IOWA STATE UNIVERSITY
Digital Repository

Creative Components

Iowa State University Capstones, Theses and Dissertations

Fall 2019

Gene Network Reconstruction with c-level Partial Correlation Graph

Hao Wang
Iowa State University, halewang@iastate.edu

Follow this and additional works at: https://lib.dr.iastate.edu/creativecomponents

 Part of the Applied Statistics Commons, Biostatistics Commons, Microarrays Commons, Probability Commons, Statistical Methodology Commons, and the Statistical Models Commons

Recommended Citation

Wang, Hao, "Gene Network Reconstruction with c-level Partial Correlation Graph" (2019). *Creative Components*. 462.
https://lib.dr.iastate.edu/creativecomponents/462

This Creative Component is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Creative Components by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

Gene Co-expression Network Reconstruction with c-level Partial Correlation Graph

Hao Wang
Department of Statistics, Iowa State University
and
Peng Liu
Department of Statistics, Iowa State University
and
Yumou Qiu
Department of Statistics, Iowa State University
and
Chong Wang
Department of Statistics, Iowa State University

November 20, 2019

Abstract

A key aim in system biology is to understand molecules' structural and functional processes in a living cell. With the development of high-throughput technologies, quantitative methods can be applied on large scale 'omics' datasets. Due to the nature of intricate relationships of all molecules in a cell, network-based methods have become a popular approach to reconstruct gene-gene, gene-protein, and protein-protein interactions. Among different network approaches, Gaussian Graphical Model shows advantages in reconstructing gene co-expression networks because it is able to capture the direct association between genes with partial correlations. However, estimating and infering partial correlations under high-dimensional setting are very challenging. A method utilizing penalized partial correlations called exact hypothesis testing for shrinkage based Gaussian graphical models (Shrunk MLE) is able to overcome the high-dimension problem. However, the statistical inference of such penalized partial correlations is not satisfying. In this project, a novel network inference method, named c-level Partial Correlation Graph (c-level PCG), is applied on the gene expression dataset to model gene-gene direct association. It overcomes the ill-condition of p greater than n and successfully infers estimated partial correlation with false discovery rate controlled. Compared to Shrunk MLE, c-level PCG is able to achieve much higher statistical power and control the false discovery rate at the same time, according to our simulation studies.

Keywords: Gene Co-expression Network, Gaussian Graphical Model, Partial Correlation, Statistical Inference, FDR control

1 Introduction

In system biology, a main interest, yet a challenge, is discovering and understanding the complex functional interactions between all molecules at the level of the cell Barabasi & Oltvai (2004), Boccaletti (2010). Generally, networks can be used to describe complex social, physical or chemical systems Albert & Barabási (2002). In the context of system biology, the complex system in a cell describes a network in which each molecule is a node and an edge between two molecules stands for they participate in any biochemical process together. More specifically, cellular networks include gene co-expression networks Tieri et al. (2019), gene regulatory networks Emmert-Streib et al. (2014), and protein-protein interaction networks De Las Rivas & Fontanillo (2010). These network analyses can help researchers determine how genes and proteins behave and interact in a cell, locate biomarkers, and detect functional modules. In this study, we will focus on gene co-expression networks. Formally, the gene co-expression network is defined as the undirected network (without considering the direction of a edge) where the edge exists if two genes have palpable co-expression association Tieri et al. (2019).

Recently, high-throughput genomics technologies such as DNA microarray Heller (2002) and next-generation sequencing Ansorge (2009) provide large-scale and high-quality genomics datasets. A typical gene expression dataset is a matrix with n columns and p rows, where each row represents a molecule (usually a gene or protein) and each column is an observation under certain experimental condition. Such a gene expression dataset is often extremely high-dimensional ($p >> n$), which is referred to as "large p with small n" scenario. Despite previous theoretical or experimental studies of inner interaction webs in a cell, quantitative methods can now be applied on genomics datasets to systematically identify working patterns among genes, proteins, and other molecules.

