Title
Problems in Network Modeling: Estimating Edges and Community Detection

Author
Wang, Ying Xiang

Publication Date
2015

Peer reviewed|Thesis/dissertation

Problems in Network Modeling: Estimating Edges and Community Detection

by

Ying Xiang Wang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Haiyan Huang, Co-chair
Professor Peter J. Bickel, Co-chair
Professor Cari Kaufman
Professor Lewis J. Feldman

Spring 2015

Problems in Network Modeling: Estimating Edges and Community Detection

Abstract

Problems in Network Modeling: Estimating Edges and Community Detection

by

Ying Xiang Wang

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Haiyan Huang, Co-chair

Professor Peter J. Bickel, Co-chair

Networks pervade many disciplines of science as a way of analyzing complex systems with interacting components. The problem of network modeling is often two-fold. First, the relationships between pairs of nodes, if not directly observed, have to be estimated from data. Based on the estimated (or given) network topology, various statistical and computational tools can then be applied to extract interesting patterns such as the presence of communities. In this thesis we explore studies related to both parts of the problem.

We first discuss two studies in the context of gene regulatory networks, where the goal is to infer gene interactions using expression data. With the advent of high-throughput technologies making large-scale gene expression data readily available, developing appropriate computational tools to infer gene interactions has been a major challenge in systems biology. The two studies differ in their considerations of how genes behave across the given samples. The first method applies to the case of large heterogenous samples, where the patterns of gene association may change or only exist in a subset of all the samples. We propose two new gene coexpression statistics based on counting local patterns of gene expression ranks to take into account the potentially diverse nature of gene interactions. In particular, one of our statistics is designed for time-course data with local dependence structures, such as time series coupled over a subregion of the time domain. We provide asymptotic analysis of their distributions and power, and evaluate their performance against a wide range of existing coexpression measures on simulated and real data. Our new statistics are fast to compute, robust against outliers, and show comparable if not better general performance.

In comparison, the second study goes beyond pairwise gene relationships to higher level group interactions, but requiring similar gene behaviors across all the samples. We introduce a new method for estimating group interactions using sparse canonical correlation analysis (SCCA) coupled with repeated random partition and subsampling of the gene expression dataset. By considering different subsets of genes and ways of grouping them, our interaction measure can be viewed as an aggregated estimate of partial correlations of different orders. Our approach is unique in evaluating conditional dependencies when the correct

dependent sets are unknown or only partially known. As a result, a gene network can be constructed using the interaction measures as edge weights and gene functional groups can be inferred as tightly connected communities from the network. Comparisons with several popular approaches using simulated and real data show our procedure improves both the statistical significance and biological interpretability of the results. In addition to achieving considerably lower false positive rates, our procedure shows better performance in detecting important biological pathways.

Moving onto general networks, we then discuss model selection for the stochastic block model (SBM), which is a popular tool for community detection. We consider an approach based on the log likelihood ratio statistic and analyze its asymptotic properties under model misspecification. We show the limiting distribution of the statistic in the case of underfitting is normal and obtain its convergence rate in the case of overfitting. These conclusions remain valid in the regime where the average degree grows at a polylog rate. The results enable us to derive the correct order of the penalty term for model complexity and arrive at a likelihood-based model selection criterion that is asymptotically consistent. In practice, the likelihood function can be estimated by more computationally efficient variational methods, allowing the criterion to be applied to moderately large networks.

To my parents, for their unconditional love and support

Contents

List of Figures

List of Tables

Acknowledgments

I am deeply grateful to my advisors, Haiyan Huang and Peter Bickel, for their constant encouragement, endless patience and support, and for being profoundly inspiring role models I will always look up to in my future career. I am also extremely grateful to Lewis Feldman, Keni Jiang, and Michael Waterman, with whom I had the honor and great pleasure to collaborate on several projects. Many thanks to Lewis Feldman and Cari Kaufman for kindly being on my qualifying and dissertation committees and providing constructive feedback. I would also like to thank Terry Speed for always generously sharing his time and expertise, and for further inspiring and cementing my interest in the fields of statistics and computational biology.

I am also grateful to Yun Song, who had always generously and warmly welcomed me to his research group, and to Anand Bhaskar and Matthias Steinrücken – being able to work with them on two projects was a great experience and learning opportunity.

Having chosen Berkeley for graduate school is one of the best decisions I have made in my life, and this experience was made special by many people in this wonderful department. I am forever indebted to Howard D'Abrera, whose amazing cooking and generous supplies of Tim Tams rescued me on countless occasions and made me feel I was not far from home – thank you for teaching me the importance of giving. I am also thankful to La Shana Porlaris and Mary Melinn for their help with administration, which made life much easier for me. I feel extremely fortunate to have the friendship of Riddhipratim Basu, Daisy Huang, Christine Kuang, Hongwei Li, Sujayam Saha, Funan Shi, Siqi Wu, and Angie Zhu. I will always look back to the great times we shared in the department corridors and outside school with fondness and nostalgia.

My initial apprehension about living in a country far away from home vanished after I became housemates with Jasmine Nirody and Miklos Racz, although the amazing dynamics we shared should largely be credited to their warm-heartedness and tolerance. I would like to thank them for putting up with my occasional eccentricity, for never failing to lend a listening ear and riding with me through the difficult times for all these years. I cannot overstate how much I cherish the times we spent together, and how much I learnt from them about the importance of hard work, perseverance, and always staying curious and driven about science.

To my special friends from Sydney, Maryanne Large, Anna Wang, and Xin Zhang, I feel very grateful that the physical distance did not set us apart emotionally. I would like to thank them for providing the support network that I knew I could always trust and rely on, and all the healing vibes that kept me afloat on stormy days.

Lastly, I want to thank my parents for giving me unlimited freedom to pursue my career path. This dissertation is dedicated to their unconditional love and support.

Chapter 1

Introduction

Network modeling has attracted increasing research attention in the past few decades as the amount of data on complex systems accumulates at an unprecedented rate. Many complex systems in science and nature consist of interacting individual components which can be represented as nodes with connecting edges in a network. Network modeling has found numerous applications in the studies of friendship networks in sociology, the Internet and the World Wide Web in information technology, predator-prey interactions in ecology, and protein-protein interactions, gene regulatory networks and many other biochemical networks in biology. More examples of networks and their basic properties can be found in [68].

Close examinations of these networks can reveal important knowledge about the nature of the individual nodes, their connections, and most crucially, interesting connection patterns such as communities, where groups of nodes exhibit high internal connectivity. Community structure provides a natural division of the network into subunits with certain traits. In social networks, they often arise based on people's common interests and geographic location. The World Wide Web forms communities or hubs based on the content of the web pages. In gene networks, communities correspond to genes with related functional groupings, many of which may operate in the same biological pathway.

The study of community structure relies on knowing the relationships between pairs of nodes. In some cases, such as social networks, this type of data is directly observable using online social network services. Also, the availability of fast web search algorithms means probing web links is relatively inexpensive. However, in many other cases, especially biochemical networks, direct observation of protein or gene relationships by experimental approaches is cost-prohibitive given that the typical size of the networks is in the tens of thousands. The amount of proteins present or the expression level of genes, on the other hand, is easier to measure and can be regarded as sets of covariates associated with the nodes. Therefore the analysis of this type of networks is essentially a two-fold problem. First, a network of relations needs to be learned using the covariates associated with each node. Only then one can apply community detection methods to identify tightly knit clusters. In this thesis, we investigate this composite problem. We first present two studies in Chapters 2 and 3 in the context of gene regulatory networks, where the goal is to infer gene interactions

using expression data. Moving onto general networks, we then discuss the problem of model selection for the stochastic block model (SBM), which is a popular tool for community detection, in Chapter 4. The notations used will be chapter-specific.

Inferring gene networks using expression data

Rapid advances in genomic technology have generated an enormous wealth of data on which mathematical and statistical tools can be applied to infer qualitative and quantitative relationships between DNA, RNA, proteins and other cellular molecules. Such a process of reconstructing biochemical networks using genomic data, also known as network inference or reverse engineering, has helped to elucidate the nature of complex biological processes and disease mechanisms in a variety of organisms, bringing us one step closer to understanding how genetic blueprints combined with non-genetic, environmental factors influence the characteristics of a living system.

At a high level, genes, proteins or other metabolites can be conceptualized as nodes and their interactions as edges in a graph. In metabolic networks, reactions are represented as directed edges pointing from reaction substrates to products. While metabolic networks tend to focus on proteins or protein-complexes functioning as enzymes, general protein-protein interaction (PPI) networks are undirected graphs where an edge indicates physical binding between two proteins.

At a more fundamental level, understanding biological processes requires understanding gene regulatory networks since all proteins are encoded by genes. In such a network, transcription factors (TFs), RNA and other small molecules act as regulators to activate or repress the expression levels of genes. Thus gene interactions can occur in the form of direct physical binding of proteins (TFs) to their target sequences, but in a broader sense also include indirect interactions when the expression of a gene influence the expressions of others with regulations caused by one or more intermediaries. Although experimental evidence can be gathered to search for and verify gene interactions, computational tools utilizing gene expression data offer a much more time and cost efficient way to reconstruct these networks. In the past decade, good quality gene expression data have been made readily available in the form of microarray or RNA-seq data.

Chapters 2 and 3 focus on the problem of reconstructing gene networks using expression data. Gene expression data has the form of a matrix with p genes and their expression levels measured under n experimental conditions. A typical feature of this type of data is their high dimensionality with p much larger than n, posing many estimation and computation challenges. Depending on the nature of the data given, we discuss two methods with different modeling paradigms. The method in Chapter 2 applies to the case with reasonably large but heterogeneous samples, where the patterns of gene association may change or only exist in a subset of all the samples. We propose new gene coexpression statistics based on counting local patterns of gene expression ranks to take into account the potentially diverse nature of gene interactions. This Chapter is based on [100]. In Chapter 3, we discuss a method

that goes beyond pairwise gene relationships to higher level group interactions, but requires similar gene behaviors across all the samples. This Chapter is adapted from [101]. We provide theoretical analyses for both methods and compare them to other popular approaches using simulated and real data, thus demonstrate they lead to better general performance and capture important biological features that are missed by the other methods.

Community structures in networks

One important feature of many of the aforementioned networks is the presence of communities, where a number of nodes form a densely connected subgraph and have sparser connections with the rest of the network. Community or module detection is of great importance in analyzing detailed architectures of networks. For example, in biochemical networks identifying groups of molecules performing a specific cellular function is a key issue in systems biology. In a PPI network, highly connected nodes are often proteins interacting as part of a complex or other functional modules, which are fundamental in cellular functions and have been shown to play an important role in disease pathologies ([59, 86]). In gene networks, genes modules are likely to have related biological functions or participate in the same biological pathway.

Given communities can be considered as tightly connected subgraphs, numerous heuristic algorithms have been proposed for community detection. For example, in the context of gene networks, [10] developed CAST, an algorithm that constructs one cluster at a time by adding and dropping genes iteratively according to a similarity measure. CLICK, proposed by [84], assumes edge weights between all pairs of genes follow a mixture normal distribution with a higher mean for within-cluster edges. The parameters and cluster memberships are estimated using EM methods. This idea that nodes have different connectivities depending on their cluster memberships is adopted in a more general random graph model known as the stochastic block model (SBM).

The SBM, proposed by [38] in social science, is one of the simplest random graph models incorporating community structures. It assigns each node a latent discrete block variable and the connectivity levels between nodes are determined by their block memberships. SBMs and the concept of communities as modularities have been applied in ([35, 40]) to recover community structures in biochemical networks. However, this model sometimes oversimplifies the structures of real networks, as we also demonstrate in an application in Chapter 3. Other variants have been proposed, including the degree-corrected SBM [45] relaxing the within-block degree homogeneity constraint and overlapping SBM [1] allowing a node to be in multiple blocks. These models have been applied to model real networks in social science and biology [1, 45].

One important advantage of a generative model is that it allows us to study the problem of community modeling from a theoretical perspective. Much research effort has been devoted to the problems of estimating the latent block memberships and model parameters of a SBM, including modularity [69] and likelihood maximization [11, 4], variational methods [40, 54],

spectral clustering [79, 30], belief propagation [24] to name but a few. The asymptotic properties of some of these methods have also been studied [11, 79, 14, 12]. However, these methods require knowing (or knowing at least a suitable range for) K, the number of blocks, a priori. Less attention has been paid to the problem of selecting K.

For general networks this corresponds to the issue of determining the number of communities, which remains a challenging open problem. Recursive approaches have been adopted to extract [114] or partition [13] one community sequentially, while using optimization strategies or hypothesis testing to decide whether the process should be stopped at one stage. A more general sequential test for comparing a fitted SBM against alternative models with finer structures is proposed in [56]. Conceptually these approaches are more appealing for networks with a hierarchical structure. In other cases, it would be more desirable to be able to compare different community numbers directly. A few likelihood-based model selection criteria have been proposed [40, 54, 81]. From an information-theoretic perspective, [72] proposed a criterion based on minimum length description. These approaches circumvent the difficulty of analyzing the likelihood directly by using variational approximations or assuming the node labels are fixed and using plug-in estimates obtained from other inference algorithms. Furthermore, the asymptotic studies of these criteria examining their large-sample performance remain incomplete. Empirically, a network cross-validation method has been investigated in [17]. Chapter 4 of thesis focuses on developing a likelihood-based model selection criterion for SBM. Our approach is based on the log likelihood ratio statistic and its asymptotic properties under model misspecification. We show the criterion is asymptotically consistent and valid for networks with average degrees growing at a polylog rate. This Chapter is based on the manuscript [99].

Chapter 2

Gene coexpression measures in large heterogenous samples using count statistics

2.1 Overview

As mentioned in Chapter 1, a major challenge in systems biology is the understanding of the intricate interactions and functional relationships between genes and their regulation targets. As advances in high-throughput technologies lead to the generation of enormous amounts of genomic data, the last decade has witnessed a rapidly increasing effort to develop computational tools to reconstruct gene relationships based on a wide range of "omic" data available, in particular transcriptomic or expression data. Coexpression methods are one of the earliest tools used for such a purpose. Appropriate statistical means assessing the dependence between the expression levels (e.g. linear coexpression) of two genes provide a way to identify their potential functional interaction. Coexpression analysis has been routinely used for functional gene annotation ([117, 32]) and more importantly as a measure of edge weights for reconstructing gene networks ([89, 112, 9, 111, 21]).

The problem of finding gene coexpression is closely related to that of detecting bivariate association between two vectors. Since the work by [27], the Pearson correlation has been adopted as the most widely used coexpression measure ([89, 107, 92]) for its straightforward conceptual interpretation and computational efficiency. However, it is also known that the Pearson correlation is unsuitable for capturing nonlinear relationships and susceptible to high false discovery rates. Another class of coexpression methods is based on mutual information (MI) ([91, 22, 9, 64]), which measures general statistical dependence rather than a specific type of bivariate association. The computation of MI involves discretization of the data and tuning parameters, and obtaining p-values requires computationally intensive permutation tests. The practical benefits and shortcomings of MI compared to correlation based methods are still under investigation ([91, 22, 88]). More comparisons of different coexpression

measures and the coexpression networks constructed can be found in [51] and [3].

In the broader statistical literature, other methods available for quantifying bivariate associations include the Renyi correlation ([76]) measuring the correlation between two variables after suitable transformations; various regression based techniques ([88]); Hoeffding's D ([37]) and distance covariance (dCov, [93]) for general statistical dependence. These methods are not widely adopted in genomic applications yet. More recently, [77] proposed the maximal information coefficient (MIC) as an extension of MI, but was shown to have inferior power to dCov ([85]) and MI ([50]) in various simulated scenarios.

Most of the methods mentioned so far, perhaps with the exception of MIC, do not specifically target dependence relationships that can be local in nature and often assume the data are random samples from a common distribution in the theoretical analysis. However, real gene interactions may change as the intrinsic cellular state varies or only exist under a specific cellular condition. Furthermore, with data integration now being a routine approach to combat the curse of dimensionality, samples from different experimental conditions or tissue types are likely to prescribe different gene relationships and thus create more complex situations for detecting gene interactions. For instance, a protein that positively regulates expression in one context may act as a repressor in another (e.g., MECP2 ([15])), or a gene may participate in either neural development or hematopoiesis depending on tissue type (e.g. EBF1 ([66, 113])). One possible approach to discern local gene interactions is biclustering ([19, 62]), which simultaneously clusters the genes and samples in an expression matrix. However, most biclustering techniques are restricted to detecting simple subclasses of linear associations. On the algorithm side, the optimizations of most criteria for measuring the quality of given biclusters can only be achieved locally and their global behaviors are hard to characterize. Most algorithms also involve a number of tuning parameters with little guidance on how to choose them.

Motivated by these observations, we propose two new coexpression measures based on matching patterns of local expression ranks using count statistics. Our robust statistics specifically take into account the local nature of gene associations while being general enough to detect other common types of dependence relationships. In particular, one of our statistics is designed for time-course data with local dependence structures, such as time series that are coupled over a subregion of the time domain. This is a unique feature compared to other popular coexpression measures. The statistics are fast to compute and we provide theoretical analysis of their asymptotic properties. We demonstrate their applicability via comparisons to a comprehensive list of existing methods on simulated and real data. Our new methods show better precision, and have the important ability to detect subtle gene relationships that are easily missed by other methods.

2.2 Definitions and asymptotic properties

For a heterogeneous set of samples with potentially changing gene interactions, we can define a general coexpression measure by aggregating the interactions across all subsamples of size

$k \leq n$. For genes x and y with expression levels from n samples $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, we consider

$$W = \sum_{1 \leq i_1 < \cdots < i_k \leq n} F(x_{i_1}, \ldots, x_{i_k}; y_{i_1}, \ldots, y_{i_k}), \tag{2.1}$$

where $F(\cdot;\cdot)$ is an interaction measure on local expression profiles $(x_{i_1}, \ldots, x_{i_k})$ and $(y_{i_1}, \ldots, y_{i_k})$ from a set of k samples. We choose $F(\cdot;\cdot)$ to be an indicator function comparing the rank patterns of the subsequences $(x_{i_1}, \ldots, x_{i_k})$ and $(y_{i_1}, \ldots, y_{i_k})$. Depending on the nature of the expression data studied, we define two corresponding count statistics.

1. When dealing with time-course data, it is sensible to preserve the order of the samples and consider only interactions within contiguous subsequences. Thus our first measure W_1 is defined as

$$\begin{aligned} W_1 = \sum_{i=1}^{n-k+1} & \mathbb{I}\left(\phi(x_i, \ldots, x_{i+k-1}) = \phi(y_i, \ldots, y_{i+k-1})\right) \\ & + \mathbb{I}\left(\phi(x_i, \ldots, x_{i+k-1}) = \phi(-y_i, \ldots, -y_{i+k-1})\right), \end{aligned} \tag{2.2}$$

   where $\mathbb{I}(\cdot)$ is an indicator function and $\phi(\cdot)$ is the rank function. That is, $\phi(\cdot)$ returns the indices of elements in the subvector after they have been sorted in an increasing order within themselves. W_1 counts the number of contiguous subsequences of length k with matching and reverse rank patterns, indicating positive and negative associations respectively.

2. When the order of the samples is not particularly meaningful (e.g. non time series data), we consider a more general count that includes all subsequences of length k. Define W_2 as

$$\begin{aligned} W_2 = \sum_{1 \leq i_1 < \cdots < i_k \leq n} & \mathbb{I}(\phi(x_{i_1}, \ldots, x_{i_k}) = \phi(y_{i_1}, \ldots, y_{i_k})) \\ & + \mathbb{I}(\phi(x_{i_1}, \ldots, x_{i_k}) = \phi(-y_{i_1}, \ldots, -y_{i_k})) \end{aligned} \tag{2.3}$$

   It is easy to see that W_2 is equal to the number of increasing (and decreasing) subsequences of length k in a suitably permuted sequence. Suppose σ is a permutation that sorts the elements of $\mathbf{y}$ in an increasing order. Let $\mathbf{z} = \sigma(\mathbf{x})$ be that permutation applied to $\mathbf{x}$, W_2 can be rewritten as

$$W_2 = \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbb{I}(z_{i_1} < \cdots < z_{i_k}) + \mathbb{I}(z_{i_1} > \cdots > z_{i_k}). \tag{2.4}$$

Here we give a simple example of the two counts. Suppose $\mathbf{x} = (1, 3, 4, 2, 5)$, $\mathbf{y} = (1, 4, 5, 2, 3)$, and we are interested in computing W_1 and W_2 counts for $k = 3$. For W_1, there

are three possible positions to start a contiguous subsequence of length 3, and only the ones starting at position 1 and 2 have the same ranks in $\mathbf{x}$ and $\mathbf{y}$. There are no pairs of contiguous subsequences of length 3 with reverse ranks. Hence $W_1 = 2$. To compute W_2, first sort $\mathbf{y}$ in an ascending order with permutation σ. Applying σ to $\mathbf{x}$ we have $\sigma(\mathbf{x}) = (1, 2, 5, 3, 4)$. The total number of increasing subsequences of length 3 in $\sigma(\mathbf{x})$ is 5, and there are no decreasing subsequences of length 3. Hence $W_2 = 5$.

Both counts are symmetric with respect to $\mathbf{x}$ and $\mathbf{y}$ and efficient to compute. As shown in the following lemma, counting W_1 has a running time of $O(k(\log k)n)$, while counting W_2 takes $O(kn \log n)$ time using dynamic programing and binary indexed trees.

Lemma 2.2.1. *Computing W_1 and W_2 takes $O(k(\log k)n)$ and $O(kn \log n)$ time respectively.*

Proof. Computing W_1 involves ranking and comparing the elements of vectors of length k $O(n)$ number of times, thus the running time is $O(k(\log k)n)$.

W_2 counts the total number of subsequences of length k with matching or reverse rank patterns. For any pair of subsequences with matching rank patterns, permuting the two subsequences simultaneously to sort one of them in an increasing order will also sort the other one in an increasing order. Using this observation, let σ be the permutation that sorts $\mathbf{y}$ in an increasing order and $\mathbf{z} = \sigma(\mathbf{x})$ be that permutation applied to $\mathbf{x}$. Then W_2 is the number of increasing (and decreasing) subsequences of length k in $\mathbf{z}$. To compute W_2, it suffices to consider counting the increasing subsequences. One obvious solution is dynamic programing. Let `dp[i,l]` be the number of increasing subsequences of length `l` ending at position `i`, then the matrix `dp[i,l]` can be updated as follows.

```
Initialize dp[i,l] = 0; dp[i,1] = 1
for i = 2 to n
     for j = 1 to i-1
     if z[i] > z[j]
          for l = 2 to k
          dp[i,l] += dp[j,l-1]
```

The final answer is obtained by summing `dp[i,k]` over i. It is easy to see this has a running time of $O(kn^2)$. Note that in the second loop the only entries involved in the update are `z[j]` whose ranks are smaller than that of `z[i]`. Therefore by first ranking the elements in $\mathbf{z}$, a binary indexed tree structure can be implemented to perform the sum and update efficiently, reducing the running time to $O(kn \log n)$ ([29]). □

Asymptotic distributions

We can derive the asymptotic distributions of W_1 and W_2 for different regimes of k assuming the following.

(I). The two sequences $\mathbf{x}$ and $\mathbf{y}$ are independent and have no ties.

(II). At least one of $\mathbf{x}$ and $\mathbf{y}$ has an exchangeable distribution.

Note that (II) implies the ranks of the expression vector with an exchangeable distribution is a random permutation of $\{1, 2, \ldots, n\}$.

The Stein and Chen-Stein approximations ([90, 18]) give us the following two asymptotic regimes for W_1.

Theorem 2.2.2. *For $n \to \infty$, $k \geq 3$ and $k/(\log n)^\alpha \to 0$ for some $\alpha < 1$,*

$$T_1 := \frac{W_1 - \mu_{1,n}}{\sigma_{1,n}} \xrightarrow{D} N(0,1), \tag{2.5}$$

where $\mu_{1,n} = 2(n-k+1)/k!$, $\sigma_{1,n}^2 = Var(W_1)$.

For $n \to \infty$, $\log n/k = O(1)$,

$$d_{TV}(W_1, Z) \to 0, \tag{2.6}$$

where $Z \sim Poisson(\mu_{1,n})$ and d_{TV} is the total variation distance.

Throughout the rest of the Chapter C and C_i denote positive constants which may be different at each appearance. Without loss of generality assume $\mathbf{x}$ satisfies the assumption that it has an exchangeable distribution. Then the ranks of any subsequence of $\mathbf{x}$ can be treated as a random permutation. Denote

$$\begin{aligned} \mathbb{I}_i^+ &= \mathbb{I}(\phi(x_i, \ldots, x_{i+k-1}) = \phi(y_i, \ldots, y_{i+k-1})), \\ \mathbb{I}_i^- &= \mathbb{I}(\phi(x_i, \ldots, x_{i+k-1}) = \phi(-y_i, \ldots, -y_{i+k-1})), \\ \mathbb{I}_i &= \mathbb{I}_i^+ + \mathbb{I}_i^-. \end{aligned} \tag{2.7}$$

We have

$$\begin{aligned} &\mathbb{E}(\mathbb{I}_i^+) \\ =&\sum_{\mathbf{w}} \mathbb{P}(\phi(x_i, \ldots, x_{i+k-1}) = \mathbf{w} \mid \phi(y_i, \ldots, y_{i+k-1}) = \mathbf{w}) \mathbb{P}(\phi(y_i, \ldots, y_{i+k-1}) = \mathbf{w}) \\ =&\frac{1}{k!} \sum_{\mathbf{w}} \mathbb{P}(\phi(y_i, \ldots, y_{i+k-1}) = \mathbf{w}) \\ =&\frac{1}{k!} \end{aligned} \tag{2.8}$$

by the independence assumption and the fact that there is only one way to arrange a list of numbers in a given order. Clearly also $\mathbb{E}(\mathbb{I}_i^-) = 1/k!$. In the next lemma, we characterize the behavior of the cross terms $\mathbb{E}(\mathbb{I}_i^+ \mathbb{I}_j^+)$.

Lemma 2.2.3. *1. When $|j - i| \geq k$, $\mathbb{I}_i^+$ and $\mathbb{I}_j^+$ are independent. So are $(\mathbb{I}_i^+, \mathbb{I}_j^-)$ and $(\mathbb{I}_i^-, \mathbb{I}_j^-)$.*

2. *When* $|j-i| = k-l$ *with* $1 \le l \le k-1$,

$$\frac{1}{(2k-l)!} \le \mathbb{E}(\mathbb{I}_i^+ \mathbb{I}_j^+) \le \frac{\binom{2k-2l}{k-l}}{(2k-l)!}. \tag{2.9}$$

The same conclusions hold for $(\mathbb{I}_i^-, \mathbb{I}_j^-)$, $1 \le |j-i| < k$, *and* $(\mathbb{I}_i^+, \mathbb{I}_j^-)$, $(\mathbb{I}_i^-, \mathbb{I}_j^+)$, $|i-j| = k-1$.

3. $\mathbb{E}(\mathbb{I}_i^+ \mathbb{I}_j^-) = \mathbb{E}(\mathbb{I}_i^- \mathbb{I}_j^+) = 0$ *for* $1 \le |i-j| < k-1$.

Proof. Note that conditioning on the sequence $\mathbf{y}$,

$$\begin{aligned}\mathbb{E}(\mathbb{I}_i^+ \mathbb{I}_j^+) = \sum_{\mathbf{w},\mathbf{v}} &\mathbb{P}(\phi(x_i, \ldots, x_{i+k-1}) = \mathbf{w}, \phi(x_j, \ldots, x_{j+k-1}) = \mathbf{v}) \\ &\times \mathbb{P}(\phi(y_i, \ldots, y_{i+k-1}) = \mathbf{w}, \phi(y_j, \ldots, y_{j+k-1}) = \mathbf{v}).\end{aligned} \tag{2.10}$$

For $|j-i| \ge k$, the subsequences $(x_i, \ldots, x_{i+k-1})$ and $(x_j, \ldots, x_{j+k-1})$ do not overlap. Thus their local rank patterns are independent, each having probability $1/k!$ for a given order.

$$\begin{aligned}\mathbb{E}(\mathbb{I}_i^+ \mathbb{I}_j^+) &= \left(\frac{1}{k!}\right)^2 \sum_{\mathbf{w},\mathbf{v}} P(\phi(y_i, \ldots, y_{i+k-1}) = \mathbf{w}, \phi(y_j, \ldots, y_{j+k-1}) = \mathbf{v}) \\ &= \left(\frac{1}{k!}\right)^2 = \mathbb{E}(\mathbb{I}_i^+)\mathbb{E}(\mathbb{I}_j^+).\end{aligned} \tag{2.11}$$

For $j - i = k - l < k$ (assuming WLOG $j > i$), $(x_i, \ldots, x_{i+k-1})$ and $(x_j, \ldots, x_{j+k-1})$ form a contiguous subsequence $x_i, \ldots, x_j, \ldots, x_{j+k-1}$. Suppose $\phi(x_i, \ldots, x_{j+k-1}) = (u_1, \ldots, u_{2k-l})$, then

$$\begin{aligned}\phi(u_1, \ldots, u_k) &= (w_1, \ldots, w_k), \\ \phi(u_{k-l+1}, \ldots, u_{2k-l}) &= (v_1, \ldots, v_k), \\ \phi(u_{k-l+1}, \ldots, u_k) &= \phi(w_{k-l+1}, \ldots, w_k) = \phi(v_1, \ldots, v_l) \\ &:= (o_1, \ldots, o_l), \qquad \text{say.}\end{aligned} \tag{2.12}$$

Focusing on the overlapping part (u_{k-l+s}) for $1 \le s \le l$, the numbers of elements smaller than u_{k-l+s} in the subsequences $(u_1, \ldots, u_k)$, $(u_{k-l+1}, \ldots, u_{2k-l})$ and $(u_{k-l+1}, \ldots, u_k)$ are $w_{k-l+s} - 1$, $v_s - 1$, and $o_s - 1$, respectively. Given the overall rank u_{k-l+s} in the sequence $(u_1, \ldots, u_{k-l+1}, \ldots, u_k, \ldots, u_{2k-l})$, we have

$$u_{k-l+s} - 1 = (w_{k-l+s} - 1) + (v_s - 1) - (o_s - 1), \tag{2.13}$$

since the elements in the overlapping part are counted twice. In other words, the overlapping part (u_{k-l+s}) for $1 \le s \le l$ is fixed, and there are at most $\binom{2k-2l}{k-l}$ ways of arranging the rest $2k-2l$ numbers. Thus we arrive at the upper bound in (2.9). The lower bound is trivial. The same arguments hold for $(\mathbb{I}_i^-, \mathbb{I}_j^-)$, $1 \le |j-i| < k$, and $(\mathbb{I}_i^+, \mathbb{I}_j^-)$, $(\mathbb{I}_i^-, \mathbb{I}_j^+)$, $|i-j| = k-1$.

Lastly, for $1 \le |i-j| < k-1$, $\mathbb{E}(\mathbb{I}_i^+ \mathbb{I}_j^-) = \mathbb{E}(\mathbb{I}_i^+ \mathbb{I}_j^-) = 0$ since no such arrangements of the elements are possible. □

Let N_i denote the dependency neighborhood of $\mathbb{I}_i$, the next lemma tries to bound a key quantity in the variance calculation.

Lemma 2.2.4. *For all $k \geq 3$,*

$$4(n-2k+2)\left(\sum_{l=2}^{k-1}\frac{1}{(2k-l)!}+\frac{2}{(2k-1)!}\right) \leq \sum_{i=1}^{n-k+1}\sum_{j\in N_i\setminus\{i\}}\mathbb{E}(\mathbb{I}_i\mathbb{I}_j) \leq \frac{C(n-k+1)}{(k+1)!} \quad (2.14)$$

for some $C > 0$.

Proof. First note that

$$\begin{aligned}\sum_{i=1}^{n-k+1}\sum_{j\in N_i\setminus\{i\}}\mathbb{E}(\mathbb{I}_i\mathbb{I}_j) &= 2\sum_{i=1}^{n-k+1}\sum_{j\in N_i\setminus\{i\}}\mathbb{E}(\mathbb{I}_i^+\mathbb{I}_j^+) + 2\sum_{i=1}^{n-k+1}\sum_{|j-i|=k-1}\mathbb{E}(\mathbb{I}_i^+\mathbb{I}_j^-)\\ &\leq 8(n-k+1)\sum_{l=1}^{k-1}\gamma_l \end{aligned} \quad (2.15)$$

by (2.9), where

$$\gamma_l = \frac{\binom{2k-2l}{k-l}}{(2k-l)!}. \quad (2.16)$$

It remains to bound $\sum_{l=1}^{k-1}\gamma_l$. Taking the ratio of successive terms,

$$\begin{aligned} r_l = \frac{\gamma_{l+1}}{\gamma_l} &= \frac{\binom{2k-2l-2}{k-l-1}}{(2k-l-1)!}\cdot\frac{(2k-l)!}{\binom{2k-2l}{k-l}}\\ &= \frac{(k-l)^2(2k-l)}{(2k-2l)(2k-2l-1)}\\ &= \frac{(k-l)(2k-l)}{2(2k-2l-1)}, \quad l=1,\ldots,k-2. \end{aligned} \quad (2.17)$$

For all $k \geq 3$, there exists positive constant C_1 and C_2 (independent of k) such that

$$C_1 k \leq r_l \leq C_2 k, \quad l = 1,\ldots,k-2. \quad (2.18)$$

Therefore $\sum_{l=1}^{k-1}\gamma_l$ is upper bounded by

$$\begin{aligned}\sum_{l=1}^{k-1}\gamma_l &\leq \gamma_{k-1}\sum_{l=0}^{k-2}\left(\frac{1}{C_1 k}\right)^l\\ &= \gamma_{k-1}\cdot\frac{1-\left(\frac{1}{C_1 k}\right)^{k-1}}{1-\frac{1}{C_1 k}}\\ &\leq \frac{C}{(k+1)!}\end{aligned} \quad (2.19)$$

for some $C > 0$. Equations (2.19) and (2.15) give the required upper bound.

For the lower bound, it is easy to see

$$\begin{aligned}\sum_{i=1}^{n-k+1}\sum_{j\in N_i\setminus\{i\}}\mathbb{E}(\mathbb{I}_i\mathbb{I}_j) &= 2\sum_{i=1}^{n-k+1}\sum_{j\in N_i\setminus\{i\}}\mathbb{E}(\mathbb{I}_i^+\mathbb{I}_j^+) + 2\sum_{i=1}^{n-k+1}\sum_{|j-i|=k-1}\mathbb{E}(\mathbb{I}_i^+\mathbb{I}_j^-)\\ &\geq 2\left[2(n-k+1)-2(k-1)\right]\left(\sum_{l=1}^{k-1}\frac{1}{(2k-l)!}+\frac{1}{(2k-1)!}\right)\\ &\geq 4(n-2k+2)\left(\sum_{l=2}^{k-1}\frac{1}{(2k-l)!}+\frac{2}{(2k-1)!}\right) \qquad (2.20)\end{aligned}$$

by the lower bound in (2.9). □

With the above bounds we can now prove Theorem 2.2.2.

