i<j\leq p} \theta_{ij} \sqrt{X_i}\sqrt{X_j} - \sum_{i=1}^{p} \log(X_i!)).$$

3.2.4 A unified representation

Let $T(X)$ and $B(X)$ be the sufficient statistic and the base measure respectively. All three modified Poisson-type graphical models can be described in the following generalized joint distribution,

$$\mathbb{P}_{\psi,\mathbf{\Theta}}(X) \propto \exp(\sum_{i=1}^{p} \psi_i T(X_i) + \sum_{1\leq i<j\leq p} \theta_{ij} T(X_i)T(X_j) + \sum_{i=1}^{p} B(X_i)). \tag{3.2}$$

Table 3.1: Sufficient statistics, base measures and domain of X_i in the three models.

Model	$T(X)$	$B(X)$	Domain of X_i
TPGM	X	$-\log(X!)$	$\{0, 1, ..., D_i\}$
SPGM	$S(X)$	$-\log(X!)$	{0,1,...}
SqrtPGM	$\sqrt{X}$	$-\log(X!)$	{0,1,...}

The corresponding sufficient statistic, base measure, domain of X_i for each model are summarized in Table 3.1. Although so far we only define those θ_{ij} for which $i < j$, we set $\theta_{ij} = \theta_{ji}$ to ease our notation whenever $\theta_{ij}, i > j$ is used hereafter.

3.3 Statistical inference of modified Poisson-type graphical models

We first introduce a general two-step procedure to obtain each de-biased estimator $\tilde{\theta}_{ij}$ of conditional dependency between variables X_i and X_j, with applications to three modified Poisson-type graphical models specified later. The goal is to achieve the desired asymptotic normality $(nF_{ij})^{1/2}(\tilde{\theta}_{ij} - \theta_{ij}) \to \mathcal{N}(0, 1)$ with a bounded variance $(F_{ij})^{-1}$ as $(n, p) \to +\infty$ under certain sparsity condition of the graph. In addition, we also introduce a global test to discover the entire graph structure.

3.3.1 The general framework

The first step is to provide a globally good initial estimator $\hat{\theta}_{ij}$ of θ_{ij}, and the second step is to correct the potential bias of $\hat{\theta}_{ij}$ via a variant of LDPE approach (Zhang and Zhang, 2014) to obtain the final estimator $\tilde{\theta}_{ij}$.

Step 1 (Initialization): From the joint distribution (3.2), we can obtain that the conditional distribution of the random variable X_i given all other random variables $X_{-i} =$

$(X_1, ..., X_{i-1}, X_{i+1}, ..., X_p)^\intercal$ belongs to the univariate exponential family. More specifically, the log-likelihood function $\log(\mathbb{P}_{\eta_i}(X_i|X_{-i}))$ can be written as $T(X_i)\mu_i + B(X_i) - f(\mu_i)$ with the parameters $\eta_i = (\psi_i, \theta_i) = (\psi_i, \theta_{i1}, \theta_{i2}, ..., \theta_{i(i-1)}, \theta_{i(i+1)}, ..., \theta_{ip})^\intercal \in \mathbb{R}^p$ and the sufficient statistic $T(X_i)$. In the above equation, we have $\mu_i = \psi_i + \sum_{j \neq i} \theta_{ij} T(X_j)$, and $f(\mu_i)$ is the log-normalization term. To ease notations, we introduce $X^* = (1, X^\intercal)^\intercal$, and $X^*_{-i} = (1, X_1, ..., X_{i-1}, X_{i+1}, ..., X_p)^\intercal$ denotes the subvector of X^* with X_i removed. Similarly we denote $T(X^*_{-i}) = (1, T(X_1), T(X_2), ..., T(X_{i-1}), T(X_{i+1}), ..., T(X_p))^\intercal$. Therefore, we have a simple notation of $\mu_i = T(X^*_{-i})^\intercal \eta_i$.

Due to the sparse structure of a biological network, the whole parameter set $\mathbf{\Theta}$ is commonly assumed sparse in the sense that $\theta_{ij} = 0$ for most pairs of (i, j). Thus, it is natural to estimate $\mathbf{\Theta}$ by solving p ℓ_1-penalized node-wise regressions with $i = 1, 2, ..., p$ based on the conditional distribution $\mathbb{P}_{\eta_i}(X_i|X_{-i})$. Suppose that $X^{(1)}, X^{(2)}, ..., X^{(n)}$ are denoted as n i.i.d. samples from the joint distribution $\mathbb{P}_{\psi, \mathbf{\Theta}}(X)$. $\theta_i \in \mathbb{R}^{p-1}$ can be estimated by solving the following convex optimization problem

$$\hat{\eta}_i = (\hat{\psi}_i, \hat{\theta}_i) = \arg\min_{\eta_i} \{l(\eta_i; \{X^{(k)}\}_{k=1}^n) + \lambda_i ||\theta_i||_1\}, \tag{3.3}$$

where λ_i is a tuning parameter, and the negative joint log-likelihood function $l(\eta_i; \{X^{(k)}\}_{k=1}^n)$ $= -\sum_{k=1}^n \log(\mathbb{P}_{\eta_i}(X_i^{(k)}|X_{-i}^{(k)}))$ takes the form with $\mu_i^{(k)} = T(X_{-i}^{*(k)})^\intercal \eta_i$

$$l(\eta_i; \{X^{(k)}\}_{k=1}^n) = -\sum_{k=1}^n (T(X_i^{(k)})\mu_i^{(k)} + B(X_i^{(k)}) - f(\mu_i^{(k)})).$$

Of note, we only penalize θ_i instead of entire η_i. If one has certain prior knowledge of the biological network such as group or order structure, then the generic ℓ_1 penalty can be replaced by group Lasso or fused Lasso. To demonstrate the general purpose, we only use generic ℓ_1 in our algorithm.

High-dimensional generalized linear model theory suggests that the estimator $\hat{\theta}_i$ has good statistical properties in a global sense under certain regularity conditions. Indeed, the existing method for estimation of entire graph took this approach with theoretical justifications (Yang et al., 2013; Inouye et al., 2016). However, this step itself is not sufficient for our inference purpose due to the bias incurred from the ℓ_1 penalty.

Step 2 (Likelihood-based Bias Correction): In this step, we take a variant of LDPE

approach to correct the bias of $\hat{\theta}_{ij}$ obtained from (3.3) for each pair (i, j) with $i < j$.

The original LDPE (Zhang and Zhang, 2014) can be seen as an extension of the least squares estimator in the classical theory of linear model to the high dimensional settings. We first briefly review the intuition before LDPE. For a low-dimensional linear model with $n < p$, $Y = \mathbf{Z}\beta + \epsilon \in \mathbb{R}^n$, where $Y = (Y^{(1)}, ..., Y^{(n)})^\intercal, \epsilon = (\epsilon^{(1)}, ..., \epsilon^{(n)})^\intercal, \beta = (\beta_1, ..., \beta_p)^\intercal$ and the jth column of $\mathbf{Z}$ is $Z_j = (Z_j^{(1)}, ..., Z_j^{(n)})^\intercal$, the least squares estimator of β_j can be written as a linear projection of Y onto the orthogonal complement of the column space of $\mathbf{Z}_{-j}$. In other words, with a score vector $V = (v_1, v_2, ...v_n)^\intercal$, we have

$$\tilde{\beta}_j^{proj} = \frac{\sum_{k=1}^n v_k Y^{(k)}}{\sum_{k=1}^n v_k Z_j^{(k)}} = \beta_j + \frac{\sum_{k=1}^n v_k \epsilon^{(k)}}{\sum_{k=1}^n v_k Z_j^{(k)}} + \sum_{l \neq j} \frac{\sum_{k=1}^n v_k Z_l^{(k)} \beta_l}{\sum_{k=1}^n v_k Z_j^{(k)}},$$

and when $V = Z_j^\perp$, the third term vanishes, resulting in the desired least squares estimator $\tilde{\beta}_j^{proj} = \beta_j + \sum_{k=1}^n v_k \epsilon^{(k)} / (\sum_{k=1}^n v_k Z_j^{(k)})$. However, in high-dimensional case with $p > n$ and $\mathbf{Z}$ in general position, the orthogonal complement of the column space of $\mathbf{Z}_{-j}$ vanishes and thus the ideal score vector is undefined as $Z_j^\perp = 0$. Following the linear-based projection idea but with a general nonzero score vector V, the third term in the decomposition above presents a nonzero bias. Although we do not know the exact bias term as β is unknown, this analysis of the linear estimator suggests a one-step bias correction with an initial estimator $\hat{\beta}$,

$$\tilde{\beta}_j = \tilde{\beta}_j^{proj} - \sum_{l \neq j} \frac{\sum_{k=1}^n v_k Z_l^{(k)} \hat{\beta}_l}{\sum_{k=1}^n v_k Z_j^{(k)}} = \hat{\beta}_j + \frac{\sum_{k=1}^n v_i (Y^{(k)} - Z^{(k)\intercal} \hat{\beta})}{\sum_{k=1}^n v_k Z_j^{(k)}}.$$

Therefore, with a globally good initial estimator $\hat{\beta}$ and a well-chosen score vector V, it is expected that the bias due to the third term becomes negligible, resulting in an asymptotically normal estimator $\tilde{\beta}_j$.

Since LDPE was originally introduced in linear model, for our model we first linearize the node-wise regression using initial estimators. The parameter of interest θ_i is encoded in μ_i, which corresponds to the sufficient statistic $T(X_i)$. For this reason, we expand the conditional expectation of $T(X_i)$ given X_{-i}, which equals the first derivative of $f(\mu_i)$. To further ease our notations, we denote the first and second derivatives of $f(\cdot)$ by $\dot{f}(\cdot)$ and $\ddot{f}(\cdot)$ respectively. Then at the population level, we have the following decomposition

$$T(X_i) = \mathbb{E}_{\eta_i}(T(X_i)|X_{-i}) + \epsilon_i = \dot{f}(\mu_i) + \epsilon_i = \dot{f}(T(X_{-i}^*)^\intercal \eta_i) + \epsilon_i, \tag{3.4}$$

where ϵ_i has zero mean given X_{-i}. Since $\hat{\eta}_i$ is a globally good estimator of η_i obtained in (3.3), we may take a local Taylor expansion of $\dot{f}(\mu_i)$ about $\hat{\mu}_i$ with $\hat{\mu}_i = T(X^*_{-i})^\intercal \hat{\eta}_i$, that is, $\dot{f}(\mu_i) = \dot{f}(\hat{\mu}_i) + \ddot{f}(\hat{\mu}_i) T(X^*_{-i})^\intercal (\eta_i - \hat{\eta}_i) + Re$, where Re denotes the remainder term. By rearranging terms in above equation, we have the following linearized version of (3.4),

$$T(X_i) - \dot{f}(\hat{\mu}_i) + \ddot{f}(\hat{\mu}_i) T(X^*_{-i})^\intercal \hat{\eta}_i = \ddot{f}(\hat{\mu}_i) T(X^*_{-i})^\intercal \eta_i + (Re + \epsilon_i).$$

We are in the position to apply the projection-based idea to the regression above with i.i.d. observations. Specifically, to obtain a better estimator of θ_{ij} $(i < j)$ given some initial estimator $\hat{\eta}_i = (\hat{\psi}_i, \hat{\theta}_i)$, one needs to find an appropriate score vector $V = (v_1, v_2, ..., v_n)^\intercal \in \mathbb{R}^n$ and apply a one-step bias correction from $\hat{\eta}_i$ as follows

$$\tilde{\theta}_{ij} = \hat{\theta}_{ij} + \frac{\sum_{k=1}^n v_k (T(X_i^{(k)}) - \dot{f}(\hat{\mu}_i^{(k)}))}{\sum_{k=1}^n v_k \ddot{f}(\hat{\mu}_i^{(k)}) T(X_j^{(k)})} \quad (1 \le i < j \le p). \tag{3.5}$$

With some algebra, it is easy to see that the decomposition of the estimation error for $\tilde{\theta}_{ij}$ becomes

$$\begin{aligned} \tilde{\theta}_{ij} - \theta_{ij} =& \frac{\frac{1}{n}\sum_{k=1}^n v_k \epsilon_i^{(k)}}{\frac{1}{n}\sum_{k=1}^n v_k \ddot{f}(\hat{\mu}_i^{(k)}) T(X_j^{(k)})} + \frac{\frac{1}{n}\sum_{k=1}^n v_k Re^{(k)}}{\frac{1}{n}\sum_{k=1}^n v_k \ddot{f}(\hat{\mu}_i^{(k)}) T(X_j^{(k)})} \\ &+ \frac{\frac{1}{n}\sum_{k=1}^n v_k \ddot{f}(\hat{\mu}_i^{(k)}) T(X^{*(k)}_{-\{i,j\}})^\intercal (\eta_{i,-j} - \hat{\eta}_{i,-j})}{\frac{1}{n}\sum_{k=1}^n v_k \ddot{f}(\hat{\mu}_i^{(k)}) T(X_j^{(k)})}. \end{aligned} \tag{3.6}$$

The first term in the right-hand side of (3.6) is denoted as the error term, and the second and the third terms can be regarded as the bias terms. Intuitively, to achieve the inference purpose, we need to pick an V such that the bias terms are asymptotically negligible with respect to the error term while the error term has asymptotic normality with root-n consistency.

To achieve our goal discussed in last paragraph, we look for a population version of V first. Denote $\langle a, b \rangle = \mathbb{E}(a \ddot{f}(\mu_i) b)$. To have a centered asymptotic normality for the first (error) term, it suffices to pick V as any function of X_{-i} as ϵ_i has mean zero given X_{-i}. Indeed, for such a choice, we have $\mathbb{E}(V\epsilon_i) = 0$ with variance $\mathrm{Var}(V\epsilon_i) = \langle V, V \rangle$. Consequently, the entire first term has the desired asymptotic normality. We leave the mathematical derivation of these facts in Appendix B.3. Besides, for the second (bias) term, we expect that with a reasonable choice of V, this term itself is small since it contains a remainder

term Re from the second order Taylor expansion. It remains to find a specific V under this constraint (that is, V is a measurable function of X_{-i}) so that the third (bias) term is negligible. To this end, ideally one needs that $\langle V, T(X_{-\{i,j\}})\rangle$ is a zero vector, where $X_{-\{i,j\}} = (X_1, ..., X_{i-1}, X_{i+1}, ..., X_{j-1}, X_{j+1}, ..., X_p)^\intercal$ and $T(X_{-\{i,j\}}) \in \mathbb{R}^{p-2}$ is defined accordingly. Then it is reasonable to expect the third term is small given that $\hat{\eta}_i$ is a globally good estimator.

The major novelty of our method is on the choice of score vector V. Intrinsic to the graphical model joint distribution (3.2), we propose to choose the population V based on the conditional expectation of X_j given $X_{-\{i,j\}}$ with respect to the inner product $\langle a, b\rangle$ as follows

$$\begin{aligned} V &= T(X_j) - \frac{\mathbb{E}_{\eta_i,\eta_j}(T(X_j)\ddot{f}(\mu_i)|T(X_{-\{i,j\}}))}{\mathbb{E}_{\eta_i,\eta_j}(\ddot{f}(\mu_i)|T(X_{-\{i,j\}}))} \\ &:= T(X_j) - g(T(X_{-\{i,j\}}), \eta_i, \eta_j). \end{aligned} \tag{3.7}$$

It is worthwhile to point out that the conditional expectation function $g(\cdot)$ depends on unknown parameters only through η_i and η_j. In particular, μ_i is known given η_i. By our choice, one can check that $\langle V, m(T(X_{-\{i,j\}}))\rangle = 0$ for any measurable function $m(\cdot)$. Thus, we have achieved that $\langle V, T(X_{-\{i,j\}})\rangle$ is a zero vector, and at the population level, the third (bias) term in (3.6) becomes zero.

Remark 1. *Our choice of the score V is new and intrinsic to the joint likelihood of the specific graphical model. Other methods of bias correction for GLMs were discussed in literature, e.g., van de Geer et al. (2014). The difference is that our construction of V relies on the explicit knowledge of joint conditional distribution of $T(X_j)$ given all other covariates $T(X_{-\{i,j\}})$ in which the conditional expectation of $T(X_j)$ is a non-linear function of $T(X_{-\{i,j\}})$. In contrast, the method proposed in van de Geer et al. (2014) does not impose the specific conditional likelihood pattern but essentially assumes certain linear sparsity structure among all covariates, and thus the proposed score vector is linear. We emphasize that this linear sparsity structure is invalid in general in our graphical model settings. For the reasons above, we call this step of our method the likelihood-based bias correction.*

In the end, given the population expression of V in (3.7), we need to represent the empirical element v_k in the score vector V. Denote the oracle score of the kth observation

as $v_k^{(o)} = T(X_j^{(k)}) - g(T(X_{-\{i,j\}}^{(k)}), \eta_i, \eta_j)$. Here we call $v_k^{(o)}$ the oracle score since η_i, η_j are unknown to us. Those points where $T(X_{-\{i,j\}}^{(k)})$ has explained most variability of $T(X_j^{(k)})$ would receive scores with a small magnitude, and thus play a less significant role in our method. Intuitively, we expect that the first term in the right-hand side of (3.6) dominates $\tilde{\theta}_{ij} - \theta_{ij}$ with our choice of V. One can show that if ignoring the minor difference between $\ddot{f}(\mu_i)$ and $\ddot{f}(\hat{\mu}_i)$, then the asymptotic variance of this first term is F_{ij}^{-1}, where

$$F_{ij} = \mathbb{E}_{\eta_i,\eta_j}((T(X_j) - g(T(X_{-\{i,j\}}), \eta_i, \eta_j))^2 \ddot{f}(\mu_i)) = \langle V, V \rangle.$$

We leave its mathematical derivation in Appendix B.3. Thanks to the globally good estimators $\hat{\eta}_i$ and $\hat{\eta}_j$ obtained from Step 1, it is natural for us to finally use the plugged-in estimator of the oracle, $v_k = T(X_j^{(k)}) - g(T(X_{-\{i,j\}}^{(k)}), \hat{\eta}_i, \hat{\eta}_j)$ in the bias correction step (3.5). We defer the specification of complete implementations in Section 3.4.

Intuitively, we expect that our choice of the non-linear score vector leads to the following asymptotic normality under some regularity conditions

$$\sqrt{nF_{ij}}(\tilde{\theta}_{ij} - \theta_{ij}) \to \mathcal{N}(0, 1).$$

While we do not have access to F_{ij} due to its dependence on unknown parameters, it is natural to replace it by the empirical estimator $\frac{1}{n}\sum_{k=1}^n v_k^2 \ddot{f}(\hat{\mu}_i^{(k)})$. Therefore, we expect the following asymptotic normality result

$$\left(\sum_{k=1}^n v_k^2 \ddot{f}(\hat{\mu}_i^{(k)})\right)^{1/2} (\tilde{\theta}_{ij} - \theta_{ij}) \to \mathcal{N}(0, 1). \tag{3.8}$$

3.3.2 Applications to three modified Poisson-type graphical models

We apply the proposed general framework of statistical inference to the three modified Poisson-type graphical models described in Section 3.2. Our current method for modified-Poisson graphical models can be treated as an extension of Li et al. (2016), which only considered Ising graphical model, a special case of TPGM.

Each node-wise regression in Step 1 for all three models relies on the conditional distribution $\mathbb{P}_{\eta_i}(X_i|X_{-i})$. A more specific representation is provided as

$$\mathbb{P}_{\eta_i}(X_i|X_{-i}) = \frac{\exp\left[T(X_i)(\psi_i + \sum_{j\neq i}\theta_{ij}T(X_j)) + B(X_i))\right]}{\sum_{m=0}^{D_i}\exp\left[T(m)(\psi_i + \sum_{j\neq i}\theta_{ij}T(X_j)) + B(m)\right]}, \tag{3.9}$$

where the corresponding sufficient statistic and the base measure for each model are referred to Table 3.1. The threshold D_i for each X_i is finite in TPGM, while its value becomes $+\infty$ in the other two models.

The bias correction in Step 2 needs the knowledge of $f(\mu_i)$ which is the denominator of the right-hand side of (3.9) for the three models. Specifically, the expression of $f(\mu_i)$ for each of three models is shown in Table 3.2, and the details of corresponding $\dot{f}(\mu_i)$ and $\ddot{f}(\mu_i)$ are referred to Tables B.3 and B.4 in Appendix B.4. Moreover, the expression of v_k in each model is based on the function $g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$. In general, the expression of $g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$ can be presented as

$$\begin{aligned} g(T(X_{-\{i,j\}}), \eta_i, \eta_j) &= \frac{\mathbb{E}_{\eta_i,\eta_j}\left[T(X_j)\ddot{f}(\mu_i)|T(X_{-\{i,j\}})\right]}{\mathbb{E}_{\eta_i,\eta_j}\left[\ddot{f}(\mu_i)|T(X_{-\{i,j\}})\right]} \\ &= \frac{\sum_{k_2=0}^{D_j}(T(k_2)\cdot\ddot{f}(\theta_{ij}T(k_2) + T(X^*_{-\{i,j\}})^\intercal\eta_{i,-j})\cdot Q)}{\sum_{k_2=0}^{D_j}(\ddot{f}(\theta_{ij}T(k_2) + T(X^*_{-\{i,j\}})^\intercal\eta_{i,-j})\cdot Q)} \end{aligned}$$

with

$$\begin{aligned} Q = \sum_{k_1=0}^{D_i} &\exp(T(k_1)T(X^*_{-\{i,j\}})^\intercal\eta_{i,-j} + T(k_2)T(X^*_{-\{i,j\}})^\intercal\eta_{j,-i} \\ &+ B(k_1) + B(k_2) + \theta_{ij}T(k_1)T(k_2)), \end{aligned}$$

where $\eta_{i,-j}$ is the subvector of η_i with θ_{ij} removed and $\eta_{j,-i}$ is the subvector of η_j with θ_{ji} removed. Specific $g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$ for each model is summarized in Table 3.3, and the details of corresponding Q are shown in Table B.5 in Appendix B.4.

3.3.3 Multiple testing with false discovery rate control

If the structure of an overall graph is paid attention to, then there involves a multiple testing problem for all θ_{ij}'s

$$H_0 : \theta_{ij} = 0 \quad \text{vs.} \quad H_1 : \theta_{ij} \neq 0 \quad (1 \leq i < j \leq p) \tag{3.10}$$

Table 3.2: Details of $f(\mu_i)$ in the three models.

Model	$f(\mu_i)$
TPGM	$\log(\sum_{m=0}^{D_i} \exp(m\mu_i - \log(m!)))$
SPGM	$\log(\sum_{m=0}^{+\infty} \exp(S(m)\mu_i - \log(m!)))$
SqrtPGM	$\log(\sum_{m=0}^{+\infty} \exp(\sqrt{m}\mu_i - \log(m!)))$

that tests all pairs simultaneously. One of the most popular large-scale multiple testing procedures is the false discovery rate (FDR) analysis (Benjamini and Hochberg, 1995). It is well known that the false discovery rate $\mathrm{FDR}(t) = \mathbb{E}(\mathrm{FDP}(t))$ is the expectation of false discovery proportion (FDP), which is defined as

$$\mathrm{FDP}(t) = \frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \geq t\}}{\max\{\sum_{1\leq i<j\leq p} I\{|\hat{T}_{ij}| \geq t\}, 1\}}, \tag{3.11}$$

where $\hat{T}_{ij}$ is some generic test statistic for each individual hypothesis with a given threshold level t, $\mathcal{H}_0 = \{(i,j) : i < j, \theta_{ij} = 0\}$ denotes the set of true nulls (i.e., the edge set E), the numerator is the total number of false positives, and the denominator is the total number of rejections. The numerator in (3.11) is generally unknown, but under certain mild sparsity assumption of the underlying graph, one can estimate it by $2(1-\Phi(t))(p^2-p)/2$ as suggested in Liu (2013), where $\Phi(\cdot)$ is a standard normal CDF.

The test statistic in our case is $\hat{T}_{ij} = (\sum_{k=1}^{n} v_k^2 \ddot{f}(\hat{\mu}_i^{(k)}))^{\frac{1}{2}} \tilde{\theta}_{ij}$, a standardized version of $\tilde{\theta}_{ij}$. Following the idea in Liu (2013), we set a pre-defined level of FDR as $0 < \alpha < 1$ and choose the threshold of the test statistic as

$$\hat{t} = \inf\left\{0 \leq t \leq 2\sqrt{\log p} : \frac{2(1-\Phi(t))(p^2-p)/2}{\max\{\sum_{1\leq i<j\leq p} I\{|\hat{T}_{ij}| \geq t\}, 1\}} \leq \alpha\right\}. \tag{3.12}$$

Table 3.3: Details of $g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$ in the three models.

Model	$g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$
TPGM	$\frac{\sum_{k_2=0}^{D_j}(k_2 \cdot \ddot{f}(\theta_{ij}k_2 + X^{*\intercal}_{-\{i,j\}}\eta_{i,-j}) \cdot Q)}{\sum_{k_2=0}^{D_j}(\ddot{f}(\theta_{ij}k_2 + X^{*\intercal}_{-\{i,j\}}\eta_{i,-j}) \cdot Q)}$
SPGM	$\frac{\sum_{k_2=0}^{+\infty}(S(k_2) \cdot \ddot{f}(\theta_{ij}S(k_2) + S(X^*_{\{-i,j\}})^\intercal \eta_{i,-j}) \cdot Q)}{\sum_{k_2=0}^{+\infty}(\ddot{f}(\theta_{ij}S(k_2) + S(X^*_{-\{i,j\}})^\intercal \eta_{i,-j}) \cdot Q)}$
SqrtPGM	$\frac{\sum_{k_2=0}^{+\infty}(\sqrt{k_2} \cdot \ddot{f}(\theta_{ij}\sqrt{k_2} + \sqrt{X^*_{-\{i,j\}}}^\intercal \eta_{i,-j}) \cdot Q)}{\sum_{k_2=0}^{+\infty}(\ddot{f}(\theta_{ij}\sqrt{k_2} + \sqrt{X^*_{-\{i,j\}}}^\intercal \eta_{i,-j}) \cdot Q)}$

We reject H_0 in (3.10) if $|\hat{T}_{ij}| \geq \hat{t}$. If no $\hat{t}$ can be obtained, we set $\hat{t} = 2\sqrt{\log p}$ as under null the distribution of each $\hat{T}_{ij}$ is expected to be close to a standard normal such that the largest magnitude of those $(p^2 - p)/2$ statistics is no larger than $2\sqrt{\log p}$ with probability going to 1. Although we do not provide any theory, we comment that with the constraint $t \leq 2\sqrt{\log p}$ in (3.12), the weak dependency among all $\hat{T}_{ij}$'s will not influence the FDR control asymptotically. For further theoretical justification, please refer to Liu (2013).

3.4 Implementations for graph inference

3.4.1 Algorithm

Step 1 involves a node-wise ℓ_1-penalized regression for each node X_i on all other nodes X_{-i}, see Ravikumar et al. (2010), Yang et al. (2013). Here, the intercept ψ_i is excluded from the penalization. The total computational complexity of **Step 1** is essentially equivalent to solving $O(p)$ ℓ_1-penalized regression problems. Each problem can be solved efficiently using the proximal gradient descent. Set $T(\boldsymbol{X}^*)$ as the $n \times (p+1)$ matrix with the kth row being $T(X^{*(k)})^\intercal$ for $k = 1, ..., n$. In addition, all the regressions rely on a single matrix with

the ijth element being the inner product between the ith and jth columns of $T(\boldsymbol{X}^*)$ which includes $O(np^2)$ operations. The pre-calculation of this matrix can help avoid its repetitive calculation.

The bias correction for the parameter set $\boldsymbol{\Theta}$ in **Step 2** has a total of $O(p^2)$ loops, and each loop for $\hat{\theta}_{ij}$ involves the calculation of inner products $\sum_{r\neq\{i,j\}}\hat{\theta}_{ir}T(X_r^{(k)})$ and $\sum_{r\neq\{i,j\}}\hat{\theta}_{jr}T(X_r^{(k)})$ for all $k=1,...,n$. The naïve matrix calculation tends to increase the computational complexity to $O(np^3)$. To simplify the computational steps, we pre-calculate the inner product between each $T(X_{-i}^{*(k)})$ and the initial estimators $\hat{\eta}_i$ and save the value in a prediction matrix which can be repetitively used for the inner product calculation within each loop. It can be seen that the pre-calculation of the prediction matrix helps reduce the computational complexity of these inner products to $O(np^2)$.

Besides the aforementioned implementations with high computational convenience, all the algorithms are achieved with the `Rcpp` library. Due to the lack of closed-form expressions for the normalization terms in the conditional distributions of SPGM and SqrtPGM, the numerical approximations that require a summation from zero to a large number are highly involved with many loop operations. The usage of `Rcpp` library, which incorporates the efficient `C++` code under the `R` environment, helps lower the computational burden for loops. In the end, we summarize all the steps of our two-step inference method in Algorithm 1.

3.4.2 Selection of tuning parameters

The tuning parameter λ_i in (3.3) controls the neighborhood sparsity of each node X_i or the number of edges extending out from X_i, so we need to select a sequence of λ_i with $i=1,2,...,p$ for initial estimators in **Step 1**. According to the different purpose of inference, we provide two ways for selection of tuning parameters.

We at first focus on the inference of each individual θ_{ij}. The extended BIC (EBIC) criterion has been well studied under the regime of high-dimensional graphical models (Barber and Drton, 2015). We write EBIC for each regression as follows,

$$\text{EBIC}_\gamma(J) = 2l(\eta_i; \{X^{(k)}\}_{k=1}^n) + |J|(\log(n) + 2\gamma\log(p-1)), \tag{3.13}$$

where $l(\eta_i; \{X^{(k)}\}_{k=1}^n) = -\sum_{k=1}^n \log(\mathbb{P}_{\eta_i}(X_i^{(k)}|X_{-i}^{(k)}))$, $|J|$ is the cardinality of $J = \{j : j \neq i \text{ and } \hat{\theta}_{ij} \neq 0\}$, and some universal $\gamma \geq 0$. Following the suggestion in Barber and Drton (2015), we set $\gamma = 0.5$ as the default value in real implementations and select the tuning parameters that minimize (3.13).

For multiple testing, the tuning parameters are chosen as in Liu (2013) to guarantee $2(1 - \Phi(t))(p^2 - p)/2$ as close to $\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \geq t\}$ as possible. We leave further details in Appendix B.5.

Algorithm 1 Statistical inference of the modified Poisson-type graphical models

- **Step 1: Initialization**
  1. Pre-calculate and save the inner product matrix, where the ijth element denotes the inner product between the column i and column j of $T(\boldsymbol{X}^*)$.
  2. For each X_i, $i = 1, 2, ..., p$, do node-wise ℓ_1-penalized regression (3.3).
  3. Obtain each initial estimator $\hat{\theta}_{ij}$ for **Step 2**.
- **Step 2: Likelihood-based Bias Correction**
  1. Pre-calculate and save the $n \times p$ prediction matrix $\boldsymbol{M}$, where each element $\hat{\mu}_i^{(k)}$ denotes the inner product of $T(X_{-i}^{*(k)})$ and $\hat{\eta}_i$ with $k = 1, 2, ..., n$ and $i = 1, 2, ..., p$.
  2. For each $i = 1, 2, ..., p-1$, do:
     (a.) Calculate $\dot{f}(\hat{\mu}_i^{(k)})$ and $\ddot{f}(\hat{\mu}_i^{(k)})$ in (3.5) with $k = 1, ..., n$.
     (b.) With each fixed i, for each $j = i+1, i+2, ..., p$, do:
        (i.) Calculate $q_1^{(k)} = \hat{\mu}_i^{(k)} - \hat{\theta}_{ij} T(X_j^{(k)})$ and $q_2^{(k)} = \hat{\mu}_j^{(k)} - \hat{\theta}_{ji} T(X_i^{(k)})$ with $k = 1, ..., n$.
        (ii.) Plug $q_1^{(k)}$ and $q_2^{(k)}$ into (3.7) to obtain the score vector V.
        (iii.) Generate the final estimator $\tilde{\theta}_{ij}$ in (3.5).
  3. Estimate the standard deviation, the 95% confidence interval, the p-value and the z-score for each $\tilde{\theta}_{ij}$.

To ensure the validity of EBIC to select tuning parameters, we further performed a comprehensive study of hyperparameter selection. We compared inferred networks between the proposed method and the sole estimation procedure with only node-wise ℓ_1-penalized

regressions using EBIC and cross validation under both simulation settings and the real data application in Sections 3.5-3.6. It is intriguing to notice that the proposed method is robust to different hyperparameter selection methods, while the sole estimation is very sensitive to various model selection criteria. Moreover, the proposed method can reach a better balance between false and true discovers than the sole estimation based on different hyperparameter selection methods. More details are left in Appendix B.6.

3.5 Simulations

To show the validity of the proposed two-step procedure, we evaluated its performance from two-folds: 1. Asymptotic normality; 2. False discovery rate control for multiple testing. We considered four different graph settings: (a) the Chain graph with two consecutive nodes arranged to be connected, (b) the Grid graph (4-nearest neighbor graph) with nodes arranged to a lattice with maximal degree $d = 4$, (c) the Erdős-Rényi (E-R) random graph with average node degree $d = 4$, and (d) the Scale-free network (Barabási and Albert, 1999). We generated random samples from the three modified Poisson-type models via Gibbs sampling (Zhang et al., 2017). The first 5000 draws were discarded in the burn-in period. Then, we took one sample every 100 draws to guarantee independence. In addition, we also compared the proposed method to the popular Gaussian graphical model estimation with FDR control using Lasso (GFC_L) (Liu, 2013) by evaluating their performance on simulated RNA-seq data.

3.5.1 Asymptotic normality

The lattice size of Grid graph is $\sqrt{p} \times \sqrt{p}$ here. For each given graph, we generated 100 data sets with $n = 300$, $p = 100$ and 400 respectively from the three models with each of nonzero entries drawn randomly from the set $(-0.4, -0.3, -0.2, -0.1, 0.1, 0.2, 0.3, 0.4)$. Other parameter details in the three models are described as below,

- TPGM: The intercept term $\psi_i = 0$, and the threshold value $D_i = 3$.

- SPGM: The intercept term $\psi_i = -0.5$, and two threshold values $D_{i0} = 3$ and $D_{i1} = 6$.
- SqrtPGM: The intercept term $\psi_i = 0$.

The proposed estimates of each pairwise parameter were obtained based on Algorithm 1 with EBIC criterion for selection of tuning parameters in all the graph settings. Figure 3.1 shows the histograms of randomly selected entries that cover all possible values of true parameters from the three modified Poisson-type graphical models under high-dimensional settings with $p = 400$ for Scale-free graph. Each red curve is denoted as the approximate Gaussian density of a particular entry. It can be seen that the histograms of each entry match with the corresponding normal distribution very well. The histograms for Scale-free graph with $p = 100$ and the other three graph settings are referred to Figures B.5-B.10 in Appendix B.7.1. Similarly, all the histograms of estimated entries are also in good accordance with their corresponding normal distributions.

The $(1 - \alpha)$ confidence interval for each θ_{ij} can be derived straightforwardly from the asymptotic normality in (3.8):

$$(\tilde{\theta}_{ij} - z_{\alpha/2}(\sum_{k=1}^{n} v_k^2 \ddot{f}(\hat{\mu}_i^{(k)}))^{-1/2}, \quad \tilde{\theta}_{ij} + z_{\alpha/2}(\sum_{k=1}^{n} v_k^2 \ddot{f}(\hat{\mu}_i^{(k)}))^{-1/2}),$$

where $z_{\alpha/2}$ is the z-score with the right tail probability equal to $\alpha/2$, i.e., $P(\mathcal{N}(0,1) > z_{\alpha/2}) = \alpha/2$.

In addition, we also evaluated the performance of empirical coverage probabilities of the 95% confidence intervals of θ_{ij}'s to demonstrate the validity of our inference results. Considering the sparse structures of both graph settings, we separated all θ_{ij}'s into two sets: the edge S_0 and non-edge S_0^c:

$$S_0 = \{(i,j) : \theta_{ij} \neq 0\}, \qquad S_0^c = \{(i,j) : \theta_{ij} = 0\}.$$

Then, based on all the estimates $\tilde{\theta}_{ij}$'s, the average empirical coverage probabilities of the 95% confidence intervals were evaluated in S_0 and S_0^c respectively. Table 3.4 reports the medians (standard deviations) of average empirical coverage probabilities of the 95% confidence intervals over 100 replications for $p = 400$. As we can see, all results are close to 0.95, the target confidence level. Additional simulation results towards individual inference

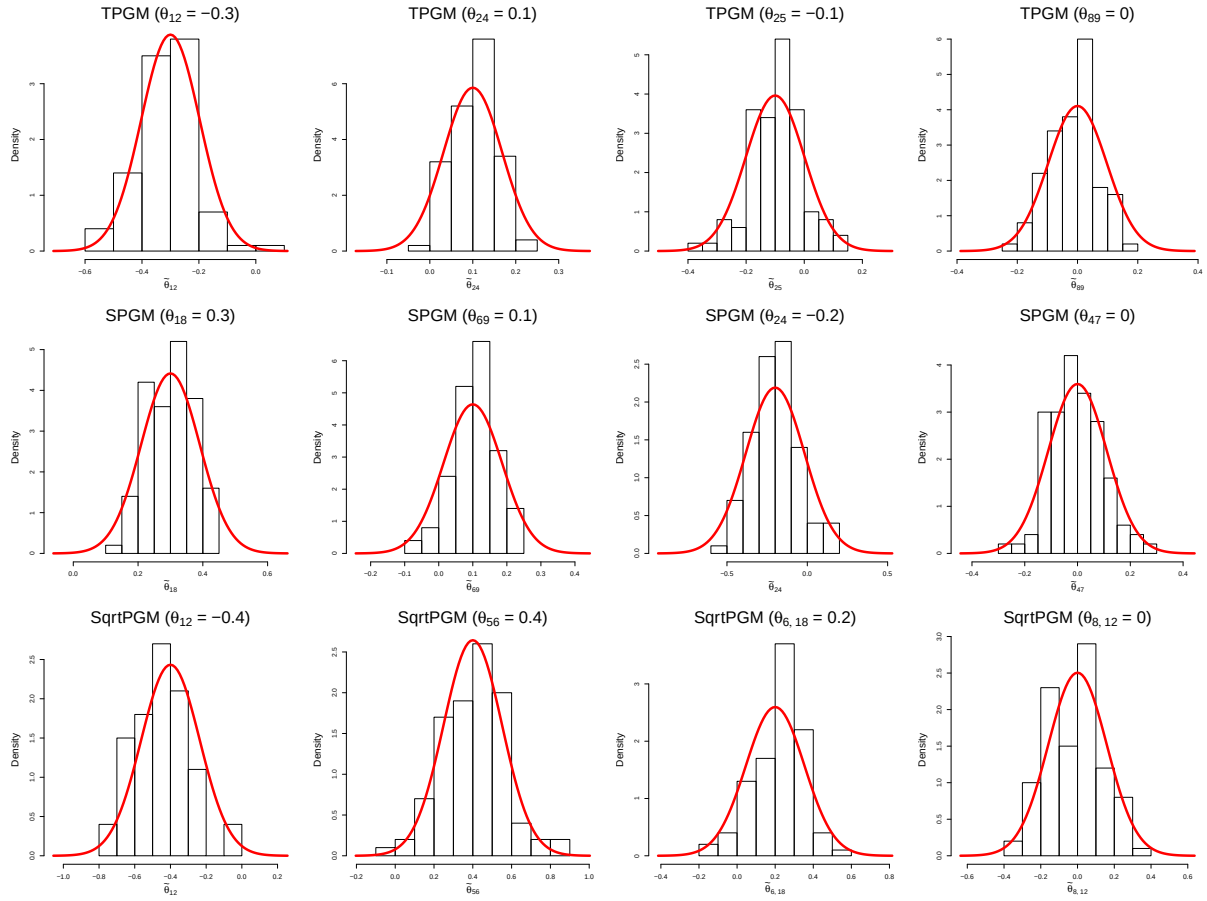

Figure 3.1: Histograms of the estimated pairwise entries for $p = 400$ from the three models in Scale-free graph.