Many statistical methods have been developed as a means of reconstructing biological networks. Overall, reconstructed biological co-expression networks can be grouped into three categories: correlation based network, Gaussian graphical model, and information theory based network Yu et al. (2013), Wang & Huang (2014). The Pearson correlation is the most simple and common probabilistic dependency. Correlation based networks use the Pearson correlation to measure the marginal association between genes. This leads to

the limitation that direct interaction between genes cannot be distinguished. Also, using the Pearson correlation can make every pair of genes connected since Pearson correlation is very likely to be different from 0 with a small sample size. A remedy is using hard or soft thresholding to select those pairs of genes with stronger association Zhang & Horvath (2005). But the selection of thresholding is lack of statistical justification. Unlike methods based on the Pearson correlation, information theory based networks use mutual information to measure the gene association, which is a more generalized measure of statistical dependency Margolin et al. (2006), Meyer et al. (2007). Although mutual information can detect nonlinear dependencies, it still cannot measure direct associations between genes Meyer et al. (2007). Some methods extend mutual information to conditional mutual information to distinguish the indirect and direct association. However, a review paper Soranzo et al. (2007) compared different methods of re-constructing gene co-expression network and concluded that Gaussian Graphical Model beats other methods in terms of performance versus runtime. In short, Gaussian graphical model shows its advantage because it models direct association by conditional correlation or partial correlation. Such probabilistic dependency captures the linear relationship between any pair of genes conditioning on all other genes, meaning eliminating other effects of other genes on desired edge.

To cover the basic theory of Gaussian graphical model, we will start with an undirected graph $G = (V, E)$, where V is a set of nodes and E is a set of edges. Let $X = (X_1, X_2, X_3, ..., X_p)$ denote the random vector associated with p genes, where each X_j is an individual gene. In Gaussian graphical model, X is assumed to have multi-normal distribution with mean zero and covariance matrix Σ. Observation $X_{(i)} = (X_{(i)1}, X_{(i)2}, ..., X_{(i)p})$, $i = 1, 2, ..., n$, is sampled from $N(0, \Sigma)$. A Gaussian graphical model is represented by the corresponding partial correlation matrix P.

Theoretically, there are two approaches to compute the partial correlation. First one is using the precision matrix Ω, which is simply the inverse of covariance matrix Σ (i.e. $\Omega = \Sigma^{-1}$). Then, the partial correlation between variable X_i and variable X_j, denoted as ρ_{ij} is defined by $\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} I(i \neq j) + I(i = j)$. The second approach uses regressions. The partial correlation between X_i and X_j given the other $p-2$ controlling variables $X_{-(i,j)}$ (the set of variables from X_1 to X_p except X_i and X_j), written ρ_{ij}, is the Pearson correlation

between the residuals ϵ_i and ϵ_j resulting from the linear regression of X_i with $X_{-(i,j)}$ and of X_j with $X_{-(i,j)}$.

However, under the high-dimensional setting, sample covariance matrix is singular. Even when p is smaller but not much smaller than n, sample covariance matrix is invertible but estimated partial correlations can be vary inaccurate. The regression approach also has similar problems. When dimension is high, estimating all partial correlations requires fitting $p \times (p-1)/2$ regressions, which is computationally inefficient. The regression on $p-2$ covariates with small sample size n can have very poor performance as well. A natural idea to overcome the high-dimensional challenge is to include penalty. Shrinkage based sample covariance estimator Ledoit & Wolf (2004) and penalized regressions based partial correlation estimator Peng et al. (2009) are two general solutions. However, these shrinkage or penalization based methods lack of statistical inference.

An existing method, exact hypothesis testing for shrinkage based Gaussian graphical models (Shrunk MLE) Bernal et al. (2019) can overcome these challenges. Shrunk MLE estimates precision matrix by directly penalize sample covariance matrix to p by p identity matrix to make it singular Schäfer & Strimmer (2005). The correct null distribution of the penalized partial correlation (shrunk partial correlation) was then proposed and eventually $p * (p-1)/2$ estimates can be tested simultaneously with Benjamini-Hochberg correction Benjamini & Hochberg (1995). This method well improved the limitation of using traditional null distribution of partial correlation by developing the exact null distribution of shrunk partial correlation and successfully reduced the number of false positives. However, after multiple testing correction, Shrunk MLE is too conservative to give us many significant edges.