Proof of Theorem 2.2.2. In order to use Stein's method for normal approximation, we first give a lower bound of the variance. Note that

$$\begin{aligned}\sigma_{1,n}^2 &= \sum_{i=1}^{n-k+1}\sum_{j\in N_i}(\mathbb{E}(\mathbb{I}_i\mathbb{I}_j)-(\mathbb{E}\mathbb{I}_i)(\mathbb{E}\mathbb{I}_j))\\ &= \sum_{i=1}^{n-k+1}\sum_{j\in N_i\setminus\{i\}}\mathbb{E}(\mathbb{I}_i\mathbb{I}_j)+\frac{2(n-k+1)}{k!}-\sum_{i=1}^{n-k+1}\sum_{j\in N_i}\mathbb{E}(\mathbb{I}_i)\mathbb{E}(\mathbb{I}_j)\\ &\geq 4(n-2k+2)\left(\sum_{l=2}^{k-1}\frac{1}{(2k-l)!}+\frac{2}{(2k-1)!}\right)+\frac{2(n-k+1)}{k!}-\frac{4(n-k+1)(2k-1)}{(k!)^2}. \qquad (2.21)\end{aligned}$$

by (2.14). For k such that $k/n \to 0$, when n is sufficiently large, $\sigma_{1,n}^2$ is lower bounded by the dominating terms

$$\begin{aligned}\sigma_{1,n}^2 &\geq C_1\left(4n\left(\sum_{l=2}^{k-1}\frac{1}{(2k-l)!}+\frac{2}{(2k-1)!}\right)+\frac{2n}{k!}-\frac{4n(2k-1)}{(k!)^2}\right)\\ &= \frac{2C_1n}{k!}\left(2\left(\frac{1}{k+1}+\frac{1}{(k+2)(k+1)}+\cdots+\frac{2}{(2k-1)\cdots(k+1)}\right)+1-\frac{2(2k-1)}{k!}\right)\\ &\geq \frac{C_2n}{k!} \qquad (2.22)\end{aligned}$$

for some $C_1, C_2 > 0$ and all $k \geq 3$. One version of Stein's method gives the following error bound for normal approximation ([80]),

$$d_W(T_1,Y) \leq \frac{D^2}{\sigma_{1,n}^3}\sum_{i=1}^{n-k+1}\mathbb{E}|\mathbb{I}_i-2/k!|^3+\frac{\sqrt{26}D^{3/2}}{\sqrt{\pi}\sigma_{1,n}^2}\sqrt{\sum_{i=1}^{n-k+1}\mathbb{E}|\mathbb{I}_i-2/k!|^4} \qquad (2.23)$$

where d_W is the Wasserstein metric (minimal L_1-metric), $Y \sim N(0,1)$ and $D = \max_i N_i = 2k-1$. This can be further bounded by

$$\begin{aligned} & C_1 \cdot \frac{D^2\mu_{1,n}}{\sigma_{1,n}^3} + C_2 \cdot \frac{D^{3/2}\mu_{1,n}^{1/2}}{\sigma_{1,n}^2} \\ \leq & C \cdot \frac{D^2\mu_{1,n}}{\sigma_{1,n}^3} \\ \leq & C \cdot \frac{k^2\sqrt{k!}}{\sqrt{n}} \to 0 \end{aligned} \tag{2.24}$$

using (2.22) for $k/(\log n)^\alpha \to 0$, $\alpha < 1$.

The Chen-Stein method yields the following error bound for Poisson approximation,

$$\begin{aligned} d_{TV}(W_1, Z) &\leq \min\{1, \mu_{1,n}^{-1}\} \left(\sum_{i=1}^{n-k+1} \sum_{j \in N_i} \mathbb{E}(\mathbb{I}_i)\mathbb{E}(\mathbb{I}_j) + \sum_{i=1}^{n-k+1} \sum_{j \in N_i \setminus \{i\}} \mathbb{E}(\mathbb{I}_i \mathbb{I}_j) \right) \\ &\leq \sum_{i=1}^{n-k+1} \sum_{j \in N_i} \mathbb{E}(\mathbb{I}_i)\mathbb{E}(\mathbb{I}_j) + \sum_{i=1}^{n-k+1} \sum_{j \in N_i \setminus \{i\}} \mathbb{E}(\mathbb{I}_i \mathbb{I}_j) \\ &\leq \frac{4(n-k+1)(2k-1)}{(k!)^2} + \frac{C(n-k+1)}{(k+1)!} \\ &\leq \frac{C(n-k+1)}{(k+1)!} \end{aligned} \tag{2.25}$$

for some $C > 0$ and k sufficiently large. For k growing fast enough such that $\mu_{1,n} = O(1)$, the above bound goes to 0. In particular, using Stirling's approximation one can show in the regime $\log n/k = O(1)$ this condition is satisfied. □

While the properties and asymptotic distribution of the longest increasing subsequence in a random permutation have been much studied and the statistic itself has been used in a number of applications ([60, 7, 2, 6]), not so much attention has been paid to increasing subsequences of length k. Here we use the results in [74] and the Stein approximation to derive a central limit theorem for W_2 for k growing sufficiently slowly. The key of the proof lies in obtaining a good upper and lower bound on the variance of W_2.

Theorem 2.2.5. *For $n \to \infty$, $k \geq 3$ and $k/(\log n)^\alpha \to 0$ for some $\alpha < 1$,*

$$T_2 := \frac{W_2 - \mu_{2,n}}{\sigma_{2,n}} \xrightarrow{D} N(0,1), \tag{2.26}$$

where $\mu_{2,n} = 2\binom{n}{k}/k!$ and $\sigma_{2,n}^2 = Var(W_2)$.

Again assuming $\mathbf{x}$ has an exchangeable distribution, the permuted sequence $\sigma(\mathbf{x})$ also has an exchangeable distribution, and its ranks can be treated as a random permutation. For notational simplicity, take $\mathbf{z}$ as a random permutation of $\{1, \ldots, n\}$. For integers $\{i_1, \ldots, i_k\}$ satisfying $1 \leq i_1 < \cdots < i_k \leq n$, define indicators $\mathbb{I}^+_{i_1,\ldots,i_k}(\mathbf{z})$ such that

$$\mathbb{I}^+_{i_1,\ldots,i_k}(\mathbf{z}) = \begin{cases} 1 & (i_1, \ldots, i_k) \text{ is a subsequence of } \mathbf{z}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.27}$$

Similarly define

$$\mathbb{I}^-_{i_1,\ldots,i_k}(\mathbf{z}) = \begin{cases} 1 & (i_k, \ldots, i_1) \text{ is a subsequence of } \mathbf{z}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.28}$$

Then W_2 can be written as the sum of

$$W_2 = \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbb{I}_{i_1,\ldots,i_k}(\mathbf{z}), \tag{2.29}$$

where

$$\mathbb{I}_{i_1,\ldots,i_k}(\mathbf{z}) = \mathbb{I}^+_{i_1,\ldots,i_k}(\mathbf{z}) + \mathbb{I}^-_{i_1,\ldots,i_k}(\mathbf{z}). \tag{2.30}$$

It is easy to see that if $\{i_1, \ldots, i_k\} \cap \{j_1, \ldots, j_k\} = \emptyset$, $\mathbb{I}_{i_1,\ldots,i_k}(\mathbf{z})$ and $\mathbb{I}_{j_1,\ldots,j_k}(\mathbf{z})$ are independent. The variance of W_2 becomes

$$\begin{aligned} &\mathrm{Var}(W_2) \\ =& \sum_{\{i_1,\ldots,i_k\} \cap \{j_1,\ldots,j_k\} \neq \emptyset} \{\mathbb{E}(\mathbb{I}_{i_1,\ldots,i_k}(\mathbf{z})\mathbb{I}_{j_1,\ldots,j_k}(\mathbf{z})) - \mathbb{E}(\mathbb{I}_{i_1,\ldots,i_k}(\mathbf{z}))\mathbb{E}(\mathbb{I}_{j_1,\ldots,j_k}(\mathbf{z}))\} \\ =& 2 \sum_{\{i_1,\ldots,i_k\} \cap \{j_1,\ldots,j_k\} \neq \emptyset} \mathbb{E}(\mathbb{I}^+_{i_1,\ldots,i_k}(\mathbf{z})\mathbb{I}^+_{j_1,\ldots,j_k}(\mathbf{z})) \\ &+ 2 \sum_{\{i_1,\ldots,i_k\} \cap \{j_1,\ldots,j_k\} \neq \emptyset} \mathbb{E}(\mathbb{I}^+_{i_1,\ldots,i_k}(\mathbf{z})\mathbb{I}^-_{j_1,\ldots,j_k}(\mathbf{z})) - \frac{4D\binom{n}{k}}{(k!)^2}, \end{aligned} \tag{2.31}$$

since $\mathbb{E}(\mathbb{I}^+(z_{i_1}, \ldots, z_{i_k})) = 1/k!$. Here D is the size of the dependency neighborhood and equals $\binom{n}{k} - \binom{n-k}{k}$.

The sum of the first cross terms can be written as (Proposition 2 in [74])

$$\begin{aligned} &\sum_{\{i_1,\ldots,i_k\} \cap \{j_1,\ldots,j_k\} \neq \emptyset} \mathbb{E}(\mathbb{I}^+_{i_1,\ldots,i_k}(\mathbf{z})\mathbb{I}^+_{j_1,\ldots,j_k}(\mathbf{z})) \\ =& \sum_{j=1}^{k} \binom{n}{2k-j} \frac{1}{(2k-j)!} A(k-j, j), \end{aligned} \tag{2.32}$$

where

$$A(N,j) = \sum_{\substack{\sum_{r=0}^{j} l_r = N \\ \sum_{r=0}^{j} m_r = N}} \prod_{r=0}^{j} \left(\frac{(l_r + m_r)!}{l_r! m_r!} \right)^2. \tag{2.33}$$

We will be using the following fact about the constants $A(N,j)$ from Lemma 3 in [74].

Fact 2.2.6. *For sufficiently large k, there exists $C > 0$ such that*

$$A(k-1,1) \geq C k^{1/2} \binom{2k-2}{k-1}^2. \tag{2.34}$$

It is easy to see for all $k \geq 2$, $A(k-1,1) > \binom{2k-2}{k-1}^2$.

The sum of the second cross terms reduces to

$$\sum_{|\{i_1,\ldots,i_k\} \cap \{j_1,\ldots,j_k\}|=1} \mathbb{E}(\mathbb{I}^+_{i_1,\ldots,i_k}(\mathbf{z}) \mathbb{I}^-_{j_1,\ldots,j_k}(\mathbf{z})),$$

since when the size of the intersection is greater than one, it is impossible to find a permutation $\mathbf{z}$ satisfying both conditions specified by the indicators. Using arguments similar to the proof of Proposition 2 in [74], we can show

$$\sum_{|\{i_1,\ldots,i_k\} \cap \{j_1,\ldots,j_k\}|=1} \mathbb{E}(\mathbb{I}^+_{i_1,\ldots,i_k}(\mathbf{z}) \mathbb{I}^-_{j_1,\ldots,j_k}(\mathbf{z})) = \binom{n}{2k-1} \frac{1}{(2k-1)!} B(k), \tag{2.35}$$

where

$$B(k) = \sum_{\substack{l_0 + l_1 = k-1 \\ m_0 + m_1 = k-1}} \binom{l_0 + m_0}{l_0} \binom{l_1 + m_1}{l_1} \binom{l_0 + m_1}{l_0} \binom{l_1 + m_0}{l_1} \tag{2.36}$$

Now we can obtain a lower bound on the variance and use the Stein method to prove Theorem 2.2.5.

Proof of Theorem 2.2.5. From equations (2.31), (2.32) and (2.35), we have

$$\begin{aligned} \frac{\sigma^2_{2,n}}{\mu^2_{2,n}} &\geq \frac{\binom{n}{2k-1}(k!)^2}{2\binom{n}{k}^2 (2k-1)!} \left(A(k-1,1) + B(k) \right) - \frac{D}{\binom{n}{k}} \\ &\geq \frac{k^2}{2n} \cdot \left(1 - \frac{k-1}{n-k+1} \right)^{k-1} \binom{2k-1}{k-1}^{-2} \left(A(k-1,1) + B(k) \right) - \frac{D}{\binom{n}{k}}. \end{aligned} \tag{2.37}$$

For $k \to \infty$ and $k = o(n^{1/2})$, it is easy to check $D/\binom{n}{k} = O(k^2/n)$. Applying Fact 2.2.6,

$$\begin{aligned}\frac{\sigma_{2,n}^2}{\mu_{2,n}^2} &\geq C \cdot \frac{k^{5/2}}{2n}\left(1 - \frac{k-1}{n-k+1}\right)^{k-1}\left[\frac{(2k-2)\cdots k}{(2k-1)\cdots(k+1)}\right]^2 + O(k^2/n) \\ &= C \cdot \frac{k^{5/2}}{2n}(1+O(k^2/n))\left(\frac{k}{2k-1}\right)^2 + O(k^2/n) \\ &\geq C \cdot \frac{k^{5/2}}{n} \end{aligned} \tag{2.38}$$

for some $C > 0$ and sufficiently large k and n. Applying the bound from the Stein method as in equation (2.23), we have

$$\begin{aligned} d_W(T_2, Y) &\leq C_1 \cdot \frac{D^2\mu_{2,n}}{\sigma_{2,n}^3} + C_2 \cdot \frac{D^{3/2}\mu_{2,n}^{1/2}}{\sigma_{2,n}^2} \\ &\leq C_1 \cdot \frac{k^{1/4}(k!)^2}{n^{1/2}} + C_2 \cdot \frac{k^{1/2}(k!)^{3/2}}{n^{1/2}} \to 0 \end{aligned} \tag{2.39}$$

for $k/(\log n)^\alpha \to 0$.

For k fixed, $D/\binom{n}{k} \leq k^2/(n-k+1) + o(1/n)$. (2.37) becomes

$$\begin{aligned}\frac{\sigma_{2,n}^2}{\mu_{2,n}^2} &\geq \frac{k^2}{2n}(1+O(1/n))\binom{2k-1}{k-1}^{-2}(A(k-1,1)+B(k)) - \frac{k^2}{n-k+1} + o(1/n) \\ &= \left\{\frac{1}{2}\left(A(k-1,1)+B(k)\right)\binom{2k-1}{k-1}^{-2} - 1\right\}\frac{k^2}{n} + o(1/n) \\ &:= C(k) \cdot \frac{k^2}{n} + o(1/n), \quad \text{say.} \end{aligned} \tag{2.40}$$

When $k = 3$, we can check that $C(3) > 0$ and thus $\sigma_{2,n}^2/\mu_{2,n}^2 \geq C/n$. For other fixed k, the same order lower bound holds. Applying (2.23),

$$d_W(T_2, Y) \leq O(n^{-1/2}) \to 0. \tag{2.41}$$

□

Asymptotic power

Next we analyze the power of T_1 and T_2 under specific alternative distributions. The first scenario we consider is related to time-course data, where the temporal order of $\mathbf{x}$ and $\mathbf{y}$ are preserved in subsequence analysis.

Theorem 2.2.7. *Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be two time series with n observations, m of which are perfectly coupled in the sense that $\phi(x_i, \ldots, x_{i+m-1}) = \phi(y_i, \ldots, y_{i+m-1})$. As $n \to \infty$, $m \to \infty$,*

1. *T_1 goes to infinity in the following regimes:*
   - *For fixed k, if $m \sim a_1 n$, $a_1 > 2/k!$, then $T_1 = \Omega(\sqrt{n})$.*
   - *For $k \to \infty$ and $k/(\log n)^\alpha \to 0$, $\alpha < 1$,*
     - *if $m \geq a_2 \cdot \frac{n}{k!}$, $a_2 > 2$, then $T_1 = \Omega(\sqrt{n/k!})$;*
     - *if $m \sim a_3 n$, $a_3 \in (0, 1]$, then $T_1 = \Omega(\sqrt{nk!})$.*
2. *T_2 goes to infinity in the following regimes:*
   - *For fixed k, if $m \sim b_1 n$, $b_1^k > 2/k!$, then $T_2 = \Omega(\sqrt{n})$.*
   - *For $k \to \infty$ and $k/(\log n)^\alpha \to 0$, $\alpha < 1$,*
     - *if $m \geq \frac{en}{k}$, then $T_2 = \Omega(\sqrt{n/k^{3/2}})$;*
     - *if $m \sim b_2 n$, $b_2 \in (0, 1]$, then $T_2 = \Omega(b_2^k k! \sqrt{n/k^{5/2}})$.*

Here $\Omega(\cdot)$ denotes asymptotic lower bound.

Remark 2.2.8. *In the regimes above, using T_1 and T_2 as statistics both lead to rejection of the null hypothesis with probability one. We also observe that for both T_1 and T_2, large k leads to better power in the sense that i) the statistics have a better convergence rate when m grows as a fraction of n; ii) a smaller lower bound on m can be achieved, consequently tolerating more noise in the data, while maintaining the power of the tests going to 1. Comparing T_1 and T_2, T_1 has better power in the sense that i) T_1 has a better rate of convergence in the regime of $k \to \infty$ and m growing as a fraction of n; ii) T_1 allows for a smaller lower bound on m in the other regimes while maintaining the power going to 1.*

The next scenario we consider is when $\mathbf{x}$ and $\mathbf{y}$ follow a perfect functional relationship with d strictly monotonic pieces. This is a reasonable subclass of general functional relationships to study since most smooth function can be approximated by piecewise strictly monotonic functions. In this case the order of the data does not have to be preserved, making T_1 a less suitable statistic than T_2. We only analyze the power of T_2.

Theorem 2.2.9. *$\mathbf{y} = f(\mathbf{x})$ for $\mathbf{x} \overset{iid}{\sim} Unif(0, 1)$, f can be decomposed into a fixed number of d strictly monotonic pieces which have lengths $\ell_1, \ldots, \ell_d$ when projected on to the x-axis. As $n \to \infty$,*

- *For fixed k, if $d^{k-1} < k!/2$, then $\mathbb{P}(T_2 \geq C\sqrt{n}) \to 1$;*
- *For $k \to \infty$ and $k/(\log n)^\alpha \to 0$, $\alpha < 1$, then $\mathbb{P}(T_2 \geq C\sqrt{n/k^{5/2}} k!/d^{k-1}) \to 1$*

for some constant $C > 0$.

Remark 2.2.10. *In the regimes above, the power of the statistic T_2 approaches 1. Large k and smaller d lead to better convergence rates and thus better power. Having fewer monotonic pieces implies there are more uninterrupted counts in each piece contributing to W_2.*

To prove Theorems 2.2.7 and 2.2.9, first we prove a lemma upper bounding the variances of T_1 and T_2.

Lemma 2.2.11.
- $\sigma_{1,n}^2 = O(n)$ *for fixed* k; $\sigma_{1,n}^2 = O(n/k!)$ *for* $k \to \infty$ *and* $k/(\log n)^\alpha \to 0$.
- $\sigma_{2,n}^2 = O(n^{2k-1})$ *for fixed* k; $\sigma_{2,n}^2 = O(\mu_{2,n}^2 k^{5/2}/n)$ *for* $k \to \infty$ *and* $k/(\log n)^\alpha \to 0$.

Proof. By the upper bound in (2.14),

$$\begin{aligned}\sigma_{1,n}^2 &= \sum_{i=1}^{n-k+1} \sum_{j \in N_i \setminus \{i\}} \mathbb{E}(\mathbb{I}_i \mathbb{I}_j) + \frac{2(n-k+1)}{k!} - \sum_{i=1}^{n-k+1} \sum_{j \in N_i} \mathbb{E}(\mathbb{I}_i)\mathbb{E}(\mathbb{I}_j) \\ &\leq \frac{C(n-k+1)}{(k+1)!} + \frac{2(n-k+1)}{k!} \\ &= \begin{cases} O(n) & \text{for fixed } k; \\ O(n/k!) & \text{for } k \to \infty,\ k/(\log n)^\alpha \to 0. \end{cases} \end{aligned} \tag{2.42}$$

To bound $\sigma_{2,n}^2$, first note that $B(k) \leq A(k-1,1)$ for all $k \geq 2$. This holds because for every pair of (l_0, l_1) and (m_0, m_1) such that $l_0 + l_1 = k-1$ and $m_0 + m_1 = k-1$, we have

$$\binom{l_0+m_0}{l_0}^2 \binom{l_1+m_1}{l_1}^2 + \binom{l_0+m_1}{l_0}^2 \binom{l_1+m_0}{l_1}^2 \geq 2\binom{l_0+m_0}{l_0}\binom{l_1+m_1}{l_1}\binom{l_0+m_1}{l_0}\binom{l_1+m_0}{l_1}.$$

By equations (2.31), (2.32) and (2.35),

$$\begin{aligned}\sigma_{2,n}^2 &= 2\sum_{j=1}^{k} \binom{n}{2k-j} \frac{1}{(2k-j)!} A(k-j,j) + 2\binom{n}{2k-1} \frac{1}{(2k-1)!} B(k) - \frac{4D\binom{n}{k}}{(k!)^2} \\ &\leq 4\sum_{j=1}^{k} \binom{n}{2k-j} \frac{1}{(2k-j)!} A(k-j,j) - \frac{4D\binom{n}{k}}{(k!)^2} \\ &= O\left(\frac{\mu_{2,n}^2 k^{5/2}}{n}\right) \end{aligned} \tag{2.43}$$

by Theorem 1 in [74]. The first part of the lemma holds since $\mu_{2,n} = O(n^k)$ for k fixed. □

Proof of Theorem 2.2.7. It is easy to see the count W_1 is bounded below by $m-k+1$. By the first part of Lemma 2.2.11,

$$\begin{aligned} T_1 &\geq \frac{m-k+1-\mu_{1,n}}{\sigma_{1,n}} \\ &\geq C\sqrt{n}\left(\frac{m}{n} - \frac{2}{k!}\right), \end{aligned} \tag{2.44}$$

for some $C > 0$, fixed k and m, n sufficiently large. In this case, m has to grow at the same rate as n, that is $m \sim a_1 n$ and $a_1 > 2/k!$. It follows then $T_1 = \Omega(\sqrt{n})$.

When $k \to \infty$ and $k/(\log n)^\alpha \to 0$, for n large enough,

$$\begin{aligned} T_1 &\geq C\sqrt{\frac{n}{k!}}\left(\frac{k!(m-k+1)}{n-k+1} - 2\right) \\ &\geq C\sqrt{\frac{n}{k!}}\left(\frac{a_2 n - k! \cdot k + k!}{n-k+1} - 2\right) \\ &= \Omega\left(\sqrt{\frac{n}{k!}}\right) \end{aligned} \tag{2.45}$$

for $m \geq a_2 n/k!$, $a_2 > 2$. If m grows at the rate of $a_3 n$, $a_3 \in (0, 1]$,

$$\begin{aligned} T_1 &\geq C\sqrt{nk!}\left(\frac{m-k+1}{n-k+1} - \frac{2}{k!}\right) \\ &\geq C\sqrt{nk!}a_3 = \Omega(\sqrt{nk!}). \end{aligned} \tag{2.46}$$

Similarly, the count W_2 is lower bounded by $\binom{m}{k}$, using the second part of Lemma 2.2.11, for fixed k,

$$\begin{aligned} T_2 &\geq C \cdot \frac{\binom{m}{k} - 2\binom{n}{k}/k!}{n^{k-1/2}} \\ &= C\sqrt{n}\left(\frac{m \cdots (m-k+1)}{n \cdots (n-k+1)} - \frac{2}{k!}\right) \\ &\geq C\sqrt{n}\left(\left(\frac{m}{n}\right)^k - \frac{2}{k!}\right) \end{aligned} \tag{2.47}$$

for sufficiently large m and n. m has to grow at the rate of $b_1 n$ for the lower bound to go to infinity, and $b_1^k > 2/k!$. We have $T_2 = \Omega(\sqrt{n})$.

When $k \to \infty$ and $k/(\log n)^\alpha \to 0$, again by Lemma 2.2.11,

$$\begin{aligned} T_2 &\geq C\sqrt{\frac{n}{k^{5/2}}}\left(k!\frac{m \cdots (m-k+1)}{n \cdots (n-k+1)} - 2\right) \\ &\geq C\sqrt{\frac{n}{k^{5/2}}}\left(k!\left(\frac{m}{n}\right)^k - 2\right) \\ &\geq C\sqrt{\frac{n}{k^{3/2}}}, \end{aligned} \tag{2.48}$$

for $m \geq en/k$. When $m \sim b_2 n$, $b_2 \in (0, 1]$,

$$\begin{aligned} T_2 &\geq C\sqrt{\frac{n}{k^{5/2}}}\left(k!\left(\frac{m}{n}\right)^k - 2\right) \\ &\geq Cb_2^k k!\sqrt{\frac{n}{k^{5/2}}}. \end{aligned} \tag{2.49}$$

□

Proof of Theorem 2.2.9. Let $n_1, \ldots, n_d$ denote the number of points in $(\mathbf{x}, \mathbf{y})$ falling on to each monotonic piece, then W_2 is lower bounded by $\sum_{t=1}^{d} \binom{n_t}{k}$. For fixed d and k, ,

$$\frac{\sum_{t=1}^{d} n_t \cdots (n_t - k + 1)}{n \cdots (n - k + 1)} \xrightarrow{P} \sum_{t=1}^{d} \ell_t^k \tag{2.50}$$

Since by Lemma 2.2.11,

$$T_2 \geq C\sqrt{n} \left(\frac{\sum_{t=1}^{d} n_t \cdots (n_t - k + 1)}{n \cdots (n - k + 1)} - \frac{2}{k!} \right) \tag{2.51}$$

for some $C > 0$, it follows

$$\mathbb{P}\left(T_2 \geq C\sqrt{n}(d^{-(k-1)} - 2/k!)\right) \to 1 \tag{2.52}$$

using Hölder's inequality and the fact $\sum_{t=1}^{d} \ell_t = 1$. Thus T_2 is lower bounded by $C\sqrt{n}$ with probability tending to 1 when $d^{k-1} < k!/2$.

When $k \to \infty$ and $k/(\log n)^\alpha \to 0$, it is easy to check

$$\frac{n_t \cdots (n_t - k + 1)}{n \cdots (n - k + 1)} \cdot \left(\frac{n}{n_t} \right)^k \xrightarrow{P} 1. \tag{2.53}$$

Also,

$$\begin{aligned} &\mathbb{P}\left(\left| \left(\frac{n_t}{n\ell_t} \right)^k - 1 \right| \geq \epsilon \right) \\ \leq &\mathbb{P}\left(\frac{n_t}{n\ell_t} - 1 \geq (1+\epsilon)^{1/k} - 1 \right) + \mathbb{P}\left(\frac{n_t}{n\ell_t} - 1 \leq (1-\epsilon)^{1/k} - 1 \right) \\ \leq &\exp(-2n\ell_t^2((1+\epsilon)^{1/k} - 1)^2) + \exp(-2n\ell_t^2((1-\epsilon)^{1/k} - 1)^2) \to 0 \end{aligned} \tag{2.54}$$

by Hoeffding's inequality. It follows then

$$\sum_{t=1}^{d} \frac{n_t \cdots (n_t - k + 1)}{n \cdots (n - k + 1)} \cdot \left(\sum_{t=1}^{d} \ell_t^k \right)^{-1} \xrightarrow{P} 1. \tag{2.55}$$

Now noting that

$$T_2 \geq C\sqrt{\frac{n}{k^{5/2}}} \left(k! \frac{\sum_{t=1}^{d} n_t \cdots (n_t - k + 1)}{n \cdots (n - k + 1)} - 2 \right), \tag{2.56}$$

we have

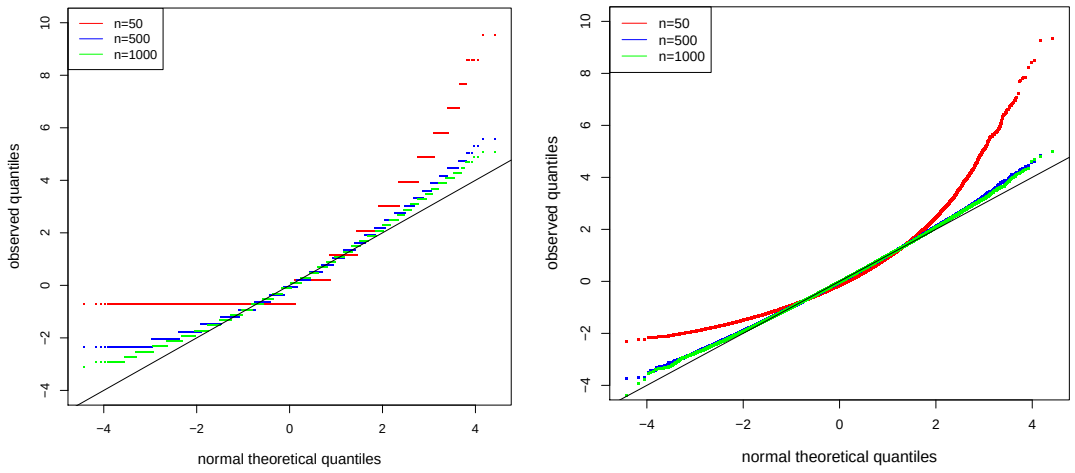
$$\mathbb{P}\left(T_2 \geq C \frac{k!}{d^{k-1}} \sqrt{\frac{n}{k^{5/2}}} \right) \to 1 \tag{2.57}$$

again by Hölder's inequality. □

2.3 Simulations

Numerical experiments supporting the asymptotics

We will first show the behaviors of the statistics conform to their derived asymptotics. Throughout the rest of the Chapter, the variances of W_1 and W_2 were estimated by Monte Carlo experiments. Figure 2.1 shows the convergence of the empirical quantiles of T_1 and T_2 toward the theoretical standard normal quantiles as n increases. Note due to the fact that T_1 can only take $n - k + 2$ possible values, it is easy to produce ties. To examine the asymptotic power of the two statistics under alternative distributions described previously, we generated data that i) were partially coupled time series with the length of dependence $m = n/10$; ii) followed an exact functional relationship with six monotonic pieces, and computed the average power at 5% significance level over 500 iterations. The results for different k and n are shown in Table 2.1. As predicted by the theoretical analysis, larger k results in better power and T_1 is more powerful than T_2 on the time-course data. In all the cases, as n increases the power tends to 1. The table also displays the average power for the corresponding null distributions of i) and ii) when the two data vectors are independent. The values are centered around 0.05 as expected.



(a) (b)

Figure 2.1: Empirical quantiles for the standardized counts (a) T_1 and (b) T_2 for $n = 50$, 500 and 1000, $k = 5$, from 10^5 simulated random permutations.

k/n	100	200	300	400	500
3	0.332	0.562	0.690	0.812	0.902
4	0.636	0.976	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1

(a) Power of T_1

k/n	100	200	300	400	500
3	0.070	0.050	0.056	0.038	0.036
4	0.060	0.044	0.030	0.054	0.052
5	0.038	0.074	0.064	0.058	0.060
6	0.052	0.042	0.064	0.046	0.076

(b) Power of T_1

k/n	100	200	300	400	500
3	0.302	0.504	0.658	0.796	0.848
4	0.340	0.568	0.726	0.844	0.908
5	0.360	0.650	0.798	0.892	0.952
6	0.392	0.734	0.882	0.924	0.982

(c) Power of T_2

k/n	100	200	300	400	500
3	0.048	0.054	0.076	0.052	0.050
4	0.068	0.060	0.044	0.072	0.040
5	0.042	0.056	0.040	0.070	0.038
6	0.050	0.068	0.056	0.046	0.046

(d) Power of T_2

k/n	100	200	300	400	500
3	0.516	0.784	0.884	0.952	0.972
4	0.710	0.952	0.996	1	1
5	0.866	0.992	1	1	1
6	0.946	1	1	1	1

(e) Power of T_2

k/n	100	200	300	400	500
3	0.058	0.048	0.062	0.044	0.052
4	0.058	0.062	0.044	0.060	0.052
5	0.078	0.040	0.060	0.062	0.030
6	0.070	0.060	0.064	0.062	0.054

(f) Power of T_2

Table 2.1: Power at 5% significance level for different choices of k and n when $\mathbf{x}$ and $\mathbf{y}$ are: (a), (c) two independent AR(1) time series (with coefficients 0.1 and -0.2 respectively) but $(x_1, \ldots, x_m) = (y_1, \ldots, y_m)$ with $m = n/10$; (e) $x_i \overset{iid}{\sim} Unif(0,1)$, $y_i = \cos(6\pi x_i)$. The right panel shows the power under the corresponding null distributions: (b), (d) $\mathbf{x}$ and $\mathbf{y}$ are two independent AR(1) time series (with coefficients 0.1 and -0.2 respectively); (f) $\mathbf{x}$ and $\mathbf{y}$ are iid $Unif(0,1)$.

Numerical examples examining statistical power

To investigate the power of our statistics in more realistic settings, we considered four types of bivariate relationships, all of which are illustrative of gene coexpression relationships likely to exist in an expression dataset. It is essential to include a linear type of relationship since pairwise gene relationships detected by current analyses are still predominantly linear. As an example of non-monotonic associations, we considered a quadratic relationship. The cross-shaped relationship may occur when two genes switch from activators to repressors across different tissue types or treatment conditions, or simply due to the changes in intrinsic cellular state ([58]). These relationships have also been used as illustrative scenarios in [77] and [50] in the context of general statistical dependence. An important additional example we considered here pertains to the case of genes with time-course data. We simulated two time series which were coupled over subregions of the time domain. The robustness of the

statistics was tested against outliers – a ubiquitous feature of biological data. Descriptions of the parameters used for each type of relationship are provided in Table 2.2.

	x_i	y_i	η_i
Linear	$x_i \overset{iid}{\sim} N(0,1)$	$y_i = x_i + 2e_i$	$\eta_i \overset{iid}{\sim} N(0,5)$
Quadratic	$x_i \overset{iid}{\sim} N(0,1)$	$y_i = x_i^2 + 2e_i$	$\eta_i \overset{iid}{\sim} N(0,5)$
Cross	$x_i \overset{iid}{\sim} N(0,1)$	$y_i = \begin{cases} \frac{1}{2} + x_i + e_i \text{ with probability } \frac{1}{2}, \\ \frac{3}{2} - x_i + e_i \text{ with probability } \frac{1}{2}. \end{cases}$	$\eta_i \overset{iid}{\sim} N(0,3)$
Partially coupled time series	$x_i \sim AR(1)$ with coefficient 0.1	$y_i = \begin{cases} x_i + e_i,\ i \in [1,30] \\ -x_i + e_i,\ i \in [101,120] \\ AR(1) \text{ with coefficient -0.2,} \\ \text{independent of } x_i, \text{ otherwise.} \end{cases}$	$\eta_i \overset{iid}{\sim} N(0,3)$

Table 2.2: Parameters for generating the four types of relationships. 2000 datasets were generated for every scenario with $i \in \{1, \ldots, 220\}$, $e_i \overset{iid}{\sim} N(0,1)$ for the first three relationships and $e_i \overset{iid}{\sim} N(0, 0.5)$ for the time-course relationship. Outliers were created by randomly choosing a fraction of the data and replacing e_i with η_i.

We compared the power of T_1 and T_2 to seven other popular measures of dependence (the Pearson, Spearman, Renyi correlations, Hoeffding's D, dCov, MI and MIC). We chose $k = 5$ for T_1 and T_2 guided by the log value of the sample size 220. The results from other values of k are provided in Figure 2.3. We note that the influence of k on the power of T_2 is negligible. While the choice of k has a bigger effect on the power of T_1 due to a smaller number of possible values for the counts, the conclusions we draw from qualitative comparisons with the other measures do not change.

The power values of various statistics computed under four types of dependence relationship are shown in Figure 2.2. Unsurprisingly, the Pearson and Spearman correlations can only detect the linear relationship, with the Pearson correlation being more sensitive to outliers. Across the first three types of dependence, T_2, Hoeffding's D, MI, dCov and Renyi are the only statistics maintaining reasonable power throughout. Of these statistics, Renyi and MI have the best performance on the quadratic relationship, but are underpowered on the linear relationship. For the linear scenario, we also computed a variant of T_1 and T_2 counting only the matching rank patterns (omitting the reverse patterns), which are denoted T_1^+ and T_2^+ in the plot. These unidirectional counts provide a way to significantly improve the power when the monotonicity of the relationship is known. In fact, T_2^+ demonstrates the best power while remaining robust to outliers. We note that T_2 has a higher power than all the other statistics on the cross relationship. T_1 does not perform well on the first three types of relationship as it is designed for data with a temporal order.

T_1 and T_2 are the only statistics showing significant power on the time-course data. Without respecting the order of the data points, the scatter plot shows no obvious association pattern, making it difficult for the other measures to detect the dependence structure. T_1 demonstrates a slightly better power than T_2.

We remark here that although other dependence relationships were tested in [77] and [50], most of these are rarely observed in real gene coexpression patterns. Such examples include sinusoidal, circular and checkerboard relationships. For the former two examples, we expect the power of T_2 to be affected by the noise level and the frequency of the sinusoidal wave. As discussed in Theorem 2.2.9, the power of T_2 is boosted by having uninterrupted counts from monotonic pieces of the association pattern. Since the checkerboard pattern is not piecewise monotonic, we do not expect T_2 to detect this type of relationships.

2.4 Real Data Examples

In this section we evaluate the performance of our new statistics on two gene expression datasets: i) the classic yeast gene expression dataset from [89] and ii) a collection of microarray data for *Arabidopsis* tissues downloaded from NCBI GEO.