Table 3.4: Medians (standard deviations) of empirical coverage probabilities of the 95% confidence intervals in S_0 and S_0^c with $p = 400$.

	S_0				S_0^c			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 300$,	$p = 400$			
TPGM	0.9436	0.9352	0.9284	0.9311	0.9511	0.9508	0.9497	0.9501
	(0.0118)	(0.0087)	(0.0081)	(0.0117)	(0.0010)	(0.0010)	(0.0012)	(0.0009)
SPGM	0.9499	0.9153	0.8910	0.9135	0.9528	0.9536	0.9519	0.9505
	(0.0118)	(0.0087)	(0.0100)	(0.0109)	(0.0010)	(0.0011)	(0.0011)	(0.0009)
SqrtPGM	0.9524	0.9512	0.9512	0.9549	0.9512	0.9501	0.9501	0.9510
	(0.0110)	(0.0087)	(0.0087)	(0.0108)	(0.0009)	(0.0009)	(0.0010)	(0.0008)

with $p = 100$, a smaller sample size ($n = 150$ and 100) and using simulation settings in Section 3.5.2 are summarized in Tables B.11-B.14 in Appendix B.7.2.

3.5.2 False discovery rate control for multiple testing

To evaluate the performance of our estimates for multiple testing with false discovery rate (FDR) control, we considered the four graphs with a two-block structure. More specifically, the first half of nodes form one block, leaving the remaining nodes as another block. Two cases were evaluated: $p = 200$ and 400. The detailed parameter settings are described as below,

- TPGM: Each of non-zero entries is randomly drawn: (i.) either -0.3 or 0.3 in Block 1; (ii.) either -0.4 or 0.4 in Block 2. For both blocks, each intercept term $\psi_i = -0.5$, and each threshold value $D_i = 3$.
- SPGM: Each of non-zero entries is randomly drawn: (i.) either -0.3 or 0.3 in Block 1; (ii.) either -0.4 or 0.4 in Block 2. For two blocks, the intercept term ψ_i is: (i.) -0.5 for

Chain and Scale-free graphs; (ii.) -1 for Grid and E-R graphs. Two threshold values $D_{i0} = 2$ and $D_{i1} = 5$.

- SqrtPGM: Each of non-zero entries is randomly drawn: (i.) either -0.6 or 0.6 in Block 1; (ii.) either -0.9 or 0.9 in Block 2. The intercept term $\psi_i = 0$.

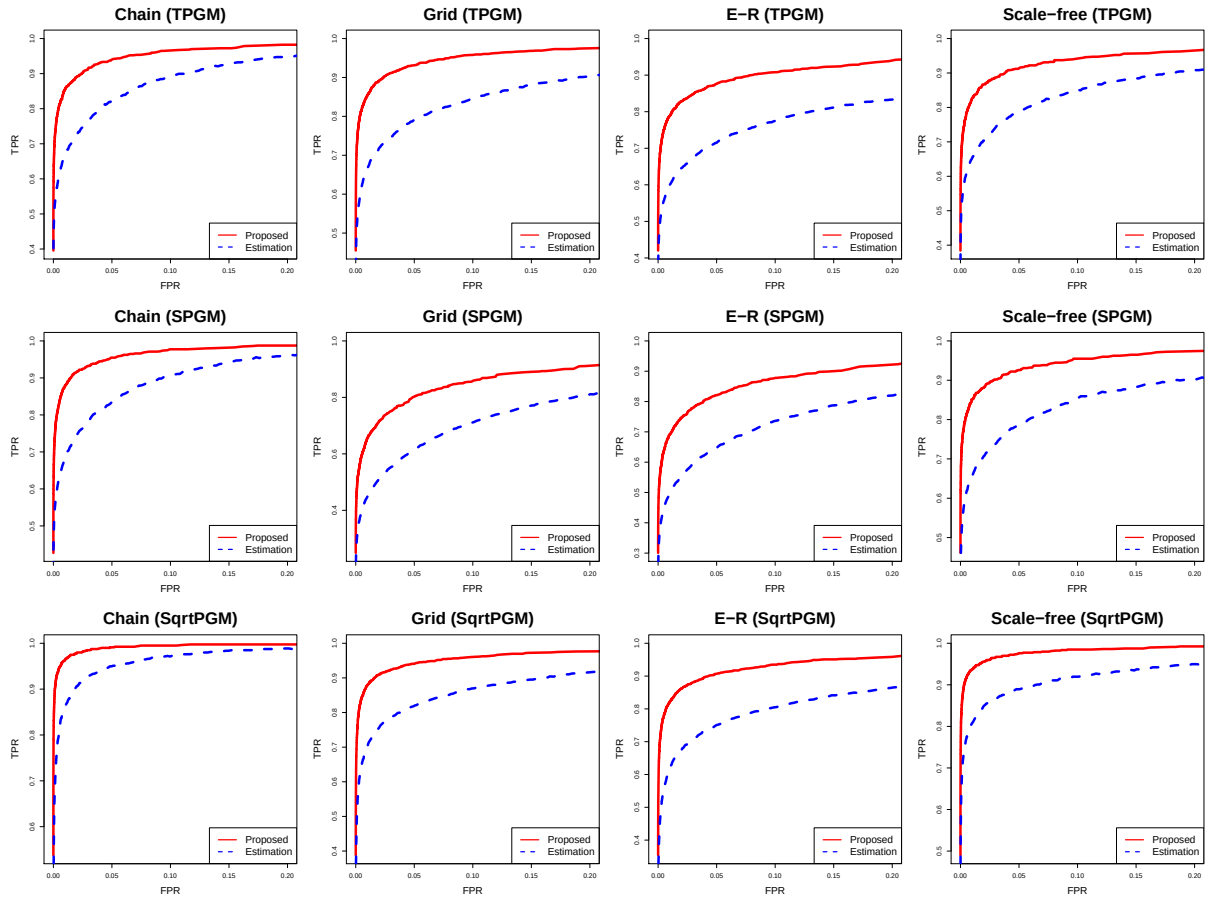

Figure 3.2: ROC curves based on TPRs and FPRs for the proposed inferential procedure and the sole estimation in the case of $p = 400$.

For each of the two cases, we generated 100 data sets with $n = 400$. We investigated the performance of our procedure by evaluating true positive rate (TPR) and false positive rate (FPR) over a range of FDR control levels. Here, we used the tuning selection scheme described in Appendix B.5 for multiple testing. To compare, we also applied the sole estimation procedure with node-wise ℓ_1-penalized regressions in each same data set through a

range of regularization parameters. The medians of TPRs and FPRs at each cut-off over 100 replications from the two procedures are presented in the receiver operating characteristic (ROC) curves for $p = 400$, as shown in Figure 3.2. It can be seen that all curves from the proposed inferential procedure lie above the ones from the sole estimation and show noticeably better performance in detecting true conditional dependency while simultaneously maintaining false discovers at a low level. ROC curves for $p = 200$ which share similar patterns are shown in Figure B.11 in Appendix B.7.3.

Furthermore, we report the medians (standard deviations) of empirical FDRs with pre-specified levels 0.1 and 0.2 for both $p = 200$ and 400 in Table 3.5. The medians (standard deviations) of their corresponding power values are shown in Table 3.6. The empirical FDRs are well controlled at the desired levels with a relatively good performance of power. Additional simulation results towards global inference with a smaller sample size ($n = 150$) are summarized in Tables B.15-B.16 in Appendix B.7.3.

3.5.3 Evaluation on simulated RNA-seq data

We further evaluated the performance of the proposed method by comparing it to GFC_L on simulated RNA-seq data. Even though there is a broad existing literature on how to simulate an RNA-seq data set, see Gerard (2020) for example, there is barely any method that can incorporate the structure of conditional dependence among genes. Therefore, we proposed a new procedure to simulate an RNA-seq data set by incorporating conditional dependence among genes as below,

(a) **(Incorporation of conditional dependence)** A simulated normalized count-valued RNA-seq data set with n samples and p genes is generated via Gibbs sampling from SqrtPGM with Scale-free graph.

(b) **(Pseudo-random number addition)** A randomly generated number from uniform distribution between 0 and 1 is added to each element of the simulated count data to ensure randomness.

(c) **(Inverse power transform)** Inverse power transform is performed on the data matrix with some value of β from a real data application.

Table 3.5: Medians (standard deviations) of empirical false discovery rates.

	$\alpha = 0.1$				$\alpha = 0.2$			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 400,\ p = 200$				
TPGM	0.0892	0.0948	0.0939	0.0939	0.1777	0.1794	0.1792	0.1856
	(0.0277)	(0.0188)	(0.0189)	(0.0246)	(0.0397)	(0.0260)	(0.0277)	(0.0338)
SPGM	0.0858	0.0840	0.0964	0.0938	0.1744	0.1623	0.1760	0.1895
	(0.0259)	(0.0209)	(0.0222)	(0.0253)	(0.0327)	(0.0261)	(0.0299)	(0.0361)
SqrtPGM	0.0884	0.0903	0.0955	0.0956	0.1762	0.1784	0.1721	0.1810
	(0.0237)	(0.0323)	(0.0277)	(0.0262)	(0.0327)	(0.0318)	(0.0368)	(0.0346)
				$n = 400,\ p = 400$				
TPGM	0.0940	0.1007	0.1120	0.0937	0.1866	0.1931	0.2052	0.1939
	(0.0142)	(0.0128)	(0.0187)	(0.0225)	(0.0205)	(0.0204)	(0.0213)	(0.0251)
SPGM	0.0998	0.1154	0.1159	0.1054	0.1852	0.2032	0.2145	0.2092
	(0.0212)	(0.0173)	(0.0141)	(0.0242)	(0.0218)	(0.0231)	(0.0201)	(0.0249)
SqrtPGM	0.0986	0.0907	0.0997	0.0976	0.2016	0.1818	0.1885	0.2021
	(0.0197)	(0.0117)	(0.0123)	(0.0176)	(0.0280)	(0.0174)	(0.0152)	(0.0254)

Table 3.6: Medians (standard deviations) of power values for corresponding FDR control levels.

	$\alpha = 0.1$				$\alpha = 0.2$			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 400,\ p = 200$				
TPGM	0.7222	0.7116	0.7867	0.7727	0.7828	0.7619	0.8329	0.8131
	(0.0236)	(0.0215)	(0.0197)	(0.0230)	(0.0234)	(0.0204)	(0.0161)	(0.0215)
SPGM	0.7172	0.4881	0.4746	0.7449	0.7677	0.5556	0.5278	0.7879
	(0.0226)	(0.0256)	(0.0228)	(0.0247)	(0.0241)	(0.0238)	(0.0241)	(0.0238)
SqrtPGM	0.8838	0.7460	0.6613	0.8939	0.9192	0.8069	0.7295	0.9242
	(0.0238)	(0.0227)	(0.0336)	(0.0198)	(0.0197)	(0.0218)	(0.0327)	(0.0167)
				$n = 400,\ p = 400$				
TPGM	0.6357	0.7131	0.6463	0.6131	0.6910	0.7592	0.6920	0.6633
	(0.0177)	(0.0114)	(0.0123)	(0.0164)	(0.0183)	(0.0113)	(0.0122)	(0.0168)
SPGM	0.6646	0.4347	0.5087	0.6796	0.7186	0.4908	0.5538	0.7236
	(0.0163)	(0.0161)	(0.0157)	(0.0166)	(0.0188)	(0.0150)	(0.0154)	(0.0179)
SqrtPGM	0.8492	0.6966	0.6467	0.8065	0.8920	0.7652	0.7107	0.8492
	(0.0185)	(0.0161)	(0.0141)	(0.0185)	(0.0157)	(0.0146)	(0.0138)	(0.0185)

(d) **(Final count generation)** Elements are rounded down to its nearest integer to obtain the final simulated RNA-seq data set.

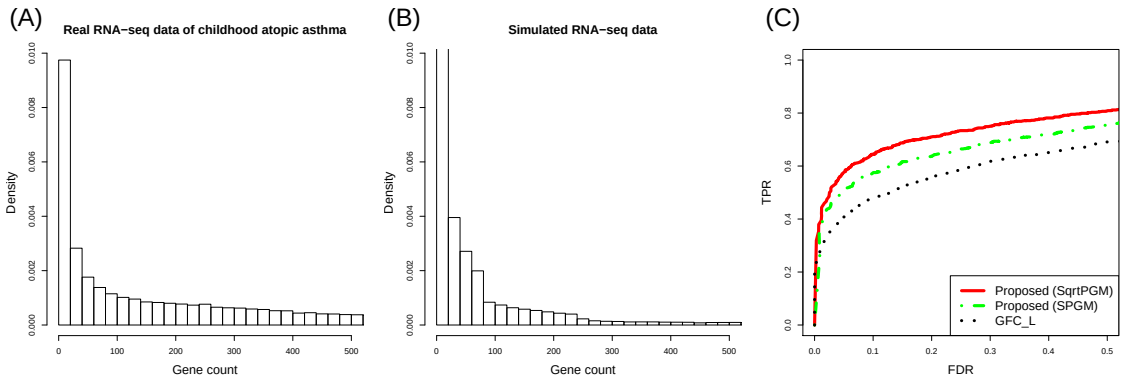

Figure 3.3: (A) Histogram of real RNA-seq data of childhood atopic asthma. (B) Histogram of typical simulated RNA-seq data. (C) ROC-type curves based on TPRs and FDRs for the proposed inferential procedure (SqrtPGM and SPGM) and GFC_L on simulated RNA-seq data.

Here, we considered the two-block Scale-free graph in Section 3.5.2 to depict the conditional dependence among genes because a graph with power-law topology generally illustrates the structure of a real biological network (Barabási and Albert, 1999), and SqrtPGM was adopted in (a) due to its more flexibility than SPGM and TPGM which usually need pre-defined thresholds. The procedure was motivated by the pre-processing steps in Allen and Liu (2013) on original RNA-seq data which mainly rely on a power transform X^{β} for $0 < \beta < 1$ to make count values close to some Poisson-type distribution. We intended to perform inverse power transform in (c) with the value of β that can be borrowed from a real data application. For example, the pre-processing on our motivating count-valued RNA-seq data of childhood atopic asthma returned $\beta = 0.2517$, so we took $\beta = 0.2517$ here.

We used the proposed procedure to generate 100 simulated RNA-seq data sets with $n = 300$ and $p = 400$. Figures 3.3(A) and 3.3(B) demonstrate the histograms of count values from the real data set of childhood atopic asthma and a typical simulated RNA-seq data set. As it can be seen, the distribution shape of simulated RNA-seq data is quite

close to that of a real data set which illustrates decaying proportions of large count values. Before implementation of the proposed method, we took a power transformation on the simulated data sets with $\beta = 0.2517$ and rounded down each element of data matrices to its nearest integer to obtain the normalized count-valued data sets. Before implementation of GFC_L, we used a log and nonparanormal transformation (Liu et al., 2009) to continuize and gaussianize the simulated data sets (Jia et al., 2017). We implemented our proposed method using SqrtPGM and SPGM because they are more general Poisson-type distributions than TPGM. For SPGM, we naturally set 0 as the lower bound D_{i0} and count maximum of each column as the upper bound D_{i1}. GFC_L was implemented with the `R` package `SILGGM` (Zhang et al., 2018b).

The evaluation of methods depends on the performance of TPR over a range of varying FDR control levels. The medians of TPRs and FDRs over 100 replications from our approach (SqrtPGM and SPGM) and GFC_L are illustrated in the ROC-type curves in Figure 3.3(C). Both curves from our approach with SqrtPGM and SPGM lie above the one from GFC_L, which indicates that our proposed method is noticeably more capable of capturing built-in features than GFC_L while controlling FDRs around same levels. Furthermore, we also reported all the medians (standard deviations) of empirical FDRs and corresponding power values with pre-specified levels $\alpha = 0.1$ and 0.2 in Table 3.7. The power values from the proposed method with SqrtPGM and SPGM are both greater than those from GFC_L while all the empirical FDRs are very similar. Therefore, our method can capture the built-in features better than GFC_L in terms of all the results from simulated RNA-seq data. Additional comparison with a smaller sample size ($n = 150$) is shown in Table B.17 in Appendix B.7.4.

3.6 Application to RNA-seq data of childhood allergic asthma

We applied our proposed approach to the motivating RNA-seq gene expression data that illustrates the count-valued transcripts of genes from the nasal epithelial cells of $n = 157$ children (62 females and 95 males) with allergic asthma in Puerto Ricans. These children

Table 3.7: Medians (standard deviations) of empirical FDRs and power values from our proposed method (SqrtPGM and SPGM) and GFC_L on simulated RNA-seq data with FDR controlled at levels $\alpha = 0.1$ and 0.2.

Proposed (SqrtPGM)		Proposed (SPGM)		GFC_L	
FDR	Power	FDR	Power	FDR	Power
		$\alpha = 0.1$			
0.1023	0.6445	0.1143	0.5842	0.0993	0.4786
(0.0181)	(0.0259)	(0.0191)	(0.0224)	(0.0250)	(0.0237)
		$\alpha = 0.2$			
0.1965	0.7085	0.2295	0.6508	0.2018	0.5553
(0.0234)	(0.0240)	(0.0208)	(0.0224)	(0.0248)	(0.0236)

have an average age of 15.3 years with a median total IgE (Immunoglobulin E) of 372 IU/mL, which is high due to atopic asthma. More detailed demographic information of these children is deferred to Table B.1 in Appendix B.1. Before using the proposed approach, we normalized the RNA-seq data following the pre-processing steps described in Allen and Liu (2013). The pre-processing steps adjust sequencing depth for all the genes and filter out the genes with low inter-sample variance. Besides, the possible overdispersion and the batch effects in the data have also been adjusted. The normalization was implemented with the `processSeq` function in the `R` package `XMRF` (Wan et al., 2016). After pre-processing, the normalized data is more approachable to a Poisson-type distribution than the original one, see the comparison of histograms in Figure B.12 in Appendix B.8.1. The normalized data includes $p = 500$ genes.

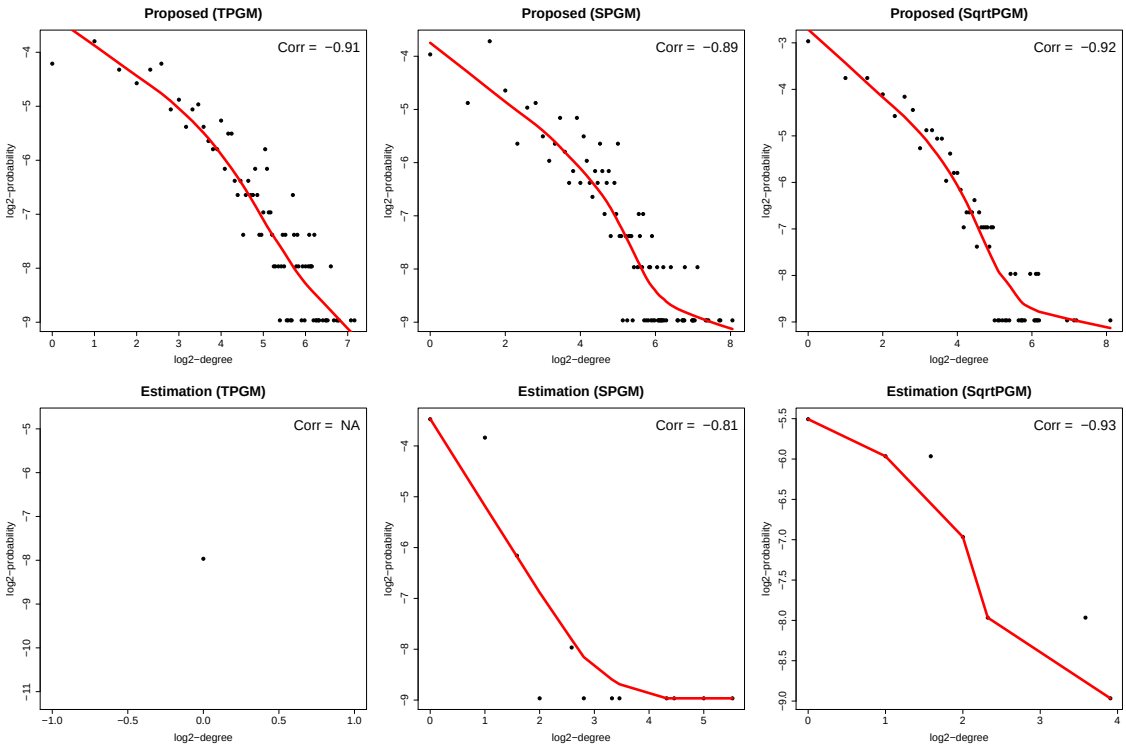

Figure 3.4: The log 2-log 2 plots of degree distribution for the inferred networks (NA: not available).

We inferred gene network using the proposed method in the three models with FDR control at level 0.001. We naturally set count maximum of each column as the upper bound D_i in TPGM and D_{i1} in SPGM respectively, and 0 as the lower bound D_{i0} in SPGM. As comparison studies, we also constructed gene network using only the estimation results from **Step 1** of the procedure based on EBIC criterion. It is well known that biological network usually has a scale-free (or power-law) pattern (Barabási and Albert, 1999; Almaas and Barabási, 2006), that is, $p(\lambda) \propto \lambda^{-\alpha}$, where λ and $p(\lambda)$ are denoted as node degree and its corresponding probability respectively, and α is a positive number. The overall network structure was then evaluated by measuring the correlation between the log 2 of node degree and the log 2 of its corresponding probability. A correlation closer to -1 indicates a better conformation to the power law. Figure 3.4 illustrates the log 2-log 2 plots of node degree distribution for inferred networks and their corresponding correlation measurements.

As it can be seen, the correlation values based on the proposed inferential procedure are all around -0.9 and much closer to -1 in TPGM and SPGM, while the values are comparable with the sole estimation in SqrtPGM. Although the correlation values are still good, the sole estimation generally leads to a much sparser network and fails to capture complex co-expression structures, particularly for TPGM which has a maximum node degree of 1 and barely demonstrates any informative interactions. Additional evaluations of the overall network structure based on the two graphical model methods under normal assumption are shown in Figure B.13 in Appendix B.8.2. Due to their failure to conform the power law with an unreliable inferred network structure, we did not include them for further analysis.

In addition to evaluating the overall network structure, we also studied community structure of all the inferred networks using the eigenspectrum of the modularity matrix (Newman, 2006) so as to explore important gene pathways within the identified gene modules to atopic asthma. Besides the aforementioned methods with three modified Poisson-type models, we included GFC_L and the nonparanormal SKEPTIC estimator (Liu et al., 2012) as comparison studies. To ensure the fairness in comparison, we extracted the same 500 genes from the original data and made a log and nonparanormal transformation to continuize and gaussianize the count values according to Jia et al. (2017) before the use of GFC_L. The nonparanormal transformation was achieved with the `huge.npn()` function in the `R` package `huge` (Zhao et al., 2012), and GFC_L was implemented with the `R` package `SILGGM` with FDR control at level 0.001. For nonparanormal SKEPTIC, we obtained Spearman's rho statistics from the original count-valued data with 500 genes by using the `huge.npn()` function. Then, the graphical Lasso was implemented to estimate networks. The resulting estimated graph was finally selected by the EBIC criterion. Table 3.8 demonstrates the identified big gene modules with a size of at least 30 genes from the inferred gene networks. It can be seen that the proposed method successfully detects 2 to 4 big gene modules among three models, while the sole estimation in TPGM and SqrtPGM, GFC_L and nonparanormal SKEPTIC fail to identify informative ones.

Table 3.8: The big gene modules identified by different approaches (NA: no modules with a size of at least 30 genes available).

Method	Size of big modules	Number of big modules
Proposed (TPGM)	312, 169	2
Proposed (SPGM)	229, 75, 120	3
Proposed (SqrtPGM)	48, 164, 120, 114	4
Estimation (TPGM)	NA	0
Estimation (SPGM)	49, 32	2
Estimation (SqrtPGM)	NA	0
GFC_L	NA	0
Nonparanormal SKEPTIC	NA	0

We further performed gene pathway enrichment analysis on the identified big modules in Table 3.8 using `ToppGene Suite` (Chen et al., 2009) with FDR control at level 0.05 based on the Benjamini-Hochberg (B-H) procedure, see Table B.22 in Appendix B.8.8 for complete results. From those modules identified by the proposed inferential procedure, we found some pathways that are shared within three models and critical to atopic asthma, for example, metal sequestration by antimicrobial proteins and FasL/CD95L signaling. The antimicrobial activity of S100A8/A9 proteins can induce a metal-witholding response by starving pathogens with metal nutrients in inflamed upper airway due to the chronic autoimmune diseases like asthma, according to Van Crombruggen et al. (2016). The potential role of Fas and its ligand (FasL) signaling pathway in asthma has been intensively studied. For example, the resistance of T helper type 2 (Th2) cells to normal degree of apoptosis induced by Fas and its ligand prolongs or delays resolution of inflammation in atopic asthma (Potapinska and Demkow, 2009). Williams et al. (2018) has recently demonstrated that non-apoptotic Fas signaling on T cells promotes resolution of Th2-mediated airway inflammation. More

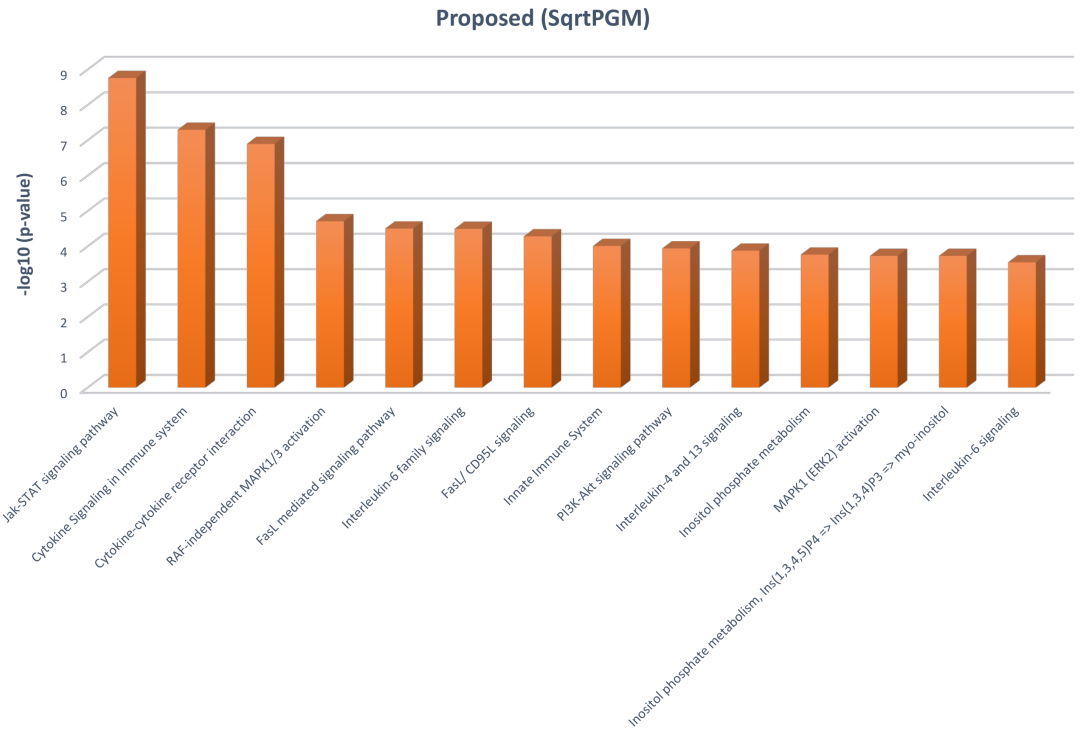

Figure 3.5: Some enriched pathways from the proposed inferential procedure in SqrtPGM.

interestingly, we also noticed some unique pathways enriched from the different modules in the three models, as shown in Figure 3.5 for SqrtPGM and Figure B.14 in Appendix B.8.3 for TPGM and SPGM. For example, CLEC7A/inflammasome pathway from TPGM, known as dectin-1 and a major receptor to β-glucans (an important group of allergens from house dust mites) (Hadebe et al., 2018), has recently been shown to protect against asthma and allergies (Gour et al., 2018). TRAIL signaling pathway from SPGM has appeared to have a detrimental role in allergic asthma by upregulating inflammation and immune responses, in terms of Braithwaite et al. (2018). More unique pathways were enriched from SqrtPGM, such as JAK-STAT signaling pathway, Interleukin-4 and 13 (IL-4/IL-13) signaling pathway, and Interleukin-6 (IL-6) signaling pathway. JAK-STAT signaling pathway has been shown to play an important role in the development of atopic asthma by differentiating Th2 cells from naïve T cells (Vale, 2016) and regulating the level of IgE (Zhang et al., 2018a). IL-4/IL-13 signaling pathway is central for IgE regulation, and genetic alterations in this pathway reveals its significance to the development of childhood atopic asthma (Kabesch et al., 2006). IL-6 signaling pathway has been evidenced to play an active role in pathogenesis of asthma, thus IL-6 can be a potential target for its treatment (Rincon and Irvin, 2012). However, the identified gene modules from the sole estimation are not capable of reflecting critical gene pathways about allergic asthma compared with the proposed inferential procedure. The gene interactions of the module in SqrtPGM with enriched JAK-STAT signaling pathway, as well as their corresponding interactions in TPGM and SPGM, are further presented in Figure B.15 in Appendix B.8.4.

Then, we investigated interactions among the 12 genes included in the JAK-STAT signaling pathway which is the most significant enriched pathway from SqrtPGM and also the one enriched using a total of 500 genes with FDR control at the 0.05 level, see Table B.21 in Appendix B.8.7. Targeting this pathway will be therapeutically effective on asthma pathology (Vale, 2016). The inferred gene interactions from our procedure, the sole estimation, GFC_L and nonparanormal SKEPTIC are demonstrated at a fixed panel in Figure 3.6. As we can see, the sole estimation, GFC_L and nonparanormal SKEPTIC can barely detect any informative interactions except the one between IL6 and CSF3. Conversely, our procedure is capable of identifying more meaningful interactions related to atopic asthma in addition

to the one between IL6 and CSF3, for example, IL6R and IL6ST from TPGM, and CSF3 and CSF3R from SqrtPGM. The activation of IL6R requires an association with IL6ST so as to regulate the immune response. CSF3R, which is associated with asthma, is known as the receptor for CSF3 and should be involved in granulopoiesis during the inflammatory process. According to a more recent study in Wang et al. (2019), CSF3 is identified as a major effector that promotes infection-dependent transition to severe asthma, and inhibition of CSF3R can be a potential strategy for preventing the pathological inflammation. These suggest that the sole estimation, GFC_L and nonparanormal SKEPTIC may neglect important functional relationships between genes closely related to atopic asthma.

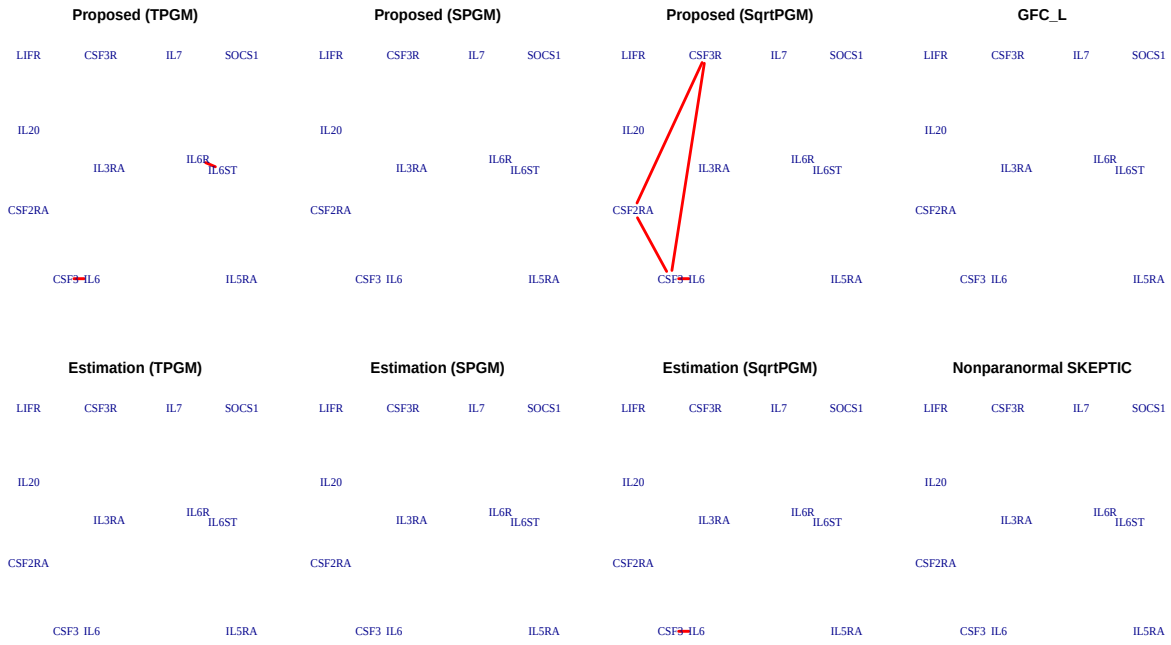

Figure 3.6: The inferred interactions of genes within the JAK-STAT signaling pathway.

Last but not least, the inferred networks from our approach can capture important hub genes that are closely associated with asthma and allergy, for example, NTRK2 (21 and 50 connections to other genes in SPGM and SqrtPGM respectively) and GSN (27, 23 and 71 connections to other genes in TPGM, SPGM and SqrtPGM respectively). The two genes are also listed as the top differentially expressed genes in Forno et al. (2020) and are well

replicated by the two external cohorts from Giovannini-Chami et al. (2012) and Yang et al. (2017). However, the sole estimation, GFC_L and nonparanormal SKEPTIC fail to identify these important hub genes.

In summary, our refined inference is more useful compared to the sole estimation, GFC_L and nonparanormal SKEPTIC. It not only reveals more significant pathways related to atopic asthma, but also captures more complex gene co-expression structures and important hub genes. The sole estimation and the inferential methods based on Gaussian and nonparanormal graphical models may lead to information loss and are less powerful to obtain informative disease-relevant results. Therefore, our procedure can be potentially useful for new treatment development in atopic asthma. To further demonstrate the advantages of our proposed method, we performed additional analysis of GFC_L on transcript per million (TPM) values from RNA-seq data and an additional comparison of methods on a well-characterized data set with some established "ground truth", see Appendices B.8.5 and B.8.6 for more details.

3.7 Conclusion and discussion

We have developed a novel procedure for statistical inference of three modified Poisson-type graphical models which provides reliable confidence intervals and p-values of pairwise edge and desirable false discovery rate control of multiple edges to tailor the network analysis of non-negative, discrete and high-dimensional data. The procedure essentially relies on the intrinsic property of graphical models and is different from the existing regression-based bias correction. Compared to the sole estimation approach, the proposed method is robust to different hyperparameter selection criteria, which results in its noticeably better performance in inferring a more biologically meaningful network by identifying more true signals while simultaneously controlling false discovers at a reasonably low level. Compared to the application of graphical model methods under normal and nonparanormal assumptions, the proposed method tends to reveal more biological meaningful networks and is more capable of capturing important gene interactions with less information loss. From Yang et al. (2013), they mentioned another modified Poisson-type model called quadratic Pois-

son graphical model (QPGM). However, unlike the desired Poisson tail, QPGM is more like Gaussian distribution and has Gaussian-esque thin tail. Due to this major drawback, we do not consider QPGM here.

The proposed method can be applied to more different types of omics data even though it is mainly motivated by the count-valued RNA-seq, for example, DNA copy number variation (CNV) data and single nucleotide polymorphism (SNP) data for genomics. In practice, TPGM is more suitable for the context with a relatively small range of discrete values such as CNV or SNP data. For RNA-seq data which generally has much larger discrete values, we recommend to explore SPGM or SqrtPGM first because they are more general Poisson-type distributions and allow a broader set of feasible parameters for pairwise conditional dependence than TPGM. Indeed, when the upper bound D_i of TPGM becomes larger, the behavior of TPGM tends to be closer to original PGM, which suffers from the limitation of negative pairwise dependency. With sufficient computational resource, we suggest to explore all three models by comparing the results of overall network inference, gene modularity detection, and gene network construction for important pathways according to different purpose of each study.