To overcome mentioned limitations above, a novel regression based partial correlation inference method, c-level partial correlation graph (c-level PCG) Qiu & Zhou (2018), can be applied to reconstruct gene co-expression network.

In this study, we implemented c-level Partial Correlation Graph and applied it on gene expression datasets. Under the assumption that real gene co-expression network is very sparse Barabasi & Oltvai (2004) and there exists gene modules Hartwell et al. (1999), we justified c-level PCG outperform Shrunk MLE with scientifically simulated data. The study

revealed that reconstructed gene co-expression networks by c-level PCG are accurate and informative.

The paper is organized as follows. In the methods section, two methods discussed in this study were briefly reviewed. In the results section, several settings of simulating synthetic networks were introduced and performance of two methods on simulated data were compared. Results of applying c-level PCG on a real dataset were also presented in this section. Finally, conclusion, limitations and future works will be discussed in the discussion section.

2 Methods

Suppose that we have a gene expression dataset with p genes and n observations under n different experimental conditions. Let y_{ij} denote the measurement of expression of j^{th} gene from i^{th} sample, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$. Assume that $y_i = (y_{i,1}, ..., y_{i,p})^T$ are independent and identically distributed (i.i.d.) random vectors with mean zero and covariance $\Sigma = \{\sigma_{ij}\}_{p*p}$. Let $\Omega = \Sigma^{-1} = \{\omega_{ij}\}_{p\times p}$ be the corresponding precision matrix.

2.1 Shrunk MLE

To estimated the precision matrix Ω, Bernal et al. (2019) took over the shrinkage-based estimator of sample covariance matrix of p genes with n observations ($p >> n$). The shrinkage-based estimator has the form

$$\hat{\Sigma} = S_\lambda = (1 - \lambda)S + \lambda T$$

where S is sample covariance matrix, T is the p by p non-singular matrix and λ is the shrinkage parameter. This estimator simply penalizes the sample covariance matrix to a non-singular matrix with the assumption that there are very few genes are really connected. Matrix T can have many different forms, such as an identity matrix or a diagonal matrix with diagonal elements equal to sample variance. λ is estimated by minimizing the mean square error of S_λ. A analytical expression of optimal λ with different choice of T can be found in Schäfer & Strimmer (2005). Then estimated partial correlation $\hat{\rho}_{ij}$ is equal to $-\frac{\hat{\omega}_{ij}}{\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}} I(i \neq j) + I(i = j)$, where $\{\hat{\omega}_{ij}\}$=$\hat{\Omega} = {S_\lambda}^{-1}$.

To infer the estimated partial correlation, previously Schäfer & Strimmer (2005) studied the distribution of the estimator. Estimated partial correlation is assumed to have a mixture density of the form

$$f(\hat{\rho}) = \pi_0 f_0(\hat{\rho}) + (1 - \pi_0) f_1(\hat{\rho}),$$

where π_0 is the proposion of the null edges, $f_0(\hat{\rho})$ is the probability density under the null ($\rho = 0$), and $f_1(\hat{\rho})$ is the probability density for the alternative ($\rho \neq 0$). According to simulation studies (for small λ) the distribution of the estimated 'shrunk' partial correlation is close to the standard distribution of estimated partial correlation (i.e. without shrinkage) as

$$f_0(\hat{\rho}) = \frac{1}{Beta(\frac{1}{2}, \frac{k-1}{2})}(1 - \hat{\rho}^2)^{(k-3)/2}.$$

And to make the mixture density simple, f_1 is assumed to be the density function of $U(-1, 1)$. Parameter k and π_0 are estimated by maximizing the mixture likelihood with estimated shrunk partial correlations. Inference is then based on empirical null fitting or corrected p-values for multiple testing, where p-values are calculated with estimated null density. More detailed explanation about empirical null fitting can be found in Efron (2004), Efron (2005), which can be viewed as the posterior probability of null edges given the estimate (i.e. $Prob(Null|\hat{\rho})$).