Yeast cell cycle data

The yeast expression data was accessed from http://genome-www.stanford.edu/cellcycle/ and contains the expression levels of 6178 genes from four reasonably long time-course experiments: alpha factor release (18 time points), cdc 15 (24 time points), cdc 28 (17 time points) and elutriation (14 time points). We linearly interpolated some missing data if a point had the two adjacent time points belonging to the same experiment with no missing values. We focused on the coexpression of 133 transcription factors (TFs) with no missing data after interpolation. Using all the statistics discussed in simulations, we computed 133×133 coexpression matrices and compared them to a total of 428 curated genetic and physical interactions from BioGrid. Since the data has a number of ties, we added small random perturbations for the computation of T_1 and T_2 and took the final results as the maximum counts over 50 iterations.

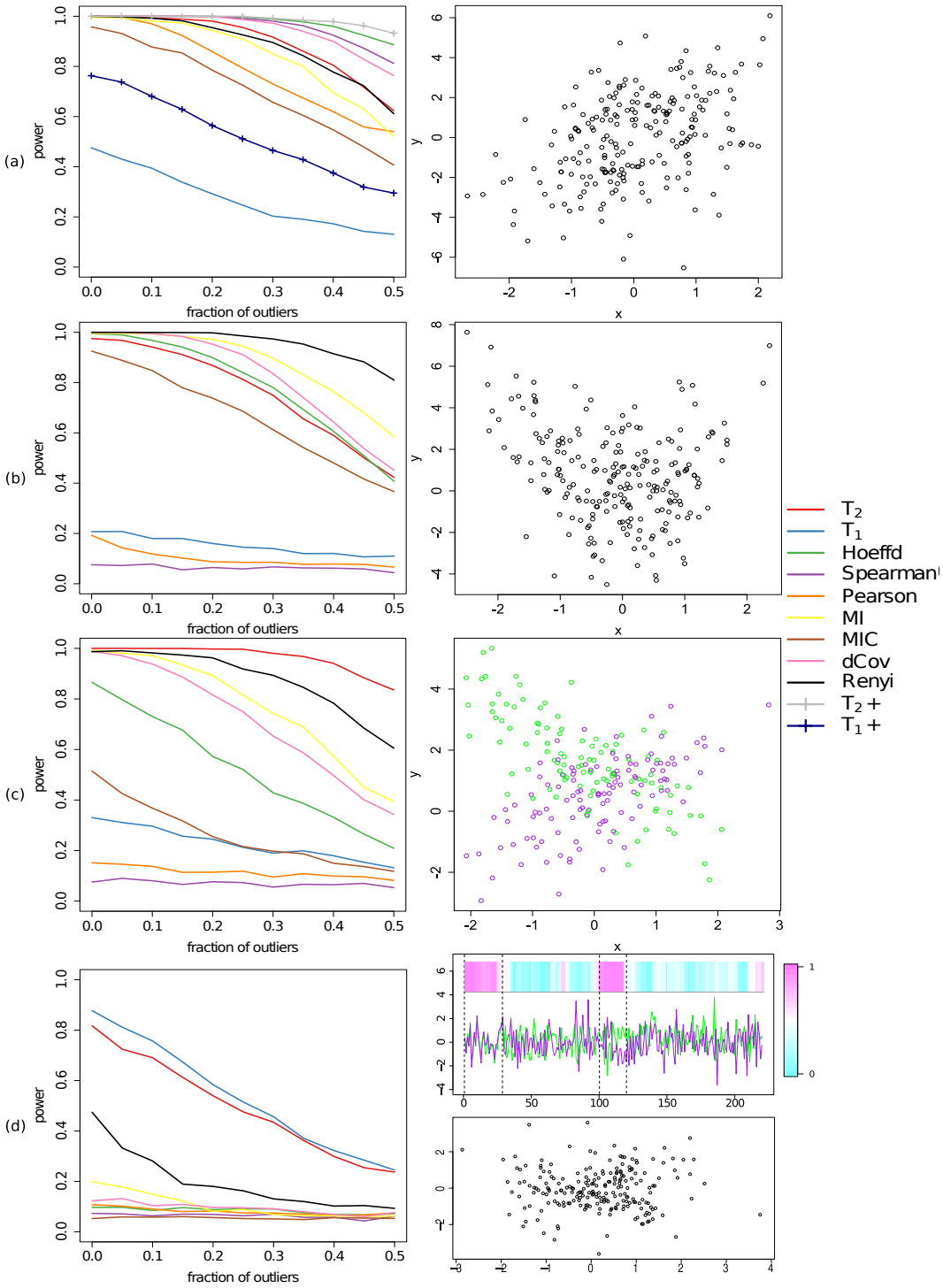


Figure 2.2: The power of various statistics rejecting at 5% significance level as level of contamination by outliers increases when the bivariate data follow (a) a linear relationship; (b) a quadratic relationship; (c) a cross-shaped relationship; (d) two partially coupled time series. The heat map in (d) shows the absolute values of the Pearson correlations calculated at each time point including its neighboring 15 points.

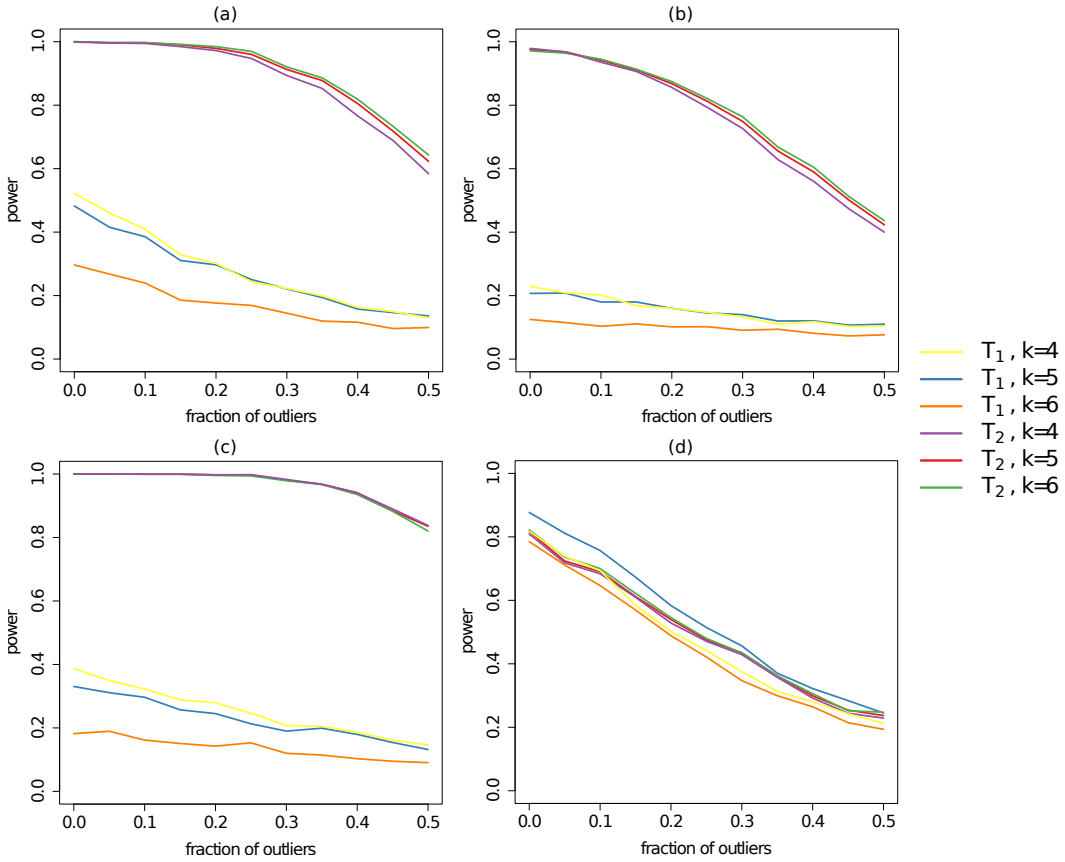


Figure 2.3: The power of T_1 and T_2 for various k values rejecting at 5% significance level as level of contamination by outliers increases when the bivariate data have (a) a linear relationship; (b) a quadratic relationship; (c) a cross-shaped relationship; (d) are two partially coupled time series.

As we expected T_1 to be more suitable for time-course data than T_2, we examined the interactions identified by T_1 more closely. These interactions reveal the ability of T_1 to capture important bivariate associations missed by the other methods. Figure 2.4 shows two pairs of TFs (BAS1 vs GCN4; MSN2 vs YAP1) whose coexpression strengths were consistently ranked among the top 10 and top 20 by T_1 with $k = 7$, but assigned very low rankings by all the other methods. Both pairs correspond to previously reported genetic interactions curated in BioGrid. However, their scatter plots show no obvious trends or dependence patterns, highlighting the importance of preserving the temporal order of the data. More specifically, [43] showed in gene deletion studies that Gcn4p and Bas1p are involved in cooperative transcriptional regulation of the ADE3 gene, which encodes an essential regulon enzyme for the biosynthesis of several amino acids. They were also able to confirm direct binding and occupancy of the promoter region of ADE3 by these two TFs. MSN2 and YAP1 are both activators required for oxidative stress tolerance and there is a partial overlap between their H_2O_2-inducible regulons ([36]). Studies using epistatic miniarray profiles ([115, 8]) have shown double mutations in MSN2 and YAP1 lead to severe fitness defect.

Table 2.3 shows the number of known interactions between TFs among strongly coexpressed pairs as ranked by various statistics. As T_1 led to many ties, the cutoffs were chosen to include the entire stretches of gene pairs with the same statistic values. Overall T_1 (with various choices of k) and the Pearson correlation have the largest number of overlap with the known interactions, with T_1 being the better of the two at most cutoffs. These are followed by T_2 and the Renyi correlation.

	$k=6$				$k=7$			$k=8$			$k=9$		
Top rank	4	7	16	31	4	11	22	5	14	44	3	11	37
Pearson	0	2	2	3	0	2	2	1	2	6	0	2	4
Spearman	0	1	2	2	0	1	2	0	1	2	0	1	2
Hoeffding's D	1	1	1	2	1	1	2	1	1	2	0	1	2
MI	0	1	1	1	0	1	1	1	1	1	0	1	1
MIC	1	1	1	1	1	1	1	1	1	2	1	1	2
dCov	1	1	1	2	1	1	1	1	1	2	1	1	2
Renyi	0	0	2	2	0	2	2	0	2	3	0	2	2
T_1	0	1	3	3	1	3	4	1	3	6	0	1	5
T_2	0	1	2	3	0	2	2	0	2	3	0	2	3

Table 2.3: Number of known interactions in highly ranked coexpression pairs by various statistics. A range of k values were tested for T_1, and $k = 7$ for T_2.

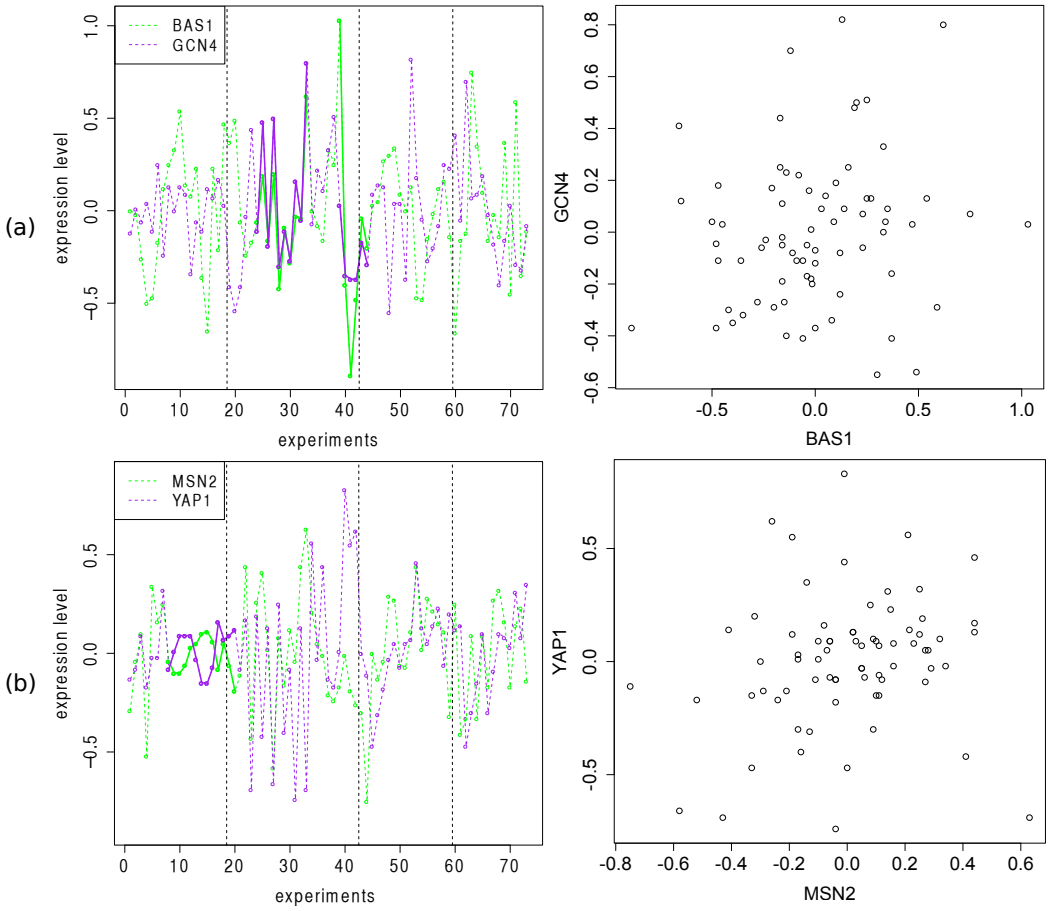


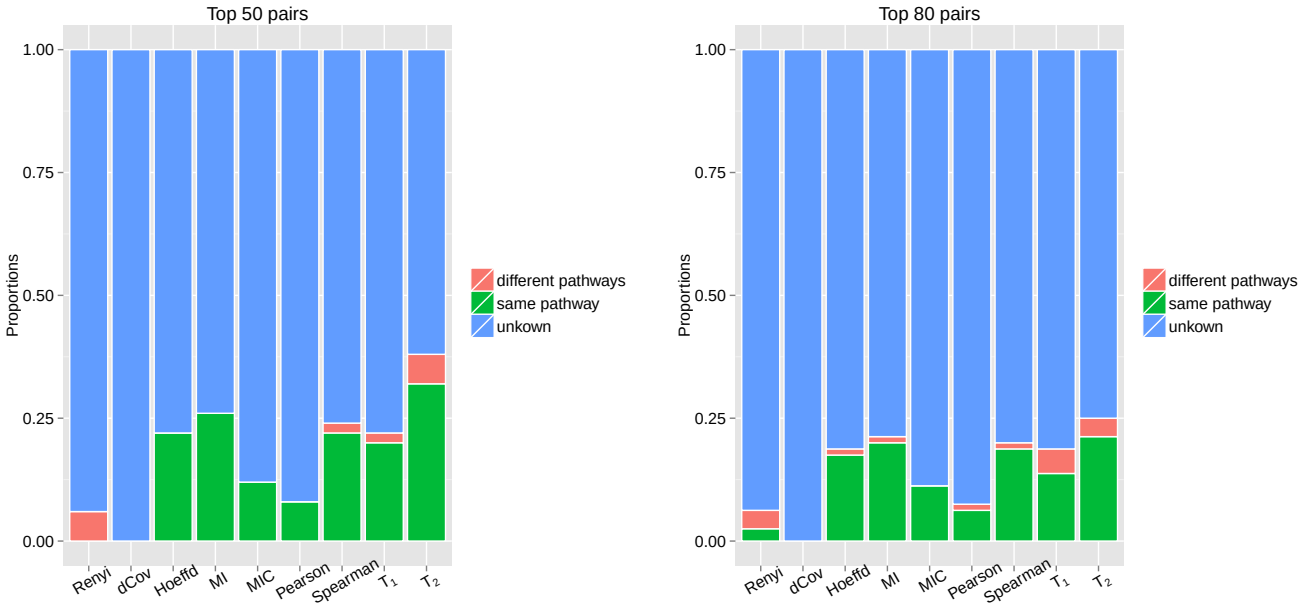
Figure 2.4: Expression levels of (a) BAS1 and GCN4; (b) MSN2 and YAP1 in four time-course experiments (boundaries indicated by the dashed lines). The solid lines highlight regions contributing to the counts in T_1.

Arabidopsis microarrays

We integrated data from 13 microarray experiments to create a meta-data with 220 samples for 22810 *Arabidopsis* genes. The samples were harvested from shoot tissues and different regions of root tissues subject to various stress experiments including salt, low pH, and sulfur deficiency treatments. From [49], we downloaded a list of genes involved in the glucosinolates biosynthesis pathway in addition to the 30 pathways in [51] to comprise a total of 510 unique pathway genes. We computed the pairwise coexpressions between these pathway genes and all the genes available to test the performance of various measures on distinguishing genes in the same pathway. Our selection of k was guided by the log value of the total sample size which is approximately 5. The results presented here were obtained by setting $k = 5$ for T_1 and $k = 9$ for T_2. As expected, T_2 is not sensitive to the choice of k and the results below

remain stable for a range of k.

Figure 2.5 shows the proportions of gene pairs i) in the same pathway; ii) in two different pathways and iii) containing one non-pathway gene among the top 50 and 80 pairs as ranked by all the methods. T_2 achieves the best pathway enrichment followed by MI, the Spearman correlation, Hoeffding's D and T_1. As the samples are not composed of long time-course data, it is not surprising that T_1 is a less ideal statistic than T_2. dCov and Renyi are among the worst performing methods with almost no pairs in the same pathway, despite their good performance in simulations. Extending the cutoffs to examine more highly ranked pairs, in Figure 2.6 the same trend continues for the best four methods until around the top 700 pairs, after which they start to become indistinguishable. dCov remains the bottom of the list.



(a) (b)

Figure 2.5: Number of gene pairs in the same pathway (green), in different pathways (red) and containing a non-pathway gene (blue) among (a) the top 50 pairs and (b) the top 80 pairs as ranked by each method.

Figure 2.7 shows two examples where the gene pairs are in the same pathway, but their coexpression values remain significant at 5% level after Bonferroni correction only under T_2. Some of the sample points are color coded according to their tissue types or treatments to highlight the different patterns of association they exhibit and the lack of a consistent global structure. T_2 is more powerful in this situation due to its definition.

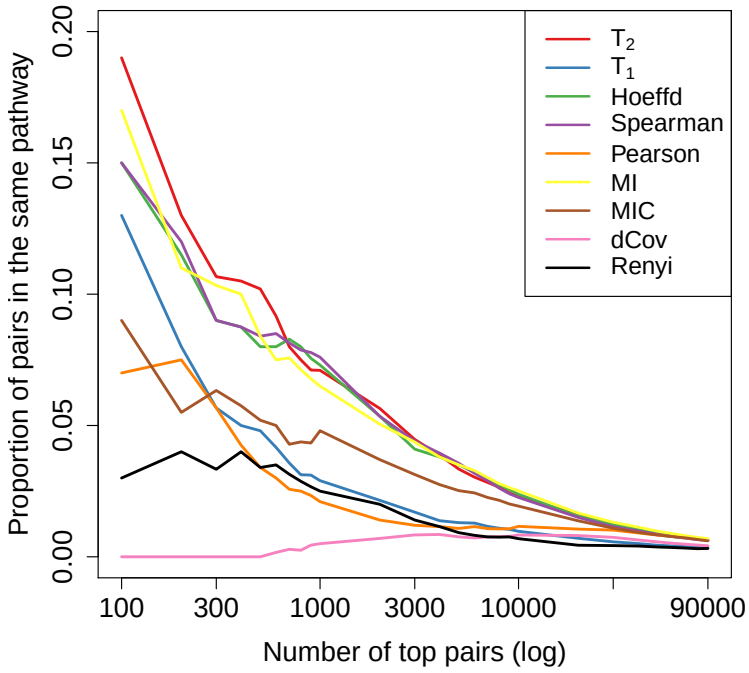


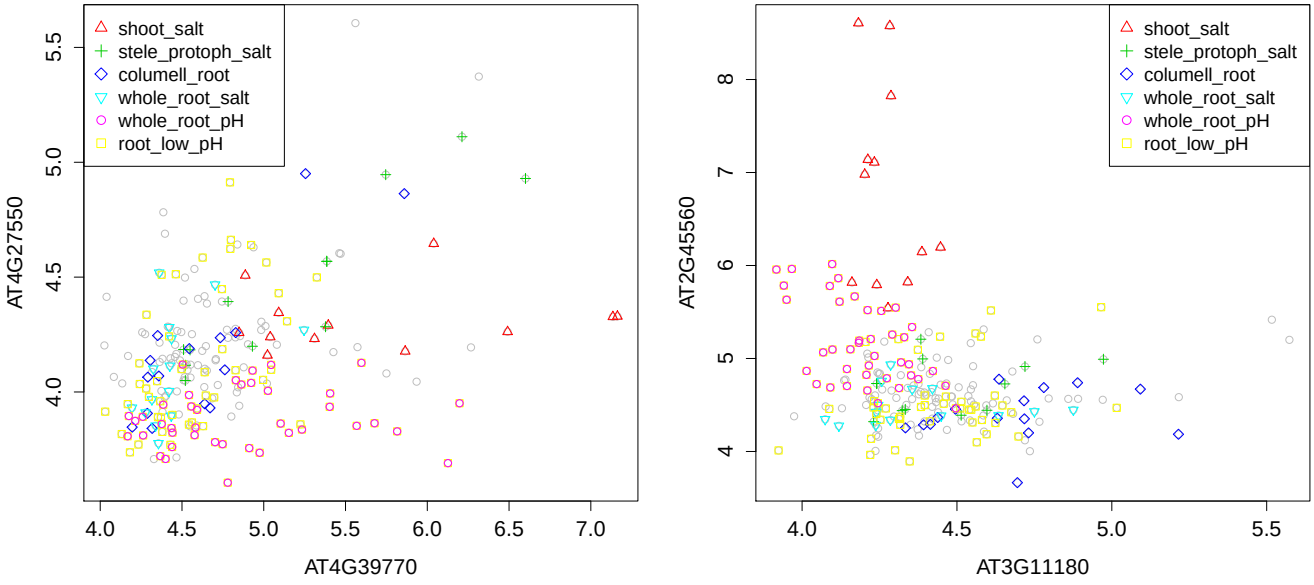
Figure 2.6: Proportion of gene pairs in the same pathway as a function of the number of highly ranked pairs chosen.

A closer look at the types of relationships detected by T_2 and its closest competitor MI reveals MI is underpowered on linear relationships with outliers, an issue also reported by [88]. Examples shown in Figure 2.8 for two pairs of genes in the same pathway, where the bulk of the samples follow a linear trend but they failed to be identified by MI at an unadjusted significance level of 5%. On the other hand, both pairs were assigned significant p-values by T_2 and other statistics including the Pearson and Spearman correlations.

2.5 Discussion

Statistically, the problem of discovering gene coexpression boils down to detecting bivariate associations between gene expression profiles. In this Chapter we propose two new statistics capable of detecting local dependence structures within expression data, motivated by the observation that real gene relationships may have disparaging behaviors in large heterogeneous samples. The statistics are fast to compute, and their asymptotic distributions under the null assumption of independence and exchangeable sample distribution can be derived.

As demonstrated in both simulation and the yeast cell cycle data, T_1 specializes in de-



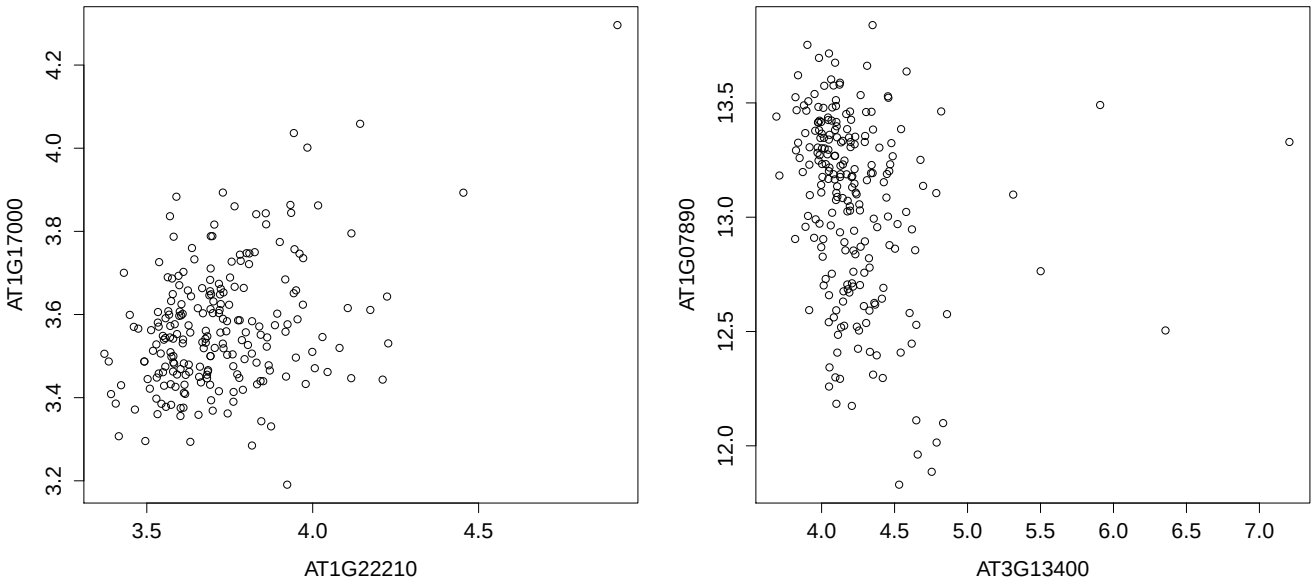
(a) (b)

Figure 2.7: Expression levels of two gene pairs in the same pathway with some samples color coded according to their tissue types or treatments: 1) shoot tissues under salt stress; 2) stele and protophloem cells under salt stress; 3) columella root cap under salt stress, low pH and sulfur deficiency; 4) whole root under salt stress; 5) whole root under low pH; 6) other root cells under low pH.

tecting local associations in time-course data. In particular, when such associations are not visible within the global association pattern, T_1 offers an attractive alternative to other commonly used coexpression measures. The statistic T_2, which considers more general local patterns of dependence, is effective on a variety of functional and nonfunctional relationships. However, as T_2 relies on counts from monotonic sub-patterns, it is sensitive to noise on high frequency sinusoidal relationships.

Both statistics involve a tuning parameter k. Some discussions regarding its choice have been given, but more thorough studies investigating how it affects the performance of the statistics in relation to the structure of data would be desirable.

Our definitions and asymptotic analyses of the two unnormalized counts W_1 and W_2 naturally open room for further investigation. Modifying the current definitions to account for ties in the data would be an important extension to pursue. At a more fundamental level, other choices of the interaction measure $F(\cdot;\cdot)$ in (2.1) would be interesting to explore. For instance, we can consider relaxing the exact pattern matches to fuzzy matches, or replacing



(a) (b)

Figure 2.8: Expression levels of two gene pairs in the same pathway showing a linear relationship with outliers which were not identified as statistically significant by MI.

the indicator function itself with a correlation-based statistic. In terms of asymptotics, it would be of theoretical interest to study the limiting distribution of W_2 for k beyond the log regime. In practice, there often exist inherent dependence structures among the gene samples, especially in time-course data. Thus removing the exchangeability assumption in the analysis of the null distributions would improve computational accuracy of the p-values. In particular, it would be interesting to consider an autoregressive model for W_1 and examine the relationship between the model coefficients and the co-occurrence of rank patterns. Alternatively, it would also be interesting to study the sample dependence directly by reversing the roles of genes and samples and applying a similar technique.

Chapter 3

Inferring gene-gene interactions and functional modules using sparse canonical correlation analysis

3.1 Overview

The coexpression measures discussed in Chapter 2 estimate marginal dependencies only considering pairwise relationships. However, in a real biological pathway, a gene can interact with a group of genes but their marginal relationships may remain weak. Such higher-level interactions (i.e. gene group interactions) are better modeled by Gaussian graphical models (GGM) due to its interpretation in terms of conditional correlations. Under the assumption of multivariate normality of gene expression vectors, the GGM uses the inverse of the gene covariance matrix (or precision matrix), as a measure for gene associations. This approach is closely related to the concept of partial correlations: the (i, j)-th element in the precision matrix is proportional to the partial correlation between gene i and j conditional on the rest of the genes. To address the "curse of dimensionality" (the number of genes being much larger than the number of samples) in estimating the precision matrix, one can exploit the belief that gene networks are inherently sparse and reframe the problem of estimating partial correlations in a penalized regression setting ([65, 73]). More studies on estimating sparse precision matrix in high dimensional GGMs can be found in e.g. [82], [31] and [116].

Despite their attractive theoretical properties, these partial correlation based methods still have limitations in their estimation methods. In the current literature, partial correlation is usually calculated conditioned on either all of the available genes or a more or less arbitrary subset of them that may contain noisy (biologically unrelated) genes. [33] reported that conditioning on all genes simultaneously can introduce spurious dependencies which are not from a direct causal or common ancestors effect. To alleviate this concern, there are alternative approaches using lower order partial correlations ([58, 33, 63, 103, 104]) which condition on one or two other genes. However, these methods come at a cost of lowering the

sensitivity for inferring higher level gene associations and do not necessarily eliminate the effect of noisy genes. [49] proposed to minimize the impact of noisy genes by conditioning on a small set (3-5 genes) of "seed genes" (i.e. known pathway genes). However, such prior biological information is not always available, especially in exploratory studies.

In this Chapter we tackle the problem of estimating gene relationships when the correct conditional set for partial correlation is unknown. Compared to the type of data considered in Chapter 2, we require similar gene behaviors across all the samples. We introduce a new method of inferring the strength of gene group interactions using sparse canonical correlation analysis (SCCA) with repeated random partition and subsampling of the gene expression dataset. There has been a growing interest in applying SCCA to genomic datasets ([98, 71, 105, 55]) in the context of studying relationships between two or more sets of variables, such as gene expression levels, copy numbers and other phenotype variations, with measurements taken from the same sample. One novelty of our method lies in the application of SCCA to a single dataset facilitated by a random partition scheme. By randomly separating the genes into two groups, SCCA searches for a strong linear relationship between a small set of genes, e.g. 5-20 genes, from both groups of genes (e.g. 500-2000 genes in total). Through multiple rounds of random partition, this SCCA approach, reframed in a linear regression setting, gives estimates proportional to partial correlations conditioned on different sets of signal genes (with noisy genes eliminated through sparsity). The subsampling procedure analyzes different subsets of the genes at a time and enables simultaneous identification of multiple interacting groups with different signal strengths. Using this construction, we build an edge weight matrix for the whole gene network whose interaction measure reflects an aggregated estimate of partial correlations of different orders. Our approach is flexible and can be adapted to work with or without prior biological knowledge.

3.2 Methods

As mentioned in the Overview, the conditional correlation interpretation of partial correlation suggests it is a more appropriate framework for modeling higher level interactions in gene networks, provided the conditional computation is carried out properly. In this section, we discuss some of the limitations of the partial correlation approach that arise due to its reliance on the correct selection of conditional sets of genes and how our SCCA based approach circumvents this difficulty. We then give a detailed description of our new method of estimating an edge weight matrix using SCCA with subsampling.

Method motivation

Recall that when the gene expression levels follow a multivariate normal distribution, for a set of genes W, the partial correlation between genes i and j can be expressed as

$$\rho_{ij} = \text{cor}(i, j|W\backslash\{i, j\}) = \begin{cases} -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}, & i \neq j \\ 1, & i = j, \end{cases} \tag{3.1}$$

where ω_{ij} are elements in the precision matrix $(\Sigma^G)^{-1}$ with Σ^G being the gene covariance matrix of the set W (see e.g. [26]). Genes i and j being conditionally independent is equivalent to the corresponding partial correlation and element in the precision matrix being zero.

As pointed out in [33] and [49], the selection of a proper set of genes on which the correlation in (3.1) is conditioned determines the effectiveness of using partial correlation to measure gene interactions. The inclusion of noisy (biologically unrelated) genes in the set $W\backslash\{i, j\}$ may introduce spurious dependencies and consequently false edges in the estimated network. The use of partial correlation may also prove problematic when W contains multiple pathways. Here is a minimal example: suppose the set W has two pathways $\{x, y, z\}$ and $\{u, v\}$ and two independent noisy genes p and q, with expression relationships

$$z = x + y + \epsilon_1 u + \epsilon_2 v + \epsilon_3 p, \quad u = \delta_1 x + \delta_2 y + \delta_3 z + \delta_4 q + v, \tag{3.2}$$

where ϵ_i and δ_j are small constants so that the dependencies between the two pathways are negligible, and gene v is independent of genes x and y. Computing the partial correlations, we have the desired dependencies:

$$\begin{aligned} &\text{cor}(z, x|W\backslash\{z, x\}) = \text{cor}(z, y|W\backslash\{z, y\}) = 1, \\ &\text{cor}(u, v|W\backslash\{u, v\}) = 1, \end{aligned}$$

but also some spurious ones:

$$\text{cor}(u, x|W\backslash\{u, x\}) = \text{cor}(u, y|W\backslash\{u, y\}) = \text{cor}(u, z|W\backslash\{u, z\}) = 1.$$

Using these partial correlations to construct an edge weight matrix would imply the two pathways are fully connected. The proper calculation should condition only on genes in the same pathway, but such information is usually hard to obtain in practice. Alternatively, a more appropriate edge weight measure can take into account the magnitude of the linear coefficients in (3.2) so that it reflects the amount of contribution each gene makes to a pathway and the two-block nature of the network. Recall that in a regression setting, the regression coefficients are multiplicative functions of the corresponding partial correlations. In this sense, the coefficients encompass more information and provide a better resolution on gene relationships than the partial correlations alone.

Motivated by these observations, we propose a new way to assess gene group interactions. In particular, we aim to identify strong linear relationships possessed by a small subset of the

candidate genes. We make direct use of the linear coefficients found by SCCA when applied to two randomly partitioned gene groups. With repeated random partition on subsampled gene sets, an edge weight matrix built by the average SCCA coefficients over iterations reflects an aggregated level of direct or partial gene interactions. Sparsity is imposed to reduce dimensionality and in particular in the example above, ensures the mixing of the two pathways is negligible on average.

Review of sparse canonical correlation analysis and its implementation

Let $\boldsymbol{X} \in \mathbb{R}^{n \times q_1}$ be a matrix comprised of n observations on q_1 variables, and $\boldsymbol{Y} \in \mathbb{R}^{n \times q_2}$ a matrix comprised of n observations on q_2 variables. CCA introduced by [39] involves finding maximally correlated linear combinations between the two sets of variables. More explicitly, one seeks to find $\boldsymbol{\alpha} \in \mathbb{R}^{q_2}$ and $\boldsymbol{\beta} \in \mathbb{R}^{q_1}$ that solve the optimization problem

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^T \Sigma_{YX} \boldsymbol{\beta} \quad \text{subject to } \boldsymbol{\alpha}^T \Sigma_{YY} \boldsymbol{\alpha} = 1, \boldsymbol{\beta}^T \Sigma_{XX} \boldsymbol{\beta} = 1, \tag{3.3}$$

where $\Sigma_{(\cdot,\cdot)}$ represent the correlation matrices. Note that provided the variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ have nonzero variances, this is equivalent to the usual CCA formulation in terms of covariance matrices.

In practice the population correlations are replaced with their sample counterparts. That is, $S_{YX} = \boldsymbol{Y}^T \boldsymbol{X}/(n-1)$, $S_{XX} = \boldsymbol{X}^T \boldsymbol{X}/(n-1)$ and $S_{YY} = \boldsymbol{Y}^T \boldsymbol{Y}/(n-1)$, assuming the columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ have been centered and scaled. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be the weight vectors solving the optimization problem

$$\max_{\boldsymbol{a},\boldsymbol{b}} \boldsymbol{a}^T S_{YX} \boldsymbol{b} \quad \text{subject to } \boldsymbol{a}^T S_{YY} \boldsymbol{a} = 1,\ \boldsymbol{b}^T S_{XX} \boldsymbol{b} = 1 \tag{3.4}$$

for sample correlations.