4.0 R-CoLaR: Robust Convex Program with Group-Lasso Refinement for Sparse CCA: Application to Heavy-Tailed CITE-Seq Data

4.1 Introduction

Recent technological advances have allowed high-throughput measurements for different layers of a biological system and generated large amount of multi-omics data, for example, DNA copy number variation (CNV) data and single nucleotide polymorphism (SNP) data for genomics, DNA methylation data for epigenomics, RNA-seq data for transcriptomics, and protein data for proteinomics. Integrative analysis combining data from different omic levels helps to better elucidate their interrelation and joint influences on the disease processes (Sun and Hu, 2016).

Canonical correlation analysis (CCA) (Hotelling, 1936) is an important statistical method in multivariate analysis to explore the association between two sets of variables. On the population level, for two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, CCA aims to identify the canonical coefficient vectors $u_j \in \mathbb{R}^p$ and $v_j \in \mathbb{R}^q$ recursively that maximize the canonical correlation $\lambda_j = \text{Corr}(u_j^\intercal X, v_j^\intercal Y)$ between canonical variables $u_j^\intercal X$ and $v_j^\intercal Y$ based on the following criterion:

$$\begin{aligned}(u_j, v_j) = \arg\max_{u,v} \quad & u^\intercal \Sigma_{xy} v \\ \text{s.t.} \quad & u^\intercal \Sigma_x u = v^\intercal \Sigma_y v = 1; \\ & u^\intercal \Sigma_x u_l = 0, v^\intercal \Sigma_y v_l = 0, \forall 1 \le l \le j-1,\end{aligned} \tag{4.1}$$

where $\Sigma_x = \text{Cov}(X, X)$, $\Sigma_y = \text{Cov}(Y, Y)$ and $\Sigma_{xy} = \text{Cov}(X, Y)$. Under low-dimensional settings, the problem (4.1) is straightforward to be solved by the singular value decomposition (SVD) on $\Sigma_x^{-1/2}\Sigma_{xy}\Sigma_y^{-1/2}$. But in genomic study, omics data sets are generally high-dimensional with large p or q. Thus, this SVD approach for classical CCA may not work under these high-dimensional settings due to the ill-defined inverse of covariance matrices Σ_x^{-1} and Σ_y^{-1}.

To cope with high-dimensional omics data analysis, sparse CCA has been extensively studied within the recent ten years by imposing sparsity constraints on the structure of canonical coefficient vectors, which leads to more interpretable results with a small set of nonzero coordinates (Witten et al., 2009; Witten and Tibshirani, 2009; Hardoon and Shawe-Taylor, 2011; Chen et al., 2019; Gao et al., 2015, 2017; Suo et al., 2017). There are two main streams of sparse CCA approaches. One line is based on Witten et al. (2009) and Suo et al. (2017). Witten et al. (2009) developed a penalized matrix decomposition approach for sparse CCA under the assumption of $\Sigma_x = I_p$ and $\Sigma_y = I_q$, which instead is unsuitable for more general structures of Σ_x and Σ_y. To circumvent this major drawback, Suo et al. (2017) proposed to solve sparse CCA to permit general structures of Σ_x and Σ_y using the linearized alternating direction method with multipliers (ADMM) following the alternating minimization approach in Witten et al. (2009). From here on, we name this approach as SCCA-ADMM. Another line is based on Chen et al. (2019), Gao et al. (2015) and Gao et al. (2017) which also laid theoretical foundation of sparse CCA. Chen et al. (2019) proposed a CAPIT (standing for canonical correlation analysis via precision adjusted iterative thresholding) method to estimate sparse canonical coefficient vectors under some known structures of precision matrices (the inverse of covariance matrices). However, structures of covariance or precision matrices are generally unknown. To relax this condition, Gao et al. (2017) proposed CoLaR (standing for Convex program with group-Lasso Refinement) to solve sparse CCA in two stages without prior knowledge of covariance or precision matrices. In addition to theoretical and methodological development, there is also a broad existing literature about applications of sparse CCA in genomic research. Witten et al. (2009) and Witten and Tibshirani (2009) applied sparse CCA to study the relationship between gene expression and CNV data on a same set of subjects for the breast cancer and the diffuse large B-cell lymphoma (DLBCL). Parkhomenko et al. (2009) used sparse CCA to identify sets of genes that are correlated with SNPs for the chronic fatigue syndrome (CSF). Lê Cao et al. (2009) performed sparse CCA on gene expression profiles of 60 human tumor cell lines from two different platforms. Furthermore, Safo et al. (2018) applied sparse CCA to study the association between DNA methylation and gene expression profiles for the breast cancer. However, all the aforementioned methods and applications of sparse CCA which depend on

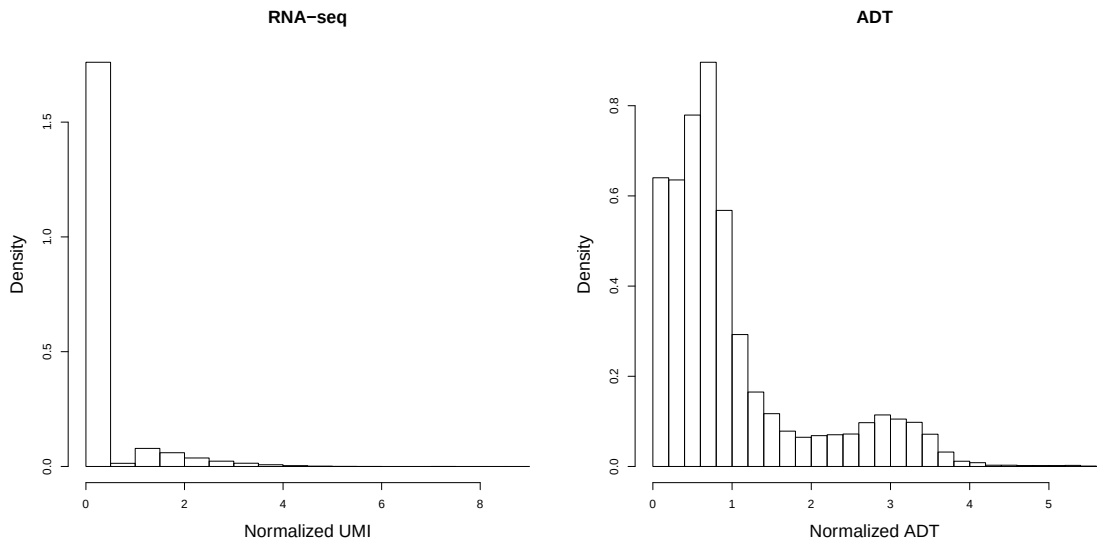

Figure 4.1: Histograms of normalized UMI and ADT counts for CITE-seq data of a MALT tumor.

sample covariance estimation of Σ_x, Σ_y and Σ_{xy} are not robust to data with heavy-tailed distributions.

Data with heavy-tailed distributions is actually common in genomic study. One typical example is the data from cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) which allows RNA-sequencing and information on cell surface proteins with available antibodies simultaneously at a single-cell level. The abundance of RNA and surface proteins is quantified by counts of unique molecular identifiers (UMI) and antibody-derived tags (ADT) respectively for a same set of single cells. The availability of CITE-seq data has paved a new way to learn protein-RNA correlation at a single-cell level, which aids in the identification of novel tumor subtypes and also permits the discovery of rare subpopulations of cells (Tirosh and Suvà, 2019). Figure 4.1 presents the histograms of normalized UMI and ADT from RNA-seq and surface proteins on a same set of single cells for a CITE-seq data set of a dissociated extranodal marginal zone B-cell lymphoma (MALT: mucosa-associated lymphoid tissue). As it can be seen, even after normalization using the standard workflow of Seurat (Stuart et al., 2019), both normalized UMI and ADT counts still have heavy-tailed

distributions. Due to the heavy-tailed nature of CITE-seq data, sample covariance matrices may not be robust, and the existing sparse CCA approaches may lead to inaccurate results and fail to identify important features.

In this chapter, we propose a novel procedure named with robust CoLaR (R-CoLaR) which consists of two parts that tailor to heavy-tailed settings such as CITE-seq data. The first part is to robustify sample covariance using the cutting-edge tail robustness estimation (Catoni, 2016; Avella-Medina et al., 2018; Ke et al., 2019), and the second part is to modify the convex optimization of original CoLaR by replacing sample covariance estimators with these robustified ones. Compared to existing sparse CCA methods, R-CoLaR maintains the interpretability of sparse CCA and overcomes their constrains under heavy-tailed distributions.

The chapter is organized as follows. Section 4.2 contains a detailed description for the procedure of R-CoLaR. Section 4.3 demonstrates the validity and noticeable advantages of R-CoLaR over existing sparse CCA approaches such as CoLaR and SCCA-ADMM using both simulation studies and application to the CITE-seq data of a MALT tumor. We finally conclude with discussion in Section 4.4.

4.2 Methods

Let $\boldsymbol{X} = (X_1, ..., X_n)^\intercal$ be one omics data set with n subjects and p dimensions and $\boldsymbol{Y} = (Y_1, ..., Y_n)^\intercal$ be another omics data set with q dimensions from the same group of subjects. $\boldsymbol{X}$ and $\boldsymbol{Y}$ consist of i.i.d. rows (or copies) of $X_k = (X_{k1}, ..., X_{kp})^\intercal \in \mathbb{R}^p$ and $Y_k = (Y_{k1}, ..., Y_{kq})^\intercal \in \mathbb{R}^q$ respectively with $k = 1, 2, ..., n$.

As mentioned in Section 4.1, the existing methods of sparse CCA (Gao et al., 2017; Suo et al., 2017) are not suitable for data with heavy-tailed distributions due to non-robustness of their sample covariance estimators. Therefore, our proposed R-CoLaR is to robustify sample covariance estimators and make them as an integral part to modify the convex optimization of original CoLaR. The procedure consists of two parts. The first part is to provide robust estimation for covariance matrices of $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_y$ and $\boldsymbol{\Sigma}_{xy}$. The second part is to plug these

robust estimated covariance matrices into the framework of CoLaR to solve the sparse CCA problem. From now on, we denote $\boldsymbol{C} = (\boldsymbol{X}, \boldsymbol{Y})$ as the n by $(p+q)$ combined data set of $\boldsymbol{X}$ and $\boldsymbol{Y}$. The covariance matrix of compound data is

$$\boldsymbol{\Sigma}_C = \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy}^\intercal & \boldsymbol{\Sigma}_y \end{pmatrix}. \tag{4.2}$$

4.2.1 Robustification of covariance matrix

In the first part, we follow the two tail robustness estimation approaches to robustify the sample covariance matrix proposed by Ke et al. (2019): 1. Element-wise truncated estimator; 2. Huber-type M-estimator.

The element-wise truncated estimator is to truncate data so as to eliminate effects of heavy-tailed noises, so each resulting estimator has sub-Gaussian tails. Its implementation depends on the truncation operator

$$\psi_\tau(w) = (|w| \wedge \tau)\,\mathrm{sign}(w), \qquad w \in \mathbb{R}, \tag{4.3}$$

where τ is a robustification parameter and $|w| \wedge \tau = \min(|w|, \tau)$. The idea behind element-wise truncated estimator is essentially based on the truncation of the U-statistic of sample covariance matrix of $\boldsymbol{C}$ which can be represented as

$$\hat{\boldsymbol{\Sigma}}_C = \frac{1}{N}\sum_{i=1}^{N}(E_i E_i^\intercal)/2, \qquad N = n(n-1)/2, \tag{4.4}$$

where $\{E_1, E_2, ..., E_N\} = \{C_1 - C_2, C_1 - C_3, ..., C_{n-1} - C_n\}$ denotes the difference between each pair of different samples of the compound data set $\boldsymbol{C}$. The use of U-statistic is free of mean estimation, which avoids the error of mean estimators under heavy-tailed distributions and also avoids additional steps to estimate mean using existing robust methods that further increase both statistical variability and computational complexity. The truncation operator in (4.3) is applied to each element of $\hat{\boldsymbol{\Sigma}}_C$ in (4.4) to obtain the truncated estimator of covariance matrix

$$\hat{\boldsymbol{\Sigma}}_C^{\mathcal{T}} = (\hat{\sigma}_{kl}^{\mathcal{T}})_{1\le k,l\le(p+q)}, \qquad \hat{\sigma}_{kl}^{\mathcal{T}} = \frac{1}{N}\sum_{i=1}^{N}\psi_{\tau_{kl}}(E_{ik}E_{il}/2), \tag{4.5}$$

and $\hat{\mathbf{\Sigma}}_x^{\mathcal{T}}$, $\hat{\mathbf{\Sigma}}_y^{\mathcal{T}}$ and $\hat{\mathbf{\Sigma}}_{xy}^{\mathcal{T}}$ are the corresponding element-wise truncated estimators of $\mathbf{\Sigma}_x$, $\mathbf{\Sigma}_y$ and $\mathbf{\Sigma}_{xy}$ in (4.2).

Another estimation approach is the Huber-type M-estimator which depends on the Huber loss (Huber, 1964)

$$L_\tau(w) = \begin{cases} w^2/2, & if \quad |w| \le \tau, \\ \tau|w| - \tau^2/2, & if \quad |w| > \tau, \end{cases} \tag{4.6}$$

where τ is also a robustification parameter like the one in (4.3). The Huber loss can be regarded as a specific type of truncation and is closely related to the truncation operator in (4.3). Similar to the idea of element-wise truncated estimator, the Huber loss in (4.6) is applied to solve an M-estimation problem for each element of covariance matrix such that

$$\hat{\mathbf{\Sigma}}_C^{\mathcal{H}} = (\hat{\sigma}_{kl}^{\mathcal{H}})_{1\le k,l\le (p+q)}, \qquad \hat{\sigma}_{kl}^{\mathcal{H}} = \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^{N} L_{\tau_{kl}}(E_{ik}E_{il}/2 - \theta), \tag{4.7}$$

where $\hat{\mathbf{\Sigma}}_x^{\mathcal{H}}$, $\hat{\mathbf{\Sigma}}_y^{\mathcal{H}}$ and $\hat{\mathbf{\Sigma}}_{xy}^{\mathcal{H}}$ are the corresponding Huber-type M-estimators of $\mathbf{\Sigma}_x$, $\mathbf{\Sigma}_y$ and $\mathbf{\Sigma}_{xy}$ in (4.2).

How to calibrate the robustification parameter τ is critical. On the one hand, a small constant level of τ results in non-negligible bias of estimators. For example, for Huber loss with $\tau = 0$, the population minimizer of (4.7) becomes the median of $E_{ik}E_{il}/2$, which is different from the mean of $E_{ik}E_{il}/2$ for most distributions. On the other hand, a very large τ leads to non-robust estimators which fail to concentrate tightly around the true expectation. How to choose τ is one recent focus in the field of theoretical statistics which aims to reach an optimal bias-robustness tradeoff with finite-sample concentration for large covariance estimators (Avella-Medina et al., 2018).

For element-wise truncated estimators, it turns out that an "ideal" choice of τ_{kl} should adapt to the sample size n and the dimension $(p+q)$ such that $\tau_{kl} \asymp v_{kl}\sqrt{\frac{n}{\log(p+q)}}$ as shown in Avella-Medina et al. (2018) and Ke et al. (2019), where $v_{kl} = \mathbb{E}(Z_i^2)$ denotes the second moment of Z_i. Ke et al. (2019) further derived a sharp constant in front of the order of τ_{kl} and introduced a data-driven procedure to automatically tune the robustification parameters,

which in practice avoids extensive computational cost incurred by cross validation. More specifically,

$$\tau_{kl} = v_{kl}\sqrt{\frac{m}{2\log(p+q)+t}}, \tag{4.8}$$

where $m = \lfloor n/2 \rfloor$ (the greatest integer not exceeding $n/2$), and the adjustable parameter t indicates the confidence level $1 - 2e^{-t}$ at which the truncated estimator is concentrated around the true covariance, see Theorem 3.1 of Ke et al. (2019) for further details. A naïve estimator of v_{kl} is $(1/N)\sum_{i=1}^{N} Z_i^2$, but it tends to overestimate the true value under heavy-tailed distributions. Therefore, a more reasonable and robust estimator is $(1/N)\sum_{i=1}^{N}\psi_{\tau_{kl}}^2(Z_i) = (1/N)\sum_{i=1}^{N}(Z_i^2 \wedge \tau_{kl}^2)$, which is based on the robustification parameter τ_{kl} using the truncation operator in (4.3). For this reason, we can finally obtain (4.9) by plugging this estimator into (4.8) as below,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{(Z_i^2 \wedge \tau_{kl}^2)}{\tau_{kl}^2} = \frac{2\log(p+q)+t}{m}, \qquad \tau_{kl} > 0, \tag{4.9}$$

where $\{Z_1, ..., Z_N\} = \{E_{1k}E_{1l}/2, ..., E_{Nk}E_{Nl}/2\}$. An adaptive estimator $\hat{\tau}_{kl}$ is obtained by solving (4.9). It is easy to see that there is a unique solution of τ_{kl} in (4.9) if $2\log(p+q)+t < (m/N)\sum_{i=1}^{N} I\{Z_i \neq 0\}$.

Based on the similar idea for element-wise truncated estimators, a data-driven estimator $\hat{\tau}_{kl}$ for Huber-type M-estimators (Ke et al., 2019) can be obtained by solving both θ and τ_{kl} together using the following system of equations

$$f_1(\theta, \tau_{kl}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\{(Z_i-\theta)^2 \wedge \tau_{kl}^2\}}{\tau_{kl}^2} - \frac{2\log(p+q)+t}{n} = 0, \tag{4.10}$$

$$f_2(\theta, \tau_{kl}) = \sum_{i=1}^{N}\psi_{\tau_{kl}}(Z_i - \theta) = 0, \tag{4.11}$$

where $\theta \in \mathbb{R}$ denotes $\mathbb{E}(Z_i)$, the first moment (or mean) of Z_i. Different from (4.9), (4.10) actually depends on the truncated estimator for the variance of Z_i rather than its second moment. (4.11) is to obtain a good estimator of $\mathbb{E}(Z_i)$ even under heavy-tailed distributions based on the robustification parameter τ_{kl} using the truncation operator in (4.3). Through a similar argument like that in Wang et al. (2020), it can be shown that there is a unique solution of τ_{kl} in (4.10) if $2\log(p+q)+t < (n/N)\sum_{i=1}^{N} I\{Z_i \neq 0\}$ and a unique solution of θ

in (4.11) for any $\tau_{kl} > 0$. Specifically, we can iteratively solve $f_1(\theta^{(s-1)}, \tau_{kl}^{(s)})$ and $f_2(\theta^{(s)}, \tau_{kl}^{(s)})$ in (4.10) and (4.11) for $s = 1, 2, ...$ by starting with the initial $\theta^{(0)} = (1/N)\sum_{i=1}^{N} Z_i$ until both $\theta^{(s)}$ and $\tau_{kl}^{(s)}$ converge.

Finally, for the adjustable parameter t in (4.9) and (4.10), the choice of $t = \log(n)$ is recommended by Ke et al. (2019) in practical use, but its value can be customized to accommodate different situations.

Both Huber-type M-estimators and estimators defined via element-wise truncation calibrate the robustification parameter τ_{kl} for a bias-robustness tradeoff to achieve tail robustness. Different from truncation-based estimators which truncate around zero as shown in (4.5), M-estimators truncate around the true expectation as shown in (4.7). In a finite-sample case, Huber-type M-estimators can outperform the element-wise truncation-based estimators due to smaller bias. But this subtle difference between the two types of estimators becomes insignificant when the sample size n becomes larger.

4.2.2 Robust sparse CCA

In the second part, we modify the two-stage framework of CoLaR proposed by Gao et al. (2017) by replacing the sample covariance estimators with the robust covariance estimators introduced in Section 4.2.1. In practice, we need to guarantee the positive definiteness of $\hat{\mathbf{\Sigma}}_x^{\mathcal{H}}$ (or $\hat{\mathbf{\Sigma}}_x^{\mathcal{T}}$) and $\hat{\mathbf{\Sigma}}_y^{\mathcal{H}}$ (or $\hat{\mathbf{\Sigma}}_y^{\mathcal{T}}$) for feasible solutions when working on the convex optimization problems of CoLaR. Since $\hat{\mathbf{\Sigma}}_x^{\mathcal{H}}$ (or $\hat{\mathbf{\Sigma}}_x^{\mathcal{T}}$) and $\hat{\mathbf{\Sigma}}_y^{\mathcal{H}}$ (or $\hat{\mathbf{\Sigma}}_y^{\mathcal{T}}$) may not be positive definite particularly under high-dimensional settings, we instead use their nearest positive definite matrix with distance characterized by Frobenius norms (Higham, 2002) before plugging them into problems.

The idea of CoLaR is to solve problems of sparse CCA in two stages to finally extract adaptive estimators of $U = [u_1, u_2, ..., u_r]$ and $V = [v_1, v_2, ..., v_r]$ (two collections of canonical coefficient vectors) which achieve desirable estimation rates for U and V separately. For convenience, we use Huber-type M-estimators $\hat{\mathbf{\Sigma}}_x^{\mathcal{H}}$, $\hat{\mathbf{\Sigma}}_y^{\mathcal{H}}$ and $\hat{\mathbf{\Sigma}}_{xy}^{\mathcal{H}}$ in the following elaboration. The corresponding covariance estimators can be easily replaced by element-wise truncated estimators $\hat{\mathbf{\Sigma}}_x^{\mathcal{T}}$, $\hat{\mathbf{\Sigma}}_y^{\mathcal{T}}$ and $\hat{\mathbf{\Sigma}}_{xy}^{\mathcal{T}}$ as well.

The purpose of initial stage is to deal with the following original sparse CCA problem with sparsity constraints:

$$\begin{aligned} &\max_{U\in\mathbb{R}^{p\times r},V\in\mathbb{R}^{q\times r}} \operatorname{Tr}(U^\intercal \hat{\mathbf{\Sigma}}_{xy}^{\mathcal{H}} V) \\ &\text{s.t.} \qquad U^\intercal \hat{\mathbf{\Sigma}}_{x}^{\mathcal{H}} U = V^\intercal \hat{\mathbf{\Sigma}}_{y}^{\mathcal{H}} V = I_r, \quad |S_u| \le s_u, \quad |S_v| \le s_v, \end{aligned} \tag{4.12}$$

where S_u and S_v are the indices of nonzero rows in U and V respectively, and $|S_u| \le s_u$ and $|S_v| \le s_v$ represent that the sizes of S_u and S_v can be even smaller than some sparsity levels $s_u \le p$ and $s_v \le q$. However, the program (4.12) is non-convex and computationally infeasible with ℓ_0 norms. Therefore, instead of solving (4.12) directly, the initial stage is to view U and V as a whole and provide a joint estimator of $F = UV^\intercal$ by solving a convex relaxation program of (4.12):

$$\begin{aligned} &\max_{F\in\mathbb{R}^{p\times q}} \operatorname{Tr}(\hat{\mathbf{\Sigma}}_{xy}^{\mathcal{H}\intercal} F) - \rho||F||_1 \\ &\text{s.t.} \qquad ||\hat{\mathbf{\Sigma}}_{x}^{\mathcal{H}\frac{1}{2}} F \hat{\mathbf{\Sigma}}_{y}^{\mathcal{H}\frac{1}{2}}||_* \le r, \quad ||\hat{\mathbf{\Sigma}}_{x}^{\mathcal{H}\frac{1}{2}} F \hat{\mathbf{\Sigma}}_{y}^{\mathcal{H}\frac{1}{2}}||_{\text{op}} \le 1, \end{aligned} \tag{4.13}$$

where $\operatorname{Tr}(.)$ denotes the trace of a matrix, and ρ is a penalty parameter that controls sparsity. $||.||_*$ and $||.||_{\text{op}}$ are denoted as the nuclear norm and the operator norm respectively. As we can see from (4.13), the nuclear and operator norms are bounded by r and 1, and the program is convex which can be solved efficiently by ADMM algorithm. However, the joint estimator $\hat{F}$ may still neglect some separate intrinsic structures of U and V. Therefore, we need another stage to further refine the current results. In spite of the drawback of $\hat{F}$, it is still a globally good estimator spanned by the initial estimated canonical coefficient vectors $\hat{U}^{(0)}$ and $\hat{V}^{(0)}$ (the first r left and right singular vectors from $\hat{F}$), which is sufficient to achieve the final estimators of U and V in the refined stage.

The purpose of refined stage is thus to improve $\hat{U}^{(0)}$ and $\hat{V}^{(0)}$ from the initial stage so as to get final estimators $\hat{U}$ and $\hat{V}$ by taking the separate structures of U and V into consideration. With the knowledge of $\hat{U}^{(0)}$ and $\hat{V}^{(0)}$, the refined stage aims to solve the sparse CCA problem:

$$\begin{aligned} &\min_{U\in\mathbb{R}^{p\times r}} \operatorname{Tr}(U^\intercal \hat{\mathbf{\Sigma}}_{x}^{\mathcal{H}} U) - 2\operatorname{Tr}(U^\intercal \hat{\mathbf{\Sigma}}_{xy}^{\mathcal{H}} \hat{V}^{(0)}) \\ &\text{s.t.} \qquad |S_u| \le s_u. \end{aligned} \tag{4.14}$$

and

$$\begin{aligned} &\min_{V\in\mathbb{R}^{q\times r}} \operatorname{Tr}(V^\intercal \hat{\mathbf{\Sigma}}_y^{\mathcal{H}} V) - 2\operatorname{Tr}(V^\intercal \hat{\mathbf{\Sigma}}_{xy}^{\mathcal{H}} \hat{U}^{(0)}) \\ &\text{s.t.} \qquad |S_v| \le s_v. \end{aligned} \tag{4.15}$$

The programs (4.14) and (4.15) can be viewed as a regression interpretation of CCA with sparsity constraints for U and V. To obtain $\hat{U}$ (or $\hat{V}$) is equivalent to solve a multivariate regression problem when $\hat{V}^{(0)}$ (or $\hat{U}^{(0)}$) is known. If we view the rows of U (or V) as groups, then the regression interpretation of CCA leads to the following convex relaxation problem of (4.14) with group-Lasso:

$$\min_{U\in\mathbb{R}^{p\times r}} \operatorname{Tr}(U^\intercal \hat{\mathbf{\Sigma}}_x^{\mathcal{H}} U) - 2\operatorname{Tr}(U^\intercal \hat{\mathbf{\Sigma}}_{xy}^{\mathcal{H}} \hat{V}^{(0)}) + \rho_u \sum_{j=1}^{p} ||U_{j\cdot}||, \tag{4.16}$$

where $\sum_{j=1}^{p} ||U_{j\cdot}||$ is the sum of the ℓ_2 norm of the rows of U as the group sparsity penalty, and ρ_u is a tuning parameter that controls sparsity of U. Similar to (4.16) for U, we have the following problem with group-Lasso for V:

$$\min_{V\in\mathbb{R}^{q\times r}} \operatorname{Tr}(V^\intercal \hat{\mathbf{\Sigma}}_y^{\mathcal{H}} V) - 2\operatorname{Tr}(V^\intercal \hat{\mathbf{\Sigma}}_{xy}^{\mathcal{H}\intercal} \hat{U}^{(0)}) + \rho_v \sum_{j=1}^{q} ||V_{j\cdot}||, \tag{4.17}$$

where $\sum_{j=1}^{q} ||V_{j\cdot}||$ is the sum of the ℓ_2 norm of the rows of V, and ρ_v is another tuning parameter that controls sparsity of V. The final estimators $\hat{U}$ and $\hat{V}$ of U and V are derived after solving problems (4.16) and (4.17) and normalizing the results with Huber-type robust covariance estimators $\hat{\mathbf{\Sigma}}_x^{\mathcal{H}}$ and $\hat{\mathbf{\Sigma}}_y^{\mathcal{H}}$ respectively.

4.3 Results

4.3.1 Simulation study

We evaluated the validity of R-CoLaR based on its accuracy of estimation and performance of feature selection under two cases: 1. single pair of canonical coefficient vectors; 2.

two pairs of canonical coefficient vectors among four different distributions.

For both cases, we considered the following compound covariance matrix

$$\boldsymbol{\Sigma}_C = \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy}^\intercal & \boldsymbol{\Sigma}_y \end{pmatrix},$$

where $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}$ has a `Toeplitz` structure: $\boldsymbol{\Sigma} = (\sigma_{ij})$ with $\sigma_{ij} = 0.5^{|i-j|}$ for all i, j, and $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_x U \Lambda V^\intercal \boldsymbol{\Sigma}_y$ includes two collections of canonical coefficient vectors $U = [u_1, u_2, ..., u_r]$ and $V = [v_1, v_2, ..., v_r]$ and the ordered canonical correlations $\Lambda = \text{diag}(\lambda_1, ..., \lambda_r)$ with $1 > \lambda_1 \geq ... \geq \lambda_r > 0$. It can be seen that $r = 1$ and 2 for cases of single pair and two pairs of canonical coefficient vectors respectively. Moreover, the number of row sparsity was chosen to be 5 for both U and V, and their positions were randomly sampled from the rows of U and V respectively. Each nonzero value was then chosen randomly from the uniform distribution between -2 and 2. In the end, U and V were normalized with respect to $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$ to guarantee $U^\intercal \boldsymbol{\Sigma}_x U = V^\intercal \boldsymbol{\Sigma}_y V = I_r$.

Based on the aforementioned true settings for both cases, we generated 100 data sets of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. The details to simulate each data set of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are as follows. We at first generated an $n \times (p + q)$ data matrix $\boldsymbol{D}_1$ where each element is i.i.d. with a certain distribution. Then, we scaled $\boldsymbol{D}_1$ to $\boldsymbol{D}_2$ using true means and true standard deviations. Finally, $\boldsymbol{D}_2$ was further rescaled to obtain the final data set $\boldsymbol{D}_3 = \boldsymbol{D}_2(\boldsymbol{\Sigma}_C)^{1/2}$. The first p columns of $\boldsymbol{D}_3$ represent a data set of $\boldsymbol{X}$, while the remaining columns consist of a data set of $\boldsymbol{Y}$. Here, we set $n = 150$, $p = 200$ and $q = 200$. We considered four distributions same as those in Ke et al. (2019) to generate $\boldsymbol{D}_1$:

- Normal distribution. Each element of $\boldsymbol{D}_1$ is i.i.d. with the standard normal distribution.
- Student's t distribution. $\boldsymbol{D}_2 = \boldsymbol{D}_1/\sqrt{3}$, where each element of $\boldsymbol{D}_1$ is i.i.d. and follows Student's t distribution with 3 degrees of freedom.
- Pareto distribution. $\boldsymbol{D}_2 = 2(\boldsymbol{D}_1 - 3/2)/\sqrt{3}$, where each entry of $\boldsymbol{D}_1$ is i.i.d. with Pareto distribution with shape parameter 3 and scale parameter 1.
- Log-normal distribution. $\boldsymbol{D}_2 = (\boldsymbol{D}_1 - e^{0.5})/\sqrt{e^2 - e}$, where each element of $\boldsymbol{D}_1$ is i.i.d. with Log-normal distribution with parameters $\mu = 0$ and $\sigma = 1$.

4.3.1.1 Case I: single pair of canonical coefficient vectors For the single case, the true pair of canonical coefficient vectors is denoted as (u_1, v_1), and the rank of $\mathbf{\Sigma}_{xy}$ is $r = 1$. We set the true canonical correlation $\lambda_1 = 0.9$ here.

We denoted R-CoLaR$^{\mathcal{H}}$ with Huber-type M-estimators $\hat{\mathbf{\Sigma}}_x^{\mathcal{H}}$, $\hat{\mathbf{\Sigma}}_y^{\mathcal{H}}$ and $\hat{\mathbf{\Sigma}}_{xy}^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ with element-wise truncated estimators $\hat{\mathbf{\Sigma}}_x^{\mathcal{T}}$, $\hat{\mathbf{\Sigma}}_y^{\mathcal{T}}$ and $\hat{\mathbf{\Sigma}}_{xy}^{\mathcal{T}}$ as covariance estimation respectively. When performing R-CoLaR, we used the `R` function `nearPD()` to compute their nearest positive definite approximation to guarantee the positive definiteness of $\hat{\mathbf{\Sigma}}_x^{\mathcal{H}}$, $\hat{\mathbf{\Sigma}}_y^{\mathcal{H}}$, $\hat{\mathbf{\Sigma}}_x^{\mathcal{T}}$ and $\hat{\mathbf{\Sigma}}_y^{\mathcal{T}}$. In the second part to perform sparse CCA, the tuning parameter in the initial stage was set at $\rho = 0.5\sqrt{\log(p \vee q)/n}$ where $p \vee q = \max(p, q)$, and the tuning parameters in the refined stage were set at $\rho_u = b\sqrt{(r + \log(p))/n}$ and $\rho_v = b\sqrt{(r + \log(q))/n}$ with $b = 0.5$. To compare with R-CoLaR, we performed original CoLaR with sample covariance estimators $\hat{\mathbf{\Sigma}}_x$, $\hat{\mathbf{\Sigma}}_y$ and $\hat{\mathbf{\Sigma}}_{xy}$ using default tuning parameters $\rho = 0.5\sqrt{\log(p \vee q)/n}$ and $\rho_u = \rho_v = b\sqrt{(r + \log(p \vee q))/n}$ from the authors' package with $b = 0.5$. We also performed SCCA-ADMM (Suo et al., 2017) with tuning parameters $\tau_u = b_1\sqrt{\log(p)/n}$ and $\tau_v = b_1\sqrt{\log(q)/n}$ with $b_1 = 0.5$ as another comparison study.

We at first evaluated the accuracy of estimated pair of canonical coefficient vectors $(\hat{u}_1, \hat{v}_1)$ from all methods based on two different types of errors:

- Error 1: $\min(||\hat{a} - a||_2^2, ||\hat{a} + a||_2^2)$ with $\hat{a} = \hat{u}_1/||\hat{u}_1||_2$ (or $\hat{v}_1/||\hat{v}_1||_2$) and $a = u_1/||u_1||_2$ (or $v_1/||v_1||_2$) (Suo et al., 2017; Chen et al., 2019).
- Error 2: $\inf_{w \in \{-1,1\}} ||\mathbf{\Sigma}^{1/2}(\hat{z}w - z)||_F^2$ with $\mathbf{\Sigma} = \mathbf{\Sigma}_x$ (or $\mathbf{\Sigma}_y$), $\hat{z} = \hat{u}_1$ (or $\hat{v}_1$) and $z = u_1$ (or v_1) (Gao et al., 2017), where $||.||_F$ is the Frobenius norm.

Error 1 essentially measures the sin angle expanded by the estimator $\hat{u}_1$ (or $\hat{v}_1$) and the ground truth u_1 (or v_1) with normalized lengths. Error 2 depicts the prediction loss which is the expected squared error for predicting canonical variables $\boldsymbol{X}u_1$ (or $\boldsymbol{Y}v_1$) using $\boldsymbol{X}\hat{u}_1$ (or $\boldsymbol{Y}\hat{v}_1$). Due to the capability of capturing important features from data for robust sparse CCA and sparse CCA, we then evaluated the performance of feature selection on the estimated pair of canonical coefficient vectors $(\hat{u}_1, \hat{v}_1)$. We made an additional thresholding step on $(\hat{u}_1, \hat{v}_1)$ to keep those non-zero entries with absolute magnitudes greater than 1×10^{-4}. In the ground truth (u_1, v_1) of our simulation settings, the numbers of non-zero and zero elements

are 5 and 195 respectively for both u_1 and v_1. Three metrics were used to measure the performance of identifying true signals and simultaneously controlling false discovers,

- True positive (TP): the number of true signals (or non-zero elements) in $\hat{u}_1$ (or $\hat{v}_1$) compared to u_1 (or v_1).
- True negative (TN): the number of true non-signals (or zero elements) in $\hat{u}_1$ (or $\hat{v}_1$) compared to u_1 (or v_1).
- The Matthews correlation coefficient (MCC): $= \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP}) \times (\mathrm{TP}+\mathrm{FN}) \times (\mathrm{TN}+\mathrm{FP}) \times (\mathrm{TN}+\mathrm{FN})}}$, where FP = false positives (the number of false signals) and FN = false negatives (the number of false non-signals). It is a more robust metric to the case with class imbalance (e.g. the number of true signals is much less than that of non-signals), and its value lies in the interval between -1 and 1. A value of 1 indicates a perfect identification, while a value of -1 demonstrates a totally wrong identification. A value of 0 illustrates a random guess.

Tables 4.1-4.4 report medians of all metrics from different methods based on the four distributions. Here, we independently summarized the results from initial stages of R-CoLaR and CoLaR as well. Table 4.1 shows the results of comparison on Normal distribution. R-CoLaR$^{\mathcal{H}}$ (initial) and R-CoLaR$^{\mathcal{T}}$ (initial) show very similar performance to CoLaR (initial) on all the metrics. Likewise, the results from R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ are similar to CoLaR. These methods after the refined stage further reduce two errors and improve the identification of TPs while slightly sacrificing the performance of TNs. SCCA-ADMM is slightly better than R-CoLaR and CoLaR on two types of errors. Under our expectation, R-CoLaR illustrates equivalently good performance as CoLaR and SCCA-ADMM on Normal distribution. From Tables 4.2-4.4 under settings of heavy-tailed distributions, R-CoLaR shows noticeable advantages over CoLaR and SCCA-ADMM. As it can be seen, Errors 1 and 2 even from R-CoLaR$^{\mathcal{H}}$ (initial) and R-CoLaR$^{\mathcal{T}}$ (initial) are significantly lower than those from CoLaR (initial), CoLaR and SCCA-ADMM. The values of Error 1 from CoLaR (initial), CoLaR and SCCA-ADMM are all close to 2 (the maximum value), implying that the estimated canonical coefficient vectors are nearly orthogonal to the ground truth. R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ further improve performance of errors. The refined stage in the

original CoLaR fails to reduce errors under these heavy-tailed distributions. Moreover, R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ further manifest their advantages in identifying true positives on Student's t, Pareto and Log-normal distributions, as shown in Tables 4.2-4.4. CoLaR and SCCA-ADMM barely detect important true signals, while both R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ are capable of identifying 4 out of 5 true positives and maintaining true negatives close to 195 as well. In terms of the TNs, it is clearly to notice that the original CoLaR after the refined stage does not control false discovers very well. In addition, the MCCs of R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ are obviously better than those of CoLaR and SCCA-ADMM which are very close to 0 and equivalent to a random guess in feature selection.