However the estimated partial correlations involve the shrinkage paramter λ. Assuming the null density of $\hat{\rho}$ to have a standard form can make the p-values sub-optimal. So Bernal et al. (2019) improved this inference procedure by proposing the exact null distribution of shrinkage based partial correlation estimator. They proposed that the exact distribution of shrunk partial correlation is

$$f_0^{\lambda}(\hat{\rho}_{\lambda}) = \frac{((1-\lambda)^2 - \hat{\rho}_{\lambda}^2)^{(k-3)/2}}{Beta(\frac{1}{2}, \frac{k-1}{2})(1-\lambda)^{(k-2)}}.$$

In such way, the parameters k and λ are estimated by maximizing null density with simulated shrunk partial correlations where λ is the same analytically calculated one. p-values can be find by the estimated null distribution and multiple testing correction can be applied to control false discovery rate.

2.2 c-level PCG

Starting with the definition of partial correlation calculated by regressions mentioned in the Introduction and based on Lemma 1 in Peng et al. (2009), the partial correlation can be expressed via only p node-wise regressions. Let Y_{-j} denote the $p-1$ dimensional random vector $(y_1, ..., y_{j-1}, y_{j+1}, ..., y_p)^T$. Theoretically,

$$y_{j_1} = \alpha_{j_1,0} + \sum_{j_2 \neq j_2} \alpha_{j_1,j_2} y_{j_2} + \epsilon_{j_1},\ j_1 = 1, 2, ..., p.$$

The regression error ϵ_{j_1} is uncorrelated with Y_{-j1} if and only if $\alpha_{j_1,j_2} = -\frac{\omega_{j_1,j_2}}{\omega_{j_1,j_1}}$ for all $j_2 \neq j_1$. We can then write the variance of ϵ_{j_1} and covariance of ϵ_{j_1} and ϵ_{j_2} in terms of ω's

$$Var(\epsilon_{j_1}) = \frac{1}{\omega_{j_1,j_1}},\ \text{and } Cov(\epsilon_{j_1}, \epsilon_{j_2}) = \frac{\omega_{j_1,j_2}}{\omega_{j_1,j_1}\omega_{j_2,j_2}} = -\frac{\rho_{j_1,j_2}}{(\omega_{j_1,j_1}\omega_{j_2,j_2})^{1/2}}.$$

Let $\epsilon = (\epsilon_1, ..., \epsilon_p)^T$ and $V = Cov(\epsilon) = \{v_{j_1,j_2}\}_{p\times p}$. The partial correlation can be expressed as $\rho_{j_1,j_2} = -\frac{v_{j_1,j_2}}{\sqrt{\omega_{j_1,j_1}\omega_{j_2,j_2}}},\ j_1 \neq j_2$.

In practice, to get estimated partial correlation, node-wide regressions are fitted by lasso and tunning parameter λ is pre-specified as $\sqrt{2 \times log(p)/n}$. Let $\hat{\epsilon}_i = (\hat{\epsilon}_{i,1}, ..., \hat{\epsilon}_{i,p})^T$ be the residuals of the i^{th} observation and $\tilde{V} = \{\tilde{v}_{j_1,j_2}\}$ be the sample covariance of the residuals, where $\tilde{v}_{j_1,j_2} = \sum_{i=1}^n \frac{\hat{\epsilon}_{i,j_1}\hat{\epsilon}_{i,j_2}}{n}$. Although $\sum_{i=1}^n \frac{\epsilon_{i,j_1}\epsilon_{i,j_2}}{n}$ is an unbiased estimator of v_{j_1,j_2}, replacing $\epsilon_{i,j}$ by $\hat{\epsilon}_{i,j}$ will incur a bias term. In Qiu & Zhou (2018), a novel estimator of v_{j_1,j_2} was proposed by the authors with the form

$$\hat{v}_{j_1,j_2} = \begin{cases} -\frac{1}{n}\sum_{i=1}^n (\hat{\epsilon}_{i,j_1}\hat{\epsilon}_{i,j_2} + \hat{\alpha}_{j_1,j_2}\hat{\epsilon}_{i,j_2}^2 + \hat{\alpha}_{j_2,j_1}\hat{\epsilon}_{i,j_1}^2),\ j_1 \neq j_2 \\ \hat{v}_{j_1,j_2} = \sum_{i=1}^n \frac{\hat{\epsilon}_{i,j_1}\hat{\epsilon}_{i,j_2}}{n},\ j_1 = j_2. \end{cases}$$