For high throughput biological data, q_1 and q_2 are typically much larger than n. It is thus natural to impose sparsity on $\boldsymbol{a}$ and $\boldsymbol{b}$, and this can be done by including (typically convex) penalty functions in (3.4). A number of studies ([98, 106, 71]) have proposed various methods for formulating the penalized optimization problem and obtaining sparse solutions. Here we adopt the diagonal penalized CCA criterion given by [106], which treats the covariance matrices in (3.4) as diagonal and relaxes the equality constraints for convexity:

$$\max_{\boldsymbol{a},\boldsymbol{b}} \boldsymbol{a}^T \boldsymbol{Y}^T \boldsymbol{X} \boldsymbol{b} \quad \text{subject to } \boldsymbol{a}^T \boldsymbol{a} \le 1,\ \boldsymbol{b}^T \boldsymbol{b} \le 1,\ p_1(\boldsymbol{a}) \le c_1,\ p_2(\boldsymbol{b}) \le c_2, \tag{3.5}$$

where p_1 and p_2 are convex penalty functions. In this paper, we consider an L_1 penalty and solve the above optimization using the modified NIPALS algorithm proposed by [55], which is reported to yield better empirical performance than Witten et. al. (2009)'s algorithm. The modified NIPALS algorithm performs penalized regressions iteratively on $\boldsymbol{X}$ and $\boldsymbol{Y}$ with the penalty functions $p_{\lambda_1}(\cdot) = \lambda_1 \|\cdot\|_1$ and $p_{\lambda_2}(\cdot) = \lambda_2 \|\cdot\|_1$. This is an equivalent formulation to iteratively optimizing (3.5) using the bounded constraints.

It is important to note that one more complication arises when SCCA is applied to gene expression data. In CCA, the estimation of the correlation matrix using sample correlations requires the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ have independent rows. However, given a gene expression matrix with genes in columns and experiments in rows, it is often the case that row-wise and column-wise dependencies co-exist. Row-wise dependencies, or experiment dependencies, can be defined as the dependencies in gene expression between experiments due to the similar or related cellular states induced by the experiments ([95]). When unaccounted for, they can introduce redundancies that overwhelm the important signals and lead to inaccurate estimates of gene correlation matrix. To decouple the effect of experiment dependencies from the estimation of gene correlations, we apply the *Knorm* procedure from [95]. The *Knorm* model assumes a multiplicative structure for the gene-experiment interactions, and iteratively estimates the gene covariance matrix and experiment covariance matrix through a weighted correlation formula. In addition, row subsampling and covariance shrinkage are used to ensure robust estimation.

Constructing edge weight matrix by SCCA with repeated random partition and subsampling

Suppose an observed dataset contains measurements of the expression levels of p genes in n experiments, where each experiment has a small number of replicates. We next describe our new procedure of computing edge weights that reflect gene group interactions in the gene network.

Summary of procedure

Step (i): *Data normalization by Knorm.* A gene expression matrix $\boldsymbol{Z}_b$ of dimension $n \times p$ can be generated from the full dataset by sampling one replicate from each experiment. Using the *Knorm* model in [95], we normalize $\boldsymbol{Z}_b$ as

$$\boldsymbol{Z}_b^* = (\hat{\Sigma}^E)^{-1/2}(\boldsymbol{Z}_b - \hat{\boldsymbol{M}}), \tag{3.6}$$

where $\hat{\boldsymbol{M}}$ is the estimated mean matrix and $\hat{\Sigma}^E$ is the estimated experiment correlation matrix.

Step (ii): *Subsampling.* For each normalized expression matrix $\boldsymbol{Z}_b^*$, sample (without replacement) a fixed fraction s, say 70%, of the genes to obtain an $n \times sp$ submatrix $\boldsymbol{Z}_b^{\text{sub}}$.

Step (iii): *SCCA with random partition on the subsampled matrix.* For each partition t, randomly split the columns (genes) of $\boldsymbol{Z}_b^{\text{sub}}$ into two groups of equal size (more explanation given in the remarks below) to form $\boldsymbol{X}_{b,t}^{\text{sub}}$ and $\boldsymbol{Y}_{b,t}^{\text{sub}}$. Run SCCA on $\boldsymbol{X}_{b,t}^{\text{sub}}$ and $\boldsymbol{Y}_{b,t}^{\text{sub}}$: find sparse weight vectors $\boldsymbol{a}_{b,t}^{\text{sub}}$ and $\boldsymbol{b}_{b,t}^{\text{sub}}$ using the modified NIPALS algorithm ([55]) with the L_1 penalty and tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, the choice of which will be discussed in Section 3.3.

Let $\boldsymbol{c}_{b,t}$ be the list of the absolute values $|\boldsymbol{a}_{b,t}^{\text{sub}}|$ and $|\boldsymbol{b}_{b,t}^{\text{sub}}|$ ordered according to the gene list. For the genes not included in the subsampled matrix, the corresponding values in $\boldsymbol{c}_{b,t}$ are set to 0. Average over all the partitions to obtain the average weights $\bar{\boldsymbol{c}}_b$. Define edge weight matrix $\boldsymbol{A}_b = \bar{\boldsymbol{c}}_b \bar{\boldsymbol{c}}_b^T$, setting $\text{diag}(\boldsymbol{A}_b) = 0$ to exclude self loops.
Step (iv): *Repeat steps (ii) and (iii) B times.* Define $\bar{\boldsymbol{A}} = 1/B \sum_{b=1}^{B} \boldsymbol{A}_b$ and normalize by the maximum value in $\bar{\boldsymbol{A}}$.

As will be demonstrated in Section 3.3, $\bar{\boldsymbol{A}}$ defined above exhibits a natural block structure when there is one or multiple functional groups. Here are more remarks on our procedure to construct $\bar{\boldsymbol{A}}$:

1. Step (i) can be skipped when dependencies between experimental conditions are weak and not of concern.

2. Step (ii) subsampling is necessary if we aim to identify multiple functional groups (that may overlap) simultaneously. As there will be multiple groups with strong interactions, not all of them can be detected unless different subsets of genes are considered. More discussion about the subsampling step and the choice of subsampling levels is given below.

3. During the random partition in step (iii), the two sets of genes do not have to be exactly equal in size, but they need to be comparable in order to maximize the chance of separating any gene functional group of interest into two sets.

4. Through multiple rounds of random partition, SCCA gives estimates in a regression setting proportional to partial correlations conditioned on different sets of signal genes. Overall subsampling and random partition enable us to consider different subsets of the genes and ways to group them. Thus the elements in $\bar{\boldsymbol{A}}$ can be interpreted as an aggregated measure of partial correlations of different orders as the algorithm steps through different conditional sets of genes.

5. As we search through different subsets of genes, different signal groups are identified depending on the strengths of linear associations in the subset. As will be shown empirically in Section 3.3, the averaged result leads to the formation of a distinct block structure with different connectivities in the matrix.

Our procedure is flexible and can be modified easily to incorporate the following variants:

1. If prior knowledge is available on a pathway of interest, e.g. it is known in advance that some genes are actively involved in that pathway, one may focus on the identification of the gene group related to this pathway first and incorporate the prior knowledge by lowering the penalties associated with those known pathway genes in the SCCA algorithm. Examples involving using prior knowledge of pathway genes can be found in Section 3.3.

2. If the interest is to identify disjoint gene groups and running time is not a concern, we can run the whole procedure iteratively with no subsampling, each time identifying one dominating signal group and removing it from the subsequent analysis.

Asymptotic behavior of our procedure

Here we first show asymptotically the validity of our procedure by considering a simple case where there exists only one functional group and all the other genes are uncorrelated. Due to this simplification, no subsampling is needed, and the use of CCA without sparsity suffices since in the asymptotics we consider the regime of n (number of experiments) going to infinity with p (number of genes) fixed. Without loss of generality, in the entire gene set $G = \{1, 2, \ldots, p\}$ let the first k genes $K = \{1, 2, \ldots, k\}$ form one pathway.

For every partition t, let $\boldsymbol{a}_t$ and $\boldsymbol{b}_t$ be the solutions to (3.4) and $\boldsymbol{c}_t$ be the list of the absolute values $|\boldsymbol{a}_t|$ and $|\boldsymbol{b}_t|$ ordered according to the gene list. Assuming $\boldsymbol{Z}$ follows a multivariate normal distribution and the inverse covariance matrix has a diagonal block structure, we have the following proposition regarding the asymptotic difference between the values of $\{c_{i,t}, i \in K\}$ and $\{c_{j,t}, j \notin K\}$ averaged over t. For convenience suppose p is even and denote $q = p/2$.

Proposition 3.2.1. *Let $\bar{\boldsymbol{c}} = \sum_{t=1}^{N} \boldsymbol{c}_t/N$, where N is the number of partitions, then given $1 < k < q$,*

$$\lim_{N\to\infty} \lim_{n\to\infty} (\min_{i\in K} \bar{c}_i - \max_{j\notin K} \bar{c}_j) = D \tag{3.7}$$

for some positive constant D.

We give the proof of proposition 3.2.1 with a lower bound on D that quantifies the asymptotic difference in the assigned weights between functional group genes and noisy genes. The separation in $\bar{\boldsymbol{c}}$ implies the genes in the graph characterized by the edge weight matrix $\bar{\boldsymbol{A}} = \bar{\boldsymbol{c}}\bar{\boldsymbol{c}}^T$ can be grouped into different clusters based on their connectivity.

We first present the assumptions and proofs needed to establish Proposition 3.2.1. Let $\boldsymbol{Z} \in \mathbb{R}^{n\times p}$ represent an expression matrix with p genes and n experiments, with centered and scaled columns. We have the following assumptions regarding the distribution of $\boldsymbol{Z}$.

Assumption 3.2.2. *$\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$, where $\boldsymbol{z}_i$ are iid $p-$dimensional normal random variables with mean $\mathbf{0}$ and correlation matrix Σ that is invertible.*

Assumption 3.2.3. *The matrix $\Omega = \Sigma^{-1}$ is a diagonal block matrix,*

$$\Omega = \begin{pmatrix} \Omega_1 & \mathbf{0}_{k\times(p-k)} \\ \mathbf{0}_{(p-k)\times k} & \Omega_2 \end{pmatrix}, \tag{3.8}$$

where $\Omega_2 = diag(1, \ldots, 1)$.

Remark 3.2.4. *Note that the diagonal block structure of* Ω *in Assumption 3.2.3 is mirrored in its inverse* Σ*, that is*

$$\Sigma = \begin{pmatrix} \Sigma_1 & \mathbf{0}_{k\times(p-k)} \\ \mathbf{0}_{(p-k)\times k} & \Sigma_2 \end{pmatrix}, \tag{3.9}$$

where $\Sigma_1 = \Omega_1^{-1}$ *and* $\Sigma_2 = diag(1, \ldots, 1)$*. The structure of the correlation matrix implies dependencies only exist among pathway genes.*

Partition the index set G into two sets J_1 and J_2 of equal size. Let $I_1 = J_1 \cap K$ and $I_2 = J_2 \cap K$, that is, I_1 and I_2 represent the corresponding partition on the pathway gene set. For convenience, assume the indices in J_1, J_2, I_1 and I_2 are ordered. Compose submatrix $\boldsymbol{X}$ by selecting columns of $\boldsymbol{Z}$ whose indices lie in the set J_1. Similarly compose submatrix $\boldsymbol{Y}$ based on the index set J_2. In the population case CCA requires finding $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ that solves the optimization problem (3.3). Note that when Σ_{YX} is a nonzero matrix, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are uniquely determined up to a sign. To eliminate this indeterminacy we require $\alpha_1 > 0$, $\beta_1 > 0$.

We also assume the following is true regarding the singular value decomposition of Σ_{I_1,I_2} for every partition.

Assumption 3.2.5. *For any partition, the nonzero singular values of* Σ_{I_1,I_2} *are all distinct.*

Remark 3.2.6. *Assumption 3.2.5 is equivalent to requiring the corresponding submatrix* Σ_{YX} *has distinct nonzero singular values. This assumption is common in literature for the purpose of establishing asymptotic theory for PCA or CCA.*

Since in practice one always aims to solve the sample case (3.4), we first need to establish the asymptotic properties of $\boldsymbol{a}$ and $\boldsymbol{b}$ for a given partition.

Lemma 3.2.7. *As* $n \to \infty$*,*

(i) If $I_1 = \emptyset$*, then* $a_i \leq 1 + o_P(1)$ *for* $k+1 \leq i \leq q$ *and* $b_i \leq 1 + o_P(1)$ *for* $1 \leq i \leq q$*. Similar conclusions hold for the case* $I_2 = \emptyset$*.*

(ii) If $I_1 \neq \emptyset$ *and* $I_2 \neq \emptyset$*, we have* $\boldsymbol{a} \xrightarrow{P} \boldsymbol{\alpha}$ *and* $\boldsymbol{b} \xrightarrow{P} \boldsymbol{\beta}$*.*

Proof of Lemma 3.2.7. We first show the constraints on $\boldsymbol{a}$ and $\boldsymbol{b}$ imply they are bounded with probability one. Let $\hat{\lambda}_i$ be the eigenvalues of S_{YY} and λ_i be the eigenvalues of Σ_{YY}, then

$$\hat{\lambda}_i \xrightarrow{a.s.} \lambda_i \tag{3.10}$$

follows from the fact that

$$S_{YY} \xrightarrow{a.s.} \Sigma_{YY}. \tag{3.11}$$

Writing $S_{YY} = U\text{diag}(\hat{\lambda}_i)U^T$, where U is an orthogonal matrix,

$$\begin{aligned} \boldsymbol{a}^T S_{YY}\boldsymbol{a} &= \boldsymbol{a}^T U\text{diag}(\hat{\lambda}_i)U^T\boldsymbol{a} \\ &= \sum_{i=1}^{q} \hat{\lambda}_i (a_i')^2 = 1, \end{aligned} \tag{3.12}$$

with $(a'_1, \ldots, a'_q)^T = U^T(a_1, \ldots, a_q)^T$. Noting that $\|\boldsymbol{a}'\|_2 = \|\boldsymbol{a}\|_2$ and $\lambda_i > 0$, one can conclude that $\boldsymbol{a} = O_P(1)$. Thus

$$\boldsymbol{a}^T S_{YY} \boldsymbol{a} = \boldsymbol{a}^T \Sigma_{YY} \boldsymbol{a} + o_P(1) = 1, \tag{3.13}$$

and (i) follows from the structure (3.9) of Σ. The same argument applies to $\boldsymbol{b}$.

In the case $I_1 \neq \emptyset$ and $I_2 \neq \emptyset$, $\text{rank}(\Sigma_{YX}) \geq 1$. One can show the convergence holds using Assumption 3.2.5, the fact that $\sqrt{n}(S_{(\cdot,\cdot)} - \Sigma_{(\cdot,\cdot)})$ has a limiting normal distribution and following the arguments in [5]. □

We then proceed to prove Proposition 2.1.

Proof of Proposition 2.1. Consider the following possible partition configurations. (For convenience, the dependency of the coefficients on partition t is suppressed.)

Case (i) $I_2 = \emptyset$.

The probability of this configuration is

$$P_0 = \mathbb{P}(\{I_2 = \emptyset\}) = \frac{\binom{p-k}{q}}{\binom{p}{q}} \cdot \frac{1}{2} \tag{3.14}$$

with $q = p/2$. By Lemma 3.2.7, $c_i \leq 1 + o_P(1)$ as $n \to \infty$ for $i \notin K$.

Case (ii) $|I_2| = 1$, $|I_1| = k - 1$.

This happens with probability $\frac{kq}{q-k+1} P_0$. Assume without loss of generality $I_1 = \{1, \ldots, k-1\}, I_2 = \{k\}$. Partition the pathway correlation matrix Σ_1 and its inverse Ω as

$$\Sigma_1 = \begin{pmatrix} \Sigma_{I_1,I_1} & \Sigma_{I_1,k} \\ \Sigma_{k,I_1} & 1 \end{pmatrix}, \Omega = \begin{pmatrix} \Omega_{I_1,I_1} & \Omega_{I_1,k} \\ \Omega_{k,I_1} & \omega_{k,k} \end{pmatrix}. \tag{3.15}$$

It is easy to see in the population case the solution to (2.3) has the form $\boldsymbol{\alpha} = (\alpha_1, \boldsymbol{\alpha}_2)$ with $\alpha_1 \in \mathbb{R}, \boldsymbol{\alpha}_2 = \mathbf{0}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ with $\boldsymbol{\beta}_1 \in \mathbb{R}^{k-1}, \boldsymbol{\beta}_2 = \mathbf{0}$. Furthermore, α_1 and $\boldsymbol{\beta}_1$ solve the optimization problem

$$(\alpha_1, \boldsymbol{\beta}_1) = \arg\max_{\alpha_1, \boldsymbol{\beta}_1} \alpha_1 \Sigma_{k,I_1} \boldsymbol{\beta}_1 \quad \text{subject to } \alpha_1^2 = 1, \boldsymbol{\beta}_1^T \Sigma_{I_1,I_1} \boldsymbol{\beta}_1 = 1. \tag{3.16}$$

The above condition implies $\alpha_1 = 1$, $\boldsymbol{\beta}_1 = (1/\rho)\Sigma_{I_1,I_1}^{-1} \Sigma_{I_1,k} \alpha_1$ and $\rho^2 = \Sigma_{k,I_1} \Sigma_{I_1,I_1}^{-1} \Sigma_{I_1,k}$, where ρ is the maximal correlation. Noting that

$$\Omega_{I_1,k} = -\Sigma_{I_1,I_1}^{-1} \Sigma_{I_1,k} \omega_{k,k} \tag{3.17}$$

and

$$\omega_{k,k} = (1 - \Sigma_{k,I_1} \Sigma_{I_1,I_1}^{-1} \Sigma_{I_1,k})^{-1}, \tag{3.18}$$

we can write $\boldsymbol{\beta}_1$ as

$$\boldsymbol{\beta}_1 = -\frac{\Omega_{I_1,k}}{\sqrt{\omega_{k,k}^2 - \omega_{k,k}}}. \tag{3.19}$$

Generalizing this to any partition resulting in $|I_2| = 1$ and using the convergence in Lemma 3.2.7, as $n \to \infty$, $c_i = o_P(1)$ for $i \notin K$ and $c_i = C_1 + o_P(1)$ for $i \in K$, where

$$C_1 = \begin{cases} 1 & \text{if } I_2 = \{i\} \\ \frac{|\omega_{i,j}|}{\sqrt{\omega_{j,j}^2 - \omega_{j,j}}} & \text{if } I_2 = \{j\},\ j \neq i. \end{cases} \tag{3.20}$$

Case (iii) $|I_1| > 1$ and $|I_2| > 1$.

By the same argument as in Case (ii), $c_i = o_P(1)$ for $i \notin K$, and $a_i = \alpha_i + o_P(1)$ and $b_i = \beta_i + o_P(1)$ for $i \in K$ with $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ solving the sub-problem

$$\begin{aligned} (\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1) &= \arg\max_{\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1} \boldsymbol{\alpha}_1 \Sigma_{I_2, I_1} \boldsymbol{\beta}_1 \\ &\qquad \text{subject to } \boldsymbol{\alpha}_1^T \Sigma_{I_2, I_2} \boldsymbol{\alpha}_1 = 1, \boldsymbol{\beta}_1^T \Sigma_{I_1, I_1} \boldsymbol{\beta}_1 = 1. \end{aligned} \tag{3.21}$$

Combining results from the above discussion,

$$\lim_{N\to\infty} \lim_{n\to\infty} \bar{c}_i \leq 2P_0, \quad i \notin K; \tag{3.22}$$

$$\begin{aligned} \lim_{N\to\infty} \lim_{n\to\infty} \bar{c}_i \geq & \left(1 + (k-1) \min_{1 \leq j \neq i \leq k} \left\{ \frac{|\omega_{i,j}|}{\sqrt{\omega_{j,j}^2 - \omega_{j,j}}} \right\}\right) \\ & \times \frac{q}{q-k+1} 2P_0, \quad i \in K. \end{aligned} \tag{3.23}$$

The proposition holds with

$$D \geq 2P_0 \left(\frac{k-1}{q-k+1} + (k-1) \min_{1 \leq j \neq i \leq k} \left\{ \frac{|\omega_{i,j}|}{\sqrt{\omega_{j,j}^2 - \omega_{j,j}}} \right\} \frac{q}{q-k+1} \right) > 0 \tag{3.24}$$

□

To further understand the asymptotic behavior of our procedure in general cases when multiple functional groups exist, we present an example that consists of two (disjoint) groups of interacting genes and other unrelated genes. We show a theoretical derivation of $\bar{\boldsymbol{A}} = 1/B \sum_{b=1}^{B} \boldsymbol{A}_b = 1/B \sum_{b=1}^{B} \bar{\boldsymbol{c}}_b \bar{\boldsymbol{c}}_b^T$ for this example in detail to highlight and explain the role of subsampling. We can see that with subsampling, the limiting $\bar{\boldsymbol{A}}$ (when $n \to \infty$) exhibits a natural block structure corresponding to the two gene groups, thus extending the validity of proposition 3.2.1. The ideas underlying the analytical derivation in this simple example are straightforward and directly applicable to general cases, though the computations involved would be very tedious. Note that the analytical computations look tedious even in this small example.

Suppose there are 20 genes in total with two independent functional groups of size 3 each. Let $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T \in \mathbb{R}^{n \times 20}$ represent an expression matrix with $\boldsymbol{z}_i \sim iid$ normal variables with mean 0 and correlation matrix Σ, where $\Sigma = diag(\Sigma_1, \Sigma_2, 1, \ldots, 1)$ with

$$\Sigma_1 = \begin{pmatrix} 1 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1 & 0 \\ 1/\sqrt{2} & 0 & 1 \end{pmatrix}$$

and

$$\Sigma_2 = \begin{pmatrix} 1 & 0.4 & 0.8 \\ 0.4 & 1 & 0.3 \\ 0.8 & 0.3 & 1 \end{pmatrix}.$$

Note that the genes in the first group have a perfect linear relationship $z_{i,1} = \frac{1}{\sqrt{2}} z_{i,2} + \frac{1}{\sqrt{2}} z_{i,3}$ while the second group does not.

We first compute the asymptotic value of $\boldsymbol{A}$ with no subsampling as the number of partitions goes to infinity using the population correlation matrix Σ (i.e. assuming we have infinite observations). Since no subsampling is involved, there is only one such edge weight matrix. The asymptotic values of the edge weight matrix can be calculated by summing the weights from CCA under different partition configurations, weighted by their respective probabilities. Here are more algebraic details for the computation of $\boldsymbol{A}$. For every possible partition, split the index set $\{1, 2, \ldots, 20\}$ into two sets J_1 and J_2 of equal size. Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be weight vectors solving

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \arg\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^T \Sigma_{J_1, J_2} \boldsymbol{\beta} \qquad \text{subject to } \boldsymbol{\alpha}^T \Sigma_{J_1, J_1} \boldsymbol{\alpha} = 1, \boldsymbol{\beta}^T \Sigma_{J_2, J_2} \boldsymbol{\beta} = 1.$$

When Σ_{J_1, J_2} is a zero matrix, we adopt the convention which randomly chooses one gene in J_1 and one gene in J_2 and assigns them weight 1. As in the paper, suppose $\boldsymbol{c}$ is the list of the absolute values $|\boldsymbol{\alpha}|$ and $|\boldsymbol{\beta}|$ ordered according to the gene list. For example, for the partition placing the 20 genes in J_1 and J_2 with genes 1, 4, 5, 6 in J_1 and genes 2, 3 in J_2, $\boldsymbol{c} = (1, 1/\sqrt{2}, 1/\sqrt{2}, 0, \ldots, 0)$. Now define the average weight vector $\bar{\boldsymbol{c}} = 1/N \sum_{t=1}^{N} \boldsymbol{c}_t$, where N is the number of random partitions. $\bar{\boldsymbol{c}} \to \mathbb{E}(\boldsymbol{c})$ as $N \to \infty$ and $\mathbb{E}(\boldsymbol{c})$ can be computed explicitly from the four cases listed in the calculation part below. Then asymptotically,

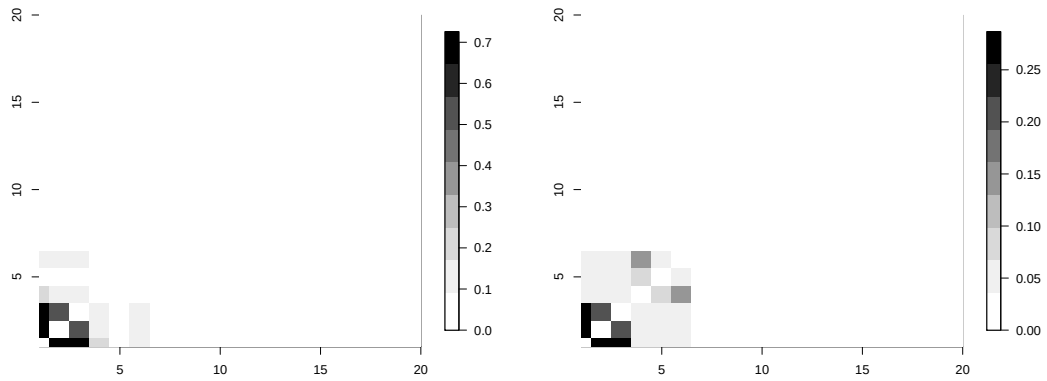
$$\boldsymbol{A} = \bar{\boldsymbol{c}} \bar{\boldsymbol{c}}^T \to \mathbb{E}(\boldsymbol{c}) \mathbb{E}(\boldsymbol{c})^T. \tag{3.25}$$

The asymptotic values (setting the diagonal to 0, without normalization) are plotted in Figure 3.1 (a). We see that without subsampling, the second group is completely overwhelmed by the first group (as demonstrated in configurations of type (iv)). Note also the signal strength within the second group is weaker than that of the interaction between the two groups. When agglomerative clustering is applied, the genes in group two will be merged with group one before merging among themselves, making it very difficult to identify the second group.

With the help of subsampling, we hope to create more subsamples in which the second group dominates the first, thus enhancing its signal strength in $\bar{\boldsymbol{A}}$ when the averages are taken over different subsamples. In this small example, it is possible to compute the asymptotic value of $\bar{\boldsymbol{A}}$ as the number of random partitions and the number of subsamples go to infinity, again assuming we have infinite obeservations which allow us to use the population correlation matrix. Averaging over all subsamples b,

$$\begin{aligned}\bar{\boldsymbol{A}} &= 1/B \sum_{b=1}^{B} \bar{\boldsymbol{c}}_b \bar{\boldsymbol{c}}_b^T \\ &= 1/B \sum_{b=1}^{B} \left(1/N \sum_{t=1}^{N} \boldsymbol{c}_{b,t} \right) \left(1/N \sum_{t=1}^{N} \boldsymbol{c}_{b,t} \right)^T \\ &\to \mathbb{E}_b \left(\mathbb{E}_t \bar{\boldsymbol{c}}_b \right) \left(\mathbb{E}_t \bar{\boldsymbol{c}}_b \right)^T ,\end{aligned} \tag{3.26}$$

as $N \to \infty$ and $B \to \infty$, where $\mathbb{E}_t$ and $\mathbb{E}_b$ denote expectation taken with respect to random partition and subsampling, respectively. The asymptotic values of $\bar{\boldsymbol{A}}$ (setting diagonal to zero, without normalization) are plotted in Figure 3.1 (b). We have set the subsampling level to 70% and more details of the calculations can be found below. The comparison with Figure 3.1 (a) demonstrates theoretically subsampling helps the identification of the weaker group.



(a) (b)

Figure 3.1: Asymptotic values of $\bar{\boldsymbol{A}}$ with (a) no subsampling and (b) 50% subsampling.

The theoretical analysis of the role of subsampling for multiple pathways under general settings can be carried out in a similar fashion. However, since one needs to consider all the possible subsamples and their corresponding partition configurations, the process is rather tedious even when the gene groups are very small as shown by the toy example above. We

deem the full proof out of scope of the current paper. We remark here that subsampling is only necessary when we would like to identify multiple functional groups simultaneously. In practice, if running time is not a concern, we can always run the whole procedure iteratively with no subsampling, each time identifying one dominating signal group and removing it from the subsequent analysis.

Detailed Calculations

We first show the calculations for the case with no subsampling. Let $K_1 = \{1, 2, 3\}$ and $K_2 = \{4, 5, 6\}$ denote the indices of the genes in these two functional groups. For each partition, let $I_{i,j} = K_i \cap J_j$ for $i, j = 1, 2$. Consider the following scenarios:

(i) $|I_{1,j}| = 3$ and $|I_{2,l}| = 3$, $j, l \in \{1, 2\}$. This happens with probability

$$\frac{2 \cdot \binom{14}{10} + 2 \cdot \binom{14}{7}}{\binom{20}{10}},$$

and since the cross correlation matrix is zero in this case, we randomly choose two genes to assign weight 1.

(ii) $|I_{1,j}| = 1$ and $|I_{2,l}| = 3$, $j, l \in \{1, 2\}$. The probability of this type of partition is

$$\frac{2 \cdot 3 \cdot \binom{14}{6} + 2 \cdot 3 \cdot \binom{14}{9}}{\binom{20}{10}}.$$

Clearly $c_i = 0$ for $4 \leq i \leq 20$. Depending on which gene is grouped into $I_{1,1}$ (or $I_{1,2}$),

$$(c_1, c_2, c_3) = \begin{cases} (1, 1/\sqrt{2}, 1/\sqrt{2}) & \text{if gene 1 is chosen,} \\ (\sqrt{2}, 1, 1) & \text{if gene 2 is chosen,} \\ (\sqrt{2}, 1, 1) & \text{if gene 3 is chosen,} \end{cases}$$

all of which are equally likely.

(iii) $|I_{1,j}| = 3$ and $|I_{2,l}| = 1$, $j, l \in \{1, 2\}$, which has probability

$$\frac{2 \cdot 3 \cdot \binom{14}{6} + 2 \cdot 3 \cdot \binom{14}{9}}{\binom{20}{10}}.$$

In this case, $c_i = 0$ for $1 \leq i \leq 3$ and $7 \leq i \leq 20$. Depending on which gene is selected by $I_{2,1}$ (or $I_{2,2}$),

$$(c_4, c_5, c_6) = \begin{cases} (1, 0.215, 0.914) & \text{if gene 4 is chosen,} \\ (1.107, 1, 0.138) & \text{if gene 5 is chosen,} \\ (1.012, 0.030, 1) & \text{if gene 6 is chosen,} \end{cases}$$

all of which are equally likely.

Type of subsample	Probability
$\lvert S_1\rvert = 3,\ \lvert S_2\rvert = 3$	0.0775
$\lvert S_1\rvert = 3,\ \lvert S_2\rvert = 2$	0.1550
$\lvert S_1\rvert = 2,\ \lvert S_2\rvert = 3$	0.1550
$\lvert S_1\rvert = 3,\ \lvert S_2\rvert = 1$	0.0775
$\lvert S_1\rvert = 2,\ \lvert S_2\rvert = 2$	0.2324
$\lvert S_1\rvert = 1,\ \lvert S_2\rvert = 3$	0.0775
$\lvert S_1\rvert = 3,\ \lvert S_2\rvert = 0$	0.0094
$\lvert S_1\rvert = 2,\ \lvert S_2\rvert = 1$	0.0845
$\lvert S_1\rvert = 0,\ \lvert S_2\rvert = 3$	0.0094
$\lvert S_1\rvert = 2,\ \lvert S_2\rvert = 0$	0.0070
$\lvert S_1\rvert = 1,\ \lvert S_2\rvert = 1$	0.0211
$\lvert S_1\rvert = 0,\ \lvert S_2\rvert = 2$	0.0070
$\lvert S_1\rvert = 1,\ \lvert S_2\rvert = 0$	0.0011
$\lvert S_1\rvert = 0,\ \lvert S_2\rvert = 1$	0.0011
$\lvert S_1\rvert = 0,\ \lvert S_2\rvert = 0$	0.0000

Table 3.1: Different subsamples created

(iv) $|I_{1,j}| = 1$ and $|I_{2,l}| = 1$, $j, l \in \{1, 2\}$, which has probability

$$\frac{2 \cdot 3 \cdot 3 \cdot \binom{14}{8} + 2 \cdot 3 \cdot 3 \cdot \binom{14}{7}}{\binom{20}{10}}.$$

Both functional groups have been split up by the partition, resulting in the cross correlation matrix having two non-zero diagnal blocks. It is easy to see the genes in the first group are collinear, thus possessing a stronger linear relationship than the second group. The non-zero block associated with the first group produces the largest singular value, and it follows that the values in $\boldsymbol{c}$ are the same as in case (ii).

Combining the above four cases gives the values in Figure 3.1 (a).

With subsampling, let S denote the indices of the selected genes. Further denote $S_1 = K_1 \cap S$ and $S_2 = K_2 \cap S$. We can create subsamples listed in Table 3.1. For each type of subsample listed, one can carry out the same computation for $\mathbb{E}(\bar{\boldsymbol{c}}_b)$ by considering all the possible partitions for a subsample b.

Identify community structures given the edge weight matrix $\bar{\boldsymbol{A}}$

To demonstrate that $\bar{\boldsymbol{A}}$ possesses advantages over traditional approaches in identifying gene functional modules, subsequent analysis of $\bar{\boldsymbol{A}}$ based on community detection tools is needed. Many methods are available in this field. In particular, clustering has been a popular and well

studied technique. [46, 96, 41] provide general reviews of various clustering techniques, and reviews with more specific focus on gene expression data can be found in [25, 42, 47]. Variants of spectral clustering are also widely explored for detecting communities in sparse networks ([75]). Viewing gene relationships as edges in a graph, a natural approach is to consider functional modules as tightly connected subgraphs. Genes with related functionalities are expected to have dense connections, whereas biologically unrelated (noisy) genes may be only sparsely connected. The Stochastic Block Model (SBM) builds a general probabilistic graph model based on such an assumption that nodes (genes) have different connectivities depending on their block memberships.

Below we introduce two popular community detection tools, SBM and hierarchical clustering (HC), which we will use in later simulation and real data analysis to dissect gene interaction groups from $\bar{\boldsymbol{A}}$. As we have mentioned, there are many other choices for performing this task. The structure of $\bar{\boldsymbol{A}}$ itself may also imply some methods are more suitable than others. In this paper, it is not our intention to suggest or evaluate the best community detection tools that should be applied to $\bar{\boldsymbol{A}}$. Here we are presenting SBM and HC just as two illustrative approaches.

The SBM, formally introduced by [38], generalizes the Erdős-Rényi model and defines a family of probability distributions for a graph. Here is a detailed model definition.

Definition 3.2.8. *A SBM is a family of probability distributions for a graph with node set* $\{1, 2, \ldots, p\}$ *and* Q *node blocks defined as follows.*

1. *Let* $\boldsymbol{C} = (C_1, C_2, \ldots, C_p)$ *denote the set of labels such that* $C_i = k$ *if the node* i *belongs to block* k.

$$\boldsymbol{C} \overset{i.i.d}{\sim} \textit{Multinomial}(\boldsymbol{\gamma}),$$

   where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_Q)$ *is the vector of porportions.*

2. *Let* $\boldsymbol{\pi} = (\pi_{lk})_{1 \le l,k \le Q}$ *be a symmetric matrix of block dependent edge probability matrix and* $\boldsymbol{A}$ *be the adjacency matrix. Conditioned on the block labels* $\boldsymbol{C}$, $(\boldsymbol{A}_{ij})$ *for* $i < j$ *are independent, and*

$$P(\boldsymbol{A}_{ij}|\boldsymbol{C}) = P(\boldsymbol{A}_{ij} = 1 | C_i = l, C_j = k) = \pi_{lk}.$$

Discretizing $\bar{\boldsymbol{A}}$ defined in Section 3.2 into a 0-1 matrix, the class labels and the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$ are estimated using the psuedo-likelihood algorithm by [4]. The unconditional version of the algorithm fits the conventional SBM above, while the conditional version takes into account the variability of node degrees within blocks ([45]). Potential functional groups are identified as classes having large diagonal entries in $\boldsymbol{\pi}$.