In summary, R-CoLaR is equivalently good as CoLaR and SCCA-ADMM on Normal distribution and particularly superior than these sparse CCA approaches on the heavy-tailed distributions in terms of both accuracy and feature selection under the case of single pair of canonical coefficient vectors.

Table 4.1: Method comparison under the single case on Normal distribution.

	R-CoLaR$^{\mathcal{H}}$ (initial)	R-CoLaR$^{\mathcal{H}}$	R-CoLaR$^{\mathcal{T}}$ (initial)	R-CoLaR$^{\mathcal{T}}$	CoLaR (initial)	CoLaR	SCCA-ADMM
			$\hat{u}_1$				
Error 1	0.2024	0.1034	0.1767	0.0865	0.1788	0.0873	0.0207
Error 2	0.1593	0.0739	0.1417	0.0661	0.1424	0.0658	0.0343
TP	3	4	4	4	4	4	4
TN	195	186	195	185	195	185	192
MCC	0.7707	0.4774	0.7707	0.4581	0.7707	0.4581	0.6665
			$\hat{v}_1$				
Error 1	0.4025	0.0694	0.3614	0.0579	0.3646	0.0577	0.0343
Error 2	0.3895	0.0664	0.3467	0.0563	0.3484	0.0567	0.0354
TP	2	4	2	4	2	4	5
TN	195	188.5	195	189	195	189	192
MCC	0.6276	0.5834	0.6276	0.5837	0.6276	0.5955	0.7228

Table 4.2: Method comparison under the single case on Student's t distribution.

	R-CoLaR$^{\mathcal{H}}$ (initial)	R-CoLaR$^{\mathcal{H}}$	R-CoLaR$^{\mathcal{T}}$ (initial)	R-CoLaR$^{\mathcal{T}}$	CoLaR (initial)	CoLaR	SCCA-ADMM
			$\hat{u}_1$				
Error 1	0.2790	0.0999	0.2511	0.0893	2.0000	1.9594	2.0000
Error 2	0.2485	0.0889	0.2142	0.0802	1.3086	1.3224	1.4129
TP	3	4	3	4	0	1	0
TN	195	192	195	188	192.5	172	182
MCC	0.7707	0.6210	0.7707	0.5233	-0.0114	0.0031	-0.0367
			$\hat{v}_1$				
Error 1	0.4721	0.1006	0.4721	0.0810	2.0000	1.9951	2.0000
Error 2	0.4952	0.1097	0.4464	0.0906	1.3246	1.3532	1.3686
TP	1	4	1	4	0	1	0
TN	195	192	195	190	193	172.5	183
MCC	0.4427	0.6665	0.4427	0.6210	-0.0114	-0.0229	-0.0386

Table 4.3: Method comparison under the single case on Pareto distribution.

	R-CoLaR$^{\mathcal{H}}$ (initial)	R-CoLaR$^{\mathcal{H}}$	R-CoLaR$^{\mathcal{T}}$ (initial)	R-CoLaR$^{\mathcal{T}}$	CoLaR (initial)	CoLaR	SCCA-ADMM
			$\hat{u}_1$				
Error 1	0.7073	0.2002	0.7826	0.2794	2.0000	1.9999	2.0000
Error 2	0.7480	0.3219	0.5925	0.3210	1.3419	1.3351	1.4013
TP	2	4	2	3	0	0.5	0
TN	195	194	194	189	190.5	177	188
MCC	0.6276	0.7707	0.4427	0.5111	-0.0161	-0.0246	-0.0269
			$\hat{v}_1$				
Error 1	0.4721	0.2500	0.4721	0.2699	2.0000	2.0000	2.0000
Error 2	0.6870	0.3999	0.5905	0.3062	1.2734	1.2754	1.3105
TP	1	3	1	4	0	0	0
TN	195	193.5	194.5	189	192	179	188
MCC	0.4427	0.6634	0.4427	0.5111	-0.0161	-0.0327	-0.0305

Table 4.4: Method comparison under the single case on Log-normal distribution.

	R-CoLaR$^{\mathcal{H}}$ (initial)	R-CoLaR$^{\mathcal{H}}$	R-CoLaR$^{\mathcal{T}}$ (initial)	R-CoLaR$^{\mathcal{T}}$	CoLaR (initial)	CoLaR	SCCA-ADMM
			$\hat{u}_1$				
Error 1	0.4275	0.1486	0.3640	0.1784	2.0000	1.9943	2.0000
Error 2	0.4504	0.1745	0.3139	0.1371	1.3293	1.3125	1.3537
TP	2	4	2	4	0	1	0
TN	195	193.5	195	189	191.5	173	185
MCC	0.6276	0.7228	0.6276	0.5510	-0.0114	0.0083	-0.0327
			$\hat{v}_1$				
Error 1	0.4721	0.1790	0.4721	0.1426	2.0000	1.9975	2.0000
Error 2	0.5653	0.2210	0.4572	0.1494	1.3645	1.3666	1.3626
TP	1	4	1.5	4	0	1	0
TN	195	193	195	189	190.5	173	185
MCC	0.4427	0.6665	0.4427	0.5243	-0.0161	0.0010	-0.0348

4.3.1.2 Case II: two pairs of canonical coefficient vectors

For the multiple case, the collections of true canonical coefficient vectors are denoted as $U = [u_1, u_2]$ and $V = [v_1, v_2]$, and the rank of $\mathbf{\Sigma}_{xy}$ is $r = 2$. We set the true canonical correlations $\lambda_1 = 0.9$ and $\lambda_2 = 0.8$. Both R-CoLaR and CoLaR were implemented using the same procedure and tuning parameters as those in Section 4.3.1.1 for the single case. Here, we did not include SCCA-ADMM as a comparison because it was mainly developed for the single case.

Similar to the single case, we at first evaluated the accuracy of estimated canonical coefficient vectors $\hat{U} = [\hat{u}_1, \hat{u}_2]$ and $\hat{V} = [\hat{v}_1, \hat{v}_2]$ using the following two different types of errors which generalize their versions in Section 4.3.1.1 to matrix cases:

- Error 1: $||\hat{A}\hat{A}^\intercal - AA^\intercal||_F^2$, where $\hat{A}$ represents r left singular vectors after SVD of $\hat{U}$ (or $\hat{V}$), and A denotes r left singular vectors after SVD of U (or V) (Suo et al., 2017; Chen et al., 2019) with $r = 2$ here.
- Error 2: $\inf_{W \in O(r)} ||\mathbf{\Sigma}^{1/2}(\hat{Z}W - Z)||_F^2$ with $\mathbf{\Sigma} = \mathbf{\Sigma}_x$ (or $\mathbf{\Sigma}_y$), $\hat{Z} = \hat{U}$ (or $\hat{V}$) and $Z = U$ (or V) (Gao et al., 2017), where $O(r)$ denotes the set of all $r \times r$ orthogonal matrices with $r = 2$ here.

Error 1 measures the difference between the subspace spanned by canonical coefficient vectors $\hat{A}$ and A. Error 2 is the prediction loss where we need to find a W from all possible 2×2 orthogonal matrices that minimizes its quantity. Then, we evaluated the performance of feature selection on the estimated canonical coefficient vectors $\hat{U}$ and $\hat{V}$. We also kept only the non-zero elements whose absolute magnitudes are greater than 1×10^{-4}. Unlike the single case in Section 4.3.1.1, we here measured the rows of $\hat{U}$ and $\hat{V}$ with slightly different definitions of TP and TN. Three metrics were also adopted to measure the performance of identifying true signals while simultaneously controlling false discovers,

- TP: the number of true non-zero rows (e.g. both $\hat{u}_1$ (or $\hat{v}_1$) and $\hat{u}_2$ (or $\hat{v}_2$) are non-zero) in $\hat{U}$ (or $\hat{V}$) compared to those in U (or V).
- TN: the number of true zero rows (e.g. either $\hat{u}_1$ (or $\hat{v}_1$) or $\hat{u}_2$ (or $\hat{v}_2$) is zero) in $\hat{U}$ (or $\hat{V}$) compared to those in U (or V).
- MCC: $= \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP}) \times (\mathrm{TP}+\mathrm{FN}) \times (\mathrm{TN}+\mathrm{FP}) \times (\mathrm{TN}+\mathrm{FN})}}$.

In the ground truth of U and V, the number of true non-zero and zero rows are 5 and 195 respectively.

Tables 4.5-4.8 report medians of all the metrics from R-CoLaR and CoLaR on four different distributions for the multiple case. From Table 4.5, it is expected to see that all the metrics in terms of accuracy and feature selection are quite similar among R-CoLaR$^{\mathcal{H}}$ (initial), R-CoLaR$^{\mathcal{T}}$ (initial) and CoLaR (initial) on Normal distribution. Likewise, R-CoLaR$^{\mathcal{H}}$, R-CoLaR$^{\mathcal{T}}$ and CoLaR show similar results as well. R-CoLaR shows significant advantages over CoLaR on heavy-tailed distributions, as shown in Tables 4.6-4.8. In terms of accuracy, Errors 1 and 2 from R-CoLaR$^{\mathcal{H}}$ (initial) and R-CoLaR$^{\mathcal{T}}$ (initial) are noticeably lower than those from CoLaR (initial) and CoLaR. The values of Error 1 from CoLaR (initial) and CoLaR are all close to 4 (the maximum value), implying that the subspace spanned by the estimated canonical coefficient vectors is nearly orthogonal to the ground truth. Moreover, R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ after the refined stage further improve accuracy and significantly lower the two errors. On asymmetric heavy-tailed distributions like Pareto and Log-normal, R-CoLaR$^{\mathcal{H}}$ with Huber-type M-estimators can be even more accurate than R-CoLaR$^{\mathcal{T}}$ with truncation-based estimators. This may be due to that those asymmetric distributions could amplify the subtle difference between the two types of estimators discussed at the end of Section 4.2.1. In terms of feature selection, R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ are capable of consistently identifying 4 or 5 out of 5 true positives while at the same time maintaining high true negatives as well, and they are more powerful than R-CoLaR$^{\mathcal{H}}$ (initial) and R-CoLaR$^{\mathcal{T}}$ (initial) in identifying true signals particularly on Pareto and Log-normal distributions. Both CoLaR (initial) and CoLaR fail to identify all the true positives, and CoLaR also leads to high false positives in terms of low values of TNs. As for MCCs, R-CoLaR is obviously better than CoLaR whose values are very close to 0 and equivalent to a random guess, and R-CoLaR (initial) can be better than R-CoLaR due to its more balanced TPs and TNs. Moreover, it is worthwhile to notice that R-CoLaR$^{\mathcal{H}}$ can be even better on Pareto and Log-normal distributions considering both accuracy and feature selection.

To summarize, R-CoLaR is equivalently good as CoLaR on Normal distribution and particularly superior than CoLaR on the heavy-tailed distributions in terms of both accuracy and feature selection. Under the case of two pairs of canonical coefficient vectors, R-CoLaR$^{\mathcal{H}}$

with Huber-type M-estimators shows even better performance on asymmetric heavy-tailed distributions.

Table 4.5: Method comparison under the multiple case on Normal distribution.

	R-CoLaR$^{\mathcal{H}}$ (initial)	R-CoLaR$^{\mathcal{H}}$	R-CoLaR$^{\mathcal{T}}$ (initial)	R-CoLaR$^{\mathcal{T}}$	CoLaR (initial)	CoLaR
			$\hat{U}$			
Error 1	0.9343	0.2361	0.7409	0.1828	0.7505	0.1814
Error 2	0.4905	0.1216	0.3781	0.0934	0.3829	0.0942
TP	5	5	5	5	5	5
TN	194	180	195	181	195	181
MCC	0.8921	0.4804	0.9013	0.4942	0.9105	0.4942
			$\hat{V}$			
Error 1	0.8740	0.2470	0.6797	0.1915	0.6833	0.1910
Error 2	0.4659	0.1289	0.3636	0.0983	0.3698	0.0987
TP	5	5	5	5	5	5
TN	194	180	194.5	179	195	180
MCC	0.8921	0.4804	0.9105	0.4675	0.9105	0.4804

4.3.2 Application to CITE-seq data of a MALT tumor

We applied R-CoLaR, CoLaR and SCCA-ADMM to the CITE-seq data mentioned in Section 4.1 from a MALT tumor stained with 17 TotalSeq-B antibodies, including CD3, CD4, CD8a, CD14, CD15, CD16, CD56, CD19, CD25, CD45RA, CD45RO, PD-1, TIGIT, CD127, IgG2a, IgG1 and IgG2b. The estimated first pairs of canonical coefficient vectors were compared among all the approaches.

The data set was downloaded from the 10X Genomics website: `https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/malt_10k_protein_v3`.

Table 4.6: Method comparison under the multiple case on Student's t distribution.

	R-CoLaR$^{\mathcal{H}}$ (initial)	R-CoLaR$^{\mathcal{H}}$	R-CoLaR$^{\mathcal{T}}$ (initial)	R-CoLaR$^{\mathcal{T}}$	CoLaR (initial)	CoLaR
			$\hat{U}$			
Error 1	1.4155	0.3520	1.1726	0.3211	3.9964	3.9775
Error 2	0.8879	0.2302	0.7028	0.1961	2.4166	2.4738
TP	4	5	4	5	1	2
TN	195	189	194	181	184	151
MCC	0.8408	0.6637	0.7949	0.4942	0.0014	0.0363
			$\hat{V}$			
Error 1	1.2797	0.3308	1.0849	0.2737	3.9960	3.9736
Error 2	0.8733	0.2050	0.6728	0.1508	2.5746	2.5592
TP	4	5	5	5	1	3
TN	195	187	194	182	181	146.5
MCC	0.8921	0.6073	0.7845	0.5092	0.0478	0.0832

Table 4.7: Method comparison under the multiple case on Pareto distribution.

	R-CoLaR$^{\mathcal{H}}$ (initial)	R-CoLaR$^{\mathcal{H}}$	R-CoLaR$^{\mathcal{T}}$ (initial)	R-CoLaR$^{\mathcal{T}}$	CoLaR (initial)	CoLaR
			$\hat{U}$			
Error 1	2.0726	1.0778	2.7174	2.2171	4.0000	3.9957
Error 2	1.9533	0.8978	2.2754	1.9973	2.7471	2.7162
TP	2	4	2	4	0	2
TN	195	193	193	181	184.5	157
MCC	0.6276	0.7845	0.4346	0.3850	-0.0161	0.0273
			$\hat{V}$			
Error 1	2.0635	0.9775	2.5425	2.2802	4.0000	3.9980
Error 2	1.8959	0.8535	2.2120	1.9350	2.6165	2.5952
TP	2	4	2	4	0	1
TN	195	193	193	181	188	161.5
MCC	0.6276	0.7845	0.5072	0.3683	-0.0198	-0.0010

Table 4.8: Method comparison under the multiple case on Log-normal distribution.

	R-CoLaR$^{\mathcal{H}}$ (initial)	R-CoLaR$^{\mathcal{H}}$	R-CoLaR$^{\mathcal{T}}$ (initial)	R-CoLaR$^{\mathcal{T}}$	CoLaR (initial)	CoLaR
			$\hat{U}$			
Error 1	1.7294	0.4070	2.2188	2.0928	4.0000	3.9901
Error 2	1.1631	0.2785	1.6271	1.5221	2.6824	2.6590
TP	3	5	3	5	0	2
TN	195	191	193	177.5	185	153
MCC	0.7707	0.7377	0.6307	0.3794	-0.0161	0.0533
			$\hat{V}$			
Error 1	1.5007	0.3786	2.2376	2.0671	4.0000	3.9926
Error 2	1.0827	0.2817	1.6175	1.4833	2.6753	2.6696
TP	3	5	3	5	0	2
TN	195	191	193	178	183.5	151
MCC	0.7707	0.7377	0.6118	0.4079	-0.0161	0.0401

Following a standard workflow of pre-processing from Seurat (Stuart et al., 2019), we filtered cells that have unique gene counts over 2500 or less than 200 and that have $> 5\%$ mitochondrial counts. Then, we took log-normalization of UMI counts for the single-cell RNA-seq and extracted top 1000 highly variable genes based on the algorithm implemented by Seurat that accounts for mean-variance relationship (Stuart et al., 2019). Unlike the typical log-normalization, the ADT data was normalized by a centered log-ratio for each antibody on the same set of cells as recommended by Stuart et al. (2019). In the end, the normalized RNA-seq data with $n = 736$ cells and $p = 1000$ most highly variable genes and the normalized ADT data with $q = 17$ antibodies on the same set of cells were used for further analysis.

When implementing R-CoLaR$^{\mathcal{H}}$, R-CoLaR$^{\mathcal{T}}$ and CoLaR, we used the same tuning parameter $\rho = 0.5\sqrt{\log(p \vee q)/n}$ in the initial stage. In the refined stage, we used five-fold cross validation to select a value of b from the candidate set $b = \{0.5, 1, 1.5, 2\}$ for $\rho_u = b\sqrt{(r + \log(p))/n}$ and $\rho_v = b\sqrt{(r + \log(q))/n}$ of R-CoLaR and for the common penalty $\rho_u = \rho_v = b\sqrt{(r + \log(p \vee q))/n}$ of original CoLaR with $r = 1$. Similarly, the common value of b_1 for penalty levels $\tau_u = b_1\sqrt{\log(p)/n}$ and $\tau_v = b_1\sqrt{\log(q)/n}$ of SCCA-ADMM was also selected via five-fold cross validation from the candidate set $b_1 = \{0.5, 1, 1.5, 2\}$. Specifically, four folds were used as a training set $(\boldsymbol{X}^{(l)}_{\text{train}}, \boldsymbol{Y}^{(l)}_{\text{train}})$, and one fold was used as a testing set $(\boldsymbol{X}^{(l)}_{\text{test}}, \boldsymbol{Y}^{(l)}_{\text{test}})$ with $l = 1, 2, ..., 5$. The first pair of canonical coefficient vectors $(\hat{u}^{(l)}_1, \hat{v}^{(l)}_1)$ was obtained from the training set. Then, the value of b (or b_1) was selected as the one that maximizes the canonical correlation on the projected data $(\boldsymbol{X}^{(l)}_{\text{test}}\hat{u}^{(l)}_1, \boldsymbol{Y}^{(l)}_{\text{test}}\hat{v}^{(l)}_1)$ with the testing set. The canonical correlations of R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ were calculated from the robust covariance of $(\boldsymbol{X}^{(l)}_{\text{test}}\hat{u}^{(l)}_1, \boldsymbol{Y}^{(l)}_{\text{test}}\hat{v}^{(l)}_1)$ using the Huber-type M-estimator and the element-wise truncated estimator respectively. Finally, the estimated $(\hat{u}_1, \hat{v}_1)$ whose non-zero elements have absolute magnitudes greater than 1×10^{-4} was obtained from the selected b and b_1 using the complete data.

Table 4.9 demonstrates the number of genes and ADTs identified from all the aforementioned approaches. It can be seen that R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ can generally detect more genes and ADTs compared to CoLaR and SCCA-ADMM, and SCCA-ADMM is particularly conservative in identifying functionally important genes. Specific lists of ADTs and genes

Table 4.9: The identified number of genes and ADTs from all the approaches.

	Number of genes	Number of ADTs
R-CoLaR$^{\mathcal{H}}$	40	7
R-CoLaR$^{\mathcal{T}}$	35	8
CoLaR	35	6
SCCA-ADMM	20	5

for all the approaches are shown in Tables 4.10 and 4.11 respectively, where boldface with an asterisk represents a gene or an ADT that is not identified by CoLaR or SCCA-ADMM, and boldface with two asterisks is denoted as a gene or an ADT that fails to be identified by both CoLaR and SCCA-ADMM.

In terms of ADTs on cell surface in Table 4.10, R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ are capable of identifying the PD-1 protein which is expressed on the cell surface of T lymphocytes. Ample evidence has confirmed that markedly elevated PD-1 levels in tumor-infiltrating T cells inhibit immune responses, and clinical PD-1 blockade is a promising immunotherapy for B-cell lymphomas (Xu-Monette et al., 2018). However, SCCA-ADMM fails to detect this important protein. Furthermore, R-CoLaR can identify the TIGIT and the CD45RA proteins that are important to B-cell lymphomas, while both CoLaR and SCCA-ADMM are not capable of detecting them. The TIGIT protein has also been shown to mark intratumoral T cells, and it is identified as a frequently expressed coinhibitory receptor of PD-1. Thus, TIGIT and PD-1 coblockade deserves to be further studied to promote antitumor responses for B-cell non-Hodgkin lymphoma, according to Josefsson et al. (2019). The CD45RA protein has been shown to typically exhibit positive immunophenotypes in the tumor cells among patients diagnosed as low-grade MALT lymphoma of the nasopharynx (El-Banhawy and El-Desoky, 2005). Therefore, R-CoLaR$^{\mathcal{H}}$ and R-CoLaR$^{\mathcal{T}}$ are more powerful in capturing functionally important proteins related to MALT or other types of B-cell lymphomas in addition to the

ones captured by CoLaR and SCCA-ADMM.

In terms of genes in Table 4.11, all the methods can detect ITM2A that regulates im-

Table 4.10: Lists of identified ADTs from all the approaches (Boldface with an asterisk: not appeared in CoLaR or SCCA-ADMM; Boldface with two asterisks: not appeared in both CoLaR and SCCA-ADMM).

ADT			
R-CoLaR$^{\mathcal{H}}$	R-CoLaR$^{\mathcal{T}}$	CoLaR	SCCA-ADMM
CD3	CD3	CD3	CD3
CD4	CD4	CD4	CD4
CD19	CD19	CD19	CD19
CD45RO	CD45RO	CD45RO	CD45RO
PD-1*	**PD-1***	PD-1	CD127
TIGIT**	**TIGIT****	CD127	
CD127	CD127		
	CD45RA**		

mune response and has high expression in CD4+ T cells. The gene has also recently been shown as generally co-expressed with PD-1 (Andor et al., 2019). However, R-CoLaR can identify more important genes related to MALT lymphoma such as HSP90AA1, RGCC and PIK3R1, compared to SCCA-ADMM. HSP90AA1, which belongs to the HSP90 family of genes, closely interacts with BCL-6 gene (Cerchietti et al., 2009) that is suggested as a marker for transformation of MALT lymphoma (Flossbach et al., 2011) and that is believed to facilitate the proliferation of CD19+ B lymphocytes and the differentiation of CD4+ T cells (Crotty et al., 2010). RGCC, which may contribute to regulate the cell cycle of CD4+ T cells (Tegla et al., 2015), enhances the activity of CDK1 gene that is believed to be actively associated with the evolution of *H.pylori*-associated gastritis to MALT lymphoma in the modulation of cellular death by apoptosis, cellular proliferation and transformation (Banerjee et al., 2000). PIK3R1 is an important gene within the PI3K/AKT pathway which

is negatively regulated by PD-1 activity in malignant T cells (Wartewig and Ruland, 2019). Then, compared to CoLaR, R-CoLaR$^{\mathcal{H}}$ can further identify ANXA1 gene which plays a role in increasing proliferation and activation of CD3 T cells (D'Acquisto et al., 2007). It is also a specific marker of hairy cell leukemia (Falini et al., 2004) and can be helpful to distinguish it from other B-cell lymphomas like MALT. Moreover, R-CoLaR$^{\mathcal{H}}$ is capable of identifying additional important genes like GZMK and HLA-DQB1 compared to both CoLaR and SCCA-ADMM. GZMK is over-expressed in cytotoxic T cells with CD3 or CD4 antibodies, which reflects a host anti-lymphoma response (Schuhmacher et al., 2016). HLA-DQB1, which binds peptides derived from antigens and presents them on the cell surface to be recognized by CD4 T cells, is shown to have a high prevalence of mutation among patients with MALT lymphoma according to Filip et al. (2018). Therefore, R-CoLaR can detect more functionally important genes associated with cell surface proteins for MALT and other B-cell lymphomas than CoLaR and SCCA-ADMM.

Taken together, the proposed method R-CoLaR is better in interpreting the gene-protein relationship for the mechanism of MALT or other types of B-cell lymphomas than CoLaR and SCCA-ADMM.

4.4 Conclusion and discussion

We have developed a novel robust sparse CCA procedure R-CoLaR that extends its application to data with heavy-tailed distributions. In simulation studies, compared to existing sparse CCA methods such as CoLaR and SCCA-ADMM, R-CoLaR shares similar performance under non-heavy-tailed distributions and shows noticeable advantages with lower errors of estimated canonical coefficient vectors and more accurate feature selection of nonzero coordinates under heavy-tailed distributions. In application of the CITE-seq data, R-CoLaR is more powerful in identifying sets of genes correlated with cell surface proteins, which interprets the mechanism of MALT better than CoLaR and SCCA-ADMM. In practical use, R-CoLaR$^{\mathcal{H}}$ with Huber-type M-estimators is recommended in a case with relatively small n due to its even better performance with finite sample sizes. When the sample size n becomes

larger, R-CoLaR$^{\mathcal{T}}$ with element-wise truncated estimators is a better option because it is more computationally efficient with no need of mean estimation.

Table 4.11: Lists of identified genes from all the approaches (Boldface with an asterisk: not appeared in CoLaR or SCCA-ADMM; Boldface with two asterisks: not appeared in both CoLaR and SCCA-ADMM).

Gene			
R-CoLaR$^{\mathcal{H}}$	R-CoLaR$^{\mathcal{T}}$	CoLaR	SCCA-ADMM
IGKC, CD79A	IGKC, CD79A	IGKC, CD79A	IGKC, CD79A
IGHM, HLA-DRA	IGHM, HLA-DRA	IGHM, HLA-DRA	IGHM, HLA-DRA
GZMK**, CD74	CD74, CD83	CD74, CD83	CD74, ANXA1
ANXA1*, **IGHA1****	**NEAT1***, **IGLC2***	NEAT1, IGLC2	CD83, HLA-DPA1
CD83, **NEAT1***	HLA-DPA1, MS4A1	HLA-DPA1, MS4A1	MS4A1, HLA-DPB1
IGLC2*, HLA-DPA1	HLA-DPB1, IL32	HLA-DPB1, IL32	IL32, RGS1
MS4A1, HLA-DPB1	**YBX3***, **GRASP***	YBX3, GRASP	ITM2A, TRBC1
HLA-DQB1**, IL32	RGS1, **MAF***	RGS1, MAF	IL7R, GAPDH
YBX3*, **GRASP***	**HSP90AA1***, ITM2A	HSP90AA1, ITM2A	SRGN, TSPYL2
RGS1, **MAF***	**BATF***, TRBC1	BATF, TRBC1	KLF6, TRBC2
HSP90AA1*, ITM2A	**BANK1***, IL7R	BANK1, IL7R	
BATF*, TRBC1	**RTKN2***, **CD48***	RTKN2, CD48	
BANK1*, IL7R	GAPDH, SRGN	GAPDH, SRGN	
RTKN2*, **CD48***	TSPYL2, **RGCC***	TSPYL2, RGCC	
GAPDH, SRGN	**FKBP5***, **H2AFZ***	FKBP5, H2AFZ	
TSPYL2, **RGCC***	**PIK3R1***, KLF6	PIK3R1, KLF6	
FKBP5*, **CARD16****	**IL6ST***, TRBC2	IL6ST, TRBC2	
H2AFZ*, **PIK3R1***	**TRAT1***	TRAT1	
KLF6, **IL6ST***			
TRBC2, **TRAT1***			

5.0 Discussion and Future Works

5.1 Discussion

Gene co-expression network analysis and multi-omics study are two popular types of studies in biological research. The high dimensions, the discreteness and even the heavy-tailed distributions of omics data have posed great challenges in both computation and methodology. To address the computational challenge, we have developed an efficient and unified package `SILGGM` for statistical inference of high-dimensional Gaussian graphical model in large-scale gene network analysis in Chapter 2. To further address the challenge from discrete and non-negative omics data in biological network analysis, we have proposed a novel two-step procedure for statistical inference of high-dimensional modified Poisson-type graphical models in Chapter 3. To cope with high-dimensional omics data even with the heavy-tailed phenomenon in multi-omics study, we have proposed a novel robust sparse CCA procedure named with R-CoLaR in Chapter 4.

5.2 Future works

For Chapter 2, we will add parallel computing to `SILGGM` so as to allow users to use multiple clusters for bigger data analysis since the droplet-based single-cell technology will further increase the sample size (Macosko et al., 2015). In addition, the new feature for the rigorous statistical inference of high-dimensional multiple gene networks is another potential extension of our package because differential gene network analysis among different cell types or cells of multiple individuals is being paid more attention to.

For Chapter 3, there are several limitations of our proposed method which need future study. On the one hand, the proposed method is not symmetric between i and j in estimating each θ_{ij}, and generally depends on the ordering of variables. One can naively apply a sample splitting scheme for symmetrization. More specifically, we randomly split the data into two

halves. Then for each fixed pair $i < j$, we fit the first half of the data into our method to obtain estimator $\tilde{\theta}_{ij}$, and then apply the second half to our method with i and j switched to obtain $\tilde{\theta}_{ji}$. The final asymptotically normal estimator is the average of these two independent estimators $\tilde{\theta}_{ij}^{sym} = (\tilde{\theta}_{ij} + \tilde{\theta}_{ji})/2$. However, both that sample splitting scheme only uses part of the data for inference and that the result depends on the random split of the data make it less preferred in practice. Some preliminary analysis suggests that sample splitting is not necessary for asymptotic normality of $\tilde{\theta}_{ij}^{sym}$ but the dependency between $\tilde{\theta}_{ij}$ and $\tilde{\theta}_{ji}$ obtained with the same entire samples requires a refined theoretical analysis. We thus leave it as a future work. On the other hand, our method allows only a single discrete-type data set as an input. Due to the increasing popularity of multi-omics study, the integrative network analysis of multi-layered data sets with both continuous and discrete values is a promising future direction. To this end, we will further expand our procedure to more generalized or mixed-type exponential family graphical models as a future work.

For Chapter 4, there are several potential extensions of R-CoLaR that need future study. On the one hand, we intend to incorporate phenotype (e.g. disease status, cell status etc.) of subjects into the current framework of R-CoLaR as a guidance for results also associated with phenotype of each observation. On the other hand, R-CoLaR can be extended to a multiple framework that allows more different types of omics data for analysis. In this sense, how to determine the canonical correlation with more than two data sets will be also an interesting problem that needs to be addressed.

Appendix A Supplement to Chapter 2

A.1 Theoretical procedures of each method included in the package SILGGM

Without loss of generality, we assume that $\boldsymbol{X} = (X_1, ..., X_p)$ is an $n \times p$ matrix, where each row vector $(X_{k1}, ..., X_{kp})^\intercal$ for $1 \leq k \leq n$ follows a p-dimensional independently and identically multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ. The precision matrix is denoted as $\Omega = (\omega_{ij})_{p\times p} = \Sigma^{-1}$, where $i, j = 1, 2, ..., p$.

A.1.1 The bivariate nodewise scaled Lasso

B_NW_SL (Ren et al., 2015) is to make inference on each ω_{ij} with $i \neq j$. Based on the bivariate conditional normal distribution with index set $A = \{i, j\}$,

$$\boldsymbol{X}_A | \boldsymbol{X}_{A^c} \sim \mathcal{N}\left(-\Omega_{A,A}^{-1}\Omega_{A,A^c}, \Omega_{A,A}^{-1}\right), \qquad \Omega_{A,A} = \begin{pmatrix} \omega_{ii} & \omega_{ij} \\ \omega_{ji} & \omega_{jj} \end{pmatrix}, \tag{A.1}$$

the bivariate nodewise scaled Lasso regression of the two variables in A against the other variables in A^c is proposed,

$$\underset{\beta\in\mathbb{R}^{p-2},\sigma>0}{\arg\min} \left\{ \frac{||\boldsymbol{X}_m - \boldsymbol{X}_{A^c}\beta||^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \sum_{k\in A^c} \frac{||X_k||}{\sqrt{n}} |\beta_k| \right\}, \quad m \in A = \{i, j\}. \tag{A.2}$$

Here, each run of scaled Lasso regression is tuning-free, and the tuning parameter is taken as $\lambda = \sqrt{2\log(p/\sqrt{n})/n}$. The estimated residual $\hat{\epsilon}_A = \boldsymbol{X}_A - \boldsymbol{X}_{A^c}\hat{\beta}_A$ can be obtained once $\hat{\beta}_A$ is estimated from (A.2). Then, Ω for variables i and j can be estimated:

$$\hat{\Omega}_{A,A} = \begin{pmatrix} \hat{\omega}_{ii} & \hat{\omega}_{ij} \\ \hat{\omega}_{ji} & \hat{\omega}_{jj} \end{pmatrix} = \left(\frac{1}{n}\hat{\epsilon}_A^\intercal \hat{\epsilon}_A\right)^{-1}, \quad A = \{i, j\}. \tag{A.3}$$

Under the minimal sparseness assumption $s = o(\sqrt{n}/\log(p))$, each estimator $\hat{\omega}_{ij}$ has been shown asymptotically normal and efficient,

$$\sqrt{n(\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2)^{-1}}\,(\hat{\omega}_{ij} - \omega_{ij}) \xrightarrow{D} \mathcal{N}(0, 1). \tag{A.4}$$

According to (A.4), we can estimate the corresponding p-value and confidence interval of each ω_{ij}. The naïve implementation of the procedure requires $O(p^2)$ runs of scaled Lasso regression, but the total number of runs of regression can be reduced to $O(sp)$ in terms of the comments in Ren et al. (2015) and the implementation in Wang et al. (2016).

A.1.2 The de-sparsified nodewise scaled Lasso

D-S_NW_SL (Janková and van de Geer, 2017) is based on the p runs of nodewise scaled Lasso regression for i^{th} variable against all the other variables i^c,

$$\underset{\beta_i \in \mathbb{R}^{p-1}, \sigma > 0}{\arg\min} \left\{ \frac{||X_i - \boldsymbol{X}_{i^c}||^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \sum_{k \in i^c} |\beta_{ik}| \right\}. \tag{A.5}$$

Again, the tuning parameter for each run of regression is taken as $\lambda = \sqrt{2\log(p/\sqrt{n})/n}$. Unlike the previous B_NW_SL, the procedure deals with the coefficients rather than the regression noise. If $\hat{\beta}_i$'s are the estimated coefficients from (A.5), we can define $\hat{\sigma}_i^2 = ||X_i - \boldsymbol{X}_{i^c}\hat{\beta}_i||^2/n$, $\tilde{\sigma}_i^2 = \hat{\sigma}_i^2 + \lambda\hat{\sigma}_i||\hat{\beta}_i||_1$ and $\hat{B}_i = (-\hat{\beta}_{i,1}, ..., -\hat{\beta}_{i,i-1}, 1, -\hat{\beta}_{i,i+1}, ..., -\hat{\beta}_{i,p})^\intercal$. Then, the i^{th} column of Ω can be estimated:

$$\hat{\omega}_i = \hat{B}_i/\tilde{\sigma}_i^2. \tag{A.6}$$

However, it is well known that the initial estimators in (A.6) have bias, so the authors have proposed a bias correction procedure on $\hat{\omega}_{ij}$. Under the Karush-Kuhn-Tucker (KKT) conditions, the desparsified (or de-biased) estimator $\check{\omega}_{ij}$ is

$$\check{\omega}_{ij} = \hat{\omega}_{ij} + \hat{\omega}_{ji} - \hat{\omega}_i^\intercal \hat{\Sigma} \hat{\omega}_j, \tag{A.7}$$

where $\hat{\Sigma} = \boldsymbol{X}^\intercal \boldsymbol{X}/n$.

Under the minimal sparseness assumption $s = o(\sqrt{n}/\log(p))$, each de-biased estimator $\check{\omega}_{ij}$ has achieved the asymptotically efficient result with

$$\sqrt{n(\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2)^{-1}} \, (\check{\omega}_{ij} - \omega_{ij}) \xrightarrow{D} \mathcal{N}(0, 1). \tag{A.8}$$

According to (A.8), the corresponding p-value and confidence interval can be estimated for each ω_{ij}.

A.1.3 The de-sparsified graphical Lasso

D-S_GL (Janková and van de Geer, 2015) also depends on a bias correction procedure which is very similar to the one in D-S_NW_SL. However, the initial estimator $\Omega = (\hat{\omega}_{ij})_{p\times p}$ here is obtained by solving a graphical Lasso optimization problem for Ω:

$$\arg\min_{\Omega} \left\{ \mathrm{Tr}(\Omega^{\intercal}\hat{\Sigma}) - \log\det(\Omega) + \lambda||\Omega||_{1,\mathrm{off}} \right\}. \tag{A.9}$$

Even though (A.9) is not tuning-free, the tuning parameter can be taken as $\lambda = \sqrt{\log(p)/n}$ according to the suggestion in Janková and van de Geer (2015). Then, with the same idea of bias correction in D-S_NW_SL, the desparsified (or de-biased) estimator $\check{\Omega} = (\check{\omega}_{ij})_{p\times p}$ is

$$\check{\Omega} = 2\hat{\Omega} - \hat{\Omega}\hat{\Sigma}\hat{\Omega}. \tag{A.10}$$

Under the minimal sparseness assumption $s = o(\sqrt{n}/\log(p))$, each de-biased estimator $\check{\omega}_{ij}$ achieves the same asymptotically efficient result as the one in (A.8).