Then the partial correlation between gene j_1 and j_2 is estimated by

$$\hat{\rho}_{j_1,j_2} = -\hat{v}_{j_1,j_2} \times \sqrt{\hat{\omega}_{j_1,j_1}\hat{\omega}_{j_2,j_2}} = -\frac{\hat{v}_{j_1,j_2}}{\sqrt{\hat{v}_{j_1,j_1}\hat{v}_{j_2,j_2}}}.$$

Inference of estimated partial correlation is based on the uncertainty of the estimator and false discovery rate control. Variance of the estimator is $nVar(\hat{\rho}_{j_1,j_2}) = \kappa(1-\rho_{j_1,j_2}^2)^2\{1+o(1)\}$, where $\kappa = E(\epsilon_j^4)/[3E^2(\epsilon_j^2)]$. Let $\tilde{\rho}_{j_1,j_2} = \hat{\rho}_{j_1,j_2} I(|\hat{\rho}_{j_1,j_2}| > 2[log(p)/n]^{1/2})$. $nVar(\hat{\rho}_{j_1,j_2})$ is estimated by $\hat{\kappa}[1-\tilde{\rho}_{j_1,j_2}^2]^2$, where $\hat{\kappa} = \frac{n}{3p}\sum_{j=1}^p \frac{\sum_{i=1}^n \hat{\epsilon}_{i,j}^4}{(\sum_{i=1}^n \hat{\epsilon}_{i,j}^2)^2}$ Then adaptive

thresholding estimator for partial correlation matrix is proposed as

$$\hat{\rho}^{(t)}_{j_1,j_2}(\tau) = \hat{\rho}_{j_1,j_2} I[|\hat{\rho}_{j_1,j_2}| > \tau(1-\tilde{\rho}^2_{j_1,j_2})\{\hat{\kappa} log(p)/n\}^{1/2}].$$

Particularly, adaptive thresholding estimator for c-level graph has form of

$$\hat{E}_c = [(j_1,j_2) : |\hat{\rho}_{j_1,j_2}| > c + \tau(1-\tilde{\rho}^2_{j_1,j_2})\{\hat{\kappa} log(p)/n\}^{1/2}]$$

where c is used to construct different hypothesis tests. For example, if c is set to be 0.25, then c-level PCG is actually conducting hypothesis test with null hypothesis $H_0 :\ \rho <= 0.25$. When c is set to be 0, the corresponding hypothesis test is the common one testing if ρ is significant different from zero or not. Since other methods can only infer $H_0 : \rho = 0$, we set c to be 0 as well.

To choose the threshold parameter τ, the authors controlled the False Discovery Proportion at a desired level, where FDP is the number of false positives over the number of discovery. False Discovery Rate, FDR, is the expectation of FDP. Let $\#\{A\}$ denote the size of a set A and $\bar{A}$ denote the complementary set of A and $FDP_c(\tau)$ can be written as $\frac{\#\{\bar{E}_c(\tau)\}FPR_c(\tau)}{max[1,\#\{\hat{E}_c(\tau)\}]}$ ($FPR_c(\tau) = \frac{\text{FP}}{N}$). The authors show that the numerator $\#\{\bar{E}_c(\tau)\}FPR_c(\tau)$ is bounded by a function of τ, denoted as $B(\tau)$. Detailed explanation can be found in Qiu & Zhou (2018). Then FDP is purely based on known quantities and unknown τ. For a sequence of candidate τ in (0,2] and to control FDR at α we could choose $\tau_{FDP} = inf\{\tau \in (0,2] : \frac{B(\tau)}{max[1,\#\{\hat{E}_c(\tau)\}]} <= \alpha\}$.