Agglomerative HC is another widely used non-model-based technique for extracting communities, especially in the study of social networks ([83]). In our application, we adopt the Ward's distance ([102]) for the computation of merging costs. Let g_i be the nodes, the

distance between two clusters M_1, M_2 is defined as

$$\begin{aligned} d(M_1, M_2) &= \frac{n_1 n_2}{n_1 + n_2} \|m_1 - m_2\|^2 \\ &= \frac{1}{2(n_1 + n_2)} \sum_{i,j \in M_1 \cup M_2} \|g_i - g_j\|^2 - \frac{1}{2n_1} \sum_{i,j \in M_1} \|g_i - g_j\|^2 \\ &\quad - \frac{1}{2n_2} \sum_{i,j \in M_2} \|g_i - g_j\|^2 \end{aligned}$$

where n_1 and n_2 denote the sizes of M_1 and M_2, m_1 and m_2 are the cluster centers of M_1 and M_2 respectively. A natural way to define the square of the pairwise distance is $\|g_i - g_j\|^2 = 1 - \bar{\boldsymbol{A}}_{ij}$ for $i \neq j$, and zero otherwise. Since Ward's method minimizes the increase in the within group sum of squares at each merging and tends to merge clusters that are close to each other and small in size, a small cluster that manages to survive a long distance before coalescing is likely to be a tight cluster, indicating the genes it contains have high connectivity with each other. Thus at an appropriately chosen cutoff level Q, we identify the smallest few clusters as potential functional groups.

Both SBM and HC require a priori knowledge of the number of clusters Q, and the proper selection of Q remains an open problem in literature. For SBM, we refer to some discussions in [40] and [16]. For HC, a common way to choose the cutoff Q is to set it as the number just before the merging cost starts to rise sharply. Due to the scale and complexity of a typical gene expression dataset, this criterion is not very applicable. In this paper, for the HC approach we choose Q empirically based on the sizes of the clusters each Q produces. That is, Q is increased incrementally until small clusters start to emerge. A comparison between SBM and HC can be found in Section 3.3. Although we note here fitting the SBM requires the input of a binary adjacency matrix and the process of thresholding will incur information loss compared to HC.

Flow chart summarizing the whole procedure

A comprehensive summary of the whole procedure including the tuning parameters needed in constructing $\bar{\boldsymbol{A}}$ and illustrative subsequent analysis of $\bar{\boldsymbol{A}}$ is provided in Figure 3.2.

3.3 Results

In this section we evaluate the performance of the proposed method and other approaches using simulated and real microarray datasets. In particular, we compare the quality of the estimated gene functional groups resulting from different ways of computing edge weights, and the two methods of community detection (SBM and HC) discussed in Section 3.2. We use *precision* and *recall*, defined as *precision* = *TP/(TP+FP)* and *recall* = *TP/(TP+FN)*, as measures for evaluating classification performance. Here *TP* is the number of true positive

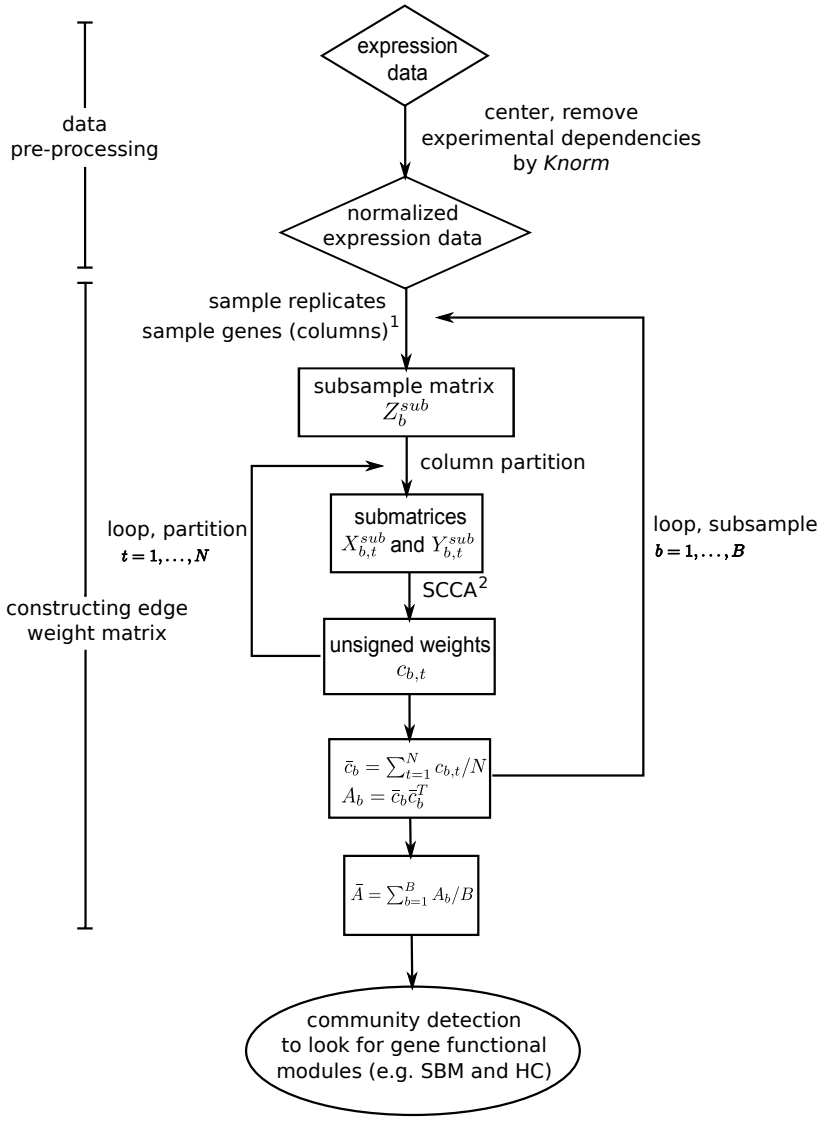


Figure 3.2: Flow chart summarizing the whole procedure. Each numeric superscript in the diagram indicates the need for tuning parameters: 1. subsampling level, 2. penalty parameter $\boldsymbol{\lambda}$.

findings of functional group genes, *FP* is the number of false positives and *FN* is the number of false negatives. In the context of this study, they can be regarded as a measure of exactness and completeness of our search results, respectively. The problems of choosing appropriate proportion of subsampling and $\boldsymbol{\lambda}$ for sparsity are also discussed.

Simulation

Generation of simulation datasets

We simulate a microarray dataset consisting of $p = 150, 300$ or 500 genes and $n = 30$ experiments, with 5 replicates for each experiment. To make the data more realistic, we introduce experiment dependencies, multiple functional groups and random noise. The simulation parameters are generated as follows:

(i) Experiment correlation matrix, Σ^E. For illustrative purpose, we set the experiment correlation matrix to have 0, 33 and 67% dependencies. In the case of a 33% dependency, for example, 33% of the experiments have high dependencies (correlation between 0.5 and 0.6) while the remaining experiments are uncorrelated with one another.

(ii) Gene correlation matrix, Σ^G. In each dataset, we introduce one or two functional groups with 15 genes in each. Genes in the same group are correlated, having either high correlations (0.5 - 0.6) or low correlations (0.1 - 0.2) with the other genes, and otherwise they are not.

Using the above parameters, we generate the simulation data as follows. First, we generate a 30×500 gene expression matrix $\boldsymbol{Z}$, with $vec(\boldsymbol{Z}^T)$, from a multivariate normal distribution with mean zero and a covariance matrix $\Sigma^G \otimes \Sigma^E$. To introduce linear relationships, within each group we take linear combinations of some genes to replace their original values. Using the final 30×500 gene expression matrix, we add random noise with a small SD (e.g. 0.01) to each row to generate the 5 replicates for each experiment.

Estimated $\bar{\boldsymbol{A}}$ and tuning parameter selection

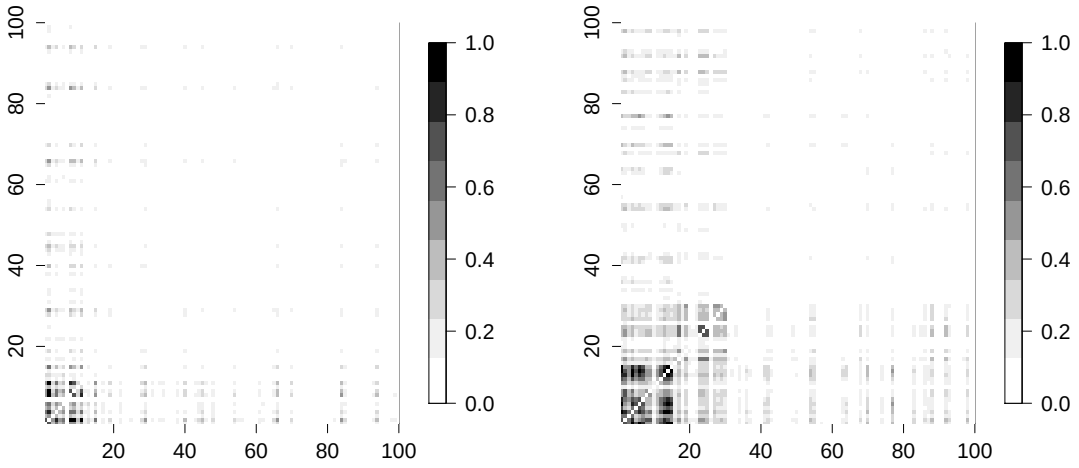
Figure 3.3 shows the heatmaps of the matrix $\bar{\boldsymbol{A}}$ for two datasets with different numbers of functional groups. For visual clarity, the genes are ordered according to their true group memberships. In both cases, the matrix demonstrates a clear block structure. In particular, in the two-group case both pathways are visible although the first one is more prominent. We remark here that the difference in signal strength between the two pathways is introduced by chance variation during data generation and the use of subsampling is necessary for the identification of the weaker group. Although we present results obtained with a subsampling level of 70%, a range of reasonable subsampling levels can be chosen without significantly affecting the final results (further analysis below). The other tuning parameter $\boldsymbol{\lambda}$ is chosen such that the matrix $\bar{\boldsymbol{A}}$ displays optimal contrast between the pathway and non-pathway groups, and we shall use this as a guidance for assessing the quality of $\bar{\boldsymbol{A}}$ and selecting $\boldsymbol{\lambda}$.

Among the common approaches for the selection of optimal tuning parameters, cross-validation based methods are used in [98], [71] and [55]. However, all of their methods involve dividing a sample into multiple sets which is impractical for datasets with only a few tens of observations. [105] proposed an alternative permutation-based method which estimates the p-value of the maximal correlation found by performing SCCA on permuted samples. Due to the large number of partitions and subsamplings required in our method, this approach would be very computationally expensive. Instead we measure the effectiveness of $\boldsymbol{\lambda}$ using

the entropy of $\bar{\boldsymbol{A}}$, defined as

$$H(\boldsymbol{A}) = -\sum_{i<j, \boldsymbol{A}_{ij}>0} (\boldsymbol{A}_{ij}/S_{\boldsymbol{A}}) \log(\boldsymbol{A}_{ij}/S_{\boldsymbol{A}}), \tag{3.27}$$

where $S_{\boldsymbol{A}} = \sum_{i<j} \boldsymbol{A}_{ij}$. The entropy quantifies the sharpness of its distribution and thus is indicative of the signal intensity. Figure 3.4 plots the contours of $H(\bar{\boldsymbol{A}})$ for the same two datasets used in Figure 3.3. Regions with low entropy correspond to $\boldsymbol{\lambda}$ leading to a matrix with better signal intensity.



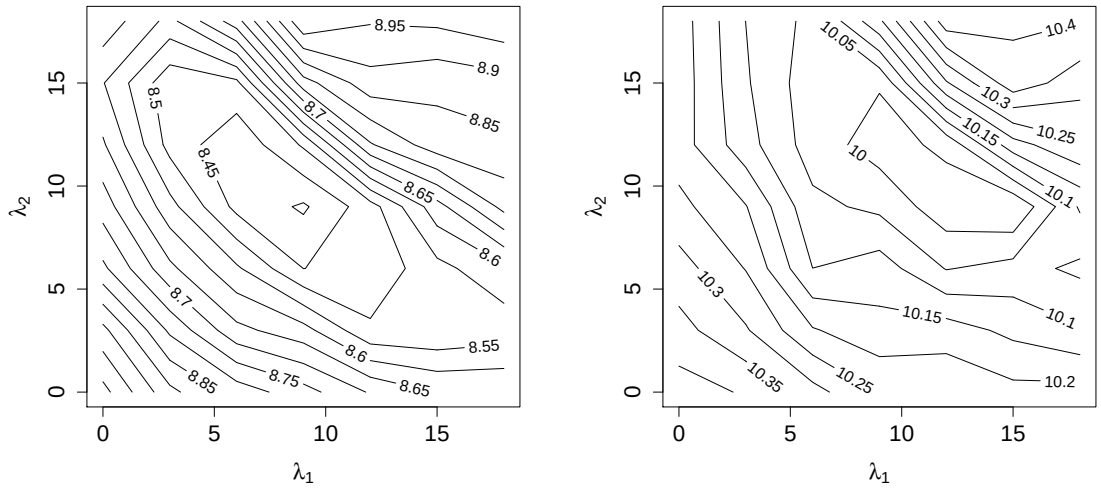
(a) (b)

Figure 3.3: Heatmaps of the matrix $\bar{\boldsymbol{A}}$ using datasets with (a) $p = 150$, 0% experiment dependency, one functional group, subsampling level 70% and $(\lambda_1, \lambda_2) = (9, 9)$; (b) $p = 300$, 0% experiment dependency, two functional groups, subsampling level 70% and $(\lambda_1, \lambda_2) = (9, 15)$. For clarity, only the first 100×100 entries are shown and the functional groups are placed at positions 1-15 and 16-30, respectively.

Performance comparison

Figure 3.5 compares the classification performance of our methods, *scca.sbm* and *scca.hc*, with four correlation-based methods, *pearson.hc*, *pearson.sbm*, *module.dynamic* and *module.hybrid*. The methods are named by cross-mixing the following to allow for comparisons in the two-stage procedure.

scca: Calculate $\bar{\boldsymbol{A}}$'s with $\boldsymbol{\lambda} \in \{9, 12, \ldots, 27\}^2$ and select 10 of these with the smallest entropy values. The final cluster membership (after community detection) is decided by a



(a) (b)

Figure 3.4: Contour plots of the entropy of the upper triangular entries of $\bar{\boldsymbol{A}}$ on the grid $(\lambda_1, \lambda_2) \in \{0, 3, \ldots, 18\}^2$ using datasets with (a) p=150, 0% experiment dependency, one functional group and subsampling level 70%; (b) p=300, 0% experiment dependency, two functional groups and subsampling level 70%.

majority vote based on the selected $\bar{\boldsymbol{A}}$'s so only stable clusters and cluster members are chosen.

pearson: Pearson's correlation matrix after the data is normalized using equation (3.6) and *Knorm* estimates.

module: Transformed Pearson's correlation matrix used in [52].

sbm: Fit a SBM on a discretized edge weight matrix (at level $\{0.3, 0.4, \ldots, 0.8\}$) using the unconditional pseudo-likelihood algorithm in [4] with $Q = 2$ (or 3) initialized by spectral clustering with perturbation. Select the cluster with the highest internal connectivity based on the estimates.

hc: HC with the Ward's distance and cut the dendrogram when clusters of size less than 25 start to appear as the number of clusters Q increases. The choice of this upper bound is based on the size of the cluster selected in *scca.sbm*, and a range of reasonable numbers can be used without affecting the final results.

dynamic, *hybrid*: HC with dendrogram cutting methods in the R package `dynamicTreeCut` ([53]).

Figure 3.5 plots the average *precision* and *recall* of the above six methods calculated on 10 simulation datasets for each level of experiment dependency. It can be seen that using our SCCA approach to compute edge weights in general leads to higher *precision* across

all experiment dependency levels. Of the two ways of community identification, *scca.hc* produces higher *precision* than *scca.sbm* at comparable *recall* levels.

Table 3.2 shows the same performance measures obtained from datasets containing two independent functional groups for *scca.hc*, *pearson.hc*, *module.dynamic* and *module.hybrid*. The numbers are averages from 10 simulation datasets for each level of experiment dependency. Similar to the one-group case, we choose the smallest Q that produces two clusters of size less than 25 as the cutoff in HC. We remark here that when multiple groups are present, *scca.sbm* tends to detect only the strongest signal group while failing to pick up the weaker one. This can be explained by considering the within-class homogeneity assumption in the SBM model and noting that the degree distribution is often less homogeneous in the weaker signal group (see e.g. Figure 3.3). Neither is the conditional pseudo-likelihood algorithm in [4] sensitive enough to detect the finer distinctions. Results from *pearson.sbm* are also omitted as they are very noisy. In all the cases, *scca.hc* demonstrates the best *precision* at comparable, if not better *recall*.

Table 3.2: Classification performance of different methods using datasets with $p = 500$, two pathway groups, subsampling level 70%, and various levels (0%, 33% and 67%) of experiment dependency.

	Pathway 1					
	0%		33%		67%	
	Precision	Recall	Precision	Recall	Precision	Recall
scca.hc	0.861	0.533	0.831	0.441	0.811	0.433
pearson.hc	0.238	0.233	0.497	0.427	0.471	0.393
module.dynamic	0.718	0.3	0.742	0.333	0.764	0.38
module.hybrid	0.439	0.407	0.544	0.447	0.453	0.385
	Pathway 2					
	0%		33%		67%	
	Precision	Recall	Precision	Recall	Precision	Recall
scca.hc	0.808	0.487	0.890	0.489	0.833	0.420
pearson.hc	0.438	0.387	0.323	0.307	0.460	0.273
module.dynamic	0.758	0.4	0.808	0.347	0.8	0.4
module.hybrid	0.565	0.473	0.529	0.387	0.455	0.46

Subsampling levels

Using the same simulated dataset that produced the heatmaps in Figure 3.5 of the paper, we performed the calculation of $\bar{\boldsymbol{A}}$ again at (a) 50% subsampling level and (b) 95% subsampling level. There are two gene groups at positions 1-15 and 16-30, respectively.

As can be seen in Figure 3.6, when almost no subsampling is applied, $\bar{\boldsymbol{A}}$ is predominantly expressing signals from the first group, while the signal intensity of the second group is weaker

than that of the between-group interaction (due to the product definition of A) and only marginally stronger than the background noise. This is because genes in the first group possess stronger linear relationships and the weights in SCCA are preferentially assigned to them under most partitions. Thus the only way to simultaneously capture both groups under our framework is to consider different subsamples whereby the first group does not dominate all the time. A low subsampling level, however, also increases the chance of selecting only noise genes in the subsample. As a result, more background noise is introduced in the final output and the signal ratio is diminished (Figure 3.6 (a)). The choice of subsampling level is a trade-off between group structure complexity and signal to noise ratio. More complex group structures require lower subsampling levels. On the other hand, if the ratio of signal genes to noise genes is small, we need to apply higher subsampling levels to ensure the results do not include too much noise.

Overlapping functional groups

In practice it is often the case there are genes actively participating in multiple pathways. Our edge weight matrix would reflect the overlapping structure as long as the overlapping genes possess strong direct or partial correlations with other genes in those pathways. At the second step, these overlapping blocks can be detected by, for example, fitting an overlapping SBM ([1]). To test our method's performance under the overlapping setting, we simulate a dataset with 150 genes and 30 experiments with 5 replicates each. Genes 1-15 form one functional group and genes 11-25 form the second group. The two groups overlap by 5 genes and genes in the same group have correlations around 0.5-0.6. Figure 3.7 shows the heatmaps of $\bar{\boldsymbol{A}}$ when the procedure is run at 50% level of subsampling, with the matrix demonstrating the desired overlapping block structure.

Incorporating prior knowledge

For illustration, we incorporate prior knowledge for two simulated datasets used in the computation of Table 1 in the paper. Both have 500 genes, 30 experiments with five replicates each. The first set has the first 15 genes forming a functional group. Randomly selecting four genes in the group as prior knowledge and reducing their penalties by half, the procedure retains a perfect precision of 1, and recall improves from 0.533 to 0.733. The second set has two functional groups of size 15 each (genes 1-15 and genes 16-30). Randomly choosing four genes in the first group as prior knowledge, the recall of the first group improved from 0.533 to 0.733. Doing the same for the second group, the recall increased from 0.467 to 0.667.

Application to real data

We tested the performance of our procedure by applying it to *Arabidopsis thaliana* microarray expression data retrieved from AtGenExpress (http://www.arabidopsis.org/servlets/TairObject?type=expression_set&id=1007966941). Analyzed dataset included expression

measurements collected from shoot tissues subject to oxidation stress for 22810 genes under 13 experiment conditions with two replicates for each experiment. In these experiments, the plants were treated with methyl viologen (MV), which led to the formation of reactive oxygen species (ROS). Various studies have shown that depending on the type of ROS, a different biological response is provoked. Thus by focusing on the ROS induced by MV, we were able to show and validate that the results of our pathway gene search were supported, in part, by other already published ROS-related microarray experiments.

A subset of all 22810 genes was selected for analysis based on the following criteria. (i) The experiment variance of the gene exceeds 0.1. An unvarying expression profile suggests the gene has an activity level unaltered by the particular stress condition and hence is unlikely to be part of any stress-induced pathway. The inclusion of such genes may cause problems in covariance estimation as well. We also removed genes with a suspiciously high experiment variance as it could suggest inaccuracy in measurements. (ii) The discrepancy between the two replicates is smaller than 2 for each experiment. This ensures only genes with consistent measurements are included in our analysis. (iii) The minimum expression level exceeds 7. More active genes are likely to possess stronger signals, making our search easier. This requirement further trims down the dataset to a smaller size more desirable for our procedure. We note here that the inclusion of (iii) is optional — if running time is not a concern, the minimum expression level could be either lowered or entirely removed. The final subset for analysis contained 2718 genes.

Potential functional groups were found by *scca.hc*. Due to the complexity and noise level of the dataset, we did not expect the entropy (3.27) to have a clean-cut unimodal distribution. Furthermore, the presence of many groups with varying signal strengths implies each may need a different optimal $\boldsymbol{\lambda}$ for detection. For example, strong groups are likely to require more regularization, or in other words, larger $\boldsymbol{\lambda}$. For this reason, we performed our search in multiple stages starting from large $\boldsymbol{\lambda}$ for stronger groups to smaller $\boldsymbol{\lambda}$ for weaker ones. At every stage, the groups found were removed from the original set before proceeding to the next stage. The upper bound on $\boldsymbol{\lambda}$ was found by increasing $\boldsymbol{\lambda}$ until the entropy stabilized. Searching down from this upper bound, we chose $\boldsymbol{\lambda}$ from three grids: $\{90, 100, 110\}^2$, $\{60, 70, 80\}^2$ and $\{30, 40, 50\}^2$. The cutoff level Q in HC was increased incrementally until at least five clusters of size less than 30 appeared. A reasonable range of numbers can be used to choose the cutoff and our results are not very sensitive to the choice of this number. The full procedure produced 13 groups of genes, the full list of which including annotations can be found in the supplementary information of [101].

To test the biological significance of all 13 groups found (i.e., whether there is a functional relationship between genes within the various groups), we first examined for enrichment of gene product properties, collectively designated gene ontology (GO) annotations, within each group using information available at The Arabidopsis Information Resource (http://www.arabidopsis.org/tools/bulk/index.jsp). We determined that 8 out of 13 groups were highly enriched with genes having the same GO annotation and calculated their p-values using Fisher's exact test to compare with the counts obtained from the full analyzed dataset (Table 3.3).

Table 3.3: GO enrichment of groups

Group ID	Enriched GO term	Number of genes with enriched terms	P-values
1	Chloroplast organellar gene	10 out of 15[1]	1.10×10^{-4}
2	Phenylpropanoid-flavonoid biosynthesis	3 out of 4	6.65×10^{-7}
3	Glucosinolate biosynthsis	7 out of 7	1.95×10^{-14}
4	Chloroplast organellar gene	3 out of 3	7.83×10^{-3}
5	Ribosome	10 out of 15	7.20×10^{-13}
8	Ribosome	5 out of 6	8.31×10^{-8}
10	Photosystem I or II	8 out of 10	2.87×10^{-14}
12	Endomembrane system	3 out of 4	2.35×10^{-3}

In addition to the GO enrichment approach for validating the groups, and in order to support the biological significance of the groups found, we also evaluated other forms of evidence. We were able to determine that for several groups, that the genes placed in the groups encode for known pathways. For example, group 2 genes encode steps in the phenylpropanoid-flavonoid (FB) biosynthesis pathway, and group 3 genes encode for steps in the glucosinolate (GSL) biosynthesis pathway. Both are well-studied secondary metabolic pathways. Flavonoids are compounds of diverse biological activities such as anti-oxidants, functioning in UV protection, in defense, in auxin transport inhibition, and in flower coloring ([34, 67, 94, 109]), and GSLs are sulfur-rich amino acid-containing compounds which become active in response to tissue damage, and believed to offer a protective function ([87, 97, 110]). A considerable number of genes in both pathways are induced by broad environmental stresses, and regulated at the transcriptional level. Based on the lists of genes associated with these two pathways reported in [49], our analyzed dataset contained 13 FB pathway genes and 26 GSL pathway genes. The precisions of our search are 75% and 100%, respectively.

In order to assess the likelihood that genes in the remaining groups could also encode steps within specific pathways, we reviewed microarray data from plants subjected to other forms of oxidative stress (these experiments are similar to the experiment from which our dataset using MV was obtained). Using this approach we found that genes in each of the additional seven groups (1, 4, 5, 8, 9, 11, 12) were strongly associated in these independent experiments (Table 3.4).

[1] 4 out of the 10 chloroplast genes are mitochondrial organellar genes.

Table 3.4: Groups that show co-expression in other oxidative-stress-inducing conditions

Group ID	Experimental (oxidative-stress-inducing) conditions					
	Number of genes with changed expression	SOD knockdown[2]	*apx1* exposed to high light[3]	Ozone[4]	*alx8*[5]	High light (time course)[6]
1	11 out of 15	Up regulated[7]				
4	2 out of 3	Up regulated	Up regulated	Up regulated		
12	4 out of 4	Up regulated	Down regulated[8]	Down regulated	Up regulated	
5	15 out of 15					Co-expression[9]
8	6 out of 6					Co-expression
11	8 out of 8					Co-expression
9	5 out of 5					Co-expression

Of all the groups found, groups 6, 7 and 13 remain uncharacterized in the literature. Nonetheless, using CoExSearch (part of the ATTD-II database (http://atted.jp/top_search.shtml #CoexVersion)), all four genes in group 7 were correlated to some degree with abiotic stress conditions. We also found these genes were common anoxia-repressed genes ([61]). The lack of complete characterization for these groups in the current literature leaves potential scope for further biological examination.

For comparison we applied *pearson.hc*, *module.dynamic* and *module.hybrid* to the same data. As the simulation study suggests the latter two methods in general have better per-

[2]The thylakoid-bound Cu/Zn superoxide dismutases (Cu/ZnSOD, At2g28190) knockdown mutant was compared to wild type plants ([78])

[3]The knockout cytosolic ascorbate peroxidase (apx1, At1g07890) mutant was exposed to high light, and compared to untreated *apx1* plants. ([23])

[4]Arabidopsis seedlings were exposed to 200 ppb ozone for 1 h. (http://affymetrix.arabidopsis.info /narrays/experimentpage.pl? experimentid526)

[5]*alx8*(At5g63980) mutant, which has half H2O2 of that in wild type, was compared to wild type. ([28])

[6]Leaves from 4-week-old plants were exposed to high light for 0.75, 1.5, 3, and 6 h. ([48])

[7]Gene expression is coordinately increased as a result of the specific experimental condition.

[8]Gene expression is coordinately decreased as a result of the specific experimental condition.

[9]Gene expression changes coordinately (up or down) throughout the time course.

Table 3.5: GO enrichment of groups — first cut

Group ID	Enriched GO term	Number of genes with enriched terms	P-values
9	Cell wall	16 out of 81	4.46×10^{-6}
10	Defense response	29 out of 78	1.58×10^{-2}
11	Phenylpropanoid-flavonoid biosynthesis	11 out of 76	5.42×10^{-12}

Table 3.6: GO enrichment of groups — second cut

Group ID	Enriched GO term	Number of genes with enriched terms
62	NA	0 out of 6
63	Chloroplast	4 out of 6
64	Located in plasma membrane	2 out of 5
65	Located in plasma membrane	3 out of 5
66	Pyridoxine biosynthetic process	2 out of 5

formance than *pearson.hc*, particularly in the multi-group case, we will present the results from these two methods. In order to compare with our results, we chose two cuts of the dendrogram such that the first cut produced the same number of groups as our method, and the second one led to groups with sizes comparable to ours. The first cut resulted in 13 groups with sizes ranging from 60 to 293. We picked three most promising groups based on their annotations and the GO analysis is summarized in Table 3.5. Although all of them have statistically significant p-values, their precisions are quite low. In particular, group 11 contains our group 2 as a subset and includes 11 genes (out of 76) in the FB pathway and 5 genes are in isoprenoid biosynthesis pathway. These two pathways are derived from different initial precursors and and known to be unrelated. We note here that at this cut level, the GSL pathway cannot be identified by the method. The second cut produces 66 groups with sizes from 5 to 81. We picked five small groups for analysis and only one group with genes localized in chloroplast has significant GO enrichment (Table 3.6). Even so, these genes are unlikely to be functionally related. The comparison suggests our method can achieve better precision and lead to more biologically meaningful groupings of genes.

Effects of tuning parameters

To systematically study the effects of different tuning parameters on the identification of gene functional groups, we perform sensitivity analysis for different choices of subsampling levels and penalty parameter $\boldsymbol{\lambda}$ using both the simulated and real data discussed above. For the sake of completeness, we also compare tuning parameters from the HC and SBM procedures. Overall our results are reasonably stable for a range of $\boldsymbol{\lambda}$ values. Further stability can be

achieved by pooling results from different $\boldsymbol{\lambda}$. As expected, the choice of subsampling level is more important when there exist multiple functional groups. Our results suggest levels between 50% and 80% can all be considered in practice. For community detection, HC is more robust than SBM in the sense that the classification results are not sensitive to the cutoff chosen. The results are summarized in the tables in Appendix.

3.4 Discussion

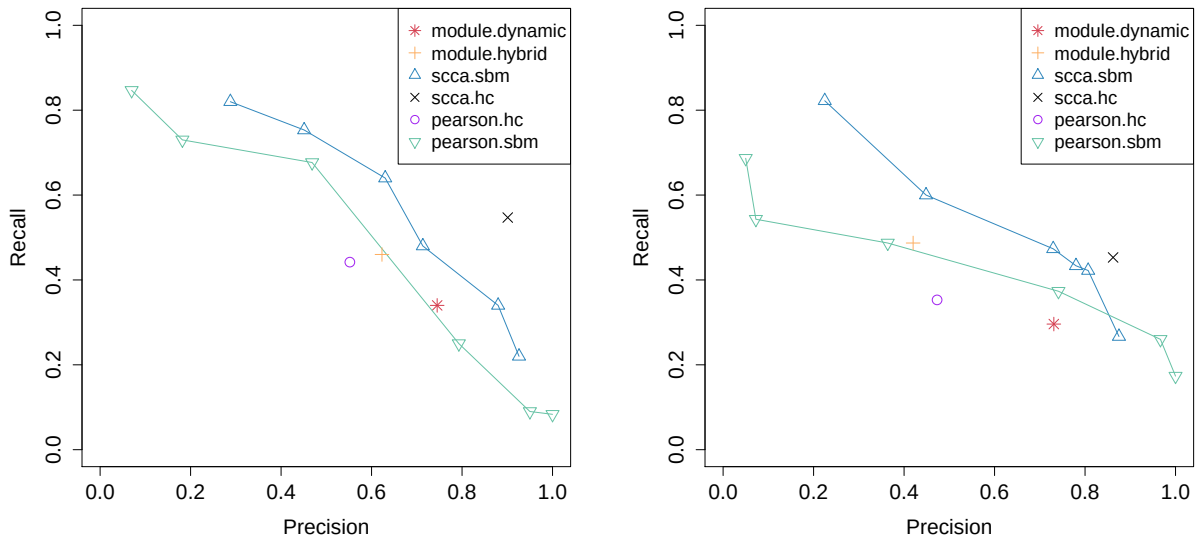
In this paper, we focus on the problem of estimating gene group interactions in gene networks, where data are given in the form of nodes and their associated covariates and estimation of the true network is a challenging task. We propose a new method to construct an edge weight matrix for the full network by applying SCCA to sampled subsets of genes with random partitioning. To evaluate the quality of the constructed network, subsequent analysis of the community structures is applied to identify potential gene functional groups. Although the work is presented under the setting of gene networks, we believe our approach can be generally applicable to answer similar questions in other biochemical networks and even networks in other fields that are sparse and have similar covariate features.

Compared to other popular ways of measuring gene interactions, our SCCA approach is more conceptually appealing. By seeking maximally correlated sets of genes among randomly sampled subsets, this approach provides an aggregated measure of gene partial correlations when the correct conditional set is unknown and thus gives us a better chance of capturing group interactions. As demonstrated in both simulation and real data applications, one of the main attractions of our procedure is its high *precision*. Although it does not seem to greatly improve *recall*, this is not a huge drawback in light of the search algorithm by [49]. Given the accuracy of our search results in general, one can use these identified genes as "seed genes" to initiate a more complete search and expand on the current lists.

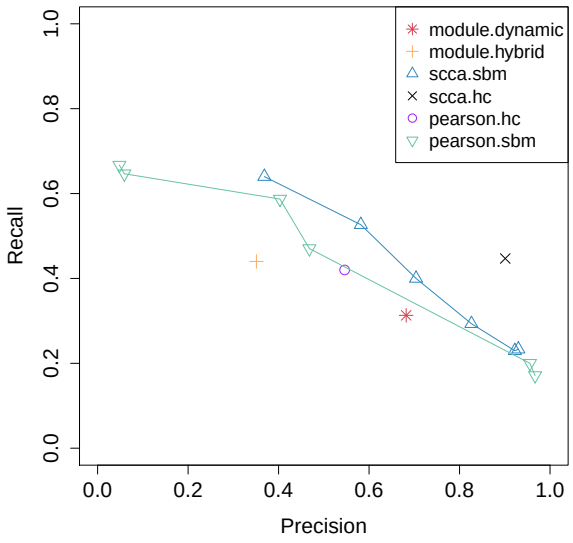
Our approach can be modified to handle other practical situations. When it is known in advance that some genes operate in the same functional group, one may incorporate the prior knowledge by lowering the penalties associated with those genes in the SCCA algorithm. Although we have focused on the case with disjoint functional groups, our method of constructing an edge weight matrix is still applicable to the overlapping case as long as the shared genes possess strong direct or partial interactions with all the other functional genes. However, a different community detection method (e.g. mixed membership SBM, [1]) should be applied to identify the overlapping structures.

The core of our procedure consists of an implementation of SCCA by LASSO regression, and this naturally opens room for further investigation. For example, it would be interesting to find out if using other penalty functions yields different results; more importantly, whether SCCA can be implemented using a different optimization criterion or a more efficient algorithm to lessen the computational cost of our procedure. In the theoretical aspect, it would be desirable to incorporate sparsity into our asymptotic analysis.

On the community detection side, although we used SBM and HC as examples, there are many other available methods to be further explored, especially their properties in relation to the edge weight matrix $\bar{\boldsymbol{A}}$. The use of SBM and HC also gives rise to other interesting extensions. As noted in Section 3.3, conventional SBM does not perform well when there are multiple groups, which is mainly caused by the heterogeneity of node degrees. However, fitting a degree-corrected model using the conditional pseudo-likelihood algorithm does not seem offer significant improvement. It would be desirable to carry out further study on the theoretical properties of the degree-corrected SBM and characterize its identifiability problem. Another possible extension is to modify these algorithms to take weighted adjacency matrices without discretization. Developing a practical but more systematic way of choosing the cutoff level for HC also invites future study.