A.1.4 The Gaussian graphical model (GGM) estimation with false discovery rate (FDR) control using scaled Lasso or Lasso

While the previous three methods are originally developed for individual inference of each ω_{ij}, GFC_SL or GFC_L (Liu, 2013) is proposed particularly for global inference of all ω_{ij}'s. The approach is based on a bias correction procedure on the sample covariance of residuals between each pair of variables i and j. In order to obtain the estimators of residuals, the first step of the method needs p runs of nodewise scaled Lasso regression same as those in (A.5) or nodewise Lasso regression for i^{th} variable against all the other variables i^c as below,

$$\arg\min_{\beta_i\in\mathbb{R}^{p-1}} \left\{ \frac{||X_i - \boldsymbol{X}_{i^c}\beta_i||^2}{2n} + \lambda_i \sum_{k\in i^c} |\beta_{ik}| \right\}. \tag{A.11}$$

If the estimated coefficients $\hat{\beta}_i$'s are obtained from (A.5) or (A.11), then we can obtain the estimated residual $\hat{\epsilon}_i = X_i - \boldsymbol{X}_{i^c}\hat{\beta}_i$ and the estimated sample covariance of residuals between

$(i,j)^{th}$ pair of variables $\hat{r}_{ij} = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_{ki}\hat{\epsilon}_{kj}$. The second step is to make a bias correction on $\hat{r}_{ij}$ to obtain

$$T_{ij} = \frac{1}{n}\left(\sum_{k=1}^{n} \hat{\epsilon}_{ki}\hat{\epsilon}_{kj} + \sum_{k=1}^{n} \hat{\epsilon}_{ki}^2\hat{\beta}_{ji} + \sum_{k=1}^{n} \hat{\epsilon}_{kj}^2\hat{\beta}_{i(j-1)}\right), 1 \leq i < j \leq p \tag{A.12}$$

and construct a new test statistic

$$\hat{T}_{ij} = \sqrt{\frac{n}{\hat{r}_{ii}\hat{r}_{jj}}}T_{ij} \tag{A.13}$$

for the multiple testing

$$H_0 : \omega_{ij} = 0 \quad \text{vs}. \quad H_1 : \omega_{ij} \neq 0. \tag{A.14}$$

Under the null hypothesis in (A.14) and the same minimal sparseness assumption as before, (A.13) has an asymptotically normal result with

$$\hat{T}_{ij} \xrightarrow{D} \mathcal{N}(0,1). \tag{A.15}$$

Since GFC_SL or GFC_L is developed for global inference, another main component of this method is to provide a novel framework for FDR control that has been theoretically proved valid in high-dimensional settings. It is well known that the false discovery proportion (FDP) with a threshold t can be written as

$$\text{FDP}(t) = \sum_{(i,j)\in H_0} I\left\{|\hat{T}_{ij}| \geq t\right\} / \max\left\{\sum_{1\leq i<j\leq p} I\{|\hat{T}_{ij}| \geq t\}, 1\right\}. \tag{A.16}$$

To control FDR needs to control (A.16) since we have $E(\text{FDP}(t)) = \text{FDR}(t)$. The numerator of (A.16) is generally unknown, but according to Liu (2013), the author has proved that

$$\sum_{(i,j)\in H_0} I\{|\hat{T}_{ij}| \geq t\} \approx 2(1-\Phi(t))(p^2-p)/2, \tag{A.17}$$

where $\Phi(.)$ is a standard normal cumulative distribution function. Therefore, we can choose the threshold

$$\hat{t} = \inf\left\{0 \leq t \leq 2\sqrt{\log p} : \frac{2(1-\Phi(t))(p^2-p)/2}{\max\left\{\sum_{1\leq i<j\leq p} I\{|\hat{T}_{ij}| \geq t\}, 1\right\}} \leq \alpha\right\}, \quad 0 \leq \alpha \leq 1, \tag{A.18}$$

where α is a pre-defined level of FDR control. We reject H_0 in (A.14) if $|\hat{T}_{ij}| \geq \hat{t}$.

As an alternative to the tuning-free scaled Lasso regression, each run of (A.11) for X_i against $\boldsymbol{X}_{i^c}$ requires a selection of the tuning parameter $\lambda_i = \delta\sqrt{\hat{\sigma}_{ii}^2 \log(p)/n}$, where $\hat{\sigma}_{ii}^2 = \sum_{k=1}^n X_{ki}^2/n$, with a data-driven choice of δ from 0 to 2. The following data-driven scheme is used based on the result in (A.17):

$$\begin{aligned} \delta &= \hat{l}/N, \\ \hat{l} &= \underset{0\leq l\leq 2N}{\arg\min} \sum_{k=3}^{9} \left(\frac{\sum_{1\leq i<j\leq p} I\{|\hat{T}_{ij}(l/N)| \geq \Phi^{-1}(1-k/20)\}}{k(p^2-p)/20} - 1 \right)^2. \end{aligned} \tag{A.19}$$

Here, $N = 20$ is set by default, and a different value of N can be set up in practice.

Since the previous three methods have asymptotically normal results in terms of (A.4) and (A.8), the FDR framework described in (A.17) and (A.18) can also be applied to them by replacing $\hat{T}_{ij}$ with a different test statistic based on $\hat{\omega}_{ij}$ or $\hat{\omega}_{ij}$. Therefore, the implementations of B_NW_SL, D-S_NW_SL and D-S_GL are allowed for global inference as well.

A.2 The graph settings for time evaluation in simulation studies

A.2.1 Time evaluation on the GGM estimation with FDR control using Lasso

- **Band graph**: a p by p precision matrix $\Omega = (\omega_{ij})_{p\times p}$ with $\omega_{i,i+1} = \omega_{i+1,i} = 0.6$, $\omega_{i,i+2} = \omega_{i+2,i} = 0.3$ and the other off-diagonal elements $\omega_{ij} = 0$ for $|i-j| \geq 3$. The diagonal entries of Ω are $\omega_{ii} = 1$ for $i = 1, 2, 3, ..., p$. The expected node degree of the graph is 4.
- **Hub graph**: an initial p by p matrix $\Omega' = (\omega_{ij})_{p\times p}$ with $\omega_{ij} = \omega_{ji} = 0.5$ for $i = 10(r-1)+1$, $10(r-1)+2 \leq j \leq 10(r-1)+10$ and $1 \leq r \leq p/10$ and the other off-diagonal entries of 0. The diagonal entries of Ω' are $\omega_{ii} = 1$ for $i = 1, 2, 3, ..., p$. To make the matrix positive definite, the final precision matrix is $\Omega = \Omega' + (|\lambda_{\min}| + 0.05)I_p$, where $\lambda_{\min}$ is the minimum eigenvalue of Ω', and I_p is a p by p identity matrix. For $p/10$ variables or nodes in the graph, the expected node degree is 10.

- **E-R graph**: an initial p by p matrix $\Omega^{'} = (\omega_{ij})_{p\times p}$ with each offdiagonal entry $\omega_{ij} = \omega_{ji} = \mu_{ij} * \phi_{ij}$, where μ_{ij} is a uniform random variable between 0.4 and 0.8, and ϕ_{ij} is a Bernoulli random variable (1 means success and 0 means failure) with the success probability of $\min(0.05, 5/p)$. The diagonal entries of $\Omega^{'}$ are $\omega_{ii} = 1$ for $i = 1, 2, 3, ..., p$. To make the matrix positive definite, the final precision matrix is $\Omega = \Omega^{'} + (|\lambda_{\min}| + 0.05)I_p$, where $\lambda_{\min}$ is the minimum eigenvalue of $\Omega^{'}$, and I_p is a p by p identity matrix. The expected node degree is 5 if $p \geq 100$; otherwise, it is $0.05p$.

A.2.2 Time evaluation on the bivariate nodewise scaled Lasso

- **E-R graph**: an initial p by p matrix $\Omega^{'} = (\omega_{ij})_{p\times p}$ with each offdiagonal entry $\omega_{ij} = \omega_{ji}$ = a value randomly picked from the set $\{0.3, 0.6, 1\}$, where the probability of each $\omega_{ij} = \omega_{ji} \neq 0$ is π. The diagonal entries of $\Omega^{'}$ are all set as 4. Then, all the elements including the diagonals in the bottom right block with a size of $p/2 \times p/2$ in $\Omega^{'}$ are multiplied by 2. The final precision matrix is now denoted as Ω. The expected node degree of the graph is πp.

A.3 Testing on the accuracy of individual inference

In this section, we present the details of evaluation on the accuracy of individual inference of each ω_{ij} (or gene pair) for the cases with $n = 800$, $p = 5000$ and $n = 800$, $p = 10000$. We considered three graph settings described as below:

- **Band graph**: a p by p precision matrix $\Omega = (\omega_{ij})_{p\times p}$ with $\omega_{i,i+1} = \omega_{i+1,i} = 0.6$, $\omega_{i,i+2} = \omega_{i+2,i} = 0.3$ and the other off-diagonal elements $\omega_{ij} = 0$ for $|i - j| \geq 3$. The diagonal entries of Ω are $\omega_{ii} = 1$ for $i = 1, 2, 3, ..., p$. The expected node degree of the graph is 4.
- **E-R graph**: an initial p by p matrix $\Omega^{'} = (\omega_{ij})_{p\times p}$ with each offdiagonal entry $\omega_{ij} = \omega_{ji} = \mu_{ij} * \phi_{ij}$, where μ_{ij} is a uniform random variable between 0.4 and 0.8, and ϕ_{ij} is a Bernoulli random variable (1 means success and 0 means failure) with the success probability of

$\min(0.05, 5/p)$. The diagonal entries of $\Omega^{'}$ are $\omega_{ii} = 1$ for $i = 1, 2, 3, ..., p$. To make the matrix positive definite, the final precision matrix is $\Omega = \Omega^{'} + (|\lambda_{\min}| + 0.05)I_p$, where $\lambda_{\min}$ is the minimum eigenvalue of $\Omega^{'}$, and I_p is a p by p identity matrix. The expected node degree of the graph is 5 for $p = 5000$ and 10000.

- **Scale-free graph**: By using the preferential attachment scheme, we started with a single node (or gene) and no edges in the first time step. Then, in each time step, a new gene is added, and the newly-added gene initiates an edge to one of the old genes. An old gene i is selected based on the probability $p(i) \propto d(i)^{0.01} + 1$, where $d(i)$ is the node degree of gene i in the current time step and 0.01 is the power of the preferential attachment. Therefore, the total number of edges in the entire generated graph is given by $p-1$. The above procedure is achieved by the implementation of the function `barabasi.game()` in the `R` package `igraph`. Therefore, we generated a p by p adjacency matrix $A = (a_{ij})_{p\times p}$ with each off-diagonal element $a_{ij} = 1$ if there is a non-zero partial correlation between gene i and j; otherwise, $a_{ij} = 0$. The diagonal elements of A are all equal to 0. Then, we generated an initial p by p matrix $\Omega^{'} = (\omega_{ij})_{p\times p}$ and set any off-diagonal element $\omega_{ij} = 0.3$ if its corresponding $a_{ij} = 1$. To make the matrix positive definite, the final precision matrix is $\Omega = \Omega^{'} + (|\lambda_{\min} + 0.2|)I_p$, where $\lambda_{\min}$ is the minimum eigenvalue of $\Omega^{'}$, and I_p is a p by p identity matrix. The following histograms in Figure A.1 show that the node degree distribution of Scale-free graph for $p = 5000$ and $p = 10000$ follows a power law. The expected node degree of the graph is around 2 for $p = 5000$ and 10000.

Under each of the three graph settings, we simulated 100 data sets. We customized GFC_L to be implemented among 5 candidates of tuning parameters for tuning selection, and the other approaches in `SILGGM` were run with default parameters. We set a pre-specified level of 0.05 on the estimated p-value of each ω_{ij}. In terms of the estimated p-values of all ω_{ij}'s in an entire graph, the mean of the estimated Type I error under the 0.05 level and the corresponding mean of the estimated Type II error over the 100 replications for Band graph, E-R graph and Scale-free graph are reported in Table A.1. The results indicate that all the approaches control the Type I error well in these large scales ($p = 5000$ and 10000) for individual testing on each gene pair. Also, a non-zero partial correlation can be correctly identified in the case of either $p = 5000$ or 10000 since the corresponding Type II error for

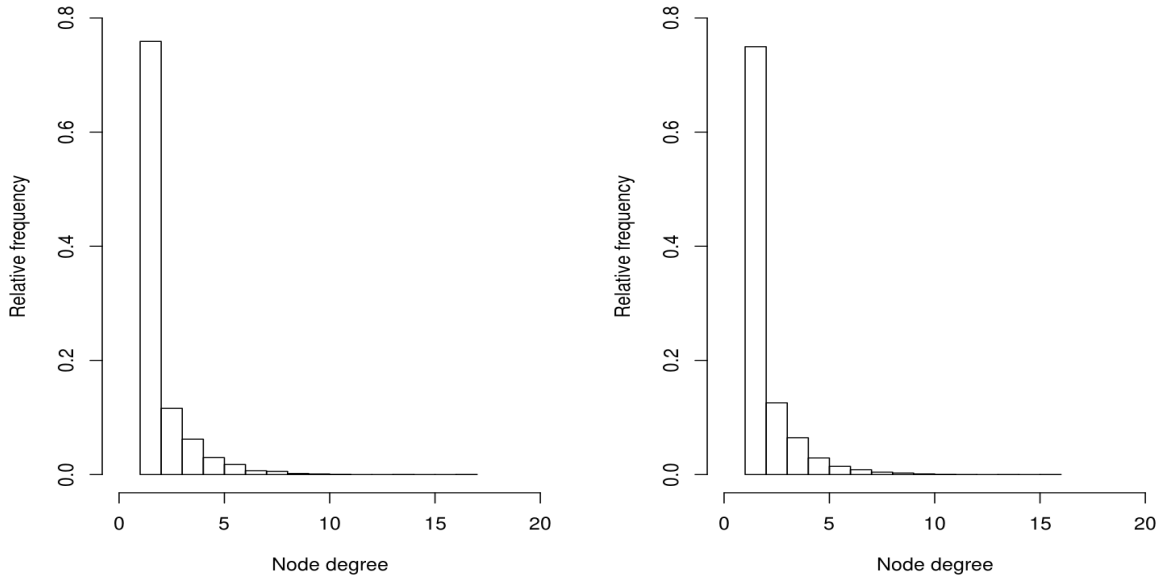

Figure A.1: Histograms of node degrees of Scale-free graph. The left plot illustrates the case of $p = 5000$, and the right plot shows the node degree distribution when $p = 10000$.

all the simulation settings are around 0.

The validation with Type I and Type II errors for individual inference of whether a known zero or a non-zero partial correlation can be correctly identified based on the information of p-values implies no differences among all the approaches. To make a further comparison for individual inference, we then evaluated the average empirical coverage probabilities for the 95% confidence intervals of the ω_{ij}'s for the "non-zero partial correlation" set S_0 (a set of all pairs with non-zero ω_{ij}'s) and the "zero partial correlation" set S_0^c (a set of all pairs with zero ω_{ij}'s) respectively.

Table A.1: Type I and II errors of all the methods under three graph settings.

	Type I error (0.05 level)			Type II error		
	Band	E-R	Scale-free	Band	E-R	Scale-free
			$n = 800, \quad p = 5000$			
B_NW_SL	0.0496	0.0496	0.0495	0	9.4×10^{-4}	5.6×10^{-5}
D-S_NW_SL	0.0228	0.0280	0.0427	0	1.6×10^{-3}	6.0×10^{-5}
D-S_GL	0.0006	0.0315	0.0415	0	8.0×10^{-4}	5.8×10^{-5}
GFC_SL	0.0501	0.0501	0.0501	0	9.3×10^{-4}	5.4×10^{-5}
GFC_L	0.0503	0.0501	0.0501	0	1.3×10^{-3}	4.8×10^{-5}
			$n = 800, \quad p = 10000$			
B_NW_SL	0.0496	0.0496	0.0495	0	6.0×10^{-4}	7.0×10^{-5}
D-S_NW_SL	0.0221	0.0276	0.0432	0	1.1×10^{-3}	8.0×10^{-5}
D-S_GL	0.0002	0.0300	0.0431	0	5.6×10^{-4}	8.0×10^{-5}
GFC_SL	0.0501	0.0501	0.0501	0	6.0×10^{-4}	7.0×10^{-5}
GFC_L	0.0502	0.0501	0.0501	0	8.9×10^{-4}	6.0×10^{-5}

Based on the same 100 replications, we report the mean of 100 estimated average coverage probabilities of the 95% confidence intervals of ω_{ij}'s in S_0 and S_0^c respectively for Band graph, E-R graph and Scale-free graph in Table A.2.

Table A.2: Average empirical coverage probabilities of the 95% confidence intervals in S_0 and S_0^c under three graph settings.

	S_0			S_0^c		
	Band	E-R	Scale-free	Band	E-R	Scale-free
			$n = 800,\ p = 5000$			
B_NW_SL	0.9505	0.8588	0.9330	0.9504	0.9504	0.9505
D-S_NW_SL	0.7864	0.9454	0.9459	0.9772	0.9720	0.9573
D-S_GL	0.5355	0.7967	0.9354	0.9994	0.9685	0.9585
			$n = 800,\ p = 10000$			
B_NW_SL	0.9496	0.8452	0.9361	0.9504	0.9504	0.9505
D-S_NW_SL	0.7368	0.9448	0.9467	0.9779	0.9724	0.9568
D-S_GL	0.5538	0.7801	0.9397	0.9998	0.9700	0.9569

Since GFC_SL or GFC_L provides no confidence intervals, we included the other three approaches here. As it can be seen, the results of empirical coverage probabilities in S_0^c coincide the ones in Type I error rates, and they are all good with our desired level 0.95. For Scale-free graph with $p = 5000$ and 10000, the empirical coverage probabilities of the three methods in S_0 are all around 0.95 as well. However, there are some differences in S_0 for Band graph and E-R graph. For Band graph, B_NW_SL particularly outperforms D-S_NW_SL and D-S_GL since its empirical coverage probabilities in S_0 are well around the desired level, while the empirical coverage probabilities of D-S_NW_SL in S_0 are less than 0.80, and the results of D-S_GL in S_0 are around 0.55. For E-R graph, D-S_NW_SL is the best one with the empirical coverage probabilities in S_0 close to the desired level, but the differences in results among the three methods are much less significant than the ones in Band graph. The empirical coverage probabilities of B_NW_SL in S_0 are still around 0.85, and the results of D-S_GL can be around 0.80 as well.

According to the results from the three graph settings, the overall performance of the confidence intervals among the three methods are good since S_0^c is a major part of the sparse graph settings. But in terms of the confidence intervals in S_0 or the nonzero partial correlations, B_NW_SL and D-S_NW_SL perform better than D-S_GL. Moreover, the performance of B_NW_SL is more stable than that of D-S_NW_SL.

A.4 Testing on the accuracy of global inference

In this section, we show the details of evaluation on the accuracy of global inference which requires a simultaneous testing on all ω_{ij}'s (or gene pairs) with $H_0 : \omega_{ij} = 0$ vs. $H_1 : \omega_{ij} \neq 0$ for $1 \leq i < j \leq p$. We considered the same three graph settings as shown in Appendix A.3.

As for accuracy metrics, we used false discovery rate (FDR), power and the Matthews correlation coefficient (MCC). FDR is the expected proportion of false "discoveries" (the number of incorrect rejections on H_0's) among the total "discoveries" (the total number of rejections on H_0's). We need to control FDR to avoid the inflation of false positives through global inference. The details of the FDR procedure are referred to Appendix A.1. As the second measure, the corresponding power is to show to what extent that the total number of true non-zero partial correlations can be correctly identified through the FDR procedure. Besides FDR and the corresponding power, we also considered MCC as the third measure. Here, MCC is used to gauge how well the known zero partial correlations and the known non-zero partial correlations can be correctly identified through the FDR procedure. It is well known that MCC is even robust to class imbalances, so it tailors to our sparse graph settings which have far more the zero partial correlations than the non-zero partial correlations. Note that MCC lies in the interval between -1 and 1. A value of 1 indicates a perfect selection of all the known zero and the non-zero partial correlations, while a value of -1 implies a total disagreement between prediction and the true partial correlations. A value of 0 means a random guess. Therefore, a closer value of MCC to 1 suggests a better identification of the overall zero and non-zero partial correlations.

For each of the three graph settings, we used the same 100 simulated datasets in Appendix A.3. When implementing all the approaches, we set the argument `alpha` $= 0.05$ to denote a pre-specified level of 0.05 for FDR control. We further set the argument `global = TRUE` when implementing D-S_NW_SL, D-S_GL and B_NW_SL to include global inference. In addition, we customized GFC_L to be implemented among 5 candidates of tuning parameters for tuning selection. With the 100 simulated data sets and the pre-specified level of FDR set at $\alpha = 0.05$, the average empirical FDRs of all the graph settings for $p = 5000$ and $p = 10000$, the corresponding mean power values and the corresponding average MCCs are reported in Table A.3.

Table A.3: Empirical false discovery rates at $\alpha = 0.05$, corresponding power values and MCCs of all the methods under three graph settings.

	Band			E-R			Scale-free		
	FDR	Power	MCC	FDR	Power	MCC	FDR	Power	MCC
				$n = 800, \quad p = 5000$					
B_NW_SL	0.039	1.000	0.981	0.039	0.924	0.942	0.036	0.937	0.950
D-S_NW_SL	0.002	1.000	0.999	0.009	0.894	0.941	0.030	0.933	0.952
D-S_GL	0.010	1.000	0.995	0.015	0.920	0.952	0.023	0.931	0.954
GFC_SL	0.037	1.000	0.982	0.037	0.916	0.939	0.044	0.942	0.949
GFC_L	0.047	1.000	0.976	0.033	0.901	0.934	0.045	0.947	0.951
				$n = 800, \quad p = 10000$					
B_NW_SL	0.037	1.000	0.981	0.037	0.921	0.942	0.034	0.881	0.922
D-S_NW_SL	0.002	1.000	0.999	0.008	0.888	0.938	0.029	0.876	0.922
D-S_GL	0.014	0.999	0.993	0.012	0.911	0.948	0.025	0.874	0.923
GFC_SL	0.036	1.000	0.982	0.037	0.913	0.938	0.045	0.893	0.924
GFC_L	0.047	1.000	0.976	0.033	0.896	0.931	0.045	0.899	0.926

As we can see, the FDRs of all the methods for the three graph settings are effectively controlled below the desired 0.05 level for both $p = 5000$ and $p = 10000$. In terms of Band

graph, almost all the power values are 1 such that the nonzero partial correlations can be correctly identified under the well-controlled FDRs. MCCs of all the methods are close to 1, indicating a near-perfect identification of the overall zero and non-zero partial correlations. For E-R graph, all the approaches have comparable results of power and MCC, and the performance of B_NW_SL is slightly better considering both power and MCC. Similarly, the results of power and MCC of all the methods are also very close for Scale-free graph, and the performance of GFC_L is slightly better according to power and MCC. Even though the overall results are slightly worse for E-R graph and Scale-free graph due to their far more randomized structures than Band graph, all the methods still show high values of power and MCC even in the cases of $p = 10000$. When $p = 10000$, all the power values are around 0.90, and the MCCs are about 0.95 for E-R graph. For Scale-free graph, the power values of all the methods are almost 0.90, and the MCCs are still more than 0.92. Among the three graph settings, even though the FDRs of D-S_NW_SL and D-S_GL are controlled more conservatively below the desired level, their power values and MCCs do not suffer a noticeably negative impact, and some results are even better compared to the other approaches in some particular settings. Therefore, all the approaches have shown good performance in correctly identifying the zero and the nonzero partial correlations in a global sense even for the very high-dimensional scenarios. In addition, we also clarified the mean numbers of false positives (incorrect rejections on the true H_0's) of each approach and the corresponding mean false positive rates (the proportions of false positives among the true H_0's) in the previous benchmarking with the FDR procedure in Table A.4 based on the 100 replications. We can see from the tables that all the false positive rates are close to 0. In other words, the observed false positive numbers are acceptable given the huge number of true negatives (true H_0's).

A.5 The package installation

- Windows users should install `Rtools` before installation of this package.
- The package `SILGGM` is available on `CRAN` and can be installed using the following `R` com-

Table A.4: Numbers of false positives and false positive rates of all the methods under three graph settings.

	Number of false positives			False positive rate		
	Band	E-R	Scale-free	Band	E-R	Scale-free
			$n = 800, \quad p = 5000$			
B_NW_SL	401.8	466.1	175.5	3.2×10^{-5}	3.7×10^{-5}	1.4×10^{-5}
D-S_NW_SL	24.1	105.3	142.2	1.9×10^{-6}	8.4×10^{-6}	1.1×10^{-5}
D-S_GL	98.7	175.0	111.6	7.9×10^{-6}	1.4×10^{-5}	9.0×10^{-6}
GFC_SL	378.7	443.6	215.7	3.0×10^{-5}	3.6×10^{-5}	1.7×10^{-5}
GFC_L	493.3	384.6	225.1	4.0×10^{-5}	3.1×10^{-5}	1.8×10^{-5}
			$n = 800, \quad p = 10000$			
B_NW_SL	768.7	882.1	311.4	1.5×10^{-5}	1.8×10^{-5}	6.3×10^{-6}
D-S_NW_SL	42.3	172.7	264.4	8.5×10^{-7}	3.5×10^{-6}	5.3×10^{-6}
D-S_GL	282.5	278.7	225.1	5.7×10^{-6}	5.6×10^{-6}	4.5×10^{-6}
GFC_SL	751.8	875.0	415.6	1.5×10^{-5}	1.8×10^{-5}	8.3×10^{-6}
GFC_L	976.5	748.9	427.4	2.0×10^{-5}	1.5×10^{-5}	8.6×10^{-6}

mands:

```
install.packages("Rcpp")
install.packages("SILGGM")
```

The first line can be omitted if "install dependencies" is checked in the R package installer.

- When the source code file "SILGGM_1.0.0.tar.gz" is downloaded from CRAN, the package can also be installed:

```
install.packages("Rcpp")
install.packages(pkgs = "SILGGM_1.0.0.tar.gz",
repos = NULL, type = "source")
```

Appendix B Supplement to Chapter 3

B.1 RNA-seq gene expression data of childhood atopic asthma

The RNA-seq data set comes from the Epigenetic Variation and Childhood Asthma in Puerto Ricans (EVA-PR) study, a case-control cohort study for children or adolescents aged from 9 to 20 years recruited from Feb 12, 2014 to May 8, 2017 with a multistage probability sampling (Forno et al., 2019). The purpose of this study is to figure out how genetic or genomic features influence and lead to the development of childhood asthma or atopic asthma particularly among the Hispanic group. Before RNA extraction, a protocol in a subset of nasal samples was implemented to select CD326(+) nasal epithelial cells to alleviate potential effects of different cell types. RNA was extracted from nasal specimens from the inferior turbinate and was sequenced using paired-end reads at 75 cycles and with 80M reads per sample (Forno et al., 2020). In this analysis, we focus on $n = 157$ children with atopic asthma, defined as a doctor's diagnosis of asthma and atopy (≥ 1 positive IgE to aeroallergens), because the case group is more eligible to explain the disease mechanism. The demographic information of the 157 children with atopic asthma is detailed in Table B.1.

B.2 Notation summary for Section 3.3

For convenience, all the important notations related to η_i and μ_i that will be frequently used in this Section 3.3 are summarized in Table B.2 with their corresponding descriptions.

B.3 Properties on the population score variable V

In this section, we provide details on a few facts mentioned in Section 3.3.1.

We first show that $\mathbb{E}(V\epsilon_i) = 0$ with variance $\mathrm{Var}(V\epsilon_i) = \langle V, V\rangle$ if V is a measurable

Table B.1: Characteristics of children with atopic asthma in EVA-PR study. Numbers represent (%) for categorical variables, and mean (standard deviation) or median [interquartile range] for continuous variables.

	Children with atopic asthma
Age (years)	15.3 (2.9)
Sex, n(%)	62 (39.5%) females, 95 (60.5%) males
Race/ethnicity	100% Hispanic/Latino
Total IgE (IU/mL)	372 [208-805]
Number of specific IgE+	2 [1-3]

Table B.2: A summary of notations related to η_i and μ_i.

Notation	Description
$\eta_i = (\psi_i, \theta_i)$	ψ_i: intercept; θ_i: pairwise parameter
$\mu_i = T(X^*_{-i})^\intercal \eta_i$	Canonical parameter of the conditional distribution $\mathbb{P}_{\eta_i}(X_i \vert X_{-i})$
$f(\mu_i)$	Log-partition of $\log(\mathbb{P}_{\eta_i}(X_i \vert X_{-i}))$
$\dot{f}(\mu_i)$	Conditional expectation of $T(X_i)$ given X_{-i}
$\ddot{f}(\mu_i)$	Conditional variance of $T(X_i)$ given X_{-i}

function of X_{-i}. Indeed, since V only depends on X_{-i}, it is clear that,

$$\mathbb{E}(V\epsilon_i) = \mathbb{E}(\mathbb{E}(V\epsilon_i)|X_{-i}) = \mathbb{E}(V\mathbb{E}(\epsilon_i|X_{-i})) = 0.$$

In addition, by using the same strategy, the variance of $V\epsilon_i$ is

$$\begin{aligned}\mathbb{E}(V\epsilon_i)^2 &= \mathbb{E}(\mathbb{E}(V\epsilon_i)^2|X_{-i}) \\ &= \mathbb{E}(V^2 \cdot \mathbb{E}\{(T(X_i) - \dot{f}(T(X^*_{-i})^\intercal\eta_i)|X_{-i}) \\ &= \mathbb{E}(V^2\ddot{f}(\mu_i)) = \langle V, V\rangle,\end{aligned}$$

where we have used equation (3.4) that $\epsilon_i = T(X_i) - \dot{f}(T(X^*_{-i})^\intercal\eta_i)$ and the definition $\langle a, b\rangle = \mathbb{E}(a\ddot{f}(\mu_i)b)$.

We next show that by the choice of score variable V shown in (3.7), we have that $\langle V, m(T(X_{-\{i,j\}}))\rangle = 0$ for any measurable function $m(\cdot)$. Indeed, by the definition of the inner product, it suffices to show $\mathbb{E}(V\ddot{f}(\mu_i)|X_{-\{i,j\}}) = 0$. To see this, we note that

$$\begin{aligned}&\mathbb{E}\left(V\ddot{f}(\mu_i)|X_{-\{i,j\}}\right) \\ =\ &\mathbb{E}\left[\left(T(X_j) - \frac{\mathbb{E}_{\eta_i,\eta_j}(T(X_j)\ddot{f}(\mu_i)|T(X_{-\{i,j\}}))}{\mathbb{E}_{\eta_i,\eta_j}(\ddot{f}(\mu_i)|T(X_{-\{i,j\}}))}\right)\cdot\ddot{f}(\mu_i)|X_{-\{i,j\}}\right] \\ =\ &\mathbb{E}\left[T(X_j)\ddot{f}(\mu_i)|X_{-\{i,j\}}\right] - \mathbb{E}\left[T(X_j)\ddot{f}(\mu_i)|X_{-\{i,j\}}\right] \\ =\ &0.\end{aligned}$$

Finally, we show that if we ignore the minor difference between $\ddot{f}(\mu_i)$ and $\ddot{f}(\hat{\mu}_i)$, then the asymptotic variance of the entire first term in the decomposition (3.6) is F^{-1}_{ij} when using empirical (oracle) score vector $(v^{(o)}_1, ..., v^{(o)}_n)^\intercal$. Recall that $F_{ij} = \mathbb{E}_{\eta_i,\eta_j}((T(X_j) - \frac{\mathbb{E}_{\eta_i,\eta_j}(T(X_j)\ddot{f}(\mu_i)|T(X_{-\{i,j\}}))}{\mathbb{E}_{\eta_i,\eta_j}(\ddot{f}(\mu_i)|T(X_{-\{i,j\}}))})^2\ddot{f}(\mu_i)) = \langle V, V\rangle$. Indeed, we have shown that the population version of the numerator $V\epsilon_i$ has zero mean and variance $\langle V, V\rangle$. Consequently, by CLT, the numerator itself (standardized by multiplying $\sqrt{n}$) in (3.7) weakly converges to $N(0, \langle V, V\rangle)$ when using the empirical score vector. Therefore, our claim can be immediately obtained

by applying the Slutsky's theorem and the weak Law of Large Numbers, together with the fact that $\mathbb{E}(V\ddot{f}(\mu_i)T(X_j)) = \langle V, V\rangle$. To show the last equality, we note that,

$$\begin{aligned}
& \mathbb{E}\left(V\ddot{f}(\mu_i)T(X_j)\right) \\
= \; & \mathbb{E}\left[V\ddot{f}(\mu_i)\cdot\left(V + g\left(X_{-\{i,j\}}, \eta_i, \eta_j\right)\right)\right] \\
= \; & \mathbb{E}\left[V^2\ddot{f}(\mu_i)\right] + \mathbb{E}\left[V\ddot{f}(\mu_i)g\left(X_{-\{i,j\}}, \eta_i, \eta_j\right)\right] \\
= \; & \langle V, V\rangle + \langle V, g\left(X_{-\{i,j\}}, \eta_i, \eta_j\right)\rangle \\
= \; & \langle V, V\rangle,
\end{aligned}$$

where we used that $g\left(X_{-\{i,j\}}, \eta_i, \eta_j\right)$ is a function of $T(X_{-\{i,j\}})$.

B.4 Detailed expression of $\dot{f}(\mu_i)$, $\ddot{f}(\mu_i)$ and Q in three models

We demonstrate the detailed expression of $\dot{f}(\mu_i)$ and $\ddot{f}(\mu_i)$ in Tables B.3 and B.4 corresponding to $f(\mu_i)$ for three modified Poisson-type graphical models described in Section 3.2. The details of Q with respect to $g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$ are illustrated in Table B.5.

Table B.3: Details of $\dot{f}(\mu_i)$ in three models.

Model	$\dot{f}(\mu_i)$
TPGM	$\frac{\sum_{m=0}^{D_i} m\cdot\exp(m\mu_i - \log(m!))}{\sum_{m=0}^{D_i}\exp(m\mu_i - \log(m!))}$
SPGM	$\frac{\sum_{m=0}^{+\infty} S(m)\cdot\exp(S(m)\mu_i - \log(m!))}{\sum_{m=0}^{+\infty}\exp(S(m)\mu_i - \log(m!))}$
SqrtPGM	$\frac{\sum_{m=0}^{+\infty} \sqrt{m}\cdot\exp(\sqrt{m}\mu_i - \log(m!))}{\sum_{m=0}^{+\infty}\exp(\sqrt{m}\mu_i - \log(m!))}$

Table B.4: Details of $\ddot{f}(\mu_i)$ in three models.

Model	$\ddot{f}(\mu_i)$
TPGM	$\frac{\sum_{m=0}^{D_i} m^2 \cdot \exp(m\mu_i - \log(m!))}{\sum_{m=0}^{D_i} \exp(m\mu_i - \log(m!))} - \left(\frac{\sum_{m=0}^{D_i} m \cdot \exp(m\mu_i - \log(m!))}{\sum_{m=0}^{D_i} \exp(m\mu_i - \log(m!))}\right)^2$
SPGM	$\frac{\sum_{m=0}^{+\infty} S(m)^2 \cdot \exp(S(m)\mu_i - \log(m!))}{\sum_{m=0}^{+\infty} \exp(S(m)\mu_i - \log(m!))} - \left(\frac{\sum_{m=0}^{+\infty} S(m) \cdot \exp(S(m)\mu_i - \log(m!))}{\sum_{m=0}^{+\infty} \exp(S(m)\mu_i - \log(m!))}\right)^2$
SqrtPGM	$\frac{\sum_{m=0}^{+\infty} m \cdot \exp(\sqrt{m}\mu_i - \log(m!))}{\sum_{m=0}^{+\infty} \exp(\sqrt{m}\mu_i - \log(m!))} - \left(\frac{\sum_{m=0}^{+\infty} \sqrt{m} \cdot \exp(\sqrt{m}\mu_i - \log(m!))}{\sum_{m=0}^{+\infty} \exp(\sqrt{m}\mu_i - \log(m!))}\right)^2$

Table B.5: Details of Q for $g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$ in three models.

Model	Q
TPGM	$\sum_{k_1=0}^{D_i} \exp(k_1 X_{-\{i,j\}}^{*\intercal} \eta_{i,-j} + k_2 X_{-\{i,j\}}^{*\intercal} \eta_{j,-i} - \log(k_1!) - \log(k_2!) + \hat{\theta}_{ij} k_1 k_2)$
SPGM	$\sum_{k_1=0}^{+\infty} \exp(S(k_1) S(X_{\{-i,j\}}^*)^\intercal \eta_{i,-j} + S(k_2) S(X_{\{-i,j\}}^*)^\intercal \eta_{j,-i} - \log(k_1!) - \log(k_2!) + \theta_{ij} S(k_1) S(k_2))$
SqrtPGM	$\sum_{k_1=0}^{+\infty} \exp(\sqrt{k_1}\sqrt{X_{-\{i,j\}}^*}^\intercal \eta_{i,-j} + \sqrt{k_2}\sqrt{X_{-\{i,j\}}^*}^\intercal \eta_{j,-i} - \log(k_1!) - \log(k_2!) + \theta_{ij}\sqrt{k_1}\sqrt{k_2})$

B.5 Selection of tuning parameters for multiple testing

The multiple testing for all θ_{ij}'s with $1 \leq i < j \leq p$ simultaneously requires selection of tuning parameters $\lambda_i = \delta\sqrt{\hat{\sigma}_{ii}^2 \log(p)/n}$ for each node-wise regression of X_i on other variables X_{-i} with n samples, where $\hat{\sigma}_{ii}^2$ denotes certain estimated variance of X_i given X_{-i}, and δ is a positive constant. Therefore, each λ_i is determined by two quantities: the estimated $\hat{\sigma}_{ii}^2$ and δ, which leads to the following two-step data-driven tuning procedure. The first step is to estimate each σ_{ii}^2 from node-wise regression of X_i on X_{-i}. Based on the extended BIC (EBIC) criterion (Barber and Drton, 2015), we can obtain the estimator $\hat{\eta}_i$. Then, we estimate σ_{ii}^2 using the maximal component of $\ddot{f}(\hat{\mu}_i^{(k)})$ from n samples with $k = 1, 2, ..., n$. In other words,

$$\hat{\sigma}_{ii}^2 = \max_{1 \leq k \leq n} \ddot{f}(\hat{\mu}_i^{(k)}), \qquad \hat{\mu}_i^{(k)} = T(X_{-i}^{*(k)})^\intercal \hat{\eta}_i.$$

The second step is a data-driven procedure for selection of the constant δ following Liu (2013). To guarantee that $2(1-\Phi(t))(p^2-p)/2$ is close to $\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \geq t\}$, we select an appropriate δ between 0 and a relatively large upper bound L (e.g., $L = 2$) via the following optimization

$$\delta = \hat{l}/N,$$
$$\hat{l} = \underset{0 \leq l \leq LN}{\arg\min} \sum_{k=3}^{9} \left(\frac{\sum_{1 \leq i < j \leq p} I\{|\hat{T}_{ij}(l/N)| \geq \Phi^{-1}(1 - k/20)\}}{k(p^2-p)/20} - 1 \right)^2,$$

where $\hat{T}_{ij}(l/N)$ is the corresponding test statistic (recall that $\hat{T}_{ij} = (\sum_{k=1}^n v_k^2 \ddot{f}(\hat{\mu}_i^{(k)}))^{\frac{1}{2}} \tilde{\theta}_{ij}$) when we use $\delta = l/N$ in the chosen tuning parameter λ_i, and N is a pre-specified integer number. In our simulations and real application, we set $N = 10$. For further details, please refer to Liu (2013).