2.3 Data Simulation

To assess the performance of c-level PCG and Shrunk MLE, we need to know the true structure of a given network. However, for real dataset, it is almost impossible to learn about real interaction relationships. So, evaluation of different methods involves generating in-silico networks whose structures are known. Following the simulation setting used in Bernal et al. (2019) Schäfer & Strimmer (2005), we first used functions in an R package to simulate networks with desired proportion of real edges and then simulate multi-normal data based on simulated networks. The package is called GeneNet Schaefer et al. (2009), a popular package for inferring shrinkage based Gaussian Graphical Models. Starting from

an empty p by p matrix, a pre-specified proportion of off diagonal elements are selected and assigned with a random number from uniform distribution between -1 and 1; then diagonal element is equal to the column-sum plus a small constant (e.g. 0.0001) to make the matrix positive definite; the matrix is standardized so that all the diagonal entries equal 1 to obtain the simulated true partial correlation matrix Schäfer & Strimmer (2004). However simulating partial correlation matrix in this way has its limitations. When p becomes large, simulated partial correlations are very small.

We simulated the true partial correlation matrix in a different manner. Starting with the p by p empty precision matrix Ω, $\omega_{ij} = \omega_{ji}$ for all $i \neq j$ are equal 0 with probability $1-\epsilon_0$ and equal to a random number from $U(-1, 1)$ with probability ϵ_0, where $\epsilon_0 = 0.1 or 0.5$. All diagonal elements are set to be 1.75 to make the matrix positive definite. The the corresponding partial correlation matrix is calculated by $\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} I(i \neq j) + I(i = j)$. Besides, since it is validated to assume the biological network is very sparse Barabasi & Oltvai (2004) and modules of genes exist in the biological network, we simulated networks by setting the partial correlation matrix as block diagonal matrix to account for sparse and modular features. To simulate desired block-diagonal partial correlation structure, we started with covariance matrix Σ directly. We set the covariance matrix to contain 4 by 4 compound symetry block matrix B along the diagonal of Σ whose off-diagonal element is a random number from U(0.3,0.6). We set p equal to 100 or 200 with n equal to 60 or 80 (4 cases).

We plan to to evaluate performance of different methods by ROC curve and nominal FDR versus empirical FDR plot. In detail, we simulated the true network structure and then simulated 50 random datasets with respect to the truth. Different methods, including Schäfer & Strimmer (2005) (denoted as ENF), Bernal et al. (2019) (denoted as MLE) and Qiu & Zhou (2018) (denoted as cLevel), were applied on each random data. We ranked estimated partial correlations from the most significant to the least significant by the p-values calculated by ENF and MLE and by test statistics ($t = \frac{\text{Point Estimate - c}}{\text{Standard Error}}$) calculated by cLevel. For top k edges (k varies from 1 to $\frac{p\times(p-1)}{2}$ with step 1), we checked the true positives and false positives for every method and calcuated false positive rate and true positiive rate, where $FPR = \frac{FP}{FP+TN} = \frac{FP}{N}$ and $TPR = \frac{TP}{TP+FN} = \frac{TP}{P}$. For 50 replicates,

we collected the largest TRR at a given FPR level and averaged them. ROC curve was then plotted based on the FPR and averaged TPR. Besides, for each method, we conducted inference with controlling FDR at α level and computed the empirical FDR, where α varies from 0.001 to 0.2 with step 0.001 and FDR is equal to $\frac{FP}{\text{Discoveries}}$. Inference of ENF and MLE are based on Benjamini-Hochberg corrected p-values (q-values). 50 empirical FDR were also averaged and plotted against nomial FDR.

2.4 Escherichia Coli Microarray Data

The dataset we used in this study consists of E.coli microarray gene-expression from Schmidt-Heck et al. (2004). To assess the complex functional association between molecules during stress response, the expression of 4289 protein coding genes of the E.coli was measured by microarrays. 102 genes were selected as they differentially expressed after normalization. The dataset can be found from the *GeneNet* package. It consists of 9 observations (9 time points at 0, 8, 15, 22, 45, 68, 90, 150 and 180 min) and 102 genes.

2.5 Software

All implementations are performed with R version 3.5.2.