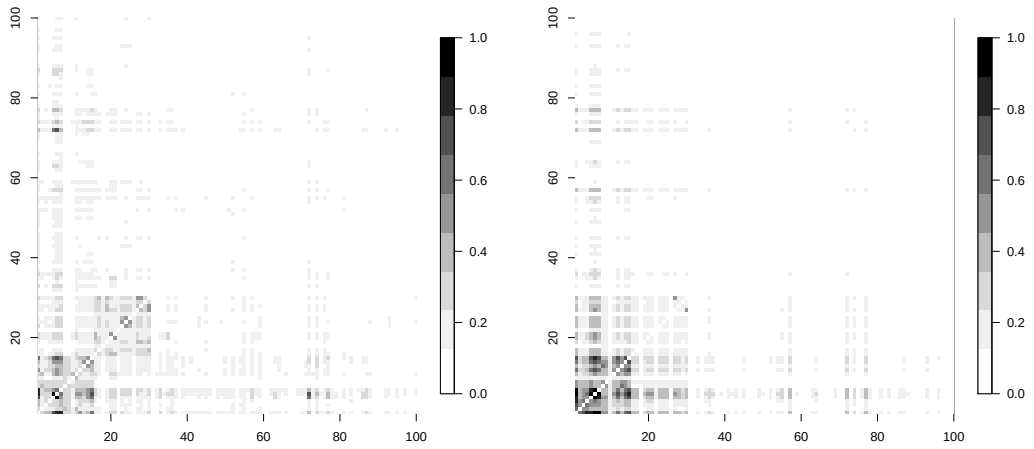


(a) (b)



(c)

Figure 3.5: Classification performance of different methods using datasets with $p = 500$, one pathway group, subsampling level 70%, and (a) 0%, (b) 33% and (c) 67% of experiment dependency. *pearson.sbm* and *scca.sbm* are applied to matrices at discretization levels $\{0.3, 0.4, \ldots, 0.8\}$ (from left to right on the curve).



(a) (b)

Figure 3.6: Heatmaps of $\bar{\boldsymbol{A}}$ at (a) 50% subsampling level and (b) 95% subsampling level.

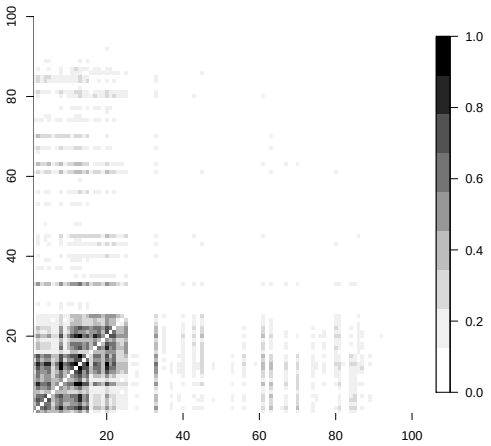


Figure 3.7: Heatmap of $\bar{\boldsymbol{A}}$ with 50% subsampling.

Chapter 4

Likelihood-based model selection for stochastic block models

4.1 Overview

In Chapter 3, we applied SBM to detect communities in gene networks. In this Chapter, we analyze this popular random graph model from a theoretical perspective and study the problem of model selection. As mentioned in Chapter 1, much effort has been devoted to estimating the model parameters, whereas the issue of choosing the block number is less explored. We directly address the challenges involved in analyzing the asymptotic distribution of the maximum log likelihood function under model misspecification. We show the log likelihood ratio statistic is asymptotically normal in the case of underfitting. Although obtaining an explicit asymptotic distribution of the statistic in the case of overfitting is much more challenging, we have still derived its order of convergence and subsequently shown these two cases of misspecification can be separated with probability tending to one. We thus propose a model selection criterion taking the form of a penalized likelihood and show it is asymptotically consistent. Our conclusions remain valid for networks with average degree growing at a polylog rate in the semi-sparse regime. Computationally the likelihood can be approximated with the variational algorithm in [40], making this approach applicable to reasonably large networks. We also provide comparisons of its performance on simulated and real networks with other model selection approaches.

4.2 Results

Preliminaries

To set the notation for this Chapter, a SBM with K blocks on n nodes is defined as follows. A vector of latent labels $Z = (Z_1, \ldots, Z_n)$ is generated with Z_i taking integer values from $[K] = \{1, \ldots, K\}$ governed by a multinomial distribution with parameters $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$.

Given $Z_i = a$, $Z_j = b$, an adjacency matrix A is generated with

$$A_{i,j}|(Z_i = a, Z_j = b) \sim \text{Bernoulli}(H_{a,b}), \qquad i \neq j.$$

We consider a symmetric A with zero diagonal entries corresponding to an undirected graph, although our arguments generalize easily to directed graphs. H is a $K \times K$ symmetric matrix describing the connectivities within and between blocks. We denote the model parameters $\theta = (\pi, H)$ and let Θ_K be the parameter space of a K-block model,

$$\Theta_K = \{\theta \mid \pi \in (0,1)^K, \sum_{a=1}^{K} \pi_a = 1, H \in (0,1)^{K \times K}\}.$$

Throughout the paper, $\theta^* = (\pi^*, H^*)$ will denote the true generative parameter giving rise to an observed A. We will further parametrize H^* by $H^* = \rho_n S^*$, where the degree density ρ_n may be $\Omega(1)$ or going to zero at a rate $n\rho_n / \log n \to \infty$. We assume $\theta^* \in \Theta_K$ and H^* has no identical columns, meaning the underlying model has K blocks and it is identifiable in the sense that it cannot be further collapsed to a smaller model. $z = (z_1, \ldots, z_n) \in [K']^n$ represents another set of labels under a K'-block model with K' not necessarily equaling K. $g(A; \theta)$ is the likelihood function describing the distribution of A with parameter $\theta \in \Theta_{K'}$ and can be written as the sum of the complete likelihood function $f(z, A; \theta)$ associated with the labels $z \in [K']^n$:

$$g(A; \theta) = \sum_{z \in [K']^n} f(z, A; \theta), \tag{4.1}$$

where $f(z, A; \theta)$ takes the form

$$\begin{aligned} f(z, A; \theta) &= \left(\prod_{i=1}^{n} \pi_{z_i} \right) \left(\prod_{i<j} H_{z_i, z_j}^{A_{ij}} (1 - H_{z_i, z_j})^{1 - A_{ij}} \right) \\ &= \left(\prod_{a=1}^{K'} \pi_a^{n_a(z)} \right) \left(\prod_{a=1}^{K'} \prod_{b=1}^{K'} H_{a,b}^{O_{a,b}(z)} (1 - H_{a,b})^{n_{a,b}(z) - O_{a,b}(z)} \right)^{1/2} \end{aligned}$$

with count statistics

$$\begin{aligned} n_a(z) &= \sum_{i=1}^{n} \mathbb{I}(z_i = a), \quad n_{a,b}(z) = \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{I}(z_i = a, z_j = b) \\ O_{a,b}(z) &= \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{I}(z_i = a, z_j = b) A_{i,j}. \end{aligned}$$

g and f are invariant with respect to a permutation on the block labels, $\tau : [K'] \to [K']$, and its corresponding permutations on the node labels z and the parameters θ. Furthermore, let

$R(z)$ be the $K' \times K$ confusion matrix whose (k, a)-th entry is

$$R_{k,a}(z, Z) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(z_i = k, Z_i = a). \tag{4.2}$$

We take a likelihood-based approach toward model selection and first investigate whether different model choices can be separated using the log likelihood ratio

$$L_{K,K'} = \log \frac{\sup_{\theta \in \Theta_{K'}} g(A; \theta)}{\sup_{\theta \in \Theta_K} g(A; \theta)}. \tag{4.3}$$

Here the comparison is made between the correct K-block model and fitting a misspecified K'-block model.

In the following sections, we analyze the asymptotic distribution of $L_{K,K'}$ for $K' \neq K$. The main focus of analysis lies in handling the sum in (4.1) which contains an exponential number of terms. It has been shown in [12] that when $\theta \in \Theta_K$, $\sup_{\theta \in \Theta_K} g(A; \theta)$ is essentially equivalent to maximizing the complete likelihood corresponding to the correct labels Z, $\sup_{\theta \in \Theta_K} f(Z, A; \theta)$. In the next section, we first show an analogous result in the case of underfitting and use it to derive the asymptotic distribution of $L_{K,K'}$.

Underfitting

We start by considering $K' = K - 1$. Intuitively, a $(K-1)$-block model can be obtained by merging blocks in a K-block model. More specifically, given the correct labels $Z \in [K]^n$ and the corresponding block proportions $p = (p_1, \ldots, p_K)$, $p_a = n_a(Z)/n$, we define a merging operation $U_{a,b}(H^*, p)$ which combines blocks a and b in H^* by taking weighted averages with proportions in p. For example, for $H = U_{K-1,K}(H^*, p)$,

$$\begin{aligned}
H_{l,k} &= H^*_{l,k} \quad \text{for } 1 \leq l, k \leq K-2; \\
H_{l,K-1} &= \frac{p_l p_{K-1} H^*_{l,K-1} + p_l p_K H^*_{l,K}}{p_l p_{K-1} + p_l p_K} \quad \text{for } 1 \leq l \leq K-2; \\
H_{K-1,K-1} &= \frac{p^2_{K-1} H^*_{K-1,K-1} + 2 p_{K-1} p_K H^*_{K-1,K} + p^2_K H^*_{K,K}}{p^2_{K-1} + 2 p_{K-1} p_K + p^2_K}.
\end{aligned} \tag{4.4}$$

A schematic representation of H is given in Figure 4.1.

For consistency, when merging two blocks (a, b) with $b > a$, the new merged block will be relabeled a and all the blocks c with $c > b$ will be relabeled $c - 1$. Using this scheme, we also obtain the merged node labels $U_{a,b}(Z)$ and merged proportions $U_{a,b}(p)$ with $[U_{a,b}(p)]_a = p_a + p_b$.

Constraining the parameters to a smaller model results in a suboptimal likelihood and its distance from the likelihood associated with the correct model can be measured by the

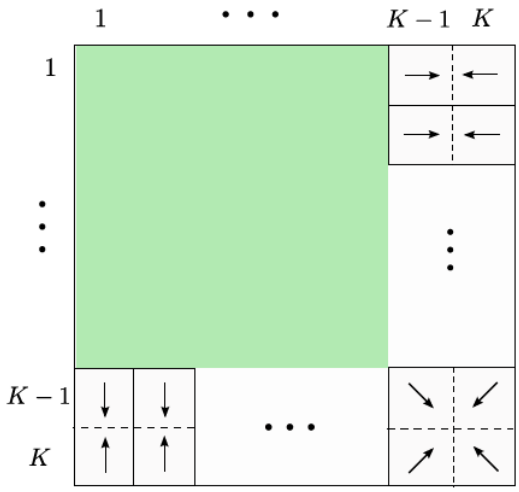


Figure 4.1: A schematic representation of how H^* is merged to give $H = U_{K-1,K}(H^*, p)$. The green area contains unchanged parameters and the arrows indicate where mergings occur.

Kullback-Leibler divergence, denoted $D_{KL}(\cdot\|\cdot)$. Let

$$\gamma_1(x) = x\log x + (1-x)\log(1-x),$$
$$\gamma_2(x) = x\log x - x.$$

and define

$$D_i(a,b) = \sum_{k,l=1}^{K-1} [U_{a,b}(\pi^*)]_k [U_{a,b}(\pi^*)]_l \gamma_i([U_{a,b}(H^*,\pi^*)]_{k,l}) \tag{4.5}$$

When $p = \pi^*$ and treating the labels Z as fixed parameters, denote $P_{A|Z,H^*}$ the probability distribution of A. Then the information loss incurred by the merging operation $U_{a,b}$ can be measured by

$$\begin{aligned} &D_{\mathrm{KL}}\left(P_{A|Z,H^*} \| P_{A|U_{a,b}(Z),U_{a,b}(H^*,\pi^*)}\right) \\ &= \begin{cases} \frac{n^2}{2}\left[\sum_{c,d=1}^K \pi_c^*\pi_d^*\gamma_1(H_{c,d}^*) - D_1(a,b)\right] + O(n), & \text{for } \rho_n = \Omega(1); \\ \frac{n^2\rho_n}{2}\left[\sum_{c,d=1}^K \pi_c^*\pi_d^*\gamma_2(H_{c,d}^*) - D_2(a,b)\right] + O(n^2\rho_n^2), & \text{for } \rho_n \to 0. \end{cases} \end{aligned} \tag{4.6}$$

Thus an optimal merging minimizing D_{KL} is essentially equivalent to maximizing $D_i(a,b)$.

We assume the following holds for θ^*:

Assumption 4.2.1. *A unique maximum exists for* $\max_{(a,b)} D_i(a,b)$.

This assumption is more of a notational convenience than necessity. From now on without loss of generality assume the maximum is achieved at $a = K-1$ and $b = K$, and denote $H' = U_{K-1,K}(H^*, \pi^*)$, $S' = H'/\rho_n$ and $Z' = U_{K-1,K}(Z)$. We also assume H' is identifiable in the sense that

Assumption 4.2.2. *H' has no identical columns.*

Thus the merged model cannot be collapsed further to a smaller model.

The next lemma argues $\sup_{\theta\in\Theta_{K-1}} g(A;\theta)$ is essentially dominated by the complete likelihood associated with the optimal merging.

Lemma 4.2.3. *Let $\mathcal{S}(z)$ be the set of labels which are equivalent up to a permutation τ, $\mathcal{S}(z) = \{\tau(z), \tau : [K-1] \to [K-1]\}$. Then*

$$\sum_{z\notin\mathcal{S}(Z')} \sup_{\theta\in\Theta_{K-1}} f(z, A;\theta) = \sup_{\theta\in\Theta_{K-1}} f(Z', A;\theta) o_P(1). \tag{4.7}$$

The proof is shown in Section 4.6.

This lemma provides a tractable bound on $\sup_{\theta\in\Theta_{K-1}} g(A;\theta)$, allowing the rest of the analysis to be carried out by usual Taylor expansion. Define

$$\mu_1(\theta^*) = \frac{1}{2}\left[D_1(K-1,K) - \sum_{c,d=1}^{K} \pi_c^*\pi_d^*\gamma_1(H^*_{c,d})\right]$$
$$\mu_2(\theta^*) = \mu_1 + \frac{1}{n}\left\{(\pi^*_{K-1}+\pi^*_K)\log(\pi^*_{K-1}+\pi^*_K) - \pi^*_{K-1}\log\pi^*_{K-1} - \pi^*_K\log\pi^*_K\right\}$$

The following theorem gives the asymptotic distribution of $L_{K,K-1}$, the proof of which is shown in Section 4.6.

Theorem 4.2.4. *Suppose the underlying model parameter generating A is $\theta^* = (\pi^*, H^*) \in \Theta_K$, then $L_{K,K-1}$ is asymptotically normal with*

$$\begin{aligned} n^{-3/2}L_{K,K-1} - \sqrt{n}\mu_1(\theta^*) &\xrightarrow{D} N(0,\sigma_1^2(\theta^*)), \qquad \text{if } \rho_n = \Omega(1); \\ \rho_n^{-1}n^{-3/2}L_{K,K-1} - \rho_n^{-1}\sqrt{n}\mu_2(\theta^*) &\xrightarrow{D} N(0,\sigma_2^2(\theta^*)), \qquad \text{if } \rho_n \to 0. \end{aligned} \tag{4.8}$$

Let $\mathcal{I}$ be the set of indices affected by the merge $U_{K-1,K}(H^*, \pi^*)$,

$$\mathcal{I} = \{(a,b) \in [K]^2 \mid K-1 \le a \le K \text{ or } K-1 \le b \le K\},$$

and $u(a)$ such that

$$u(a) = \begin{cases} a & \text{for } a \le K-2 \\ K-1 & \text{for } K-1 \le a \le K. \end{cases}$$

Define $d_i = (d_i(a,b))_{(a,b)\in\mathcal{I},a\le b}$ *as*

$$d_1(a,b) = H^*_{a,b}\log\frac{H'_{u(a),K-1}}{H^*_{a,b}} + (1-H^*_{a,b})\log\frac{1-H'_{u(a),K-1}}{1-H^*_{a,b}}$$
$$d_2(a,b) = S^*_{a,b}\log\frac{S'_{u(a),K-1}}{S^*_{a,b}} + (S'_{u(a),K-1} - S^*_{a,b}).$$

Denote $\Sigma(\pi^*)$ *the covariance matrix of a multinomial*(π^*) *distribution,* $B(x)$ *the Jacobi matrix of the vector valued function* $\xi(x_1,\ldots,x_K) = (\xi_{a,b})_{(a,b)\in\mathcal{I},a\le b}$*, where*

$$\xi_{a,b} = \begin{cases} x_a x_b \text{ for } a \neq b \\ \frac{x_a^2}{2} \text{ for } a = b. \end{cases}$$

The variance $\sigma_i(\theta^*)$ *is given by* $d_i^T B(\pi^*)\Sigma(\pi^*)B(\pi^*)^T d_i$ *for* $i = 1, 2$*.*

Remark 4.2.5. *(i) In general, underfitting a* $K^- < K$ *model will lead to the same type of limiting distribution under conditions similar to Assumptions 4.2.1 and 4.2.2, assuming the uniqueness of the optimal merging scheme and identifiability after merging. That is,*

$$\rho_n^{-1} n^{-3/2} L_{K,K^-} - \rho_n^{-1}\sqrt{n}\mu \xrightarrow{D} N(0,\sigma^2) \tag{4.9}$$

for some mean $\mu \sim C\rho_n$ *and variance* σ^2*. The proof will be similar but involve more tedious descriptions of how various merges can occur.*

(ii) The asymptotic distributions derived under the null distribution of a K*-block model suggest one might consider performing hypothesis testing directly to compare against an alternative simpler model. However, the asymptotic means depend on the true parameters, and its maximum likelihood estimate converges only at the rate* $\sqrt{n}$ *[12].*

(iii) Without Assumptions 4.2.1 and 4.2.2, it is easy to show

$$L_{K,K^-} \le -\Omega_P(n^2\rho_n), \tag{4.10}$$

where $\Omega(\cdot)$ *denotes asymptotic lower bound, using the method in proving Theorem 4.2.7.*

Overfitting

In the case of overfitting a K^+-block model with $K^+ > K$, deriving the asymptotic distribution of L_{K,K^+} is much more challenging. Intuitively, embedding a K-block model in a larger model can be achieved by appropriately splitting the labels Z and there are an exponential number of possible splits. We first show a result analogous to Lemma 4.2.3. However, the number of summands involved in $\sup_{\theta\in\Theta_{K^+}} g(A;\theta)$ remains exponential this time.

Recall that for $z \in [K^+]^n$, $R(z,Z)$ is the $K^+ \times K$ confusion matrix. We first define a subset $\mathcal{V}_{K^+} \in [K^+]^n$ such that

$$\mathcal{V}_{K^+} = \left\{ z \in [K^+]^n \mid \text{ there is at most one nonzero entry in every row of } R(z,Z) \right\}.$$

$\mathcal{V}_{K^+}$ is obtained by splitting of Z such that every block in z is always a subset of an existing block in Z. The next lemma shows it suffices to consider only the subclass of labels $\mathcal{V}_{K^+}$ in the sum $g(A;\theta)$, the proof of which is given in Section 4.6.

Lemma 4.2.6. *Suppose $\theta^* \in \Theta_K$, then*

$$\sum_{z\in[K^+]^n} \sup_{\theta\in\Theta_{K^+}} f(z,A;\theta) = (1+o_P(1)) \sum_{z\in\mathcal{V}_{K^+}} \sup_{\theta\in\Theta_{K^+}} f(z,A;\theta).$$

The lemma does not provide a direct simplification of the sum and suggests the reason why obtaining an asymptotic distribution for L_{K,K^+} is difficult. On the other hand, with appropriate concentration we can still derive the asymptotic order of the statistic.

Theorem 4.2.7. *Suppose $\theta^* \in \Theta_K$, then overfitting by a K^+-block model with $K^+ > K$ gives $L_{K,K^+} = O_P(n^{3/2}\rho_n^{1/2})$.*

The proof is provided in Section 4.6.

Model selection

The results in the previous sections lead us to construct a penalized likelihood criterion for selecting the optimal block number. The criterion is consistent in the sense that asymptotically it chooses the correct K with probability one. Define

$$\beta(K') = \sup_{\theta\in\Theta_{K'}} \log g(A;\theta) - N_{K'}B_n, \tag{4.11}$$

where B_n gives the order of the penalty term, and $N_{K'}$ is a strictly increasing sequence indexed by K' describing the complexity of the model. The optimal K_0 is such that

$$K_0 = \arg\max_{K'} \beta(K'). \tag{4.12}$$

Corollary 4.2.8. *For $K' < K$, setting $B_n = o(n^2\rho_n)$,*

$$\mathbb{P}_{\theta^*}(\beta(K') < \beta(K)) \to 1. \tag{4.13}$$

For $K' > K$, setting B_n such that $B_n n^{-3/2}\rho_n^{-1/2} \to \infty$,

$$\mathbb{P}_{\theta^*}(\beta(K') < \beta(K)) \to 1. \tag{4.14}$$

Proof. For $K' < K$, generalizing Theorem 4.2.4,

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\beta(K') < \beta(K)\right) \\
=&\mathbb{P}_{\theta^*}\left(n^{-3/2}\rho_n^{-1}\log\frac{\sup_{\theta\in\Theta_{K'}} g(A;\theta)}{\sup_{\theta\in\Theta_K} g(A;\theta)} - \sqrt{n}\rho_n^{-1}\mu \right. \\
&\qquad\qquad \left. < (N_{K'} - N_K)\frac{B_n}{n^{3/2}\rho_n} - \sqrt{n}\rho_n^{-1}\mu\right) \\
\to& 1, \qquad (4.15)
\end{aligned}$$

since $B_n = o(n^2\rho_n)$ and $-\rho_n^{-1}\mu \geq C(\theta^*)$ for some positive constant depending on θ^*. In general the same conclusion holds by Remark 4.2.5 (iii).

For $K' > K$, using Theorem 4.2.7,

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\beta(K') < \beta(K)\right) \\
=&\mathbb{P}_{\theta^*}\left(\frac{1}{n^{3/2}\rho_n^{1/2}}\log\frac{\sup_{\theta\in\Theta_{K'}} g(A;\theta)}{\sup_{\theta\in\Theta_K} g(A;\theta)} < (N_{K'} - N_K)\frac{B_n}{n^{3/2}\rho_n^{1/2}}\right) \\
\to& 1, \qquad (4.16)
\end{aligned}$$

when $B_n n^{-3/2}\rho_n^{-1/2} \to \infty$. □

Since the ratio of the upper bound $n^2\rho_n$ and the lower bound $n^{3/2}\rho_n^{1/2}$ tends to infinity, such a sequence B_n exists. Choosing B_n in this interval, we have $K_0 = K$ with probability tending to 1. However, we also note that for finite cases with moderate-sized n, $\sqrt{n}\rho_n^{-1}\mu$ in (4.15) is small, making it easy to over penalize with large B_n. At the same time, the lower bound in Theorem 4.2.7 is not tight and can be refined further.

We further assume the followings hold for tractable approximation.

Assumption 4.2.9. $\sum_{z:\in\mathcal{V}_{K^+}} \sup_{\theta\in\Theta_{K^+}} f(z, A;\theta) = O_P(e^{M_n})$, *where*

$$M_n = \max_{z\in\mathcal{V}_{K^+}} \sup_{\theta\in\Theta_{K^+}} \log f(z, A;\theta) \qquad (4.17)$$

Assumption 4.2.10. *The maximum is achieved in the set* $\mathcal{N}_{K^+} = \{z \in \mathcal{V}_{K^+} \mid n_k(z) \geq \epsilon n$ *for all* k, *for some* $\epsilon > 0, \}$.

Assumption 4.2.9 assumes a Laplace-type approximation holds for the sum, whereas Assumption 4.2.10 assumes the maximum can only be achieved on a loosely balanced block design. These assumptions together with Lemma 4.2.6 imply it remains to analyze the order of $\max_{z\in\mathcal{N}_{K^+}} \sup_{\theta\in\Theta_{K^+}} \log f(z, A;\theta)$. The following theorem shows the order of L_{K,K^+} can be refined to $O_P(1)$. The details can be found in Section 4.6.

Theorem 4.2.11. *Under Assumptions 4.2.9 and 4.2.10, L_{K,K^+} is of order $O_P(1)$ for $K^+ > K$.*

It follows then choosing B_n growing slightly faster than a constant will ensure consistency in the sense described in Corollary 4.2.8 . It can also be deduced from the proof that the order of L_{K,K^+} grows at most at the rate $(K^+ + 1)K^+/2$. Thus we choose a penalized likelihood of the following form,

$$\beta(K') = \sup_{\theta \in \Theta_{K'}} \log g(A; \theta) - \lambda \cdot \frac{K'(K'+1)}{2} \log n, \tag{4.18}$$

where the constant λ is a tuning parameter and does not affect the asymptotic properties of the criterion. It is not surprising that the penalty term has the same order as other BIC-type criteria (e.g. [40]) based on the complete likelihood assuming the node labels are fixed. Recall that in the underfitting case we have proved the likelihood is essentially equivalent to the complete likelihood corresponding to the appropriate labels. A similar equivalence also holds for the overfitting case by Lemma 4.2.6 and Assumption 4.2.9.

Approximation by variational likelihood

In practice, direct computations of the likelihood function $g(A; \theta)$ involves an exponential number of summands and quickly become intractable as n grows. In particular, the optimization over θ using the EM algorithm requires computing the conditional distribution of Z given A, which is not factorizable in this case. Variational methods tackle the true conditional distribution $f_{Z|A;\theta}$ with the mean field approximation, thus simplifying the local optimization at each iteration. The variational log likelihood $J(q, \theta; A)$ for a K'-block model is defined as

$$J(q, \theta; A) = -D_{KL}(q \| f_{Z|A;\theta}) + \log g(A; \theta), \tag{4.19}$$

where $q \in \mathcal{D}_{K'}$ is any product distribution with $q(z) = \prod_{i=1}^n q_i(z_i)$, $1 \le z_i \le K'$. The variational estimates $\hat{\theta}^{\text{VAR}}_{K'}$ is given by

$$\hat{\theta}^{\text{VAR}}_{K'} = \arg \max_{\theta \in \Theta_{K'}} \max_{q \in \mathcal{D}_{K'}} J(q, \theta; A),$$

which can be optimized using the EM algorithm in [40]. Also we note that $J(q, \theta; A)$ simplifies to

$$\begin{aligned} J(q, \theta; A) = &\sum_{i=1}^{n} \sum_{k=1}^{K'} q_i(k)(-\log q_i(k) + \log \pi(k)) \\ &+ \sum_{i<j} \sum_{k,l=1}^{K'} q_i(k) q_j(l) \left(A_{ij} \log H_{k,l} + (1 - A_{ij}) \log(1 - H_{k,l})\right) \end{aligned}$$

and hence can be easily evaluated.

We can replace the likelihood in (4.18) by the variational log likelihood J without changing its asymptotic performance. More precisely, the criterion with variational approximation

$$\beta^{\mathrm{VAR}}(K') = \sup_{\theta \in \Theta_{K'}} \sup_{q \in \mathcal{D}_{K'}} J(q, \theta; A) - \lambda \cdot \frac{K'(K'+1)}{2} \log n \tag{4.20}$$

is still asymptotically consistent. Noting that

(i) $\sup_{\theta \in \Theta_{K'}} \sup_{q \in \mathcal{D}_{K'}} J(q, \theta; A) \leq \sup_{\theta \in \Theta_{K'}} \log g(A; \theta)$;

(ii) $\sup_{\theta \in \Theta_K} \sup_{q \in \mathcal{D}_K} J(q, \theta; A) - \sup_{\theta \in \Theta_K} \log g(A; \theta) = O_P(1)$ as shown in [12],

it can be easily verified that (4.15) and (4.16) still hold.

4.3 Simulations

We first examined how well the normal limiting distribution approximated the empirical distribution of the statistic in the case of underfitting. Figure 4.2 plots the distribution of $n^{-3/2} L_{K,K-1}$ for $n = 200$ and $n = 500$ obtained from 200 replications for the following two scenarios:

(a) $K = 2$, $\pi^* = (0.4, 0.6)$, $H^* = \begin{pmatrix} 0.15 & 0.05 \\ & 0.01 \end{pmatrix}$;

(b) $K = 3$, $\pi^* = (0.4, 0.3, 0.3)$, $H^* = \begin{pmatrix} 0.2 & 0.1 & 0.1 \\ & 0.2 & 0.03 \\ & & 0.1 \end{pmatrix}$.

The log likelihoods are approximated by the variational EM algorithm initialized by regularized spectral clustering [44]. The solid curves are normal densities with mean $\mu_2(\theta^*)$ and $\sigma(\theta^*)$ given in Theorem 4.2.4. Even though the $O(n)$ term in $\mu_2(\theta^*)$ diminishes asymptotically for ρ_n going to 0 slowly, we found it essential to correct for the bias in the finite sample regimes above. In both cases, the convergence to the Gaussian shape appears faster than the convergence to the mean, and a bias exists for $n = 200$. When the network size reaches 500, the empirical distributions are well approximated by their limiting distribution. We note that the bias should not have an adverse effect on model selection since it is in the direction away from zero, making it easier to separate the two models.

Next we investigated how the success rate of the criterion (4.20) changes with respect to the tuning parameter λ. Figure 4.3 shows the fraction of the penalized likelihood selecting the correct K out of 50 trials for λ values varying between 0 and 4. The generative parameters for $K = 2$ and $K = 3$ are given in scenarios (a) and (b), and in addition a $K = 5$ model was generated with $\pi_i^* = 0.2$ for all i and the entries in H^* varying between 0.06 and 0.19. For $K = 2$ and $K = 3$, the penalized likelihood achieves reasonable success rate for λ smaller

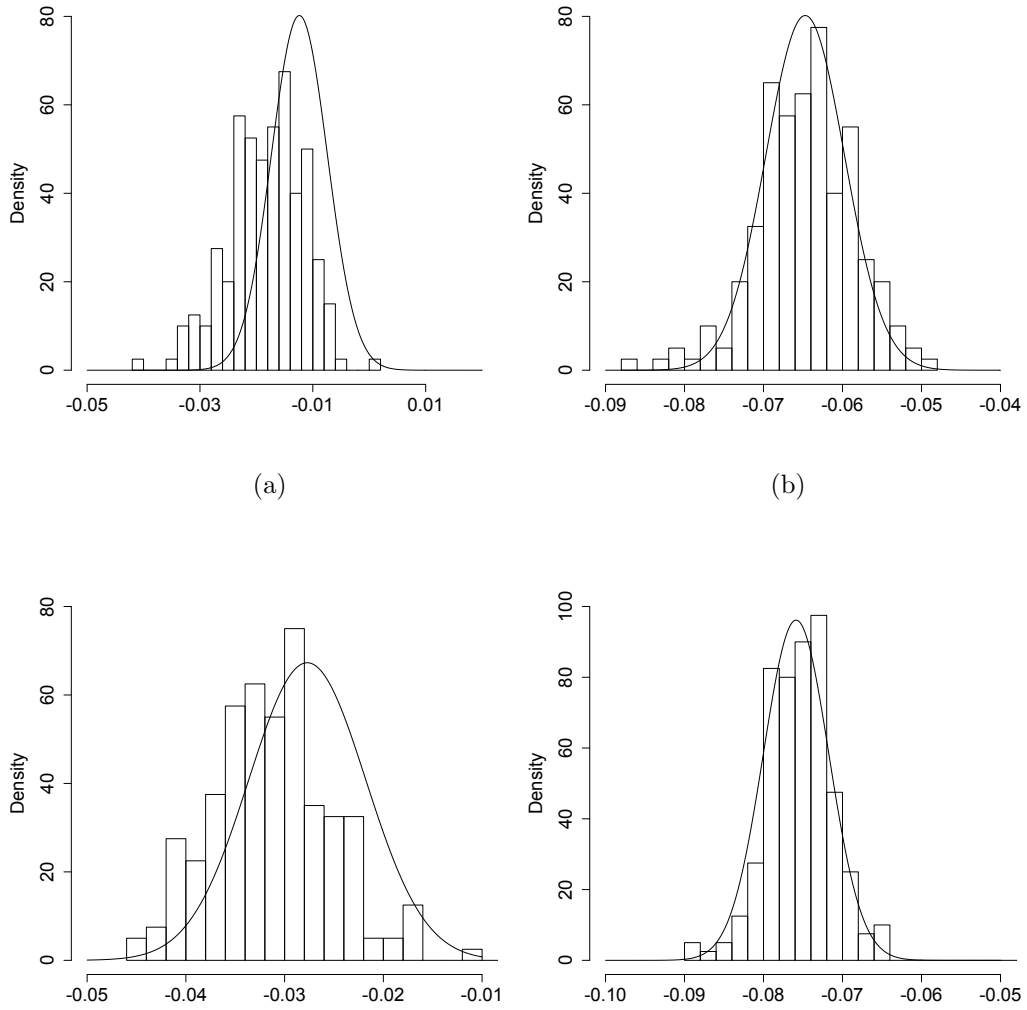
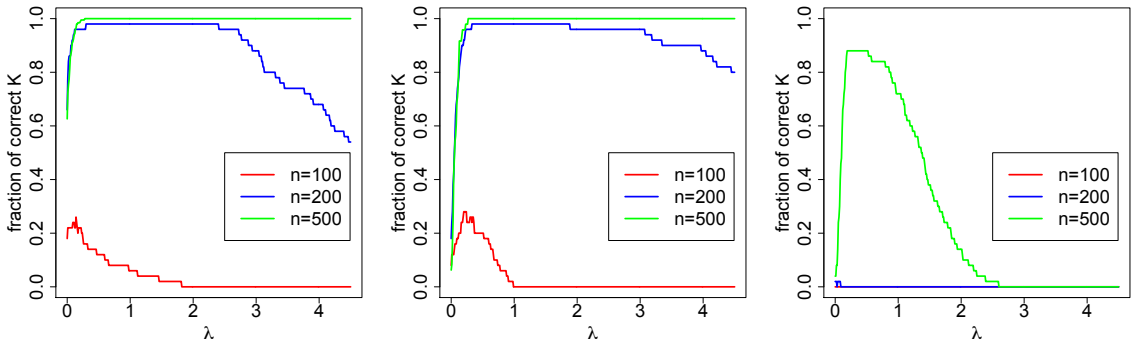


Figure 4.2: Empirical distributions of $n^{-3/2}L_{K,K-1}$ for (a), (b) K=2, π^* and H^* as described in scenario (a); (c), (d) K=3, π^* and H^* as described in scenario (b). $n = 200$ in (a) and (c); $n = 500$ in (b) and (d). The solid curves are normal densities with mean $\mu_2(\theta^*)$ and $\sigma(\theta^*)$ as given in Theorem 4.2.4.

than 3 when the network size reaches 200. When $n = 500$, the success rate appears robust to the choice of λ and is maintained at 1 for a wide range of values. For $K = 5$, however, it becomes difficult to select the correct K since the task of fitting also becomes harder as K increases.



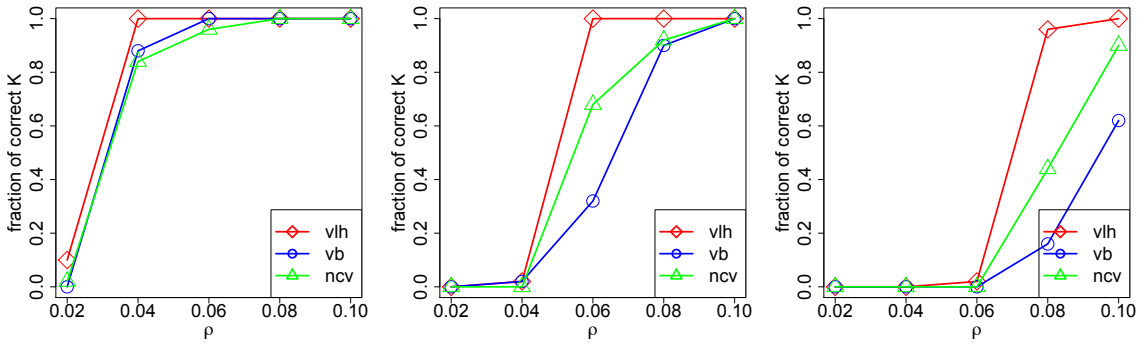
(a) (b) (c)

Figure 4.3: The fraction of the penalized likelihood with difference values of λ successfully choosing the correct K out of 50 iterations for (a) $K = 2$, π and H^* as described in scenario (a); (b) $K = 3$, π and H^* as described in scenario (b); (c) $K = 5$, $\pi_i = 0.2$ for all i, H^* with entries varying between 0.06 and 0.19.