B.6 Comparison of different hyperparameter selection methods

We further compared inferred networks on both simulation settings and the real data application in Sections 3.5 and 3.6 using cross validation as the hyperparameter selection

method.

We at first evaluated the performance of cross validation as the hyperparameter selection method under simulation settings. Similar to EBIC in individual inference, we used 10-fold cross validation to select a tuning parameter on each node-wise regression that minimizes the average negative joint conditional log-likelihood function shown in (3.3) over testing sets. For global inference with multiple testing, we also incorporated cross validation into the tuning selection scheme like EBIC so as to guarantee $2(1-\Phi(t))(p^2-p)/2$ as close to $\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \geq t\}$ as possible.

Table B.6: Medians (standard deviations) of empirical coverage probabilities of the 95% confidence intervals in S_0 and S_0^c from cross validation.

	S_0				S_0^c			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 300,\ p = 100$				
TPGM	0.9495	0.9524	0.9421	0.9495	0.9610	0.9670	0.9652	0.9602
	(0.0232)	(0.0146)	(0.0146)	(0.0244)	(0.0023)	(0.0025)	(0.0024)	(0.0028)
SPGM	0.9495	0.8836	0.9319	0.9495	0.9578	0.9643	0.9643	0.9586
	(0.0245)	(0.0194)	(0.0153)	(0.0216)	(0.0030)	(0.0026)	(0.0028)	(0.0029)
SqrtPGM	0.9495	0.9524	0.9481	0.9596	0.9551	0.9544	0.9539	0.9546
	(0.0198)	(0.0156)	(0.0161)	(0.0216)	(0.0032)	(0.0036)	(0.0033)	(0.0031)
				$n = 300,\ p = 400$				
TPGM	0.9236	0.9063	0.9209	0.9173	0.9555	0.9446	0.9459	0.9543
	(0.0134)	(0.0098)	(0.0091)	(0.0129)	(0.0010)	(0.0016)	(0.0015)	(0.0010)
SPGM	0.9373	0.8697	0.8587	0.8997	0.9624	0.9417	0.9374	0.9547
	(0.0139)	(0.0109)	(0.0125)	(0.0111)	(0.0009)	(0.0013)	(0.0015)	(0.0009)
SqrtPGM	0.9536	0.9538	0.9536	0.9524	0.9558	0.9568	0.9564	0.9553
	(0.0107)	(0.0080)	(0.0080)	(0.0112)	(0.0010)	(0.0009)	(0.0010)	(0.0010)

Based on the same simulated data sets from the graph settings for individual inference in Section 3.5.1, we report the medians (standard deviations) of average empirical coverage

probabilities of 95% confidence intervals of θ_{ij}'s in the edge set $S_0 = \{(i,j) : \theta_{ij} \neq 0\}$ and the non-edge set $S_0^c = \{(i,j) : \theta_{ij} = 0\}$ over 100 replications for $p = 100$ and 400 in Table B.6. As it can be seen, all the results from both low- and high-dimensional settings are close to 0.95, our target confidence level, and they are very similar to the ones with EBIC shown in Table 3.4. For global inference, we evaluated the performance of true positive rate (TPR) and false positive rate (FPR) over a range of false discovery rate (FDR) control levels by comparing their values to the ones with EBIC using the same simulated data sets from the graph settings in Section 3.5.2. The medians of TPRs and FPRs at each cut-off over 100 replications from the two different hyperparameter selection methods are presented in the receiver operating characteristic (ROC) curves for $p = 200$ and 400, as shown in Figures B.1 and B.2. As we can see, all the red solid curves from EBIC overlap the black dashed curves from cross validation in both low- and high-dimensional settings, which indicates that the two different hyperparameter selection methods provide equivalently good results in global inference. We further reported the medians of the empirical FDRs with pre-specified levels 0.1 and 0.2 for both $p = 200$ and 400 in Table B.7. The medians of their corresponding power values are shown in Table B.8. Like the results from EBIC shown in Tables 3.5 and 3.6, the empirical FDRs from cross validation are also well controlled at the desired levels with a relatively good performance of power. Therefore, our proposed two-step inferential procedure is robust to different hyperparameter selection methods, which is a noticeable advantage over the sole estimation approach that highly depends on a specific model selection criterion. To further illustrate this advantage, we also summarized the empirical FDRs and corresponding power values from the sole estimation approach which only involves the first step of our method using EBIC and cross validation based on the same simulated data sets in Section 3.5.2, as shown in Tables B.9 and B.10 for four graph settings. As it can be seen, the estimated networks are totally different by EBIC and cross validation in terms of distinct FDRs and powers, which indicates that the two hyperparameter selection methods generate inconsistent results in the sole estimation. The FDRs can be around level 0.1 or 0.2 from the sole estimation with EBIC, but the corresponding power values are much smaller compared to the ones from our proposed method with EBIC at the desired FDR control level $\alpha = 0.1$ or 0.2 in Table 3.6. For cross validation, even though the sole estimation approach seems

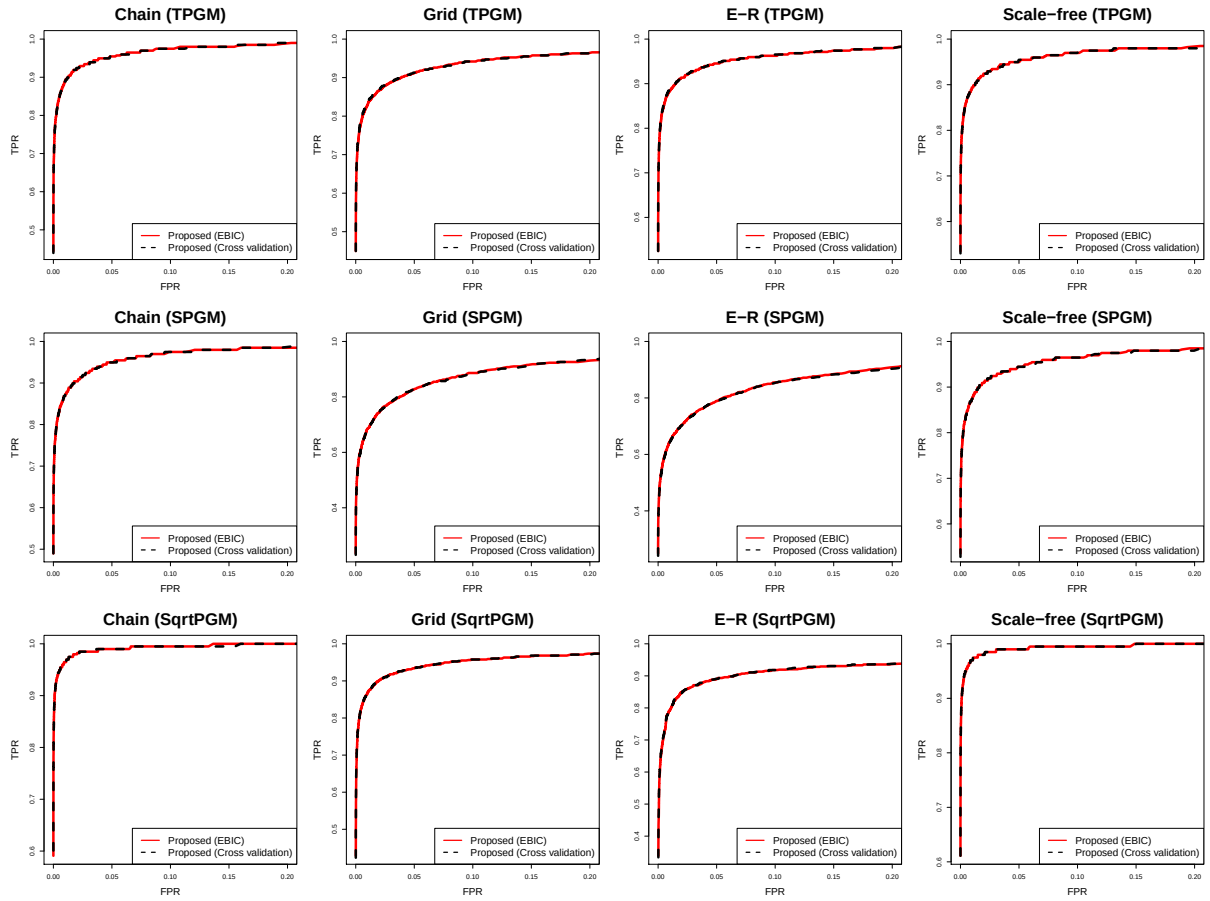

Figure B.1: ROC curves based on TPRs and FPRs for the proposed inferential procedure with EBIC and cross validation in the case of $p = 200$.

to identify more signals with high powers, it generates much more false discovers than those from the proposed method shown in Table B.7 as well. Therefore, our proposed method is capable of reaching a much better trade-off between false and true discovers than the sole estimation approach.

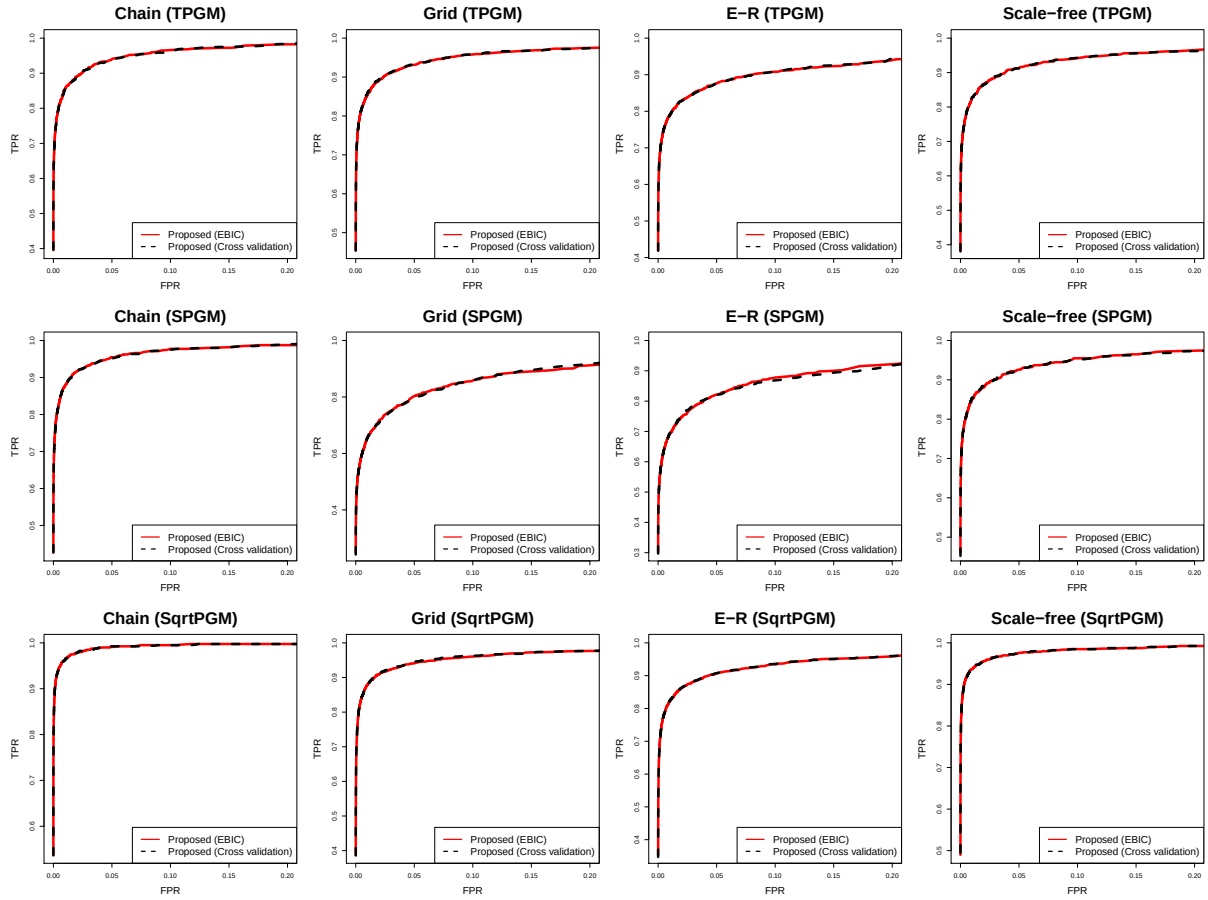

Figure B.2: ROC curves based on TPRs and FPRs for the proposed inferential procedure with EBIC and cross validation in the case of $p = 400$.

Besides the simulation settings, we also evaluated the performance of cross validation in the real data application of $n = 157$ children with atopic asthma in Puerto Rico and $p = 500$ genes. We inferred gene networks using the proposed method with cross validation on the three models with FDR control at level 0.001. As comparison studies, we also constructed gene networks using the sole estimation approach which involves only the first step of our

proposed procedure based on cross validation. Like the evaluation of the constructed networks with EBIC, we also evaluated the inferred networks with cross validation about how their patterns are close to the scale-free (or power-law) topology which generally depicts the structure of a real biological network. We can numerically measure the correlation between the $\log 2$ of node degree and the $\log 2$ of its corresponding probability. A correlation closer to -1 indicates a better conformation to the power law. Figure B.3 shows the $\log 2$-$\log 2$ plots of node degree distribution for inferred networks from cross validation and their corresponding correlation measurements.

Compared to Figure 3.4, the proposed method with either EBIC or cross validation provides biologically meaningful gene networks with correlation values around -0.9, which tends to generate less false discovers. On the contrary, the selected gene networks from the sole estimation by cross validation show very dense structures that are highly likely to include many false discovers and fail to follow the scale-free topology, and they are quite different from the ones selected from the sole estimation approach by EBIC. Again, our proposed method is robust to different hyperparameter selection methods and can consistently demonstrate inferred gene networks with a biologically meaningful structure, while the sole estimation is sensitive to a certain model selection criterion. Furthermore, the proposed method is capable of reaching a better trade-off between false and true discovers than the sole estimation which fails to detect any gene interactions with TPGM using EBIC as shown in Figure 3.4 and is very likely to incur inflated false discovers using cross validation as illustrated in Figure B.3.

In conclusion, we have uncovered two advantages of our proposed method over the sole estimation approach after a careful and comprehensive study of hyperparameter selection. First, the proposed method is robust to different hyperparameter selection methods, while the sole estimation is very sensitive to various model selection criteria. Second, the proposed method can reach a much better balance between false and true discovers than the sole estimation based on different hyperparameter selection methods.

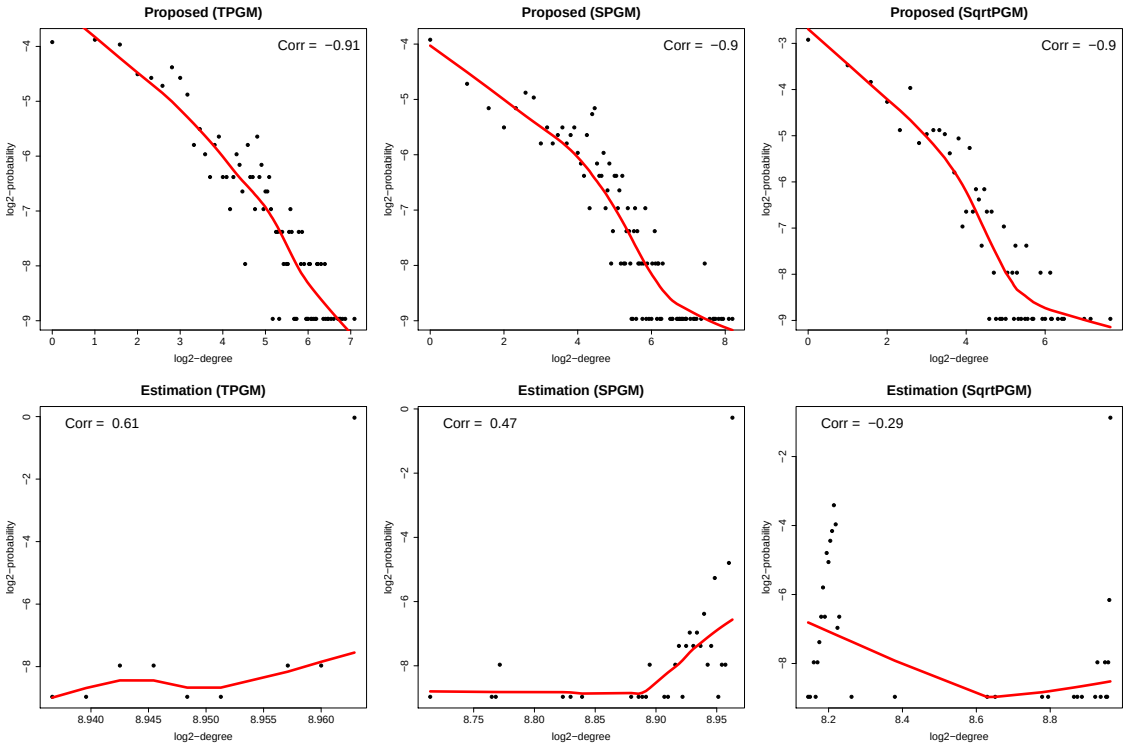

Figure B.3: The log 2-log 2 plots of degree distribution for the inferred networks from the proposed approach and the sole estimation with cross validation.

Table B.7: Medians (standard deviations) of empirical false discovery rates from cross validation.

	$\alpha = 0.1$				$\alpha = 0.2$			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 400,\ p = 200$				
TPGM	0.0918	0.0973	0.0964	0.0938	0.1825	0.1770	0.1844	0.1838
	(0.0282)	(0.0195)	(0.0195)	(0.0245)	(0.0401)	(0.0260)	(0.0277)	(0.0348)
SPGM	0.0858	0.0862	0.1007	0.0973	0.1737	0.1644	0.1815	0.1919
	(0.0261)	(0.0193)	(0.0227)	(0.0266)	(0.0326)	(0.0266)	(0.0323)	(0.0364)
SqrtPGM	0.0884	0.0900	0.0933	0.0949	0.1743	0.1794	0.1734	0.1844
	(0.0226)	(0.0327)	(0.0237)	(0.0263)	(0.0316)	(0.0324)	(0.0257)	(0.0350)
				$n = 400,\ p = 400$				
TPGM	0.0951	0.1011	0.1145	0.0986	0.1865	0.1945	0.2084	0.1964
	(0.0128)	(0.0163)	(0.0195)	(0.0198)	(0.0209)	(0.0228)	(0.0161)	(0.0247)
SPGM	0.1033	0.1154	0.1206	0.1022	0.1874	0.2063	0.2160	0.2108
	(0.0219)	(0.0196)	(0.0155)	(0.0234)	(0.0226)	(0.0233)	(0.0209)	(0.0284)
SqrtPGM	0.1018	0.0950	0.0992	0.0964	0.2022	0.1865	0.1868	0.1974
	(0.0191)	(0.0171)	(0.0138)	(0.0168)	(0.0299)	(0.0263)	(0.0193)	(0.0247)

Table B.8: Medians (standard deviations) of power values for corresponding FDR control levels from cross validation.

	$\alpha = 0.1$				$\alpha = 0.2$			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 400,\ p = 200$				
TPGM	0.7222	0.7222	0.7954	0.7753	0.7828	0.7685	0.8357	0.8131
	(0.0233)	(0.0209)	(0.0193)	(0.0228)	(0.0242)	(0.0190)	(0.0162)	(0.0209)
SPGM	0.7121	0.4868	0.4661	0.7424	0.7677	0.5556	0.5254	0.7879
	(0.0227)	(0.0257)	(0.0246)	(0.0251)	(0.0241)	(0.0255)	(0.0253)	(0.0239)
SqrtPGM	0.8838	0.7513	0.6675	0.8939	0.9192	0.8082	0.7320	0.9242
	(0.0237)	(0.0239)	(0.0333)	(0.0199)	(0.0198)	(0.0216)	(0.0313)	(0.0168)
				$n = 400,\ p = 400$				
TPGM	0.6357	0.7157	0.6543	0.6131	0.6910	0.7606	0.6969	0.6683
	(0.0185)	(0.0096)	(0.0123)	(0.0172)	(0.0193)	(0.0098)	(0.0125)	(0.0175)
SPGM	0.6658	0.4373	0.5062	0.6759	0.7198	0.4927	0.5563	0.7249
	(0.0167)	(0.0140)	(0.0172)	(0.0172)	(0.0194)	(0.0141)	(0.0141)	(0.0203)
SqrtPGM	0.8518	0.7117	0.6485	0.8053	0.8907	0.7724	0.7107	0.8455
	(0.0192)	(0.0197)	(0.0155)	(0.0200)	(0.0163)	(0.0166)	(0.0161)	(0.0193)

Table B.9: Medians (standard deviations) of empirical FDRs and power values from the sole estimation with EBIC and cross validation under Chain and Scale-free graph settings.

	Chain				Scale-free			
	EBIC		Cross validation		EBIC		Cross validation	
	FDR	Power	FDR	Power	FDR	Power	FDR	Power
				$n = 400,\ p = 200$				
TPGM	0.0814	0.5051	0.9426	0.9141	0.1042	0.6010	0.9544	0.9242
	(0.0234)	(0.0265)	(0.0052)	(0.0202)	(0.0238)	(0.0258)	(0.0039)	(0.0190)
SPGM	0.1156	0.5152	0.9617	0.9141	0.1308	0.5606	0.9738	0.9394
	(0.0289)	(0.0221)	(0.0036)	(0.0219)	(0.0227)	(0.0238)	(0.0016)	(0.0165)
SqrtPGM	0.1281	0.6364	0.9655	0.8636	0.1229	0.7121	0.9416	0.8737
	(0.0286)	(0.0306)	(0.0045)	(0.0188)	(0.0213)	(0.0332)	(0.0085)	(0.0225)
				$n = 400,\ p = 400$				
TPGM	0.0767	0.4322	0.9853	0.8643	0.0718	0.4296	0.9813	0.8731
	(0.0163)	(0.0130)	(0.0010)	(0.0186)	(0.0221)	(0.0182)	(0.0014)	(0.0209)
SPGM	0.1062	0.4523	0.9853	0.9121	0.1137	0.4548	0.9912	0.9259
	(0.0180)	(0.0185)	(0.0007)	(0.0172)	(0.0168)	(0.0147)	(0.0003)	(0.0114)
SqrtPGM	0.1034	0.5553	0.9936	0.8291	0.1036	0.5452	0.9905	0.9008
	(0.0142)	(0.0261)	(0.0002)	(0.0145)	(0.0140)	(0.0279)	(0.0004)	(0.0153)

Table B.10: Medians (standard deviations) of empirical FDRs and power values from the sole estimation with EBIC and cross validation under Grid and E-R graph settings.

	Grid				E-R			
	EBIC		Cross validation		EBIC		Cross validation	
	FDR	Power	FDR	Power	FDR	Power	FDR	Power
				$n = 400,\ p = 200$				
TPGM	0.1170	0.5238	0.9735	0.9365	0.1298	0.6297	0.9734	0.9568
	(0.0199)	(0.0194)	(0.0006)	(0.0105)	(0.0178)	(0.0169)	(0.0009)	(0.0113)
SPGM	0.1226	0.3069	0.8960	0.7950	0.1884	0.3245	0.9223	0.7942
	(0.0268)	(0.0191)	(0.0056)	(0.0212)	(0.0264)	(0.0144)	(0.0048)	(0.0198)
SqrtPGM	0.1447	0.5291	0.9492	0.8267	0.1204	0.4739	0.9536	0.7816
	(0.0599)	(0.0276)	(0.0045)	(0.0252)	(0.0207)	(0.0255)	(0.0037)	(0.0293)
				$n = 400,\ p = 400$				
TPGM	0.1017	0.5020	0.9895	0.9565	0.1081	0.4691	0.9882	0.9154
	(0.0157)	(0.0128)	(0.0001)	(0.0083)	(0.0170)	(0.0102)	(0.0002)	(0.0093)
SPGM	0.1347	0.2573	0.9772	0.7652	0.1400	0.3373	0.9768	0.8094
	(0.0211)	(0.0105)	(0.0008)	(0.0204)	(0.0210)	(0.0121)	(0.0008)	(0.0124)
SqrtPGM	0.1217	0.4459	0.9890	0.8384	0.1054	0.4237	0.9861	0.7485
	(0.0200)	(0.0177)	(0.0001)	(0.0162)	(0.0455)	(0.0187)	(0.0004)	(0.0189)

B.7 Additional simulation results

B.7.1 Additional histograms of the pairwise estimates for Chain, Grid, E-R and Scale-free graph settings

Figure B.4 shows the histograms of estimated entries with $p = 100$ for Scale-free graph, and Figures B.5-B.10 show the histograms of randomly selected pairwise estimates that cover all the possible values of true parameters from the three modified Poisson-type graphical models with $p = 100$ and 400 for Chain, Grid and E-R graph settings.

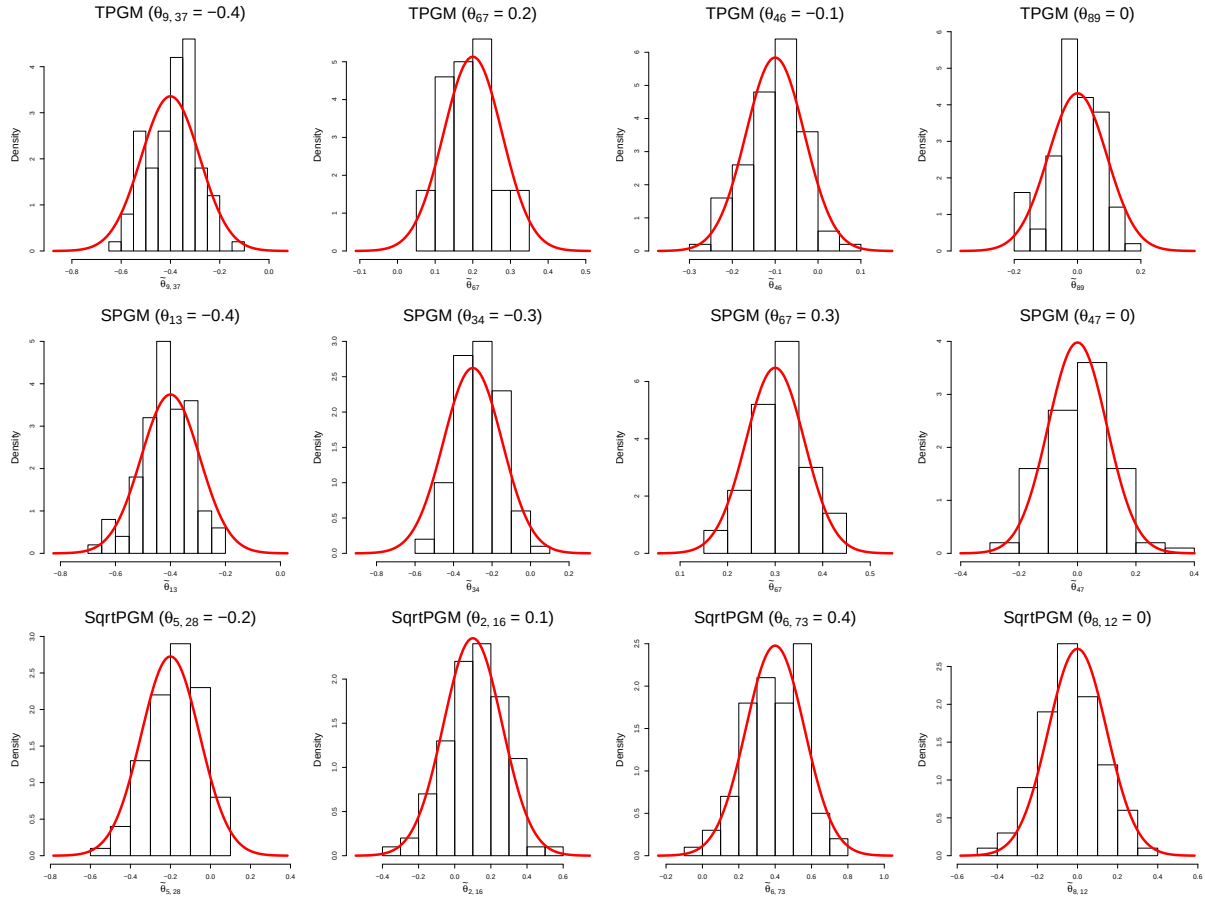

Figure B.4: Histograms of the estimated pairwise entries for $p = 100$ from the three models in Scale-free graph.

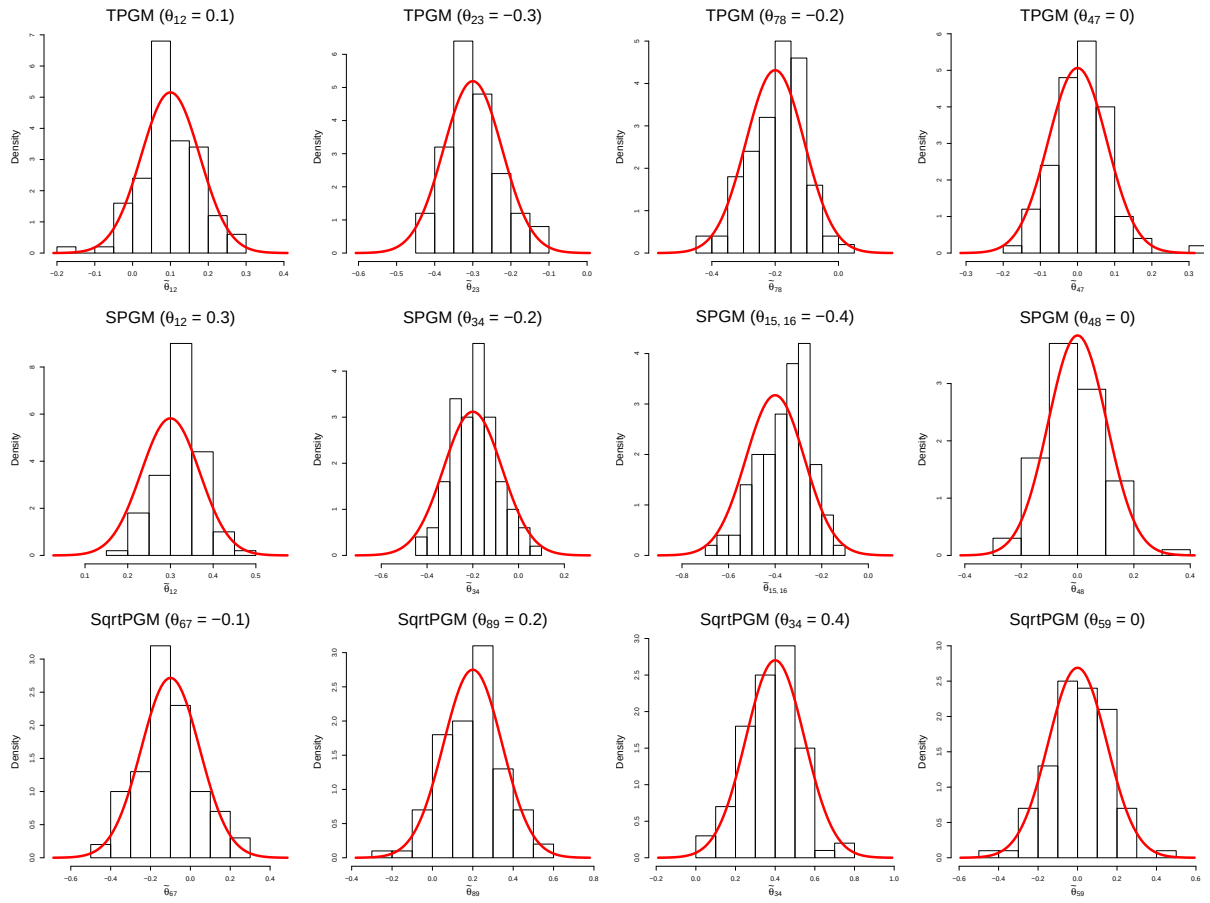

Figure B.5: Histograms of the estimated pairwise entries for $p = 100$ from the three models in Chain graph.

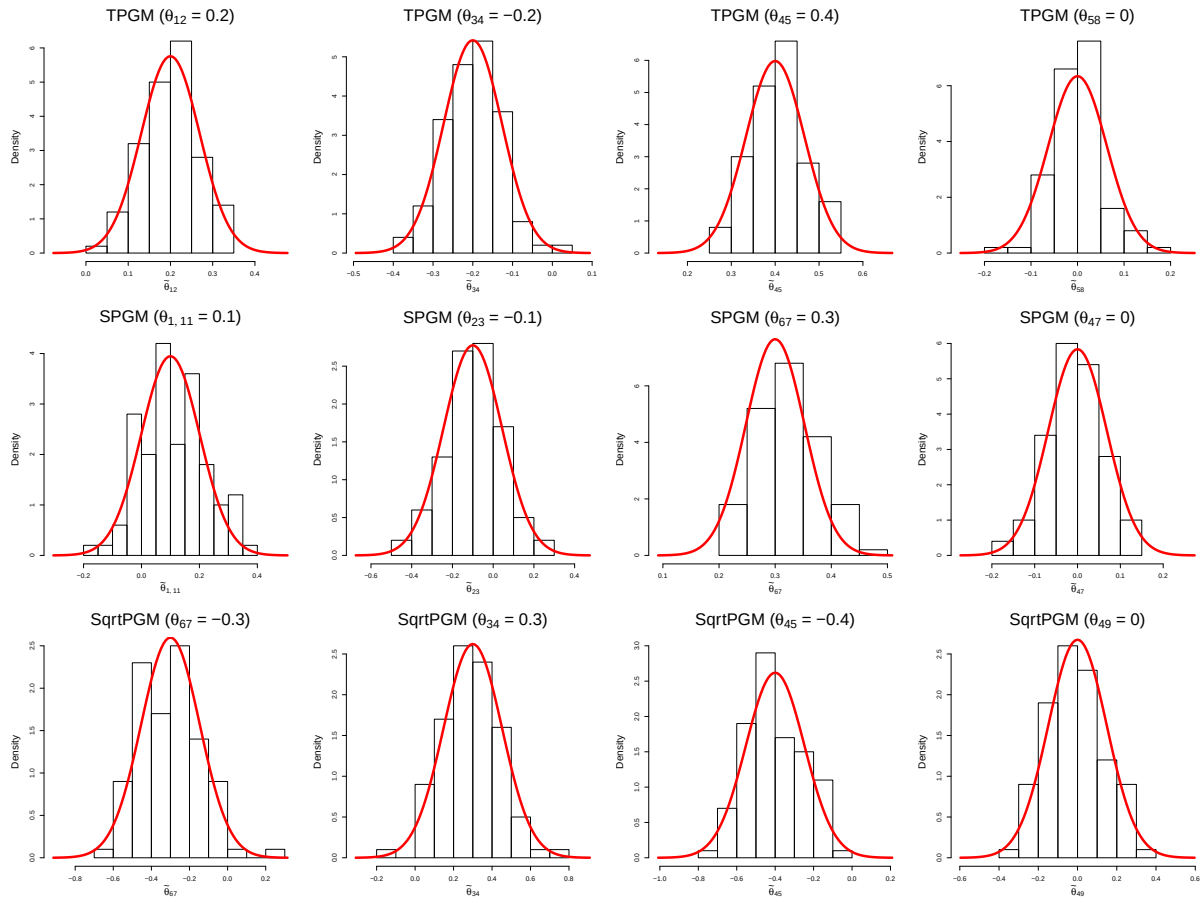

Figure B.6: Histograms of the estimated pairwise entries for $p = 100$ from the three models in Grid graph.