3 Results

3.1 Simulation Study

To consider different ways of selecting λ, we included two results based on c-level partial correlation graph. cLevel represents the inference results of c-level PCG based on theoretically selected λ, which is $\sqrt{2 \times (log(p)/n)}$. And we assumed that each gene is connected to other d genes on average. So we fixed the degree of freedom of each lasso regression to be d. d is equal to 5 in our study, which is arbitrarily selected. As a result, cLevel_new represents the results based on specifying the number of non-zero coefficients in node-wide lasso regressions. Another way to select λ is using cross validation. However, it can takes long time when number of variables is large. We will not discuss it here.

Figure 1 and Figure 2 are corresponding to true networks of random structure with 1% true edges. As we can see in Figure 1, different methods perform very similarly under such setting. And the line for ENF and the line for MLE are actually perfectly overlapped, which means they have the exactly the same performance in terms of the rank of significance. This is because MLE inherits the shrunk estimates of ENF. It improved the p-values but the order of significant estimates remains the same. Also, under this setting, cLevel also performs not bad as MLE, no matter how we select tunning parameter λ. Figure 2 tells us c-level PCG with theoretically selected λ can almost perfectly control FDR at a desired level. But c-level PCG with setting degree of freedom to be 5 fails to control FDR. Unlike c-level PCG, performance of ENF and MLE depends on n and p. When $n = 80$ and $p = 100$, both ENF and MLE controlled the FDR. But in other cases, they both failed to control FDR.

Figure 3 and Figure 4, corresponding to true networks of random structure with 5% true edges, show similar results. In terms of the rank, all these methods perform not badly. But in terms of controlling FDR, only c-level PCG with theoretically selected λ can consistently control FDR at a desired level. Interestly, when p is equal to 200 and sample size is increase, performance of ENF and MLE are actually not improved, which is oppsite to what we expect.

Figure 5 and Figure 6 are based on true network with block-diagonal structure. ROC curves show that the rank from ENF/MLE is the best, which means shrunk partial correlation can almost capture all true edges. c-level PCG with theoretically selected λ also performs well. Most top ranked infered edges are edges that really exist. But c-level PCG with setting degree of freedom to be 5 is relatively sub-optimal. Figure 6, nominal FDR versus empirical FDR plot, shows very interesting results. Two c-level PCG infered results consistently controlled FDR but ENF failed to control FDR and MLE are too conservative.

Further more, we summarized performance of four methods in terms of false positives and true positives at FDR 0.05 level. For every penal in Figure 7, point at right bottom corner means that method leads to fewer false positives and more true positives relatively. As we can see, clevel performs consistently better than others. It also supports that using theoretically selected λ is better than fixing degree of freedom.

3.2 Application on Real Dataset

Since the sample size of this real dataset is too small, we can not use theoretically selected λ to calculate coefficients of lasso regression (λ is so large that all coefficients are penalized to 0 and corresponding estimated partial correlations are actually sample correlation). We tried two ways to select tunning parameter λ. The first one is cross validation (Figure 8) and the second one is setting degree of freedom to be 7 (Figure 9). We actually tried several different numbers (3, 5, 7) and found that when degree of freedom is 7, infered network is more informative. Following results are network inference of E.coli dataset by Shrunk MLE and c-level PCG at 0.05 FDR level. Also a network constructed by Shrunk MLE without FDR controlling is also included (Figure 11).

Clearly we can find that after multiple testing correction, Shrunk MLE with Benjamini-Hochberg correction is too conservative to find many siginificant edges (Figure 10). While compared to the network constructed by thresholding unadjusted p-values by Shrunk MLE less than or equal to 0.05 (Figure 11), c-level PCG constructs the network with different structures. Shrunk MLE directly penalizes the sample covariance matrix instead of the precision matrix, which might makes the results very different from what c-level PCG suggests. More biological interpretation and justifications are needed.

4 Discussion

In this study, we applied a novel method, c-level partial correlation graph, on gene expression data. We conducted several simulation studies and confirmed that this new method outperforms other methods in terms of high statistical power and FDR controlling. We also applied c-level PCG on real dataset and found inconsistent results from previous studies which needs biological justification. We also found that c-level PCG can find more real edges in a gene network and control false positive rate at the same time.

However, the selection of tunning parameter λ still needs further discussion. Theoretically selected λ might not be meaningful when the ratio of p and n is too large and selecting λ by cross validation can be too time consuming. Also we simulated data from multivariate normal distribution but typical gene expression data does not follow normal distribution.