To see how our criterion (denoted `vlh`) compares with other existing model selection methods, we fix $\lambda = 1$ and compare its success rate with variational Bayes ([54], denoted `vb`) and the 3-fold network cross validation method in [17] (denoted `ncv`). In Figure 4.4, these methods were implemented on 50 networks of size 500 with $K = 2, 3, 4$, $H^* = \rho S^*$, and $\rho \in \{0.02, 0.04, \ldots, 0.1\}$. The average degrees of these networks range from around 12 to 75. In general, the success rate of each method decreases as the networks become sparser and the number of blocks grows. Overall `vlh` outperforms the other two methods, and although not explicitly shown, the trends remain true for λ values between 0.25 and 2.

4.4 Real world networks

We first implemented our method along with `vb` and `ncv` on nine Facebook ego networks, collected and labeled by [57]. An ego network is created by extracting subgraphs formed on the neighbors of a central (ego) node, i.e. a network of connections between the ego's friends. Any isolated node was removed before analysis. The actual sizes of the networks and the number of communities selected by the three methods are shown in Table 4.1. The second row of the table shows the number of friend circles in every network with some individuals belonging to multiple circles, but not every individual possesses a circle label. The third row of the table shows the average degree of every network, which gives us a sense of the network density. These circle numbers give partial truth on how many communities there are in the networks. Overall, the penalized likelihood and `vb` tend to produce comparable community numbers, whereas `ncv` consistently favors small community numbers. The penalized likelihood approach is reasonably robust to the choice of λ on larger networks. Both



(a) (b) (c)

Figure 4.4: Comparison of the success rates of the penalized likelihood ($\lambda = 1$, `vlh`) with variational Bayes (`vb`) and network cross validation (`ncv`) when (a) $K = 2$, $\pi = (0.4, 0.6)$; (b) $K = 3$, $\pi = (0.3, 0.3, 0.4)$; (c) $K = 4$, $\pi_i = 0.25$ for all i. In all the cases, $H^* = \rho S^*$, where $\rho \in \{0.02, 0.04, \ldots, 0.1\}$, the diagonal elements of S^* equal 2 and the off diagonal elements equal 1.

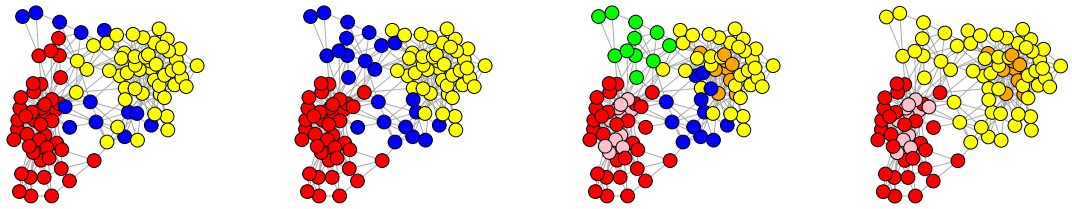
the penalized likelihood and `vb` tend to underestimate the community number on smaller networks with a large number of circles and a small average degree, reflecting the difficulty in fitting a large K-block model on small networks.

# Non-isolated vertices	333	1034	224	150	168	61	786	534	52
# Circles	24	9	14	7	13	13	17	32	17
Average degree	15	52	29	23	20	9	36	18	6
Optimal K, $(\lambda = 1/4)$	13	15	15	10	13	9	20	14	6
$(\lambda = 1/2)$	13	15	13	10	9	8	20	14	6
$(\lambda = 1)$	10	15	13	7	9	6	20	14	3
Optimal K, `vb`	11	24	16	9	11	6	25	23	6
Optimal K, `ncv`	3	6	4	2	4	2	2	2	3

Table 4.1: Facebook ego networks and the number of communities selected by the three methods, the penalized likelihood with three choices of λ.

We also experimented these methods on the political book network [70], which consists of 105 books and their edges representing co-purchase information from Amazon. Figure 4.5 (a) shows the manual labeling of the books based on their political orientations being either conservative, liberal or neutral. (b) and (c) show the community structures obtained by our method with three choices of λ. When $\lambda = 2$, the method selected $K = 3$ with the clustering of the nodes being close to the truth. With the other two smaller λ values, the

method selected $K = 6$ and the clustering further splits each of the communities obtained previously into two. `vb` found four communities but merged two clusters in (a) into one. `ncv` again produced the smallest K value with $K = 2$.



(a) (b) (c) (d)

Figure 4.5: Communities in 105 political books based on (a) manually curated ground truth; (b) penalized likelihood with $\lambda = 2$; (c) penalized likelihood with $\lambda = 1, 1/2$; (d) `vb`.

4.5 Discussion

In this paper, we have studied the problem of selecting the community number under a regular stochastic block model, allowing the average degree to grow at a polylog rate and the true block number being fixed. Using techniques similar to [12], we have shown the log likelihood ratio statistic has an asymptotic normal distribution when a smaller model with fewer blocks is specified. In the case of misfitting a larger model, we have obtained the convergence rate for the statistic. Combining these results we arrive at a likelihood-based model selection criterion that is asymptotically consistent. For finite-sized networks, we have further refined the bound for the statistic in the overfitting case under reasonable assumptions to correct for the possibility of over-penalizing.

There are a number of open problems for future work. (i) It would be desirable to have a data-driven approach to select the tuning parameter λ. Similar to other AIC and BIC-type criteria under standard models, the choice of this constant does not affect the asymptotic consistency of the criterion. Our analysis and simulation suggest small λ is often preferred to avoid over-penalizing. However, it is less clear on the real networks which λ is optimal. Some form of a cross-validation method would seem appropriate for this purpose. (ii) It would be interesting to investigate whether the results can be extended to other block model variants, such as degree-corrected SBM [45] and overlapping SBM [1]. (iii) We have performed our analysis with fixed block number as the number of nodes tends to infinity. However, in practice the number of communities is also likely to grow as a network expands [20], especially when we view block models as histogram approximations for more general models [11, 108]. [72] has provided some analysis on the maximum number of blocks detectable

for a given SBM graph with fixed labels. In general as more time-course network data become available in biology, social science, and many other domains, incorporating dynamic features of community structures into network modeling will remain an interesting direction to explore.

4.6 Proofs of lemmas and theorems

In this section, we prove all the lemmas and theorems in the main paper. Denote $\mu_n = n^2\rho_n$, the total number of edges $L = \sum_{i=1}^n \sum_{j=i+1}^n A_{i,j}$, and $N(z) = (n_{k,l}(z))_{1\le k,l\le K'}$. For two sets of labels z and y, $|z-y| = \sum_{i=1}^n \mathbb{I}(z_i \neq y_i)$. $\|\cdot\|_\infty$ denotes the maximum norm of a matrix. We abbreviate $R(z,Z)S^*R(z,Z)^T$ as $RS^*R^T(z)$. $C, C_1, C_2, \ldots$ are constants which might be different at each occurrence. The following concentration inequalities bound the variations in A and will be used throughout the section.

Lemma 4.6.1. *Suppose $z \in [K']^n$ and define $X(z) = O(z)/\mu_n - RS^*R^T(z)$. For $\epsilon \le 3$,*

$$\mathbb{P}\left(\max_{z\in[K']^n} \|X(z)\|_\infty \ge \epsilon\right) \le 2(K')^{n+2}\exp\left(-\frac{1}{4(\|S^*\|_\infty+1)}\epsilon^2\mu_n\right). \tag{4.21}$$

Let $y \in [K']^n$ be a fixed set of labels, then for $\epsilon \le 3m/n$,

$$\begin{aligned}&\mathbb{P}\left(\max_{z:|z-y|\le m} \|X(z)-X(y)\|_\infty > \epsilon\frac{m}{n}\right)\\ \le& 2\binom{n}{m}(K')^{m+2}\exp\left(-\frac{n\epsilon^2\mu_n}{4m(4\|S^*\|_\infty+1)}\right).\end{aligned} \tag{4.22}$$

Proof. The proof follows from [12] with minor modifications for general K'-block models and correcting for the zero diagonal in A. □

Recall that

$$\begin{aligned}\gamma_1(x) &= x\log x + (1-x)\log(1-x),\\ \gamma_2(x) &= x\log x - x.\end{aligned}$$

Define $F_i(M,t)$, $i = 1, 2$, as

$$F_i(M,t) = \sum_{k,l=1}^{K'} t_{k,l}\gamma_i\left(\frac{M_{k,l}}{t_{k,l}}\right), \tag{4.23}$$

Then the log of the complete likelihood can be expressed as

$$\sup_{\theta\in\Theta_{K'}} \log f(z,A;\theta) = n\sum_{k=1}^{K'}\alpha(n_k(z)/n) + \frac{n^2}{2}F_1\left(O(z)/n^2, N(z)/n^2\right), \tag{4.24}$$

where $\alpha(x) = x\log(x)$. Noting the first term is of smaller order compared to the second term, and the conditional expectation of the argument in γ_1 given Z is $[RH^*R^T(z)]_{k,l}/[R\mathbf{1}\mathbf{1}^TR^T(z)]_{k,l}$ and $[RS^*R^T(z)]_{k,l}/[R\mathbf{1}\mathbf{1}^TR^T(z)]_{k,l}$ for γ_2 (up to a diagonal difference) with fluctuation bounded by Lemma 4.6.1, we will focus on analyzing the conditional expectation

$$G_1(R(z), H^*) = \sum_{k,l=1}^{K'} [R\mathbf{1}\mathbf{1}^TR^T(z)]_{k,l}\gamma_1\left(\frac{[RH^*R^T(z)]_{k,l}}{[R\mathbf{1}\mathbf{1}^TR^T(z)]_{k,l}}\right) \quad \text{for } \rho_n = \Omega(1), \tag{4.25}$$

$$G_2(R(z), H^*) = \sum_{k,l=1}^{K'} ([R\mathbf{1}\mathbf{1}^TR^T(z)]_{k,l}\gamma_2\left(\frac{[RS^*R^T(z)]_{k,l}}{[R\mathbf{1}\mathbf{1}^TR^T(z)]_{k,l}}\right) \quad \text{for } \rho_n \to 0. \tag{4.26}$$

The following lemma shows in the case of underfitting a $(K-1)$-block model, to maximize G_i over different configurations of $R(z, Z)$ with given Z, it suffices to consider the merging scheme described in Section 4.2 by combining two existing blocks in Z.

Lemma 4.6.2. *Given the true labels Z with block proportions $p = n(Z)/n$, maximizing the function $G_1(R(z), H^*)$ over R achieves its maximum in the label set*

$$\{z \in [K-1]^n \mid \text{ there exists } \tau \text{ such that } \tau(z) = U_{a,b}(Z), 1 \le a < b \le K, \}$$

where $U_{a,b}$ merges Z_i with labels a and b.

Furthermore, suppose z_0 gives the unique maximum (up to permutation τ), for all R such that $R \ge 0, R^T\mathbf{1} = p$,

$$\left.\frac{\partial G_1((1-\epsilon)R(z_0) + \epsilon R, H^*)}{\partial \epsilon}\right|_{\epsilon=0+} < -C < 0 \tag{4.27}$$

for $\rho_n = \Omega(1)$. The same conclusions hold for $G_2(R(z), S^*)$.

Proof. Treating R as a $(K-1)\times K$-dimensional vector, it is easy to check $G_1(\cdot, H^*)$ is a convex function. Furthermore, since $R \ge 0$, $R^T\mathbf{1} = p$, the domain is part of a convex polyhedron $P_R = \{R \in \mathbb{R}^{K(K-1)} \mid R \ge 0, R^T\mathbf{1} = p\}$. Therefore the maximum is attained at the vertices of P_R, that is R^{vert} such that for every a, exactly one $R^{vert}_{k,a}$, $(1 \le k \le K-1)$ is nonzero. This is equivalent to assigning all $Z_i \in [K]$ with the same label into one group with a new label in $[K-1]$. Let $u : [K] \longrightarrow [K-1]$ be the function specified by R^{vert}, then

$$G_1(R^{vert}, H^*) = \sum_{k,l} \sum_{\substack{a\in u^{-1}(k),\\ b\in u^{-1}(l)}} p_a p_b \gamma_1\left(\frac{\sum_{\substack{a\in u^{-1}(k),\\ b\in u^{-1}(l)}} H^*_{a,b}p_ap_b}{\sum_{\substack{a\in u^{-1}(k),\\ b\in u^{-1}(l)}} p_ap_b}\right). \tag{4.28}$$

Note that there exists at least one $l \in [K-1]$ such that $|\{u^{-1}(l)\}| > 1$, and $\{u^{-1}(k), k \in [K-1]\}$ forms a partition on $[K]$. By strict convexity of γ_1 and identifiability of H^*, to

maximize G_1 it suffices to consider merging two of the labels in $[K]$ and mapping the other labels to the remaining labels in $[K-1]$ in a one-to-one relationship.

The second part of the lemma holds since it is easy to see when the maximum is unique, the derivative of the G_1 at the optimal vertex is bounded away from 0 in all directions. The same arguments apply to G_2. □

Noting that when $p = \pi^*$, G_i evaluated at $R(U_{a,b}(Z))$ is equal to D_i defined in (4.5), it is easy to see Assumptions 4.2.1 and 4.2.2 guarantees the maximum is unique. We will now prove Lemma 4.2.3.

Proof of Lemma 4.2.3. Taking the log of the complete likelihood,

$$\begin{aligned}&\sup_{\theta\in\Theta_{K-1}} \log f(z, A; \theta)\\ =&n\sum_{k=1}^{K-1}\alpha(n_k(z)/n) + \frac{n^2}{2}F_1\left(O(z)/n^2, N(z)/n^2\right).\end{aligned} \tag{4.29}$$

By concentration of p_k, it suffices to consider $\{\|p-\pi^*\|_\infty < \eta\}$, where η is small enough that Z' remains the unique maximizer of $G_1(R(z), H^*)$ and $G_2(R(z), S^*)$, and distribution conditional on Z.

Using techniques similar to [12], we prove this by considering z far away from Z' and close to Z' (up to permutation τ). Let δ_n be a sequence converging to 0 slowly. Define

$$I_{\delta_n} = \{z \in [K-1]^n : G_1(R(z), H^*) - G_1(R(Z'), H^*) < -\delta_n\}.$$

First by (4.21) in Lemma 4.6.1, for $\epsilon_n \to 0$ slowly,

$$\begin{aligned}&\left|F_1\left(O(z)/n^2, N(z)/n^2\right) - G_1(R(z), H^*)\right|\\ \le& C\cdot\sum_{k,l}\left|O_{k,l}(z)/n^2 - (RH^*R^T(z))_{k,l}\right| + O(n^{-1})\\ =&o_P(\epsilon_n)\end{aligned} \tag{4.30}$$

since γ_1 is Lipschitz on any interval bounded away from 0 and 1 and $\min H^* = \Omega(1)$. For $z \in I_{\delta_n}$ and $\rho_n = \Omega(1)$,

$$\begin{aligned}\sum_{z\in I_{\delta_n}}\sup_{\theta\in\Theta_{K-1}} e^{\log f(z,A;\theta)} &\le \sup_{\theta\in\Theta_{K-1}} f(Z', A; \theta)(K-1)^n e^{O(n)+o_P(n^2\epsilon_n)-n^2\delta_n}\\ &= \sup_{\theta\in\Theta_{K-1}} f(Z', A; \theta)o_P(1)\end{aligned} \tag{4.31}$$

choosing $\delta_n \to 0$ slowly enough such that $\delta_n/\epsilon_n \to \infty$. Similarly for $\rho_n \to 0$, define

$$J_{\delta_n} = \{z \in [K-1]^n : G_2(R(z), S^*) - G_2(R(Z'), S^*) < -\delta_n\}.$$

Note that in this case, for $\epsilon_n \to 0$ slowly,

$$\begin{aligned} & F_1\left(O(z)/n^2, N(z)/n^2\right) \\ =& 2\log \rho_n L/n^2 + \rho_n F_2\left(O(z)/\mu_n, N(z)/n^2\right) + O_P(\rho_n^2) \\ =& 2\log \rho_n L/n^2 + \rho_n G_2\left(R(z), S^*\right) + o_P(\rho_n \epsilon_n) + O_P(\rho_n^2), \end{aligned} \tag{4.32}$$

by (4.21) and the fact that γ_2 is Lipschitz on any interval bounded away from 0 and 1 and $\min S^* > 0$. Then for $z \in J_{\delta_n}$,

$$\begin{aligned} & \sum_{z \in J_{\delta_n}} \sup_{\theta \in \Theta_{K-1}} e^{\log f(z,A;\theta)} \\ \leq & \sup_{\theta \in \Theta_{K-1}} f(Z', A; \theta)(K-1)^n e^{O(n)+O_P(\mu_n\rho_n)+o_P(\mu_n\epsilon_n)-\mu_n\delta_n} \\ = & \sup_{\theta \in \Theta_{K-1}} f(Z', A; \theta) o_P(1). \end{aligned} \tag{4.33}$$

choosing $\epsilon_n \to 0$, $\delta_n \to 0$ slowly enough.

For $z \notin J_{\delta_n}$, $|G_2(R(z), H^*) - G_2(R(Z'), H^*)| \to 0$. Let $\bar{z} = \min_\tau |\tau(z) - Z'|$. Since the maximum is unique up to τ, $\|R(\bar{z}) - R(Z')\|_\infty \to 0$ and $|\sum_k \alpha(n_k(\bar{z})/n) - \sum_k \alpha(n_k(Z')/n)| \to 0$.

By (4.22),

$$\begin{aligned} & \mathbb{P}\left(\max_{z \notin \mathcal{S}(Z')} \|X(\bar{z}) - X(Z')\|_\infty > \epsilon |\bar{z} - Z'|/n \right) \\ \leq & \sum_{m=1}^{n} \mathbb{P}\left(\max_{z: z=\bar{z}, |\bar{z}-Z'|=m} \|X(z) - X(Z')\|_\infty > \epsilon \frac{m}{n} \right) \\ \leq & \sum_{m=1}^{n} 2(K-1)^{K-1} n^m (K-1)^{m+2} \exp\left(-C\frac{m\mu_n}{n}\right) \to 0. \end{aligned} \tag{4.34}$$

It follows for $|\bar{z} - Z'| = m$, $z \notin J_{\delta_n}$,

$$\begin{aligned} \left\| \frac{O(\bar{z})}{\mu_n} - \frac{O(Z')}{\mu_n} \right\|_\infty &= o_P(1)\frac{|\bar{z} - Z'|}{n} + \|RS^*R^T(\bar{z}) - RS^*R^T(Z')\|_\infty \\ &\geq \frac{m}{n}(C + o_P(1)). \end{aligned} \tag{4.35}$$

Observe $\|O(Z')/\mu_n - RS^*R(Z')\|_\infty = o_P(1)$ by Lemma 4.6.2, $N(Z')/n^2 = R\mathbf{1}\mathbf{1}^T R^T(Z') + o(1)$ on $\{\|p - \pi^*\|_\infty < \eta\}$, and $F_2(\cdot,\cdot)$ has continuous derivative in the neighborhood of $(O(Z')/\mu_n, N(Z')/n^2)$. Using (4.27) in Lemma 4.6.2,

$$\left. \frac{\partial F_2\left((1-\epsilon)\frac{O(Z')}{\mu_n} + \epsilon M, (1-\epsilon)\frac{N(Z')}{n^2} + \epsilon t\right)}{\partial \epsilon} \right|_{\epsilon = 0^+} < -\Omega_P(1) < 0$$

for (M,t) in the neighborhood of $(O(Z')/\mu_n, N(Z')/n^2)$. Hence

$$\begin{aligned} & F_2\left(O(\bar{z})/\mu_n, N(\bar{z})/n^2\right) - F_2\left(O(Z')/\mu_n, N(Z')/n^2\right) \\ \leq & -\Omega_P(1)\frac{m}{n}. \end{aligned} \tag{4.36}$$

We have

$$\begin{aligned} & \sup_{\theta\in\Theta_{K-1}} \log f(z,A;\theta) - \sup_{\theta\in\Theta_{K-1}} \log f(Z',A;\theta) \\ \leq & n\left|\sum_{k=1}^{K-1} \alpha(n_k(\bar{z})/n) - \alpha(n_k(Z')/n)\right| \\ & + n^2\left(F_1(O(\bar{z})/\mu_n, N(z)/n^2) - F_1(O(Z')/\mu_n, n(Z')/n^2)\right) \\ \leq & \left(O(n) + o_P(\mu_n) - \Omega_P(\mu_n)\right)\frac{m}{n} \\ = & -\Omega_P(\mu_n)\frac{m}{n} \end{aligned} \tag{4.37}$$

using (4.32) and (4.36). We can conclude

$$\begin{aligned} & \sum_{z\notin J_{\delta_n}, z\neq\tau(Z')} \sup_{\theta\in\Theta_{K-1}} e^{\log f(z,A;\theta)} \\ \leq & \sup_{\theta\in\Theta_{K-1}} f(Z',A;\theta) \sum_{m=1}^{n} (K-1)^{K-1} n^m (K-1)^m e^{-\Omega(\mu_n)m/n} \\ = & \sup_{\theta\in\Theta_{K-1}} f(Z',A;\theta) o_P(1) \end{aligned} \tag{4.38}$$

The bounds (4.33) and (4.38) yield (4.7). The case for $\rho_n = \Omega(1)$ can be shown in a similar way. □

Now Theorem 4.2.4 follows by Taylor expansion.

Proof of Theorem 4.2.4. First note that

$$\begin{aligned} L_{K,K-1} &= \log\frac{\sup_{\theta\in\Theta_{K-1}} g(A;\theta)}{g(A;\theta^*)} - \log\frac{\sup_{\theta\in\Theta_K} g(A;\theta)}{g(A;\theta^*)} \\ &= \sup_{\theta\in\Theta_{K-1}} \log\left[\frac{g(A;\theta)}{f(Z,A;\theta^*)}\cdot\frac{f(Z,A;\theta^*)}{g(A;\theta^*)}\right] + O_P(1) \\ &= \sup_{\theta\in\Theta_{K-1}} \log\frac{g(A;\theta)}{f(Z,A;\theta^*)} + O_P(1) \end{aligned} \tag{4.39}$$

by a consequence of Theorem 1 and Lemma 3 in [12]. Noting that $\sup_{\theta\in\Theta_{K-1}} f(Z', A; \theta)$ is uniquely maximized at (omitting the argument Z)

$$\begin{aligned}
&\hat{\pi}_a = \frac{n_a}{n} = \pi_a^* + O_P(n^{-1/2}) \text{ for } 1 \le a \le K-2, \quad \hat{\pi}_{K-1} = \frac{n_{K-1}+n_K}{n} \\
&\hat{H}_{a,b} = \frac{O_{a,b}}{n_{a,b}} = H^*_{a,b} + O_P(\sqrt{\rho_n} n^{-1}) \text{ for } 1 \le a \le b \le K-2, \\
&\hat{H}_{a,K-1} = \frac{O_{a,K-1}+O_{a,K}}{n_{a,K-1}+n_{a,K}} = H'_{a,K-1} + O_P(\sqrt{\rho_n} n^{-1}) \text{ for } 1 \le a \le K-2, \\
&\hat{H}_{K-1,K-1} = \frac{\sum_{a=K-1}^{K}\sum_{b=a}^{K} O_{a,b}}{\sum_{a=K-1}^{K}\sum_{b=a}^{K} n_{a,b}} = H'_{K-1,K-1} + O_P(\sqrt{\rho_n} n^{-1}),
\end{aligned} \tag{4.40}$$

and Assumption 4.2.2 the merged H' is identifiable, we have

$$\frac{\sup_{\theta\in\Theta_{K-1}} \sum_{z\in\mathcal{S}(Z')} f(z, A; \theta)}{\sup_{\theta\in\Theta_{K-1}} f(Z', A; \theta)} = 1 + o_P(1).$$

Combined with Lemma 4.2.3

$$\begin{aligned}
&\sup_{\theta\in\Theta_{K-1}} \log \frac{g(A;\theta)}{f(Z, A; \theta^*)} \\
=& \sup_{\theta\in\Theta_{K-1}} \log \frac{f(Z', A; \theta)}{f(Z, A; \theta^*)} + o_P(1).
\end{aligned} \tag{4.41}$$

We will check the expansion for the case $\rho_n \to 0$; the case $\rho_n = \Omega(1)$ can be shown in the

same way.

$$
\begin{aligned}
& n^{-3/2}\rho_n^{-1} \sup_{\theta\in\Omega_{K-1}} \log \frac{g(A;\theta)}{f(Z,A;\theta^*)} \\
=& n^{-3/2}\rho_n^{-1} \sup_{\theta\in\Omega_{K-1}} \log \frac{f(Z',A;\theta)}{f(Z,A;\theta^*)} + o_P(1) \\
=& n^{-3/2}\rho_n^{-1} \left\{ n\sum_{a=1}^{K-1}\alpha(\hat{\pi}_a) + \sum_{a=1}^{K-2}\sum_{b=a}^{K-2} n_{a,b}\gamma_1(\hat{H}_{a,b}) + \sum_{a=1}^{K-2}(n_{a,K-1}+n_{a,K})\gamma_1(\hat{H}_{a,K-1}) \right. \\
& \quad + \frac{1}{2}\sum_{a=K-1}^{K}\sum_{b=K-1}^{K} n_{a,b}\gamma_1(\hat{H}_{K-1,K-1}) \\
& \quad \left. - \sum_{a=1}^{K} n_a \log\pi_a^* - \frac{1}{2}\sum_{a=1}^{K}\sum_{b=1}^{K}\left(O_{a,b}\log\frac{H_{a,b}^*}{1-H_{a,b}^*} + n_{a,b}\log(1-H_{a,b}^*)\right)\right\} + o_P(1) \\
=& n^{-1/2}\rho_n^{-1}\left[\alpha(\pi_{K-1}^*+\pi_K^*) - \alpha(\pi_{K-1}^*) - \alpha(\pi_K^*)\right] \\
& \quad n^{-3/2}\rho_n^{-1}\frac{1}{2}\sum_{(a,b)\in\mathcal{I}}\left(O_{a,b}\log\frac{H'_{u(a),u(b)}(1-H_{a,b}^*)}{(1-H'_{u(a),u(b)})H_{a,b}^*} + n_{a,b}\log\frac{1-H'_{u(a),u(b)}}{1-H_{a,b}^*}\right) + o_P(1)
\end{aligned}
\tag{4.42}
$$

It is easy to see the expectation of this term is $\rho_n^{-1}\sqrt{n}\mu_2$, we have

$$
\begin{aligned}
& n^{-3/2}\rho_n^{-1} \sup_{\theta\in\Omega_{K-1}} \log \frac{g(A;\theta)}{f(Z,A;\theta^*)} - \sqrt{n}\rho_n^{-1}\mu_2 \\
=& \frac{1}{2n^{3/2}\rho_n}\sum_{(a,b)\in\mathcal{I}}\left[(O_{a,b}-\mathbb{E}(O_{a,b}))\log\frac{H'_{u(a),u(b)}(1-H_{a,b}^*)}{(1-H'_{u(a),u(b)})H_{a,b}^*} + (n_{a,b}-\mathbb{E}(n_{a,b}))\log\frac{1-H'_{u(a),u(b)}}{1-H_{a,b}^*}\right] \\
& + o_P(1) \\
=& \frac{1}{2n^{3/2}\rho_n}\sum_{(a,b)\in\mathcal{I}}\left\{ (n_{a,b}-\mathbb{E}(n_{a,b}))\left[H_{a,b}^*\log\frac{H'_{u(a),u(b)}(1-H_{a,b}^*)}{(1-H'_{u(a),u(b)})H_{a,b}^*} + \log\frac{1-H'_{u(a),u(b)}}{1-H_{a,b}^*}\right]\right\} \\
& + o_P(1) \\
\xrightarrow{D}& N(0,\sigma_2^2(\theta^*)),
\end{aligned}
\tag{4.43}
$$

where the form of $\sigma_2^2(\sigma^*)$ can be checked by Taylor expansion and the delta method.

□

Proof of Lemma 4.2.6. The proof follows using arguments similar to Lemma 4.2.3. Note that in this case $G_1(R(z),H^*)$ is maximized at any $z\in\mathcal{V}_{K^+}$ with the value $\sum_{a,b}p_ap_b\gamma_1(H_{a,b}^*)$ (or $\sum_{a,b}p_ap_b\gamma_2(S_{a,b}^*)$ for $G_2(R(z),S^*)$).

It suffices to discuss the case $\rho_n \to 0$. Denote the optimal $G^* := \sum_{a,b} p_a p_b \gamma_2(S^*_{a,b})$, define similarly to Lemma 4.2.3

$$J_{\delta_n} = \{z \in [K^+]^n : G_2(R(z), S^*) - G^* < -\delta_n\}$$

for $\delta_n \to 0$ slowly enough. It is easy to see

$$\sum_{z \in J_{\delta_n}} \sup_{\theta \in \Theta_{K^+}} f(z, A; \theta) \leq \sup_{\theta \in \Theta_{K^+}} f(z_0, A; \theta) o_P(1)$$

for any $z_0 \in \mathcal{V}_{K^+}$.

Next note that treating $R(z)$ as a vector, $\{R(z) \mid z \in \mathcal{V}_{K^+}\}$ is a subset of the union of some of the $K^+ - K$ faces of the polyhedron P_R. For every $z \notin J_{\delta_n}, z \notin \mathcal{V}_{K^+}$, let $z_\perp$ be such that $R(z_\perp) := \min_{R(z_0): z_0 \in \mathcal{V}_{K^+}} \|R(z) - R(z_0)\|_2$. $R(z) - R(z_\perp)$ is perpendicular to the corresponding $K^* - K$ face. Furthermore, this orthogonality implies the directional derivative of $G_2(\cdot, S^*)$ along the direction of $R(z) - R(z_\perp)$ is bounded away from 0. That is

$$\left.\frac{\partial G_2\left((1-\epsilon)R(z_\perp) + \epsilon R(z), S^*\right)}{\partial \epsilon}\right|_{\epsilon = 0^+} < -C$$

for some universal positive constant C. Similar to (4.37),

$$\sup_{\theta \in \Theta_{K^+}} \log f(z, A; \theta) - \sup_{\theta \in \Theta_{K^+}} \log f(z_\perp, A; \theta) \leq -\Omega_P(\mu_n)\frac{m}{n}$$
$$\sup_{\theta \in \Theta_{K^+}} f(z, A; \theta) \leq e^{-\Omega_P(\mu_n)\frac{m}{n}} \sup_{\theta \in \Theta_{K^+}} f(z_\perp, A; \theta)$$

where $|z - z_\perp| = m$. We have

$$\begin{aligned}
&\sum_{z \notin J_{\delta_n}, z \notin \mathcal{V}_{K^+}} \sup_{\theta \in \Theta_{K^+}} f(z, A; \theta) \\
\leq &\sum_{z \in \mathcal{V}_{K^+}} \sup_{\theta \in \Theta_{K^+}} f(z, A; \theta) \sum_{m=1}^{n} (K-1)^m n^m e^{-\Omega_P(\mu_n)\frac{m}{n}} \\
= &o_P(1) \sum_{z \in \mathcal{V}_{K^+}} \sup_{\theta \in \Theta_{K^+}} f(z, A; \theta).
\end{aligned}$$

Hence the claim follows. □

Proof of Theorem 4.2.7. First note

$$L_{K,K^+} = \log \frac{\sup_{\theta \in \Theta_{K^+}} g(A; \theta)}{f(Z, A; \theta^*)} + O_P(1),$$

where

$$\log \frac{\sup_{\theta\in\Theta_{K^+}} g(A;\theta)}{f(Z,A;\theta^*)} \geq \log \frac{\sup_{\theta\in\Theta_{K^+}} f(Z,A;\theta)}{f(Z,A;\theta^*)} = O_P(1). \tag{4.44}$$

Let $D(\cdot)$ be a diagonal matrix, upper bounding by the maximum,

$$\begin{aligned}
&\log \frac{\sup_{\theta\in\Theta_{K^+}} g(A;\theta)}{f(Z,A;\theta^*)} \\
\leq &\max_z \sup_{\theta\in\Theta_{K^+}} \log \frac{f(z,A;\theta)}{f(Z,A;\theta^*)} + n\log K^+ \\
= &\max_z \frac{n^2}{2}\left\{F_1\left(O(z)/n^2, N(z)/n^2\right) - F_1\left(D(p)H^*D(p), pp^T\right)\right\} + O_P(n) \\
\leq &\max_z \frac{n^2}{2}\left|F_1\left(O(z)/n^2, N(z)/n^2\right) - F_1(RH^*R^T(z), R\mathbf{1}\mathbf{1}^TR^T(z))\right| \\
&+ \max_z \frac{n^2}{2}\left[F_1(RH^*R^T(z), R\mathbf{1}\mathbf{1}^TR^T(z)) - F_1\left(D(p)H^*D(p), pp^T\right)\right] + O_P(n) \\
\leq &C\mu_n \max_z \left\|\frac{O(z)}{\mu_n} - RS^*R^T\right\|_\infty + O_P(n) \\
= &O_P(n^{3/2}\rho_n^{1/2})
\end{aligned} \tag{4.45}$$

using (4.21) in Lemma 4.6.1, and the fact that

$$\max_{z\in[K^+]^n} F_1(RH^*R^T(z), R\mathbf{1}\mathbf{1}^TR^T(z)) = F_1\left(D(p)H^*D(p), pp^T\right).$$

□

Next we prove Theorem 4.2.11.

Proof of Theorem 4.2.11. It remains to upper bound L_{K,K^+}. By Lemma 4.2.6 and Assumption 4.2.9, it suffices to consider

$$\begin{aligned}
&\max_{z\in\mathcal{V}_{K^+}} \sup_{\theta\in\Theta_{K^+}} \log f(z,A;\theta) - \sup_{\theta\in\Theta_K} \log g(A;\theta) \\
= &\max_{z\in\mathcal{V}_{K^+}} \sup_{\theta\in\Theta_{K^+}} \log f(z,A;\theta) - \log f(Z,A;\theta^*) + O_P(1).
\end{aligned}$$

It follows from the definition of $\mathcal{V}_{K^+}$ there exists a surjective function $h : [K^+] \to [K]$ describing the block assignments in $R(z, Z)$. We have

$$\begin{aligned}
&\max_{z\in\mathcal{V}_{K^+}} \sup_{\theta\in\Theta_{K^+}} \log f(z, A; \theta) - \log f(Z, A; \theta^*) \\
=&n\sum_{k=1}^{K^+} \alpha\left(n_k(z)/n\right) - n\sum_{a=1}^{K} \alpha\left(\sum_{k\in h^{-1}(a)} n_k(z)/n\right) \\
&+ \frac{1}{2}\sum_{k=1}^{K^+}\sum_{l=1}^{K^+}\left(O_{k,l}\log\frac{\hat{H}_{k,l}}{H^*_{h(k),h(l)}} + (n_{k,l} - O_{k,l})\log\frac{1-\hat{H}_{k,l}}{1-H^*_{h(k),h(l)}}\right), \qquad (4.46)
\end{aligned}$$

where $\hat{H}_{k,l} = O_{k,l}(z)/n_{k,l}(z)$. The first part of the expression is nonpositive since α is superadditive.

For $z \in \mathcal{N}_{K^+}$, $\hat{H}_{k,l} - H^*_{h(k),h(l)} = O_P(n^{-1}\rho_n^{1/2})$. Furthermore, the order is uniform since by (4.34),

$$\|X(z) - X(z_0)\|_\infty = o_P(1)$$

for any fixed $z_0 \in \mathcal{N}_{K^+}$, and all $z \in \mathcal{N}_{K^+}$, $z \notin \mathcal{S}(z_0)$. It follows by Taylor expansion that (4.46) is upper bounded by

$$\frac{1}{4}\sum_{k,l} n_{k,l}\frac{(\hat{H}_{k,l} - H^*_{h(k),h(l)})^2}{H^*_{h(k),h(l)}} + o_P(1) = O_P(1) \qquad (4.47)$$

uniformly for all $z \in \mathcal{N}_{K^+}$. The claim follows with Assumptions 4.2.9 and 4.2.10. Since (4.46) has $K^+(K^+ + 1)/2$ terms of order $O_P(1)$, it suffices to bound the model complexity term for $\sup_{\theta\in\Theta_{K^+}} \log g(A; \theta)$ by $\lambda \cdot K^+(K^+ + 1)/2 \cdot \log n$ for some constant λ.