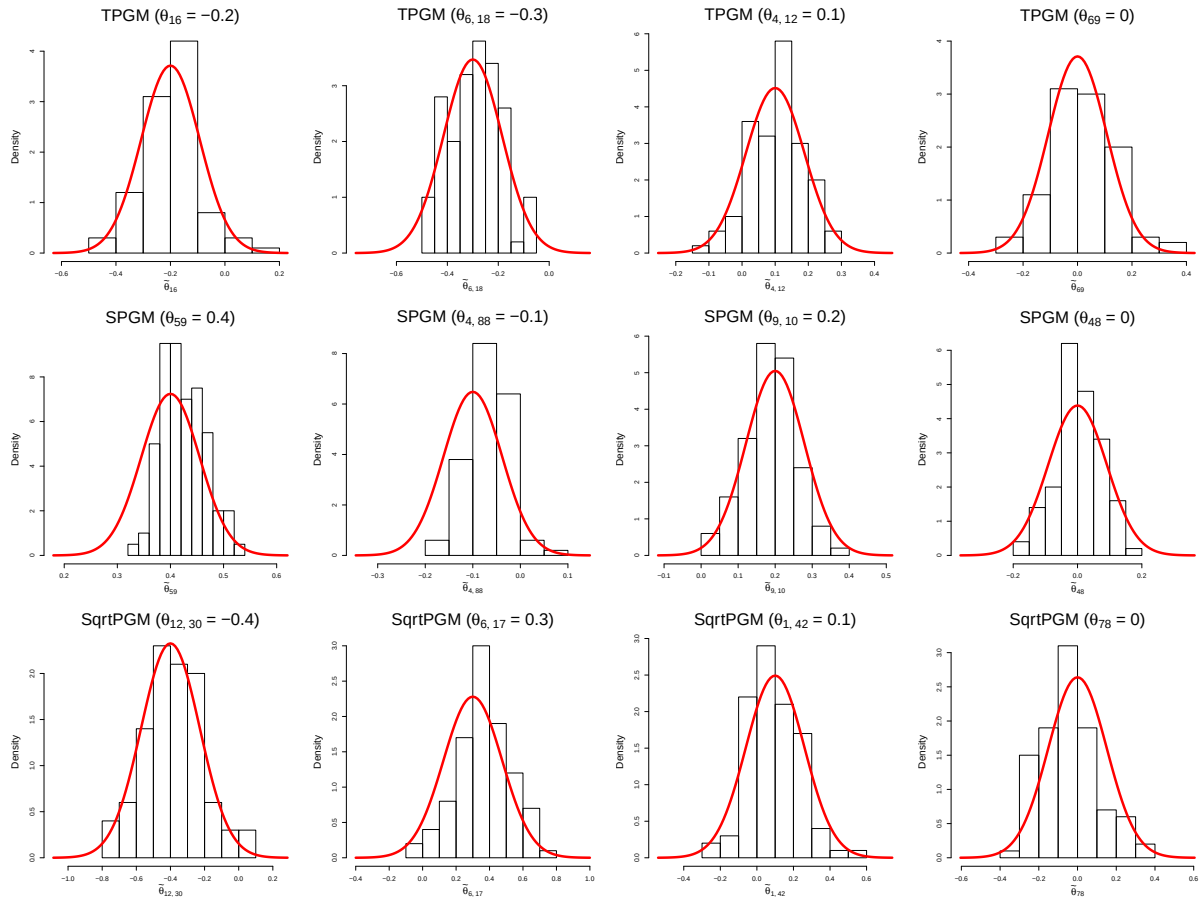

Figure B.7: Histograms of the estimated pairwise entries for $p = 100$ from the three models in E-R random graph.

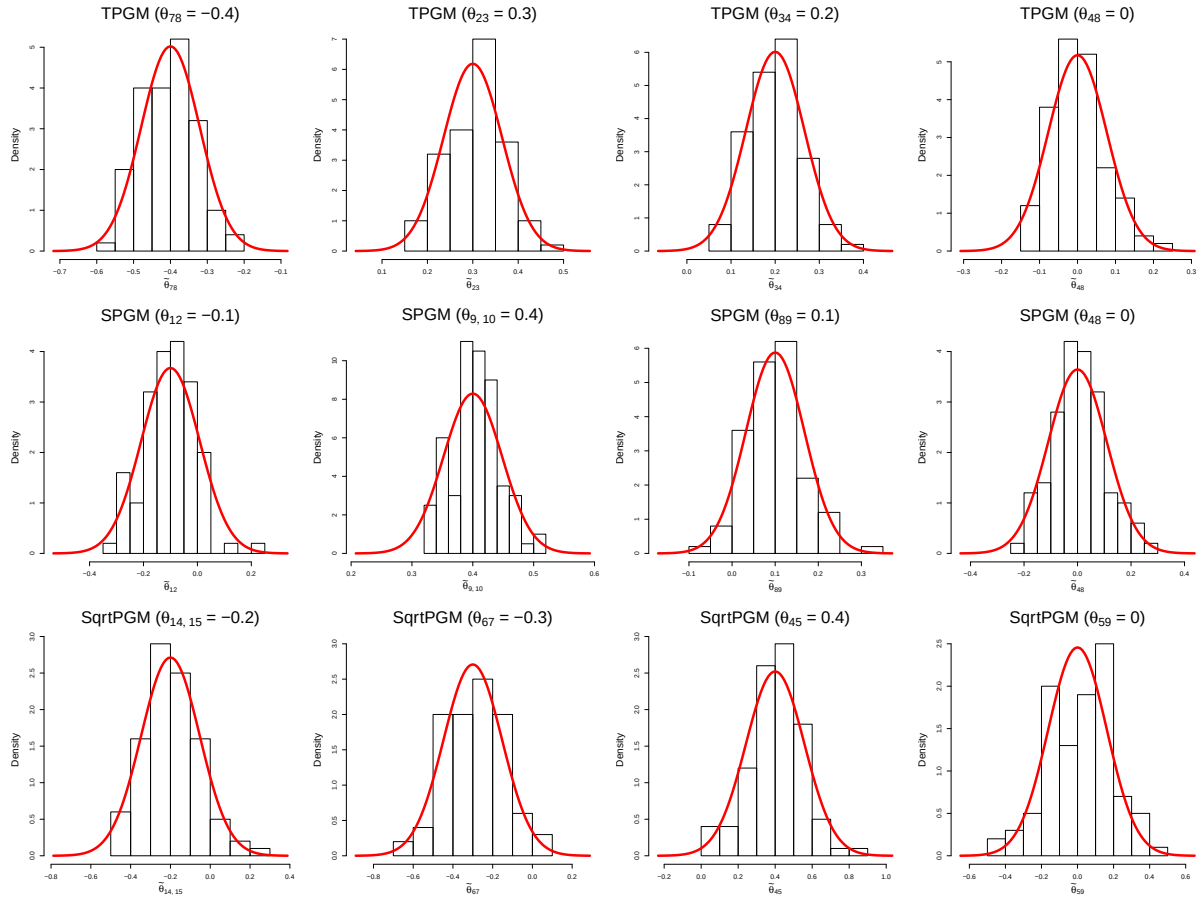

Figure B.8: Histograms of the estimated pairwise entries for $p = 400$ from the three models in Chain graph.

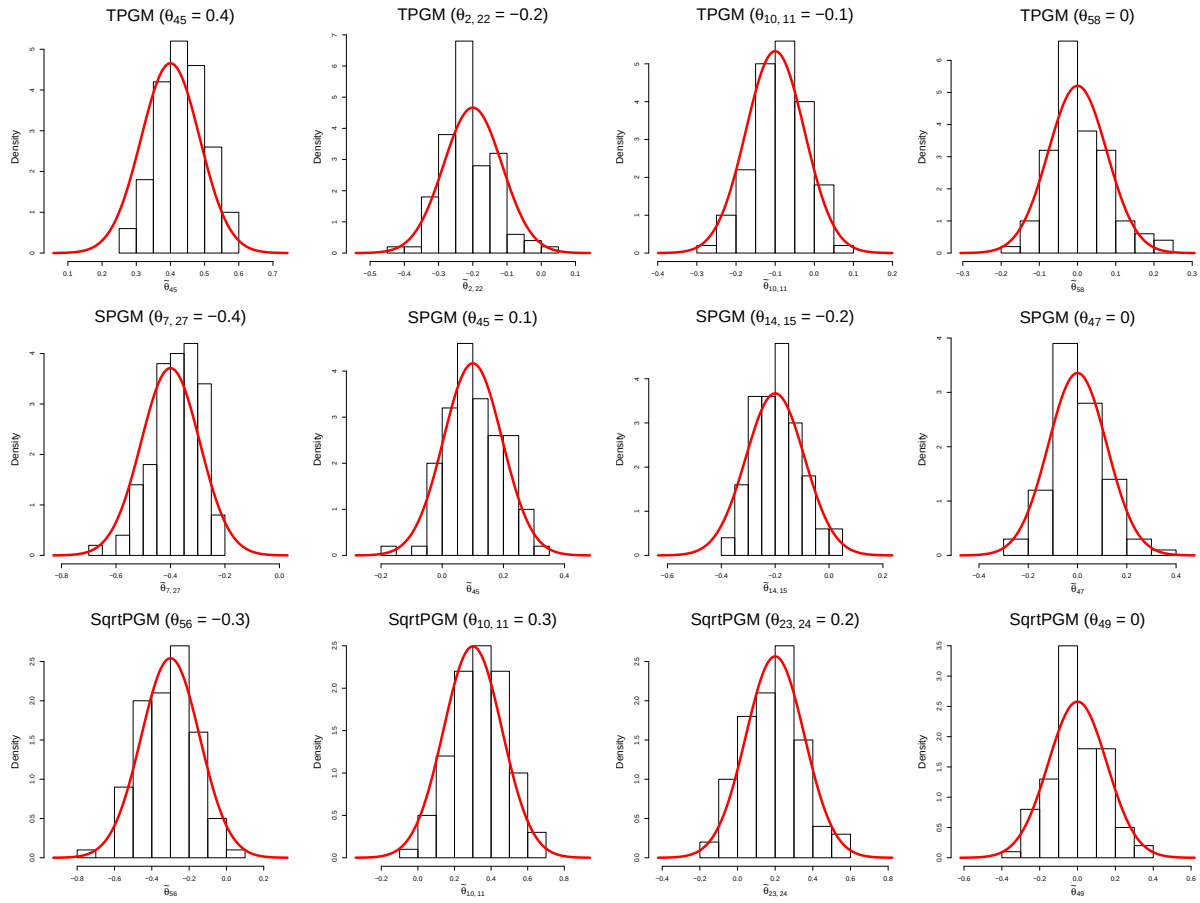

Figure B.9: Histograms of the estimated pairwise entries for $p = 400$ from the three models in Grid graph.

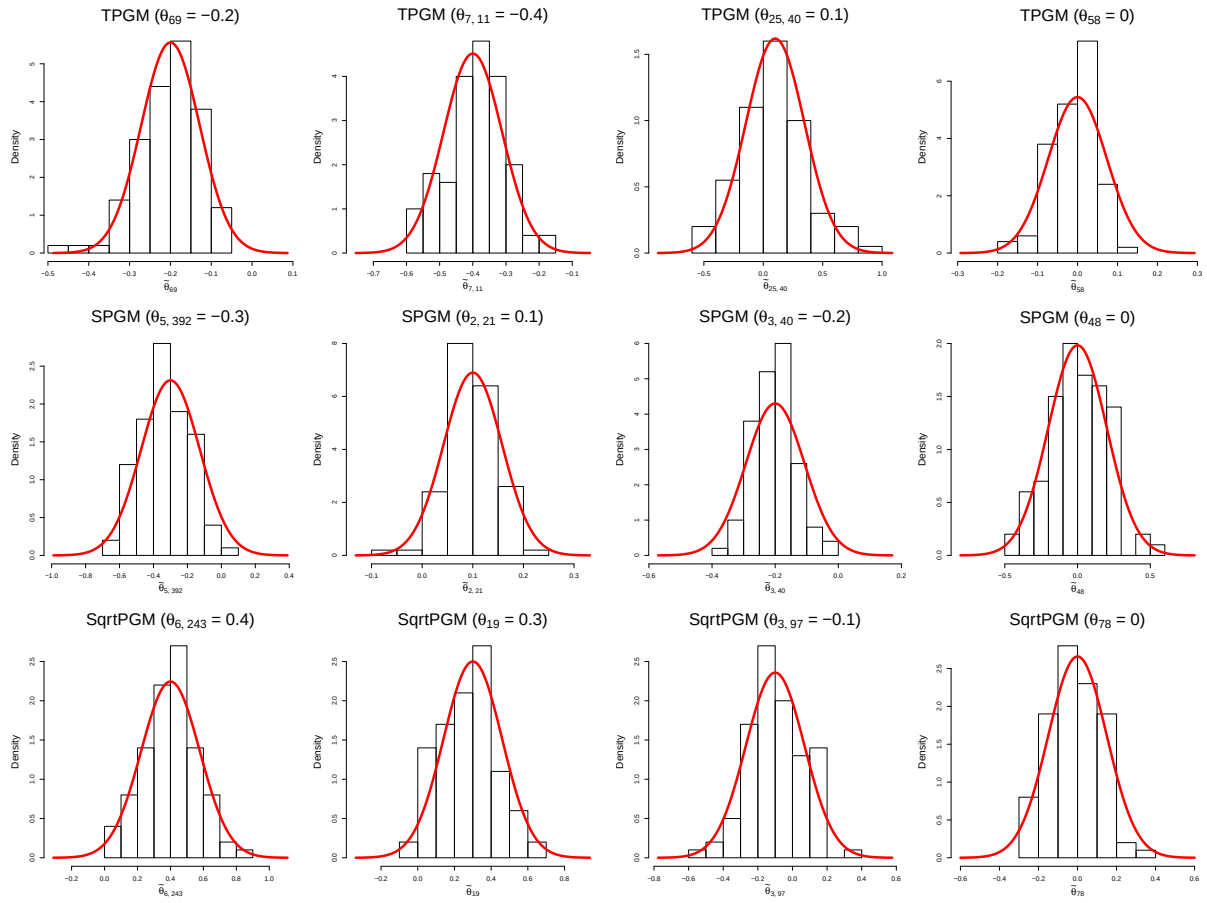

Figure B.10: Histograms of the estimated pairwise entries for $p = 400$ from the three models in E-R random graph.

B.7.2 Additional simulation on individual inference

The medians (standard deviations) of average empirical coverage probabilities of the 95% confidence intervals over 100 replications for $p = 100$ are reported in Table B.11. As we can see, all results are close to 0.95, the target confidence level.

We further considered smaller sample sizes with $n = 150$ and 100 under simulation settings in Section 3.5.1 to evaluate the performance of asymptotic normality for individual inference. In addition, we also included the simulation settings in Section 3.5.2 to show the performance of asymptotic normality for individual inference.

We report the medians (standard deviations) of average empirical coverage probabilities of 95% confidence intervals of θ_{ij}'s in the edge set $S_0 = \{(i, j) : \theta_{ij} \neq 0\}$ and the non-edge set $S_0^c = \{(i, j) : \theta_{ij} = 0\}$ over 100 replications for $p = 100$ and 400 with $n = 150$ and 100 under simulation settings in Section 3.5.1 in Tables B.12 and B.13 respectively. The medians (standard deviations) of empirical coverage probabilities of 95% confidence intervals using the simulation settings in Section 3.5.2 are reported in Table B.14. As we can see, all results are close to 0.95, the target confidence level.

Table B.11: Medians (standard deviations) of empirical coverage probabilities of the 95% confidence intervals in S_0 and S_0^c with $p = 100$.

	S_0				S_0^c			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 300$,	$p = 100$			
TPGM	0.9495	0.9365	0.9263	0.9394	0.9524	0.9525	0.9491	0.9514
	(0.0218)	(0.0170)	(0.0193)	(0.0251)	(0.0030)	(0.0033)	(0.0033)	(0.0036)
SPGM	0.9495	0.9101	0.9108	0.9394	0.9514	0.9548	0.9509	0.9506
	(0.0221)	(0.0193)	(0.0160)	(0.0203)	(0.0032)	(0.0038)	(0.0041)	(0.0034)
SqrtPGM	0.9495	0.9524	0.9481	0.9495	0.9512	0.9504	0.9491	0.9516
	(0.0199)	(0.0156)	(0.0135)	(0.0209)	(0.0033)	(0.0036)	(0.0036)	(0.0032)

B.7.3 Additional simulation on global inference

ROC curves for $p = 200$ are shown in Figure B.11. Likewise, all curves from the proposed inferential procedure lie above the ones from the sole estimation.

We further considered a smaller sample size with $n = 150$ to evaluate the performance of multiple testing with FDR control. We report the medians (standard deviations) of empirical FDRs with pre-specified levels 0.1 and 0.2 for both $p = 200$ and 400 in Table B.15. The medians (standard deviations) of their corresponding power values are reported in Table B.16. Due to the sparsity assumption with the maximum node degree $s = o(\sqrt{n}/\log(p))$ for high-dimensional statistical inference, we only included Chain and Scale-free graphs in the results. Because the maximum node degree is $s = 4$ for both Grid and E-R graph settings, the required sample sizes n need to be far larger than 150 to meet the sparsity assumption when $p = 100$ (e.g. $n = 340$) and $p = 400$ (e.g. $n = 575$). From the results in Tables B.15 and B.16, it is worthwhile to notice that the FDRs can be still controlled quite well around the desired levels even in the case of Scale-free graph with $n = 150$ and $p = 400$ among all the three modified Poisson-type graphical models. Since the biological

Table B.12: Medians (standard deviations) of empirical coverage probabilities of the 95% confidence intervals in S_0 and S_0^c with $n = 150$.

	S_0				S_0^c			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 150,\ p = 100$				
TPGM	0.9495	0.9365	0.9263	0.9394	0.9491	0.9460	0.9451	0.9486
	(0.0238)	(0.0165)	(0.0174)	(0.0248)	(0.0039)	(0.0043)	(0.0040)	(0.0040)
SPGM	0.9596	0.9100	0.9155	0.9495	0.9503	0.9537	0.9485	0.9499
	(0.0206)	(0.0196)	(0.0202)	(0.0223)	(0.0036)	(0.0042)	(0.0043)	(0.0034)
SqrtPGM	0.9596	0.9471	0.9528	0.9545	0.9511	0.9500	0.9496	0.9509
	(0.0189)	(0.0160)	(0.0154)	(0.0233)	(0.0031)	(0.0038)	(0.0032)	(0.0028)
				$n = 150,\ p = 400$				
TPGM	0.9449	0.9371	0.9322	0.9311	0.9478	0.9441	0.9441	0.9459
	(0.0096)	(0.0076)	(0.0098)	(0.0149)	(0.0012)	(0.0013)	(0.0015)	(0.0012)
SPGM	0.9474	0.9255	0.9051	0.9323	0.9507	0.9492	0.9489	0.9492
	(0.0121)	(0.0096)	(0.0092)	(0.0111)	(0.0010)	(0.0013)	(0.0016)	(0.0011)
SqrtPGM	0.9524	0.9525	0.9512	0.9549	0.9507	0.9497	0.9497	0.9503
	(0.0126)	(0.0073)	(0.0071)	(0.0091)	(0.0009)	(0.0011)	(0.0009)	(0.0011)

Table B.13: Medians (standard deviations) of empirical coverage probabilities of the 95% confidence intervals in S_0 and S_0^c with $n = 100$.

	S_0				S_0^c			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 100,$	$p = 100$			
TPGM	0.9495	0.9365	0.9316	0.9394	0.9470	0.9444	0.9430	0.9460
	(0.0225)	(0.0157)	(0.0209)	(0.0233)	(0.0036)	(0.0042)	(0.0038)	(0.0039)
SPGM	0.9596	0.9127	0.9155	0.9495	0.9489	0.9507	0.9460	0.9485
	(0.0197)	(0.0208)	(0.0186)	(0.0226)	(0.0039)	(0.0043)	(0.0043)	(0.0041)
SqrtPGM	0.9495	0.9524	0.9528	0.9596	0.9510	0.9506	0.9503	0.9505
	(0.0223)	(0.0161)	(0.0167)	(0.0217)	(0.0033)	(0.0032)	(0.0035)	(0.0032)
				$n = 100,$	$p = 400$			
TPGM	0.9511	0.9422	0.9340	0.9411	0.9454	0.9414	0.9418	0.9446
	(0.0115)	(0.0095)	(0.0097)	(0.0110)	(0.0010)	(0.0011)	(0.0020)	(0.0016)
SPGM	0.9499	0.9262	0.9105	0.9298	0.9495	0.9458	0.9459	0.9474
	(0.0119)	(0.0109)	(0.0092)	(0.0112)	(0.0013)	(0.0017)	(0.0013)	(0.0012)
SqrtPGM	0.9524	0.9519	0.9487	0.9511	0.9505	0.9497	0.9500	0.9508
	(0.0144)	(0.0067)	(0.0082)	(0.0127)	(0.0007)	(0.0012)	(0.0008)	(0.0010)

Table B.14: Medians (standard deviations) of empirical coverage probabilities of the 95% confidence intervals in S_0 and S_0^c with simulation settings in Section 3.5.2.

	S_0				S_0^c			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 400,\ p = 200$				
TPGM	0.9495	0.9234	0.9150	0.9343	0.9536	0.9541	0.9559	0.9546
	(0.0144)	(0.0133)	(0.0131)	(0.0168)	(0.0015)	(0.0019)	(0.0018)	(0.0016)
SPGM	0.9495	0.9444	0.9395	0.9192	0.9555	0.9523	0.9543	0.9549
	(0.0151)	(0.0127)	(0.0120)	(0.0159)	(0.0017)	(0.0017)	(0.0016)	(0.0016)
SqrtPGM	0.9192	0.9074	0.8784	0.9091	0.9558	0.9525	0.9502	0.9552
	(0.0190)	(0.0148)	(0.0169)	(0.0204)	(0.0016)	(0.0019)	(0.0030)	(0.0016)
				$n = 400,\ p = 400$				
TPGM	0.9510	0.9248	0.9210	0.9347	0.9532	0.9548	0.9546	0.9519
	(0.0092)	(0.0100)	(0.0098)	(0.0134)	(0.0009)	(0.0008)	(0.0015)	(0.0008)
SPGM	0.9497	0.9314	0.9251	0.9221	0.9538	0.9522	0.9536	0.9531
	(0.0127)	(0.0099)	(0.0105)	(0.0131)	(0.0009)	(0.0011)	(0.0009)	(0.0009)
SqrtPGM	0.9196	0.9090	0.8858	0.8970	0.9541	0.9521	0.9505	0.9524
	(0.0134)	(0.0121)	(0.0106)	(0.0169)	(0.0009)	(0.0008)	(0.0014)	(0.0010)

network has a scale-free pattern in our application which has about 150 observations, the simulation results here to some extent provide additional evidence that the application from our proposed method is reliable.

Table B.15: Medians (standard deviations) of empirical false discovery rates with $n = 150$.

	$\alpha = 0.1$		$\alpha = 0.2$	
	Chain	Scale-free	Chain	Scale-free
		$n = 150, \quad p = 200$		
TPGM	0.0800 (0.0449)	0.0983 (0.0373)	0.1711 (0.0501)	0.1903 (0.0438)
SPGM	0.0876 (0.0346)	0.0968 (0.0385)	0.1686 (0.0386)	0.1789 (0.0452)
SqrtPGM	0.1065 (0.0397)	0.1136 (0.0412)	0.1808 (0.0456)	0.1813 (0.0462)
		$n = 150, \quad p = 400$		
TPGM	0.0933 (0.0338)	0.1224 (0.0308)	0.1862 (0.0418)	0.2096 (0.0412)
SPGM	0.1016 (0.0313)	0.1280 (0.0294)	0.1978 (0.0335)	0.2210 (0.0333)
SqrtPGM	0.1340 (0.0286)	0.1372 (0.0401)	0.1871 (0.0315)	0.2269 (0.0388)

B.7.4 Additional simulation on simulated RNA-seq data

All the medians (standard deviations) of empirical FDRs and corresponding power values with $n = 150$ and pre-specified levels $\alpha = 0.1$ and 0.2 are reported in Table B.17. Again, the proposed method can capture the built-in features better than GFC⊥L.

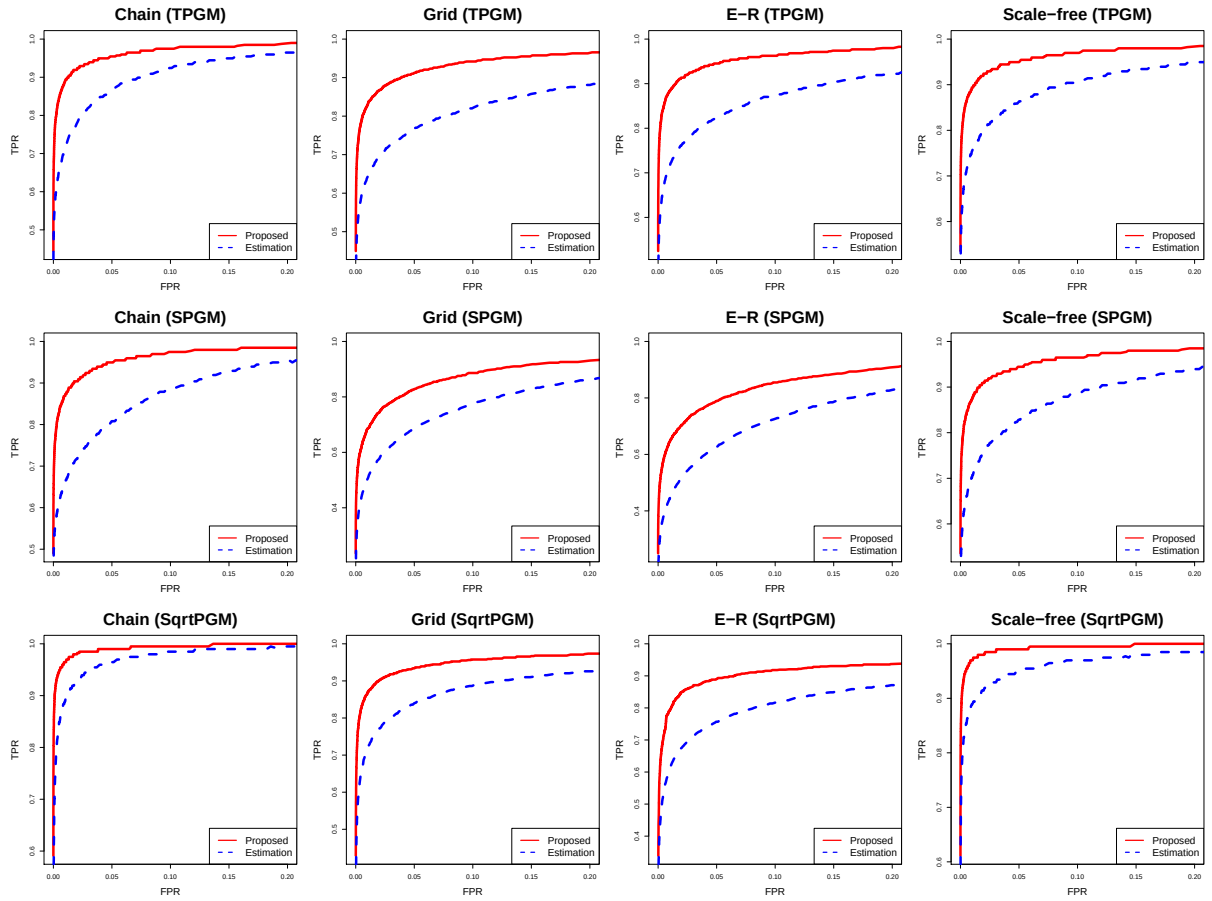

Figure B.11: ROC curves based on TPRs and FPRs for the proposed inferential procedure and the sole estimation in the case of $p = 200$.

Table B.16: Medians (standard deviations) of power values for corresponding FDR control levels with $n = 150$.

	$\alpha = 0.1$		$\alpha = 0.2$	
	Chain	Scale-free	Chain	Scale-free
		$n = 150, \quad p = 200$		
TPGM	0.2323 (0.0342)	0.3308 (0.0332)	0.2879 (0.0372)	0.3939 (0.0358)
SPGM	0.3232 (0.0265)	0.3535 (0.0343)	0.3737 (0.0311)	0.4066 (0.0335)
SqrtPGM	0.3030 (0.0346)	0.3232 (0.0385)	0.3965 (0.0384)	0.3939 (0.0406)
		$n = 150, \quad p = 400$		
TPGM	0.2098 (0.0191)	0.2048 (0.0209)	0.2563 (0.0176)	0.2475 (0.0237)
SPGM	0.2299 (0.0238)	0.2789 (0.0187)	0.2827 (0.0216)	0.3254 (0.0204)
SqrtPGM	0.2437 (0.0242)	0.2111 (0.0254)	0.3166 (0.0205)	0.2802 (0.0300)

Table B.17: Medians (standard deviations) of empirical FDRs and power values from our proposed method (SqrtPGM and SPGM) and GFC_L on simulated RNA-seq data with $n = 150$ and FDR controlled at levels $\alpha = 0.1$ and 0.2.

Proposed (SqrtPGM)		Proposed (SPGM)		GFC_L	
FDR	Power	FDR	Power	FDR	Power
		$\alpha = 0.1$			
0.1436	0.2324	0.1197	0.2111	0.1124	0.1118
(0.0438)	(0.0265)	(0.0345)	(0.0283)	(0.0466)	(0.0260)
		$\alpha = 0.2$			
0.2219	0.2927	0.1993	0.2638	0.2265	0.1709
(0.0508)	(0.0320)	(0.0345)	(0.0296)	(0.0364)	(0.0303)

B.8 Additional results in real data application

B.8.1 Comparison between original and normalized counts

The histograms of original and normalized counts of RNA-seq data are shown in Figure B.12.

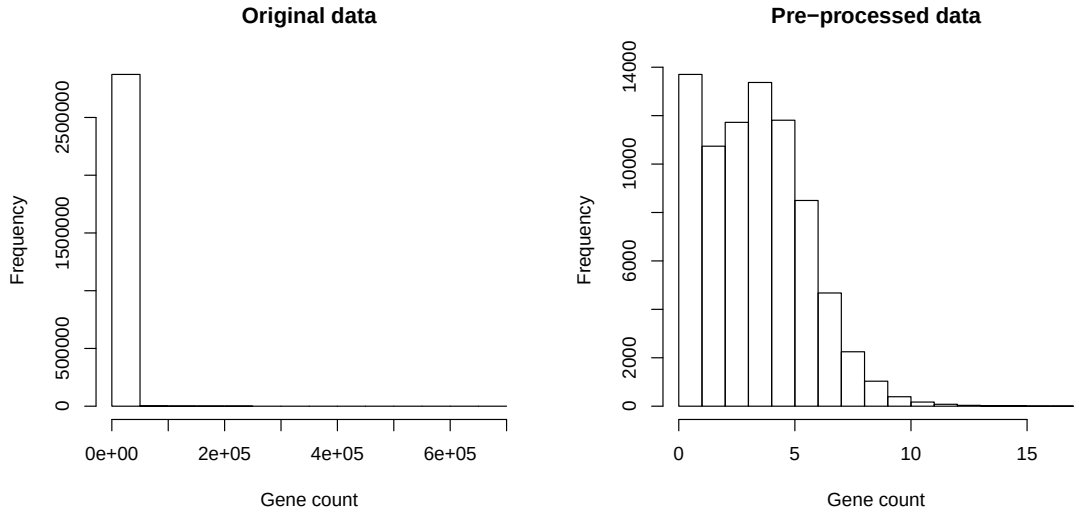

Figure B.12: Histograms of the genes for the original and the pre-process RNA-seq data.

B.8.2 Additional evaluations for the overall network structure

We further compared our proposed approach with two other approaches under normal distribution assumption: c-level partial correlation graph estimation (c-level PC) (Qiu and Zhou, 2018) and sparse partial correlation estimation (SPACE) (Peng et al., 2009) on evaluation of the overall inferred network structure.

Before performing c-level PC and SPACE, we extracted the same $p = 500$ genes from the original RNA-seq gene expression data among the $n = 157$ children with atopic asthma in Puerto Ricans and made a log plus nonparanormal transformation (Liu et al., 2009) to continuize and gaussianize the count values in terms of Jia et al. (2017). Then, we performed c-level PC and SPACE on the normalized data after transformation. For c-level PC, we imple-

mented it using the code from the author's Github repository: `https://github.com/yumouqiu/Estimating-c-level-partial-correlation`. To be consistent with our application, we chose level $c = 0$ and also determined the FDR control at a pre-specified level of 0.001. For SPACE, we implemented it using the function `space.joint()` in the `R` package `space` with the weight that is proportional to the estimated degree of each node, which would result in a preferential attachment effect and make the estimated network closer to a real biological network with a power-law (or scale-free) pattern (Barabási and Albert, 1999).

Figure B.13 illustrates the $\log 2$-$\log 2$ plots of node degree distribution based on the inferred networks from c-level PC and SPACE with their corresponding correlation measurements. As we can see, the correlations are -0.71 and -0.53 for c-level PC and SPACE respectively. The values demonstrate weaker negative linear relationship than those from the proposed approach, which provides evidence that the inferred networks from the two approaches for graphical models under normal distribution assumption do not conform the power law well.

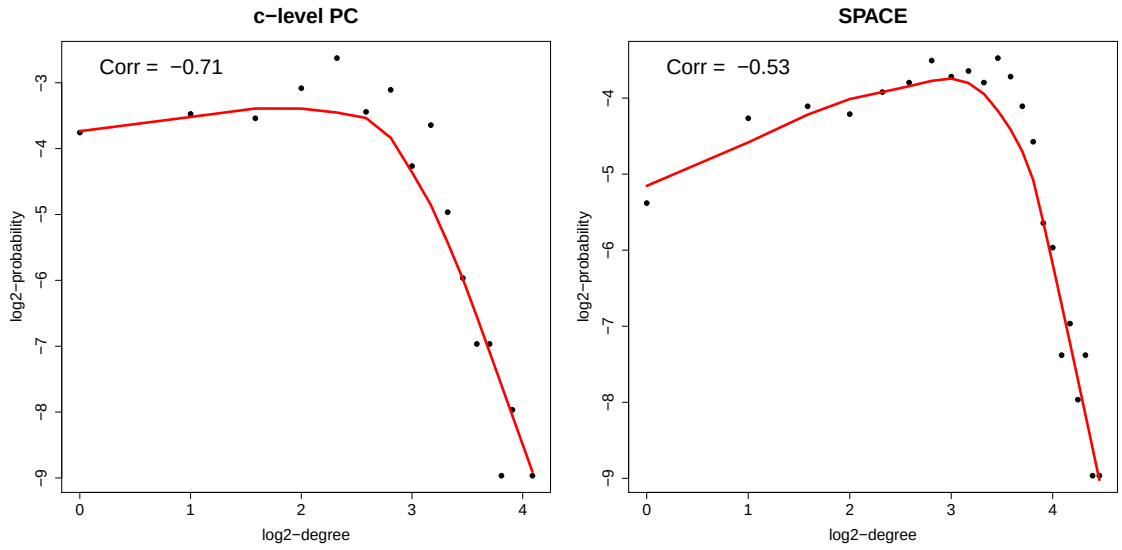

Figure B.13: The $\log 2$-$\log 2$ plots of degree distribution for the inferred networks from c-level PC and SPACE.

Figure B.14: Some enriched pathways from the proposed inferential procedure in TPGM and SPGM.

B.8.3 Some enriched pathways using the proposed method in TPGM and SPGM

Figure B.14 illustrates some enriched pathways from the modules in TPGM and SPGM.

B.8.4 Gene interactions of the module in SqrtPGM with enriched JAK-STAT signaling pathway and their corresponding interactions in TPGM and SPGM

Figure B.15 presents the gene interactions of the module in SqrtPGM where the JAK-STAT signaling pathway is enriched and their corresponding interactions in TPGM and SPGM in a fixed panel. As expected, SqrtPGM shows 77 gene connections which are more than those identified in TPGM (27 connections) and SPGM (6 connections). The gene connections in SqrtPGM include all those identified in SPGM and 14 out of 27 identified in TPGM. Moreover, it is interesting to notice two hub genes DYNLT1 and TRIM22 shown in yellow ovals in SqrtPGM, and they are both related to atopic asthma. DYNLT1 is shown to be one of the top genes to be associated with atopic dermatitis, which is closely related to atopic asthma. TRIM22, which involves in antiviral response regulated by an interferon pathway, has the strongest association with leukotriene receptor antagonist treatment to childhood asthma, according to Perez-Garcia et al. (2020).

B.8.5 Additional analysis on TPM values from the RNA-seq data of childhood atopic asthma

We evaluated the performance of GFC_L on the TPM values of the same 500 genes among 157 children with atopic asthma in Puerto Rico. Since TPM still spans a wide range of values, we took a log transformation before implementing GFC_L. To ensure the fairness in comparison, we also set FDR control at level 0.001 which is same as the one used in other approaches.

We at first evaluated the identified gene modules. Table B.18 illustrates sizes and numbers of all different identified gene modules from the inferred network of GFC_L. It can be

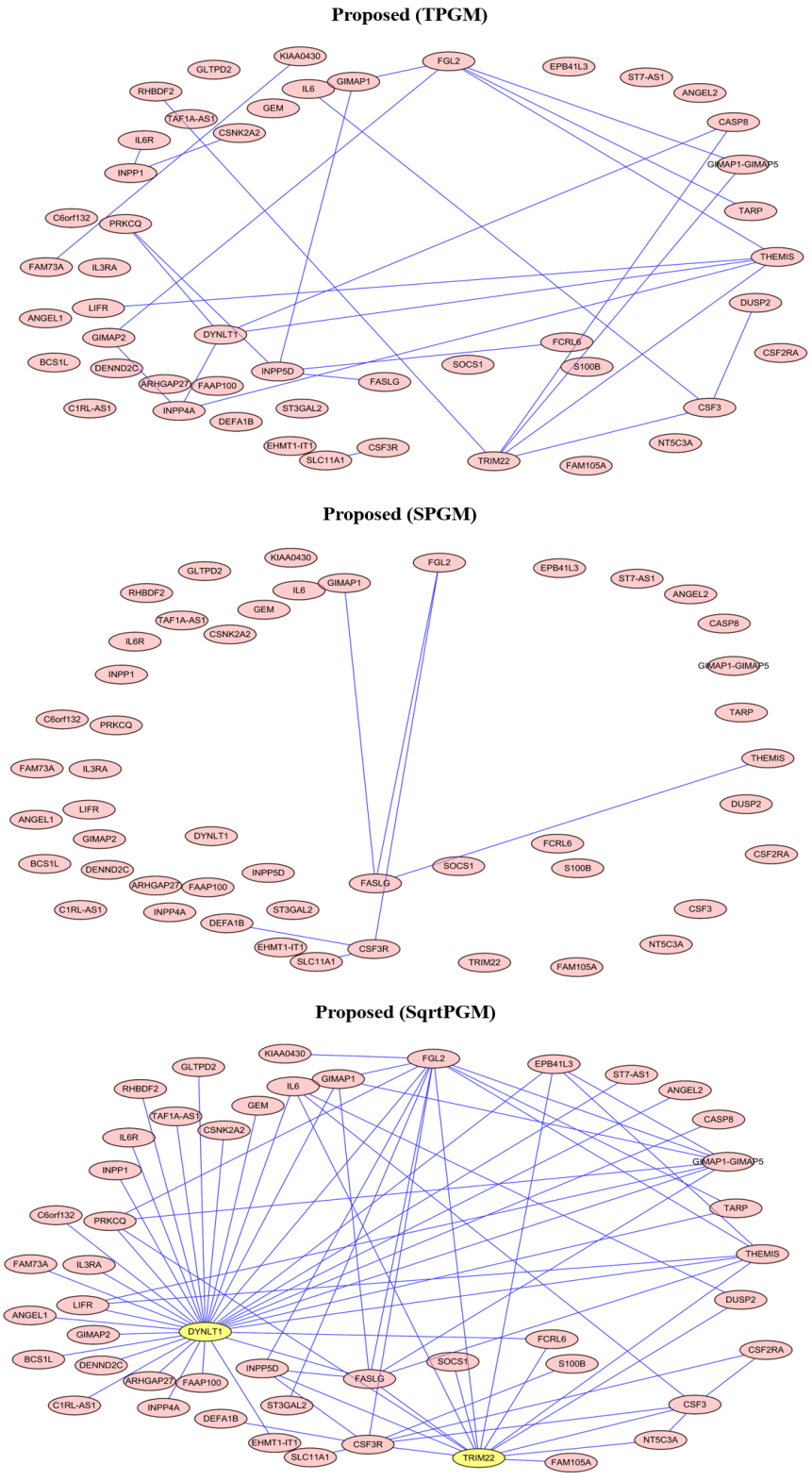

Figure B.15: Gene interactions in the module with enriched JAK-STAT signaling pathway in SqrtPGM and their corresponding interactions in TPGM and SPGM.

seen that GFC_L still fails to detect big gene modules with a size of at least 30 genes like the ones from our proposed method. These gene modules have the largest size of 5 genes. Then, we evaluated the inferred interactions from GFC_L among the 12 genes in the JAK-STAT signaling pathway. However, GFC_L is not capable of identifying important gene interactions even using TPM values.