More simulation studies are needed.

In the future, we plan to try to borrow information from the dataset to determine λ and extend the partial correlation to a more generalized statistical dependency measure. More simulation study can be done by changing the way of generating random data and increasing number of variables. Last but not least, genomic experiments are usually conducted under different conditions with several repilcates under each condition. So a more scientific way to analyze gene expression data would be using hierarchical model. In such way, gene network inference might be more precise and informative.

References

Albert, R. & Barabási, A.-L. (2002), 'Statistical mechanics of complex networks', *Reviews of modern physics* **74**(1), 47.

Ansorge, W. J. (2009), 'Next-generation dna sequencing techniques', *New biotechnology* **25**(4), 195–203.

Barabasi, A.-L. & Oltvai, Z. N. (2004), 'Network biology: understanding the cell's functional organization', *Nature reviews genetics* **5**(2), 101.

Benjamini, Y. & Hochberg, Y. (1995), '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.

Bernal, V., Bischoff, R., Guryev, V., Grzegorczyk, M. & Horvatovich, P. (2019), 'Exact hypothesis testing for shrinkage based gaussian graphical models', *Bioinformatics* .

Boccaletti, S. (2010), *Handbook on biological networks*, Vol. 10, World Scientific.

De Las Rivas, J. & Fontanillo, C. (2010), 'Protein–protein interactions essentials: key concepts to building and analyzing interactome networks', *PLoS computational biology* **6**(6), e1000807.

Efron, B. (2004), 'Large-scale simultaneous hypothesis testing: the choice of a null hypothesis', *Journal of the American Statistical Association* **99**(465), 96–104.

Efron, B. (2005), 'Local false discovery rates'.

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

Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. (1999), 'From molecular to modular cell biology', *Nature* **402**(6761supp), C47.

Heller, M. J. (2002), 'Dna microarray technology: devices, systems, and applications', *Annual review of biomedical engineering* **4**(1), 129–153.

Ledoit, O. & Wolf, M. (2004), 'A well-conditioned estimator for large-dimensional covariance matrices', *Journal of multivariate analysis* **88**(2), 365–411.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. & Califano, A. (2006), Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *in* 'BMC bioinformatics', Vol. 7, BioMed Central, p. S7.

Meyer, P. E., Kontos, K., Lafitte, F. & Bontempi, G. (2007), 'Information-theoretic inference of large transcriptional regulatory networks', *EURASIP journal on bioinformatics and systems biology* **2007**, 8–8.

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

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

Schaefer, J., Opgen-Rhein, R. & Strimmer, K. (2009), 'Genenet: modeling and inferring gene networks', *URL http://CRAN. R-project. org/package= GeneNet. R package version* **1**(4).

Schäfer, J. & Strimmer, K. (2004), 'An empirical bayes approach to inferring large-scale gene association networks', *Bioinformatics* **21**(6), 754–764.

Schäfer, J. & Strimmer, K. (2005), 'A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics', *Statistical applications in genetics and molecular biology* **4**(1).

Schmidt-Heck, W., Guthke, R., Toepfer, S., Reischer, H., Duerrschmid, K. & Bayer, K. (2004), Reverse engineering of the stress response during expression of a recombinant protein, *in* 'Proceedings of the EUNITE symposium', pp. 10–12.

Soranzo, N., Bianconi, G. & Altafini, C. (2007), 'Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data', *Bioinformatics* **23**(13), 1640–1647.

Tieri, P., Farina, L., Petti, M., Astolfi, L., Paci, P. & Castiglione, F. (2019), 'Network inference and reconstruction in bioinformatics'.

Wang, Y. R. & Huang, H. (2014), 'Review on statistical methods for gene network reconstruction using expression data', *Journal of theoretical biology* **362**, 53–61.

Yu, D., Kim, M., Xiao, G. & Hwang, T. H. (2013), 'Review of biological network data and its applications', *Genomics & informatics* **11**(4), 200.

Zhang, B. & Horvath, S. (2005), 'A general framework for weighted gene co-expression network analysis', *Statistical applications in genetics and molecular biology* **4**(1).