□

Bibliography

[1] Edoardo M. Airoldi et al. "Mixed membership stochastic blockmodels". In: *J. Mach. Learn. Res.* 9 (2008), pp. 1981–2014.

[2] David Aldous and Persi Diaconis. "Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theorem". In: *Bull Am Math Soc* 36.4 (1999), pp. 413–432.

[3] Jeffrey D Allen et al. "Comparing statistical methods for constructing large scale gene networks". In: *PLoS One* 7.1 (2012), e29348.

[4] A. A. Amini et al. "Pseudo-likelihood methods for community detection in large sparse networks". In: *The Annals of Statistics* 41 (2013), pp. 2097–2122.

[5] T. W. Anderson. "Asymptotic Theory for Canonical Correlation Analysis". In: *Journal of Multivariate Analysis* 70.1 (1999), pp. 1–29.

[6] Richard Arratia, Andrew D Barbour, and Simon Tavaré. *Logarithmic combinatorial structures: a probabilistic approach*. Vol. 1. European Mathematical Society Zürich, 2003.

[7] Jinho Baik, Percy Deift, and Kurt Johansson. "On the distribution of the length of the longest increasing subsequence of random permutations". In: *J Am Math Soc* 12.4 (1999), pp. 1119–1178.

[8] Sourav Bandyopadhyay et al. "Rewiring of genetic networks in response to DNA damage". In: *Science* 330.6009 (2010), pp. 1385–1389.

[9] K. Basso et al. "Reverse engineering of regulatory networks in human B cells". In: *Nat Genet* 37 (2005), pp. 382–390.

[10] A. Ben-Dor, R. Shamir, and Z. Yakhini. "Clustering gene expression patterns". In: *Journal of Computational Biology* 6 (1999), pp. 281–297.

[11] P. Bickel and A. Chen. "A nonparametric view of network models and Newman-Girvan and other modularities". In: *Proceedings of the National Academy of Sciences* 106 (50 2009), pp. 21068–73.

[12] Peter Bickel et al. "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels". In: *Ann. Statist.* 41.4 (2013), pp. 1922–1943.

[13] Peter J Bickel and Purnamrita Sarkar. "Hypothesis testing for automated community detection in networks". In: *arXiv preprint arXiv:1311.2694* (2013).

[14] A. Celisse, J. J. Daudin, and L. Pierre. "Consistency of maximum-likelihood and variational estimators in the stochastic block model". In: *Electronic Journal of Statistics* 6 (2012), pp. 1847–1899.

[15] Maria Chahrour et al. "MeCP2, a key contributor to neurological disease, activates and represses transcription". In: *Science* 320.5880 (2008), pp. 1224–1229.

[16] A. Channarond, J.-J. Daudin, and S. Robin. "Classification and estimation in the Stochastic Block Model based on the empirical degrees". In: *Electronic Journal of Statistics* 6 (2012), pp. 2574–2601.

[17] Kehui Chen and Jing Lei. "Network cross-validation for determining the number of communities in network data". In: *arXiv preprint arXiv:1411.1715* (2014).

[18] Louis HY Chen. "Poisson approximation for dependent trials". In: *Ann Probab* (1975), pp. 534–545.

[19] Y. Cheng and G. M. Church. "Biclustering of expression data". In: *Proc Int Conf Intell Syst Mol Biol* 8 (2000), pp. 93–103.

[20] David S Choi, Patrick J Wolfe, and Edoardo M Airoldi. "Stochastic blockmodels with a growing number of classes". In: *Biometrika* (2012), asr053.

[21] The FANTOM Consortium. "A promoter-level mammalian expression atlas". In: *Nature* 507.7493 (2014), pp. 462–470.

[22] C. O. Daub et al. "Estimating mutual information using B-spline functions — an improved similarity measure for analysing gene expression data". In: *BMC Bioinformatics* 5 (2004), p. 118.

[23] S. Davletova et al. "Cytosolic ascorbate peroxidase 1 is acentral component of the reactive oxygen gene network of Arabidopsis". In: *The Plant Cell* 17 (2005), pp. 268–281.

[24] Aurelien Decelle et al. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". In: *Phys. Rev. E* 84.6 (2011), p. 066106.

[25] P. D'haeseleer, S. Liang, and R. Somogyi. "Genetic network inference: from co-expression clustering to reverse engineering". In: *Bioinformatics* 16 (2000), pp. 707–726.

[26] D. I. Edwards. *Introduction to Graphical Modelling*. 2nd. Springer, 2000.

[27] M. B. Eisen et al. "Cluster analysis and display of genome-wide expression patterns". In: *Proc Natl Acad Sci USA* 95 (1998), pp. 14863–14868.

[28] G. M. Estavillo et al. "Evidence for a SAL1-PAP chloroplast retrograde pathway that functions in drought and high light signalingin Arabidopsis". In: *The Plant Cell* 23 (2011), pp. 3992–4012.

[29] Peter M Fenwick. "A new data structure for cumulative frequency tables". In: *Software: Practice and Experience* 24.3 (1994), pp. 327–336.

[30] Donniell E Fishkind et al. "Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown". In: *SIAM J. Matrix Anal. Appl.* 34.1 (2013), pp. 23–39.

[31] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical Lasso". In: *Biostatistics* 9 (2007), pp. 432–441.

[32] Fang-Fang Fu and Hong-Wei Xue. "Coexpression analysis identifies Rice Starch Regulator1, a rice AP2/EREBP family transcription factor, as a novel rice starch biosynthesis regulator". In: *Plant Physiol* 154.2 (2010), pp. 927–938.

[33] Alberto de la Fuente et al. "Discovery of meaningful associations in genomic data using partial correlation coefficients". In: *Bioinformatics* 20.18 (2004).

[34] C. M. Gachon et al. "Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications". In: *Plant Molecular Biology* 58.2 (2005), pp. 229–245.

[35] R. Guimerà and L. A. N. Amaral. "Functional cartography of complex metabolic networks". In: *Nature* 433 (2005), pp. 895–900.

[36] Rukhsana Hasan et al. "The control of the yeast H2O2 response by the Msn2/4 transcription factors". In: *Mol Microbiol* 45.1 (2002), pp. 233–241.

[37] Wassily Hoeffding. "A non-parametric test of independence". In: *Ann Math Stat* (1948), pp. 546–557.

[38] P. W. Holland, K. B. Laskey, and S. Leinhardt. "Stochastic blockmodels: First steps". In: *Social Networks* 5 (2 1983), pp. 109–137.

[39] H Hotelling. "Relations between two sets of variates". In: *Biometrika* 28.3/4 (1936), pp. 321–377.

[40] F. Picard J.-J. Daudin and S. Robin. "A mixture model for random graphs". In: *Statistics and Computing* 18 (2 2008), pp. 173–183.

[41] A. K. Jain, M. N. Murty, and P. J. Flynn. "Data clustering: a review". In: *ACM Computing Surveys* 31 (1999), pp. 264–323.

[42] D. Jiang, C. Tang, and A. Zhang. "Cluster analysis for gene expression data: a survey". In: *IEEE Transactions on Knowledge and Data Engineering* 16 (2004), pp. 1370–1386.

[43] Yoo Jin Joo et al. "Cooperative regulation of ADE3 transcription by Gcn4p and Bas1p in Saccharomyces cerevisiae". In: *Eukaryot Cell* 8.8 (2009), pp. 1268–1277.

[44] Antony Joseph and Bin Yu. "Impact of regularization on Spectral Clustering". In: *arXiv preprint arXiv:1312.1733* (2013).

[45] Brian Karrer and M. E. J. Newman. "Stochastic blockmodels and community structure in networks". In: *Physical Review E* 83 (2011), p. 016107.

[46] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. New York: John Wiley and Sons, 2005.

[47] G. Kerr et al. "Techniques for clustering gene expression data". In: *Computers in Biology and Medicine* 38 (2008), pp. 283–293.

[48] A. Khandelwal et al. "Arabidopsis transcriptome reveals control circuits regulating redox homeostasis and the role of an AP2 transcription factor". In: *Plant Physiology* 148 (2008), pp. 2050–2058.

[49] K. Kim et al. "Using biologically interrelated experiments to identify pathway genes in Arabidopsis". In: *Bioinformatics* 28.6 (2012), pp. 815–822.

[50] Justin B. Kinney and Gunrinder S. Atwal. "Equitability, mutual information, and the maximal information coefficient". In: *Proc Natl Acad Sci USA* 10.1073/pnas.1309933111 (2014).

[51] Sapna Kumari et al. "Evaluation of gene association methods for coexpression network construction and biological knowledge discovery". In: *PLoS One* 7.11 (2012), e50411.

[52] P. Langfelder and S. Horvath. "Eigengene networks for studying relationships between co-expression modules". In: *BMC Systems Biology* 1 (2007), p. 54.

[53] P. Langfelder, B. Zhang, and S. Horvath. "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R". In: *Bioinformatics* 24 (2008), pp. 719–720.

[54] Pierre Latouche, Etienne Birmele, and Christophe Ambroise. "Variational Bayesian inference and complexity control for stochastic block models". In: *Stat. Modelling* 12.1 (2012), pp. 93–115.

[55] W. Lee et al. "Sparse canonical covariance analysis for high-throughput data". In: *Statistical Applications in Genetics and Molecular Biology* 10.1 (2011), pp. 1–24.

[56] Jing Lei. "A Goodness-of-fit Test for Stochastic Block Models". In: *arXiv preprint arXiv:1412.4857* (2014).

[57] Jure Leskovec and Julian J Mcauley. "Learning to discover social circles in ego networks". In: *Adv. Neural Inf. Process. Syst.* 2012, pp. 539–547.

[58] Ker-Chau. Li. "Genome-wide coexpression dynamics: Theory and application". In: *Proceedings of the National Academy of Sciences* 99 (2002), 1687516880.

[59] J. Lim et al. "A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration". In: *Cell* 125 (2006), pp. 801–814.

[60] Benjamin F Logan and Larry A Shepp. "A variational problem for random Young tableaux". In: *Adv Math* 26.2 (1977), pp. 206–222.

[61] E. Loreti et al. "A Genome-Wide Analysis of the Effects of Sucroseon Gene Expression in Arabidopsis Seedlingsunder Anoxia". In: *Plant Physiology* 137 (2005), pp. 1130–1138.

[62] S. C. Madeira and A. L. Oliveira. "Biclustering algorithms for biological data analysis: a survey". In: *IEEE/ACM Trans Comput Biol Bioinform* 1 (2004), pp. 24–45.

[63] Paul Magwene and Junhyong Kim. "Estimating genomic coexpression networks using first-order conditional independence". In: *Genome Biology* 5.12 (2004), R100.

[64] A. A. Margolin et al. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context". In: *BMC Bioinformatics* 7 (2006), S7.

[65] N. Meinshausen and P. Bühlmann. "High-dimensional graphs and variable selection with the Lasso". In: *The Annals of Statistics* 34 (2006), pp. 1049–1579.

[66] A Milatovich et al. "Gene for a tissue-specific transcriptional activator (EBF or Olf-1), expressed in early B lymphocytes, adipocytes, and olfactory neurons, is located on human chromosome 5, band q34, and proximal mouse chromosome 11". In: *Mamm Genome* 5.4 (1994), pp. 211–215.

[67] M. A. Naoumkina et al. "Genome-wide analysis of phenylpropanoid defence pathways". In: *Molecular Plant Pathology* 11.6 (2010), pp. 829–846.

[68] M. E. J. Newman. *Networks: An introduction*. Oxford University Press, 2010.

[69] Mark E. J. Newman. "Modularity and community structure in networks". In: *Proc. Natl. Acad. Sci. USA* 103.23 (2006), pp. 8577–8582.

[70] Mark EJ Newman. "Finding community structure in networks using the eigenvectors of matrices". In: *Phys. Rev. E* 74.3 (2006), p. 036104.

[71] Elena Parkhomenko, David Tritchler, and Joseph Beyene. "Sparse canonical correlation analysis with application to genomic data integration". In: *Statistical Applications in Genetics and Molecular Biology* 8.1 (2009), pp. 1–34.

[72] Tiago P Peixoto. "Parsimonious module inference in large networks". In: *Phys. Rev. Lett.* 110.14 (2013), p. 148701.

[73] J. Peng et al. "Partial correlation estimation by joint sparse regression models". In: *Journal of the American Statistical Association* 104 (2009), pp. 736–746.

[74] Ross G Pinsky. "Law of large numbers for increasing subsequences of random permutations". In: *Random Struct Algorithms* 29.3 (2006), pp. 277–295.

[75] A. Ramesh et al. "Clustering context-specific gene regulatory networks". In: *Pacific Symposium on Biocomputing*. 2010, pp. 444–455.

[76] Alfred Rényi. "On measures of dependence". In: *Acta Mathematica Hungarica* 10.3 (1959), pp. 441–451.

[77] David N. Reshef et al. "Detecting Novel Associations in Large Data Sets". In: *Science* 334 (2011), pp. 1518–1524.

[78] L. Rizhsky, H. Liang, and R Mittler. "The water-water cycle is essential forchloroplast protection in the absence of stress". In: *Journal of Biological Chemistry* 278 (2003), pp. 38921–38925.

[79] K. Rohe, S. Chatterjee, and B. Yu. "Spectral clustering and the high-dimensional stochastic block model". In: *Ann. Statist.* 39 (4 2011), pp. 1878–1915.

[80] Nathan Ross et al. "Fundamentals of Stein's method". In: *Prob. Surv* 8 (2011), pp. 210–293.

[81] Diego Franco Saldana, Yi Yu, and Yang Feng. "How Many Communities Are There?" In: *arXiv preprint arXiv:1412.1684* (2014).

[82] Juliane Schäfer and Korbinian Strimmer. "An empirical Bayes approach to inferring large-scale gene association networks". In: *Bioinformatics* 21.6 (2005), pp. 754–764.

[83] John P. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, 2000.

[84] R. Sharan, A. Maron-Katz, and R. Shamir. "CLICK and EXPANDER: a system for clustering and visualizing gene expression data". In: *Bioinformatics* 19 (2003), pp. 1787–1799.

[85] Noah Simon and Robert Tibshirani. "Comment on" Detecting Novel Associations In Large Data Sets" by Reshef Et Al, Science Dec 16, 2011". In: *arXiv preprint arXiv:1401.7645* (2014).

[86] M. Soler-López et al. "Interactome mapping suggests new mechanistic details underlying Alzheimer's disease". In: *Genome Research* 21 (2011), pp. 364–376.

[87] I. E. Sønderby, F. Geu-Flores, and B. A. Halkier. "Biosynthesis of glucosinolates–gene discovery and beyond". In: *Trends in Plant Science* 15.5 (2010), pp. 283–290.

[88] Lin Song, Peter Langfelder, and Steve Horvath. "Comparison of co-expression measures: mutual information, correlation, and model based indices". In: *BMC Bioinformatics* 13.1 (2012), p. 328.

[89] P. T. Spellman et al. "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization". In: *Mol Biol Cell* 9 (1998), pp. 3273–3297.

[90] Charles Stein. "Approximate computation of expectations". In: *Lecture Notes-Monograph Series* 7 (1986), pp. i–164.

[91] R. Steuer et al. "The mutual information: detecting and evaluating dependencies between variables". In: *Bioinformatics* 18 (2002), S231–240.

[92] J. M. Stuart et al. "A gene-coexpression network for global discovery of conserved genetic modules". In: *Science* 302 (2003), pp. 249–255.

[93] Gábor J Székely, Maria L Rizzo, et al. "Brownian distance covariance". In: *Ann App Stat* 3.4 (2009), pp. 1236–1265.

[94] L. P. Taylor and E. Grotewold. "Flavonoids as developmental regulators". In: *Current Opinion in Plant Biology* 8.3 (2005), pp. 317–323.

[95] S. L. Teng and H. Huang. "A statistical framework to infer functional gene relationships from biologically interrelated microarray experiments". In: *Journal of the American Statistical Association* 104.486 (2009), pp. 465–473.

[96] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, Fourth Edition*. 4th. Academic Press, 2005.

[97] R. Verkerk et al. "Glucosinolates in Brassica vegetables: the influence of the food supply chain on intake, bioavailability and human health". In: *Molecular Nutrition and Food Research* 53.Suppl. 2 (2009), S219.

[98] Sandra Waaijenborg, Philip Verselewe de Witt Hamer, and Aeilko H Zwinderman. "Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis". In: *Statistical Applications in Genetics and Molecular Biology* 7.1 (2008), pp. 1–43.

[99] YX Wang and Peter J Bickel. "Likelihood-based model selection for stochastic block models". In: *arXiv preprint arXiv:1502.02069* (2015).

[100] YX Rachel Wang, Michael S Waterman, and Haiyan Huang. "Gene coexpression measures in large heterogeneous samples using count statistics". In: *Proceedings of the National Academy of Sciences* 111.46 (2014), pp. 16371–16376.

[101] YX Rachel Wang et al. "Inferring gene-gene interactions and functional modules using sparse canonical correlation analysis". In: *Annals of Applied Statistics* to appear (2015).

[102] Joe H. Ward. "Hierarchical grouping to optimize an objective function". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244.

[103] A. Wille et al. "Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana". In: *Genome Biology* 5.11 (2004), pp. 1–13.

[104] Anja Wille and Peter Bühlmann. "Low-order conditional independence graphs for inferring genetic networks". In: *Statistical Applications in Genetics and Molecular Biology* 5.1 (2006), p. 1.

[105] Daniela M. Witten and Robert Tibshirani. "Extensions of sparse canonical correlation analysis with applications to genomic data". In: *Statistical Applications in Genetics and Molecular Biology* 8.1 (2009), pp. 1–27.

[106] Daniela M. Witten, Robert Tibshirani, and Trevor Hastie. "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis". In: *Biostatistics* 10.3 (2009), pp. 515–534.

[107] C. J. Wolfe, I. S. Kohane, and A. J. Butte. "Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks". In: *BMC Bioinformatics* 6 (2005), p. 227.

[108] Patrick J Wolfe and Sofia C Olhede. "Nonparametric graphon estimation". In: *arXiv preprint arXiv:1309.5936* (2013).

[109] H. H. Woo, B. R. Jeong, and M. C. Hawes. "Flavonoids: from cell cycle regulation to biotechnology". In: *Biotechnology Letters* 27.6 (2005), pp. 365–374.

[110] X. Yan and S. Chen. "Regulation of plant glucosinolate metabolism". In: *Planta* 226.6 (2007), pp. 1343–1352.

[111] Yang Yang et al. "Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types". In: *Nat Commun* 5 (2014).

[112] B. Zhang and S. Horvath. "A general framework for weighted gene co-expression network analysis". In: *Stat Appl Genet Mol Biol* 4 (2005), p. 17.

[113] Fang Zhao et al. "Inhibition of p300/CBP by early B-cell factor". In: *Mol Cell Biol* 23.11 (2003), pp. 3837–3846.

[114] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. "Community extraction for social networks". In: *Proc. Natl. Acad. Sci. USA* 108.18 (2011), pp. 7321–7326.

[115] Jiashun Zheng et al. "Epistatic relationships reveal the functional organization of yeast transcription factors". In: *Mol Syst Biol* 6.1 (2010).

[116] S. Zhou et al. "High-dimensional covariance estimation based on Gaussian graphical models". In: *Journal of Machine Learning Research* 12 (2011), pp. 2975–3026.

[117] Xianghong Zhou, Ming-Chih J Kao, and Wing Hung Wong. "Transitive functional annotation by shortest-path analysis of gene expression data". In: *Proc Natl Acad Sci USA* 99.20 (2002), pp. 12783–12788.

Appendix A

Supplementary information for Chapter 3

A.1 Sensitivity analysis

Simulation

Tables A.1 and A.2 show the average *precision* and *recall* calculated for 10 simulation datasets, reflecting the sensitivity of the procedure to different parameter choices as outlined in Figure 1 of the paper. At the step of building the edge weight matrix, different choices of the penalty parameter $\boldsymbol{\lambda}$ and the subsampling level are tested. When applying HC, the dendrogram is cut when clusters of size less than n_{min} start to appear as the number of clusters increases, and a number of n_{min} values are tested. When fitting SBMs, different discretization levels and cluster number Q are compared. We omit the results using SBM for the two-pathway case since SBM does not perform well with multiple signal groups as discussed in the paper.

Table A.1: Classification performance of the procedure under different parameter settings for datasets with $p = 300$, one pathway group and 33% experimental dependency.

	HC, $n_{min} = 15$									
	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(3,3)	0.771	0.4	0.828	0.333	0.742	0.44	0.746	0.467	0.713	0.387
(3,6)	0.78	0.44	0.817	0.473	0.78	0.467	0.81	0.4	0.793	0.447
(6,6)	0.832	0.413	0.862	0.44	0.792	0.42	0.751	0.42	0.774	0.413
(6,9)	0.88	0.42	0.847	0.447	0.86	0.467	0.837	0.453	0.787	0.48
(9,9)	0.891	0.407	0.868	0.447	0.857	0.473	0.849	0.507	0.812	0.587
(9,12)	0.879	0.487	0.938	0.473	0.888	0.527	0.86	0.553	0.869	0.533
(12,12)	0.879	0.373	0.915	0.56	0.939	0.447	0.799	0.56	0.777	0.56
(12,15)	0.861	0.453	0.875	0.44	0.807	0.493	0.725	0.553	0.751	0.52
(15,15)	0.891	0.487	0.93	0.413	0.903	0.52	0.75	0.607	0.763	0.54
	HC, $n_{min} = 20$									
	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(3,3)	0.771	0.4	0.782	0.42	0.742	0.44	0.711	0.513	0.679	0.467
(3,6)	0.77	0.493	0.73	0.54	0.78	0.467	0.782	0.427	0.756	0.473
(6,6)	0.824	0.433	0.794	0.5	0.704	0.54	0.722	0.447	0.76	0.453
(6,9)	0.88	0.42	0.834	0.493	0.804	0.54	0.824	0.5	0.769	0.507

Continued on next page

Table A.1 – *Continued from previous page*

	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(9,9)	0.793	0.52	0.845	0.473	0.857	0.473	0.824	0.553	0.812	0.587
(9,12)	0.861	0.487	0.88	0.5	0.877	0.547	0.86	0.553	0.828	0.547
(12,12)	0.876	0.427	0.915	0.56	0.939	0.447	0.799	0.56	0.777	0.56
(12,15)	0.861	0.453	0.875	0.44	0.807	0.493	0.725	0.553	0.751	0.52
(15,15)	0.891	0.487	0.93	0.413	0.866	0.567	0.703	0.607	0.681	0.6
	HC, $n_{min} = 25$									
	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(3,3)	0.771	0.4	0.782	0.42	0.742	0.44	0.711	0.513	0.616	0.52
(3,6)	0.77	0.493	0.73	0.54	0.769	0.533	0.696	0.58	0.718	0.513
(6,6)	0.813	0.507	0.794	0.5	0.704	0.54	0.722	0.447	0.721	0.547
(6,9)	0.814	0.487	0.814	0.513	0.804	0.54	0.824	0.5	0.769	0.507
(9,9)	0.793	0.52	0.819	0.52	0.823	0.493	0.824	0.553	0.812	0.587
(9,12)	0.827	0.507	0.88	0.5	0.877	0.547	0.86	0.553	0.828	0.547
(12,12)	0.764	0.533	0.859	0.56	0.75	0.547	0.799	0.56	0.741	0.58
(12,15)	0.861	0.453	0.875	0.44	0.737	0.54	0.683	0.593	0.751	0.52
(15,15)	0.885	0.507	0.884	0.473	0.708	0.613	0.643	0.607	0.681	0.6
	SBM, discretization level = 0.4, $Q = 2$									
	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(3,3)	0.619	0.593	0.705	0.607	0.658	0.593	0.531	0.593	0.524	0.593
(3,6)	0.802	0.513	0.78	0.54	0.718	0.567	0.702	0.553	0.605	0.647
(6,6)	0.81	0.54	0.797	0.6	0.767	0.587	0.717	0.593	0.696	0.64
(6,9)	0.868	0.5	0.807	0.593	0.775	0.567	0.719	0.6	0.693	0.593
(9,9)	0.868	0.473	0.794	0.6	0.789	0.567	0.76	0.56	0.743	0.573
(9,12)	0.641	0.62	0.72	0.593	0.743	0.587	0.768	0.567	0.743	0.573
(12,12)	0.599	0.607	0.652	0.6	0.605	0.62	0.566	0.627	0.558	0.607
(12,15)	0.581	0.6	0.559	0.62	0.499	0.567	0.55	0.54	0.47	0.527

Continued on next page

Table A.1 – *Continued from previous page*

	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(15,15)	0.619	0.573	0.65	0.567	0.619	0.633	0.586	0.647	0.505	0.66
	SBM, discretization level = 0.4, $Q = 3$									
	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(3,3)	0.566	0.527	0.714	0.447	0.771	0.393	0.707	0.393	0.677	0.373
(3,6)	0.842	0.433	0.767	0.44	0.798	0.36	0.745	0.387	0.69	0.387
(6,6)	0.771	0.447	0.802	0.4	0.866	0.407	0.729	0.4	0.757	0.407
(6,9)	0.874	0.447	0.813	0.44	0.813	0.387	0.691	0.393	0.652	0.393
(9,9)	0.888	0.427	0.83	0.42	0.794	0.4	0.705	0.413	0.694	0.447
(9,12)	0.687	0.493	0.654	0.433	0.71	0.44	0.807	0.387	0.694	0.487
(12,12)	0.507	0.573	0.628	0.48	0.538	0.493	0.59	0.433	0.642	0.473
(12,15)	0.631	0.487	0.574	0.48	0.543	0.473	0.489	0.44	0.711	0.393
(15,15)	0.693	0.407	0.694	0.373	0.548	0.46	0.471	0.447	0.593	0.46
	SBM, discretization level = 0.6, $Q = 2$									
	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(3,3)	0.886	0.287	0.886	0.4	0.779	0.493	0.826	0.4	0.798	0.427
(3,6)	0.919	0.3	0.902	0.34	0.89	0.353	0.863	0.353	0.855	0.393
(6,6)	0.745	0.427	0.903	0.347	0.905	0.333	0.87	0.36	0.843	0.373
(6,9)	0.955	0.313	0.853	0.367	0.855	0.447	0.859	0.373	0.876	0.353
(9,9)	0.797	0.46	0.943	0.373	0.836	0.42	0.841	0.413	0.89	0.347
(9,12)	0.948	0.333	0.969	0.347	0.839	0.393	0.831	0.453	0.935	0.373
(12,12)	0.745	0.473	0.791	0.427	0.884	0.367	0.837	0.413	0.811	0.46
(12,15)	0.772	0.353	0.731	0.4	0.802	0.38	0.815	0.407	0.792	0.42
(15,15)	0.638	0.58	0.53	0.573	0.681	0.487	0.707	0.44	0.803	0.407
	SBM, discretization level = 0.6, $Q = 3$									
	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall

Continued on next page

Table A.1 – *Continued from previous page*

	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(3,3)	0.868	0.287	0.925	0.22	0.825	0.32	0.931	0.28	0.814	0.28
(3,6)	0.925	0.273	0.842	0.34	0.881	0.313	0.9	0.273	0.814	0.307
(6,6)	0.773	0.367	0.9	0.267	0.955	0.3	0.918	0.267	0.814	0.26
(6,9)	0.946	0.233	0.933	0.287	0.907	0.26	0.929	0.26	0.814	0.32
(9,9)	0.875	0.32	0.95	0.287	0.844	0.347	0.835	0.327	0.814	0.32
(9,12)	0.95	0.327	0.969	0.267	0.819	0.34	0.855	0.333	0.814	0.34
(12,12)	0.835	0.307	0.893	0.26	0.866	0.287	0.859	0.253	0.814	0.32
(12,15)	0.732	0.373	0.71	0.32	0.802	0.313	0.821	0.3	0.814	0.353
(15,15)	0.725	0.493	0.615	0.447	0.68	0.42	0.729	0.353	0.814	0.327

Table A.2: Classification performance of the procedure under different parameter settings for datasets with $p = 300$, two pathway groups and 0% experimental dependency.

	HC, $n_{min} = 15$									
	Pathway 1									
	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(6,6)	0.819	0.353	0.838	0.427	0.647	0.48	0.731	0.46	0.732	0.427
(6,9)	0.893	0.407	0.823	0.44	0.828	0.433	0.82	0.46	0.689	0.447
(9,9)	0.906	0.413	0.832	0.453	0.777	0.427	0.831	0.447	0.757	0.44
(9,12)	0.872	0.42	0.879	0.433	0.853	0.407	0.829	0.46	0.713	0.433
(12,12)	0.936	0.467	0.968	0.433	0.942	0.407	0.867	0.467	0.698	0.433
(12,15)	0.989	0.387	0.971	0.427	0.946	0.44	0.778	0.46	0.578	0.513
(15,15)	0.99	0.327	0.98	0.36	0.872	0.447	0.814	0.467	0.648	0.473
(15,18)	0.918	0.34	0.902	0.32	0.841	0.46	0.556	0.48	0.545	0.44
(18,18)	0.943	0.28	0.91	0.333	0.783	0.347	0.618	0.367	0.541	0.387
	Pathway 2									
$\boldsymbol{\lambda}$	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	

Continued on next page

Table A.2 – *Continued from previous page*

	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(6,6)	0.799	0.273	0.763	0.333	0.631	0.367	0.541	0.393	0.487	0.38
(6,9)	0.824	0.413	0.705	0.353	0.698	0.427	0.564	0.36	0.608	0.413
(9,9)	0.863	0.493	0.927	0.393	0.724	0.42	0.66	0.407	0.598	0.42
(9,12)	0.923	0.36	0.968	0.353	0.833	0.38	0.688	0.447	0.612	0.433
(12,12)	0.95	0.387	0.919	0.353	0.883	0.42	0.805	0.433	0.602	0.467
(12,15)	0.944	0.387	0.905	0.373	0.881	0.427	0.804	0.527	0.556	0.54
(15,15)	0.88	0.407	0.818	0.38	0.803	0.453	0.702	0.473	0.49	0.593
(15,18)	0.884	0.353	0.77	0.373	0.657	0.42	0.527	0.487	0.404	0.46
(18,18)	0.95	0.287	0.98	0.307	0.69	0.433	0.484	0.453	0.447	0.4
	HC, $n_{min} = 20$									
	Pathway 1									
	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(6,9)	0.825	0.447	0.823	0.473	0.825	0.447	0.82	0.46	0.67	0.453
(9,9)	0.906	0.413	0.828	0.467	0.777	0.427	0.831	0.447	0.719	0.44
(9,12)	0.868	0.44	0.864	0.46	0.838	0.447	0.829	0.46	0.713	0.433
(12,12)	0.936	0.467	0.968	0.447	0.923	0.42	0.867	0.467	0.698	0.433
(12,15)	0.989	0.387	0.951	0.447	0.917	0.46	0.778	0.46	0.578	0.513
(15,15)	0.964	0.327	0.98	0.36	0.872	0.447	0.814	0.467	0.648	0.473
(15,18)	0.878	0.373	0.868	0.333	0.841	0.46	0.556	0.48	0.545	0.44
(18,18)	0.943	0.28	0.876	0.353	0.783	0.38	0.618	0.367	0.541	0.387
	Pathway 2									
(6,6)	0.79	0.287	0.763	0.333	0.631	0.367	0.541	0.393	0.487	0.38
(6,9)	0.794	0.42	0.705	0.353	0.698	0.427	0.555	0.347	0.608	0.413
(9,9)	0.863	0.493	0.904	0.42	0.692	0.44	0.66	0.407	0.598	0.42
(9,12)	0.869	0.453	0.907	0.467	0.79	0.427	0.685	0.487	0.612	0.433
(12,12)	0.95	0.387	0.89	0.42	0.846	0.48	0.801	0.427	0.602	0.467
(12,15)	0.846	0.42	0.847	0.433	0.843	0.513	0.804	0.527	0.556	0.54

Continued on next page

Table A.2 – *Continued from previous page*

	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(15,15)	0.861	0.433	0.719	0.393	0.803	0.453	0.702	0.473	0.49	0.593
(15,18)	0.813	0.393	0.731	0.447	0.63	0.453	0.527	0.487	0.404	0.46
(18,18)	0.93	0.287	0.923	0.307	0.69	0.433	0.484	0.453	0.447	0.4
	HC, $n_{min} = 25$									
	Pathway 1									
	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
$\boldsymbol{\lambda}$	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
(6,6)	0.795	0.413	0.711	0.447	0.647	0.48	0.662	0.46	0.732	0.427
(6,9)	0.803	0.473	0.823	0.473	0.788	0.447	0.82	0.46	0.665	0.44
(9,9)	0.879	0.433	0.828	0.467	0.743	0.427	0.831	0.447	0.719	0.44
(9,12)	0.868	0.44	0.797	0.48	0.788	0.447	0.829	0.46	0.713	0.433
(12,12)	0.886	0.467	0.868	0.533	0.873	0.447	0.867	0.467	0.698	0.433
(12,15)	0.938	0.4	0.951	0.447	0.857	0.46	0.778	0.46	0.578	0.513
(15,15)	0.887	0.413	0.938	0.36	0.872	0.447	0.814	0.467	0.648	0.473
(15,18)	0.854	0.373	0.868	0.333	0.841	0.46	0.556	0.48	0.545	0.44
(18,18)	0.869	0.367	0.872	0.373	0.783	0.38	0.618	0.367	0.541	0.387
	Pathway 2									
$\boldsymbol{\lambda}$	subsample 50%		subsample 60%		subsample 70%		subsample 80%		subsample 90%	
(6,6)	0.778	0.287	0.746	0.34	0.591	0.36	0.541	0.393	0.487	0.38
(6,9)	0.778	0.427	0.677	0.427	0.67	0.433	0.555	0.347	0.608	0.413
(9,9)	0.847	0.533	0.848	0.427	0.654	0.493	0.66	0.407	0.598	0.42
(9,12)	0.781	0.54	0.846	0.513	0.752	0.433	0.685	0.487	0.612	0.433
(12,12)	0.876	0.493	0.822	0.533	0.777	0.48	0.801	0.427	0.602	0.467
(12,15)	0.755	0.447	0.758	0.507	0.81	0.513	0.761	0.527	0.556	0.54
(15,15)	0.704	0.56	0.62	0.38	0.803	0.453	0.702	0.473	0.49	0.593
(15,18)	0.71	0.433	0.725	0.467	0.63	0.453	0.527	0.487	0.404	0.46
(18,18)	0.93	0.287	0.89	0.333	0.69	0.433	0.484	0.453	0.447	0.4

Real data

Table A.3 shows the Jaccard coefficients between the groups identified during the first search stage of the Arabidopsis data and other groups produced using different choices of parameters. The Jaccard coefficient between two sets X and Y is defined as

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}, \tag{A.1}$$

which measures the degree of similarity between the two sets. As described in the paper, the first search stage corresponds to choosing $\boldsymbol{\lambda}$ from the grid $\{90, 100, 110\}^2$ and a total of four groups (Groups 1, 5, 6 and 13) were identified by our procedure. The sensitivity of the results are tested against different choices of the penalty parameter $\boldsymbol{\lambda}$ and the subsampling level. When applying HC, the cutoff level was increased incrementally until at least five clusters of size less than n_{min} appeared. We note here that n_{min} can range from 20 to 35 without altering the results in the table.

Table A.3: The Jaccard coefficients between the final groups identified and other groups produced under different parameter choices for the Arabidopsis data.

	subsample 60%				subsample 70%				subsample 80%			
$\boldsymbol{\lambda}$	Group 1	Group 13	Group 5	Group 6	Group 1	Group 13	Group 5	Group 6	Group 1	Group 13	Group 5	Group 6
(90,90)	1	0.9	0.636	0.462	1	1	0.682	0.667	0.533	1	0.7	0.667
(90,100)	0.933	0.5	0.625	0.75	1	0.9	0.722	0.75	0.533	1	0.778	0.6
(90,110)	0.6	0.455	0.438	0.75	1	1	0.824	0.857	1	1	0.778	0.444
(100,100)	0.667	0.9	0.789	0.556	1	0.9	0.875	0.857	1	1	0.733	0.75
(100,110)	1	0.9	0.5	0.667	0.733	0.5	0.824	0.714	0.767	1	0.824	0.667
(110,110)	1	1	0.409	0.5	0.467	0.9	0.65	0.545	0.767	1	0.812	0.556