Table B.18: The identified gene modules by GFC_L on TPM values.

Size of gene modules	Number of gene modules
1	436
2	19
3	7
4	0
5	1

In summary, the results from GFC_L on the TPM values provide additional convincing evidence that our proposed approach is more capable of detecting big gene modules and capturing gene interactions within important functional gene pathways related to atopic asthma.

B.8.6 Additional comparison of methods on liver cytochrome P450s

We further evaluated the validity of our proposed approach by comparing with the existing methods on a well-characterized and simpler data set with some established "ground truth". It is a count-valued RNA-seq data set for a liver cytochrome P450s subnetwork from humans with $n = 100$ samples and $p = 44$ genes, and we downloaded the data from the Supplementary Materials of Jia et al. (2017). Liver cytochrome P450s play important roles in drug metabolism, and P450 enzymes are particularly functionally included in the metabolism of various endogenous and exogenous chemicals (De Montellano, 2005). Through experimental work, Yang et al. (2010) uncovered a subnetwork of P450 regulatory system for

human liver shown in Figure 5C in their paper, where the known regulators and P450 genes are shown in blue rectangles and red ovals respectively. We can regard it as some established "ground truth". Our purpose is to evaluate how the proposed method can recover or identify the gene interactions from the "ground truth" using the count-valued RNA-seq data set with a comparison to GFC_L.

We at first evaluated the scale-free (or power-law) topology (Barabási and Albert, 1999) of inferred networks from different methods so as to ensure the reliability of network structure for further analysis. After implementing the pre-processing steps in Allen and Liu (2013) to normalize the original count-valued RNA-seq data, we performed the proposed approach using TPGM, SPGM and SqrtPGM with FDR control at levels 0.001, 0.005, 0.01, 0.05, 0.1 and 0.15. As a comparison study, we also implemented GFC_L on the normalized data after a log and nonparanormal transformation (Jia et al., 2017) on the original count values at same levels of FDR control. We further summarized the node degree distribution of all the inferred networks from different methods at each pre-specified level of FDR control. An inferred network with a better conformation to the power law illustrates its closer features to a real biological network. The power law can be described as $p(\lambda) \in \lambda^{-\alpha}$, where λ and $p(\lambda)$ are denoted as node degree and its corresponding probability, and α is a positive number. Numerically, it can be measured by the correlation between the $\log 2$ of node degree and the $\log 2$ of its corresponding probability. A correlation closer to -1 indicates a better conformation to the power law. The correlation values with respect to each level of FDR control from our proposed approach with three models and GFC_L are reported in Table B.19. Overall speaking, our proposed approach generates networks that are much more consistently follow scale-free topology than GFC_L. When pre-specified levels become 0.1 and 0.15, the correlation values from GFC_L are -0.4322 and 0.1344, which indicates that the inferred networks are highly deviated from a scale-free pattern. The correlation values from our approach with TPGM and SqrtPGM are consistently around -0.8 or -0.9. Even though the performance of SPGM is not as good as the other two models when FDR level is relaxed, its correlation value can be around -0.6 at level 0.15, which is still much better than GFC_L. We also included the results from nonparanormal SKEPTIC (Liu et al., 2012) with graphical Lasso as another comparison study, which directly estimates graphical model

structure using Spearman's rho or Kendall's tau. Because it is a sole estimation approach, we adopted the EBIC criterion (Foygel and Drton, 2010) to select an estimated graph. The correlation value from the selected graph using nonparanormal SKEPTIC is 0.2548, which indicates that the graph structure does not follow scale-free topology.

We then evaluated the identified gene interactions from inferred networks of GFC_L and the proposed approach with FDR control at level 0.001 because all of them follow scale-free topology well with negative correlation values stronger than -0.9. The four genes: AK097548s, BC019583, ENST00000301162 and NM_173466 in the "ground truth" were excluded from the study since they are non protein-coding genes, and their information is unavailable in the original data. In terms of the proposed approach, we focused on the results with SPGM due to its slightly better performance than TPGM and SqrtPGM in this application. We listed the identified gene interactions that overlap the subnetwork from Yang et al. (2010) using our proposed approach with SPGM and GFC_L in Table B.20. As it can be seen, our proposed approach with SPGM can capture most of these interactions identified from GFC_L, for example, CYP3A4 and CYP3A43, BCL6 and NCOA7, and CYP2A13 and CYP2A7. Moreover, the proposed approach with SPGM can recover more functionally important interactions between the known cytochrome P450 genes shown in red ovals of Figure 5C in Yang et al. (2010) than GFC_L, for example, CYP2C8 and CYP2C9, CYP2B6 and CYP2B7P1, and CYP2A6 and CYP2A7.

In summary, our proposed approach can generate gene networks with more consistent scale-free topology compared to the existing methods such as GFC_L and nonparanormal SKEPTIC, and it is also capable of identifying more important interactions between the known cytochrome P450 genes shown in the "ground truth" subnetwork compared to GFC_L.

Table B.19: Correlations between the log 2 of node degree and the log 2 of its corresponding probability of inferred networks.

	FDR level					
	0.001	0.005	0.01	0.05	0.10	0.15
GFC_L	-0.9304	-0.9304	-0.9304	-0.7734	-0.4322	0.1344
Proposed (TPGM)	-0.9395	-0.9000	-0.9000	-0.8788	-0.8192	-0.8041
Proposed (SPGM)	-0.9105	-0.8216	-0.7007	-0.6336	-0.5887	-0.5999
Proposed (SqrtPGM)	-0.9398	-0.9393	-0.9278	-0.9151	-0.9342	-0.8439

Table B.20: The identified gene interactions that overlap the subnetwork from Yang et al. (2010) by GFC_L and the proposed approach with SPGM.

GFC_L	Proposed (SPGM)
CYP3A4 —— CYP3A43	CYP3A4 —— CYP3A43
CYP2A7 —— CYP2A13	CYP2A7 —— CYP2A13
AKR1D1 —— GLYAT	CYP2C9 —— CYP2C8
NCOA7 —— BCL6	NCOA7 —— BCL6
ETNK2 —— NR1I2	CYP2B6 —— CYP2B7P1
	CYP2A7 —— CYP2A6
	FMO3 —— SLC10A1
	SLC10A1 —— AKR1D1

Table B.21: Enriched significant gene pathways from 500 genes.

Pathway name	P-value	FDR
Metal sequestration by antimicrobial proteins	3.73e−6	3.79e−3
FAS signaling pathway	4.68e−6	3.79e−3
Apoptotic cleavage of cellular proteins	2.35e−5	1.26e−2
HIV-I Nef: negative effector of Fas and TNF	4.50e−5	1.55e−2
FAS signaling pathway (CD95)	4.81e−5	1.55e−2
FasL/CD95L signaling	1.11e−4	2.76e−2
HIV-1 Nef: Negative effector of Fas and TNF-alpha	1.20e−4	2.76e−2
Apoptotic execution phase	1.79e−4	3.62e−2
JAK-STAT signaling pathway	2.23e−4	4.01e−2

B.8.7 Enriched significant gene pathways from 500 genes

The enriched significant gene pathways with FDR control at level 0.05 from all of the 500 genes are shown in Table B.21.

B.8.8 All the enriched significant gene pathways from the identified big gene modules

All the enriched significant gene pathways with FDR control at level 0.05 from the identified big gene modules are shown in Table B.22. Two identified big gene modules from the sole estimation with SPGM, see Table 3.8, fail to enrich any significant gene pathways based on the 0.05 level of FDR control. Therefore, we only include the significant results from our method with the three models.

Table B.22: All the enriched significant gene pathways from the big gene modules.

Proposed (TPGM) (Module 1: size = 312)		
Pathway name	P-value	FDR
Metal sequestration by antimicrobial proteins	5.66e−7	7.44e−4
Apoptotic cleavage of cellular proteins	1.62e−5	1.07e−2
Genes encoding secreted soluble factors	2.96e−5	1.30e−2
Apoptotic execution phase	9.70e−5	3.18e−2
Proposed (TPGM) (Module 2: size = 169)		
Pathway name	P-value	FDR
Inositol phosphate metabolism, Ins(1,3,4,5)P4 => Ins(1,3,4)P3 => myo-inositol	3.96e−5	1.42e−2
RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways	5.35e−5	1.42e−2
Synthesis of IP2, IP, and Ins in the cytosol	7.68e−5	1.42e−2
FAS signaling pathway (CD95)	8.80e−5	1.42e−2
HIV-I Nef: negative effector of Fas and TNF	9.43e−5	1.42e−2
HIV-1 Nef: Negative effector of Fas and TNF-alpha	1.63e−4	2.01e−2
Gonadotropin-releasing hormone receptor pathway	1.87e−4	2.01e−2
RIG-I-like receptor signaling pathway	2.31e−4	2.17e−2
TNF receptor signaling pathway	4.00e−4	3.24e−2
MAP00562 Inositol phosphate metabolism	4.31e−4	3.24e−2
Inositol phosphate metabolism	5.16e−4	3.53e−2
FasL/CD95L signaling	6.16e−4	3.87e−2
Ceramide signaling pathway	6.73e−4	3.90e−2
BBSome-mediated cargo-targeting to cilium	7.70e−4	4.14e−2
CLEC7A/inflammasome pathway	9.20e−4	4.62e−2
Phosphatidylinositol signaling system	1.04e−3	4.89e−2
Proposed (SPGM) (Module 1: size = 229)		
Pathway name	P-value	FDR
Metal sequestration by antimicrobial proteins	1.73e−7	1.64e−4

Table B.22: All the enriched significant gene pathways from the big gene modules (Continued).

Proposed (SPGM) (Module 3: size = 120)		
Pathway name	P-value	FDR
Inositol phosphate metabolism, Ins(1,3,4,5)P4 =>		
Ins(1,3,4)P3 => myo-inositol	9.13e−5	4.05e−2
Inositol phosphate metabolism	1.29e−4	4.05e−2
Synthesis of IP2, IP, and Ins in the cytosol	1.77e−4	4.18e−2
Proposed (SPGM) (Module 3: size = 120)		
Pathway name	P-value	FDR
FAS signaling pathway (CD95)	1.07e−6	3.82e−4
FAS signaling pathway	1.27e−6	3.82e−4
HIV-I Nef: negative effector of Fas and TNF	1.55e−6	3.82e−4
HIV-1 Nef: Negative effector of Fas and TNF-alpha	6.44e−5	9.55e−3
Caspase-mediated cleavage of cytoskeletal proteins	6.47e−5	9.55e−3
Caspase cascade in apoptosis	2.64e−4	3.25e−2
Caspase Cascade in Apoptosis	3.83e−4	3.53e−2
FasL/CD95L signaling	3.83e−4	3.53e−2
Apoptosis signaling pathway	4.40e−4	3.61e−2
Extrinsic apoptotic	5.72e−4	4.22e−2
TNFR1 signaling pathway	7.69e−4	4.83e−2
TRAIL signaling	7.97e−4	4.83e−2
Pyruvate metabolic	8.50e−4	4.83e−2
Proposed (SqrtPGM) (Module 1: size = 48)		
Pathway name	P-value	FDR
JAK-STAT signaling pathway	1.74e−9	7.51e−7
Signaling by interleukins	1.32e−8	2.84e−6
Cytokine signaling in immune system	5.08e−8	7.29e−6
Hematopoietic cell lineage	7.86e−8	8.47e−6

Table B.22: All the enriched significant gene pathways from the big gene modules (Continued).

Proposed (SqrtPGM) (Module 1: size = 48)		
Pathway name	P-value	FDR
Cytokine-cytokine receptor interaction	1.28e−7	1.10e−5
RAF-independent MAPK1/3 activation	1.95e−5	1.40e−3
FasL mediated signaling pathway	3.13e−5	1.73e−3
Interleukin-6 family signaling	3.20e−5	1.73e−3
FasL/CD95L signaling	5.22e−5	2.50e−3
Innate immune system	9.78e−5	4.22e−3
PI3K-Akt signaling pathway	1.14e−4	4.48e−3
Interleukin-4 and 13 signaling	1.33e−4	4.77e−3
Inositol phosphate metabolism	1.72e−4	4.90e−3
MAPK1(ERK2) activation	1.87e−4	4.90e−3
Inositol phosphate metabolism, Ins(1,3,4,5)P4 => Ins(1,3,4)P3 => myo-inositol	1.87e−4	4.90e−3
The TNF-type receptor Fas induces apoptosis on ligand binding	1.87e−4	4.90e−3
MAPK3(ERK1) activation	2.33e−4	4.90e−3
Measles	2.47e−4	4.90e−3
MAPK1/MAPK3 signaling	2.70e−4	4.90e−3
STAT3 pathway	2.84e−4	4.90e−3
Dimerization of procaspase-8	2.84e−4	4.90e−3
Regulation by c-FLIP	2.84e−4	4.90e−3
Interleukin-6 signaling	2.84e−4	4.90e−3
Synthesis of IP2, IP, and Ins in the cytosol	2.84e−4	4.90e−3
CASP8 activity is inhibited	2.84e−4	4.90e−3
Non-alcoholic fatty liver disease (NAFLD)	3.71e−4	6.15e−3
Regulation of necroptotic cell death	4.69e−4	7.48e−3
MAPK family signaling cascades	5.21e−4	7.51e−3
Regulation of hematopoiesis by cytokines	5.40e−4	7.51e−3

Table B.22: All the enriched significant gene pathways from the big gene modules (Continued).

Proposed (SqrtPGM) (Module 1: size = 48)		
Pathway name	P-value	FDR
Erythrocyte differentiation pathway	5.40e−4	7.51e−3
Inositol metabolism	5.40e−4	7.51e−3
Inositol phosphate metabolism	5.84e−4	7.81e−3
RIPK1-mediated regulated necrosis	6.16e−4	7.81e−3
Regulated necrosis	6.16e−4	7.81e−3
Ligand-dependent caspase activation	6.97e−4	8.35e−3
IL 17 signaling pathway	6.97e−4	8.35e−3
Herpes simplex infection	8.38e−4	9.66e−3
Apoptosis is mediated by caspases, cysteine proteases arranged in a proteolytic cascade	8.74e−4	9.66e−3
MAP00562 Inositol phosphate metabolism	8.74e−4	9.66e−3
3-phosphoinositide degradation	9.70e−4	1.02e−2
Superpathway of D-myo-inositol (1,4,5)-trisphosphate metabolism	9.70e−4	1.02e−2
Role of ERBB2 in signal transduction and oncology	1.18e−3	1.18e−2
IL 6 signaling pathway	1.18e−3	1.18e−2
Interleukin signaling pathway	1.24e−3	1.22e−2
IL-17 signaling pathway	1.28e−3	1.23e−2
Phosphatidylinositol signaling system	1.45e−3	1.36e−2
Apoptosis signaling pathway	1.67e−3	1.50e−2
Chagas disease (American trypanosomiasis)	1.67e−3	1.50e−2
Insulin signaling	1.77e−3	1.56e−2
Caspase activation via extrinsic apoptotic signaling pathway	1.91e−3	1.62e−2
Th17 cell differentiation	1.92e−3	1.62e−2
Pathways in cancer	1.98e−3	1.65e−2
Cytokines and inflammatory response	2.05e−3	1.66e−2
FAS signaling pathway (CD95)	2.19e−3	1.75e−2

Table B.22: All the enriched significant gene pathways from the big gene modules (Continued).

Proposed (SqrtPGM) (Module 1: size = 48)		
Pathway name	P-value	FDR
Interleukin-6 signaling	2.34e−3	1.78e−2
FAS signaling pathway	2.34e−3	1.78e−2
Activated TLR4 signaling	2.36e−3	1.78e−2
Interleukin receptor SHC signaling	2.85e−3	2.10e−2
FAS (CD95) signaling pathway	2.97e−3	2.10e−2
HIV-1 Nef: Negative effector of Fas and TNF-alpha	2.97e−3	2.10e−2
African trypanosomiasis	2.97e−3	2.10e−2
Toll like receptor 4 (TLR4) cascade	3.05e−3	2.12e−2
Interleukin-2 signaling	3.13e−3	2.14e−2
IL-2 receptor beta chain in T cell activation	3.50e−3	2.34e−2
Interleukin-3, 5 and GM-CSF signaling	3.53e−3	2.34e−2
Apoptotic cleavage of cellular proteins	3.68e−3	2.40e−2
Apoptosis	3.94e−3	2.54e−2
Graft-versus-host disease	4.06e−3	2.57e−2
Hepatitis B	4.44e−3	2.75e−2
IL6-mediated signaling events	4.46e−3	2.75e−2
Keratinocyte differentiation	5.09e−3	3.09e−2
Toll-like receptors cascades	5.26e−3	3.15e−2
G beta:gamma signaling through PI3Kgamma	5.53e−3	3.22e−2
Calcineurin-regulated NFAT-dependent transcription in lymphocytes	5.53e−3	3.22e−2
Interleukin-10 signaling	5.76e−3	3.27e−2
Malaria	5.76e−3	3.27e−2
Downstream signaling in naive CD8+ T cells	5.99e−3	3.27e−2
Synthesis of PIPs at the plasma membrane	5.99e−3	3.27e−2
Death receptor signaling	5.99e−3	3.27e−2
G-protein beta:gamma signaling	6.22e−3	3.35e−2

Table B.22: All the enriched significant gene pathways from the big gene modules (Continued).

Proposed (SqrtPGM) (Module 1: size = 48)		
Pathway name	P-value	FDR
Apoptotic execution phase	6.71e−3	3.57e−2
GPVI-mediated activation cascade	6.96e−3	3.66e−2
Legionellosis	7.21e−3	3.74e−2
Apoptosis	7.51e−3	3.84e−2
Signaling by SCF-KIT	7.57e−3	3.84e−2
Programmed cell death	7.87e−3	3.90e−2
Genes encoding secreted soluble factors	7.88e−3	3.90e−2
HIV-I Nef: negative effector of Fas and TNF	7.99e−3	3.91e−2
Axon guidance	8.38e−3	4.06e−2
IL12-mediated signaling events	8.81e−3	4.22e−2
Ion influx/efflux at host-pathogen interface	9.29e−3	4.35e−2
Activation, myristolyation of BID and translocation to mitochondria	9.29e−3	4.35e−2
Proposed (SqrtPGM) (Module 4: size = 114)		
Pathway name	P-value	FDR
Metal sequestration by antimicrobial proteins	3.66e−6	2.74e−3

Bibliography

L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Physical Review E*, 64(4):046135, 2001.

G. I. Allen and Z. Liu. A local Poisson graphical model for inferring networks from sequencing data. *IEEE Transactions on NanoBioscience*, 12(3):189–198, 2013.

E. Almaas and A.-L. Barabási. Power laws in biological networks. In *Power Laws, Scale-free Networks and Genome Biology*, pages 1–11. Springer, 2006.

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

N. Andor, E. F. Simonds, D. K. Czerwinski, J. Chen, S. M. Grimes, C. Wood-Bouwens, G. X. Zheng, M. A. Kubit, S. Greer, W. A. Weiss, et al. Single-cell RNA-Seq of follicular lymphoma reveals malignant B-cell types and coexpression of T-cell immune checkpoints. *Blood*, 133(10):1119–1129, 2019.

M. Avella-Medina, H. S. Battey, J. Fan, and Q. Li. Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, 105(2):271–284, 2018.

S. K. Banerjee, A. P. Weston, M. N. Zoubine, D. R. Campbell, and R. Cherian. Expression of cdc2 and cyclin B1 in Helicobacter pylori-associated gastric MALT and MALT lymphoma: relationship to cell death, proliferation, and transformation. *The American Journal of Pathology*, 156(1):217–225, 2000.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439): 509–512, 1999.

A.-L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

R. F. Barber and M. Drton. High-dimensional Ising model selection with Bayesian information criteria. *Electronic Journal of Statistics*, 9(1):567–607, 2015.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

A. T. Braithwaite, H. M. Marriott, and A. Lawrie. Divergent roles for trail in lung diseases. *Frontiers in Medicine*, 5:212, 2018.

A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, and G. G. Lara. Arrayexpress - a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 31(1):68–71, 2003.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

T. Cai, H. Li, J. Ma, and Y. Xia. Differential Markov random field analysis with an application to detecting differential microbial community networks. *Biometrika*, 106(2):401–416, 2019.

E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

O. Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *ArXiv Preprint ArXiv:1603.05229*, 2016.

L. C. Cerchietti, E. C. Lopes, S. N. Yang, K. Hatzi, K. L. Bunting, L. A. Tsikitas, A. Mallik, A. I. Robles, J. Walling, L. Varticovski, et al. A purine scaffold Hsp90 inhibitor destabilizes BCL-6 and has specific antitumor activity in BCL-6-dependent B cell lymphomas. *Nature Medicine*, 15(12):1369–1376, 2009.

J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, 37(suppl_2): W305–W311, 2009.

M. Chen, C. Gao, Z. Ren, and H. H. Zhou. Sparse CCA via precision adjusted iterative thresholding. *Proceedings of the Seventh International Congress of Chinese Mathematicians*, II:481–534, 2019.

A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

S. Crotty, R. J. Johnston, and S. P. Schoenberger. Effectors and memories: Bcl-6 and Blimp-1 in T and B lymphocyte differentiation. *Nature Immunology*, 11(2):114–120, 2010.

F. D'Acquisto, A. Merghani, E. Lecona, G. Rosignoli, K. Raza, C. D. Buckley, R. J. Flower, and M. Perretti. Annexin-1 modulates T-cell activation and differentiation. *Blood*, 109(3): 1095–1102, 2007.

A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

P. R. O. De Montellano. *Cytochrome P450: structure, mechanism, and biochemistry*. Springer Science & Business Media, 2005.

D. Eddelbuettel, R. François, J. Allaire, K. Ushey, Q. Kou, N. Russel, J. Chambers, and D. Bates. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8): 1–18, 2011.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

O. A. El-Banhawy and I. El-Desoky. Low-grade primary mucosa-associated lymphoid tissue lymphoma of the nasopharynx: clinicopathological study. *American Journal of Rhinology*, 19(4):411–416, 2005.

B. Falini, E. Tiacci, A. Liso, K. Basso, E. Sabattini, R. Pacini, R. Foa, A. Pulsoni, R. Dalla Favera, and S. Pileri. Simple diagnostic assay for hairy cell leukaemia by immunocytochemical detection of annexin A1 (ANXA1). *The Lancet*, 363(9424):1869–1871, 2004.

P. V. Filip, D. Cuciureanu, L. S. Diaconu, A. M. Vladareanu, and C. S. Pop. MALT lymphoma: epidemiology, clinical diagnosis and treatment. *Journal of Medicine and Life*, 11(3):187–193, 2018.

M. Filteau, S. A. Pavey, J. St-Cyr, and L. Bernatchez. Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish. *Molecular Biology and Evolution*, 30(6):1384–1396, 2013.

L. Flossbach, E. Antoneag, M. Buck, R. Siebert, T. Mattfeldt, P. Möller, and T. F. Barth. BCL6 gene rearrangement and protein expression are associated with large cell presentation of extranodal marginal zone B-cell lymphoma of mucosa-associated lymphoid tissue. *International Journal of Cancer*, 129(1):70–77, 2011.

E. Forno, T. Wang, C. Qi, Q. Yan, C.-J. Xu, N. Boutaoui, Y.-Y. Han, D. E. Weeks, Y. Jiang, F. Rosser, et al. DNA methylation in nasal epithelium, atopy, and atopic asthma in children: a genome-wide study. *The Lancet Respiratory Medicine*, 7(4):336–346, 2019.

E. Forno, R. Zhang, Y. Jiang, S. Kim, Q. Yan, Z. Ren, Y.-Y. Han, N. Boutaoui, F. Rosser, D. E. Weeks, et al. Transcriptome-wide and differential expression network analyses of childhood asthma in nasal epithelium. *Journal of Allergy and Clinical Immunology*, 2020.

R. Foygel and M. Drton. Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, pages 604–612, 2010.

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

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

C. Gaiteri, Y. Ding, B. French, G. C. Tseng, and E. Sibille. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes, Brain and Behavior*, 13(1):13–24, 2014.

C. Gao, Z. Ma, Z. Ren, and H. H. Zhou. Minimax estimation in sparse canonical correlation analysis. *The Annals of Statistics*, 43(5):2168–2197, 2015.

C. Gao, Z. Ma, and H. H. Zhou. Sparse CCA: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, 2017.

D. Gerard. Data-based RNA-seq simulations by binomial thinning. *BMC Bioinformatics*, 21:1–14, 2020.

A. Gerasch, D. Faber, J. Küntzer, P. Niermann, O. Kohlbacher, H.-P. Lenhof, and M. Kaufmann. BiNA: a visual analytics tool for biological network data. *PLoS One*, 9(2):e87397, 2014.

L. Giovannini-Chami, B. Marcet, C. Moreilhon, B. Chevalier, M. I. Illie, K. Lebrigand, K. Robbe-Sermesant, T. Bourrier, J.-F. Michiels, B. Mari, et al. Distinct epithelial gene expression phenotypes in childhood respiratory allergy. *European Respiratory Journal*, 39 (5):1197–1205, 2012.

C. Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014.

N. Gour, S. Lajoie, U. Smole, M. White, D. Hu, P. Goddard, S. Huntsman, C. Eng, A. Mak, S. Oh, et al. Dysregulated invertebrate tropomyosin–dectin-1 interaction confers susceptibility to allergic diseases. *Science Immunology*, 3(20):eaam9841, 2018.

S. Hadebe, F. Brombacher, and G. D. Brown. C-type lectins receptors in asthma. *Frontiers in Immunology*, 9:733, 2018.

D. R. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

N. J. Higham. Computing the nearest correlation matrix – a problem from finance. *IMA journal of Numerical Analysis*, 22(3):329–343, 2002.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101, 1964.

D. Inouye, P. Ravikumar, and I. Dhillon. Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive dependencies. In *International Conference on Machine Learning*, pages 2445–2453, 2016.

S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, 2014.

J. Janková and S. van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

J. Janková and S. van de Geer. Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test*, 26(1):143–162, 2017.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15:2869–2909, 2014.

B. Jia, S. Xu, G. Xiao, V. Lamba, and F. Liang. Learning gene regulatory networks from next generation sequencing data. *Biometrics*, 73(4):1221–1230, 2017.

S. E. Josefsson, K. Beiske, Y. N. Blaker, M. S. Førsund, H. Holte, B. Østenstad, E. Kimby, H. Köksal, S. Wälchli, B. Bai, et al. TIGIT and PD-1 mark intratumoral T cells with reduced effector function in B-cell non-Hodgkin lymphoma. *Cancer Immunology Research*, 7(3):355–362, 2019.

M. Kabesch, M. Schedel, D. Carr, B. Woitsch, C. Fritzsch, S. K. Weiland, and E. von Mutius. IL-4/IL-13 pathway genetics strongly influence serum IgE levels and childhood asthma. *Journal of Allergy and Clinical Immunology*, 117(2):269–274, 2006.

Y. Ke, S. Minsker, Z. Ren, Q. Sun, and W.-X. Zhou. User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3):454–471, 2019.

P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

K.-A. Lê Cao, P. G. Martin, C. Robert-Granié, and P. Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10:34, 2009.

S. Li, Z. Ren, C.-H. Zhang, and H. H. Zhou. Asymptotic normality in estimation of large Ising graphical models. *Unpublished Manuscript*, 2016.

X. Li, T. Zhao, X. Yuan, and H. Liu. The flare package for high dimensional linear regression and precision matrix estimation in R. *The Journal of Machine Learning Research*, 16:553–557, 2015.

L. Liang, N. Morar, A. L. Dixon, G. M. Lathrop, G. R. Abecasis, M. F. Moffatt, and W. O. Cookson. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Research*, 23(4):716–726, 2013.

G. Lima-Mendez and J. van Helden. The powerful law of the power law and other myths in network biology. *Molecular BioSystems*, 5(12):1482–1493, 2009.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10: 2295–2328, 2009.

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

W. Liu. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978, 2013.

S. Lysen. Permuted inclusion criterion: a variable selection technique. *Publicly Accessible Penn Dissertations*, page 28, 2009.

E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, and E. M. Martersteck. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

A. Martin, M. E. Ochagavia, L. C. Rabasa, J. Miranda, J. Fernandez-de Cossio, and R. Bringas. BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinformatics*, 11:91, 2010.

L. Mazutis, J. Gilbert, W. L. Ung, D. A. Weitz, A. D. Griffiths, and J. A. Heyman. Single-cell analysis and sorting using droplet-based microfluidics. *Nature Protocols*, 8(5):870–891, 2013.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

M. E. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

G. Pandey, O. P. Pandey, A. J. Rogers, M. E. Ahsen, G. E. Hoffman, B. A. Raby, S. T. Weiss, E. E. Schadt, and S. Bunyavanich. A nasal brush-based classifier of asthma identified by machine learning analysis of nasal RNA sequence data. *Scientific Reports*, 8:8826, 2018.

H. Pang, H. Liu, and R. Vanderbei. The fastclime package for linear programming and large-scale precision matrix estimation in R. *The Journal of Machine Learning Research*, 15:489–493, 2014.

N. N. Parikshak, M. J. Gandal, and D. H. Geschwind. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics*, 16(8): 441–458, 2015.

E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34, 2009.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

J. Perez-Garcia, E. Herrera-Luis, F. Lorenzo-Diaz, M. González, O. Sardón, J. Villar, and M. Pino-Yanes. Precision medicine in childhood asthma: Omic studies of treatment response. *International Journal of Molecular Sciences*, 21(8):2908, 2020.

O. Potapinska and U. Demkow. T lymphocyte apoptosis in asthma. *European Journal of Medical Research*, 14(4):192–195, 2009.

Y. Qiu and X.-H. Zhou. Estimating c-level partial correlation graphs with application to brain imaging. *Biostatistics*, 2018.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

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

M. Rincon and C. G. Irvin. Role of IL-6 in asthma and other inflammatory pulmonary diseases. *International Journal of Biological Sciences*, 8(9):1281, 2012.

S. E. Safo, J. Ahn, Y. Jeon, and S. Jung. Sparse generalized eigenvalue problem with application to canonical correlation analysis for integrative analysis of methylation and gene expression data. *Biometrics*, 74(4):1362–1371, 2018.

B. Schuhmacher, B. Rengstl, C. Döring, J. Bein, S. Newrzela, U. Brunnberg, H. M. Kvasnicka, M. Vornanen, R. Küppers, M.-L. Hansmann, et al. A strong host response and lack of MYC expression are characteristic for diffuse large B cell lymphoma transformed from nodular lymphocyte predominant hodgkin lymphoma. *Oncotarget*, 7(44):72197, 2016.

P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled Lasso. *The Journal of Machine Learning Research*, 14:3385–3418, 2013.

Y. V. Sun and Y.-J. Hu. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. In *Advances in Genetics*, volume 93, pages 147–190. Elsevier, 2016.

X. Suo, V. Minden, B. Nelson, R. Tibshirani, and M. Saunders. Sparse canonical correlation analysis. *ArXiv Preprint ArXiv:1705.10865*, 2017.

C. A. Tegla, C. D. Cudrici, V. Nguyen, J. Danoff, A. M. Kruszewski, D. Boodhoo, A. P. Mekala, S. I. Vlaicu, C. Chen, V. Rus, et al. RGC-32 is a novel regulator of the T-lymphocyte cell cycle. *Experimental and Molecular Pathology*, 98(3):328–337, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

I. Tirosh and M. L. Suvà. Deciphering human tumor biology by single-cell expression profiling. *Annual Review of Cancer Biology*, 3:151–166, 2019.

K. Vale. Targeting the JAK-STAT pathway in the treatment of Th2-high severe asthma. *Future Medicinal Chemistry*, 8(4):405–419, 2016.

K. Van Crombruggen, T. Vogl, C. Pérez-Novo, G. Holtappels, and C. Bachert. Differential release and deposition of S100A8/A9 proteins in inflamed upper airway tissue. *European Respiratory Journal*, 47(1):264–274, 2016.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3): 1166–1202, 2014.

G. R. Verheyen, J.-M. Nuijten, P. Van Hummelen, and G. R. Schoeters. Microarray analysis of the effect of diesel exhaust particles on in vitro cultured macrophages. *Toxicology in Vitro*, 18(3):377–391, 2004.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Y.-W. Wan, G. I. Allen, Y. Baker, E. Yang, P. Ravikumar, M. Anderson, and Z. Liu. XMRF: an R package to fit Markov networks to high-throughput genetics data. *BMC Systems Biology*, 10:69, 2016.

H. Wang, M. FitzPatrick, N. J. Wilson, D. Anthony, P. C. Reading, C. Satzke, E. M. Dunne, P. V. Licciardi, H. J. Seow, K. Nichol, et al. CSF3R/CD114 mediates infection-dependent

transition to severe asthma. *Journal of Allergy and Clinical Immunology*, 143(2):785–788, 2019.

J. Wang and M. Kolar. Inference for high-dimensional exponential family graphical models. In *Artificial Intelligence and Statistics*, pages 1042–1050, 2016.

L. Wang, C. Zheng, W. Zhou, and W.-X. Zhou. A new principle for tuning-free Huber regression. *Statistica Sinica*, 2020.

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.

T. Wartewig and J. Ruland. PD-1 tumor suppressor signaling in T cell lymphomas. *Trends in Immunology*, 40(5):403–414, 2019.

M. T. Weirauch. Gene coexpression networks for the analysis of DNA microarray data. *Applied Statistics for Network Biology: Methods in Systems Biology*, 1:215–250, 2011.

J. W. Williams, C. M. Ferreira, K. M. Blaine, C. Rayon, F. Velázquez, J. Tong, M. E. Peter, and A. I. Sperling. Non-apoptotic Fas (CD95) signaling on T cells regulates the resolution of Th2-mediated inflammation. *Frontiers in Immunology*, 9:2521, 2018.

Q. F. Wills, K. J. Livak, A. J. Tipping, T. Enver, A. J. Goldson, D. W. Sexton, and C. Holmes. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature Biotechnology*, 31(8):748–752, 2013.

D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–27, 2009.

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.

D. M. Witten, J. H. Friedman, and N. Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

Z. Y. Xu-Monette, J. Zhou, and K. H. Young. PD-1 expression and clinical PD-1 blockade in B-cell lymphomas. *Blood, The Journal of the American Society of Hematology*, 131(1): 68–83, 2018.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. On Poisson graphical models. In *Advances in Neural Information Processing Systems*, pages 1718–1726, 2013.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16:3813–3847, 2015.

I. V. Yang, B. S. Pedersen, A. H. Liu, G. T. O'Connor, D. Pillai, M. Kattan, R. T. Misiak, R. Gruchalla, S. J. Szefler, G. K. K. Hershey, et al. The nasal methylome and childhood atopic asthma. *Journal of Allergy and Clinical Immunology*, 139(5):1478–1488, 2017.

X. Yang, B. Zhang, C. Molony, E. Chudin, K. Hao, J. Zhu, A. Gaedigk, C. Suver, H. Zhong, J. S. Leeder, et al. Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome Research*, 20(8):1020–1036, 2010.

M. Yu, M. Kolar, and V. Gupta. Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems*, pages 2829–2837, 2016.

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

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

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):217–242, 2014.

N.-Z. Zhang, X.-J. Chen, Y.-H. Mu, and H. Wang. Identification of differentially expressed genes in childhood asthma. *Medicine*, 97(21):e10861, 2018a.

R. Zhang, Z. Ren, and W. Chen. SILGGM: an extensive R package for efficient statistical inference in large-scale gene networks. *PLoS Computational Biology*, 14(8):e1006369, 2018b.

Y. Zhang, Z. Ouyang, and H. Zhao. A statistical framework for data integration through graphical models with application to cancer genomics. *The Annals of Applied Statistics*, 11(1):161–184, 2017.

P. Zhao and B. Yu. On model selection consistency of Lasso. *The Journal of Machine learning research*, 7:2541–2563, 2006.

T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. The huge package for high-dimensional undirected graph estimation in R. *The Journal of Machine Learning Research*, 13:1059–1062, 2012.

G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017.

I. Zwiener, B. Frisch, and H. Binder. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS One*, 9(1):e85150, 2014.